DEDICATION

of the

REMOTE SENSING TUTORIAL

To The *Memory* of

SPACE SHUTTLE CHALLENGER (STS 51) AND COLUMBIA (STS 107) ASTRONAUTS

WHO HAVE GIVEN THEIR LIVES TO UPHOLD A NASA TRADITION:

SPACE - FOR THE BENEFIT OF ALL MANKIND

Space is an exciting place to visit and work in! Unmanned satellites are operating there all the time. But the presence of humans in space has a special allure. Since the first flights of cosmonauts (initially, Russians) and astronauts (initially, Americans), select crews of men and women have orbited the Earth and a few have landed on the Moon. This is exciting - still not routine - but unfortunately dangerous. Of the 100+ missions of the Space Transport System (STS), two of these Space Shuttles have met with disaster, with loss of all onboard. This Tutorial is dedicated with pride to those 14 individuals making up their crews.

The first disaster was the loss of the Shuttle Challenger (STS-51), which exploded (because of ice in the fuel system) upon takeoff on January 28, 1986, as seen in this photo:

The seven astronauts onboard are shown in this group picture:

The STS-51 Crew:

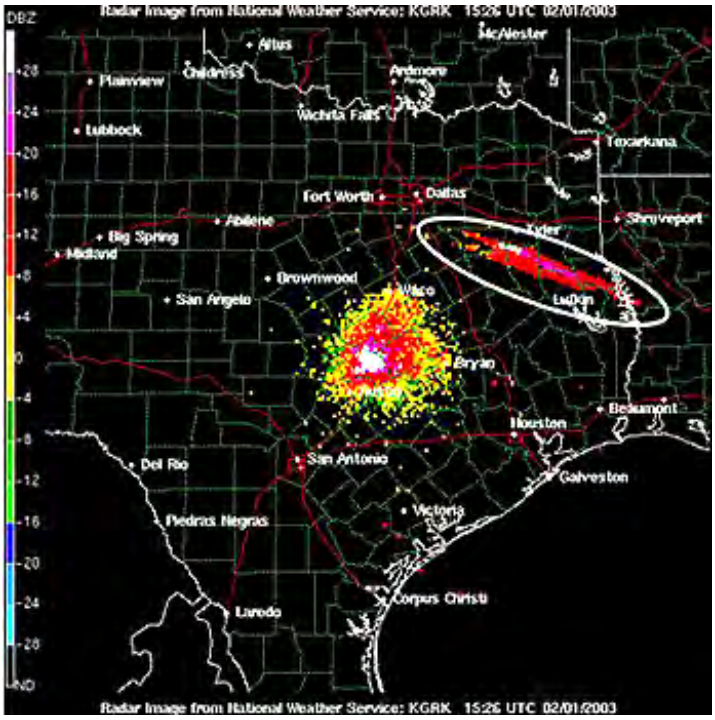
Back Row (Left to Right): ELLISON ONIZUKA, CHRISTA McCAULLIFE, GREGORY JARVIS, JUDITH RESNIK

Front Row (L to R): MICHAEL SMITH, FRANCIS SCOBEE, RONALD McNAIR

Now to the second catastrophe: Columbia was the first ever Space Shuttle to go into space, on April 12, 1981, as seen in liftoff here:

Tragically, Columbia was damaged during launch by a piece of outer covering from one of the auxiliary tanks that had broken loose and struck the Shuttle's front left edge. Upon re-entry at the end of the mission, the damaged edge allowed hot thin air to enter the wing and spread its destructive effects. This began as the spaceship passed over the West Coast of the United States. Columbia's last moments in space started at 9:00 AM EST as it began its final break up over central Texas:

Remote Sensing played a role in monitoring this tragic event. Below is a weather radar image of the stretched out debris and smoke from the exploded Shuttle Columbia extending from north-central Texas to western Louisiana. The strong signals from the Dallas-Fort Worth area are "ground clutter" caused by radar reflections from buildings.



Here is a pre-flight photo of the STS-107 crew:

The Crew From Left to Right:

Seated: RICK HUSBAND, KALPANA CHAWLA, WILLIAM McCOOL

Standing: DAVID BROWN, LAUREL CLARK, MICHAEL ANDERSON, AND ILAN RAMON

One can surmise that this thought would be the Columbia crew's earnest wish if they could communicate to us from the hereafter:

LIKE THE PHOENIX OF ANCIENT LORE, LET MANNED SPACEFLIGHT RISE AGAIN AND ASTRONAUTS CONTINUE TO EXPLORE THE SKIES OF EARTH AND THE PLANETS.

THIS HAS HAPPENED TWO AND A HALF YEARS LATER WITH THE SUCCESSFUL LAUNCH ON JULY 26, 2005 WITH

THE DISCOVERY SHUTTLE, FOLLOWING A LONG PERIOD IN WHICH SAFETY BECAME THE FOREMOST ISSUE. THIS MISSION (STS-108) IS FOCUSED ON TESTING NEW SAFETY FEATURES AS WELL AS SUPPLYING THE INTERNATIONAL SPACE STATION.

FOREWORD

By William Campbell

Throughout the years, NASA's Earth Sciences program has primarily focused on providing high quality data products to its science community. NASA also recognizes the need to increase its involvement with the general public, including areas of information and education. Many different Earth-sensing satellites, with diverse sensors mounted on sophisticated platforms, are in Earth orbit or soon to be launched. These sensors are designed to cover a wide range of the electromagnetic spectrum and are generating enormous amounts of data that must be processed, stored, and made available to the user community.

This rich source of unique, repetitive, global coverage produces valuable data and information for applications as diverse as forest fire monitoring and grassland inventory in Mongolia, early typhoon warning over the vast Pacific Ocean, flood assessment in coastal zones around the Bay of Bengal, and crop health and growth within the plains of the United States. Additionally, the commercial realm is also developing and launching various high resolution satellites and marketing these data worldwide.

The Applied Information Sciences Branch at NASA's Goddard Space Flight Center is heavily involved in technology outreach and transfer. As part of this activity, we recognized the need for a highly intuitive, easily accessible remote sensing tutorial that hopefully will serve as a primer for the new user as well as a teaching tool for the educational community. We wanted this Tutorial to provide a detailed understanding of the utility of the data in light of the fundamental principles of electromagnetic energy, especially as they relate to sensor design and function.

Enter Dr. Nicholas Short, a former NASA Goddard employee and author/editor of four NASA-sponsored books (Mission to Earth: Landsat Views the World; The Landsat Tutorial Workbook; The HCMM Anthology; and Geomorphology from Space) germane to the subject of remote sensing. We asked Nick if he would be willing to put his significant experience and talents to work to present an updated and expanded version of his past efforts. The result is this Internet website and a CD-ROM (also tied to the Internet) entitled "The Remote Sensing Tutorial". As the CD/Net versions progressed, we were joined by the Air Force Academy as a co-sponsor, followed by GST and then Goddard's EOS program in supporting the later phases of the project. We trust you will find the Tutorial informative and useful, and when you are done, please pass it on to a colleague or friend. Remember, think globally and act locally.

William J. Campbell (now retired
Head/Code 935
Applied Information Sciences Branch
NASA/Goddard Space Flight Center
Greenbelt, Maryland 20771

PREVIOUS NEXT

This first page includes various notices and comments on the origin of the Tutorial, its overall nature and contents, and suggestions for optimum use.

THE REMOTE SENSING TUTORIAL

A SURVEY OF THE SPACE AGE THAT CONSIDERS WHAT WE CAN LEARN USING SENSORS OPERATED FROM SATELLITES AND SPACECRAFT WHICH LOOK INWARD AT THE EARTH - ITS LAND, OCEANS, AND ATMOSPHERE - AND OUTWARD AT THE PLANETS, THE GALAXIES AND, GOING BACK IN TIME, THE ENTIRE UNIVERSE - THE COSMOS

PRIME DEVELOPER AND WRITER:

DR. NICHOLAS M. SHORT

(See Biographies near end of second Overview page)

THIS TUTORIAL SERVES AS A COMPREHENSIVE SURVEY OF THE SPACE PROGRAM OF THE UNITED STATES, AND WILL ALSO TREAT PROGRAMS SUPPORTED BY CANADA, EUROPE, RUSSIA, CHINA, JAPAN, INDIA AND OTHER COUNTRIES.

Before entering this Overview, ponder this slogan:

REMOTE SENSING is the BACKBONE of the SPACE PROGRAM

SO, JUST WHAT IS REMOTE SENSING? SIMPLEST ANSWER:

REMOTE SENSING REFERS TO TECHNIQUES THAT USE SENSORS TO ACQUIRE DATA/INFORMATION ABOUT OBJECTS NOT IN CONTACT WITH (THUS, DISTANT FROM) THE VIEWING INSTRUMENT

BULLETIN

THE TUTORIAL HAS NOW BE DECLARED COMPLETE AS OF JANUARY 1, 2011.

THIS MAY BE THE FINAL VERSION!

TAKE SPECIAL NOTICE OF THIS

I GET MANY REQUESTS FOR PERMISSION TO USE ILLUSTRATIONS FROM THE TUTORIAL, OR ACTUAL COPIES THEREOF. ABOUT 95% OF THOSE YOU SEE WERE EXTRACTED OFF THE INTERNET (AND I HAVE INEXACT RECORDS OF THEIR SOURCE); THE OTHER 5% CAME FROM MY PERSONAL COLLECTION WHICH I HAVE SINCE GIVEN AWAY. I AM THEREFORE UNABLE TO FILL ANY REQUESTS INCLUDING PERMISSIONS, SO PLEASE DON'T SEND ANY. AND BE AWARE THAT I (NMS) AM IN THE LATE STAGES OF DIABETES WITH SEVERE COMPLICATIONS AND MAY NOT EVEN BE AROUND IF YOU TRY TO REACH ME. BUT TRY ANYWAY (nmshort2@verizon.net), IF YOU HAVE QUESTIONS ABOUT CONTENT OR WISH TO POINT OUT ANY ERRORS. THANK YOU!

WELCOME TO THIS TUTORIAL, a training manual for learning the role of Remote Sensing - that aspect of space science and technology that relies mainly on sensors on satellites and mounted in telescopes to monitor Earth, other planetary bodies and distant stars and galaxies. (It also serves to review the so-called space programs [U.S. and international] in general and the history of space exploration in particular.) Observing the Earth will be the main focus of the Tutorial as this has the most obvious payoff for mankind. But while reaching to the edge of the Solar System and ultimately much farther out to the edge of the Universe seems mostly "academic", we shall try to demonstrate why, in the long run, those extraterrestrial endeavors that depend on remote sensing may make the greatest contributions to useful knowledge of value to humankind's future.

This Overview has several purposes: 1) To describe the contents of the entire Tutorial with suggestions on best ways to utilize it (whether accessed as a Web Site or from a CD-ROM, 2) To

synopsize the basic concepts underlying remote sensing; 3) To provide a brief synopsis of the history and uses of Remote Sensing (especially as carried out from orbiting satellites and deep space probes, and 4) To look especially at the major advances in remote sensing over the last 20 years. Elsewhere on this page, we will expand on the following thesis which is the prime reason for the importance of Remote Sensing and the raison d'etre for this Tutorial: ***Remote Sensing is the technology that is now the principal modus operandi (tool) by which (as targets or objects of surveillance) the Earth's surface and atmosphere, the planets, and the entire Universe are being observed, measured, and interpreted from such vantage points as the terrestrial surface, earth-orbit, and outer space****. The main Overview ends with a quick look at the latest products now being acquired by commercial remote sensing satellites. At the bottom of the second page are biographies and credits appropriate to the contributors. We strongly recommend that you read through this entire Overview, which may strike you as a hodgepodge of diverse topics and facts, since it will serve as a proper introduction to both the Tutorial and to the many practical ways in which Remote Sensing and allied fields contribute to gathering information about the many topics of interest to be examined in the Sections that follow.*

NOTICE: **SINCE ITS INCEPTION, THE TUTORIAL HAS BEEN CONSTRUCTED FOR SCREEN DISPLAY AT 800 BY 600 PIXELS. IN RECENT YEARS, AN INCREASING FRACTION OF THOSE WHO ACCESS IT HAVE SET THEIR DISPLAY AT HIGHER RESOLUTION. THE RESULT IS THAT THE ILLUSTRATIONS, WHICH HAD BEEN PROPERLY SIZED AT THE LOWER RESOLUTION TO FIT MUCH OF THE SCREEN WIDTH, BECOME NOTABLY SMALLER (OFTEN MAKING WORDING UNREADABLE). IF YOU HAVE A HIGHER RESOLUTION DISPLAY AND THE SIZE DECREASE IS A HINDRANCE TO USE, WE SUGGEST THAT YOU RESET RESOLUTION TO THE 800 BY 600 PIXELS LEVEL.**

SECOND NOTICE: THIS TUTORIAL WAS CONSTRUCTED USING WINDOWS XP BUT THE WRITER RECENTLY HAD ACCESS TO THE CD VERSION ON AN APPLE COMPUTER. I FOUND THAT ABOUT ONE IN EVERY TEN ILLUSTRATIONS FAILED TO COME ON SCREEN PROPERLY. IF YOU, THE READER, ARE USING APPLE, THE SAME RESULT MAY OCCUR.

ADDITIONAL NOTICE :

BECAUSE OF A MIX-UP IN RECORD KEEPING, MANY OF THE IMAGES, PHOTOS, AND ILLUSTRATIONS IN THE TUTORIAL THAT ARE NOT IN THE PUBLIC DOMAIN MAY NOT BE CREDITED, OR IF SO, ARE NOT PROPERLY CREDITED. IF YOU ARE THE SOURCE OF ANY SUCH ILLUSTRATIONS AND YOU WISH TO HAVE YOUR DESIRED CREDIT (NAME, ORGANIZATION, ETC.) APPLIED TO THE IMAGE(S), OR YOU CHOOSE NOT TO HAVE THE ILLUSTRATION(S) USED IN THIS TUTORIAL, PLEASE NOTIFY THE WRITER, NICHOLAS M. SHORT, AT THE EMAIL ADDRESS GIVEN NEAR THE BOTTOM OF THIS PAGE. SEE ALSO THE *WHAT'S NEW* PAGE IN THE 'FRONT' FOLDER.

OVERVIEW: HOW TO USE THIS TUTORIAL AND A SYNOPSIS OF THE PRINCIPLES AND APPLICATIONS OF REMOTE SENSING.

Before proceeding into the main subject of the Overview - what remote sensing is and does - we ask you to read through the next few paragraphs which provide a background into how the Tutorial came to be, how you should use it, and how it fits into a computer format. We start with 7 Notes that are informative comments.

NOTE 1: Most of the pages in the Tutorial will have a Summary for each page, bounded by blue lines, near the top.

NOTE 2: The Tutorial has been prepared for online display using the HomeSite *html* marker program; it is designed to run on the MS Internet Explorer browser. For some, the balance between text and illustration size may be best at the monitor screen setting of 600 by 800 pixels, used by the writer [NMS] to prepare the text. Also, see NOTICES above.

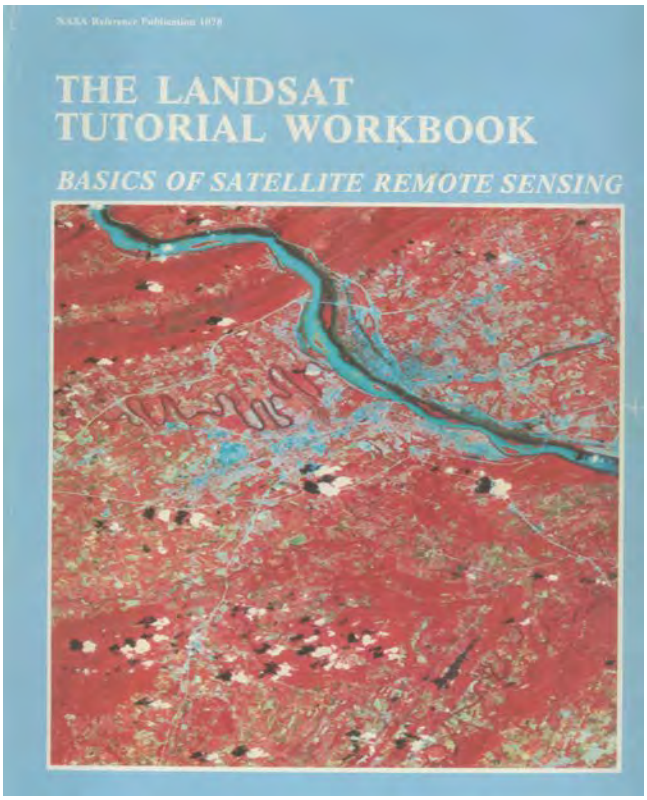
NOTE 3: There are many internal links in the Tutorial: These are cross-references that go to other pages and are indicated by blue-highlighted words such as "page #-#". They are for the most part intended to go to one specific page, on which (somewhere) is the particular image or text referred to in the starting page. To return to the original page, simply click on your browser BACK button. Similar are external links to pertinent Internet sites (regrettably, some sites may now be defunct).

NOTE 4: Some images in the Tutorial appear degraded. Many of these were downloaded off the Internet and have lost quality when reprocessed for use in the Tutorial. Others were copies from photos or other sources on the writer's scanner, and are thus also blurred somewhat.

NOTE 5: In keeping with scientific convention and the intended worldwide use of this Tutorial, we normally specify measurements in metric system units (SI), especially those for the electromagnetic spectrum and other units in physics. We will place English unit equivalents in parentheses where appropriate or to clarify, particularly when dealing with geographic parameters.)

SPECIAL NOTICE: This Tutorial was begun in 1995. Its initial content was much less than what now fills the final version. In the first few years the Tutorial was mainly an update of the principal writer's (N.M. Short) *Landsat Tutorial Workbook* (NASA SP-1078; now out of print), and other NASA publications in which he was one of the

authors. Illustrations were taken primarily from his personal collection. Over the next 14+ years a great deal of new information was gleaned from various sources - mainly from the Internet. IN FACT, this entire Tutorial has been expanded largely from Internet sources, so it is fair to say that the document is "new age" in that it owes much of its content to online material - IT HAS THUS EVOLVED INTO A TRULY INTERNET-DEPENDENT DOCUMENT. Over the years the Tutorial, has grown like "topsy" and may strike some readers as somewhat disjointed. Also, the Tutorial owes its now massive size to another influence: Originally, the intent was just to produce an updated version of the Landsat Tutorial Workbook; then, it evolved into a more general treatise on all aspects of remote sensing; but as it grew, it was decided to add material that moves beyond the realm of remote sensing per se, as for example, the subsections that provide "short courses" on Geology, Meteorology, Biology, and Astronomy, and the visual tour of the U.S. and other parts of the world that makes up Section 6. In sum, this expanded Web Tutorial has become a multipurpose book.



NOTE: Each image throughout this Tutorial will have a caption that is accessed simply by placing your mouse anywhere on the image; if that doesn't work try the lower right portion of the image.

THE REMOTE SENSING TUTORIAL (occasionally cited as RST) initially was sponsored by the now defunct Applied Information Science Branch (Code 935) at NASA's Goddard Space Flight Center; thanks is expressed to its Branch Chief, William Campbell, for the support that has made the RST viable in its early years. For a time the RST was also underwritten by the U.S. Air Force Academy, which contributed the text for Appendix A. For the past several years it was *ex officio* sponsored by the Earth Observation Systems (EOS) program at Goddard. Since 2005 the host for the Tutorial has been the Landsat Program Science Office.

At your convenience, please take time to visit two pages accessed from the buttons above. The first is a very important Dedication and Foreword. Then, read through the WHAT'S NEW text accessed by the button at the top right of this long page. That text notes that the Tutorial now includes links to several video "movies" that those (with higher speed access) can visit to learn about a variety of topics. The instructions are on this WHAT'S NEW page. That page also contains a notice about an on-going problem with image source accreditation.

As you work through these pages, you will see how users such as yourself can apply remote sensing (a term defined in connection with Question O-1, then further below, and again at the beginning of the Introduction Section). But for

the moment try this phrase as a working definition (a "mouthful"): "*The use of electronic sensors (or photography) to acquire data related to the electromagnetic spectrum in formats that can be analyzed numerically and/or that result in images (pictures) of objects/scenes located in the field of view, all of which can then be applied to the study of the land, sea, air and biotic communities that comprise Earth's environments, as well as the principal means for obtaining a deep understanding of the vital role remote sensing plays in exploring the planets and observing the stars and galaxies well out into the Cosmos*." Not only will you gain insight into past uses of aerial photography and space imagery, but you should develop skills in interpreting these visual displays and data sets by direct inspection and by computer processing.

The Tutorial has been developed for certain groups as the primary users: Faculty and students at the college level; Science teachers at the High School level; gifted or interested students mainly from the 8-12 grade levels; professionals in many fields where remote sensing comes into play, who need insights into what this technology can do for them; that segment of the educated general public who are curious about or intrigued with the many accomplishments of the space program that have utilized remote sensing from satellites, space stations, and interplanetary probes to monitor and understand surface features and processes on Earth and other bodies in the Solar System and beyond. (Most members of these user groups who access this very long Tutorial through the Internet are likely to be on fast-download lines and hence can retrieve individual pages [which can have 30 or more illustrations] rapidly enough for easy and efficient display.)

The central aim of The Remote Sensing Tutorial is to familiarize, and in so doing instruct, you as to what

remote sensing is, what its applications are, and what you need to know in order to interpret and, hopefully, use the data/information being acquired by satellite, air, and ground sensors. We try to accomplish this by presenting a very large number of remote sensing products as images which are described in a running text that explains their characteristics and utility. This Internet means of delivery of the Tutorial is thus image intensive. The abundance of pictorials becomes the principal learning device rather than the more customary dependence on textual description, supported by photographs, found in most pedagogical textbooks. The old adage that "a picture is worth a thousand words" holds especially true in remote sensing because it can convey, when accompanied by a brief textual commentary, a great deal about how remote sensing is done and the methodology/rationale by which information is gleaned from a pictorial product. The RST Internet format has one obvious advantage over standard textbooks - it can use literally hundreds of color photos and images and thus is not limited to the few permitted in those books because of cost constraints.

The Internet is a prime source for information on almost every aspect of remote sensing. Many sites offer good overviews of satellite remote sensing. One that has recently appeared, and provides an excellent synopsis of the main principles and applications, has been constructed by the Canadian Center for Remote Sensing. Click here if you want to view it now, or at your leisure. Another of special merit is the Remote Sensing Core Curriculum, project which highlights a new educational approach now under development. A comprehensive Tutorial written by S.C. Liew of the University of Singapore is worth a visit. A somewhat briefer review has been prepared by Harrison and Jupp. Several other sites that have some helpful tutorial information are: University of Colorado website; NOAA Web site; Towson St. Univ. site; OhioView site. A recent NASA-supported initiative in curriculum

development is described at the Geospatial Information Technology website of the University of Mississippi.

Another site that emphasizes remote sensing and imagery is the Eduspace program sponsored by the European Space Agency (ESA). The site can be accessed by clicking on Eduspace links. Be advised that to get into some of the features at this site, you must be able to register as a member of a teaching institution - primary through college. It has an abundance and variety of links, many of which are worth exploring at some stage in your use of this Tutorial.

For a broad perspective on how remote sensing has flourished in the last 30 years, one needs only to check out the still growing number of U.S. and International organizations - government, university, and private - that are largely concerned with various facets of remote sensing. A listing of most of these is found at this site: CCRS For those who might wish to build or expand their knowledge and background in several sciences that are relevant to remote sensing, we strongly recommend exploring the PSIGate site maintained by the University of Manchester (England) that has many useful links in Astronomy, Earth Science, and Physics.

Because of its size and the many illustrations, the Tutorial can be treated almost as a textbook. It is hoped that some teachers, especially at the college level, will elect to use the Tutorial either as a bona fide text or as a supplement. One pedagogical tool in the learning process is repetition. This tautology is deliberate: the same information in more detail, or even a repeat of an illustration, represents a) a reminder, b) a clarification, and/or c) a expansion of the ideas inherent to the information. Various topics throughout the Tutorial will follow this kind of hierarchy: 1) they appear briefly in this Overview; 2) they may be treated again, in more detail in the Introduction; 3) they may

reappear in various Sections of the Tutorial; and 4) in some instances, they may warrant elevation to a level requiring a full Section to explore. Cross-referencing by links helps to establish relationships and continuity.

One singular characteristic of the Remote Sensing Tutorial is the inclusion within the continuing text of each Section (not at the end of a chapter as is the case in most textbooks) of a series of thought or interpretive questions. The answers are accessed by clicking on the blue ANSWER button; give it a try before you look. There will normally be 10 to 40+ questions per Section. This Overview has a get-acquainted short Quiz consisting of only a half dozen questions pertaining to a set of images; its purpose is to help you decide whether you want to "get involved" in the learning experience afforded by the remainder of the Tutorial by showing you what image analysis and interpretation is all about and that your general background knowledge is probably sufficient for you to succeed in this process. There are also two "Exams" (at the close of Section 1 and Section 21) and a scene identification Quiz within Section 6 that challenges you to conduct remote sensing interpretations on images from two adjacent areas in central Pennsylvania. Lets introduce you to the type of questions to expect by asking this one right now.

O-1: Most people, even those with a good post high school education, when asked what the term "remote sensing" means to them, don't have the remotest idea. So, what do you think remote sensing is all about? Try to make up a simple definition. Then, list (mentally, or on paper) five practical applications of remote sensing as you defined it.
ANSWER

Now, following the above instructions, announcements and provisos, on to the main topics making up this

Overview.

WHAT'S IN THE TUTORIAL

In the past 30 or so years, remote sensing has become a full-fledged discipline with thousands now holding jobs related to its use/applications. Almost all universities now offer one to several courses in the field, along with related courses such as Geographic Information Systems. The Remote Sensing Tutorial is itself almost a complete course. It serves as its own textbook. But for the learner, it may only be a supplement to other sources of information.

Here is a list of well-known textbooks that detail most of the fundamentals and applications of Earth Remote Sensing:

- **Avery, T.E. and Berlin, G.L., Fundamentals of Remote Sensing and Airphoto Interpretation, 6th Ed., 2001, MacMillan Publ. Co., 472 pp.**
- **Campbell, J.B., Introduction to Remote Sensing,3rd Ed., 2002, The Guilford Press**
- **Drury, S.A., Image Interpretation in Geology, 2nd Ed., 1993, Chapman & Hall, 243 pp.**
- **Drury, S.A., A Guide to Remote Sensing: Interpreting Images of the Earth, 1998, Oxford University Press.**
- **Drury, S.A., Images of the Earth: A Guide to Remote Sensing, 2nd Ed., 1998, Oxford University Press, 212 pp.**
- **Jensen, J.R., Introductory Digital Image Processing: A Remote Sensing Perspective, 2nd Ed., 1996, Prentice-Hall.**
- **Jensen, J.R, Remote Sensing of the Environment, 2000, Prentice-Hall**
- **Kramer, H., Observation of the Earth and its Environment: Survey of Missions and Sensors, 4th Ed., 2002, Springer-Verlag**
- **Kuehn, F. (Editor), Introductory Remote Sensing Principles and Concepts, 2000, Routledge, 215 pp.**

- **Lillesand, T.M. Kiefer, R.W., and Chipman, J.W., Remote Sensing and Image Interpretation, 5th Ed., 2004, J. Wiley & Sons, 720 pp.**
- **Rees, W.G., Physical Principles of Remote Sensing, 2nd Ed., 2001, Cambridge University Press.**
- **Richards, J.A. and Jia, X, Remote Sensing Digital Image Analysis: an Introduction, 1999, Springer-Verlag**
- **Sabins, Jr., F.F., Remote Sensing: Principles and Interpretation. 3rd Ed., 1996, W.H. Freeman & Co., 496 pp.**
- **Siegal, B.S. and Gillespie, A.R., Remote Sensing in Geology, 1980, J. Wiley& Sons (especially Chapters 1 through 11)**
- **Swain, P.H. and Davis, S.M., Remote Sensing - the Quantitative Approach, 1978, McGraw-Hill Book Co.**

After this list was compiled, in mid-2007 the writer (NMS) received a copy of the following remarkable book, especially noteworthy for its numerous color illustrations: Aronson, S., Remote Sensing for GIS Managers, 2005, ESRI Press, Redlands, Calif.

An excellent blending of remote sensing imagery, ground photos, maps, and other types of geographic information is found in the *Atlas of North America: A Space Portrait of a Continent*, published by the National Geographic Society (1986). The NGS has since published a world atlas using space imagery.

Also of value are these Periodicals devoted largely to remote sensing methods and applications:

- **Canadian Journal of Remote Sensing**
- **Earth Observation Magazine**
- **Geocarto International**
- **IEEE Transactions on Geoscience and Remote Sensing.**
- **International Journal of Remote Sensing.**
- **Photogrammetric Engineering and Remote Sensing.**

- **Remote Sensing of the Environment**

Other sources of basic information about remote sensing are these books that involved the writer (NMS): the still relevant 1982 NASA Publication RP 1078: The LANDSAT TUTORIAL WORKBOOK; NASA SP-360: MISSION TO PLANET EARTH: LANDSAT VIEWS THE WORLD, co-authored with Paul D. Lowman, Jr, Stanley C. Freden, and William C. Finch, Jr (now out-of-print but in some libaries); NASA SP-465: THE HCMM ANTHOLOGY; and NASA SP-486 and GEOMORPHOLOGY FROM SPACE (co-authored with Robert Blair, Jr).

To expand upon the remarks at the beginning of the Overview, one prime purpose of this Tutorial is to be a learning resource for college students, as well as for individuals now in the work force who require indoctrination in the basics of space-centered remote sensing. In both instances the objective is to offer a background that will actually be useful in current or eventual job performance to those who may need to provide input information obtainable from remote sensing into day-to-day operations. We also think the Tutorial can be an invaluable resource for pre-college (mostly Secondary School) teachers who want to build a background in the essential contributions of the space program to society so as to better teach their students (many of whom should also be capable of working through the main ideas in the Tutorial). Our hope is that this survey of Satellite Remote Sensing will attract and inspire a few individuals from the world community who might consider a specialized career in this field or in the broader fields allied with Earth System Science (ESS) and the Environment (see below). An additional goal is to interest and inform the general public about the principles and achievements of remote sensing, with emphasis on demonstrated applications.

A solid way to appreciate how the RST goes about meeting these goals is to skim through its Table of Contents.

TABLE OF CONTENTS

Foreword

Overview of this Remote Sensing Tutorial; "Getting Acquainted" Quiz

Introduction to Remote Sensing: Technical and Historical Perspectives; Special Applications such as Geophysical Satellites, Military Surveillance, and Medical Imaging

Section:

1. Image Processing and Interpretation: Morro Bay, California; First Exam

2. Geologic Applications: Stratigraphy; Structure; Landforms

3. Vegetation Applications: Agriculture; Forestry; Ecology

4. Urban and Land Use Applications

5. Mineral and Oil Resource Exploration:

6. Flight Across the United States: Boston to San Francisco; Quiz; World Tour

7. Regional Studies: Use of Mosaics from Landsat

8. Radar and Microwave Remote Sensing

9. The Warm Earth: Thermal Remote Sensing

10. Aerial Photography as Primary and Ancillary Data

Sources

11. The Earth's Surface in 3-Dimensions: Stereo Systems and Topographic Mapping

12. The Human Remote Senser in Space: Astronaut Photography

13. Collecting Data at the Surface: Ground Truth; the "Multi" Concept; Hyperspectral Remote Sensing

14. The Water Planet: Meteorological, Oceanographic and Hydrologic Remote Sensing

15. Geographic Information Systems: The GIS Approach to Decision Making

16. Earth Systems Science; Earth Science Enterprise; and the EOS Program

17. Use of Remote Sensing in Basic Science Studies I: Mega-Geomorphology

18. Basic Science Studies II: Impact Cratering

19. Planetary Remote Sensing: The Exploration of Extraterrestrial Bodies

20. Cosmology: Remote Sensing Systems that provide observations on the Content, Origin, and Development of the Universe

21. Remote Sensing in the 21st Century

Appendix A: Modern History of Space

Appendix B: Interactive Image Processing

Appendix C: Principal Components Analysis

Appendix D: Glossary

This list does not tell the whole story. There are several mini-tutorials dealing with specific fields of Science that are embedded in individual Sections. Thus, Section 2 has a two-page survey of the principles of Geology, Section 14 includes a primer on Meteorology; Section 19 covers Planetology; Section 20 presents principles of Astronomy; Section 20 also has a page devoted to fundamentals of Biology; some of the key ideas of Physics and Chemistry are scattered throughout the Tutorial.

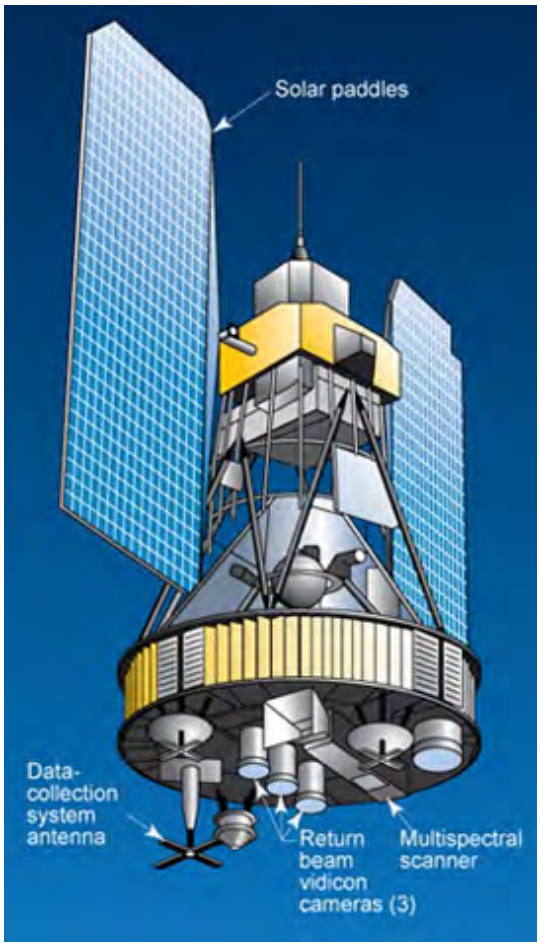
Unlike a formal course in the subject, with chapters covering principles, techniques and applications in a pedagogic and systematic way, we lead you through a series of Sections focused on one to several relevant themes and topics. Because we can represent most remote sensing data as visuals, we will our organize our instructional treatment around *illustrations*, such as space images, classifications, maps, and plots, rather than numerical data sets. These data sets are the real knowledge base for application scientists in putting this information to practical use. (Much of this material has been acquired by direct downloading off the Internet. We are grateful to the source organizations and individuals but, for the most part, we do not acknowledge each contribution per se.) Descriptions and discussions accompany these illustrations to aid in interpreting the visual concepts. "Standard" space images, particularly those from Landsat sensors, are usually the focal points of a Section, but we frequently add special computer processed renditions with ground photos that depict features in a scene and descriptive maps where appropriate.

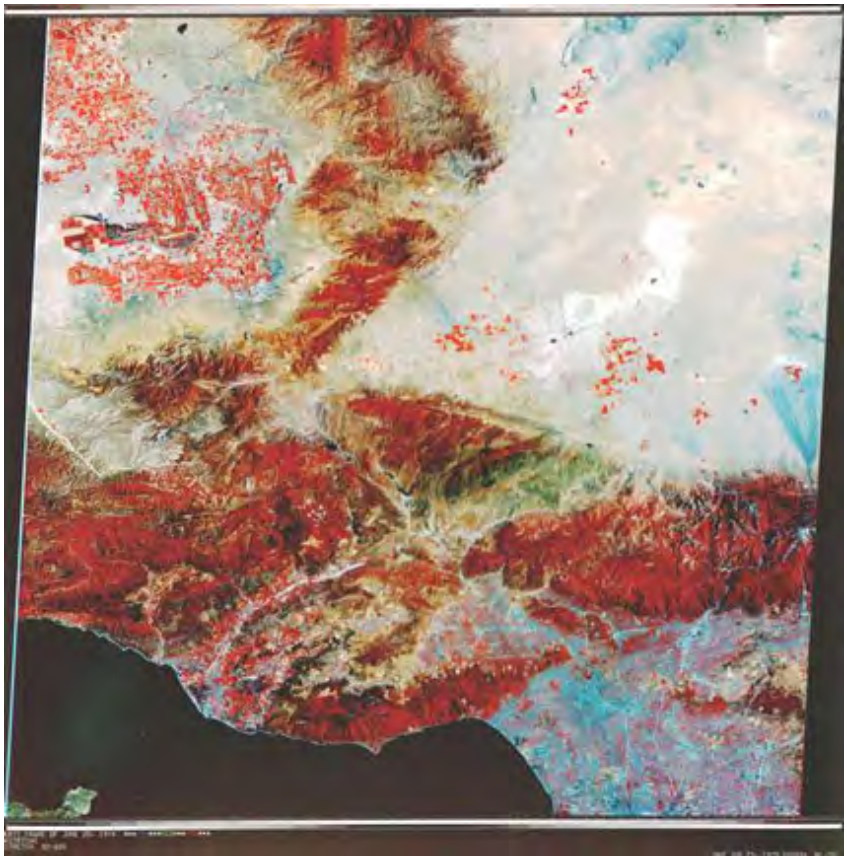
We also call out numerous links to other remote sensing sources and to various continuing or planned programs. Some of these programs are federal or international

programs such as ESE, whereas, others are programs from educational or commercial organizations that provide training and services. These links, in turn, have their own sets of links, which, as you explore them, will broaden your acquaintance with the many facets of remote sensing and its popular applications.

Perusal through the Introduction and Sections 1, 8 and 9 is the minimum effort we suggest if you want to master the basics. Section 6, which visually tours the U.S. and the rest of the world using space imagery, is worth a visit just for fun.

The Tutorial begins with an Introduction, which covers the principles of physics (especially electromagnetic radiation) underlying remote sensing, then considers the main kinds of observing platforms, and includes the history of satellite systems, with a focus on Landsat. Many of the subsequent Sections and topics center on Landsat because it continues to be a kingpin among the current remote sensing systems. Because of its pre-eminence, and also because it was the main satellite system on which the writer (NMS) worked in his 21 years at NASA's Goddard Space Flight Center (GSFC), let us take a moment now to show the satellite (originally named ERTS) and a typical color product made, in this illustration, from Landsat-1 Multispectral Scanner input:

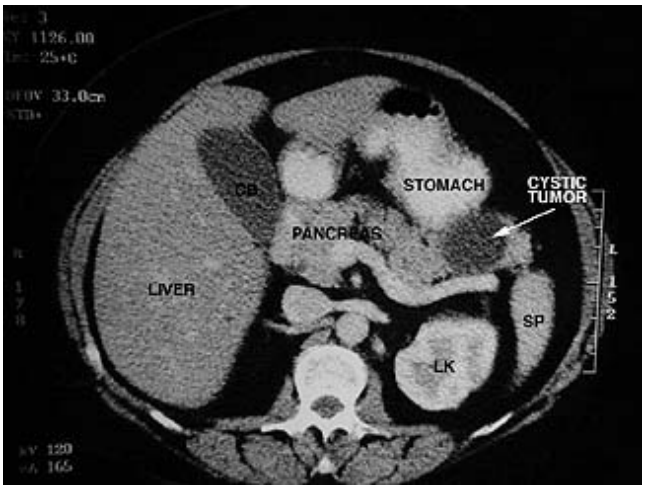




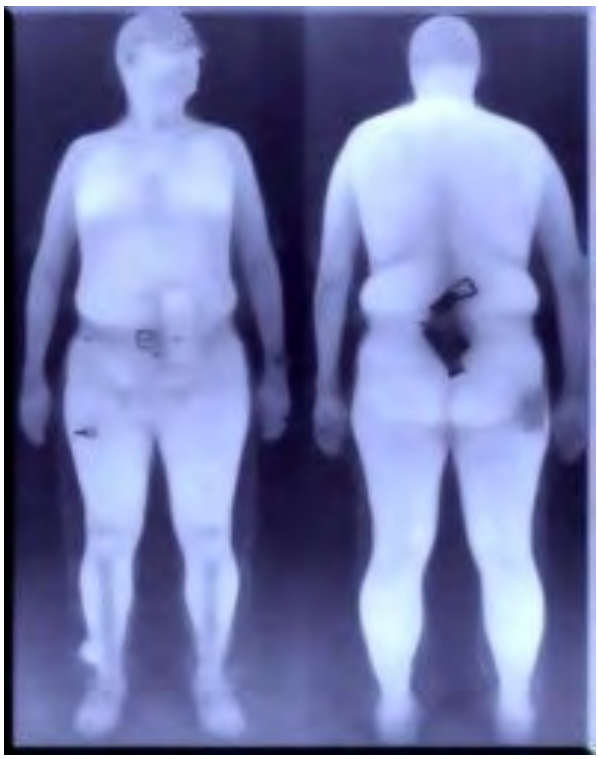
The Introduction also delves into three special topics: Use of satellites for geophysical measurements of Earth's force fields; a survey of satellite programs (military and

security agencies) employed in monitor activities detrimental to a country's safety (these are often called "spy satellites), and the applications of intruments and techniques within the purview of remote sensing that are used in medical diagnosis.

This last topic may seem a bit strange as part of this Tutorial, which deals almost entirely with remote sensing data from satellites and spacecraft that look inwardly at Earth and outward at the heavens. But, medical remote sensing (or "medical imaging") has been around for 100 years. For most people, use of medical instruments that examine the bodies of humans and their pets by means of electromagnetic radiation or force fields is the application of remote sensing of greatest personal familiarity and value in their lives. We treat this subject in three review pages in the Introduction. For now, let's just look at two examples of the sensing of the human body using X-rays. The first image shows an x-ray radiograph of a diseased lung; the second is a CAT Scan (CAT = Computer Assisted Tomography) slice through the midsection of a torso showing the labelled organs:



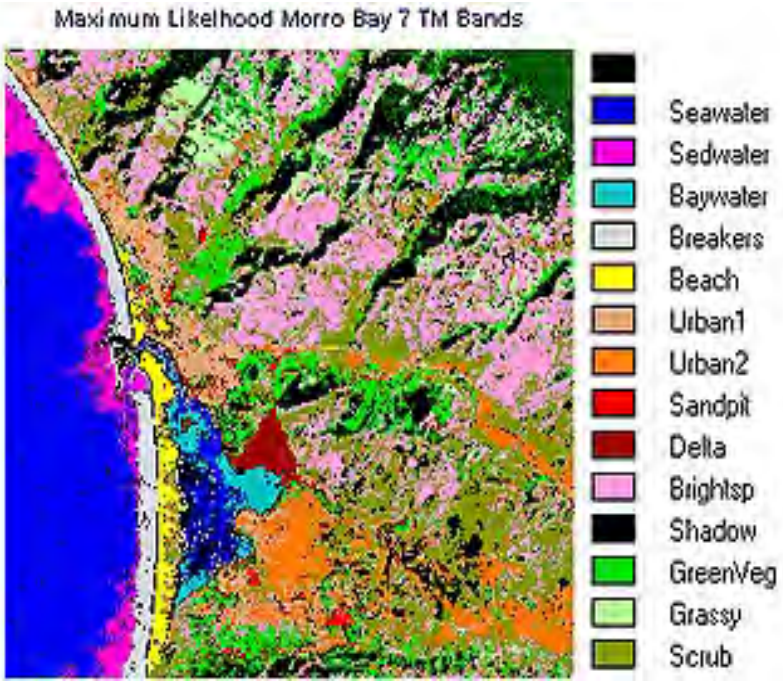
While we are on this subject of the human body, it is appropriate to mention one of the newest applications of the common mode of remote sensing that uses electromagnetic radiation. This is the "full body scanner" now in almost all (larger) airports that "searches" passengers for hidden weapons and bombs - a major tool in the continuing battle against terrorists. An image produced by the scanner appears here; the subject is revisited in the Introduction:



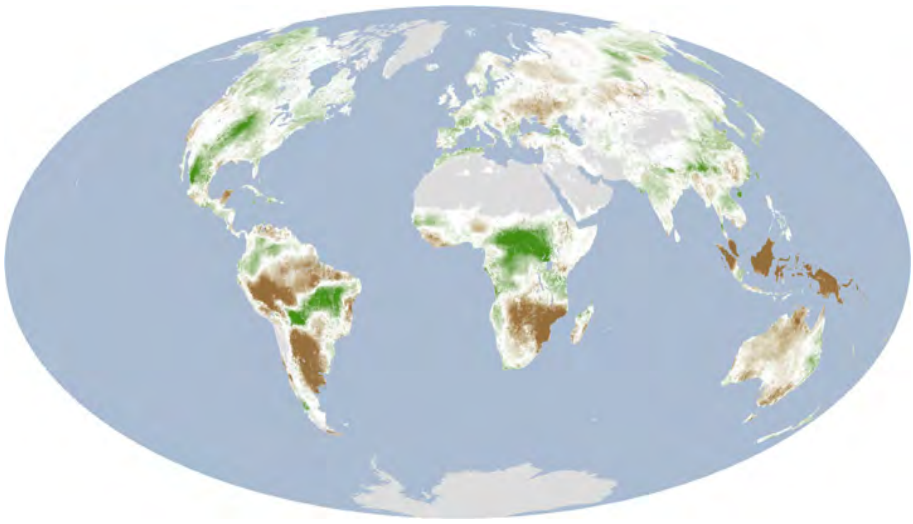
Section 1 is one of the key chapters in this Tutorial because we try to introduce most of the major concepts of image analysis and interpretation by walking you through the product types and processing outputs in common use,

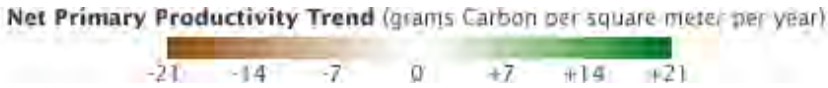
using a single subscene as the focus. That subscene is a Landsat Thematic Mapper (TM) image of Morro Bay, California. (Landsat refers to the 6 spacecraft that became the "workhorse" remote sensing system flown in space since 1972.) This is what it looks like in a false color rendition:

Images such as this are readily analyzed and interpreted by computer-based processing programs. One ultimate goal in image processing is to produce a classification map of the identifiable features or classes of land cover in a scene. In Section 1 we examine various ways of enhancing a scene's appearance and end with a *supervised classification* of the surface features we choose as meaningful to our intended use. Here is the classification of Morro Bay:



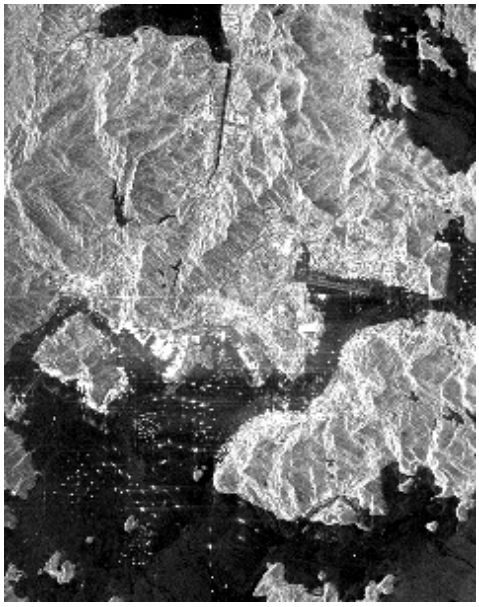
Sections 2 through 5 deal with major applications of remote sensing in Geology, Vegetation Monitoring, and Urban Development. As an example, consider the map below. It shows on a global basis the variations during the first decade of the 21st Century of general vegetation changes, both trees and grasses, in terms of Net Primary Productivity. Surprisingly, despite definitive global warming during that decade, which should lead to increases in vegetation, in the southern hemisphere large areas experienced decreases. This has implications for the amount of carbon dioxide introduced into the atmosphere which, in turn, affects warming trends. The data used in the study came from the Terra satellite's MODIS sensor.





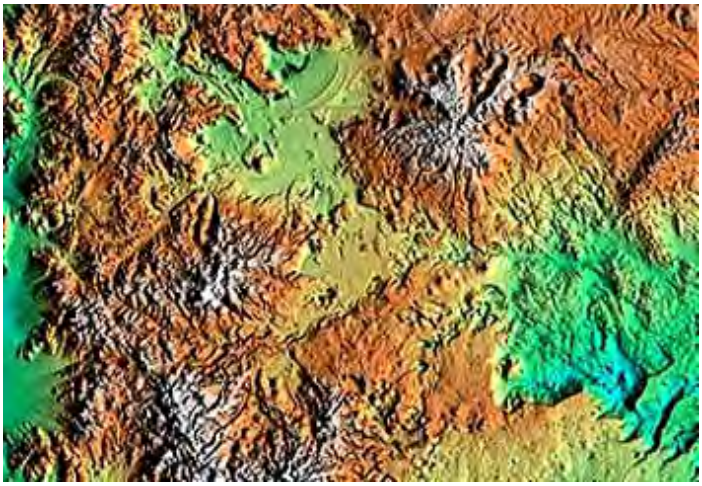
Section 6, although not essential to understanding the principles of the several types of remote sensing, deserves your attention simply because it covers a topic of general interest: familiar places associated with geographic regions and natural/manmade features in the U.S. and the rest of the World. It's a bit like a "travelogue" that takes you on an excursion first across the United States and then to a variety of locales in all continents (including the Antarctic). And it has this meritorious attribute: It can be a most effective way for high school and even college teachers involved in teaching Geography to use images and maps to describe various countries and regions of the world.

Section 8 is concerned with another mode of remote sensing, the use of radar and passive microwave. Seasat was the first civilian spacecraft that was dedicated to radar imaging. Radar has been flown several times on the U.S. Space Shuttle. This X-band Synthetic Aperture Radar (SAR) image (SIR-C mission) of Hong Kong is typical of this type of imagery:

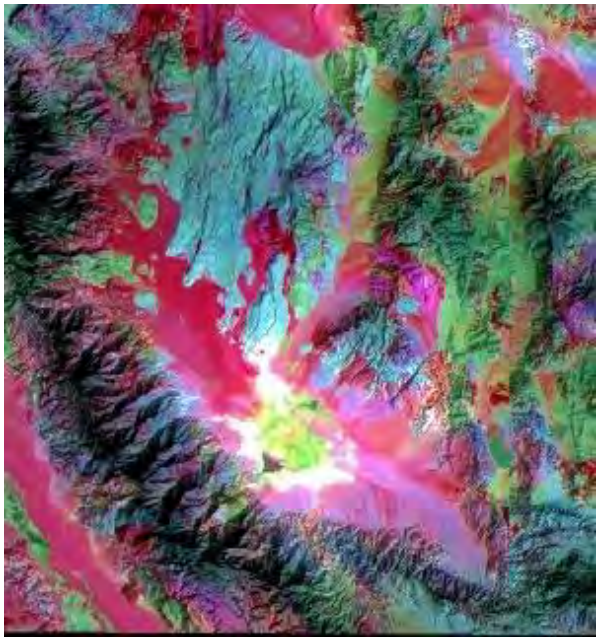


A later Shuttle flight - the Shuttle Radar Topography Mission (SRTM) - acquired both C-band and X-band

images; these were utilized in calculating topographic altitudes. This SRTM image of Patagonia, Chile is assigned colors that correspond to ranges in altitude:



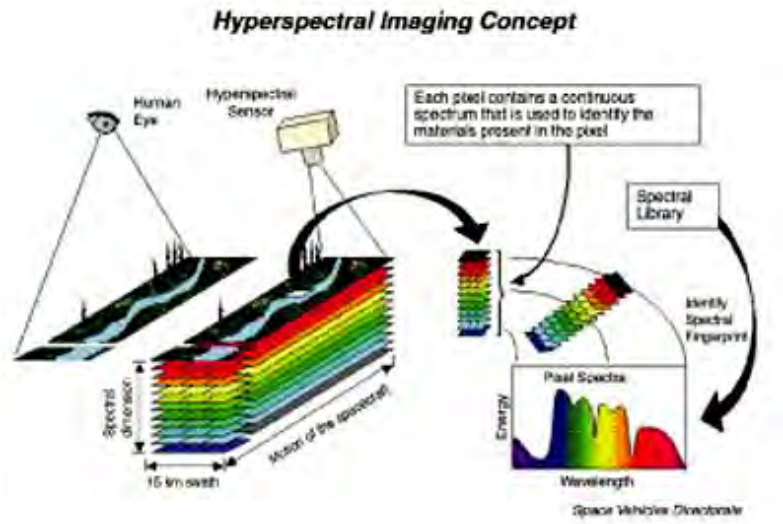
Section 9 focuses on the increasing use of thermal infrared imagery obtained both from aircraft- and spacecraft-mounted sensors that operate mainly in two spectral regions: 3-5 µm and 8-12 µm. Here we will look at two modes of operation. Both images were acquired by the ASTER instrument on NASA's Terra. The top is a 3.8 µm image taken at night, showing the coastline of Eritrea in eastern Africa. At night, water is normally warmer than much of the land, so it appears as a lighter tone than most of the land. The bottom image is a multispectral color composite of three bands in the 8-10 µm range. The area shown is the Saline Valley of eastern California (near Death Valley); most of the colors in this image can be related to rock types (silicates, carbonates, etc.).



Sections 10 and 11 touch upon some of the topics that are included in course on aerial photography. Section 12 examines the topic of photography from space platforms like the Space Shuttle conducted by onboard astronauts.

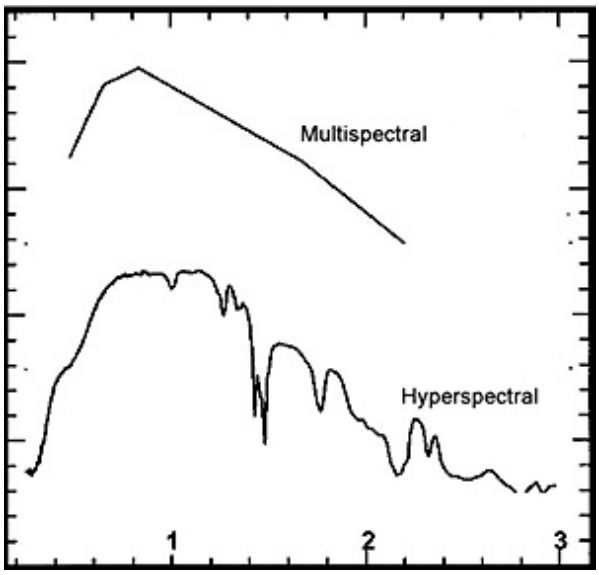
In Section 13, after a review of the methods of and necessity of "Ground Truth", you are introduced to the concepts of spectroscopy and, more particularly, hyperspectral remote sensors (this is also discussed on page Intro-24. Hyperspectral sensors are revolutionizing the ability of remote sensing to make accurate and precise measurements of individual materials (e.g., rock types; plant species) using "spectrometers" operating on the

ground, from the air, and now from space. Such a sensor is capable of imaging in narrow spectral width bands (typically 0.01 to 0.02 micrometers) over a broad, continuous range of the visible-Near Infrared spectrum. The basis for hyperspectral remote sensing is implied in this figure:



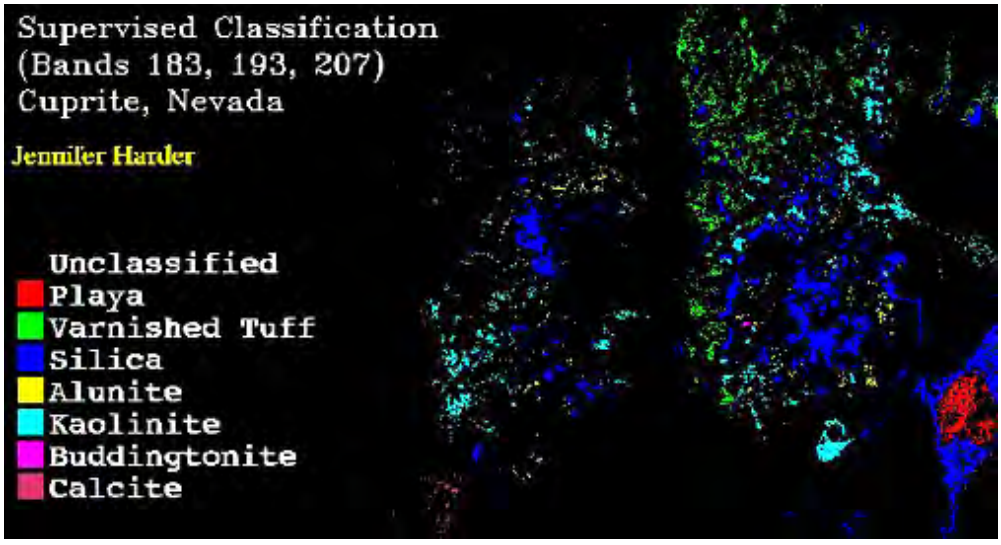
The resulting data set produces a *detailed* spectral signature (a plot of wavelengths versus some intensity function such as reflectance; see below on this page) for various individual materials (if spatial resolution is high enough) within a scene. More commonly (especially at lower resolutions) the scene contains features or classes that are a combination of several materials. Hyperspectral imagery is superior to Landsat and other modes of imaging because it can be used to better identify the individual members of those classes made up of mixed materials. Often, in the case of composition of rocks or varieties of vegetation, this can lead to much higher accuracy in separating and discriminating the components of the classes. The resulting hyperspectral signature, acquired by sensors mounted in aircraft or spacecraft, can be compared with numerous signatures for the classes/materials acquired on the ground (so-called "ground truth") on in the laboratory (this can be done automatically with signature banks stored in a computer). To appreciate this grand "leap forward", compare the two

spectral plots in this figure - the upper one is the "crude" spectral signature of a specific substance made with the 4 MSS bands on Landsat; the lower the hyperspectral equivalent signature:



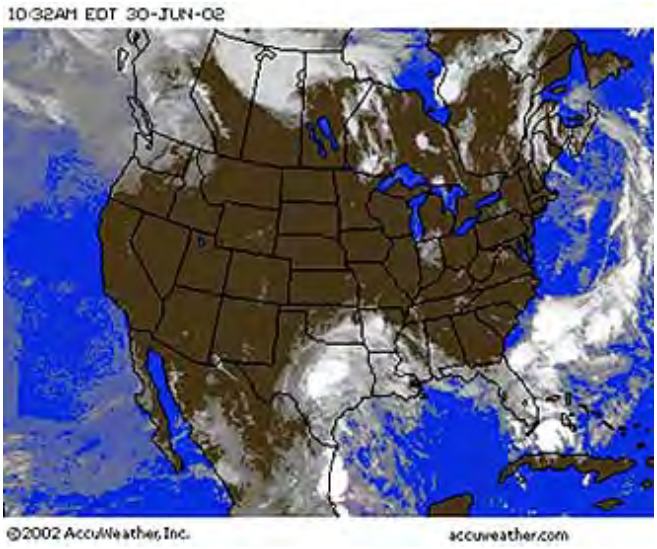
Many examples of hyperspectral signatures, images, and applications are starting to appear in the literature and on the Internet. One of the first systems used for airborne hyperspectral surveying is JPL's AVIRIS. Narrow spectral bands (equivalent to individual absorption bands in the detailed spectral signature) between 1.0 μm and 2.5 μm are particularly sensitive to key diagnostic inflections of the spectral curves obtained. Sulphides, oxides, carbonates, etc. among ore minerals and alteration products can be pinpointed by their characteristic wavelengths such that individual mineral species can be identified. Here is an AVIRIS image of part of Cuprite, NV made from 3 longer wavelength bands:

Using appropriate analytical techniques, these different minerals can be highlighted after identification at specific locations, with other materials blacked out, thus producing a mineral distribution map of minerals that are specific ore guides. Thus:

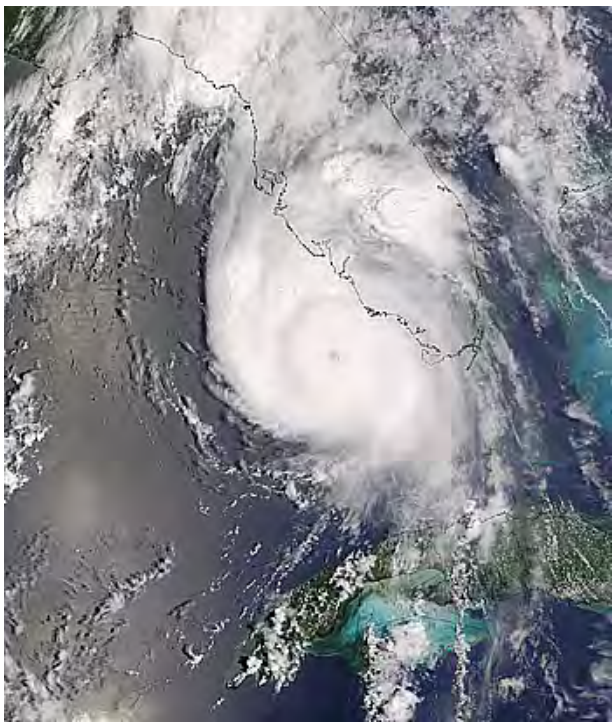


Satellites concerned with meteorological, oceanographic, and hydrologic phenomena constitute the largest number of Earth-observing platforms (Section 14). Readers of this Tutorial are certainly familiar with the Visible, Infrared, and Radar images of local, regional, continental, and hemispheric images of realtime weather systems moving in their vicinity because today's area-specific Newscasts use and show relevant images during the Weather segment of the programs. These Metsat (a general term for meteorological satellites) images are for most people the most commonly encountered satellite data presented to

the general public. For instance, this is an Accuweather image (Near Infrared) of the cloud distribution in the United States on June 30, 2002 (this weather imagery is updated hour by hour and can be easily accessed on the Internet):

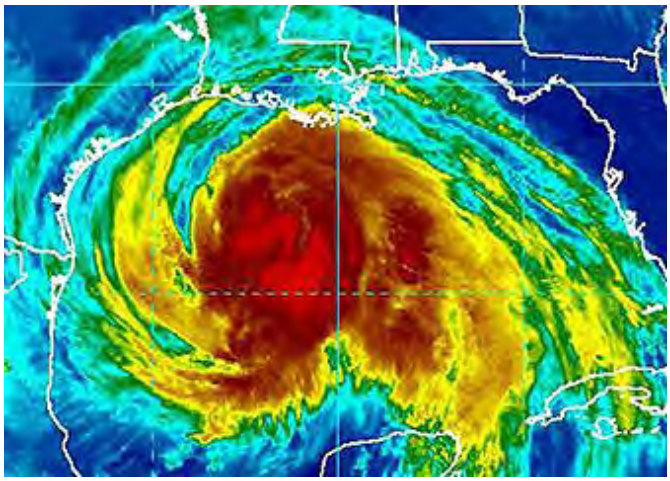


An obvious advantage, and PRIME USE, of satellite imagery is its "immediacy" of acquisition - images are available in near real time, and the areas covered can be re-examined over short and long periods. Change detection is thus one of the most common uses for repetitive satellite imagery. This is obviously important in meteorological studies but particularly so when a storm, or other weather aberrations, leads to a disaster of major proportions (thus Diaster Monitoring is a application of imagery from space). Abnormal weather events are at the top of the list for this kind of death-causing and property-destroying happening. A clear example of this was Hurricane Charley which hit the West Florida coast south of Tampa as a Category 4 (winds in excess of 120 mph) storm on August 14, 2004, killing 19 and causing more than 14 billion dollars in damage. This was one of four major hurricane to hit parts of Florida in 2004. Various satellites provided images of the storm at various stages of its advance (eventually up the East Coast into New England played out as a weak tropical disturbance). We show here a MODIS image of Charley as it hit the Florida coast.

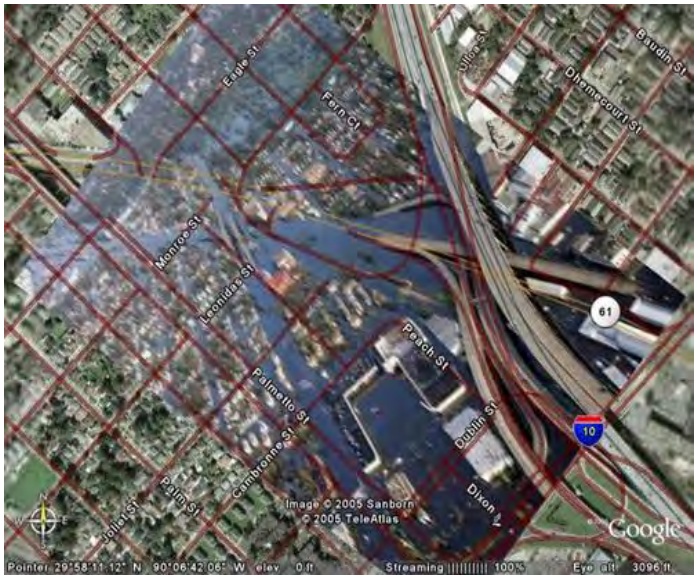


2005 has proved to be another very active hurricane season. By far the most notable was Hurricane Katrina which first hit Miami, then passed over Florida into the Gulf of Mexico turning north as a Category 5 that eventually hit Louisiana, Mississippi, and Alabama on August 29, 2005 to become probably the most destructive within the United States since meteorological records began to be kept systematically. Below is a NOAA view of the hurricane in the Gulf. But a much fuller account is given on page Page 14-10.

Hurricane Ike, which hit the Texas coast near Galveston on September 12, 2008, was a massive Category 2 storm which produced sea surge waves up to 16 feet. The storm was huge, areawise, covering most of the Gulf of Mexico. More than a million people evacuated low-lying areas, including parts of Houston that flooded. Satellite coverage of the storm was almost minute-to-minute. The surge wiped out entire communities along the coast.



Hurricanes cause flooding, which can be monitored from space quite conveniently. At one extreme is high resolution (less than one meter) imagery of small areas, such as this Google view of New Orleans during Katrina flooding obtained by the Quickbird satellite:



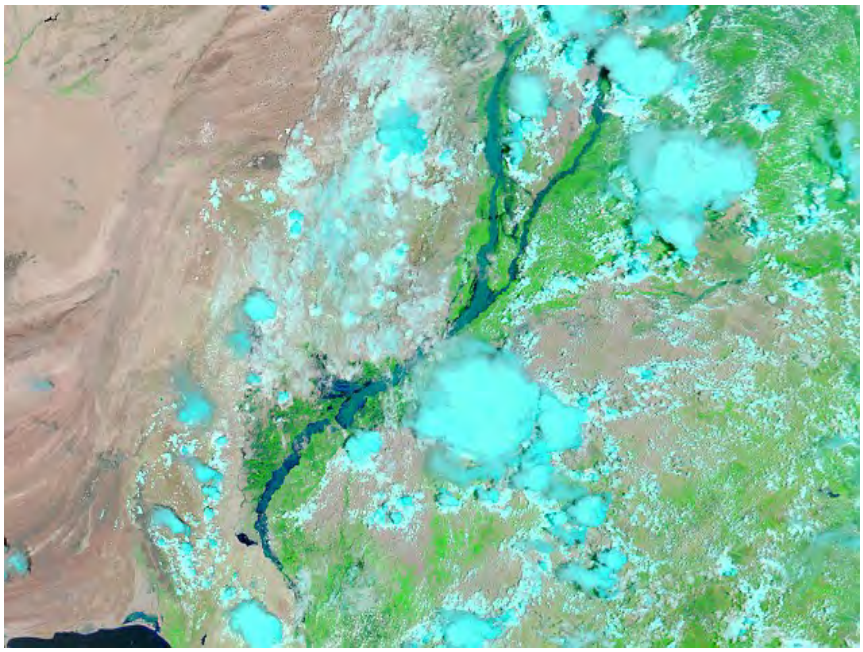
But floods often extend over large areas. This MODIS image of southern Myanmar (Burma) shows the Irawaddy delta and land to its north in a normal state (top) and in

flood after Cyclone (a hurricane) Nargis hit the Bay of Bengal in May of 2008, killing at least 10000 people. MODIS has low resolution (here, 200 meters) but covers large tracts.

April 15, 2008

May 5, 2008

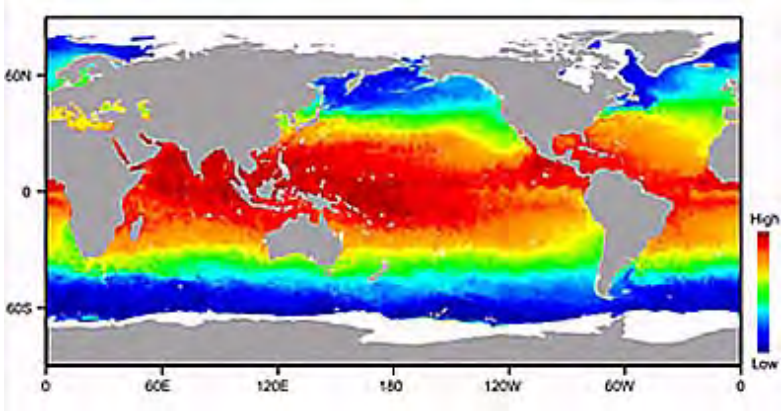
These next images capture a small part of one of the most extensive floodings in modern history. Monsoon rains deluged large areas of Pakistan in late July into August, 2010. At one stage a fifth of the country was underwater, displacing up to 20 million people and causing more than 1500 deaths. Two satellite images and two aerial photos demonstrate the extent and severity of this flooding:



© Getty Images

Another, longer time span example of systematic monitoring of weather-related change detection is *drought*. The U.S. Southwest has been experiencing an overall scarcity of rainfall, some years worse than others, for the past decade or so. This is dramatically revealed by this pair of Landsat-5 images of Lake Powell in Arizona, one in 1999 and the lower one in 2005:

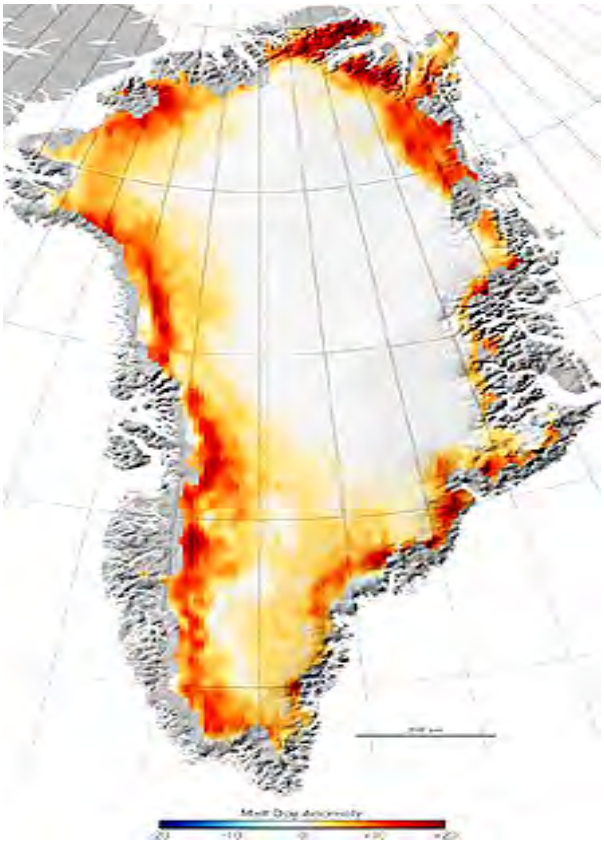
Less well known to the general public, but of great importance in understanding and predicting weather and climate on a global basis, are on-going measurements of oceanographic physical states. These, too, are investigated in Section 14 as satellites that obtain marine data are described. Here is a map of global ocean temperatures for a 3-day period in early June, 2002 as determined by Aqua, a mainstay of the Earth Observing System (EOS) program:



Section 16 also deserves your careful reading. It treats an on-going program (EOS) started in the 1980's that involves not only NASA but nearly all of the space agencies worldwide as well as environmental organizations from most of the nations now in the UN. Specifically treated are the status and results of several very sophisticated satellites - especially Terra and Aqua - that are part of the U.S.'s *Earth Science Enterprise*. This program, including satellites being launched by other countries, will peak during the first decade of the 21st Century but long range

missions extend well into the new Millenium. The fleet of satellites is dedicated to supporting a new field of science, known as *Earth System Science*. That is a multidisciplinary approach to study of Earth at a global as well as regional scales. Particularly involved are oceanographers, meteorologists/climatologists, biologist/botanists, geologists/volcanologists, environmentalists/ ecologist,physicists, chemists, and even sociologists, economists, and members of the legal profession. To learn more about these programs prior to working through Section 16, check this link: Earth Science Enterprise.

The first part of Section 16 considers the "hottest" topic going today - Climate Change and its corollary, Global Warming. Satellites, both meteorological and more generally, earth-observing play a vital role in monitoring regional conditions that provide key clues about temperature variations and their concommitant effects. As an example, here is a satellite-derived plot of melting along the periphery of Greenland in 2005:

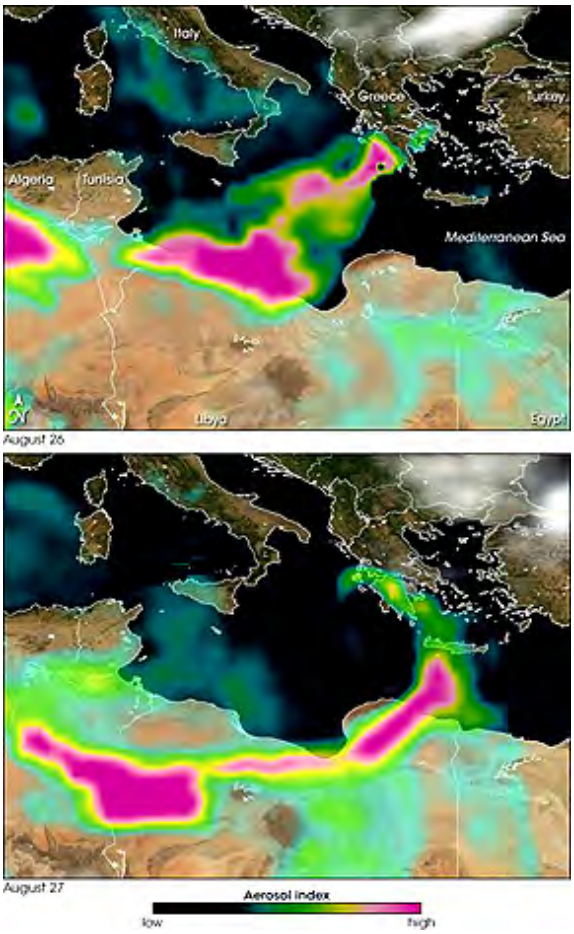


We shall see that observations made by Landsat, SPOT,

the Metsats and oceanographic satellites, Terra, Aqua, and many other satellites described both below and elsewhere in the RST have a very valuable functional asset: There are now enough of these in active orbits to cover almost the entire globe a number of times each day. (Good views are limited mainly by cloud cover.) Thus, timely (updated) information about continuing events can often be monitored successfully. Monitoring these over time illustrate one value of repetitive satellite coverage.

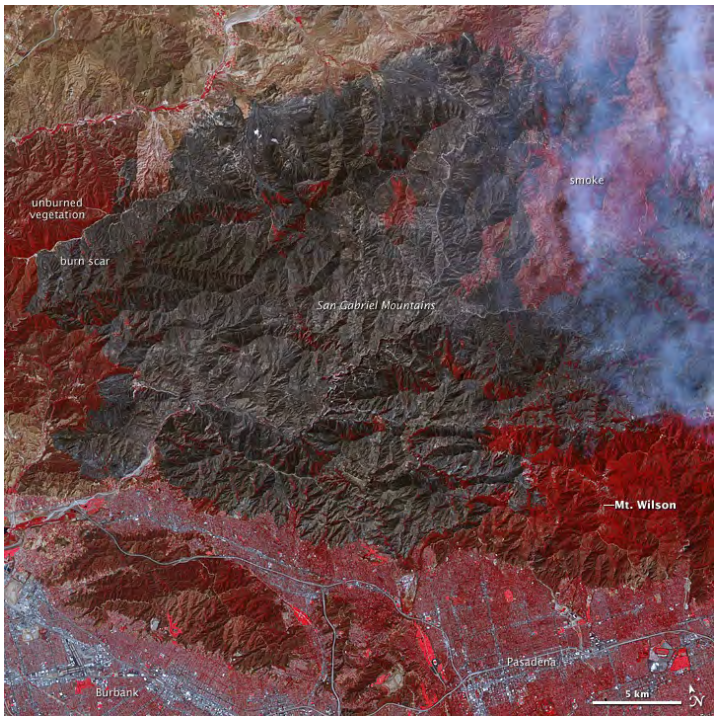
One example of monitoring short-term events that can be calamitous is coverage of dangerous fires. In August of 2007, a series of fires (most suspiciously set by arsonists) occurred over much of Greece. The region was tinder-dry owing to a severe summer drought. Whole villages were destroyed and at least 63 lives were taken. Here is a Terra MODIS image of these fires:

Terra's "sister ship", Aqua, has a sensor, OMI (Ozone Monitoring Instrument), that can image aerosols. This next illustration shows OMI's detection of smoke from the Greek fires imaged on two successive days, which exemplifies the Change Detection capabilties of earth-observing satellites:



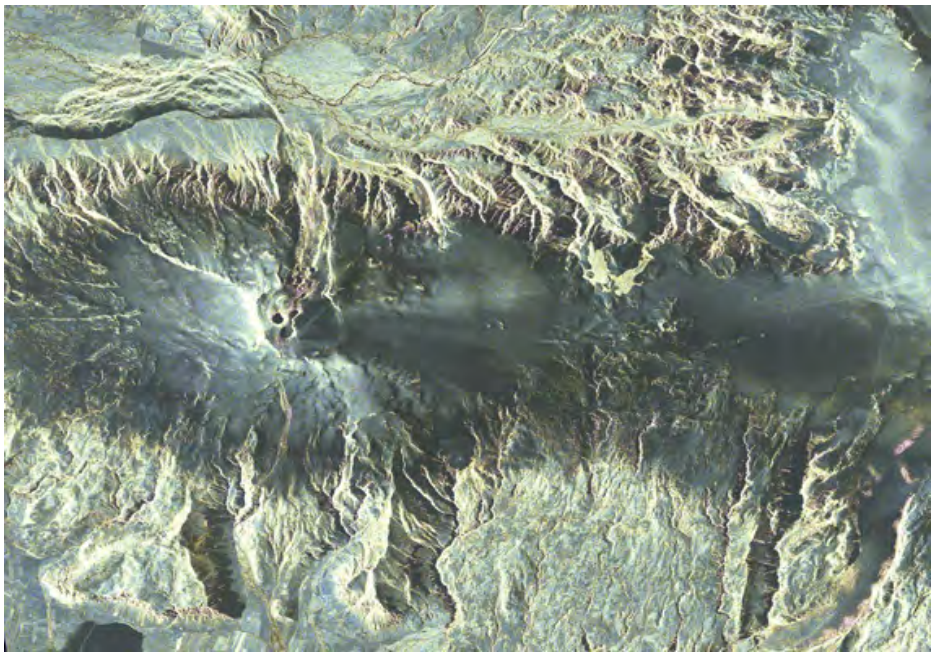
2003 started out as an average fire year until disaster struck southern California in October. Fires started both naturally and by arson in very dry forestlands and brushlands were influenced by Chinook winds (hot dry air coming south from the desert) and began to build, coalesce and move on populated areas. As of November 28, more than 950000 acres had been burned over, 4800+ homes destroyed, and at least 22 people killed. This MODIS image capture the wide extent of the blazes but does not show clearly their severity and destruction:

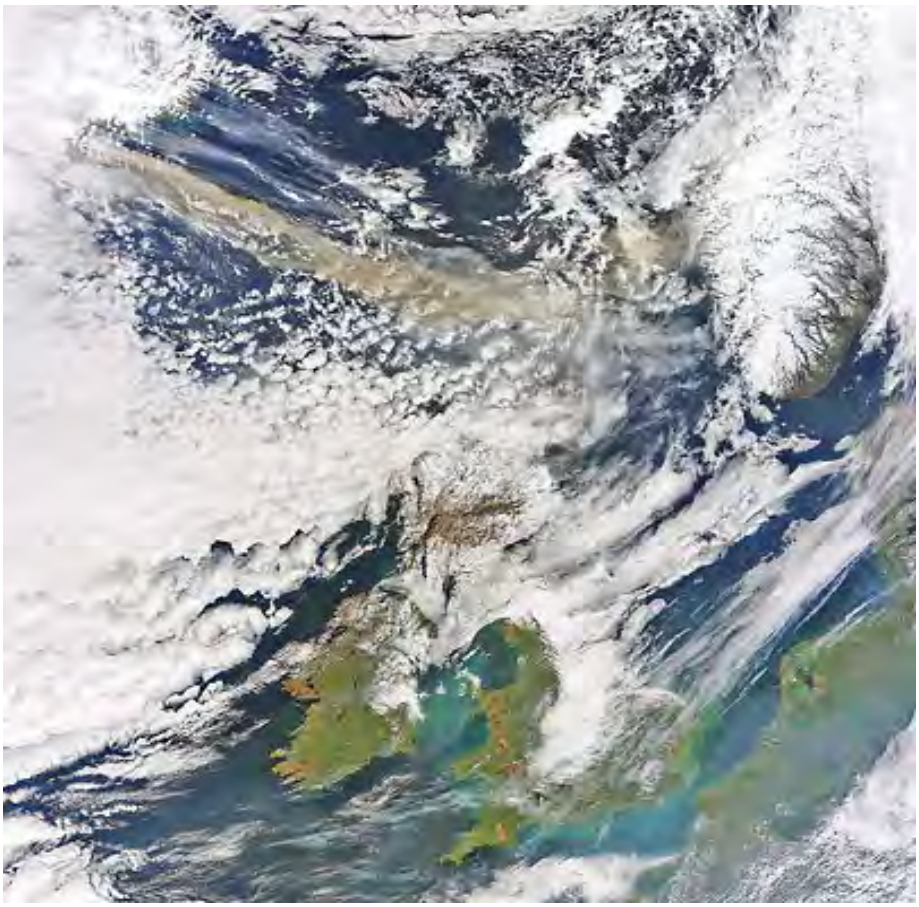
In 2009, a late summer fire in the mountains just north of Los Angeles burned more than 200000 acres. The burn scars are prominently displayed in this ASTER image;

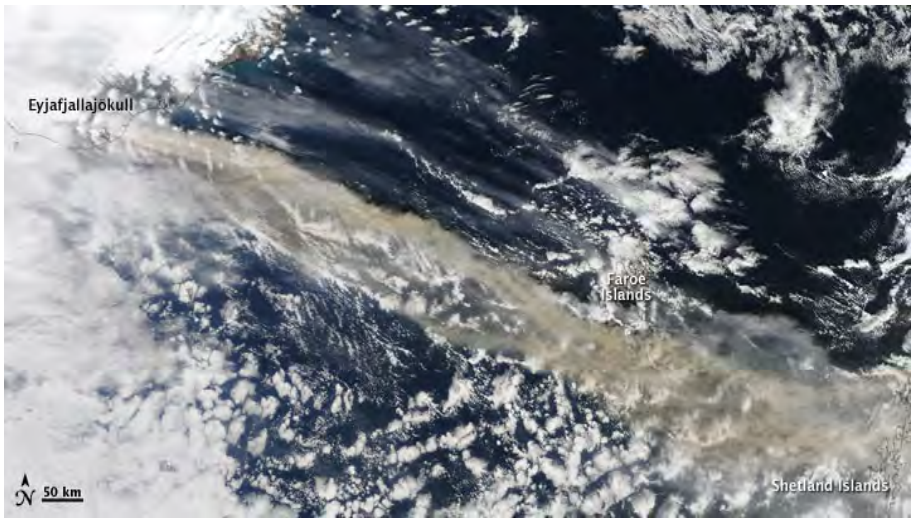


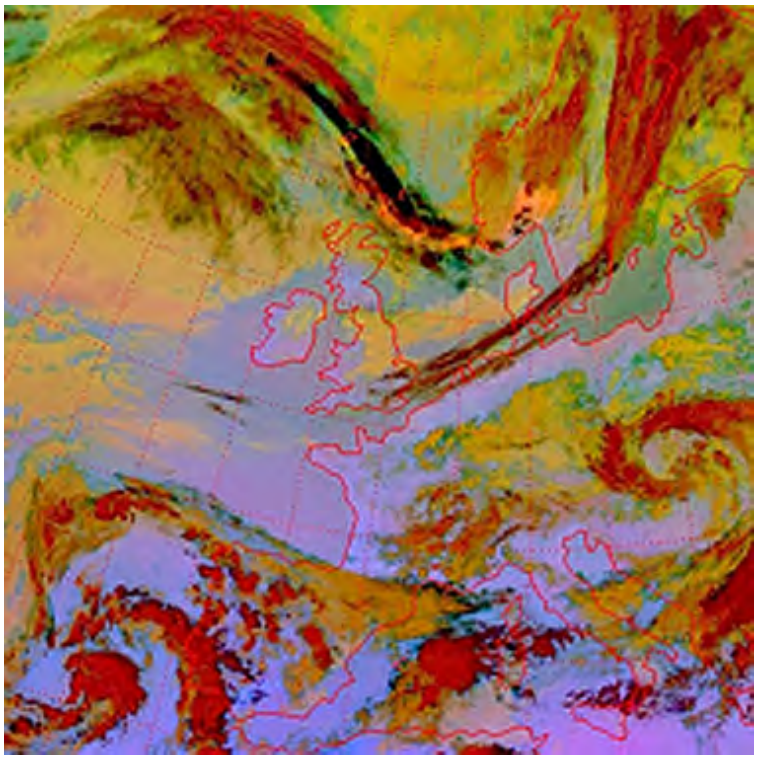
The fires described above all fall in the Disaster category. But some fires are deliberately set, and this is allowed, because they are part of a common practice followed for centuries. Thus, harvested crop stalks are burned off to prepare for the next planting (or to enrich the soil, which really doesn't happen). Or, forests can be burning as part of land reclamation or clearing. Sometimes these fires get out of hand, and grow uncontrolled; mostly, they are contained and just burn out over the area chosen for this action. Here is a MODIS image (January 2007) of a large area in southeast Asia where hundreds of set fires are visible.

Smoke of another kind can be monitored efficiently from space. In April of 2010, the subglacial volcano Eyjafjallajokul in Iceland began to erupt violently. It sent ash and gases more than 10000 meters into the stratosphere. Its plume reached the British Isles and spread by trade winds over parts of Europe. Because the fine ash can get into jet engines, the dispersed ash proved a major hazard to aircraft flying into and out of Europe. Thousands of flights had to be cancelled and many airports decided to shut down for days. Loss to the airlines was at least $200 million a day These images tell the story:

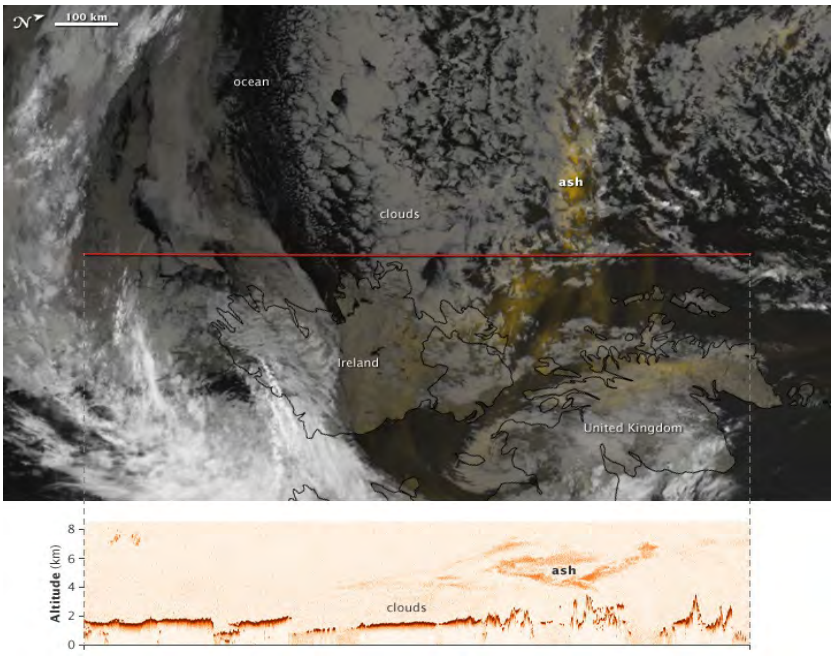





Eyjafjallajökull
Faroe Islands
Shetland Islands
N
50 km

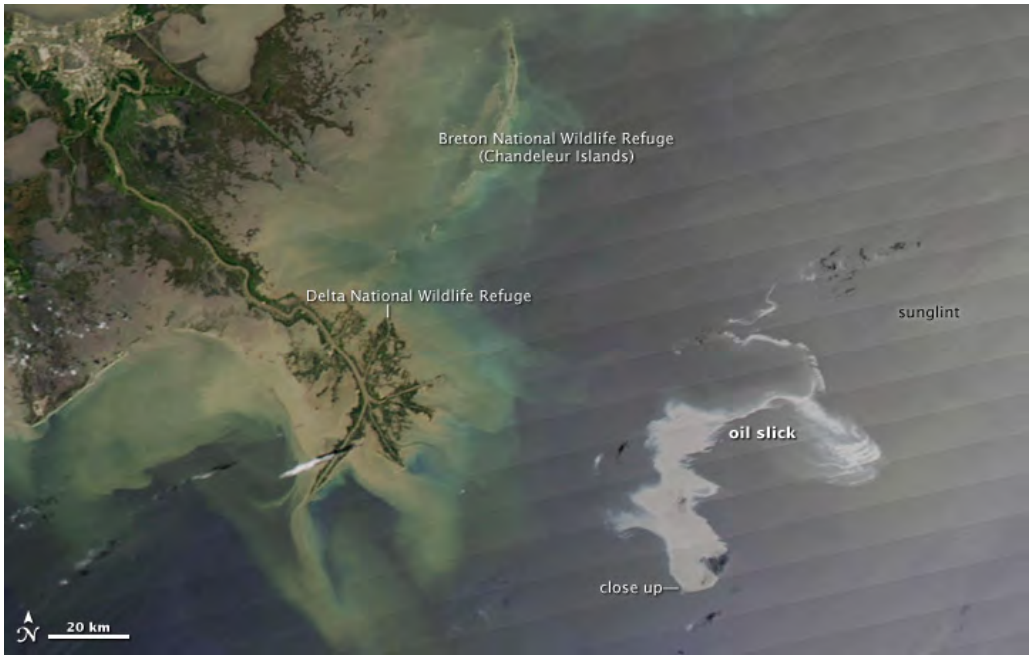


Not only photos (images) are returned from satellites, but direct measurements of physical parameters as well. For the Eyjafjallajokul event, NASA's CALYPSO satellite used its Lidar instrument to measure plume heights on May 16, 2010, leading to further airport closures. Here is an infrared image made by that satellite, beneath which is the Lidar plot of cloud heights in which the plume heights are obviously anomalous:

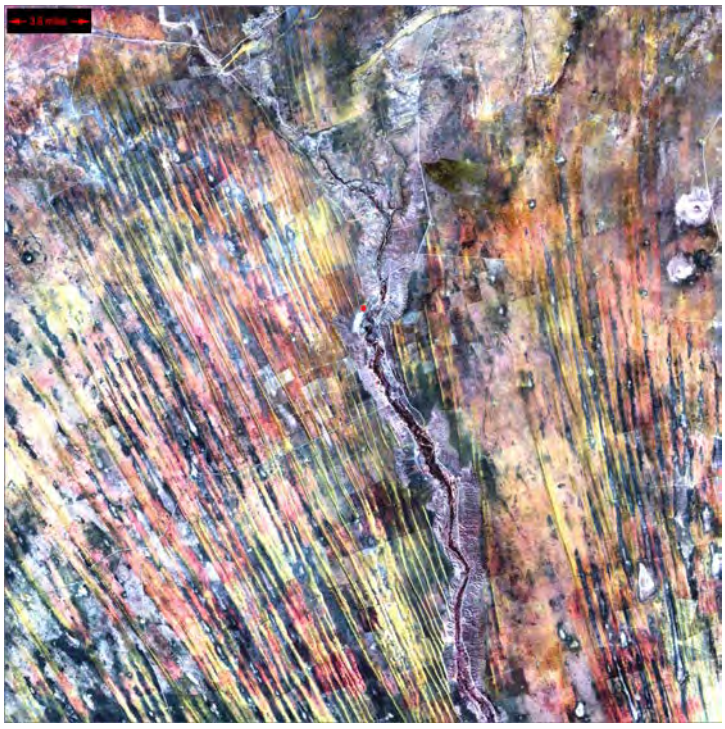


Another type of calamity readily monitored from space (provided the area is not cloud-covered) is the spill of oil

(petroleum) from a tanker ship (such as the Valdez) or from a well head. In April of 2010, a drilling platform off the Louisiana coast experienced a fiery explosion and eventually collapsed. Oil began to pour out from the floor of the Gulf of Mexico (some 5000 ft below), releasing thousands of barrels of oil each day. Rising to the surface, it collected into a slick larger than the state of Rhode Island, and driven by offshore winds hit the ecologically-sensitive coastal marshes. This MODIS image shows the slick during the first days (more on the spill on page 3-6.):



In addition to these timely applications, remote sensing is pertinent to long term basic science studies. Sections 17 (Geomorphology [Landforms]) and 18 (Impact Craters) give examples of how space imagery has been used in scientific analysis of these features on the Earth's surface. Space imagery is particularly suited to looking at landforms that extend over large areas such as dune fields, as exemplified by these parallel dunes in the desert of Namibia:



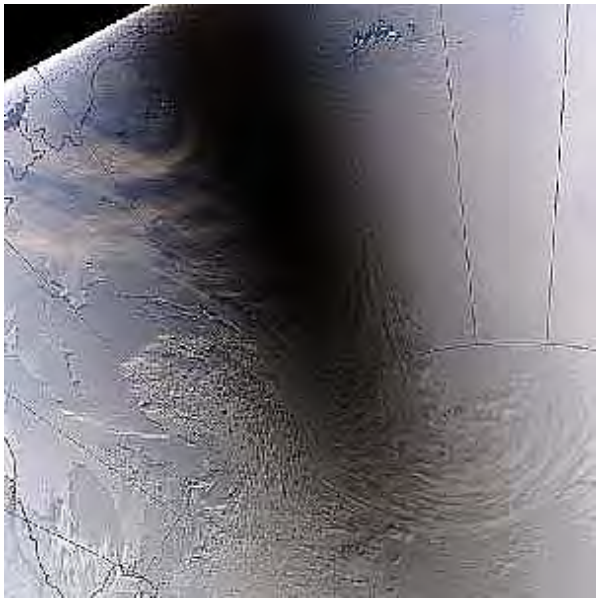
As an example of how impact craters look from space, examine this Google image of the Manicouagan crater (outlined by lakes filling its depressions outside its central peak) in Quebec, Canada:

Sections 19 and 20 - Planetary Remote Sensing, and Cosmology - also are best described as science topics. It is likely self-evident that the study of outer space - the Planets and the Cosmos - using remote sensors as the prime tool has a direct and vital bearing on how we humans need to understand the Universe beyond. One of

the most famous of all pictures taken from Space - the view of Earth from above the Moon as Apollo 8 passed overhead at Christmastime in 1968 - is reproduced here as a reminder that humankind's quest for knowledge now links our planet and to those beyond it in the Solar System. By inference this exploration hints that there are most probably other planetary systems in faraway galaxies.

A moment's digression for an oddity: On November 23, 2003, a transient incident took place that ties in both the planets and Cosmology (in the narrow sense of one star, our Sun) that also relates to the short-term change detection capability of satellite remote sensing just considered in the California fires examples above. On this date, a total eclipse of the Sun by the Moon took place in the high latitudes of the southern hemisphere. Visible from the Antarctic continent, the eclipse looked like this:

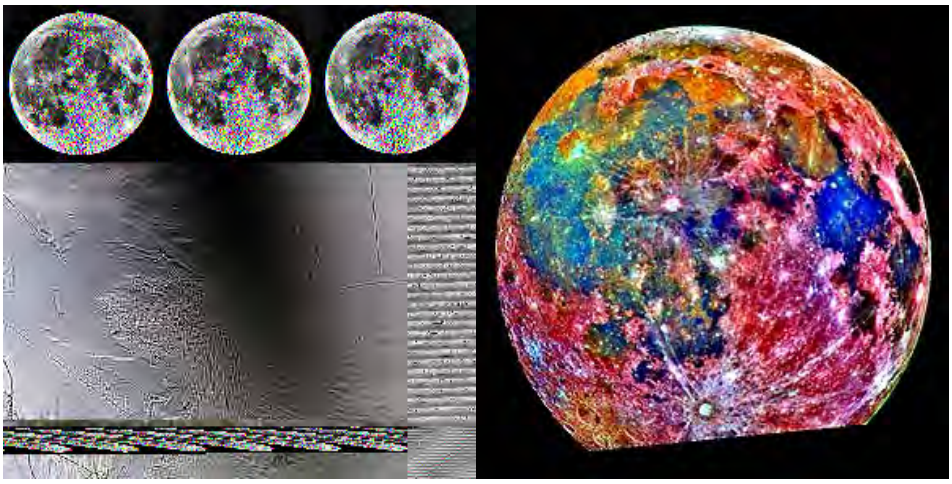
Amazingly, at the time of totality, the Aqua satellite was orbiting near the South Pole and was able to image the icy surface of the Antarctic in "real time" so as to capture the shadow caused by the Moon's blocking of sunlight as it proceeded across the continent:



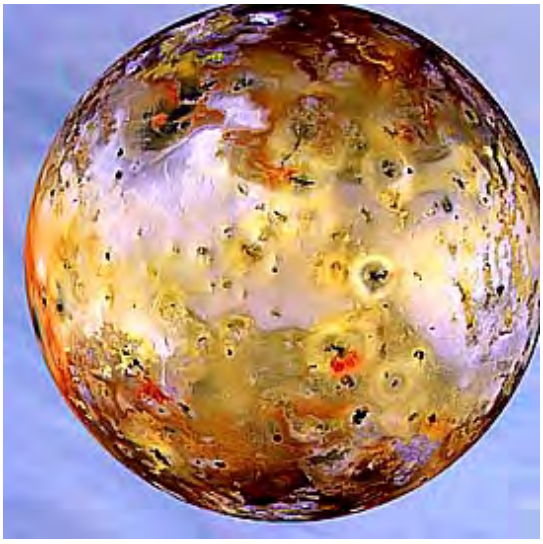
The reviews in Section 19 and 20 elucidate what Science has learned about these fascinating other worlds (planets, satellites, and asteroids) and about the stars and galaxies and their origins. While these topics seemingly stray from the main Tutorial theme focusing on the Remote Sensing of Earth, they offer an in-depth summary of the main achievements in the exploration of our Solar System and the Universe beyond. This exploration has been the centerpiece of the U.S., Russian, and now other space programs and has relied heavily on remote sensing

techniques (using not only the same wavelength intervals applied to terrestrial observations but also other regions of the EM spectrum). In fact, likely even more money has been spent on extraterrestrial remote sensing (consider the costs of Magellan, Voyager, Galileo and the Hubble amd Chandra space telescopes, and others) than on the study of the Earth that depends on unmanned satellites (although, the funding balance may be shifting with the new era of commericalization of terrestrial observations). However, interplanetary probes require high reliability so that much more money is involved in each such mission, so that comparison with Earth satellite costs is biased in that way.

In the first full decade of America's Space Program, the Kennedy commitment to land astronauts on the Moon captured this country's, and the world's, imagination as no other space adventure has matched. Exploration of the Moon is symbolic of NASA's greatest achievement. Even after the last Apollo crewmen left the lunar surface, its features have continued to be measured and analyzed. To commemorate this ongoing study of *our* satellite, shown here are two images of the Moon's front side, one just before Apollo, the other in the last decade of the 20th Century. On the left is a full view of the Moon obtained through an Earth-based telescope. On the right is a false color composite of much the same area made by sensors aboard the Galileo spacecraft as it sat in an earth-parking orbit prior to being sent on its main mission to Jupiter.



Space probes with a variety of imaging sensors have allowed planetary scientists to look closely at the Outer Planets - Jupiter, Saturn, Uranus, and Neptune - and have revealed the great variety and complexity of the many moons (satellites) around these Giant planets. To introduce the wondrous information gathered by spacecraft such as Mariner, Voyager, and Galileo, we show this full hemisphere view of Io, the innermost jovian moon. Io can be nominated as the most active, dynamic planetary body in the Solar System, if as the prime criterion *volcanism* is selected as the indicator of this status.



Section 20 considers most of the basic ideas of Astronomy and Cosmology (which, based on a Web Search, may well be the most comprehensive treatment of those two fundamental sciences now on the Internet). As a preview of the many truly beautiful, fascinating, and scientifically informative images spread throughout Section 20, we

show here a montage of what has been called planetary nebulae (a misnomer based on an earlier misconception, since these great blobs of glowing gas and dust are not the precursors of eventual planet formation but are remnants of stars that have exploded as supernovae).

In the Cosmology Section (20) (specifically, page 20-4) images of stars and galaxies made by instruments on telescopes using different intervals of the spectrum are discussed in some detail. Here we give one specific example: the Andromeda Galaxy as seen in a visible light image and an infrared image (wavelength of 175 mm), which is also then reoriented by a computer program to show it face on. The differences in information displayed and revealed are striking.

Modern History of Space, Appendix A, was prepared by staff at the Air Force Academy as part of their contribution to the Internet version. It is an exceptional review and well

worth a full read.

Appendix B has now been removed, since the PIT program it provides does not work properly for most users. See this appendix for more information.

Appendix C is a rather technical review of the concepts and underlying theory of Principal Components Analysis (PCA).

Appendix D is a fairly comprehensive Glossary. If you find a term or idea as you proceed through the Sections that may not be defined to your satisfaction, the Glossary is likely to have a concise definition to clarify the meaning.

As demonstrated earlier on this page, most Sections will have challenging and provocative questions within the text that you can use to help master some basics of image interpretation. This image is a typical example; try to respond to the question and then check the answer, so as to give you an idea of the nature of the questions you will encounter.

O-2: Below is a satellite image of a mountainous area in the state of Colorado. What do you think the white streaks are? ANSWER

With this insight into what you will encounter in the Tutorial, we move on consideration of the remaining three

topics or goals in the Overview.

THE BACKGROUND UNDERLYING REMOTE SENSING

Just what is this elusive 'remote sensing' we've been talking about? Try this general definition (a similar one is given on page 1 of the Introduction [click on this colored word to access it; then click on Back to return]): Remote Sensing involves techniques that use sensor devices to detect and record signals emanating from target(s) of interest not in direct contact (thus, at a distance) with the sensor. Let's break down the key words. Techniques range from simple visual interpretation of a sensed scene (which usually has both geometric [spatial] and geographic [locational] characteristics) carried out by one's brain to methods of analysis that utilize complex algorithms applied to digitized measurements. Sensors usually refer to systems that have optico-mechanical and electronic components - commonly sophisticated but can be as basic as a film camera. Detection implies the ability of the sensor to properly respond to the signal, and to measure its quantitative properties. Recording denotes the ability to retain the signal in a usable format that favors analysis. The signal itself can be diverse: Most commonly, it is some form of electromagnetic (EM) energy (as photons) that is expressed as radiation representing discrete wavelength intervals or bands (e.g., X-rays; visible light; radio waves) within the EM spectrum. However, remote sensing is also an appropriate term when applied to acoustical (listening) devices, to detectors that respond to magnetic force fields, and to instruments for seeing into human or animal bodies (such as CATscans). The 'target of interest' is almost self-explanatory - the words "feature", "object", "category", and "class" are descriptive. The idea behind 'not in contact' or 'at a distance' is synonomous with 'remote', in that the target is removed from physically touching the sensor. As a result, increasung distance allows a wider field of view to be

sensed (typically the sensor optics bring into focus all resolvable objects in a cone of observation) but at the expense of better resolution (showing greater detail). Most remote sensing applications involve having the sensors look down (commonly vertically) or outward (oblique views).

With this first insight in mind, consider this: Normally, we experience our world from a more or less horizontal viewpoint while living on its surface. But, under these conditions our view is usually limited to areas around our view site that fall within only a few square miles at most owing to obstructions such as buildings, trees, and topography. The total area encompassed in our vistas is considerably enlarged if we peer downward from, say, a tall building or a mountain top. This increases even more - to perhaps hundreds of square miles - as we gaze outwards from an airliner cruising above 30000 feet. From a vertical or high oblique perspective (as from a mountaintop or a skyscraper), our impression of the surface below is notably different than when we scan our surroundings from a point directly on that surface. We then see the multitude of surface features as they would appear on a thematic map in their appropriate spatial and contextual relationships. This, in a nutshell, is why remote sensing is most often practiced from platforms such as airplanes and spacecraft with onboard sensors that survey and analyze these features over extended areas from above, unencumbered by the immediate proximity of the neighborhood. It is the practical, orderly, and cost-effective way of maintaining and updating information about the world around us.

O-3: State an advantage and a disadvantage in conducting a remote sensing viewing from progressively higher altitudes. ANSWER

Until the 1960s, remote sensing was almost synonomous

with 'aerial photography', as will be evident after the next few paragraphs. Now, it is most often applied to "satellite imagery", which also is implicit in "satellite remote sensing" since the chief product of such sensing is an image or a map derived therefrom. Aerial photography is still big business but with the advent of high resolution satellite image, it has an ever lowering market share. Satellite-borne remote sensors (and comparable ones mounted in aircraft) have these major advantages over aerial photography: 1) they provide worldwide coverage almost automatically; 2) they have potentially high frequency of repeat coverage; and 3) they usually are multispectral in their design, allowing quantitative manipulations of the sensed data (which are also normally acquired in digitized formats), so that the objects in the scenes can be identified and analyzed by classification programs. This chart summarizes the main benefits of satellite remote sensing:

Below is a partial listing of some of the main applications of Remote Sensing.

1. Imaging, identification, and analysis of the features exposed on the Earth's land surface, gathering data about its landforms, geology, topography, manmade infrastructure, archeological remains, urban regions, agriculture, forests, rivers and lakes.

2. Imaging of the ocean surfaces, their physical properties, and various sea states.

3. Imaging of the seafloor and of creatures within the water (mainly by sonar).

4. Imaging of meteorological (weather) conditions and determination of various atmospheric properties.

5. Imaging of feature below the Earth's surface, using geophysical techniques.

6. Documentation of events on Earth, of landmarks, and of people, using photography and television.

7. Imaging of the internal parts of human and animal bodies by medical instruments.

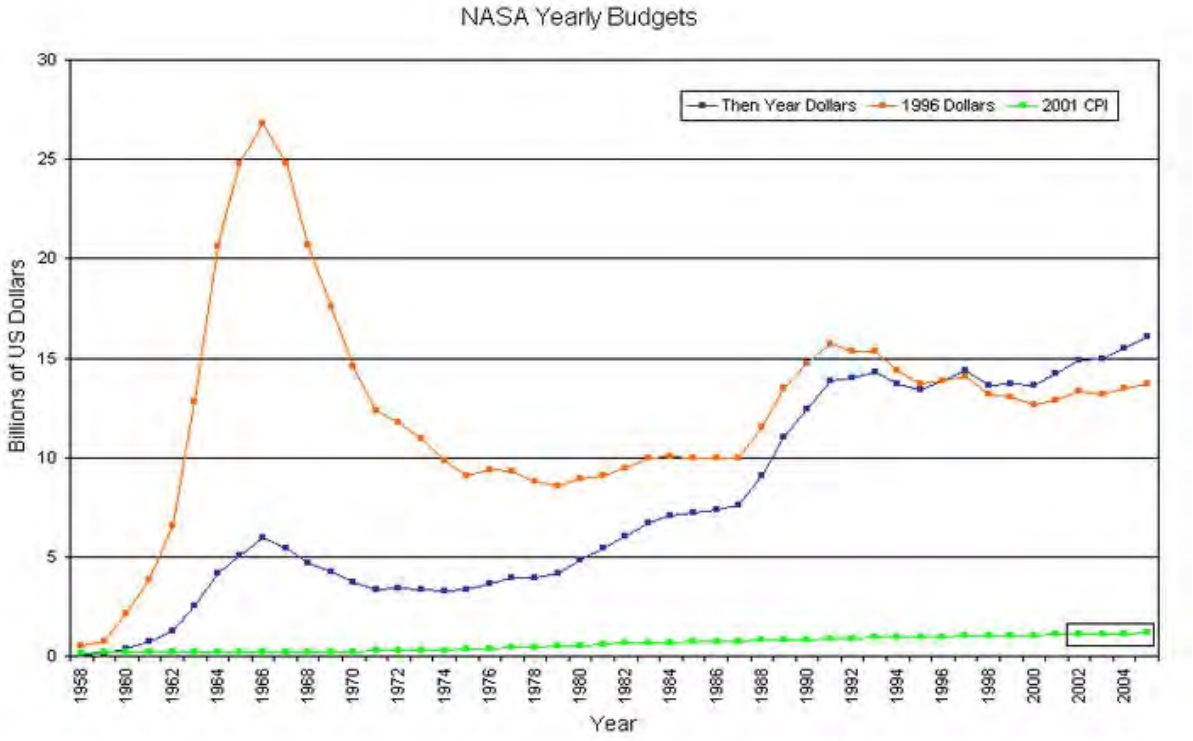
8. Imaging of the planets and moons of the Solar System; analysis of surface conditions using sensors on landers and rovers.

9. Imaging of astronomical bodies and physical materials; measurement of their spectral properties.

Now consider this very important precept or thesis spelled out in bold red letters to accentuate the importance and scope of remote sensing:

The above listing should lead you to the conclusion that remote sensing plays a vital and often prevailing role in many of mankind's activities. Most remote sensing systems are built around cameras, scanners, radiometers, Charge Coupled Device (CCD)-based detectors, radar, etc. of various kinds. Such systems are the most widely used tools (instruments) for acquiring information about Earth, the planets, the stars, and ultimately the whole Cosmos. These normally look at their targets from a distance. One

can argue that geophysical instruments operating on the Earth's surface or in boreholes are also remote sensing devices. And the instruments on the Moon's and Mars' surfaces likewise fall broadly into this category. In other words, remote sensing lies at the heart of the majority of unmanned (and as important tasks during some manned) missions flown by NASA and the Russian space agency, as well as programs by other nations (mainly, Canada, France, Germany, Italy, India, China, Japan, and Brazil) to explore both inner and outer space, from our terrestrial surface to the farthest galaxies. NASA's budget since 1958 through 2008 has been $433 billion dollars (see graph below). The budgets of the other countries involved in their own space programs is more difficult to fix but is when combined is probably similar to NASA's; thus it seems likely that at least a $trillion dollars have been invested in utilizing observations from space over the last 60 years. NASA *has spent more money* on activities (specified in the above listing) that - directly or indirectly - employ remote sensors as their primary data-gathering instruments than on those other systems operating in space (such as Shuttle/MIR/ISS and communications satellites), in which remote sensing usually plays only a subordinate role.



Add to this the idea that ground-based telescopes, photo cameras, and our eyes used in everyday life are also remote sensors, then one can rightly conclude that remote sensing is a dominant component of certain scientific and technical aspects of human activity that involve looking at and characterizing objects of interest - a subtle realization since most of us do not use the term "*remote sensing*" in our normal vocabulary.Think about the above statement for a moment and then - to further convince you of the role of remote sensing - run through this list of major (but partial) contributions from the science and technology of space: continual monitoring of agricultural productivity, deforestation, changing land use, growth of cities, effects of catastrophes (tsunamis, earthquakes, fires), short term weather behavior and long term climate changes, global warming, communications (from TV to your cell phone), military surveillance, the nature and characteristics of the planets and other bodies in our Solar System (including warnings of incoming asteroids), the astronomy and cosmology of the Universe; plus numerous "spin-offs" (computer technology has been accelerated by space needs). The essence of this list -- the space program is an

integral and intimate part of our lives. Once again, we reiterate: remote sensing has made most of this happen. However, remote sensing is utilized in various non-Space fields of application, as for example, commercial and personal photography, geophysics, and (rather loosely) medical instrumental diagnosis. Taken together, these "payoffs" in diverse utilization certainly exceed some (perhaps many) trillions of dollars in terms of investment and output. Thus:

REMOTE SENSING FROM SPACE IS NOW

BIG TIME !!!

Now, move on to the next page which provides a close overall look at the history of remote sensing and consideration of the broad principles underlying its technology.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Multispectral images are usually photo-like prints or computer monitor displays produced from input data representing segments or intervals of the electromagnetic spectrum. The process of producing multispectral photographs is explained. Hints for interpreting variations in gray levels in black and white imagery are given.

History of Remote Sensing: Multispectral Images

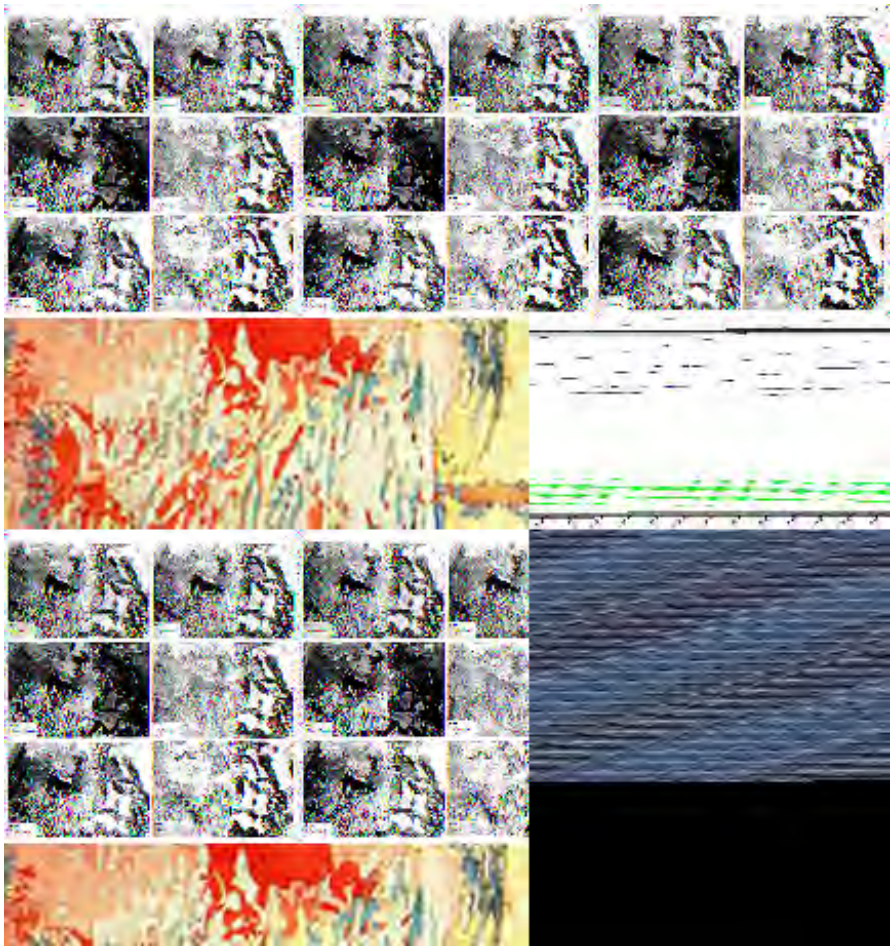
A significant advance in sensor technology stemmed from subdividing spectral ranges of radiation into bands (intervals of continuous wavelengths), allowing sensors that produce several bands of differing wavelengths to form multispectral images. This concept should be familiar to anyone who has used color filters on a photo-camera. Suppose you mount a red filter in front of the lens in a camera with black and white (b & w) negative film. (clicking on this transfers you to the next page, which describes how film works). Focused red light entering from an external object that generates red radiation passes through the filter and will activate the film, leaving numerous microflects of metallic silver, after development of the negative, wherever those light rays had struck the film; these form dark spots or patches in the negative. On printing to positive b & w paper, these dark areas in the negative prevent light from passing through; the positive film process produce light tones (a reversal) in the positive b & w print, so that red objects show bright (whitish) patterns that resemble their shapes. Conversely, the red filter absorbs light from green and blue objects, so that their (unpassed) light does not expose the negative. These areas on the film where a green object's image focuses will develop clear (no silver) in the negative and will print dark. Blue shows as bright shades when a blue filter is used. In effect, the color of an object can be identified by using a filter of that color to image it in bright tones.

Colors in color film are produced by stacking multiple layers of emulsions containing light-sensitive compounds (organic dyes) that filter out different wavelengths. In subtractive color film, the dye colors are: cyan, magenta, and yellow. Using the primary colors as reference, yellow subtracts blue, magenta subtracts green, and cyan subtracts red. So, when multicolored light enters a sensor, light from blue areas in the target or source, on striking the color film will bleach out parts of the yellow emulsion. The same pattern holds for green and red light, affecting the magenta and cyan layers, respectively. Then when white light passes through the multiple layers of the resulting transparencies (e.g., 35 mm slides), the now clear yellow areas will appear blue because the remaining cyan and magenta (so colored over these same areas) will filter out (subtract) red and green, leaving blue to display. The same reasoning holds for the other two primaries (red and green).

In color negative film from which color prints are made, the layer sensitive to red produces its complement color in the negative, which when printed onto paper produces red by leaving behind magenta and yellow dyes (from the [subtractive] color system). We won't elaborate further on the printing rationale; suffice to say that the red in print represents red from the source, green represents green, and blue represents blue. We review much more about film and camera processes in Section 10-2.

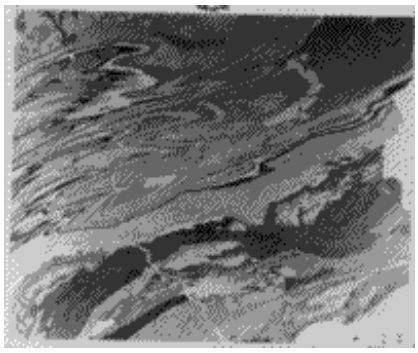
To illustrate the concepts introduced in the first paragraph, the writer (NMS) presents here an experiment done early in my process of learning remote sensing principles. I nailed the geologic map of the United States on the side of my home (done in bright afternoon sunlight, during the half of a Washington Redskins football game). I then photographed this color target (Geologic

map covering part of the western U.S.) using various filter lens, as indicated on each black and white image in the six panel illustration below it. Pick several prominent color patterns in the color version and find them in the various b & w versions, noting how the gray levels vary:

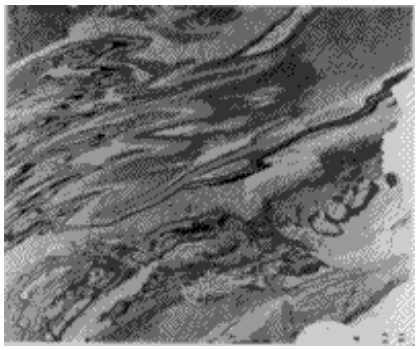


To emphasize this approach, we reproduce here the same four panels used to exemplify multispectral imagery as shown on page 34 of the Landsat Tutorial Workbook. The upper left panel shows the southeastern section of the colored geologic map of Pennsylvania.

Color Photo



Blue Filter



Green Filter

Red Filter

This map, illuminated in bright sunlight, was photographed three times on standard black and white film through three narrow band pass filters, centered on the blue, green, and red segments of the visible spectrum. Look first at the resulting black and white photo made with a blue filter. Light reflected from the bluer patterns in the map passed through the blue filter with high transmission (low to moderate absorption). The film negative receiving such light was exposed strongly (high silver density) and the positive print made therefrom showed the blue areas as light shades of gray. Conversely, the red and orange reflected from the map was highly absorbed by the blue filter, causing only slight densities in the negative and very dark shades in the print. Greens in the map show as intermediate gray shades in this blue-band print. We can apply the same reasoning to the green and red-band prints. Thus, red patterns displayed as light gray in the red-band print and moderate gray in the green-band print. Note that some colors tended to produce generally darker shades in all three filter prints, as for example, the dark reddish-brown zigzag pattern on the left side of the map. This darkening results in part from the inherent darkness (low level of saturation) and nature of the particular color brown (a mix of red and yellow with black).

I-16: **To check on your understanding of the explanation in the above paragraph, scroll upscreen until the color version of the map is at the upper left. Now, one at a time pick out about 5 to 7 points on this map where a different color is at each. It helps to choose each point where it is associated with a distinctive patter. Make an educated guess as to the level of gray (from white through various shades of gray to near black) you would predict when you locate each point again in each of the three color filter images. Derive a rule-of-thumb statement that summarizes what you would expect the gray level to be for each filter image.** **ANSWERS**

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

This brief page treats some of the basic principles underlying the standard photographic process involving exposure and development of black and white film.

Film as a Recording Medium

Black and white film uses small silver chloride crystals suspended in an emulsion, coated on a transparent backing (In the early days of photography, glass plates were used as the transparent backing). Photons striking the AgCl in a crystal grain cause ionization into Ag^+ and Cl^-. The greater the amount of light impinging on a grain, the more Ag^+ions are produced. Those crystals that have been exposed to light have their chlorine ions removed by the "Developer", leaving behind the Ag^+ now reduced to the metallic state as black silver crystals. After development the "Fixer" removes any silver chloride that was not light-activated. The small silver crystals form opaque areas on the film. Illuminated from behind, those parts that are exposed to light are dark (hence the term negative). The more light, the more silver gets left behind on the film, up to the maximum density of the film. To make a print, the process is repeated by shining light through the film onto a piece of white paper coated with a silver chloride emulsion, in effect reversing the above process so that white to light gray areas occur where little or no light passed through the negative and dark areas correspond to the clear areas in the negative (which experienced little activating light and hence no Ag concentration). Instead of a print, which is viewed as a reflected light product, the positive can be a black and white transparency, which is viewed by having transmitted light (either from sunlight or an artificial light [light table; lamp bulb]) shine through it from behind.

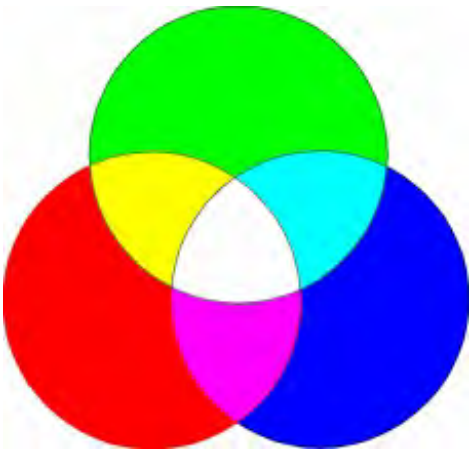
More information on the photographic process can be found on page 10-2. You may also want to visit these Wikipedia websites: film and Photography.

Primary Author: Nicholas M. Short, Sr.

The process by which natural and false color composites made from three multispectral intervals or bands by projection and superposition is discussed. One important conclusion is that when a spectral band which contains reflectances in the wavelength interval between 0.7 - 1.1 µm is projected through a red filter, healthy, active vegetation will appear bright (light tones) in a black and white image or red in a false color image.

History of Remote Sensing: Color & False Color Composites

We can use color-filtered b & w pictures as transparencies projected through color filters and superposed (registered) to produce **color composites**. To do this, imagine this setup: Pick any scene containing many features and classes of differing colors. First, replace the prints with positive transparencies (tonally analogous to prints). Work with three b & w transparencies, each representing its spectral band. Shine white light through each one mounted in its own lamp projector (total of three) on to a screen. Project the blue band b & w transparency through a blue filter, the green through green, and the red through red. Blue features on the ground are clear areas in the blue spectral band. When the blue band transparency is projected through the blue filter, the blue features will be blue on an observing screen, likewise the green band projects green objects through its filter as green, and red as red. Co-register (line up) the three projections by superimposing several distinctive patterns that are common within the photographed scene. The result will be a simulated natural color image. with any red, green, and blue objects or classes showing in the projected image as these colors respectively. Other colors present are additive mixes of two or more primaries (e.g., yellow is a mix of red and green; orange is a mix of more red and some green; white is an equal mix of all three primaries, and black is simply the absence of any colored light of any wavelength). The colors that result from combinations of blue, green, and red (the primaries) are indicated in this additive color diagram.



This way of combining individual colors to produce a composite is called *color additive viewing*. The filters and films used do not necessarily have to correspond by wavelength. For example, a green spectral band image can be projected through a blue filter and a red band through a green filter. We can modify the color assignments in making an image if one of the transparencies is the infrared band and is projected through a red filter; the result is one version of a what is called a **false color (IR) composite**. This procedure and the resulting colors that are produced are indicated in this diagram:

Note: the shading on this diagram is not as contrasty as intended (and the original figure has been lost). To clarify: 1) in the left square only the diamond is white (clear), the others are light gray; 2) in the center square the triangle and circle are white; 3) in the right diagram the cross and circle are clear. In the discussion in the next two paragraphs, one can consider the gray shades as very dark (in a transparency, would not allow light to pass through).

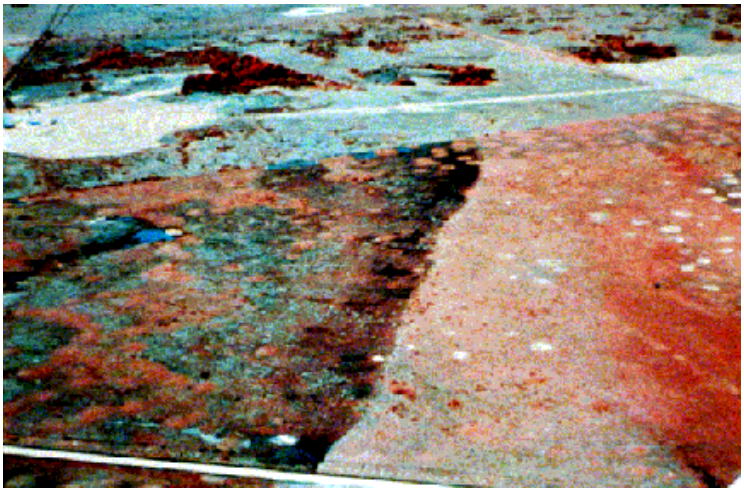
The setup is a variant of the one described in the first paragraph in which a true color image is formed. In the above figure, each of the four geometric designs in its large framing square is either clear (actually open) (shown as white inside its boundaries) and will pass all light impinging through it or is shaded gray inside which means no light passes through. Each is located in the same position in all three squares. Each square is labeled as "band filter" (corresponding to the band associated with the Multispectral Scanner on Landsat-1). Thus, the MSS Green band (bottom) shows the diamond as a positive bright figure in a black and white print images and the other shapes are dark. As applied to a real image corresponding to the MSS Green band output, this just means that objects in the scene (on the ground) that are green show up as light-toned in a print or largely clear in a transparency; blue-green and yellow-green objects (whose wavelength spreads are offset from those of a pure green object) are lighter shades of gray (less clear). The same argument applies to the Red and IR band squares, i.e., white refers to objects that give off Red and IR radiation respectively, and gray refers to objects that don't represent those wavelength bands.

The three band squares are now mounted each in its own light projector (for instance a 35 mm or a lantern slide type). Light passing through a clear design is projected on a screen surface; each projector is moved until designs with equivalent shapes are registered (superimposed). A color filter is placed in front of the lens of each projector: blue for square A, green for B, and red for C. As seen on the screen, the diamond will show as blue , the inverted triangle as green, and the + as red. The combination of green and red for the circle design (in B and C) produces yellow (in the color additive process).

To relate this to how it applies to multispectral photography, consider square A as representing a scene in which only green light is reflected from objects (not too fanciful - the Emerald City in the "Wizard of Oz" was all green). Square B represents a reflected red light only scene; square C represents a scene containing only objects that give off only infrared light. The screen receiving the projected and superimposed light would show a blue object, a green object, a red object, and a yellow object, corresponding to real world objects imaged by green, blue, and infrared filters; the projected objects would be separated in the superposition, corresponding to their initial separations in the square.

You have already seen false color images in the Overview. Let's see what a false color image looks like in the ground scene that follows. First we show the area, grassland and a field with natural shrub cover (left), as it appears in natural color in this aerial oblique view:

The companion photo below is a notably different color version (typical false color rendition) in which various kinds of vegetation display in several tones of red, pink, or yellow (the latter two may indicate a degree of stressed or unhealthy vegetation).



This type of projected color combination, yielding the false color IR composite, is ordinarily used to emphasize a property of healthy vegetation in which incident light in the range of 0.7 - 1.1 µm reflects strongly from the internal cells of plants, giving rise to bright tones in the color IR film.

I-17: **What kinds or types of materials/classes are associated with the white, the blue, the dark red, and the light red features noted in this photo. Account for the color.**
ANSWER

When this film combination is used in military reconnaissance photography, weapons camouflage - simulating vegetation - is not red in a false-color scene, because the high near-IR reflectance from healthy vegetation is absent (synthetic fabric in clothing is not bright in the near-IR; if made usually from cotton or other organic, this too is at best a subdued bright). Or, in a more familiar mode, color IR photos (and sensor-derived images that include an IR band) show a football field with natural grass in bright red as compared with artificial turf (Astroturf) in a dark, non-red tone.

Lets re-examine the square frames example above by conducting another experiment. Instead of the clear areas passing light, let them reflect light back towards a film camera. The gray areas reflect no light (ignore the areas outside the designs). The objective of this experiment is to determine what gray level wil be recorded on visible or infrared-sensitive film when the clear designs are (1) assigned a color (including IR), and then (2) viewed through some highly sensitive filter (one that passes onlylight of its stated color).

So, going back to the diagram, if we look at the reflecting diamond design in A through a green filter (the Landsat MSS case), we ask what color(s) it can have that will pass through the filter,

be recorded as very dark on a film negative, and then expressed as a white (or light gray, since the filter may also pass some adjacent wavelengths) diamond shape on a positive black and white print. The obvious answer is green. If the objects in A are any color other than green, the filter would not pass these wavelengths and the diamond would be very dark in the print. The same argument holds for B and C. C, for example, has two designs that reflect the light. When that light is passed by an infrared filter, if light tones for these designs in a b & w photo-print are the result, this would mean that both designs are brightly reflected in the IR and not in the visible range. Going back to the square A case: if the diamond were white in actual fact, not green, and a green filter were used, then only the green component of the visible spectrum would pass and the resulting print would be a low to moderate gray. (The fact that it is white would emerge from the lightness of tone in each square print [provided one was made with a blue filter].) If the diamond were black, no light would pass and the print would be dark.

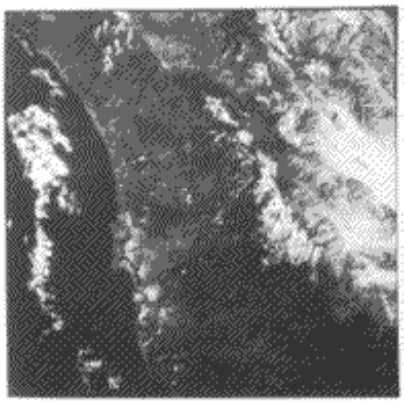
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

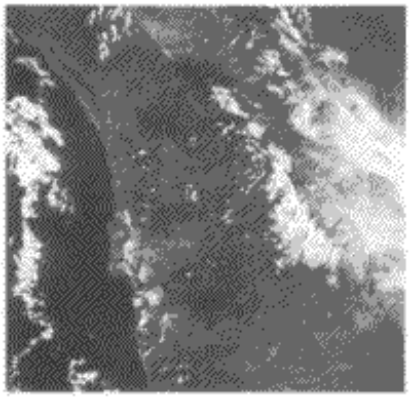
Before the first Landsat (1972), perhaps the best examples of multispectral photographs were those taken by a four camera array mounted in the Apollo 9 (1968) spacecraft window. Skylab astronauts later used a six camera array.

History of Remote Sensing: Apollo 9 Multispectral Images

The first multispectral photography from space happened during the famous 1968 Apollo 9 mission. Scientists mounted four Hasselblad cameras in a holder such that they all aimed at the same target point when an astronaut triggered their shutters simultaneously. Below are three filtered b & w photos of southern California around San Diego, which the astronauts took in the green, red, photo IR bands, and a (false) color IR picture.



Green



Red

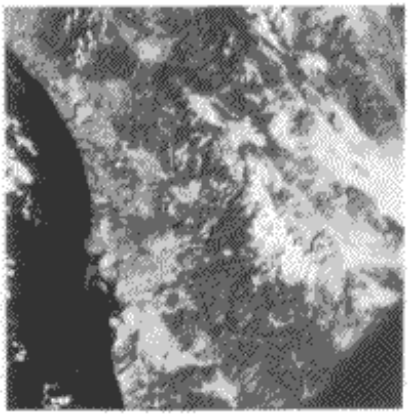
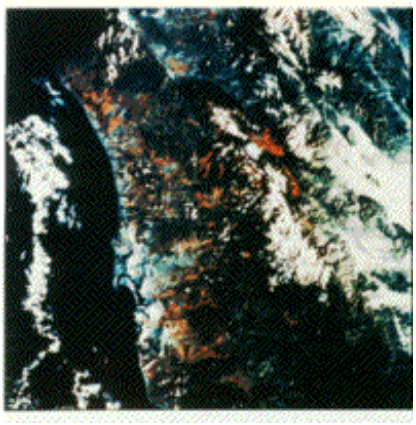


Photo IR Bands



False Color IR

I-18: **Note that clouds, snow on the mountains, and light desert soils (right center) all appear white in the false color composite image and the three black and white filter images. There are several areas of bright to medium red in the f.c. composite. Try to**

find their corresponding locations in the three filter images and be aware of their gray levels. Answers

A six-camera multispectral system flew on the Skylab spacecraft, which several crews occupied between May 1973 and February 1974. The photos they took helped show the value of multispectral photography to discipline scientists , particularly geologists, hydrologists, agronomists, foresters, and those concerned with environmental monitoring and land use and cover assessment.

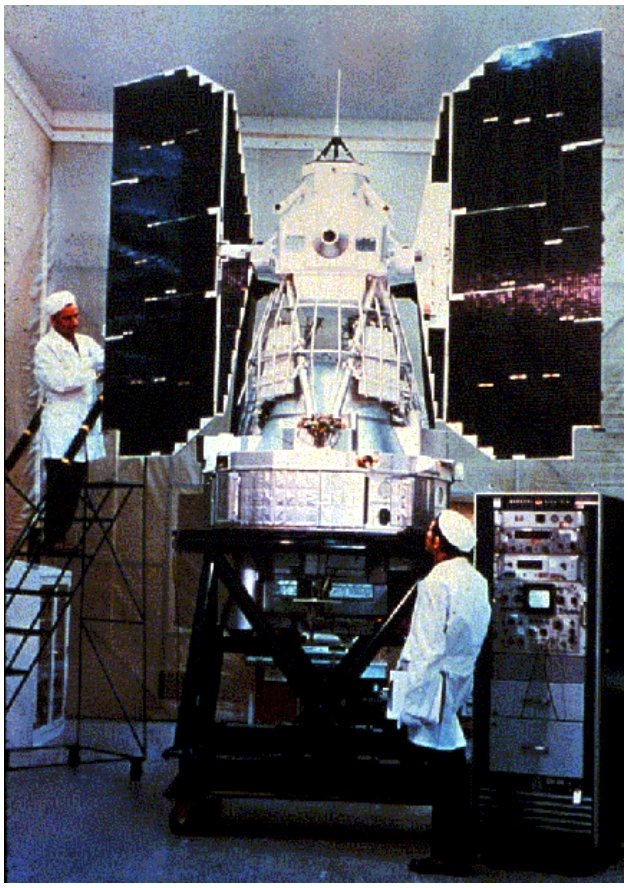
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The sources of many of the images displayed throughout this Tutorial are ERTS-1 (renamed Landsat) and subsequent Landsats, earth-observing satellites originally developed by NASA and operated by them and/or several other agencies from 1972 through the present (Landsat-7). The Landsat program is described and the orbital and instrumental parameters for this series of satellites are given. Discussion of Landsat's two prime sensors, the MSS and TM, and examples of images produced from each is deferred to the next 6 pages.

History of Remote Sensing: Landsat (ne: Earth Resources Technology Satellite)

By the late 1960s, the first unmanned satellite specifically dedicated to multispectral remote sensing entered the planning stages. NASA carefully designed and constructed, then launched ERTS-1 (Earth Resources Technology Satellite) on July 23, 1972.



No satellite program designed to monitor the Earth's surface, the features thereon, and thus the environmental states that relate to land cover, land use, and natural resources, has proved more important than ERTS (Earth Resources Technology Satellite). The key people involved in its conception, inception, and operation deserve special mention here:

William T. Pecora was Under Secretary of the Interior, Director of the U.S. Geological Survey, and Landsat program champion. Unfortunately, he died just a few days before Landsat-1 was launched.

William A. Fischer, one of the nation's best known photogeologists, had the original idea for Landsat, which he shared with Pecora.

Dr. Arch Park chaired the management team that defined the U.S. Department of Agriculture's research program for Landsat.

Dr. William Nordberg designed NASA-supported research programs for Landsat, coordinating the activities of 300 scientists from 38 countries during Landsat's first year.

John De Noyer directed the development of NASA-supported research programs for Landsat.

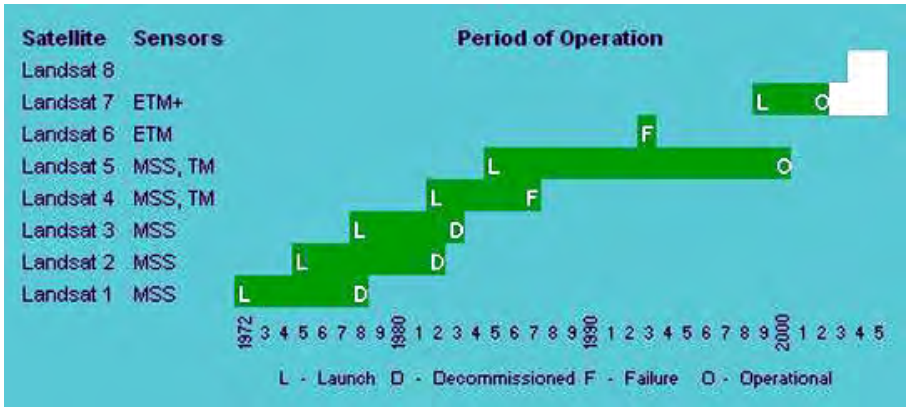
Bruton "Doc" Schardt, Landsat's first program director, worked tirelessly to ensure the simultaneous development of hardware and research programs, while resolving numerous political and budgetary issues.

Virginia Norwood, senior scientist at Hughes Aircraft Company, headed the team that developed the Landsat I MultiSpectral Scanner

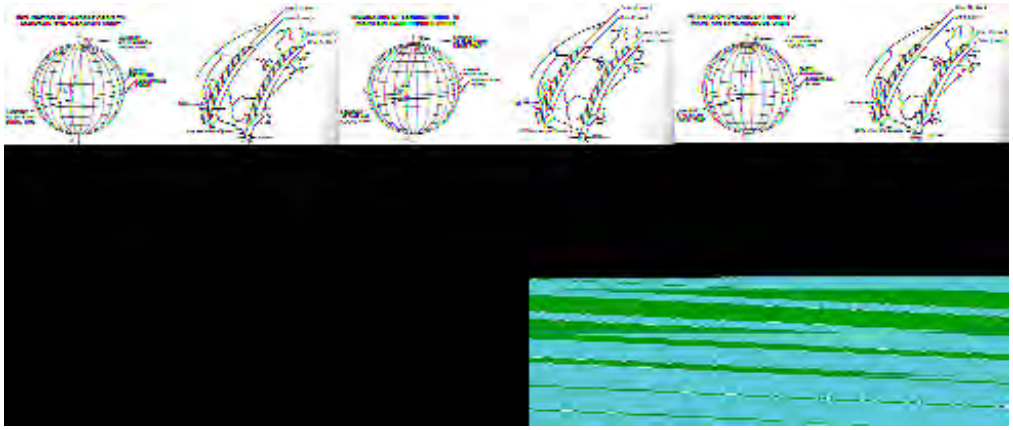
All were colleagues or acquaintances of the writer (NMS) and William Nordberg was responsible for bringing me into remote sensing and the Earth Observing program.

An interesting, "off the cuff", synopsis by Stephen Hall of the genesis and early days of the Landsat program can be visited at a USGS website called EarthShots.

Renamed Landsat, ERTS-1 was the first in this series (seven to date) of Earth-observing satellites that have permitted continuous coverage of most of Earth's surface since 1972. Launch dates are: Landsat **1**, July 23, 1972; **2**, January 22, 1975, **3**, March 5, 1978; **4** , July 16, 1982, **5**, March 1, 1984 (Landsat **6**, launched later, failed to operate); and **7**, April 15, 1999. (As of May 1999, only Landsats 5 and 7 are still operational, i.e., acquiring data; the older ones have been shut down.). To help to visualize this information, we repeat an illustration first shown on page 1 of the Overview:



The orbital conditions for Landsat 1 are depicted in this illustration and described in the next paragraph.



The first three Landsats orbited at an altitude of 570 miles (923 km); 4, 5 and 7 at 435 miles (705 km). The orbits of all Landsats are near-polar (inclined 9.09° from a longitudinal line) and Sun-synchronous (pass every time over the equator between 9:30 and 10:00 AM), making 14 passes in descending mode (southward from the North pole in the daylight mode) each day (about 103 minutes for a complete orbital circuit). After any given orbit, the spacecraft will occupy its next orbit some 1775 miles (2875 km) to the west; on the next day, the orbits are configured so that orbit 15 has displaced westward by 98 miles (159 km) at the equator. Landsats 1-3 will reoccupy almost precisely the same orbit after 252 such orbits, or 18 days later; Landsats 4, 5 and 7 reoccupy on a 16 day cycle. Under the above orbital conditions, and with an angular field of view if 11.58 ° the width of a Landsat MSS scene is 185 km (114 statute or 100 nautical miles). The continuous imagery along an orbital strip is cut (subdivided) every 185 km to produce a given image length. (Thus an image is equi-dimensional at 185 x 185 km.) These same frame dimensions hold for the Landsat Thematic Mapper (TM) images, discussed later.

The orbits of the Landsats are termed "paths" and the location of individual images along these paths are fixed by a row system. This Worldwide Reference System is described in some detail at this U.S. Geological Survey web site.

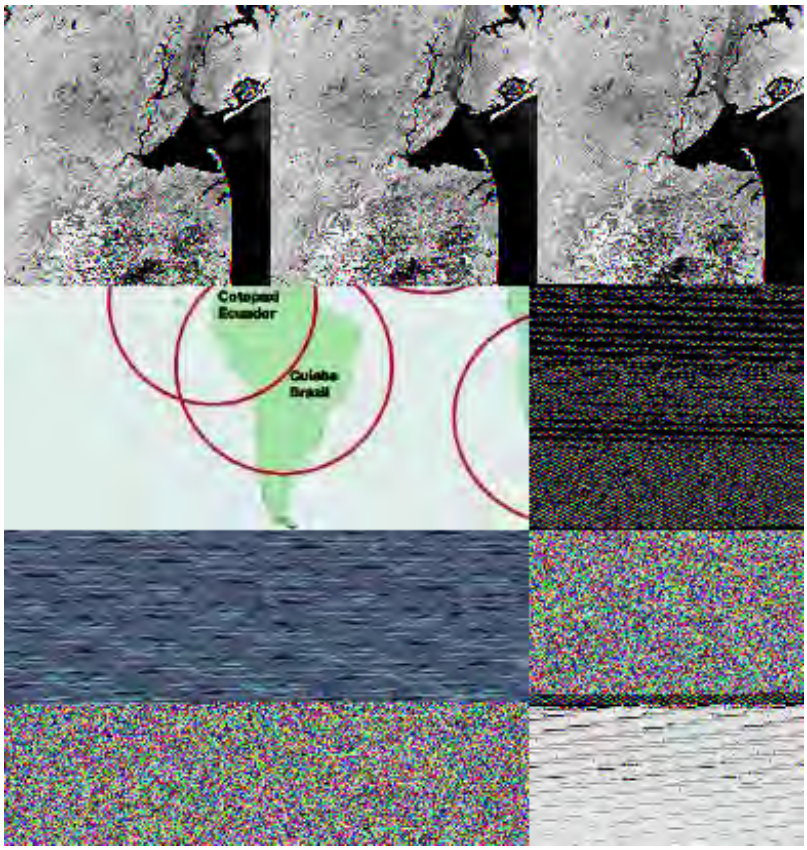
I-19 Approximately how many square miles are enclosed are enclosed in a Landsat MSS frame? Square kilometers? Acres? (Note: there are about 640 acres in a square mile.) ANSWER

All the Landsats follow a near-polar orbit (inclined about 98° to the equator; passing within 8° of the poles) and are sun-synchronous, meaning the orbit precesses about Earth at the same angular rate as Earth revolves about the Sun. Thus, it crosses the equator (traveling from North to South) each day between 9:30 and 10:00 A.M., local time. Landsats 1-3 make 14 full orbits (each successive one displaced 2875 km [1785 miles] to the west) each day (three over the U.S.), and after 252 orbits in 18 days, they repeat their previous ground tracks. Landsats 4, 5, and 7, from a lower altitude (705 km [438 miles]), cover the same ground track again every 16 days, after 233 orbits. It takes about 11,000 scenes to fully image the entire Earth's land surface (except for polar regions).

Since the first Landsats, their success and the obvious value of the information they send back have prompted many countries to set up direct-readout receiving stations so as to obtain the raw data in near real time without depending on the NASA processing centers. These stations are charted by NASA and must agree to distribute the data to users in their region. Here is a map of currently active stations:

In the U.S., the primary receiving station for Landsats 1-3 was at Goddard Space Flight Center. Later Landsats were operated by different organizations and another major receiving station was established at the United States Geological Survey's EROS Data Center in Sioux Falls, S.D. Processing at these places include reformatting of the raw data that involves orbital, geometric, and radiometric corrections. For years, the output data were distributed on 9-track Computer Compatible Tapes (CCTs) but today data for Landsat and the EOS satellites are put on 8 mm, 4 mm, DLT, CD-ROM and 3 1/2 disk storage units.

The first three Landsats used two sensor systems: the Return Beam Vidicon (RBV) and the Multi-Spectral Scanner (MSS). The RBV consisted of three TV-like cameras which used color filters to provide multispectral images in EM bands centered in the blue-green, yellow-red, and red-IR. This sensor failed early on the ERTS-1 and never came into routine use, although it flew again on Landsat-2. Below is an example of the red-IR band covering northern New Jersey and including the New York City region.

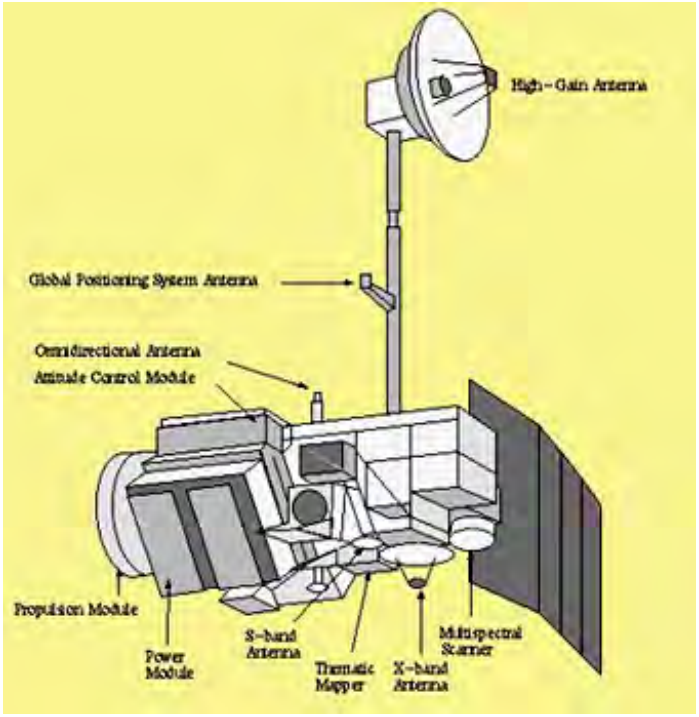


Landsat-3 carried a four-camera RBV array, with each being a panchromatic (0.505 - 0.750 µm)

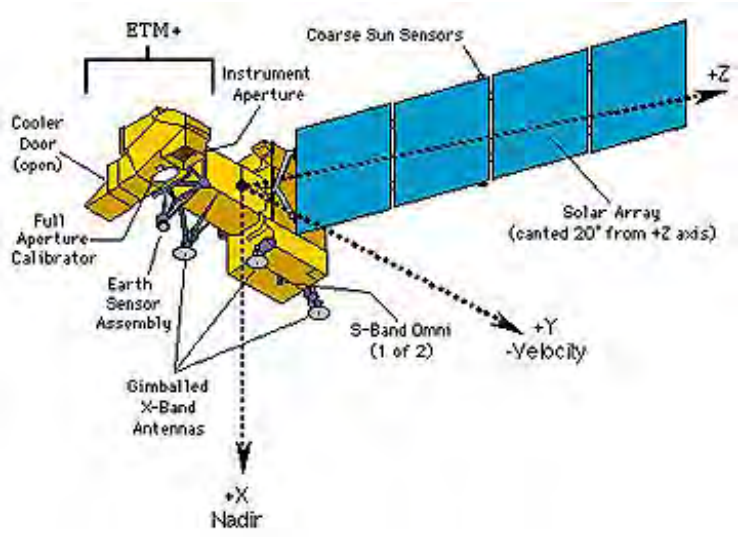
imager, that provided four contiguous images at 30 m (98ft) resolution; each image comprises approximately one-quarter of a full Landsat MSS scene). This Landsat-3 RBV scene shows Cape Canaveral, Florida where the Space Shuttle launches.

Because the RBV system seems to be redundant relative to the MSS, it was not included on Landsats after Landsat-3.

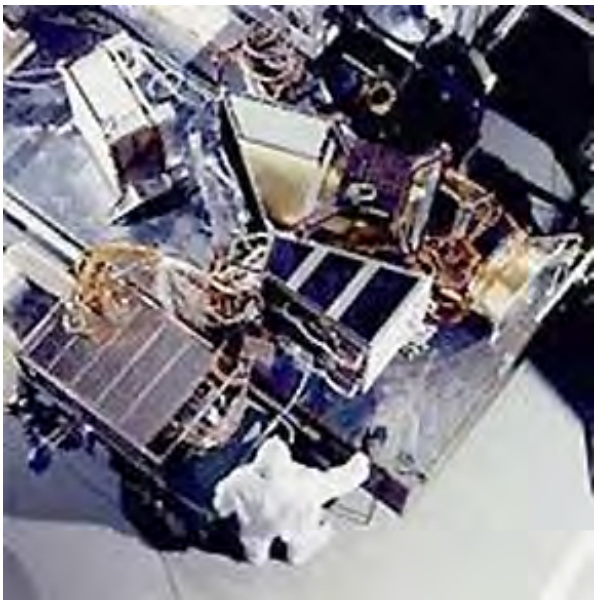
Landsats 4, 5 (and 6) carried, in addition to the MSS, a new instrument called the Thematic Mapper. Here are an artist's sketch of the Landsat 4 spacecraft and a drawing with labeled components, which point to the notable differences from the previous Landsats:



Landsat 7 mounts only a single payload, the Enhanced Thematic Mapper (ETM+). The ETM+ includes not only the 7 TM bands (the thermal band [6] has 60 m resolution; all others achieve 30 meter ground resolution), but also a panchromatic band that is capable of 15 m resolution. This artist's rendition shows that the spacecraft differs in general construction from the previous "birds" (a term of familiarity used by the launch team).



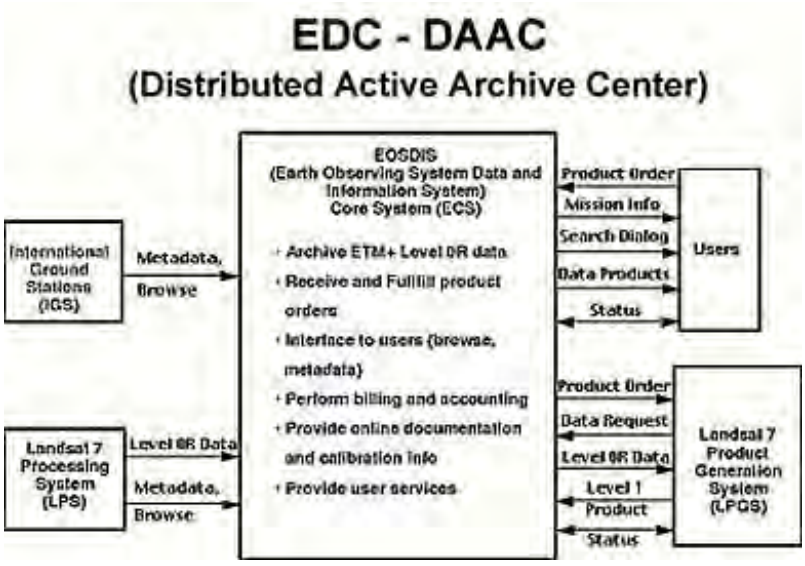
Compare this view of Landsat 7 in its assembly room with the view of Landsat 1 at the top of the page:



As of this writing Landsats-5 and 7 are still operational and returning data.

These and other earth-observing satellites send copious amounts of data to various receiving stations each day (one estimate puts it this way: The data returned over one complete 16 day repeat cycle would be enough to fill 21 Encyclopedia Brittanicas). Processing of raw data into formats preferred by the user community usually is done at the station site. The original receiving station and processing facility was at NASA's Goddard Space Flight Center, outside Greenbelt, MD. The prime facility for current data is now the LP DAAC (Land Processes Distributed Active Archive Center), a part of the U.S. Geological Survey's EROS Data Center, east of Sioux Falls, South Dakota. Its DAAC Internet home pageis worth a visit. Older Landsat data are there in part but also at other facilities.

The subject of primary data processing, archiving and distributing of Landsat material is extensive and diverse, and is deemed beyond the scope of this Tutorial. But, just to hint at the complexity, glance at this illustration - a flow chart for Landsat-7 data handling at the EROS DAAC:



The EOSDIS system now includes data distribution from Terra and Aqua, and other satellites.

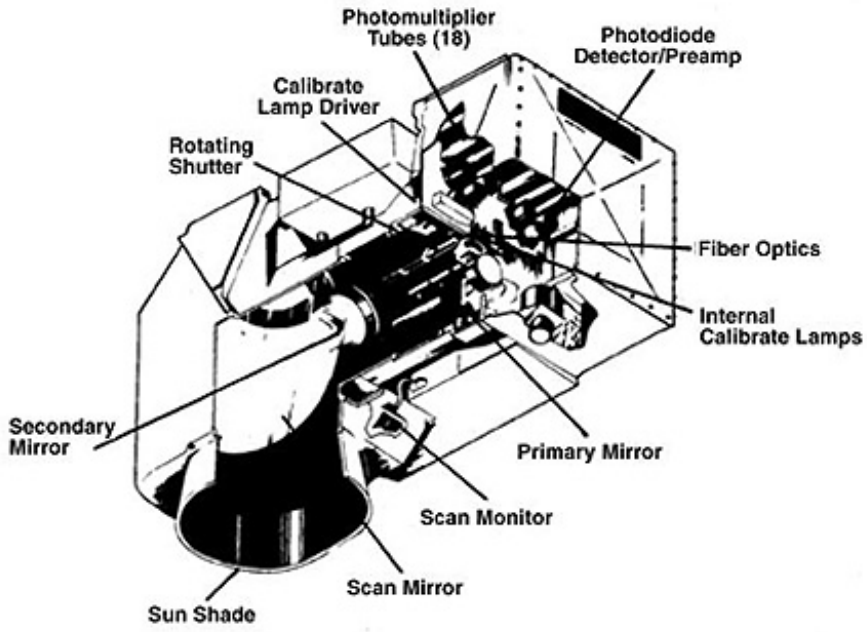
For the intrepid, a visit to the Landsat program history site developed at NASA's Ames Research Center will enlighten as to the checkered but distinguished history of this premier group of satellites.

Primary Author: Nicholas M. Short, Sr.

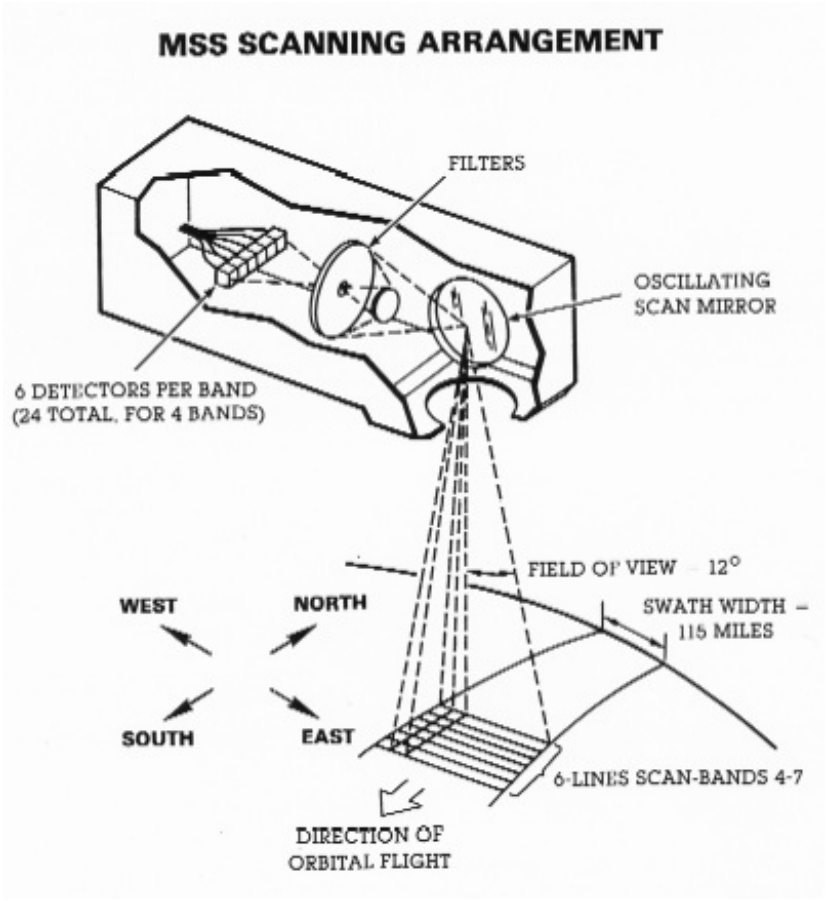
The Multispectral Scanner (MSS) has been the "workhorse" instrument on the first 5 Landsats. Its design, parameters, and operational mode are described here in some detail. This includes a first look at how the signals it creates by scanning the Earth's surface can be played back to produce a monitor display or photo image.

History of Remote Sensing: Landsat's Multi-Spectral Scanner (MSS)

The MSS instrument has operated on the first five Landsat spacecraft. Although the basics of scanning spectroradiometric sensors were reviewed earlier in this Section, because of MSS's important role in these missions which extended over 31 years some of this information is repeated and expanded on this page. This is a drawing of this venerable instrument, built by the Hughes Aircraft Corp. of Santa Barbara, CA:

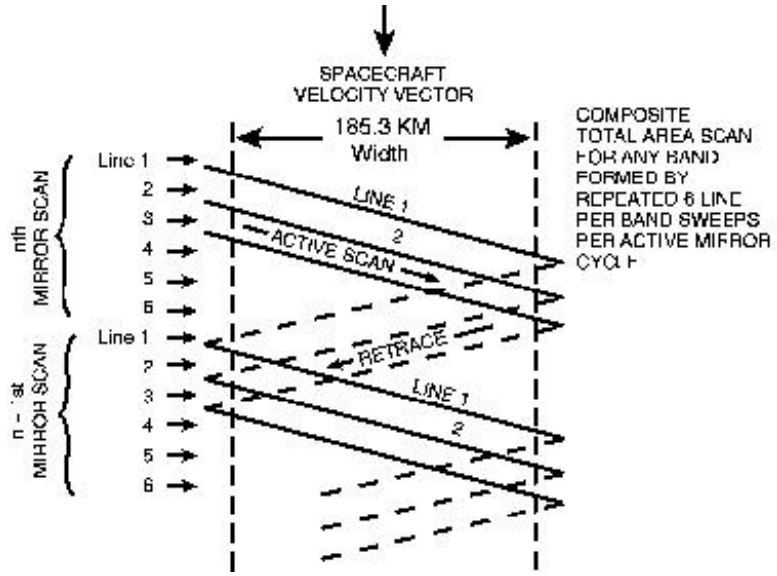


Here is a simplified model of the MSS showing how it scans the ground:



The MSS gathers light through a ground-pointing telescope (not shown). The scan mirror oscillates (1 cycle every 33 milliseconds) over an angular displacement of ± 2.89 degrees that is perpendicular to the orbital track. In the sideward (lateral) scan, the mirror covers an angle of 11.56 degrees (Angular Field of View or AFOV) that from an orbital altitude of 917 km (about ~570 miles) encompasses a swath length across the orbital track of 185 km (115 miles). During a near-instantaneous forward movement of the spacecraft (direction of orbital flight), which takes about 16 milliseconds, the mirror as it sweeps laterally (across track) is also covering a ground strip of about ~ 474 m (1554 ft) from one side of the track to the other. Said another way, this means that in the time it took to oscillate laterally, the spacecraft has advanced 454 m relative to its ground track.

Light reflected from the surface (and atmosphere) as gathered by this scan passes through an optical lens train, during which its beam is split (divided) so as to pass through 4 bandpass filters that produce images in spectral bands at **MSS 4** = 0.5 - 0.6 µm (green), **MSS 5** = 0.6 - 0.7 µm (red), **MSS 6** = 0.7 - 0.8 µm (photo-IR), and **MSS 7** = 0.8 - 1.1 µm (near-IR). (The band numbering begins with 4 because bands 1-3 were assigned to the RBV sensor.) The radiation is carried by fiber optics to six electronic detectors for each band. Bands 4 through 6 used photomultiplier tubes as detectors, and Band 7 used silicon photodiodes. Light through each filter reaches its set of six electronic detectors (24 in all, for the 4 bands) that subdivide the across-track scan into 6 parallel lines, each equivalent to a ground width of 79 m (259 ft; taken together the width covered is 6 x 259 = 1554 ft, the number stated above). The mirror movement rate (nominally, its instantaneous scan moves across the ground being imaged at a rate of 6.8 m/µsec along a scan line) is such that, at the orbital speed of 26,611 kph (16,525 mph), after the return oscillation during which no photons are collected, the next lateral swing produces a new across-track path of 6 lines (79 x 6 = 474 m) just overlapping the previous group of 6 lines. This is illustrated below:

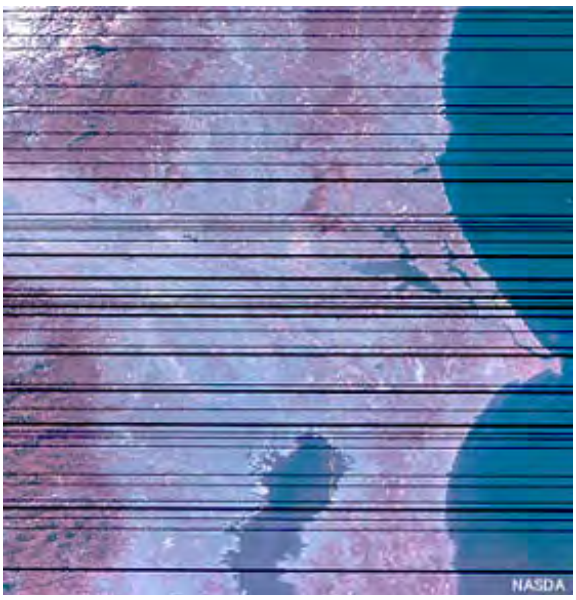


Note that the individual scan lines in this diagram are slanted relative to lines across track perpendicular to the track boundaries. These are valid traces with respect to the ground, since as the mirror moves sidewards the spacecraft is moving ever forward so that each successive moment in the scan finds its ground target (represented by the pixel) being slightly forward of the previous moment's view.

Perhaps the reader wonders, in examining the previous illustration, about the fact that no data are acquired during the return swing (other than looking then at a light source within the sensor whose known radiometric output helps to calibrate the external reflectances). Textbooks describing the operation of the MSS tend to ignore this conundrum. Here is a simple explanation: During the forward swing (for Landsat, with its southward advancing path, from west to east), the six lines are created. The reverse oscillation leads to no data. But look at the diagram. Line 1 moves left, as do the other lines. When the next forward swing occurs, line 1 is now just below line 6. The very movement of the scanned across-track scene is such that, for the scan rate involved, the next acquisition of the 6 lines starts with where the previous line 6 is located relative to the new line 1.

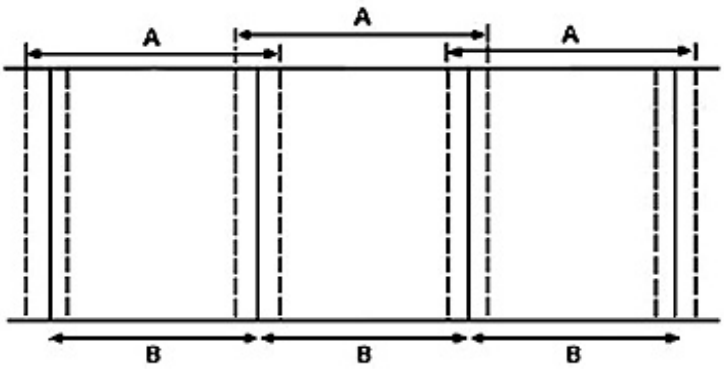
I-20: **Individual scan lines are commonly visible (stand out) in a printed or displayed image of a Landsat scene. Can you think of a technical reason why these may be seen?**
ANSWER

This question suggests that anomalous scan lines are found in individual scenes. These are usually shown in black (meaning no data received). They are frequent in this Landsat-5 image (a scanner timing failure occurred that presents this problem).



At each detector, the incoming light (photons) from the target frees electrons in numbers proportional to the number of photons striking the detector. These electrons move as a continuous current that passes through a counting system, which measures the quantity of electrons released (thus, indicating radiation intensity) during each nine microsecond detection interval. Over that minute time interval (called the *dwell time*) the advancing mirror picks up light coming from a lateral ground distance of 79 m (259 ft). The detector thus images a two-dimensional, instantaneous field of view (IFOV, usually expressed in steradians, which denotes the solid angle that subtends a spherical surface and, in scanning, connotes the tiny area, within the total area being scanned), that at any instant amounts to 0.087 mrad (milliradian, or 0.0573°). At Landsat's orbital altitude of 917 km, the effective resolving power of the instrument is based on the 79 x 79 m^2 ground equivalent (pixel) dimensions described above. Each detector is then cleared of its charge so as to produce the next batch of electrons generated from the next IFOV photon inputs during the mirror's lateral sweep. As the scanning continues through the full lateral sweep the set of all IFOV pixels in the line are rapidly read in succession. The onboard computer converts this succession of analog signals (voltages) into digital values which the onboard communication system telemeters (sends) to Earth by radio.

For each band detector, the electronic signal from this IFOV results in a single digital value (called its DN or digital number, which, for the MSS, can range from 0 - 255 [2^8]). The value relates to the proportionally averaged reflectances from all materials within the each IFOV. Since the mix of objects on the ground constantly changes, the DN numbers vary from one IFOV to the next. Each IFOV is represented in a b & w image as a tiny point of uniform gray-level tone, the pixel described earlier in this Section, whose brightness is determined by its DN value. In a Landsat MSS band image, owing to a sampling rate (every nine microseconds) effect in which there is some overlap between successive spatial intervals on the ground, a pixel has an effective ground-equivalent dimension of 79 x 57 m (259 x 187 ft) but contains the reflectances of the full 79 m^2 actually viewed. This "peculiarity", illustrated in this diagram, needs further explanation:



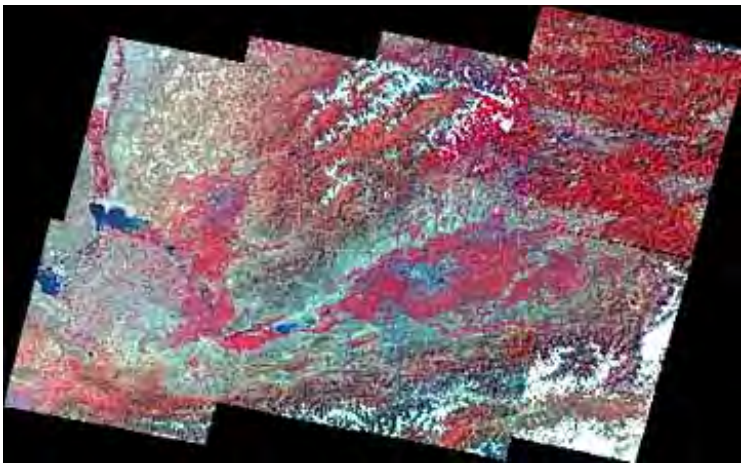
The wider rectangle (a square for the MSS), which can be designated the Ground Resolution Cell (GRC) size, is established by the IFOV of the scanner. But because the sampling interval Δt is finite, i.e., cannot be zero, the previous and next cells contribute parts of the their represented ground scene that overlap (by 11.5 m) into each individual GRC rectangle/square. This requires removal (by resampling) of the overlap effects leading to a new resolution cell that represents the actual Ground Sampled Distance (GSD). Thus, for the Landsat MSS the GRD of 79 x 79 m becomes a GSD of 79 x 57 m. Each GSD contains all the radiation sent from the GRC for each band spectral interval, integrated into single values expressed by the DNs.

The average number of pixels within a full scan line (representing 185 km) across the orbital track is 3240 (185 km/ 0.057 km). In order to image an equi-dimensional square scene, which requires 185 km of down track coverage, the average total number of lines to do this is set at 2340 (185 km/0.079 km). Each band image therefore consists of approximately (again variable) 7,581,600 (3240 x 2340) pixels - a lot to handle during computer processing, over 30 million pixels when the 4 bands are considered. The number of pixels actually does change somewhat owing

to satellite attitude (shifts in orientation (wobble) called pitch, roll, and yaw) and instrument performance that lead to slight variations in the pixel total.

Image producers can use the continuous stream of pixel values to drive an electronic device that generates a uninterrupted light beam of varying intensity, which sweeps systematically over film to produce a b & w photo image. The resulting tone variations on the image are proportional to the DNs in the array. In a different process, we can display the pixels generated from these sampling intervals as an image of each band by storing their DN values sequentially in an electronic signal array. We can then project this array line by line on to a TV monitor, and get an image made of light-sensitive spots (also called pixels) of varying brightnesses. Or, these DNs can be handled numerically, not to produce images, but to be inputs for data analysis programs (such as scene classifications as described in Section 1).

Little has been said on this page about the appearance of a basic Landsat image. This is deferred until the next page, although Landsat images have already been shown in the Overview. We comment here about several general characteristics of a Landsat image. Look at this illustration:



In this scene covering part of the southern Asian country of Kyrgyzstan, one sees four strips of imagery, joined to produce a mosaic (considered in Section 7). Each strip is a swath from one orbital pass. Notice that its imagery has a slanted appearance (assuming the vertical is true north). Why this slant: because as the MSS sensor looks down at Earth while moving in its orbit, the Earth's surface underneath has been moving from west to east owing to the planet's general rotation. At the 99° inclination (relative to longitude) of Landsat's orbit, the orbit plane precesses about the Earth at the same angular rate that the Earth moves about the Sun. In the image, each successive line slips slightly westward. The accumulation of these progressive offsets results in a figure that, for any individual scene, would be a parallelgram with inclined sides. This in part also accounts for the orbital inclination of the spacecraft, to compensate somewhat for that rotation.

We said above that a Landsat scene is produced by arbitrarily stopping it at 185 km from top to bottom. But that trimming need not happen. one can continue to produce a more continuous swath scene that is elongate in the direction of orbit. In producing an individual parallelogram scene, it is customary to have about 10% of the top consisting of the bottom 10% of the previous (more northern) scene; there is a similar 10% continuance on the bottom. There is also overlap on the left-right margins; this is called sidelap. Its amount varies with latitude. At the equator, the sidelap ranges from 7% (MSS) to 14% (TM); the amount increases going towards the poles so that at high latitudes the sidelap between adjacent scenes can be as high as 80+%. This north-south and east-west overlap allows a crude form of stereo-viewing to be possible within these margins. In practice, this stereo capability is seldom utilized.

Primary Author: Nicholas M. Short, Sr.

Here we consider how the 4 Landsat MSS bands produce multispectral images of the same area in the eastern U.S. (which you are asked to identify before reading a detailed description of its geography) as scanned in October 1972. These images are displayed side by side and then beneath them is a false color composite of the same area plus a bit more to the south. The idea behind change detection is introduced by looking at a MSS Band 7 image of the area obtained in April 1978 (Spring) and comparing it with the Fall '72 Band 7 image.

History of Remote Sensing: A Landsat Image

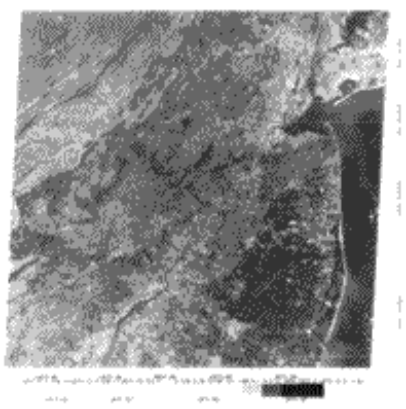
The picture below is a Landsat full image, from October 1972, shown both as individual bands and as a false color composite. We display below a succession of MSS bands 4 through 7 (bands 1-3 were assigned to the RBV), comprising images of a very well known region in the eastern U.S. Beneath these four images, is a false color composite, made from bands 4, 5, and 7, of this scene, here extended southward to include the southern tip of a peninsula that includes the town of Cape May (a clue).

Note 1: By convention, Landsat and most other satellite system images are normally oriented with North towards the top. However, because of the ninety-nine degree orbital inclination, the north direction is not vertical but is a few degrees inclined relative to the perpendicular to the top and bottom margins of the printed image.

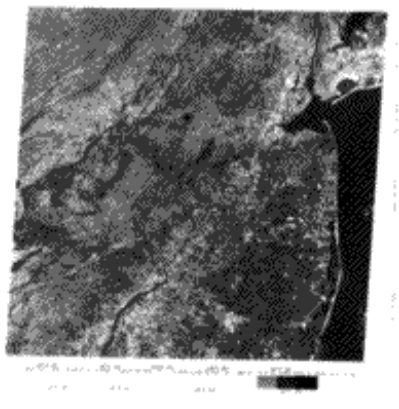
Note 2: The lettering at the bottom of each image (probably not readable on your screen) is the standard annotation placed on Landsat images produced at NASA, EROS, and most commercial facilities. The information recorded, from left to right, includes the calendar date of acquisition, the latitude-longitude coordinates of the scene principal and nadir points, the sensor type, the elevation and azimuth positions of the Sun, and the Scene (Frame) identification number (I.D.) starting with the particular Landsat (1 through 5) and ending with the specific band (or band combination if in color).

Before you look at the second paragraph beneath this image set, we challenge you to try to identify what geographic area is shown here.

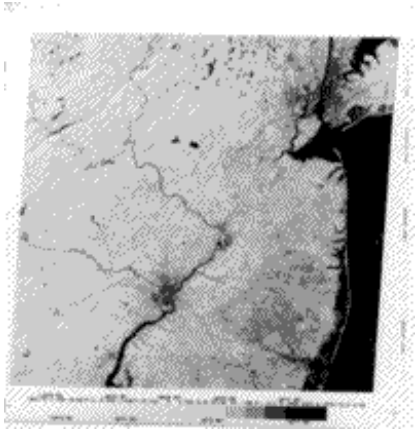
Landsat MSS October 10, 1972



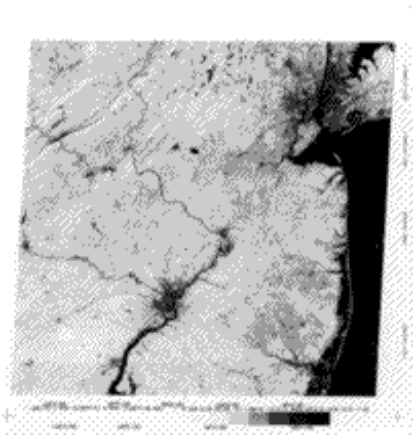
MSS Band 4



MSS Band 5

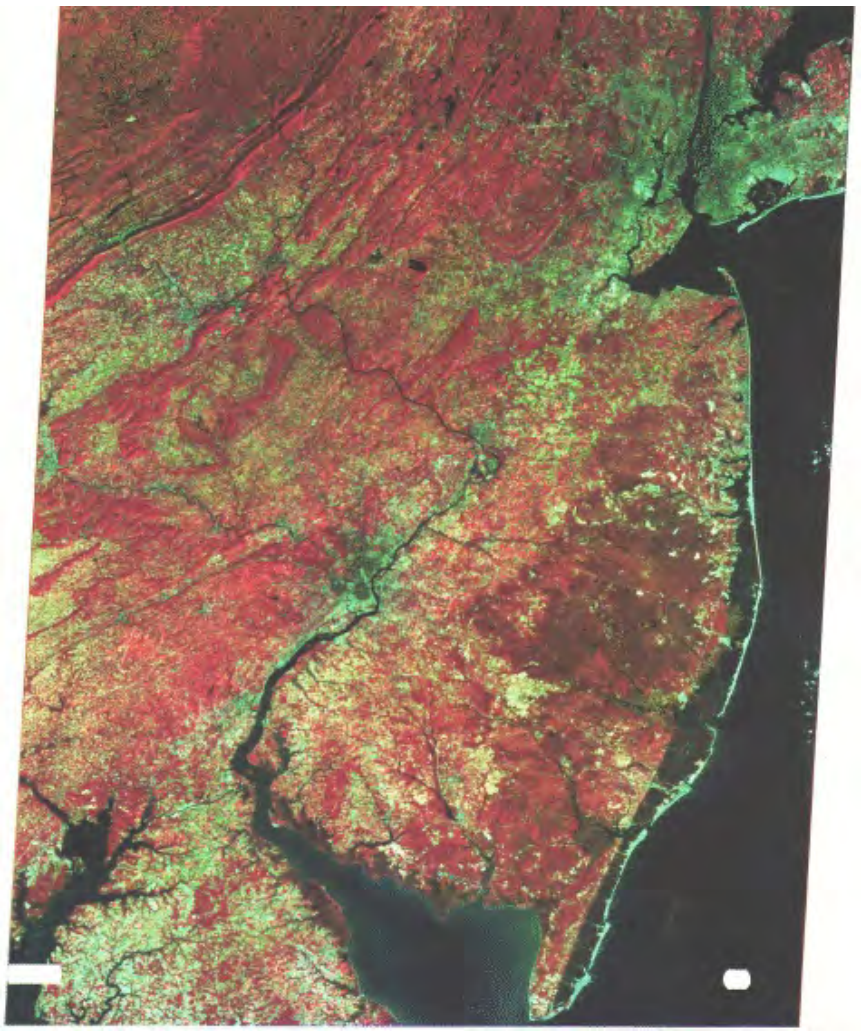


MSS Band 6



MSS Band 7

These bands plus a small part of the next image to the south joined (mosaicked) to them can be combined to make a standard false color composite using the three Band combinations as shown below the image:



MSS Band 4 = Blue

MSS Band 5 = Green

MSS Band 7 = Red

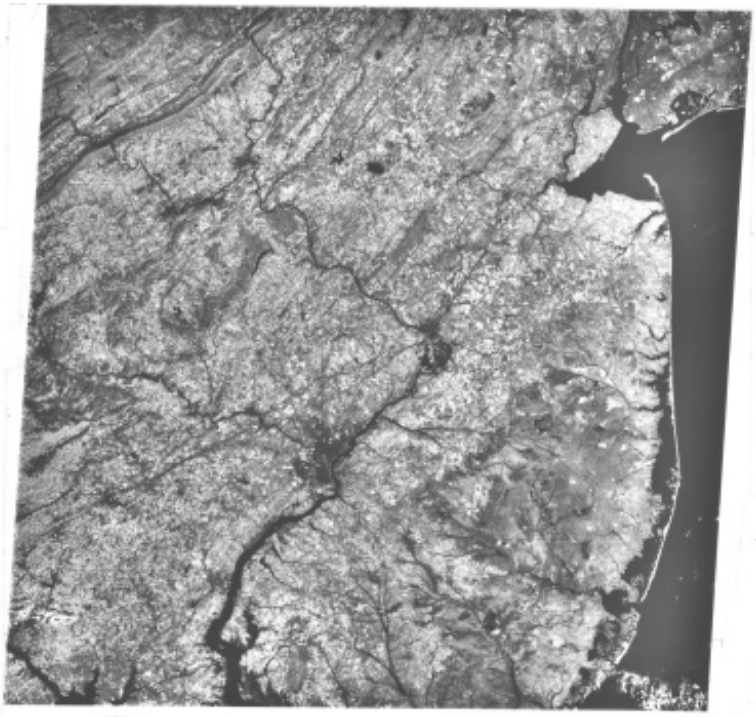
If you correctly identified the scene, then in the individual bands, you saw much of New Jersey along with New York City and the west end of Long Island in the upper right corner and Philadelphia at just left of the image center. The color composite extends the coverage to the northern Delmarva Peninsula, flanked by the northern Chesapeake Bay (bottom left) and Delaware Bay (bottom center) and Cape May (bottom right). These two urban regions, along New Jersey cities along the Hudson River, appear in light to medium gray-blue tones in Band 5 (red). In Band 7 (IR), the central areas of these metropolitan complexes are in dark tones, owing largely to the prevalence of asphalt streets and dark (usually asphalt) roofs, together with few trees and little other vegetation. In Bands 6 and 7, the urban scene contrasts with the lighter tones associated with high reflectance vegetation in the countryside. Water is dark in Bands 6 and 7 but lighter in Bands 4 and 5, in part because of silt and other sediments (more reflective). In contrast, among the brightest features in both individual-band and color composite scenes are the sandy beaches and soils comprising the ocean side of the barrier islands lining the New Jersey coast.

Vegetation in this October 10, 1972 scene is still actively growing (mostly green in natural color, but some trees are beginning to get their autumn colors), so its spectral distribution indicates an overall brightness in the Band 6 and 7 images. We can compare the dark tones of the fold belt ridges in the upper left in Band 5 with their corresponding light tones (from high reflectances related to tree leaves; this ridges are heavily forested) in Band 7. Surfaces dominated by vegetation are shown in several shades of red in the false color composite. Harvested (fallow) fields appear in blue tones, similar to those characterizing the cities. The large, darker area in New Jersey east of Philadelphia is the Pine Barrens, marked by evergreens that grow well in the sandy soils (Note the appearance of these trees in each of the four bands; in color, they are reddish brown; evergreens are bright but not as much as deciduous trees).

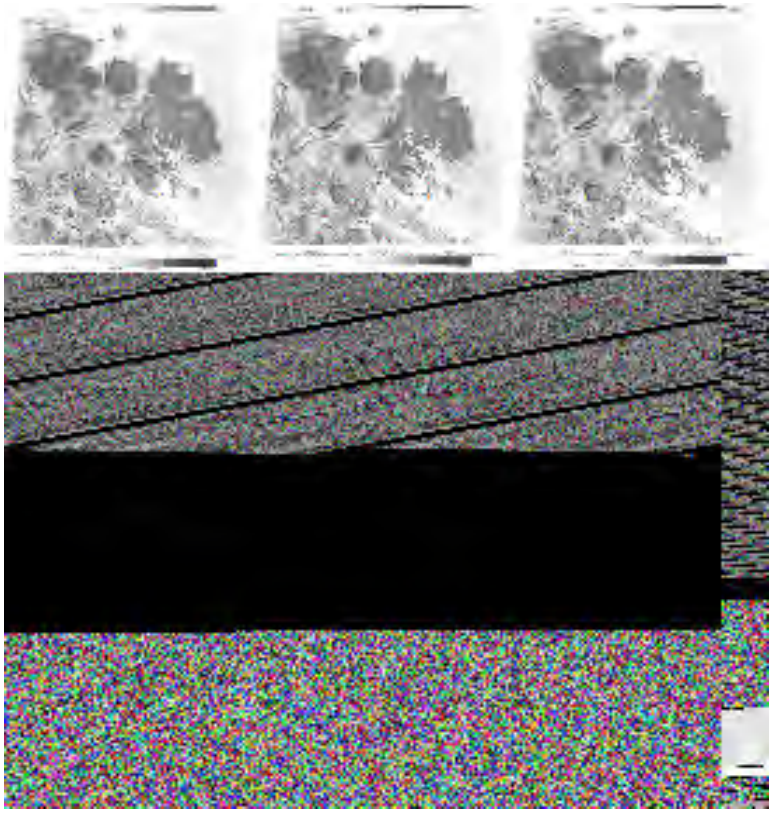
I-21: **Try your hand at picking out other major landmarks in the scene (use an atlas for help in recognizing their locations). ANSWER**

I-22: **Although it may be a "pain" scrolling up and down repeatedly, we suggest you pick out various features in each of the four MSS band images and note how they appear, in terms of gray levels (and shapes) in each band; in other words, compare. Also, describe the overall appearance of each band relative to the others. Do this mentally, or write it down if you wish. This is a worthwhile task, as it will give you a "feel" for the general nature of each of the four bands that represent the continuum of spectral intervals from 0.4 to 1.1 µm. ANSWER**

As a preview of *change detection*, compare the October 1972, Band 7 image above with the one below, which was taken on April 18, 1978. Because it's a time in the Spring in which trees and other vegetation in the eastern U.S. have not yet leafed out there are especially dark ridges whose tones result from low rock and soil reflectances that dominate the radiances because of the absence of leaves). This seasonal effect results in a reduction in brightness over much of the areas in the lowlands where the fields remain fallow prior to emergence of growing vegetation. Many of the brighter areas in this April image do correlate with some field crops that were planted earlier; small bright patches in the Pine Barrens are sand pits - highly reflective in all bands.



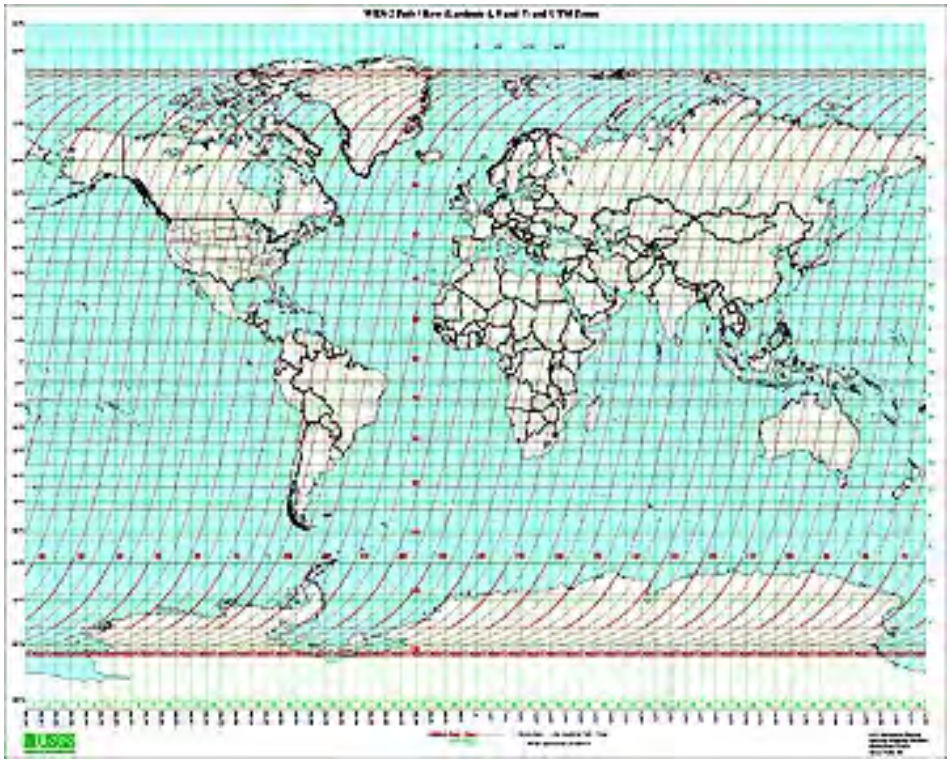
The reader may have noticed that in the two black and white Landsat-1 images near the top of this page there is a gray bar and some annotation on the bottom. This is a common format used in most renditions of early Landsat images and in some more recent versions. (Throughout the Tutorial this information is usually cropped off to reduce image size; it is almost unreadable on Internet pages anyway). We show another image (North Africa) here which contains all normally peripheral information included in photo prints of Landsat imagery:



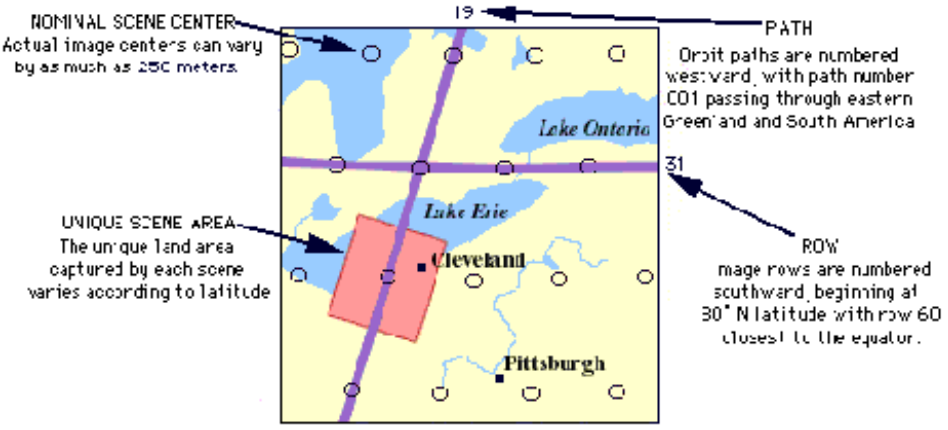
The annotation contains valuable information. Here is an enlargement of the annotation in the above image:

The first writing on the very left of the annotation row is the date of scene acquisition. This is followed by two groups of letters/numbers that define: a. the latitude-longitude coordinates of the principal point (format center), then b. latitude-longitude coordinates of scene nadir point (nadir refers to a line from the platform vertical to the ground). Next to the right is the identification of sensor (MSS) and Band (7). Then comes the elevation of the Sun (angle above horizon) and the Sun's location (azimuth defining Sun's position above the horizon geographically relative to true North), both at the time of scene acquisition. The next string of alphanumerics contains specialized information including spacecraft heading (189°), then (G), location of the receiving station (here, Goldstone in California), followed by I-N-D-IL for I = full size, N = normal processing, D = definitive accuracy for image center, 1L = 1 for linear mode for spacecraft data transmission (2 would be compressed mode), and L for low gain (H would be high gain) signal amplification level. NASA ERTS signifies that this product was made by NASA's ERTS (Landsat) satellite. The farthest right alphanumerics, E-1106-09183-7 specifies ERTS (E) followed by the mission number (1 for ERTS-1), plus 106 for the number of days after launch, -09183, the hour (09), minute (18), and second (7), all local time of observation (for scene center). The last, 01, is known as the regeneration number. Below the gray bar are three numbers that mark the longitudes (here E = east of Greenich, London) in degrees and minutes as these intersect the bottom of the image frame; these numbers appear at the top again, shifted left because of the orbital inclination.

In ordering Landsat imagery, information in the annotation can be helpful. There is another aid that users often use instead. For the Landsat program, the Worldwide Reference System (WRS) was established. This is a grid with near up and down lines called the Path trace and horizontal lines the Row coordinates. We show this map of the coordinate system (the red lines mark every 10 path or row units), realizing that on the Internet, the numbers and words appear too small to be readable:



Look at a small part of the Path-Row system that crosses this map:



For Landsats 1-3, there are 251 Paths (001 to 251), the same number of orbits needed by these spacecraft to image the Earth in one 18-day cycle. There are 120 Rows, with 1 starting near the North Pole (80°47'), 60 coinciding with the Equator, and 120 in high south polar latitude. Path 001 is set at an equatorial crossing point of 64°36'

WRS-1 refers to the coordinate set used for the first three Landsat. WRS-2 denotes use by Landsats 4, 5, and 7. Because of the lower orbital altitude, this map contains 233 Paths. When the above Ohio scene is ordered, its location can be specified as Path = 19, Row = 31. Then a time-of-year date is selected, and, usually, a limit (in percent) of cloud cover (anything below that is acceptable).

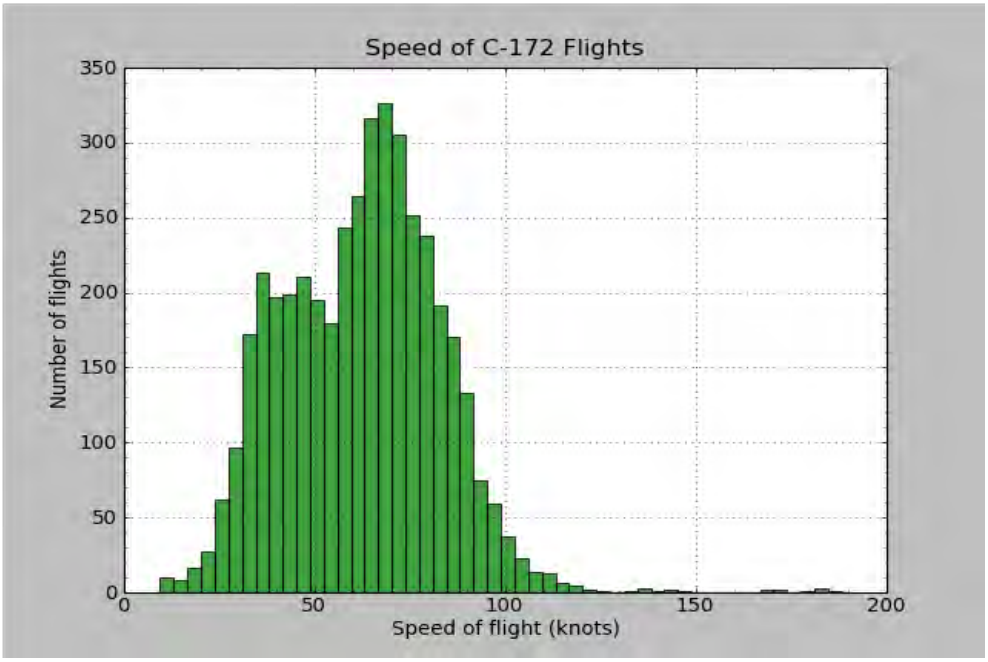
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

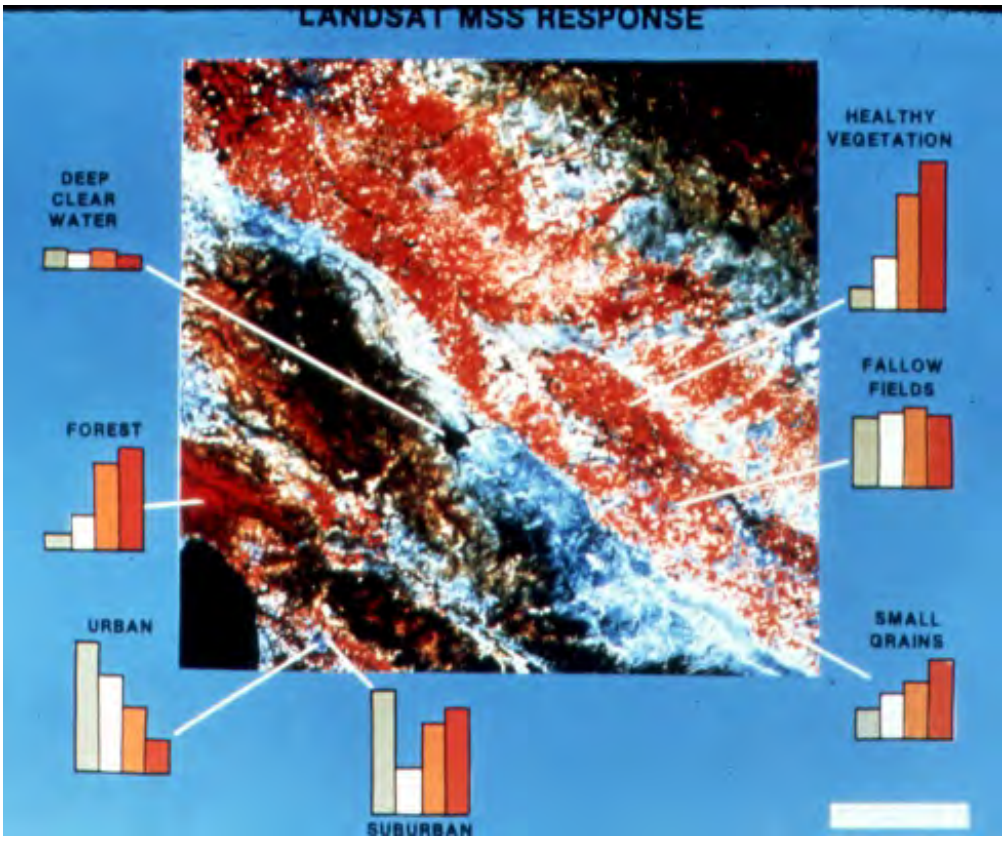
The Landsat MSS gathers radiation over spectral band widths that integrate radiation over relatively broad intervals (0.1 and 0.3 µm). Thus, instead of the spectral signatures that continuously measure spectral response in very narrow intervals, the MSS data when plotted produce histogram-like bars that are rough approximations of the signature curves. For different classes, these still yield separable patterns that facilitate identification of the materials involved. This is illustrated for a scene in central California. A table is included (by link) that specifies relative black and white gray levels and distinctive colors by which some common materials can be identified.

History of Remote Sensing: MSS Histograms

One way in which to display variation is the histogram. It simply expresses how x varies as a function of way. A typical histogram is shown below: it shows the flight history of the C-172 aircraft, namely the number of flights (ordinate) that achieved a given average speed (abscissa):



The MSS data consisting of spectral intensity for each band can be plotted as histograms. These simulate rather crude spectral signatures. Bands 4, 5, and 6 each have a bandwidth of 0.1 µm; and band 7's width is 0.3 µm. We represent band responses by bars in a histogram-like plot, in which the height of the bar signifies the relative reflectances averaged for all wavelengths within the bandwidth interval. Several crude spectral signatures (as bar histograms) are shown here, as derived from this Landsat image of a section of the U.S. West Coast.



The scene is the first color composite made from ERTS-1 digital data. This includes Monterey Bay (lower left corner), the Coast Ranges below San Francisco, the Sacramento (Great) Valley, and the western slopes of the Sierra Nevada. Notice that the pattern of each histogram set, for a particular type of surface cover, differs from the others. Thus, each class of material has a distinctive signature, roughly approximated by the 4 bars, that sets it apart. To simplify the plot, we made the width of band 7 (bar on the right) equal to the other three. Urban is most reflective in bands 4 and 5; suburban is strong in bands 4 and 7 (the latter is tied to the influence of live vegetation in lawn grass and landscaping trees). The signatures for forest and healthy croplands are similar, but the heights of the bars for bands 6 and 7 are greater for the crops. The bar heights for small grains and fallow fields are similar, but the response for bands 4 and 5 is just a bit higher than for 6 and 7. The two very dark areas in the Coast Range and in the Sierra Nevada (not shown as histograms) result from the predominance of conifers.

Next, refer to the table on the next page in which we present some criteria for using combinations of band gray tones or colors and their patterns to identify land-cover categories. Most categories are described by their MSS Bands 5 and 7, and sometimes 4, response. Again, apply these to recognize examples of any such categories in the California scene. Become familiar with this table, because we challenge you to practice this approach whenever you examine various Landsat scenes (and imagery from the other earth-observing satellites) in this Tutorial.

I-23: **Print out this table; it will serve as a handy reference for other scenes in this Tutorial. Scroll back to both the New Jersey and California scenes above and apply the table criteria (gray levels; color) to features/classes you have identified earlier to ascertain the validity of these criteria.** **ANSWER**

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Best MSS Bands for Identifying Surface Features

Item	Category	Best Bands	Salient Characteristics
a.	Clear Water	7	Black tone in black and white and color.
b.	Silty Water	4,7	Dark in 7; bluish in color.
c.	Nonforested Coastal Wetlands	7	Dark gray tone between black water and light gray land; blocky pinks, reds, blues, blacks.
d.	Deciduous Forests	5,7	Very dark tone in 5, light in 7; dark red.
e.	Coniferous Forest	5,7	Mottled medium to dark gray in 7, very dark in 5; brownish-red and subdued tone in color,
f.	Defoliated Forest	5,7	Lighter tone in 5, darker in 7 and grayish to brownish-red in color, relative to normal vegetation.
g.	Mixed Forest	4,7	Combination of blotchy gray tones; mottled pinks, reds, and brownish-red.
h.	Grasslands (in growth)	5,7	Light tone in black and white; pinkish-red.
i.	Croplands and Pasture	5,7	Medium gray in 5, light in 7, pinkish to moderate red in color depending on growth stage.
j.	Moist Ground	7	Irregular darker gray tones (broad);darker colors.
k.	Soils-bare Rock-Fallow Fields	4,5,7	Depends on surface composition and extent of vegetative cover. If barren or exposed, may be brighter in 4 and 5 than in 7, Red soils and red rock in shades of yellow; gray soil and rock dark bluish; rock outcrops associated with large land forms and structure.
1.	Faults and Fractures	5,7	Linear (straight to curved), often discontinuous; interrupts topography; sometimes vegetated.
m.	Sand and Beaches	4,5	Bright in all bands; white, bluish, to light buff.
n.	Stripped Land-Pits and Quarries	4,5	Similar to beaches – usually not near large water bodies; often mottled, depending on reclamation.
o.	Urban Areas: Commercial Industrial	5,7	Usually light toned in 5, dark in 7, mottled bluish-gray with whitish and reddish specks.
p.	Urban Areas: Residential	5,7	Mottled gray, with street patterns visible; pinkish to reddish.
q.	Transportation	5,7	Linear patterns, dirt and concrete roads light, in 5; asphalt dark in 7.

Primary Author: Nicholas M. Short, Sr.

Terrestrial geophysics involves measuring particles, fields, and radiation associated with both the solid (internal) Earth, its surface (including the hydrosphere [oceans]), and the external environment (mainly fields surrounding Earth and extending outward hundreds to thousands of kilometers. We will treat the atmosphere and hydrosphere in Section 14. On these next three pages, the subjects are the Earth's force fields (and trapped particles), its magnetic and gravitational properties, movements of the crust and lithosphere, and earthquake seismology. On this first page, the topics include the discovery of the Van Allen Belts, the ionosphere and auroras, and the development of the terrestrial magnetic field, including the implications of magnetic anomalies. Emphasis is on those satellites whose missions were primarily to gather geophysical data, from which new insights and interpretations have ensued. Also touched briefly are several of the key ideas behind the new paradigms underlying the geological model of the Earth as expressed by plate tectonics, continental drift, sea floor spreading, and subduction.

Geophysical Remote Sensing

NOTE: Many of the ideas and examples presented on this and the following two pages require some background knowledge of the basics of Geology. If desired at this time, you can review a subsection in Section 2 that covers Fundamentals of Geology.

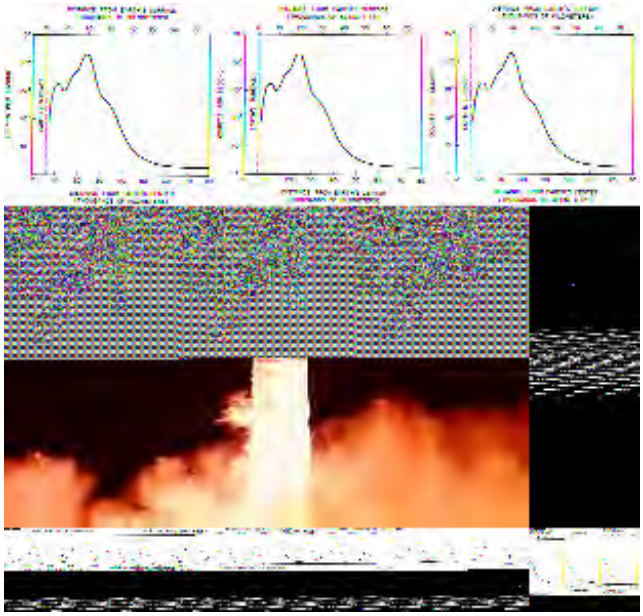
We begin this survey of Geophysical Remote Sensing with the first American triumph in space: the launch and operation of Explorer 1, on January 31, 1958, four months after the Soviet Union launched Sputnik I (see the Overview). The history-making launch of Explorer 1 riding on a Redstone-Jupiter C rocket is shown here:

A good summary of this and several of the other early U.S. launches is found at this Explorer site.

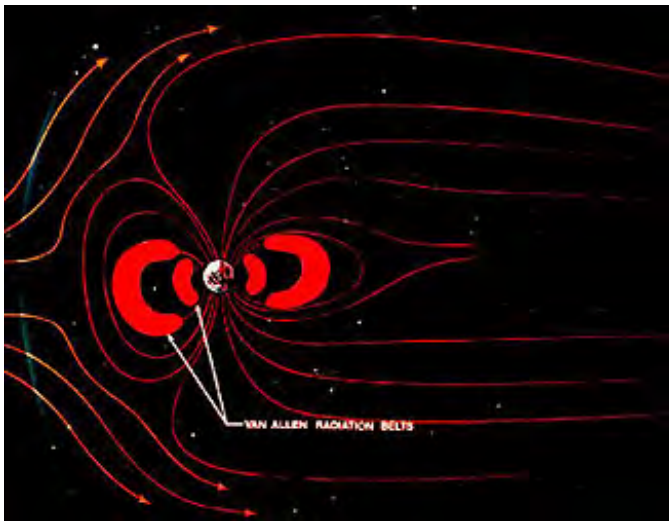
Instead of the more conventional-looking satellites (most with thrusters to adjust their orbits) that we shall depict throughout this Tutorial, Explorer 1 itself was a small rocket with its own engine. It was thus actually the fourth stage of the rocket assembly. Here is a picture (against a black background to simulate the darkness of space) of Explorer 1:

This 4th stage was 2.03 m (6.67 ft) long and 15 cm (6 inches) wide. Its bottom contained fuel for getting the entire assembly into orbit. The payload - scientific instruments including a cosmic ray detector, micrometeorite package, and several temperature sensors - was mounted in the cross-lattice frame. The top was protected by a nose cone shroud. Once in orbit, this assembly rotated rapidly. Explorer 1 functioned successfully for 3 months until its batteries gave out.

Explorer 1 made a major discovery about the interaction between the Earth's geomagnetic field and incoming charged particles from the Sun and extra-solar space. A team headed by Dr. James Van Allen put together an instrument package centered on a Geiger counter suited to measuring variations in particle radiation. As it orbited, Explorer 1 repeatedly passed through a region in which the amount of radiation increased significantly. When Pioneer 3 was placed in a highly elliptical orbit in December 1958, it was determined that there were actually two peaks in counts as shown in this plot.

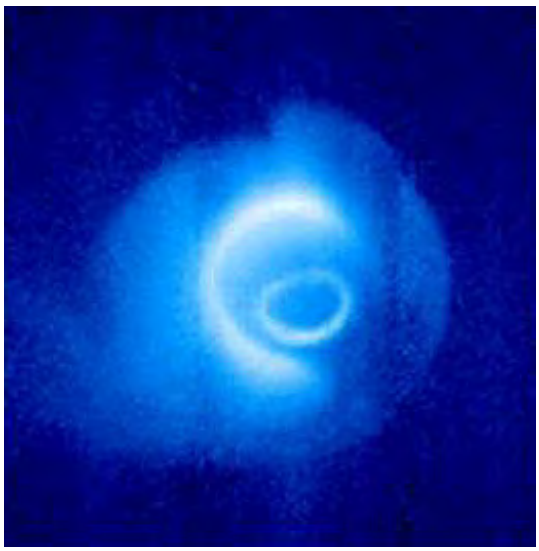


These zones of increased intensity proved to be due to trapped particles held in the Earth's magnetic field. They were concentrated in two torus or doughnut-shaped zones that were named the Inner and Outer Van Allen Belts. This illustration is a two-dimensional cutaway sketch of streamlines representing solar wind particles as they passed through Earth's magnetic field.



The Inner V.A. Belt reaches its maximum intensity at 5000 km (3000 miles) but extends inward to about 1000 km (600 miles) The Outer Belt starts at 1500 km (9300 miles) and peaks at 22000 km (15500 miles). That belt is dominated by trapped electrons from the solar wind; the Inner Belt is marked by protons brought in mainly as cosmic rays. Two important conclusions from the discovery of the Van Allen Belts: 1) they have for eons provided protection to terrestrial surface dwellers from these potentially devastating particle bombardments - a fact critical to the successful development of life on Earth; and 2) both spacecraft and humans would need to be shielded effectively when passing through the Belts (as would astronauts journeying beyond Earth for extended periods).

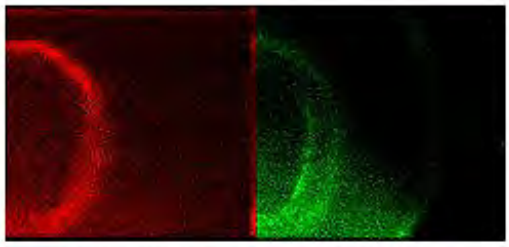
In recent years, some spectacular images of the particle-trapping fields around the entire Earth have been taken by sensors on satellites placed in highly elliptical orbits or at Lagrangian points (where Earth-Sun gravitational forces balance out) or enroute to the Sun or other distant bodies. In the top image, the Earth's radiation field in the excited atmosphere is seen in the Extreme Ultraviolet (EUV) part of the spectrum by the IMAGE (Imager for Magnetopause to Aurora Global Exploration) satellite; the shape of this field is determined largely by the Earth's magnetic field (note the trailing pattern, related to the geomagnetic tail [see below]). The arc-shaped inner belt is now called the plasmasphere. The distribution of the trapped solar particles in this image is strongly controlled by the Earth's magnetosphere (see below).



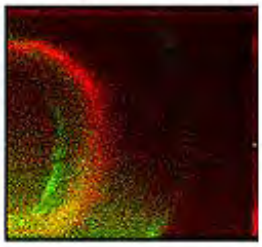
The Van Allen Belts become much weaker above 75°N and 75°S. This allows more particles to reach the the upper atmosphere and collide with oxygen, nitrogen and argon atoms in the air to generate ions that in their excited states give off constantly moving, colorful, wavy displays known as the *Aurora Borealis* in the northern hemisphere and the *Aurora Australis* in the southern hemisphere. This geophysical phenomenon occurs mainly at the higher latitudes but sometimes extends below 40°; it is seen most frequently around 70° latitudes. The image below

shows the aurora as photographed on the ground in Colorado.

A similar scene appears below, along with a panel of images showing the influence of modes of excitation:



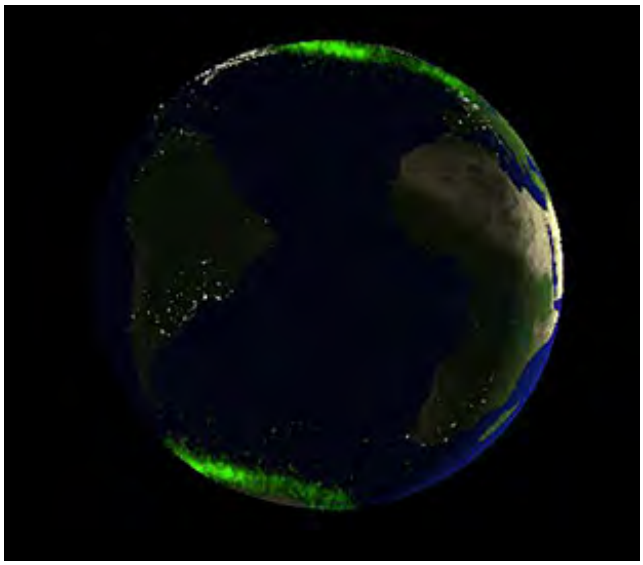
proton aurora electron aurora



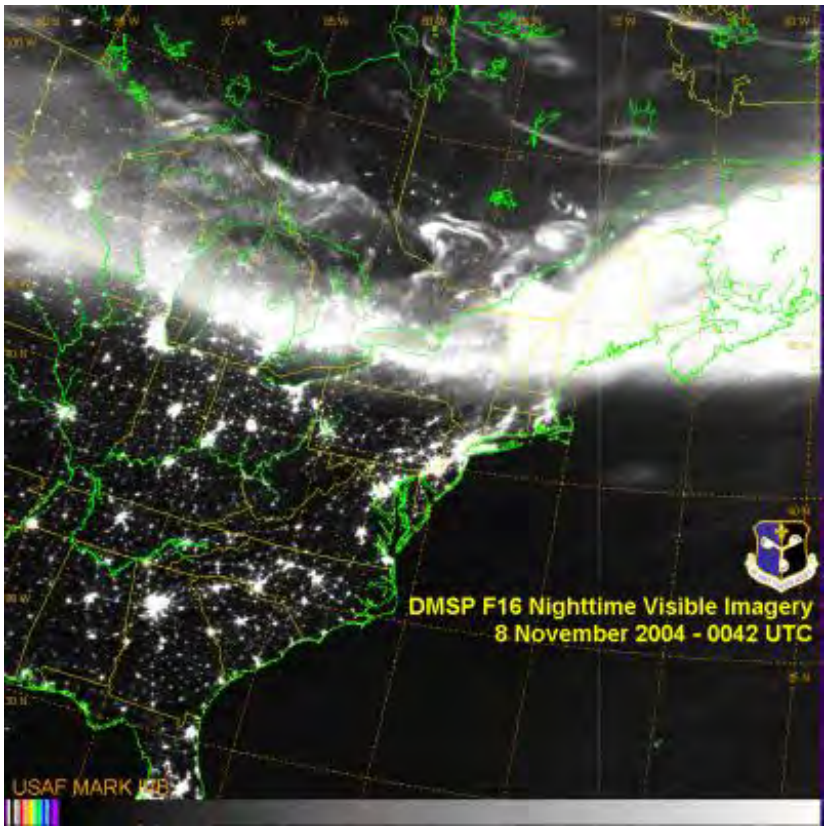
proton + electron aurora

Here is a view of the Aurora Borealis as taken by astronauts on the Space Shuttle.

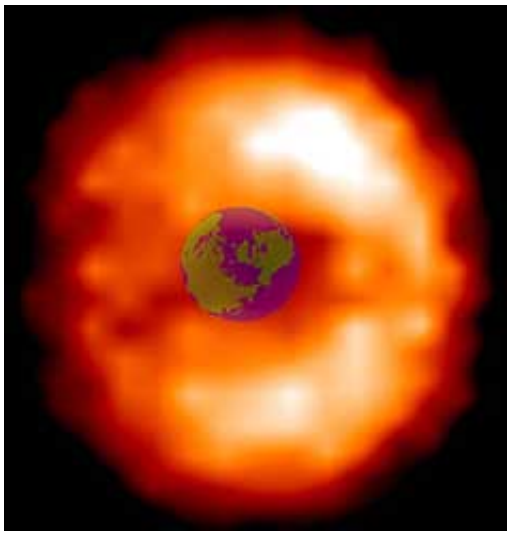
The auroras can occur simultaneously in both the northern and southern hemispheres, as depicted in this image from IMAGE at a distance of 7.2 earth radii.

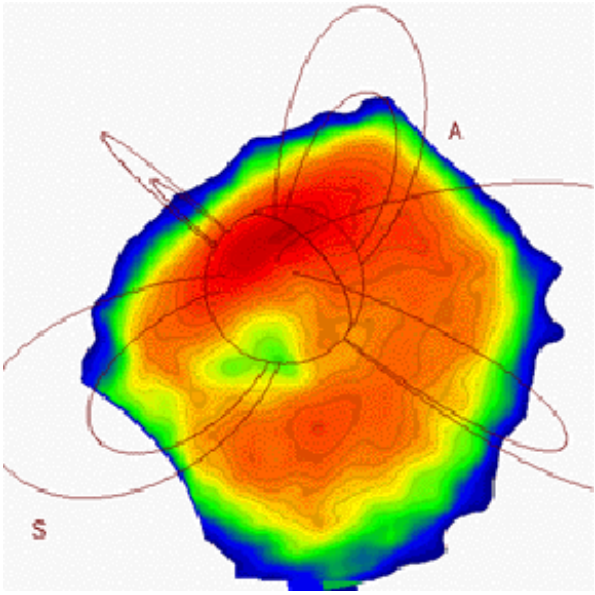


A superb aurora occurred over North America on November 7 and 8, 2004. Here is a ground scene of that event and a view from space in which the military DMSP satellite captured this aurora as a white band across the northern U.S.-southern Canada.

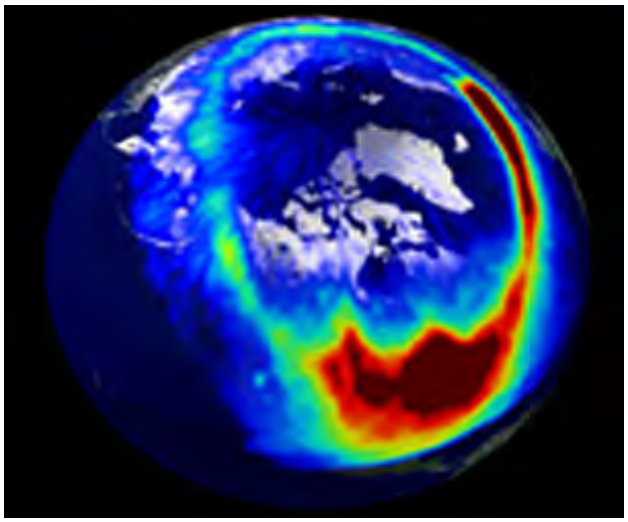


The High Energy Neutral Atom instrument on IMAGE sends back images of auroras like this one, from which intensity variations can be extracted:

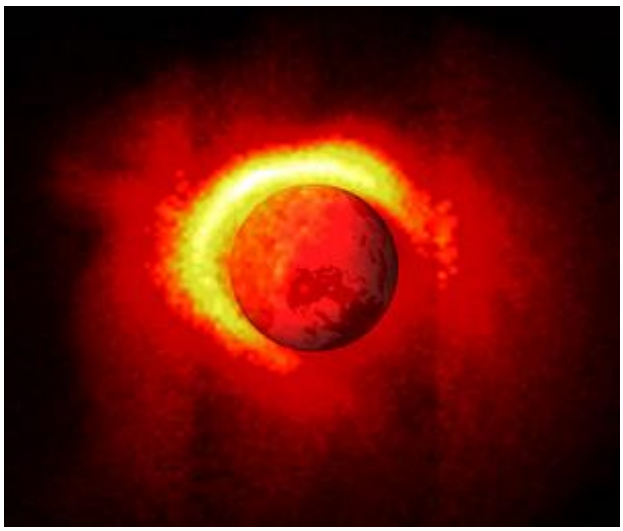




An example of symbiosis in remote sensing is this: A satellite named SOHO whose job is to monitor solar activity reported intense solar storms in mid-July. This was predicted to produce a spectacular Aurora Borealis flare-up visible as far south as the northern U.S. Unfavorable viewing conditions caused rather subdued displays. But, a NASA satellite called Polar (launched in 1996) designed to monitor just such phenomena produced this view in the visible from space:

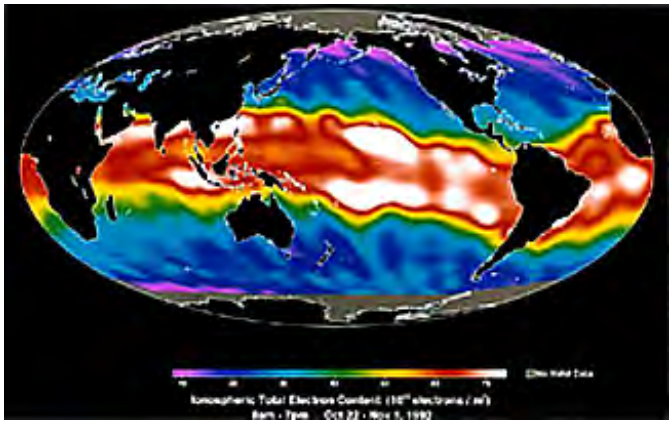


Other excitations produce displays seen in the Ultraviolet. Here is an IMAGE EUV view of excited helium extending well out beyond the Earth's atmosphere:

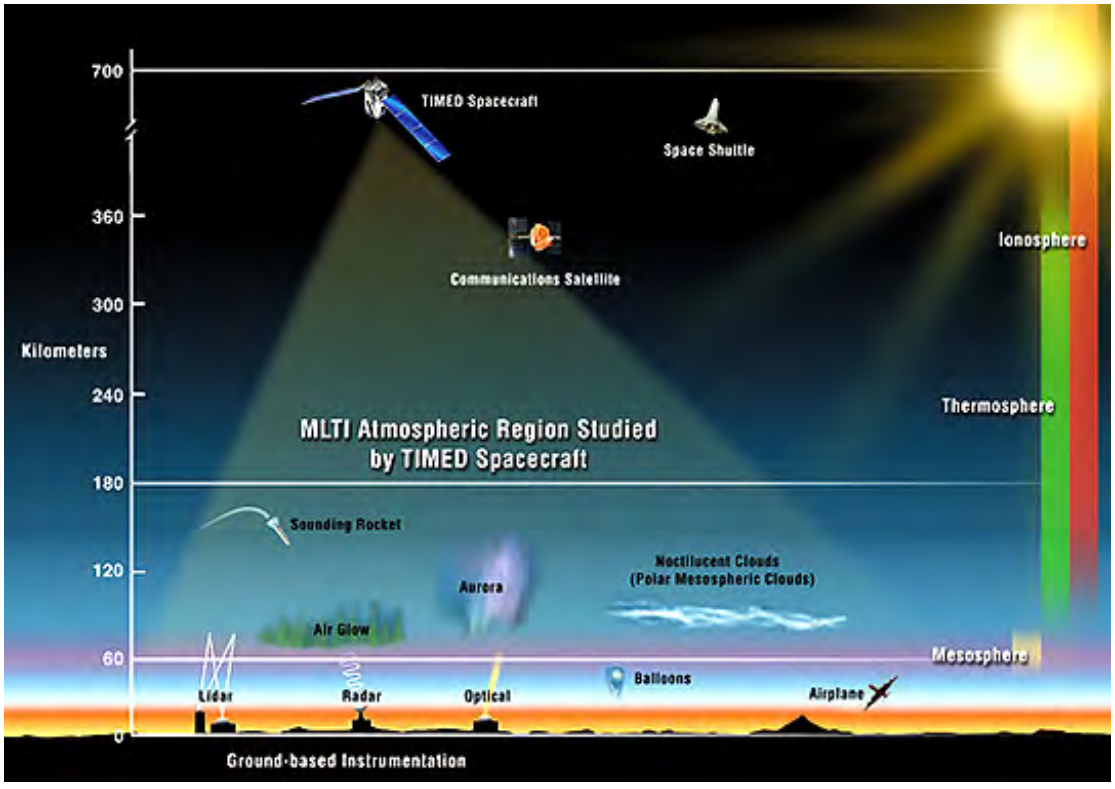


Many satellites have been placed in space by various nations to get global scale data on the

Earth's magnetosphere and its ionosphere (a globe-circling band located largely between 80-400 km [50-250 miles] above the surface, containing electrons stripped by gamma rays from nitrogen and oxygen). Here is a plot of the global distribution of the ionosphere, measured by the Jason-1 satellite (pages 8-7 and 14-12) whose prime mission is to measure Sea Surface Heights.



To better understand both the solar influence on that part of the (thin) atmosphere and the possible feedback influence of Man, more integrated data on the Mesosphere and lower Thermosphere/Ionosphere, between 40 and 110 miles (60-180 km), has been sought through the launch on December 7, 2003 of the TIMED (Thermosphere-Ionosphere-Mesosphere-Energetics-Dynamics) satellite. TIMED carries four sensors designed to measure different properties in this zone. This diagram shows its general role and objectives as it orbits at 338 miles (624 km) above Earth. Data have not yet been released for inspection by the public.



Other satellites have made countless measurements and observations of the inner atmosphere (about 97% by mass of its gases [about 78% N_2, 21% O_2, 0.93% A(rgon), 0.35% CO_2]). The atmosphere tends to be homogeneous (mixed uniformly) within the bottom (near surface) 30 km (18 miles). The upper thin atmosphere extends out to about 120 km (75 miles). At greater distances outward the gases separate into shells (the heterosphere) with the innermost being molecular nitrogen, then atomic oxygen, then atomic helium, and a thick (from 3500-10000 km [2200-6000 miles]) outermost layer of atomic hydrogen. (See Section 14 for a fuller treatment of meteorological remote sensing of the lower atmosphere.)

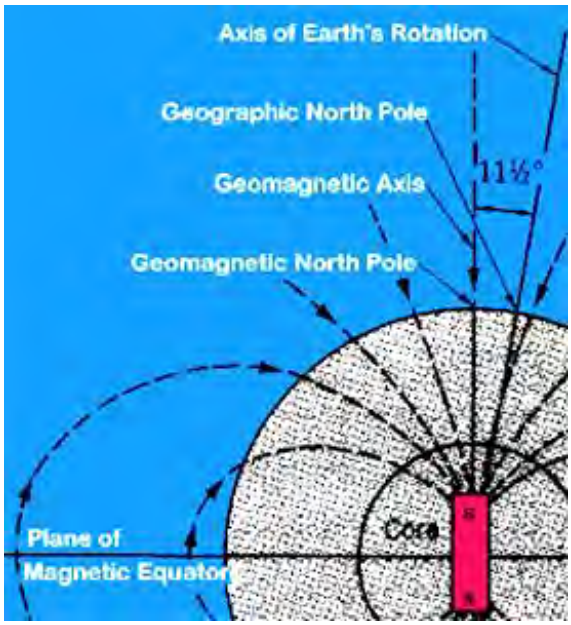
Much was learned about the Earth's natural environment above its surface - both in regards to the nature and distribution of particles and fields and to the various atmospheric zones - in the first decade of the space program. Of special note is the OGO (Orbiting Geophysical Observatory) series, of which 6 were launched in the 1960s. Three of these have nearly polar orbits and are referred to as POGOs.

Satellites have been the main tool for measuring the terrestrial geopotential fields at global scales. This includes such physical properties as regional magnetic and gravity variations or anomalies originating within the solid Earth but expressed at the surface. In particular, such properties are more difficult to measure under the oceans but data derived from satellites provide the means to determine the distribution of magnetic and gravity differences over large tracts of marine seafloor. The magnetic field varies over time but the gravity field is almost completely time invariant.

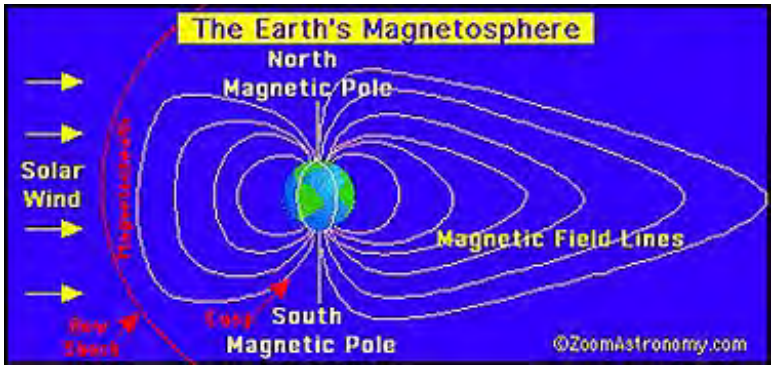
The Earth's *dipolar* geomagnetic field (first examined scientifically by W. Gilbert in 1600) results from slow motions in its fluid Outer Core (the Inner Core is very hot, under extreme pressure, but solid). The movement is powered by the Earth's rotation and by thermal gradients. Detached electrons from the Iron-Nickel core material produce electric currents that through the dynamo effect generate the magnetic field (which can be depicted by lines of force) that presently emanates from the Earth's South to its North magnetic poles. (This phenomenon of magnetic field generation was discovered by H.C. Oersted in 1819 quite serendipitously during a demonstration of electric currents to his students; the needle of a compass lying nearby moved when the current flowed, from which he deduced this cause-effect relation.)

The geomagnetic field (derivative from the lines of force data) is composed of three parts: 1) the Main Field, caused by the internal processes in the core, which accounts for more than 95% of the total field strength; 2) the External Field, resulting from processes in the ionosphere leading to a superimposed field; and 3) the Anomalous Induced Field, caused by induced (secondary) magnetism generated in iron minerals (such as magnetite and hematite) found mainly in the crust. Including in the induced field are those substances that have an inherited remanent (permanently induced) magnetism; these give rise to local variations called *anomalies* that tell geophysicists and prospectors about localized to regional concentrations of certain rock types and possible commercial minerals. The geomagnetic field strength is commonly plotted in intensity units. The basic unit is the *oersted* (1 dyne per unit pole), which is numerically equivalent to the *gauss*, the preferred term when magnetic induction is measured. A derivative unit is the γ, which is 10^{-5} gauss. In the SI units system, intensity is measured in Teslas (defined in units of Newtons/Ampere-meter) or more commonly nanoTeslas (nT; 10^{-9} Teslas); an nT is the equivalent of the γ in strength.

Although more complicated in its details, the Earth's geomagnetic field can be likened to that associated with a simple, straight bar magnet, as pictured thusly:

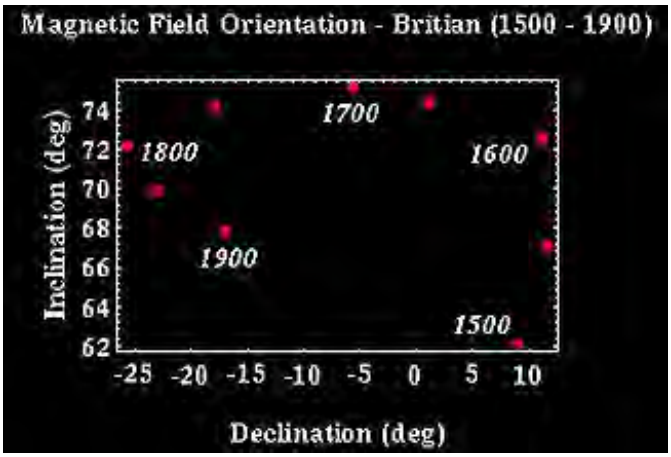


This pattern of symmetrical lines of force (the magnetic flux) is idealized to describe the concept. Actually, the main geomagnetic field is much distorted by interaction with the solar wind (a plasma of mostly charged particles spewed out from the Sun and driven outward at high speeds) which exerts a "pressure" on the field compressing it on the side facing the Sun and drawing it out in a streamlined elongation away from the Sun. In two dimensions, the modified field appears like this, as determined from satellite measurements:

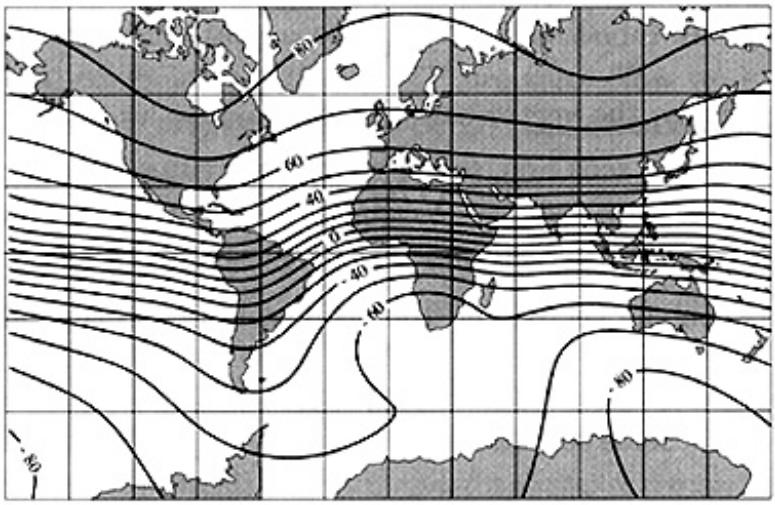


The region around Earth in which the terrestrial magnetic field is enclosed by the solar wind is termed the *geomagnetic cavity*. The outer cavity is bound by the *magnetopause*. Where the solar wind encounters the magnetosphere on the Sun-facing side, a magnetohydrodynamic *bow shock* wave is produced as much of the solar wind is deflected around the magnetopause. On that side, the region between the bow shock front and the magnetopause is called the *magnetosheath*. Behind the Earth, on its nightside, the magnetic lines of force are stretched out into the *magnetic tail* which extends for well over 150,000 km (about 100,000 miles).

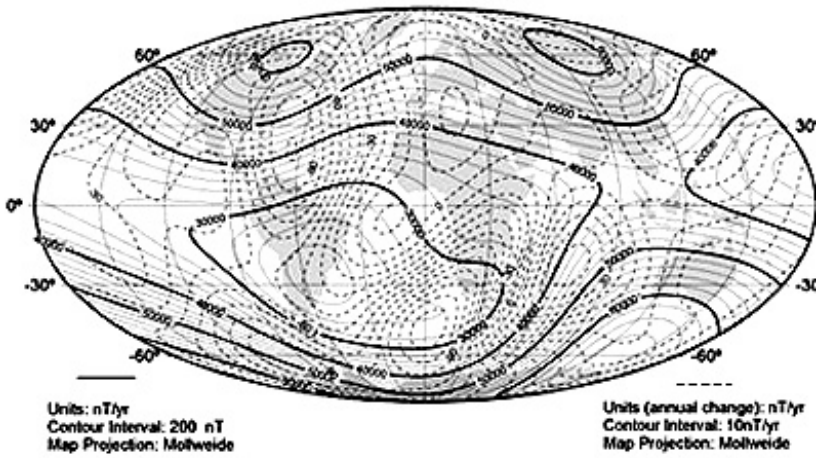
Note that the magnetic pole and the Earth's rotational pole are currently about 11° apart. Both poles wander (precess) so that this angle changes with time. The angle between true North and magnetic north, or between the geographic and magnetic meridians, is called the *magnetic declination*. The angle between the path of a magnetic line of force at some latitude and the horizontal (Earth's flat surface) is the *magnetic inclination* (it can be determined locally with a magnetic dip needle [held vertically]). The angle is zero at the magnetic equator (field line parallel to the surface, 90° at the poles, and at intermediate angles between these two reference positions. Both declination and inclination shift globally or in a region over time; long-period variations in direction and intensity are called *secular changes*. This effect is shown here for nearly four centuries of ground-based data from Great Britain:



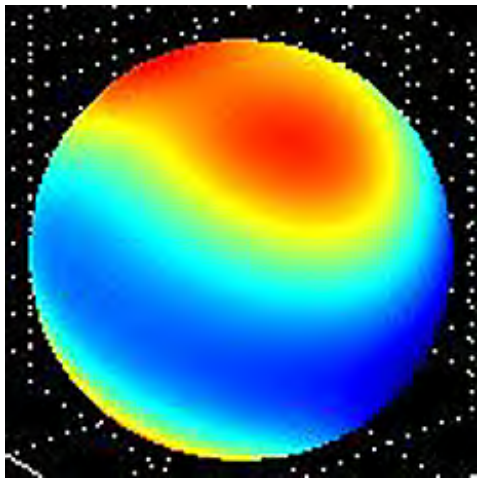
Variations in magnetic inclination, declination, total intensity, and horizontal and vertical intensity vectors can be determined from ground and space measurements. Over the years, these have become part of the geomagnetic community's data base called the International Geomagnetic Reference Field (IGRF), now updated every four years. Here is a global plot of the general geospatial distribution of magnetic inclination values for 1995.



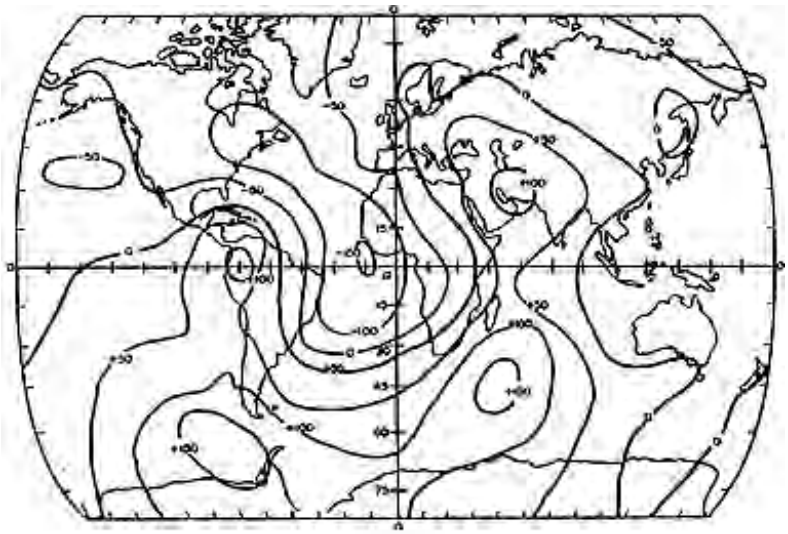
The total magnetic field as determined by satellite measurements can be shown in this global diagram, with the solid black lines representing intensity values and the dashed lines denoting typical annual variations, both plotted in units of nanoTeslas (nT).



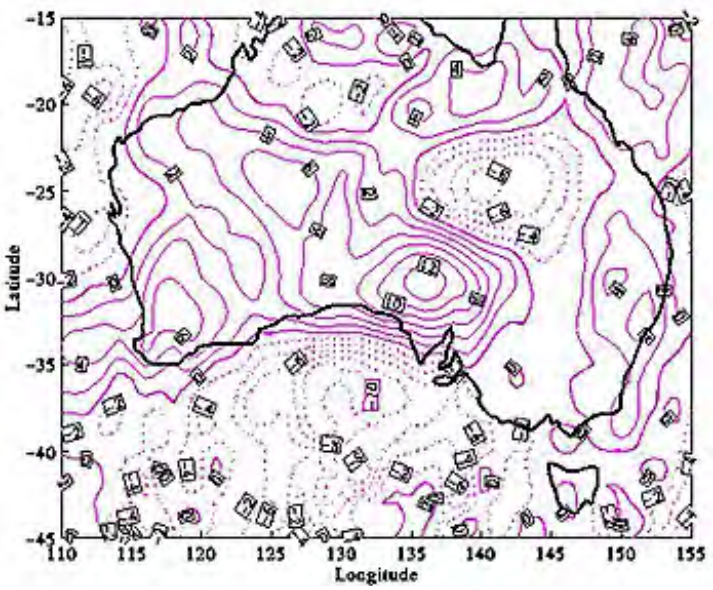
The geomagnetic measurements have confirmed and refined our knowledge of the magnitudes and distribution of global intensities. These range from about 25000 nT at the equator to 65000 nT at the poles (or from 0.25 to 0.65 gauss). This next illustration is a general depiction of this variation (low values in blue; high in red):



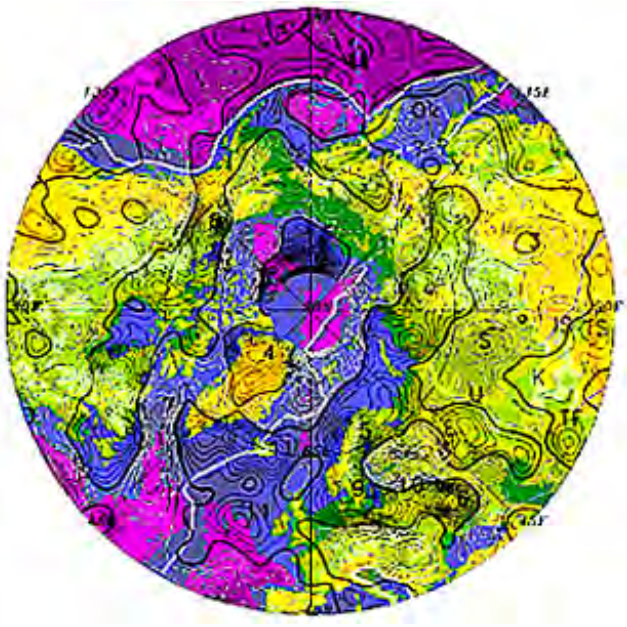
The secular variation in intensity can be as much as 150 nT per year. Here is a global map of cumulate changes during the years 1925-27 as determined from many ground stations:



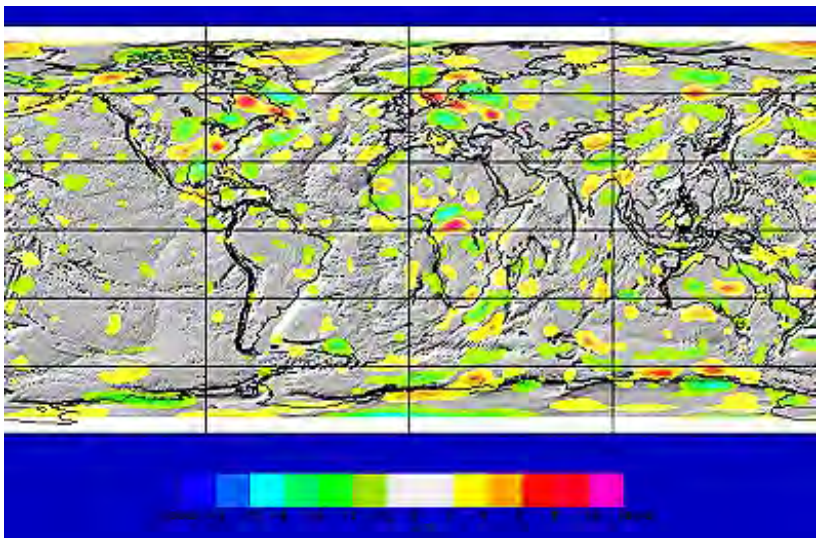
Geomagnetic measurements were among the first geophysical data acquired by early American satellites such as several Explorers, Pioneer I and V, IMP-I, and the OGO series and by such Russian satellites as the first two in the Lunik series that also went to the Moon. A major advance in surveying geomagnetic anomalies was made by Magsat, launched in 1979. Although data collection was short-lived owing to the limited duration of the mission which ended in June 1980, the spacecraft operated long enough to provide an invaluable set of measurements still being examined 20 years later. Here, for example, is a 1995-produced map of variations in intensity over most of Australia.



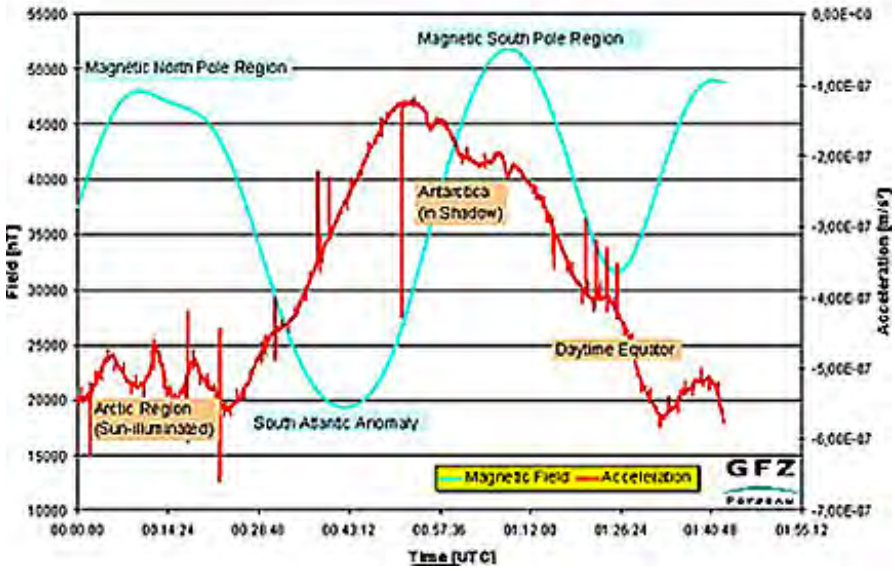
When produced as color-coded maps, Magsat-determined intensity variations are impressive in their detail. This map was made to cover the North polar region:



Major anomalies worldwide were identified by Magsat. This next map shows significant anomalies (in nT, ranging from -12 in blue to +12 in pink) plotted on a topographic/bathymetric base as a gray background with some first order surface features superimposed. One of particular interest is the Central African Anomaly (blue area low just above the red area high in the middle of Africa).



Since the Magsat benchmark mission, other satellites have, or will be, orbited to monitor and improve the geomagnetic properties measured conveniently from space. These include POGS (Polar Orbiting Geomagnetic Survey) in 1990, Oersted (a Danish mission; 1999), Champ (German; 2000), and SAC-C (Argentina; 2000). Champ, as an example, makes both magnetic and gravity (acceleration) measurements:



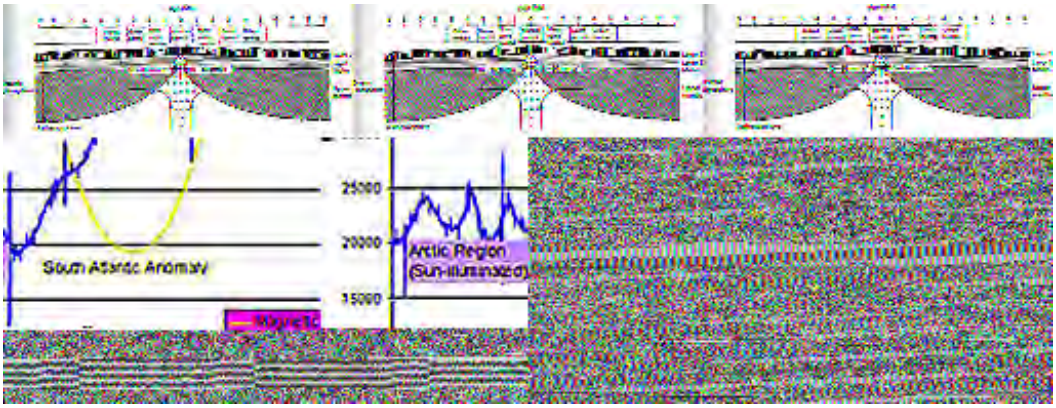
We would be remiss if the essential results of years of *paleomagnetic* studies were not included here. While satellite data support this work, they are not the prime means by which knowledge of the history of changes in the Earth's magnetic field is obtained. This has been done mainly through ground sampling of rocks collected on the continents or from the oceans' floor. This approach is not remote sensing *in strictu senso* but it does lead to reconstruction of magnetic field geometry at global scales that assumed different orientations in the past. And, it provides an "excuse" to introduce into this Tutorial a review of the now fully accepted paradigm called **Plate Tectonics** which dominates models of the Earth's geologic history and modes of operation. That will serve as background for remote sensing topics covered in several Sections dealing largely with geologic applications. (Plate Tectonics is also considered in the 'Geology Tutorial' on page 2-1a that should be looked at now if you are unfamiliar with the concept.)

The basis underlying paleomagnetism is the discovery that the magnetic field shifts its polarity with a rather irregular periodicity over tens to hundreds of thousands of years. Thus today's North geomagnetic pole (in which the magnetic lines of force enter Earth from space) has been a South pole (from which the lines emerge) many times in the past. This switching back and forth of the N and S magnetic poles relative to Earth's North rotational pole is termed *magnetic reversal*. When the two North's coincide, the present-day polarity is arbitrarily said to be normal; when the South magnetic pole is located near the North rotational pole, polarity is reverse. It is possible to determine polarity at a given time in the past (as determined by radiometric dating) by measuring the polarity of a rock sample containing ferromagnetic mineral(s). Magnetite, Fe_3O_4, is most frequently used. Magnetite is one of the first accessory minerals to crystallize in a basaltic magma/lava. As the magma temperature drops through 580° C, magnetite passes through the Curie point, at which it takes on its magnetic orientation, becoming like a dipolar "needle". During continued cooling, magnetite grains will orient their N-S polar axis, much like a compass, such that the north pole of a crystal will align with the lines of force and will point to where the geomagnetic North pole was at the time when the crystal was "frozen" in place as the magma solidified. The process by which this occurs is called *thermoremanent magnetism*.

The application is this: Continental basalts (and sometimes sedimentary rocks) containing magnetite are sampled at the place being studied, usually by coring, with the core's spatial position established by fixing its three-dimensional orientation. The (radiometrically dated) sample is then examined in the laboratory for its "fossilized" magnetic orientation (referenced to its internal field position). This determines the alignment (direction and inclination) of the permanent magnetic field imposed on the magnetite before rock consolidation at the time well into the past when the lava was extruded or sediment deposited. After the sample is placed in context with its 3-D orientation, this paleomagnetic field geometry will point to some position on the Earth (relative to today's geography) that represented an apparent location of the poles *at the time of the rock's age*. Furthermore, the polarity (normal or reversed) could also be

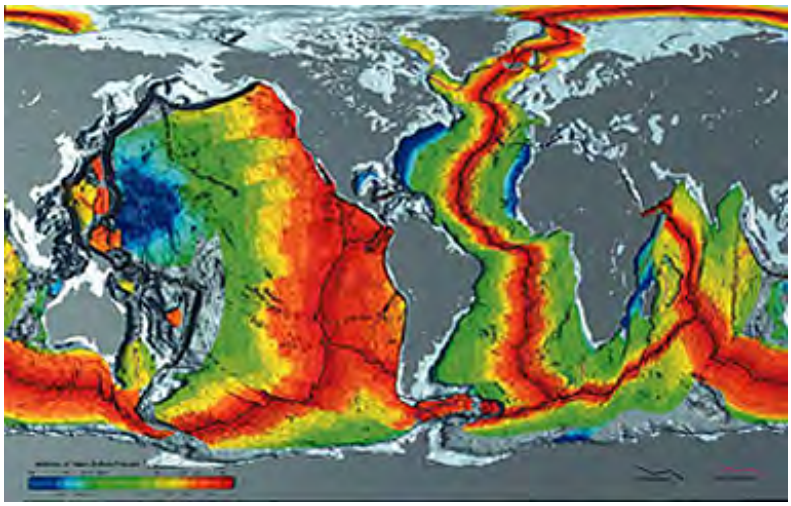
determined. The same technique, and resulting information, has been applied to oriented basaltic rocks beneath ocean floor sediments that were retrieved by submarine core drilling. Two major discoveries, vital to developing the concepts of plate tectonics and continental drift, gradually emerged through the 1960s to the present.

The first is *sea floor spreading*. Through deeper parts of the several ocean basins, volcanic upwelling has built long, relatively narrow rises or ridges which today are sites of active marine volcanism. Lavas emerging today take on the normal polarity that has persisted over the last 700,000 years. When recovered as cores, the solid basalts have magnetic orientations consistent with modern pole locations. But, as sampling moves away laterally from the ridges, the basalts become progressively older. When their polarities are determined, a pattern of alternating polarities (normal-reverse-normal-reverse-normal.....) emerges. Actually, this pattern was initially determined by geomagnetic surveys using magnetometers towed behind ships. Even more remarkably, the pattern on one side of the ridge is matched by a pattern on the opposite side. This diagram should help to visualize the observation (black is normal polarity; white is reverse):

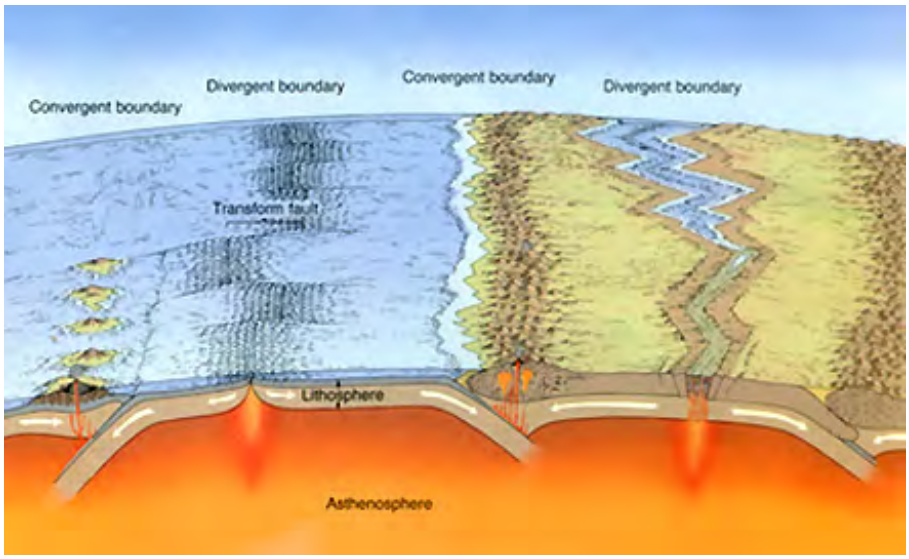


The figure shows normal-reverse patterns for the last 9 million years (horizontal top scale). The current and the three most recent preceding magnetic epochs are named. Average rates of plate movement are shown by arrows. The N-R patterns are repeated in samples from all the oceans, so that the polarity shifts are clearly global in effect. These distinctive patterns form the basis of a geomagnetic age dating method that has been proven to work well. There is enough variation and uniqueness in the patterns (each, a "fingerprint" showing some distinctive repeated normal-reverse sequence) to allow suites of rocks formed over a span of several million years or so to be singled out and separated and then mapped according to the age of each. This magnetochronology has been carried back into the Cretaceous Period some 100+ million years ago.

Enough radiometric ages collected at many locations within the sea floors of the different named oceans have been determined to allow maps of relative age based on the polarity change sequence to be produced for the entire globe. Here is one in which the youngest ages are in red and oldest in blue (thus the blue regions were formed first and have been driven away from the ridges towards plate boundaries). It follows from the defining age determinations that the seafloor is everywhere relatively young inasmuch as older basaltic units at sea floors of pre-Cretaceous age have been completely or largely destroyed by subduction (see below):



As the ages of basalt lavas were determined, and it became evident that the extruded lavas became progressively older moving towards continental land masses on either side of the ridge, a mechanism to explain these patterns was proposed independently by H. Hess of Princeton University and R. Dietz of NOAA (who was actually first, but whose seminal paper was held up by incredulous reviewers; personal comm.). Sea floor spreading postulates that lavas from the oceanic crust or mantle rise to the surface at the ridge, pour out, and push laterally against two lithospheric plates (oceanic crust plus the top of the Upper Mantle) on either side. This acts more or less continuously for long times. The plates slide along the asthenosphere (region of the Upper Mantle hot enough for the rocks to act as though viscous (i.e., "taffy" soft material), moving in a motion likened to a conveyor belt until they encounter a boundary with another plate. (There are 7 major plates and a larger number of smaller plates; see page I-1c of this Introduction or the top of page 17-3 for a global map of these plates). Several different responses are possible at these boundaries, as shown in this next diagram which summarizes much of the main idea behind *plate tectonics*.



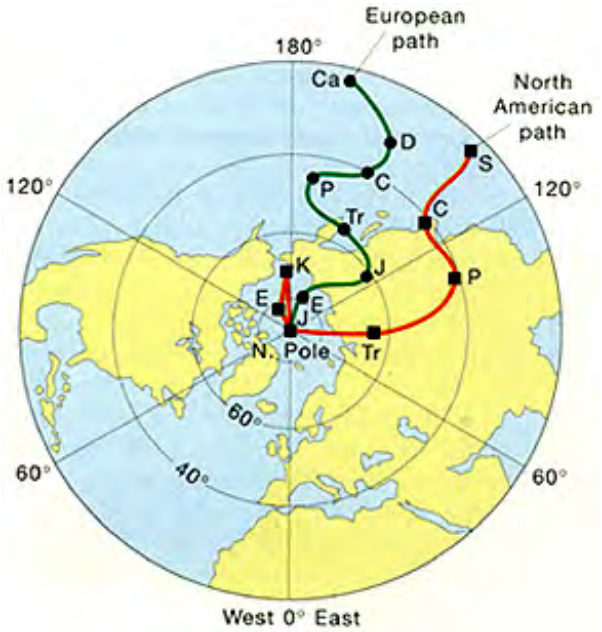
From Tarbuck and Lutgens, The Earth, 3rd Ed., 1990, Merrill Publ.

A Mid-Ocean ridge appears at the left divergent boundary, from which lithospheric plates move in opposite directions. The ridge is cut by a number of transform faults which break the plate into segments. A series of slight rises are evident. Each may be associated with a set of magnetic reversals. At the far left is one type of convergent boundary - where two oceanic plates approach each other. One plate is driven downward (subduction) beneath the other into the mantle, gradually melting as it gets deeper. Some of the molten rock reaches the surface to form a chain of volcanic islands (and sediment) called an Island Arc (Indonesia is an example). The right moving plate moves against a convergent boundary located at/near the edge of an embedded continental assemblage of rocks on its own plate. Above its subduction zone, melted rock and tectonic upwelling and compression produce folded sedimentary rocks and volcanic

and plutonic (magmas that intrude rocks above but don't reach the surface) rocks to form one or more mountain chains (the Coast Ranges and Sierra Nevada/Cascades along the western part of the U.S. are an example). In this idealized diagram, the continent-bearing plate is shown as splitting along a rift zone formed along another type of divergent boundary. The plate to its right is moving against another plate that has a second continent in its crustal portion; again mountains are developed by folding, faulting, and intrusion (the Himalayas against the Indian subcontinent are an example).

One might wonder about the mass balance associated with the plate movements. As a plate is heated up during subduction, it (at least partially) melts. Some of that now fluidized material is moved laterally back towards one or more spreading ridges. Thus, a closed cycle of moving rock and magma results. The process at a ridge is aided by 1-2 sets of circulating convection currents in the Mantle which provide the energy to power the cycle. Possibly some convection occurs above the crust-mantle boundary (the so-called Moho, abbreviated from its discover's difficultly pronounced name, the Yugoslavian seismologist A. Mohorovicik [Mo-ho-ro-vitch-ick]) and may couple with the drag applied by the upper convective movements that affect a lithospheric plate from below.

Now to the second discovery: *Polar Wandering*. When oriented rock samples of a given age are analyzed to determine where their magnetic components are pointing to the location of the magnetic pole on the Earth's sphere at the time of their formation, rocks of different ages are found to have polar field entry locations N Poles) in notably different parts of the present-day world (its geography today is the reference state). One might surmise from the geographic spread that results that the magnetic poles have wandered over much of the globe relative to its rotational poles. This next figure shows plots of magnetic pole positions on a modern geographic grid (latitude-longitude) for the northern hemisphere, with the age of the rocks involved shown by letters (denoting geologic time from oldest to youngest: Ca = Cambrian; S = Silurian; D = Devonian; C = Carboniferous; P = Permian; T = Triassic; J = Jurassic; K = Cretaceous; E = Eocene). For a table showing the geologic time scale, with assigned ages, click here).



From H. Levin, The Earth Through Time, 4th Ed., Saunders

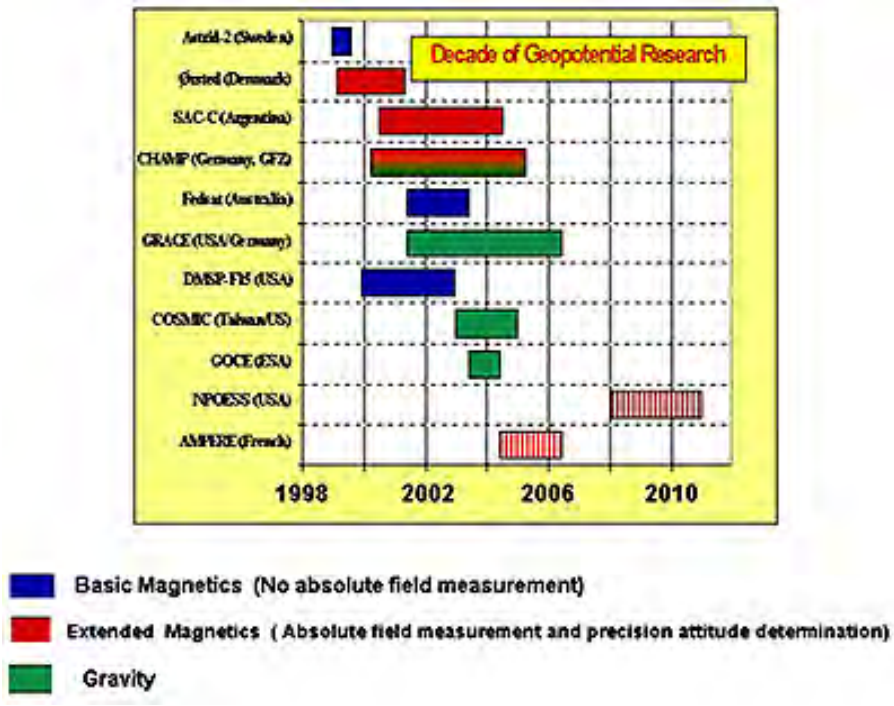
At first glance, this map seems to display a puzzling anomaly: how can now far apart continents record two sets of pole locations when at present we know that only one pole point is the norm. The poles seemingly occupied different positions from the Paleozoic onward until the Present (and still other positions in the Precambrian). One might surmise that the poles were indeed in different locations on the Earth's surface at various times in the past, but that would be hard to explain since the magnetic poles apparently have always been close to the rotational poles owing to the mechanism of magnetic field generation that assumes the electrical currents in the Outer Core move generally subparallel to the Equator. Thus, it's odd indeed to have two sets of

pole patterns in North America and Europe, moving approximately along similar paths, but separated (offset) from each other.

The answer to this duality conundrum is tied to the concept of *continental drift* . At various times in the past, the continents did not have their present shapes (their "true" edges are just beyond their sealevel outlines) or relative positions. In fact, today's continental crustal masses have been once joined into two supercontinents (derived from the Late Paleozoic megacontinent Pangaea, which has since split into Gondwana and Laurasia, each of which then divided further into the present-day continents). As a supercontinent comes apart, its new pieces move across the globe embedded in wandering crustal plates. Pangaea itself resulted from collisions of plates bearing earlier continental masses which were welded together, only to then resplit later; evidence suggests there were similar large supercontinents of different shapes/sizes and locations during Early Paleozoic and Precambrian times.

If the North American and Euro-Asian continents were relocated on the terrestrial globe back to their Early Paleozoic positions, the two curves in the above figure will almost superimpose. This is strong proof of the drift hypothesis. Thus, since these continents were then conjoined, their pole locations were the same but have now moved apart because their continents have separated, so that in today's geography the wandering paths seem to be recording two different pole location histories. Their intermediate pole positions help to locate their different relative locations at each indicated time.

As a transition to a topic - gravity - covered in the next page, we show now a summary of recently orbited and planned magnetic and gravity measuring satellites for the first decade of the 21st Century:



Information on CHAMP and COSMIC, now both in orbit, is available by clicking on these links to the Internet.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

This page is devoted to the study from space of the Earth's gravity and what departures (anomalies) from a uniform gravitational field tell us about the distribution of mass (density variations) primarily in the outer region called the lithosphere. The coverage begins with a review of the real and mathematical shapes appropriate to the whole Earth. The Earth's general figure (shape) approaches that of a pole-flattened ellipsoid. A mathematically-defined surface approximating the gravity equipotential is referred to as the geoid. Gravitational variations owing to topography, crustal structure, and rock type differences provide departures - anomalies - which can be measured from orbiting satellites. Information about these variations comes from analyzing the changes in orbital paths. Satellites have also been able to determine the variations in sea surface heights which tend to reproduce the variations in ocean floor topography. Radar and laser altimetry from these satellites can directly measure these height differences. Global geoid and topography maps have resulted from such investigations. Special attention is given to the GRACE mission which is yielding detailed measurements of gravitational field effects on the Earth's geoid, including isostatic adjustments in the crust-mantle.

In geophysics, especially that applied branch dedicated to exploration for mineral/energy resources, there are three main fields or subdivisions: Magnetics, Gravity, and Seismology. The last finds limited opportunities from space but will be touched upon here as it applies to crustal dyanamics. Most of the subject matter on this page deals with gravity.

We assume that you have some familiarity (most likely from a physics course) with the basic notion of gravity. The fundamental law, proposed by Newton, is $F = G\,(m_1 m_2/r^2) = ma$, where the two m's are masses (e.g., the Earth and you), the m without a subscript is the mass of any body at the surface or falling from above, r is the distance between the centers of masses m_1 and m_2, and G is the Universal Gravitational Constant ($G = 6.6726\ldots \times 10^{-11}$ m^3-kg^{-1}-sec^{-2}). We will not attempt much more consideration of the (physics) basics, but will start by describing some of the main ideas and terms associated with measuring gravity, gravity variations (anomalies), and techniques for obtaining data which can be reduced to a reference state (the ellipsoid) to compare with actual Earth conditions.

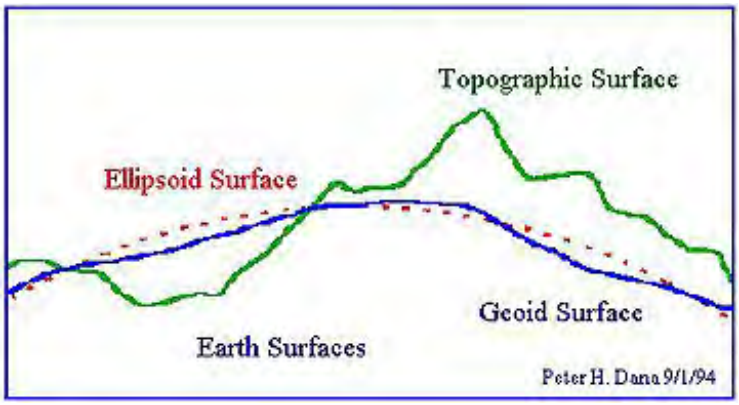
We begin by considering terrestrial gravity as it relates to the actual and the theoretical shapes or figure of the Earth - the subject matter of *Geodesy*.

Ideally, the Earth might have been a perfect sphere composed of a single substance, homogeneously distributed and with a uniform density throughout. A single value of gravity would prevail, determined by Newton's equation. The actual Earth consists of three major zones concentric about its center (where all its mass is said to concentrate for purposes of calculating its gravity): a Core (solid Inner; a thick liquid Outer Core); a Mantle (somewhat variable in mass distribution and density); and a Lithosphere (made up part of the Upper Mantle and two types of crust - oceanic (more uniform in composition) and continental (variable in thickness and notably heterogeneous). For several reasons, the strength of gravity at the Earth's real surface (the topographic surface) varies from place to place. These include the influence of different rock types (and densities), hot spot sources and mantle convection, and structural/topographic irregularities.

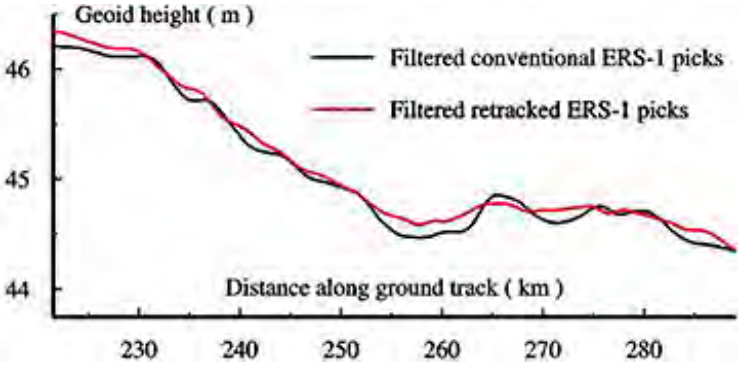
Gravity is usually measured in units of acceleration, as *gals* (or in mapping, as *milligals*) with a standard value (which has varied slightly as better measurments refine it) at 981.275 centimeters per second per second. (On the ground, the measurements are made either with a pendulum-based apparatus or, now more commonly, a gravimeter (weight on a spring).

Ignoring the surficial irregularities for the moment, the Earth's general shape departs slightly from a sphere. The Earth Ellipsoid is an *oblate ellipsoid*, with its average radius from center to equator (semi-major axis) being 6,378.16... km and center to pole (semi-minor axis) being 6,356.77... km; this amounts to a polar shortening of 1 part in 298.25. The overall ellipsoid figure is brought about mainly from centrifugal forces brought about by the Earth's rotation being strong at the equator (creating the equatorial bulge) and becoming ever weaker moving into the polar regions. The ellipsoid, which actually doesn't exist except for calculational purposes, would have a regular, smooth surface.

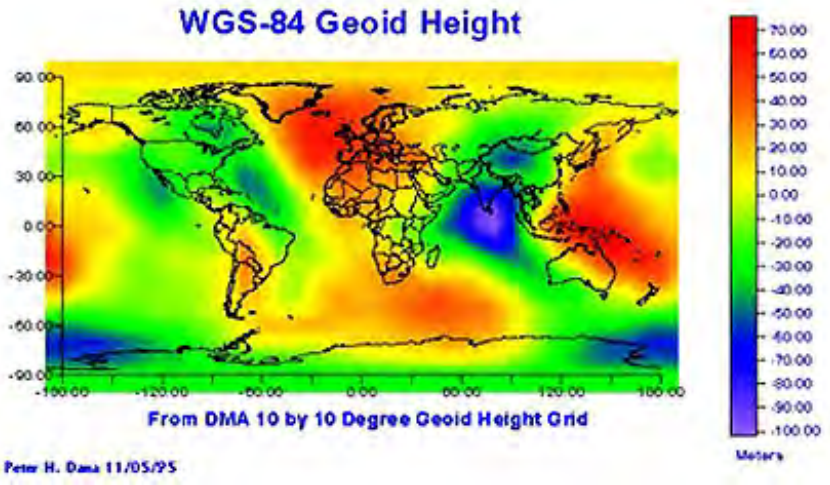
Departures from the ellipsoidal surface caused by differences in gravitational attraction brought about by variations in density produce another reference (mathematically-computed, not actual) surface called the *geoid*. The geoid has been defined as "the (undulating) surface assumed by an undisturbed (perfectly calm) top of the sea, and visualized as continuing under the continents by water filling (hypothetical) small frictionless channels (canals)." It thus closely approximates the variations in heights of existing sea level and hypothetical extensions of the sea through the continents. Related to it is the *sealevel equipotential*, a surface on which gravity is everywhere equal to its strength at mean sea level. The geoidal surface shows broad undulations related to underlying masses. This diagram may help to envision this:



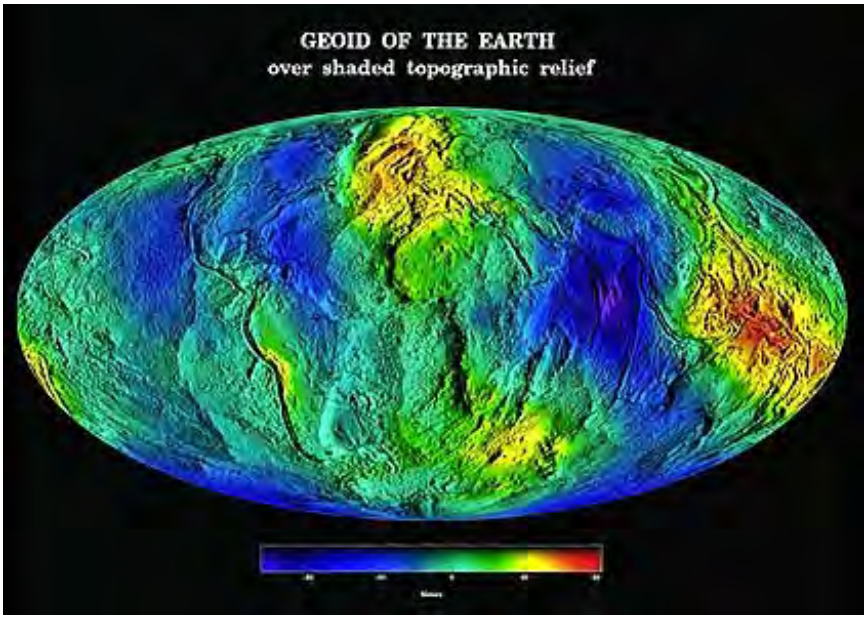
Geoidal data can be derived from satellite measurements. This happens because satellites move up and down along their orbits as they are affected by the same gravitational forces that produce the geoidal surface. An example: this plot of changes in geoid height along a 70 km track as derived from data measured by the ERS-2 satellite (see below).



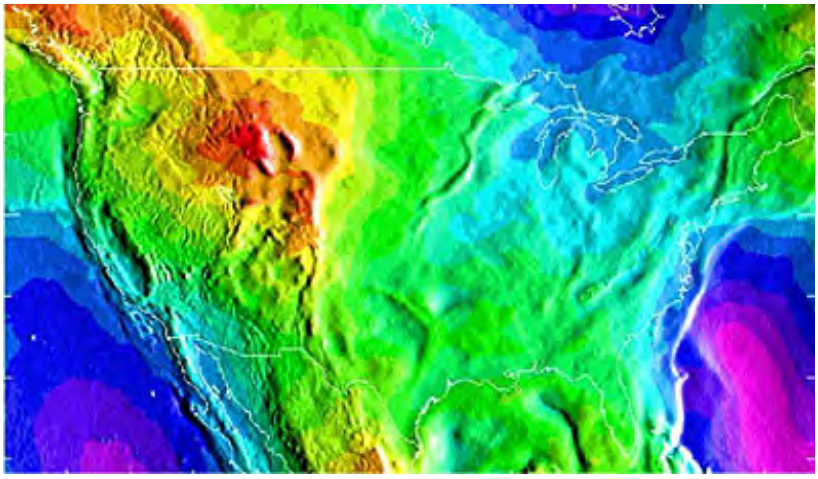
Over the years gravity measurements have led to both generalized and regional models of the geoid. Here is a colored representation of the coarse (10 ° by 10 °) geoidal surface in the standard 1984 World Geodetic System version:



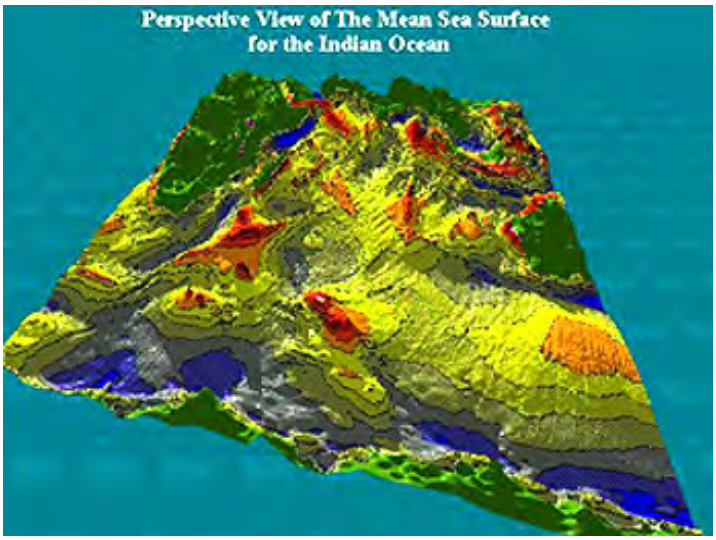
In 1996, a more detailed and definitive Earth Geoidal Model, showing the global geopotential surface (in meters), was produced by the National Imagery and Mapping Agency, NASA Goddard, and the Ohio State and Texas Universities, using data from the Topex-Poseidon, GPS, SLR, DORIS, and TDRSS satellites. These data are derived from careful measurements of perturbations of their orbital altitudes in response to localized differences in gravitational attraction due to variations in crustal mass distribution. The improved version adds information from spherical harmonics coefficient analysis and newer free air anomaly corrections



At higher resolution, this "map" shows the the geoidal surface for the United States and parts of Mexico and Canada:



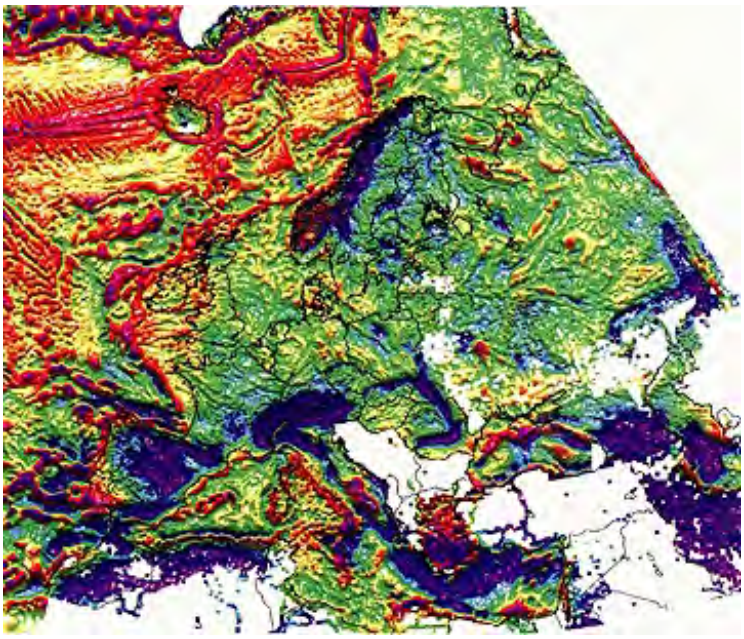
Geoidal surfaces can be calculated for ocean surfaces as well. This is a plot (strong vertical exaggeration) of part of the Indian Ocean. The irregularities are partly due to ocean floor topography (see below).



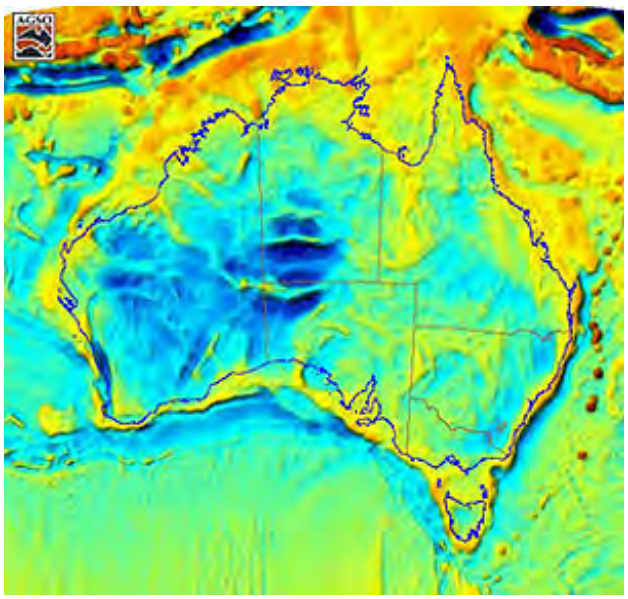
As can be seen from the above illustrations, the geoid tends to be higher in the continents and (relatively) lower in the oceans.

Gravity measurements are usually referenced to the ellipsoid, or for some puposes to the geoid. Gravitational measurements on the ground, from ships, or from space after suitable data reductions will typically give rise to values that are higher (positive anomalies) or lower (negative anomalies than the gravitational potential assigned to the equipotential surface. Several "corrections" must be applied. For instance, for a gravimeter at an elevation above sea level, the *free air correction* adjusts the gravity reading to what it ought to be at mean sea level. A typical adjustment rate is 1 milligal per 3 meters (added if above sea level; subtracted if below). Another correction is made to compensate for the slight increase in gravity going poleward (this ranges from 978.04 cm/sec/sec at the equator to 983.2 cm/sec/sec at the pole, where the instrument would be closer to the center of mass). The *Bouguer correction* removes the effect of any topographic mass (such as mountains) lying above or below the ellipsoid. When all appropriate corrections are made, the observed gravity reading is compared to the predicted or calculated reading based on a geoidal distribution. Any difference is the *Bouguer Anomaly* which is largely due to solid earth masses in the crust, or deeper, owing to different rock types and rock densities.

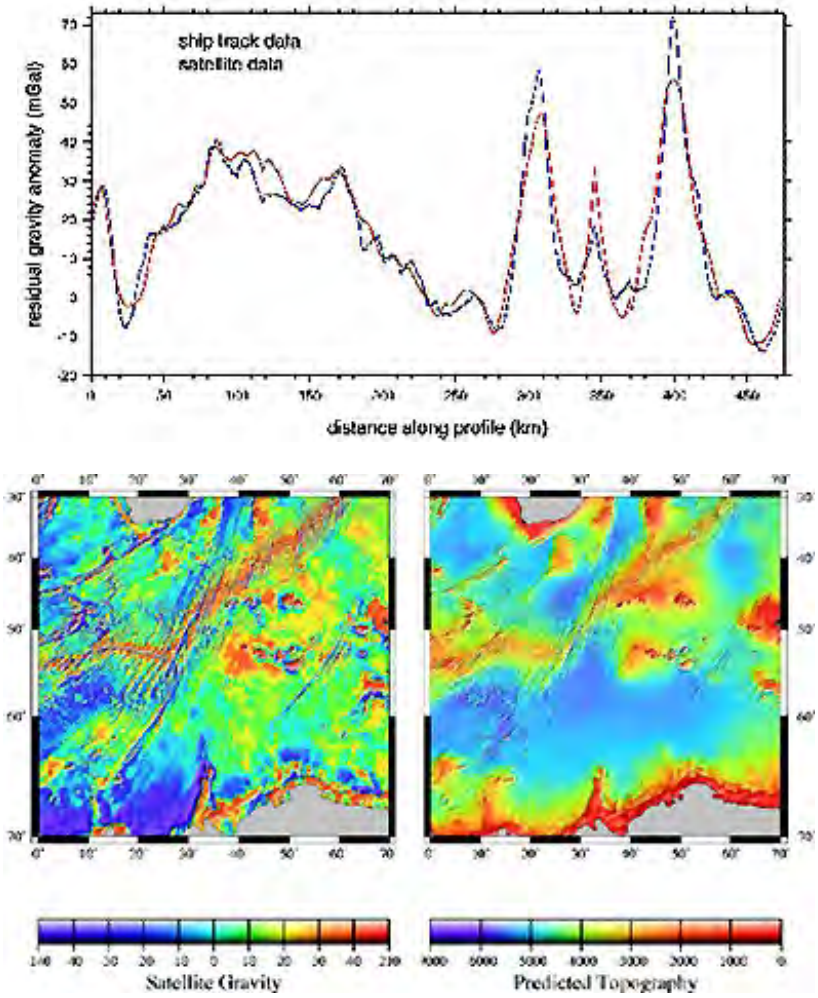
As indicated above, continental gravity surveys have made use of field instruments such as gravimeters. Much of the data have been collected for regions of interest in oil/gas exploration. Surveys can be made from low, slow flying aircraft or helicopters that make measurements with accelerometers and gradiometers. One difficulty is accurate location of the flight line; the development of the multiple satellite Global Positioning System (GPS) has improved the ability to track the flight lines with high accuracy. The data after acquisition must be processed to apply the appropriate corrections described above. Ground/aircraft gravity anomalies made from different surveys can be integrated into regional maps or even at the scale of continents. Below is a gravitational anomaly map (Bouguer on land; Free Air anomalies over water) (positive anomalies in red, with maximum at +75 milligals; negative in blue, minimum at - 75 mgal) of most of Europe, the Mediterranean, and the North Atlantic (the outlines of countries in black are there but hard to see; to get your bearings, the dark blue area near the top is Norway; a similar color below center is the Po Valley south of the Alps:



A gravity map of Australia (at lower mgal resolution) made by air/ground surveys and its surrounding oceanic floor as made with satellite data is a clear example of how both near-surface and spaceborne gravity-measuring systems can be combined to make a gravity map:

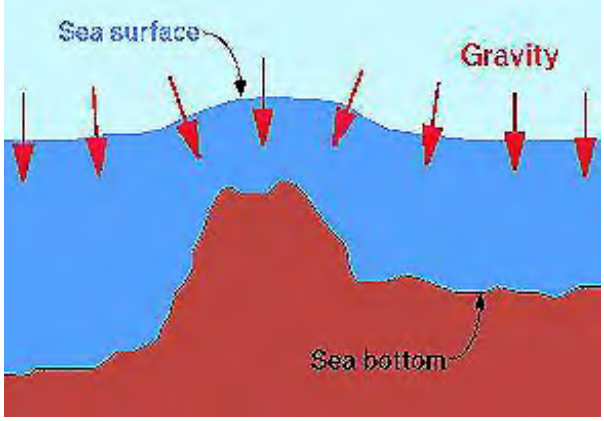


Now, on to the role that satellites play in determining Earth's gravitational field and anomalies. Two figures indicate how well *gravsats* (a term for satellites used to obtain information on gravity, introduced here to foster brevity; do not confuse it, with the name of a British Company, Gravsat, that commercially supplies satellite data) can duplicate or approximate gravity data acquired on the ground or by air. Both are in units of mgals: the top is the tracing of gravitational changes measured along a satellite track crossing the South China Sea as compared with readings from a shipborne gravimeter following the same line; the lower figure shows a gravitational anomaly map of an oceanic area on the left and a map obtained by using limited field data from a marine survey.

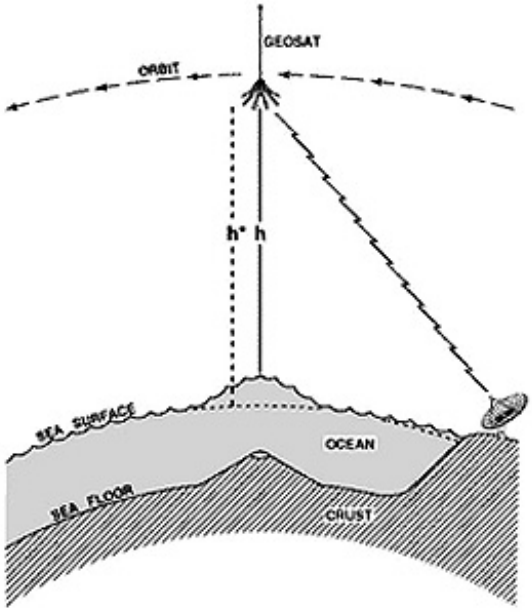


As a general statement: the prime use and successes of gravsats have been in surveying the topography of oceans and other large bodies of water. Because of their high rates of orbital velocity and other complications, satellites do not normally carry accelerometers (which have low response times) to measure direct changes in gravitational attraction along their orbital tracks. Instead, gravity variations can be calculated from changes in the position (shifts in orbital altitude) of a satellite as it orbits - these are caused by the varying force of gravity from both the transient nadir points and surrounding areas near the path trace. Tracking of radio signals (using Doppler shifts in frequency) from the satellite help to determine these variations. A more direct way is to follow the satellite's path or locate its position with a technique called satellite laser ranging (SLR). The other major approach is to measure, from the satellite itself, the changing heights of the surface (using that of sea level with reference to the geoid) by either radar altimetry or laser altimetry.

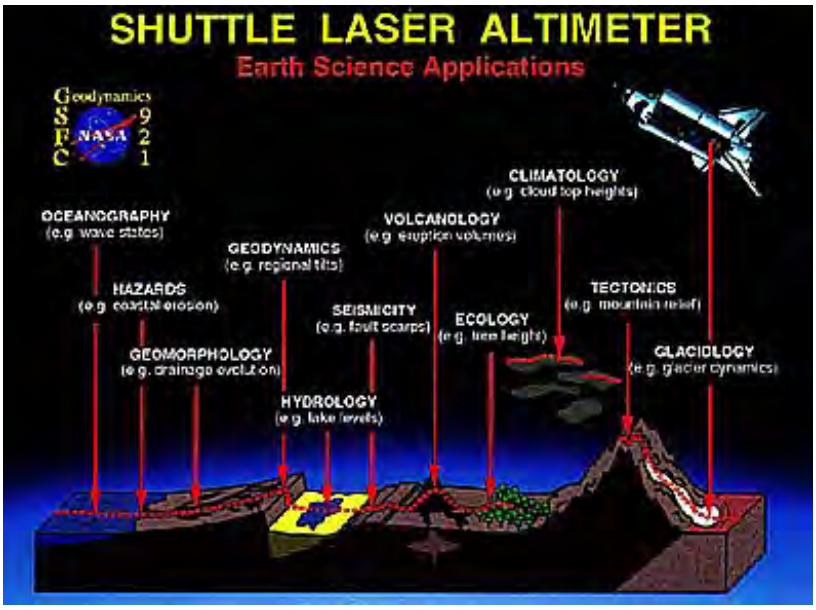
The principle behind measuring the changes in sea surface elevations (which can vary as much as 160 meters over the Earth's oceans) is evident in this diagram:



The presence of this extra mass in the seamount above the sea floor is the cause of deviation of gravitational attraction forces such that water around the seamount bunches up into a rise or mound. If instead there were a depression (e.g., oceanic trench) on the floor, the drop in gravity owing to absence of material (mass) in the trench space would cause a depression of the water at the surface. In other words, the sea surface roughly mirrors (reproduces) the topographic variations of the sea floor. Simply by measuring the distance between a satellite sending a signal (for example, a radar beam that produces returns) to the water surface, departures N from the geoid are determined, as shown in this next figure (where N = h* - h; h* is the distance to the calculated geoid for the area and h is the signal path distance of the surface bulge to the satellite).

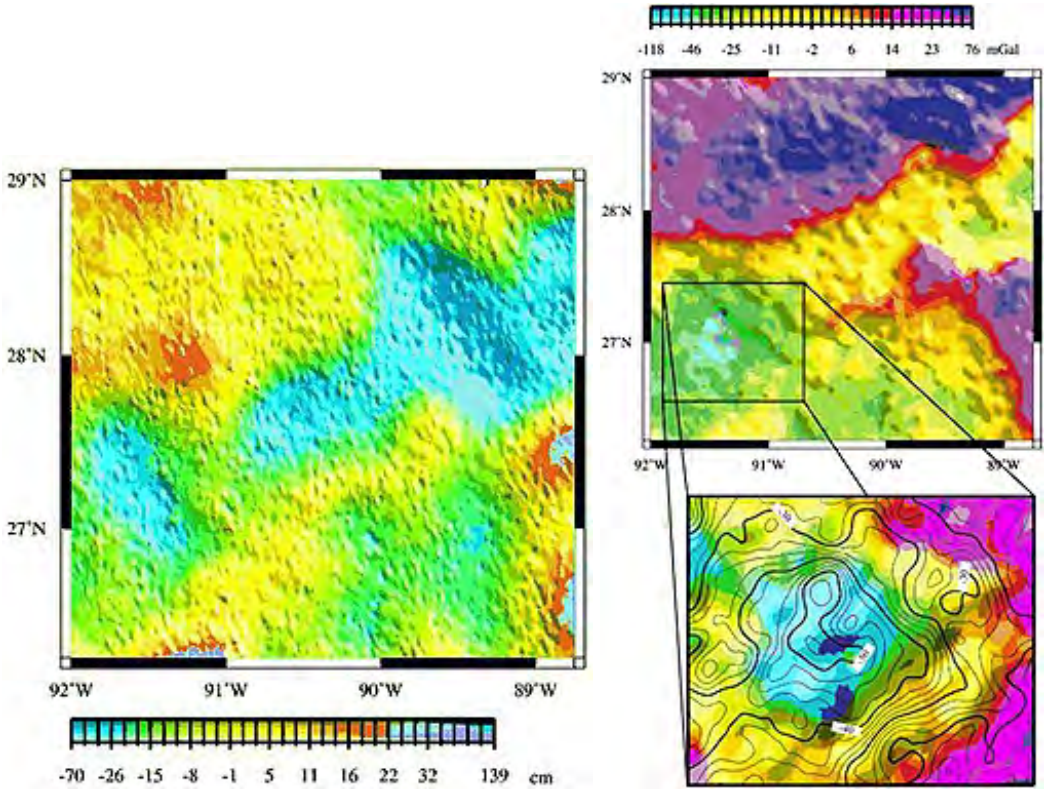


This approach was first tested by an experiment on Skylab. Proof of concept was then dramatically demonstrated by the radar altimeter on Seasat (see page 8-6). Altimetry data from this mission led marine geophysicists to construct the now famous general map of the world's ocean floors (bottom of p. 8-6), with their structural patterns reflected in their topography. Radar altimeters now operate (or have earlier) from a host of satellites including Seasat, Geosat, Topex-Poseidon, ERS-1 and -2, Geos3; new systems will be onboard Envisat and JASON-1 to be launched later. Radar interferometry is also used for sea state (surface conditions) studies (see page 8-7). Laser altimeters (which send out pulses of coherent light) are increasingly being deployed on aircraft and have flown on two Space Shuttle missions and will be on the GLAS (Geoscience Laser Altimeter System) launched in 2003. Lasers also were on a Mars spacecraft. The application of lasers from moving platforms are summarized in this diagram developed for the Shuttle Laser Altimeter (SLA) experiments:

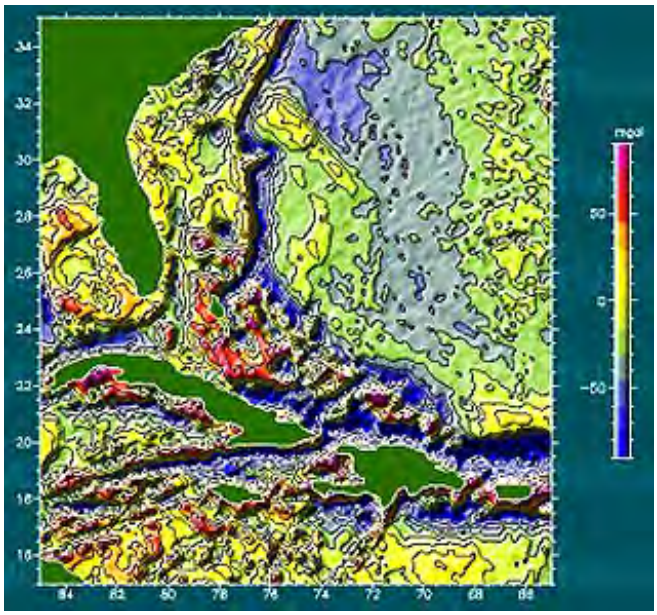


On the remainder of this page we will give some examples of ocean floor gravity anomalies and a topography interpretation that have resulted from satellite altimetry.

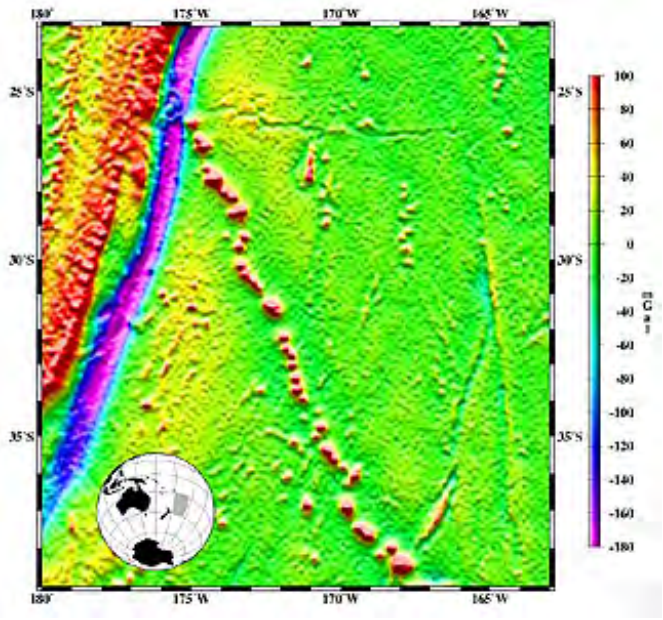
First are two maps of part of the Gulf of Mexico made from ERS-1 and Geosat data combined that are part of the study of that region in the Getech project (University of Leeds). The map at the top is a geoidal base showing mean sea level variations. Below it is the same area now converted to gravitational variations (in mgals) that indicate the influence of ocean floor configuration.



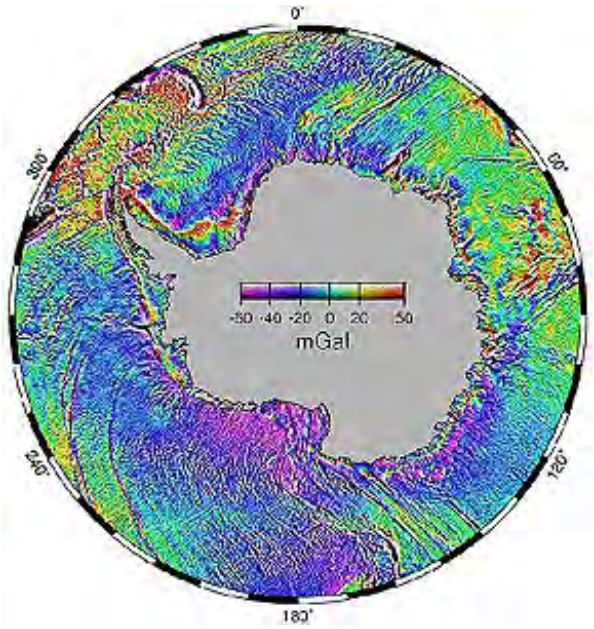
To the east is the ocean off the Florida coast, with determinations of gravity anomalies extracted from ERS-1 altimeter data.



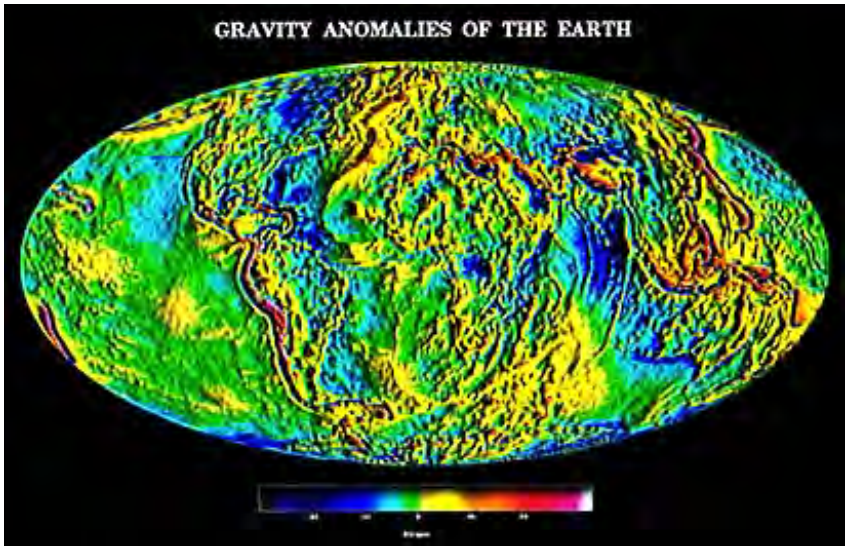
Here is a gravity anomaly map of a region in the southwest Pacific in which one of the major ocean trenches (Tonga-Kermadee, on the left) is easily recognized, as are a series of small ocean floor volcanoes making up seamounts. This map emphasizes the fact that the gravitational variations tend to define the actual topography of the crust at its interface with the ocean.



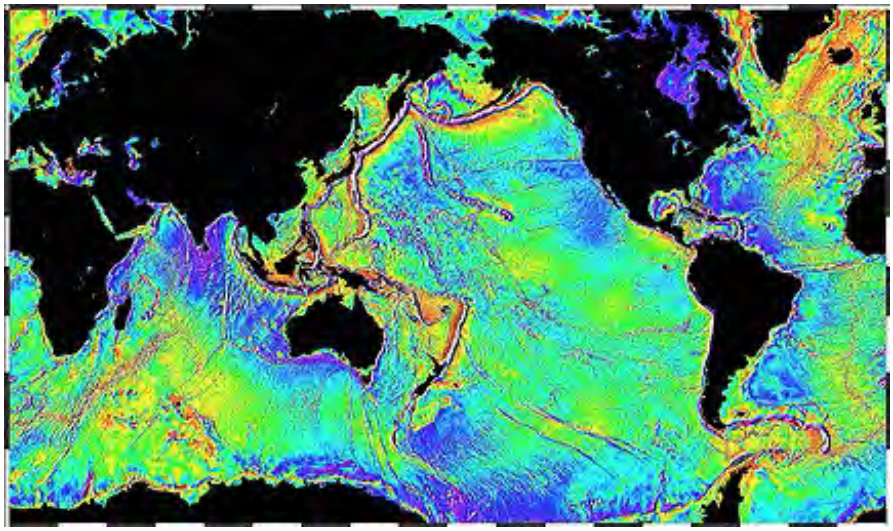
Now, examine this gravity map of the Southern Oceans around Antarctica.



Various maps of global gravity variations have been constructed for both continents and ocean floors combined of which this is typical:

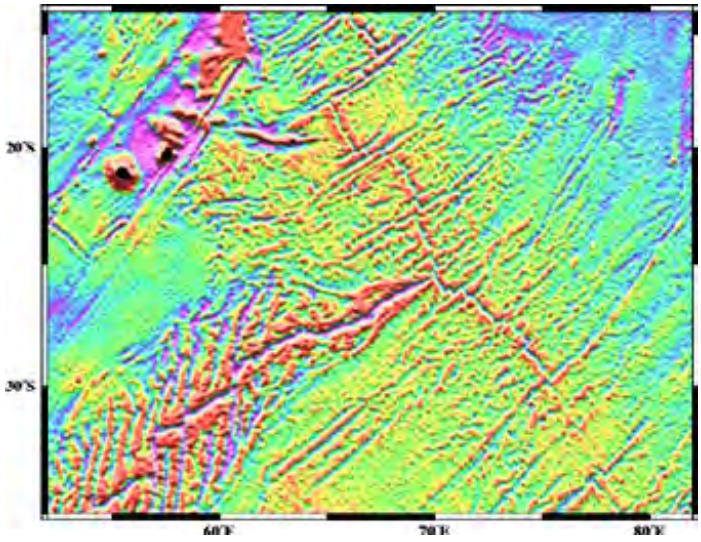


Likewise, maps showing just the sea floor gravity anomalies have been published:

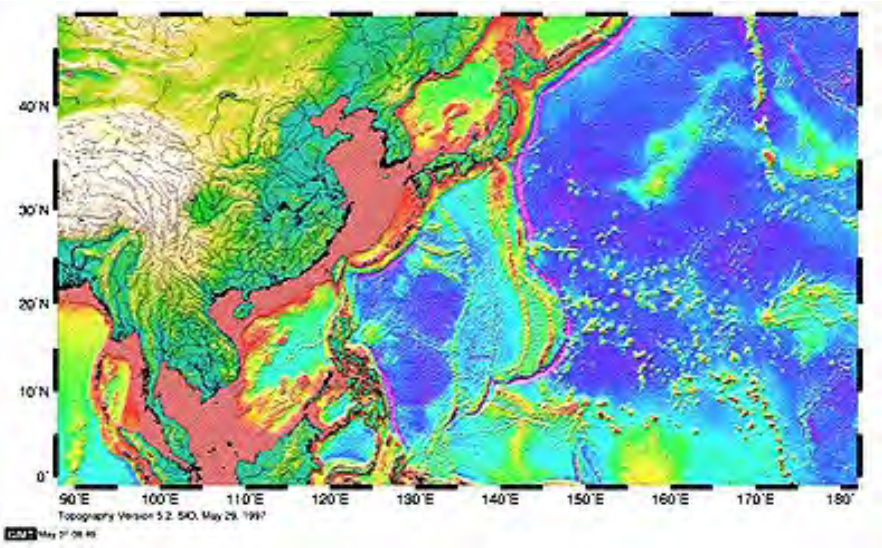


Maps of the actual ocean bottom surface (its topography) resemble the gravity maps closely since they commonly use the same color scheme (blue for low; red for high). These maps are nearly identical to bathymetric maps (those that show the thickness of the ocean water cover -

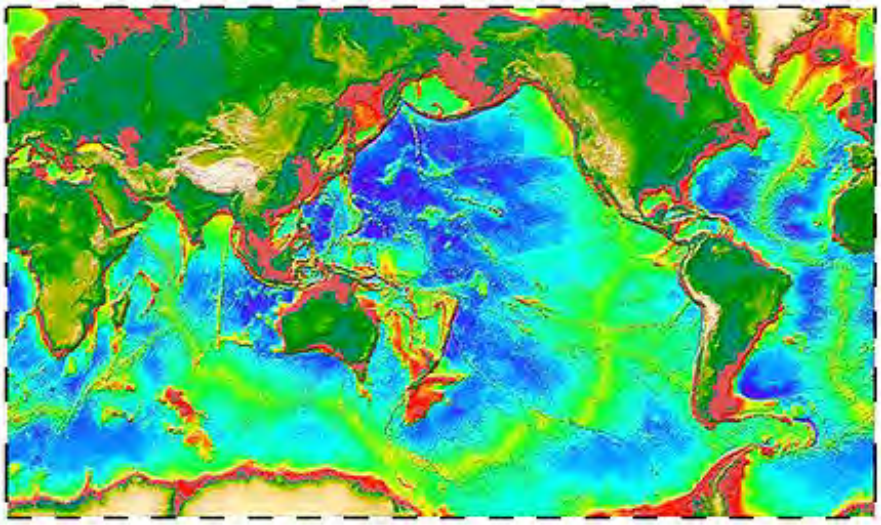
depths). Here is a topographic map of the ridge and transform fault system in part of the western Indian Ocean:



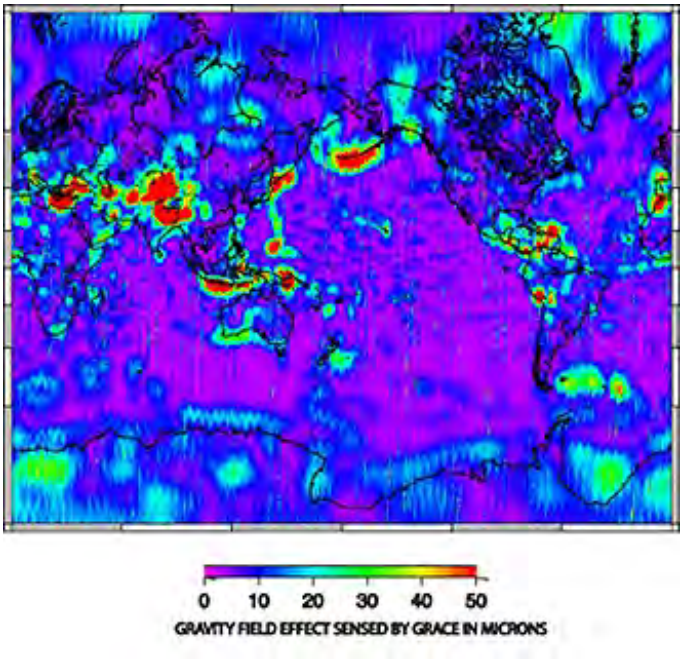
This is part of a global land and ocean topography map produced by Smith and Sandwell; it covers a 45 by 45° area that includes part of China and Southeast Asia and marine topography from north of Japan to Indonesia eastward into the western Pacific.



The full Smith and Sandwell first order topographic map (1995) for the whole world is reproduced here (ocean deeps are dark blue and continental shelves are red-orange):

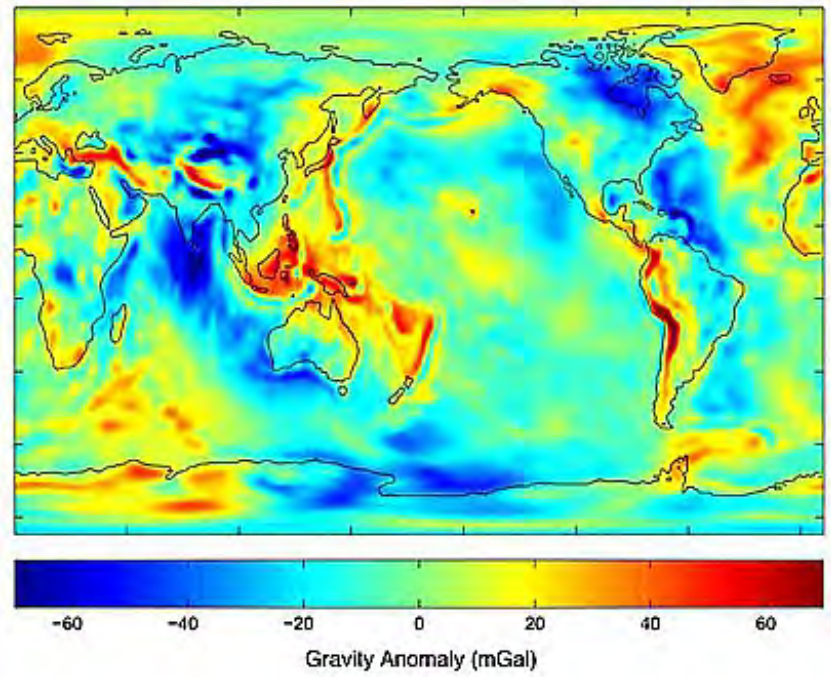


A recent geophysical-gravity satellite is GRACE (Gravity Recovery and Climate Experiment), launched on March 17, 2002. GRACE is actually a pair of satellites in the same orbit at 500 km but 220 km apart (formation flying). The precise location of each satellite is constantly determined by their radio interaction with Global Positioning Satellites (GPS). Each GRACE satellite uses microwave signals to determine very precisely the vertical distance between spacecraft and points on the ocean surface (and land as well) to get a sea surface height that is 10^3 times more accurate than previous systems. This yields a much more detailed picture of the geoid, or mean gravity field. This first global image shows a remarkable feature:



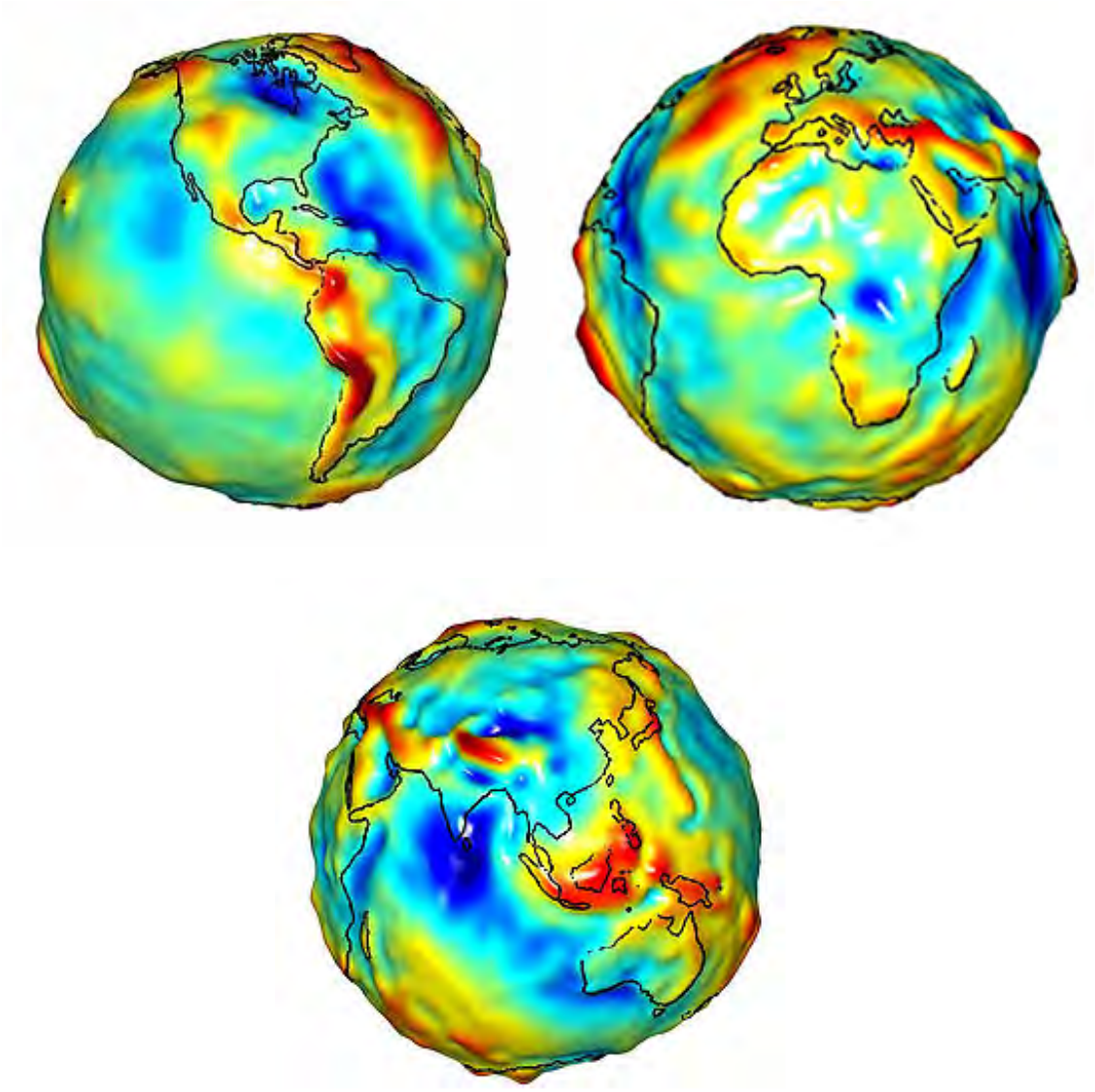
That feature is the excellent definition of subduction zones at converging plate boundaries. These are the patterns that are displayed in reds. Note especially the Aleutian chain, the Indonesian chain, and the Himalayan-Indian boundary.

In July 2003, the GRACE team released a more accurate (by a factor better than 10x) gravity map of the entire world. This incorporates the ability now to show variations at the centimeter level. Gravity Highs are in red; Lows in blue.

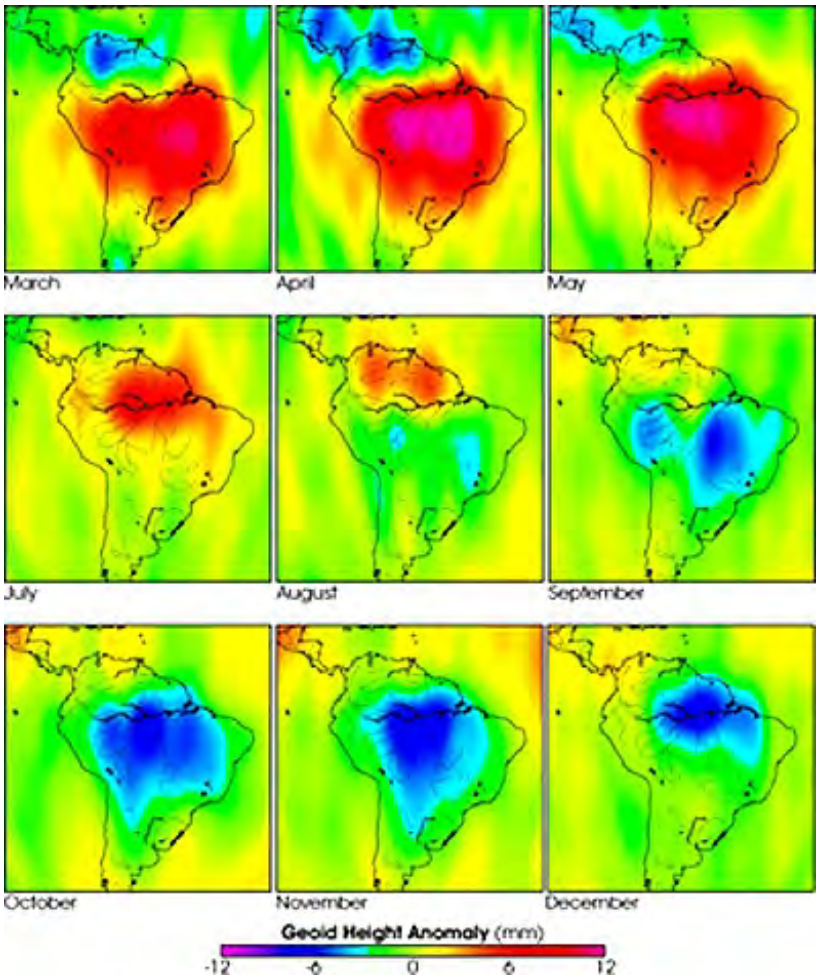


Since 2003, GRACE scientists have further massaged the data to produce regional gravity maps that show an increase in accuracy by another factor of 10. Here are two of these that show

most of the globe:

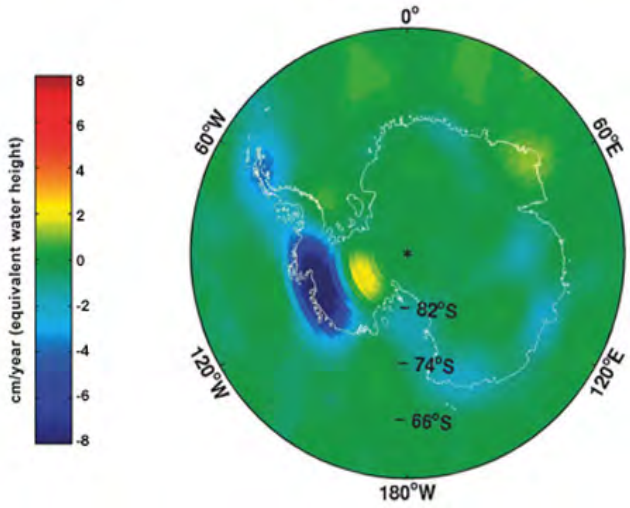


A spectacular demonstration of the value of subcontinental scale GRACE measurements over an extended time period is evident in this image of much of South America that shows small changes (millimeter scale) relative to the geoid in mean surface elevation in the Amazon Basin and beyond:



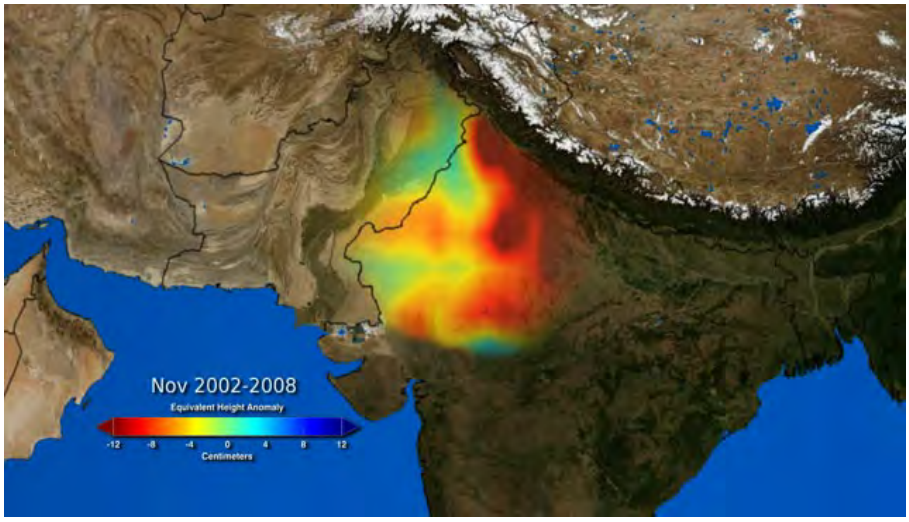
From March through about July the geoid in the Amazon is somewhat higher than average but this elevation transitions into lower heights from September through December. Such variations suggest what is known as isostatic adjustments. Isostasy refers to the tendency for the Earth's outer sphere to experience forces that cause its surface to rise or fall as loads are added or removed. The response is not instantaneoous but occurs over time. For instance, if much of a continent is covered with ice (glaciation), then the crust beneath will sink under this load; when the ice is removed that crust will rise (rebound). Part of the isostatic mechanism involves plastic flow adjustments in the upper mantle. In the South American case pictured above, the added load, leading to the blue stage, is the accumulation of water in the Amazon Basin during the rainy season. GRACE is sensitive enough to monitor the resulting elevation changes.

GRACE is capable of showing changes in the thickness of ice in the Antarctic. This map points to a decrease (in blue) involving several centimeters in a region of the ice sheet



A more subtle example of the value of GRACE is found in western India. Groundwater

withdrawal can lead to a sinking of the surface as the overburden pressure acts on empty pores. Note the reduction of surface elevation (in red) in this map:



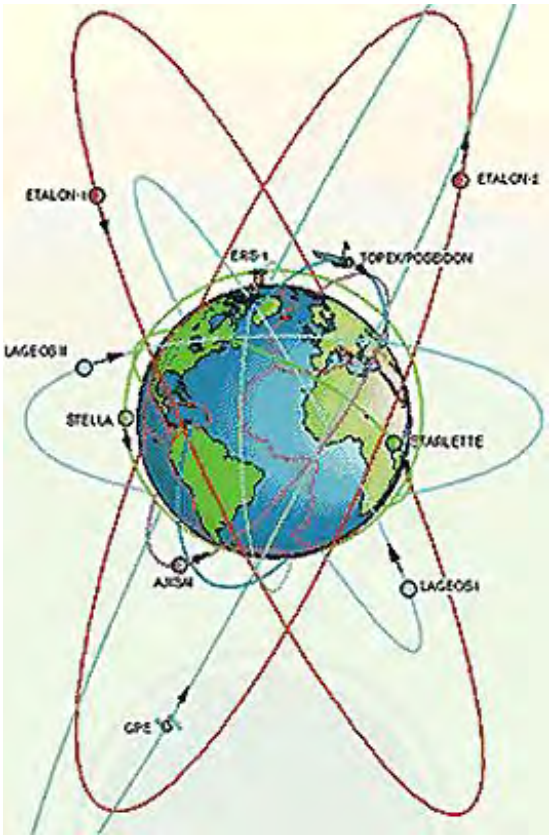
GRACE will continue to acquire high sensitivity data for up to 5 years; from the data sets gravity profiles will be constructed.Global gravity map made from GRACE microwave measurements. A good overall review of the GRACE project was presented in a JPL lecture. Access this at von Karman lectures, choosing the topic "GRACE: Gravity Recovery and Climate Recovery, March 16, 2002.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Satellites have been effective in determining the slow but continual movement of the Earth's plates, thus helping to understanding crustal dynamics and associated deformation. A technique called Satellite Laser Ranging (SLR) uses a laser beam to range a moving satellite, on which corner cube reflectors (functioning like mirrors) are embedded around its surface, reflecting that beam to its source site on the ground. From the resulting travel-time measurements. distances to the satellite can be calculated to within a few centimeters (or less). The orbit can be tracked from several laser stations. Over time (in years), different stations, on different plates or within a plate, will separate from one another according to the amount of differential movement in the plate or, more commonly, between plates. A second technique, called Very Long Baseline Interferometry, uses several radio telescopes spaced well apart to receive radio waves from distant sources (e.g., quasars); the signals can be analyzed to precisely locate each station at a given time, so that again over years plate displacements can be specified. This page also considers some elements of seismology, particular in regard to earthquakes. Radar interferometry from a satellite is capable of determining horizontal or vertical displacements of a surface following an earthquake beneath it.

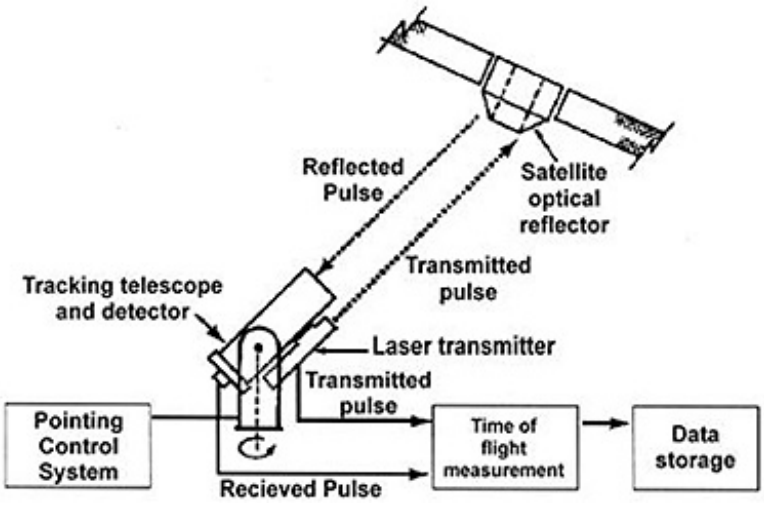
On the previous page we mentioned that sea surface heights and seafloor variations can be determined from altimeter studies. Some of the satellites dedicated to, or including systems for, altimeter studies of the sea surface and, where pertinent, the land surface are shown in this schematic summary:



These satellites have one thing in common: They all have corner-cube reflectors emplaced uniformly in spacing over their outer surface. These, and many other satellites, are used in a technology called *Satellite Laser Ranging* (SLR). This is the reverse of laser altimetry where the instrument transmitting the laser beam is on the satellite itself. Check this Website for a good review of the basic principles underlying SLR. In SLR, the laser generator is fixed at a ground

station and its beam is directed into space to intercept and track the satellite equipped with silicon glass plates (usually two per port, mounted at right angles [corner-cube configuration] so as to accept a beam from any direction and return it) in an array termed retrograde reflectors. Here is Lageos (sometimes written as LAGEOS; it stands for Laser Geodynamics Satellite), one of the first satellites dedicated to geodesic measurements, orbited at 6000 km in 1976.

This schematic diagram synopsizes the mode of operation of a typical SLR system:

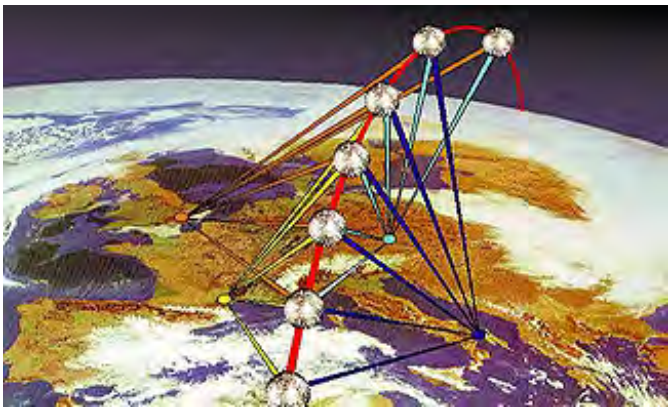


These are two of the now more than 46 active SLR stations in a global network:

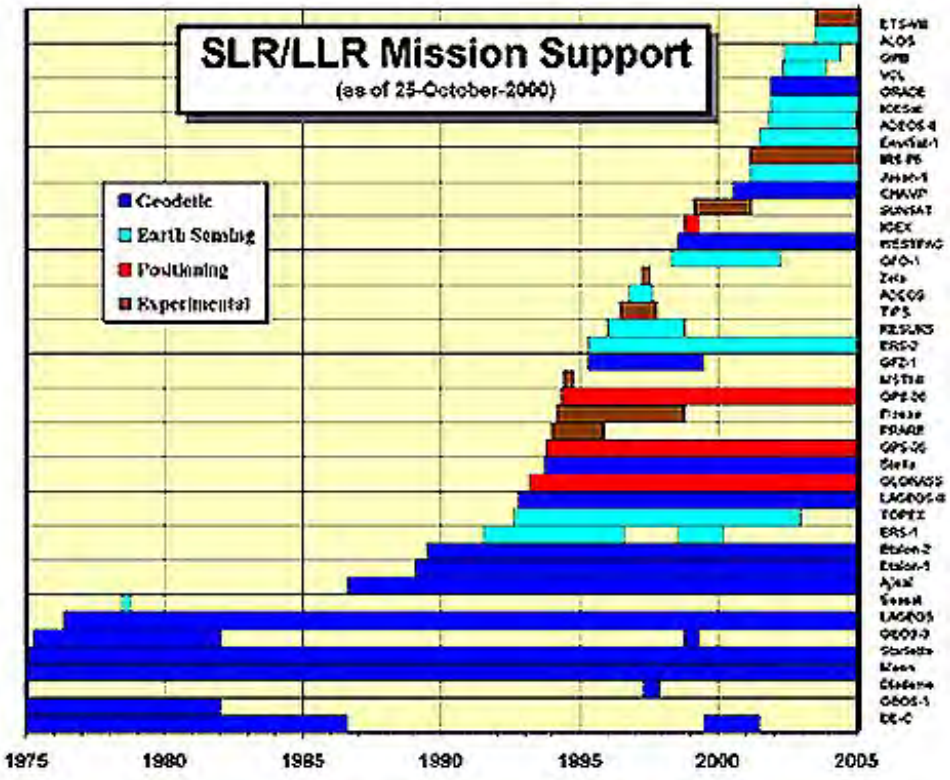
The laser beam consists of coherent light generated at the site and sent out through a telescope as bursts of very short pulses. The beam seeks out an orbiting satellite and after

finding it will continue to track the orbit for a time. A given pulse is returned to the SLR site in a very brief period of time; the time of start and the time of the same pulse's return can be very precisely determined by a maser clock (interval differences in picoseconds [10^{-12} sec] can be separated). Since the speed of light is known very precisely, the distance between the SLR station and the satellite at the instant of intercept can be determined to an accuracy of less than a meter. Because separate ephemeral (very short time duration) data can fix the changing position of the satellite quite accurately, the distance plus angle of beam projection allow the location of the transmitting station and the orbiting satellite to be well set. (As an aside, the "trick" in the whole operation is to be able to "hit" the fast-moving satellite [whose size may be a meter or the mirror-bearing part of a larger satellite may not be much bigger], and then follow it; ancillary optical and radio [S-band; 10 cm wavelength] tracking or Doppler tracking allow the orbital path to be so exactly known that its position at any moment can predicted close enough that the chances of an intercept are quite good [and, typically, one in every 5 to 10 pulses will reach a mirror].)

Commonly, a satellite is within the line of sight of several SLR stations. These may operate simultaneously such that they in effect "triangulate" on a satellite, a situation that greatly favors more accurate determinations of position, both of the satellite and the ground stations. Thus:



There are now a fair number of satellites fitted with retroreflectors suited to SLR tracking. Some of these have other prime purposes; a few are simply targets for geodetic studies. This next chart shows the past, presently functioning, and plannned satellites equipped with reflecting mirrors:



Because the lettering on the ordinate that identifies a satellite is hard to read, here is a partial listing of SLR-compatible satellites: ADEOS; ALOS; CHAMP; Envisat; ERS1 and 2; Etalon; Geos3; GLONASS; GPS35 and 36; Lageos1 and 2; Meteor3 and 6, Starlette; Stella. The designation LLR refers to Lunar Laser Ranging; three reflectors were left on the Moon's surface by Apollo astronauts and two more were on Russian probes that successfully landed; their presence has allowed much improved determinations of the lunar orbit and distance variations. A list of satellites that participate in SLR measurements is found at this Goddard Website.

Notice that the listing includes two GPSs (Global Positioning Satellites). This is an entirely separate system for location of points on the Earth or in space (it is the basis of civilian commercial hand-held instruments that allow anyone to locate himself/herself on the ground [for instance, when hiking in the wilderness] or in a car or boat). The system in likewise invaluable as a surveying method and in conducting some of the studies mentioned in the next paragraph. More is said about GPS on page 11-6 or you can find a good review online at this University of Colorado summary site. The Russians operate a similar system called GLONASS.

The use of SLR has become sophisticated enough that its data are now collected and available through the International Laser Ranging Service. From these data the International Terrestrial Reference Frame, a compilation of data on movements and relevant conditions of the land and ocean surfaces, is derived.

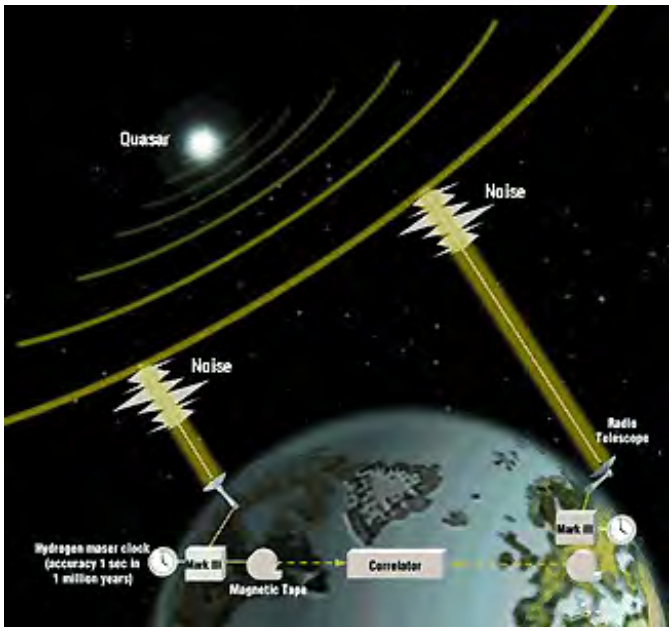
SLR has found many important applications, chief among which are: Polar motions and Earth rotations; gravitational variations of land surfaces; ocean surface configurations; ocean tides in basins; ice mass and volume; post-glacial rebound measurements; crustal deformation and tectonic plate motions.

This last category involves a major NASA, NOAA, U.S.Geological Survey program, joined by a group of international organizations that is commonly referred to as the Crustal Dynamics (now run out of Codes 920.2 and 926, NASA Goddard, check their Website); related information is found at the Geodynamics Branch site. The main objective is to measure both horizontal and vertical movements of the Earth's tectonic plates and their influence on crustal deformation. These motions, particularly at plate boundaries, amount to a few millimeters to centimeters a year. By checking periodically every several years on the precise locations of geodetic control points, mainly at specific stations established as bench sites, the actual movements of these

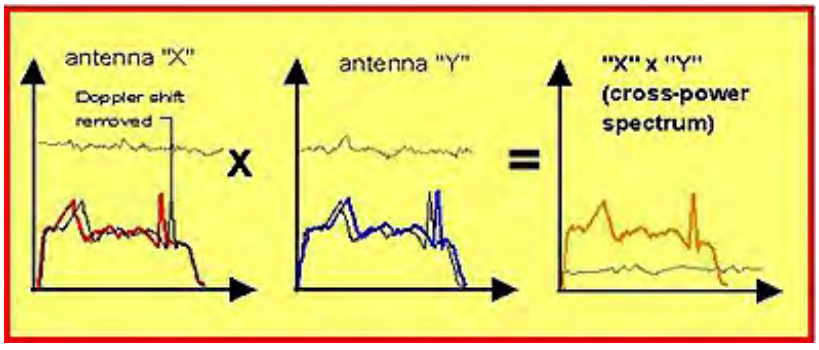
points translate into approximately the same displacements within the plates driven away from spreading ridges. Very precise measurement of these points as they shift over time has allowed reliable estimates of the movements (both direction and velocity) both within plates and at their edges. Different plates move at different rates and directions.

SLR has proved capable of high accuracy. But another method, Very Long Baseline Interferometry (VLBI), also is also able to pinpoint a station's position and changes since its previous location(s). It relies on extremely accurate analysis of radio signals from distance sources that are received simultaneously at several radio telescope stations. Some of the basics of VLBI are covered at this Goddard site. The Canadians have prepared another good review site covering Principles of Radio Interferometry; a clear overview of some general principles of interferometry is found at this Canadian site.

Some of the main ideas behind radio interferometry are summarized in the next several paragraphs. Look first at this diagram:



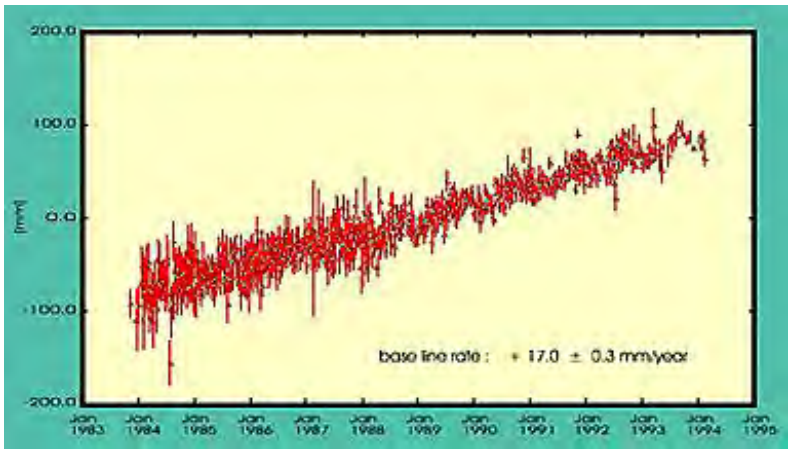
The radio wave source is usually some radio galaxy or a quasar sending strong radio signals. These, of course, travel at light speed. Two or more widely separated (very long baseline; resolution increases with length or separation) radio telescope receivers each record the signal. Each is also tied to a very accurate (hydrogen maser) clock which synchronizes the timing. Data reduction includes removing atmospheric effects and then checking time differentials for a given signal which result because one station is further away from the distant signal source than the other. The wave forms of the signals thus do not coincide precisely, i.e., they show partial to complete constructive or destructive interference (in optical interferometry, these slight differences in phase would appear as light fringes). Using a computer-based correlator, signal analysis by cross-correlation is conducted, as displayed in this diagram:



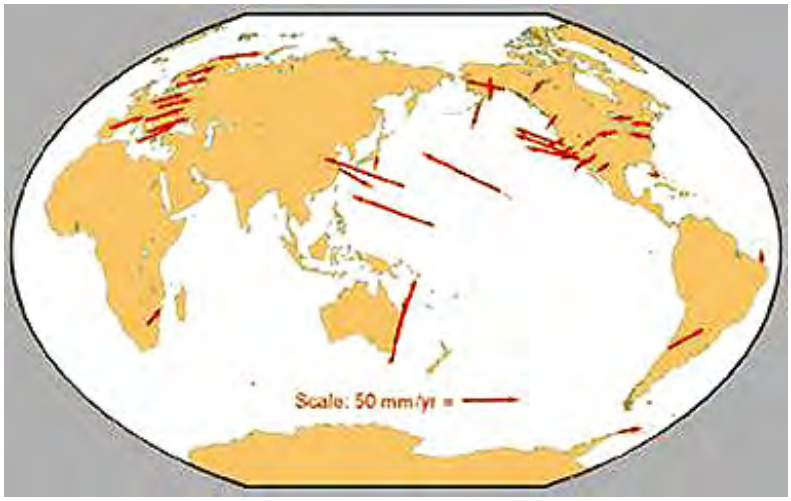
The result is a new data set from which the positions of reference points can be calculated and compared. Presently, about 40 receiver sites, mostly in North America and Europe, are

participating in this effort.

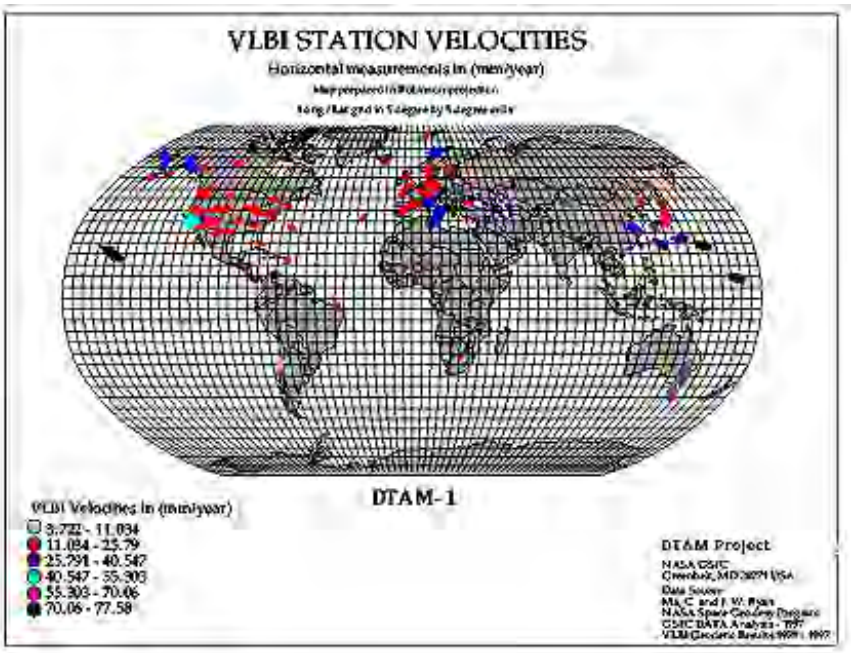
The most accurate data on displacement rates come from averaging years of measurements. For instance, this plot covers 10 years of continuing data for the increasing separation between a European and a North American station. While there is scatter (uncertainty) for each measurement, the straight line fit (17 mm/yr) is credible.



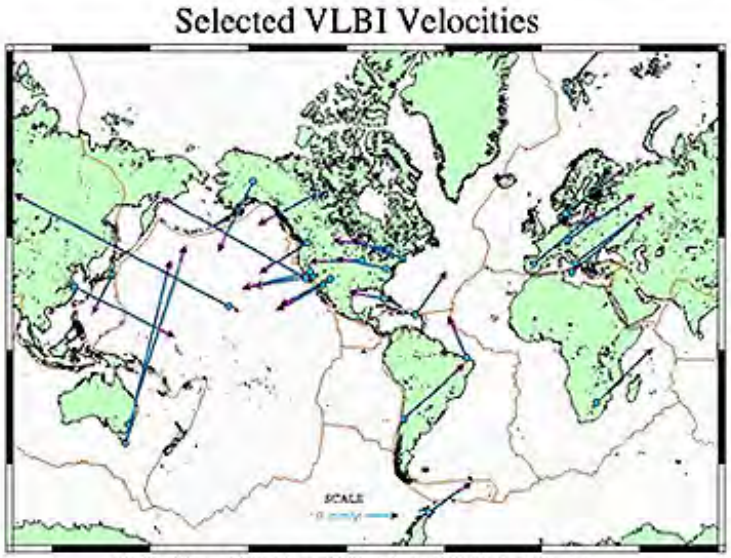
Two maps indicate the emerging patterns of relative plate motion for those plates with enough site points to indicate both magnitude and direction of movement. The first is more general:



The plate containing the European "continent" is moving northeast, as the African plate pushes against it. The Pacific plate's direction is to the northwest. Two contrasting movement patterns occur in the North American plate; most is being driven west from the Mid-Atlantic Ridge but a small segment or sliver of western California southward is moving north-northwest to northwest where complex interactions (in part related to the overriding of the Cocos plate with the eastern Pacific plate cause sliding along a series of transverse faults (e.g., San Andreas). This next diagram offers more details and shows the fastest movements now occur in California (turquoise) and in southeast Australia (purple).

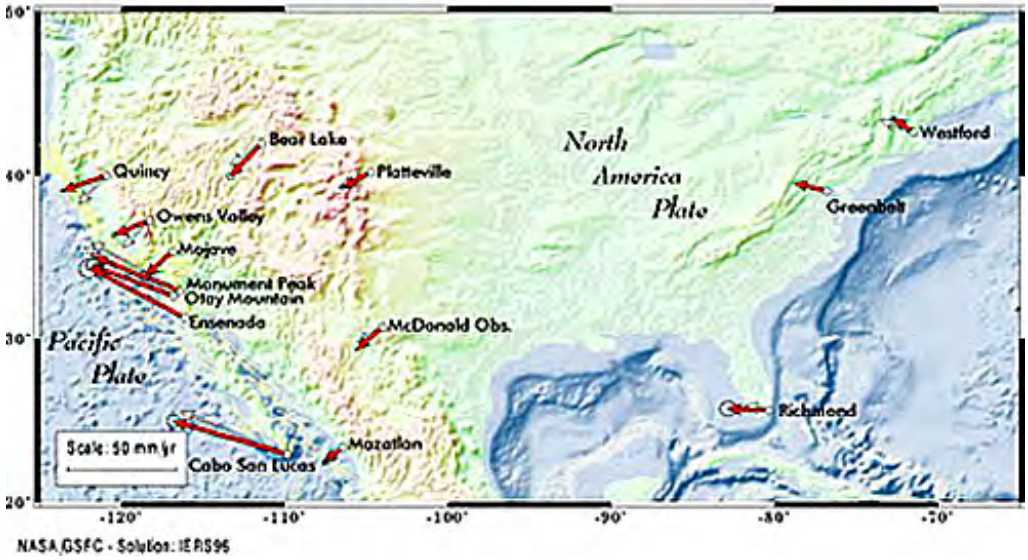


This is more evident in the following map on which velocity vectors are plotted relative to the plates involved:

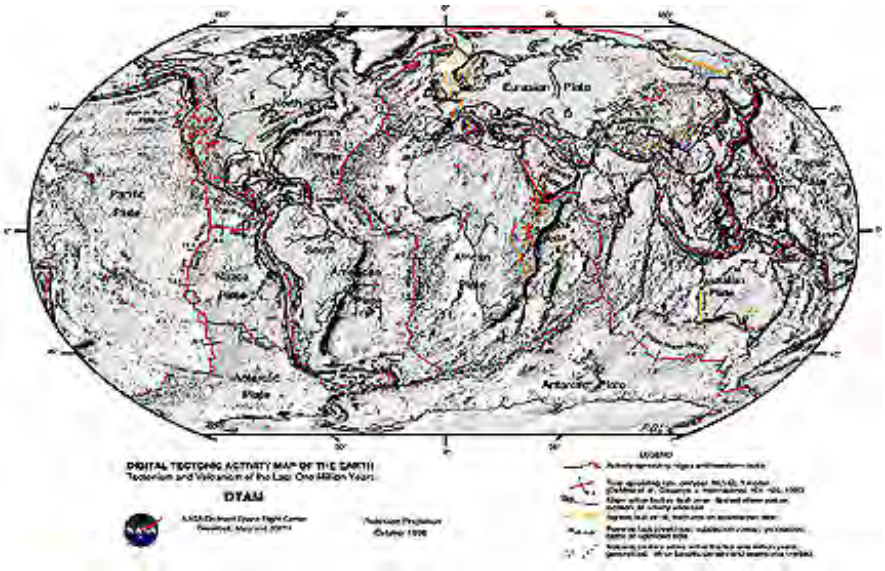


Goddard Space Flight Center VLBI solution CM 1122 version 02
NUVEL1A-NNR reference frame.

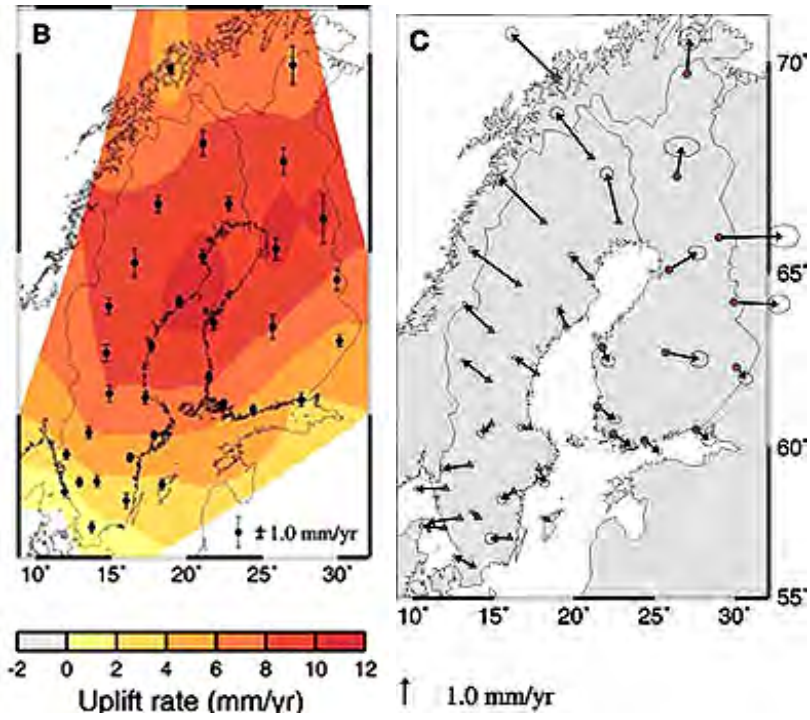
With accumulating years of measurements, the movement picture for the North American plate has improved in detail. Here is one of the recent plots concentrating on plate shifts as measured at the stations indicated:



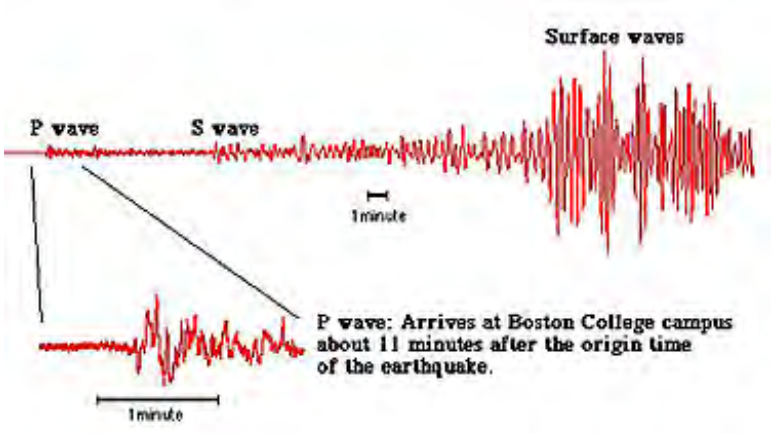
One of the by-products from NASA Goddard's Geodynamic Branch work on crustal deformation is the splendid plate tectonics map made by Dr. Paul D. Lowman, Jr. (author of Section 12 of this Tutorial). It now shows plate locations, plate velocities, and volcanic activity for the last one million years. We show this important map below but because at this smaller size it is unreadable we offer the option of looking at a large, more readable version by clicking here (to return to this page, hit your Back button).



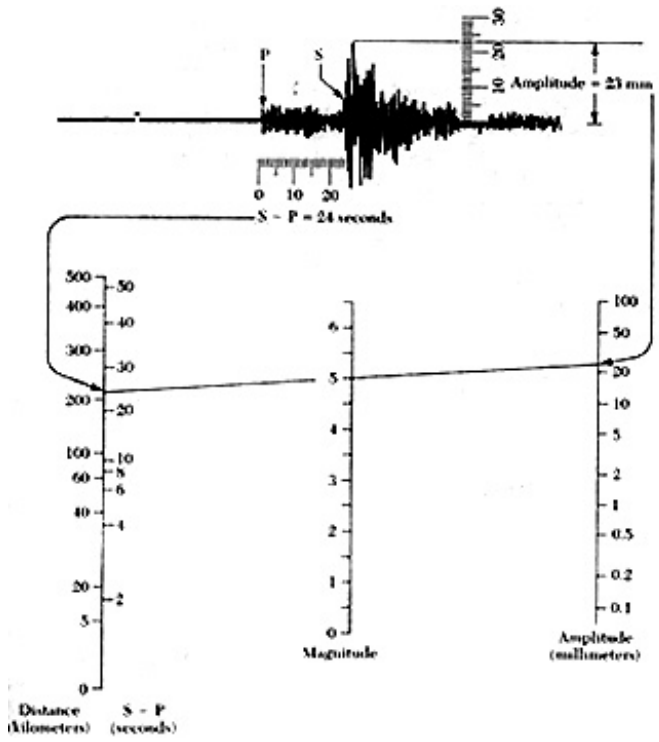
Crustal movements of another kind have now been further detailed using positional information obtained from ranging to the Global Positioning System array of satellites. Parts of the Earth's crust were depressed (pushed downwards) from the weight of several kilometers thickness of ice at various times during the Pleistocene Ice Ages. As the more rigid upper crust bends down, the more plastic lower crust/upper mantle is driven sidewards to make room. When the ice sheets melt, there is a slow, generally steady upwards rebound that is now being measured by GPS and other satellites. A group of scientists at the University of Toronto and other institutions have recently reported on results obtained for the Fennoscandinavian crust. The illustration on the left shows measured vertical rebound; that on the right indicates that these workers also detected some horizontal movement outward from a depression center.



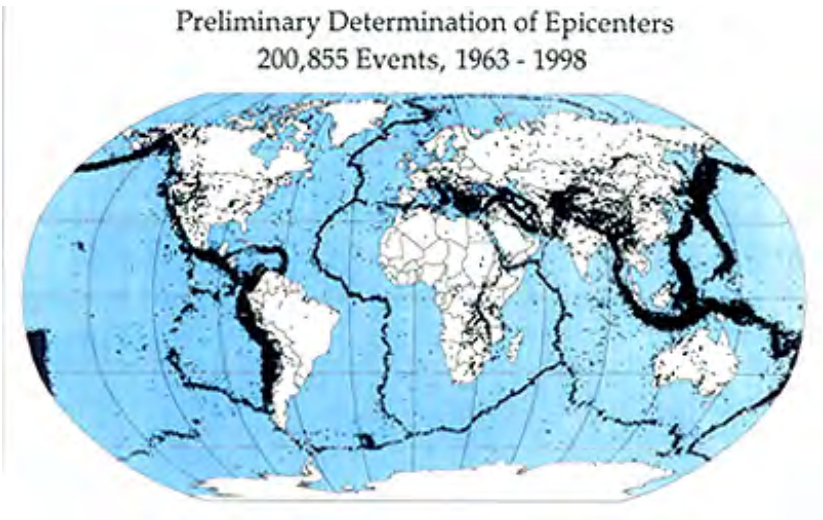
The third major subdivision of applied geophysics, Seismology, obviously does not benefit directly from air and space observations. But it is clearly a mode of remote sensing, similar to radar in measuring travel times of seismic waves. Seismometers must of necessity be on the surface (or in drill holes), preferably mounted on bedrock, to pick up the various kinds of seismic disturbances, especially from earthquakes. The seismic waves are oscillations of several types - P or primary waves (push-pull [compressional], analogous to acoustic wave motion), S or secondary waves (sinusoidal, also called shear waves), and others, such as surface Love and Rayleigh waves. Each wave type travels from its source to a seismic monitoring instrument (at a earthquake receiving station or, in geophysical surveying, in shallow holes near the point where wave-producing explosive detonations are set off) at some velocity, dependent on its depth and the seismic properties of materials along its particular path to the station. The oscillations at the station are recorded as "wiggles" or tracings on a seismogram. Here is a seismogram recorded at the Weston Station of Boston College for a recent earthquake in Turkey. The seismic waves traveled first through the crust under Turkey, then the Upper Mantle, and finally through the crust of the North American Plate under Boston.



Much can be learned from a seismogram: the distance to the earthquakes epicenter (point at the surface about its zone of origin); depth to the origin; magnitude or intensity (and kinetic energy released); duration; direction of movement; and often the type of fault along which the earthquake movement results in strain release. Each station has developed nomographs that can be used to solve for some of these parameters. Examine this one:

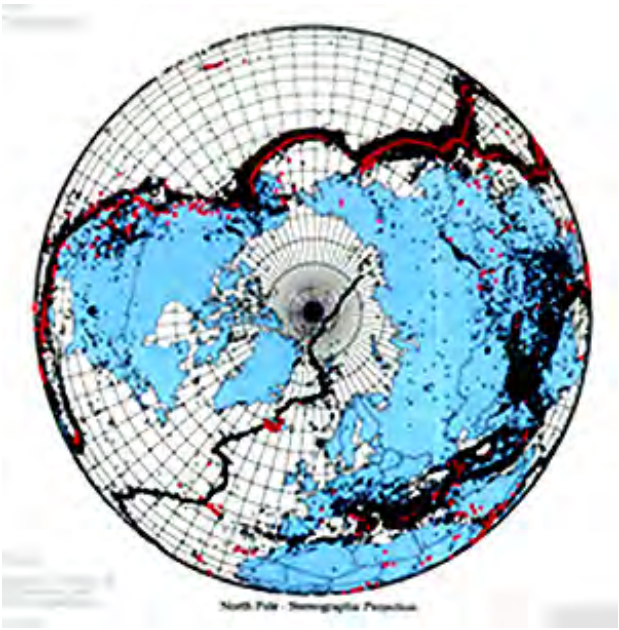


The distance to the epicenter can be in any direction (360°) from the station. But when distance information from three stations is used to draw a circle from each with the radius taken as its distance to the epicenter, there will emerge a narrow zone of intersection that defines the epicenter common to the three; thus:

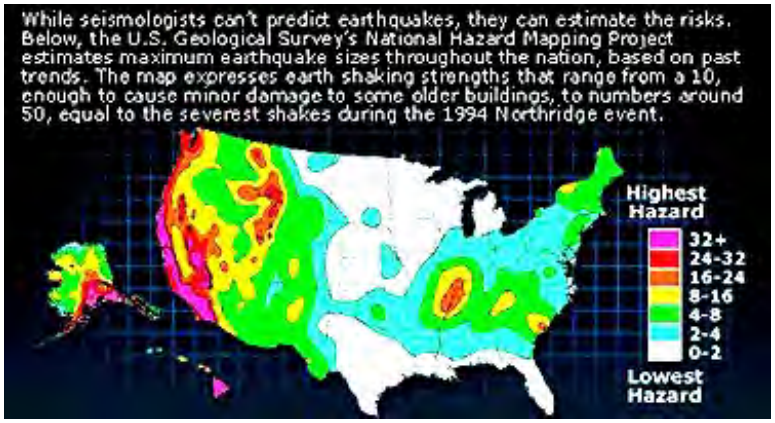


Seen from the viewpoint of looking down at the northern hemisphere (North Pole at center), and

with zones of active volcanism in red and continents in blue, the pattern again reveals that most earthquakes and volcanic eruptions occur at plate boundaries.

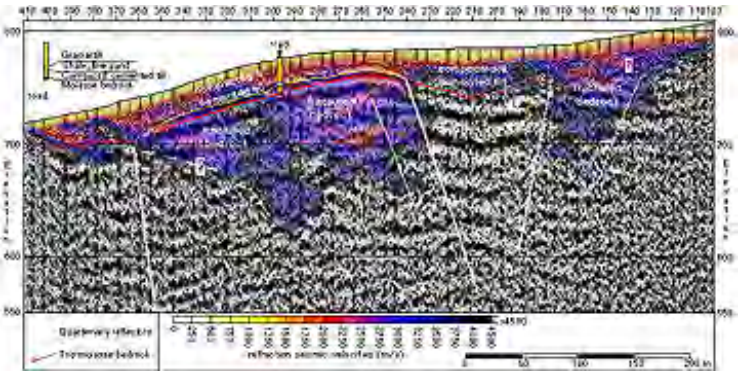


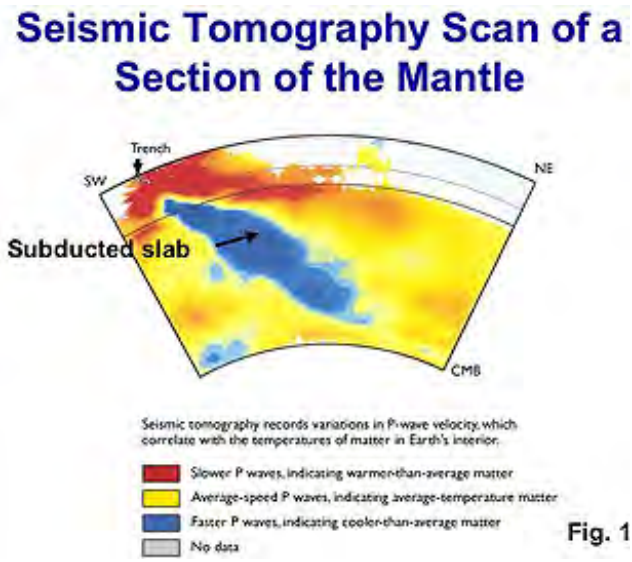
Such plots help to define zones or regions of higher earthquake activity. This distribution plus information about the nature and extent of damage allows estimates of seismic risk or likelihood of degree of destruction within some time span. This risk map has appeared in the newspaper *USA Today* shortly after an earthquake in Seattle, Washington in late February of 2001.

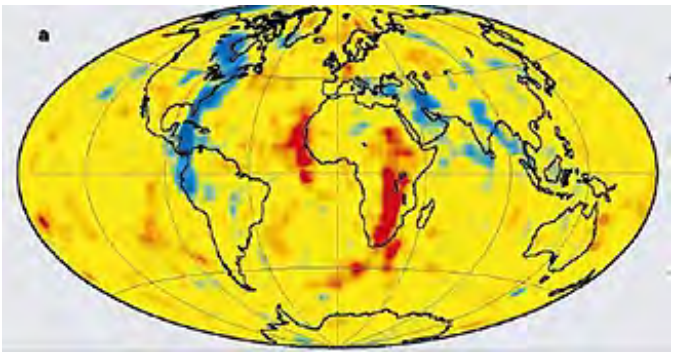


Another aspect of seismology is the technique called seismic tomography. In general, the term "tomography" refers to a representation in cross-section in which neighboring 2-dimensional cross-sections are combined to provide a 3-dimensional model. Computer-assisted tomography (CAT) as used in medical diagnosis is well-known as a non-invasive method of examining internal organs for abnormal regions (see page I-26c). X-rays or ultrasonic waves are absorbed unequally be different materials, and computer-aided tomography consists of studying the *attenuation* of X-rays, Gamma-rays, or ultrasonic waves that pass through the body in distinctly controlled planar sections. Seismic tomography uses the same principles, with the difference that the *travel-times* of the signals, rather than their attenuation, are observed. Seismic tomography involves the 3-dimensional modeling of the velocity distribution of seismic waves in the Earth. Sophisticated computer programs analyze arrival time differences to construct, in effect, slices into the Earth which when stacked together yield 3-D images.

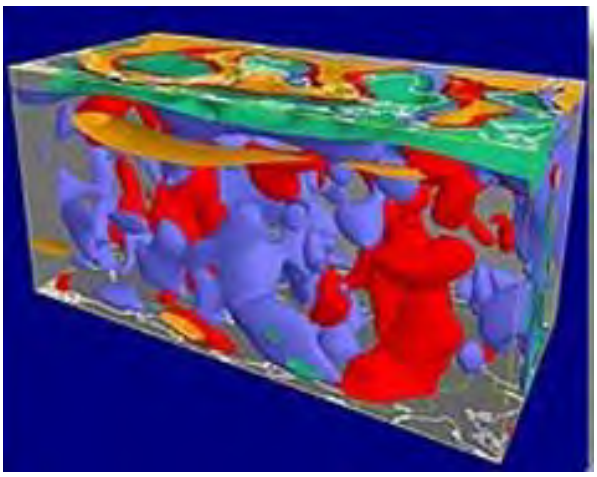
Seismic tomography is at the leading edge of research into the nature of the crust, mantle, and core of Earth. It is too specialized to be considered in this brief survey. But these several figures may add a bit to your understanding. The first shows a series of seismic reflection waves beneath a small segment of crust; materials known to be at and just below the surface are indicated







Tomographic analysis can integrate seismic velocity variations into this type of 3-D representation in which the bodies in blue and red are also indicative of lithologic materials and/or structures within the mantle:

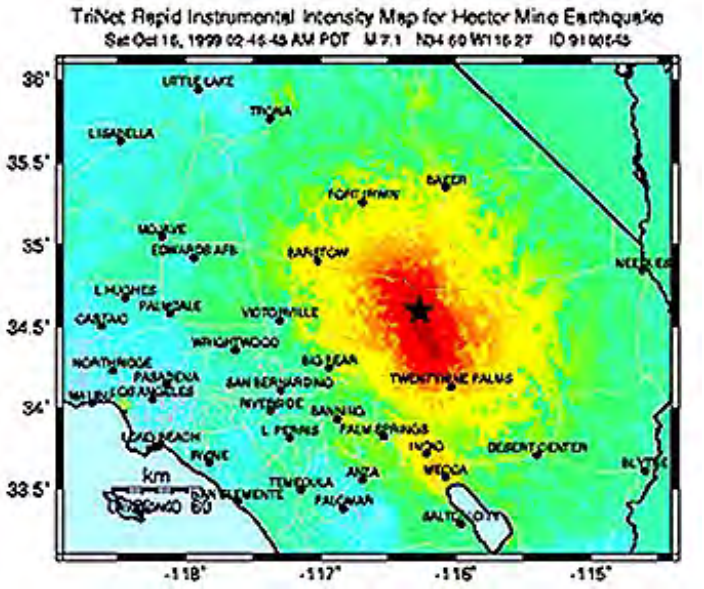


As said above, satellites cannot directly record earthquake (seismic) events. However, some seismometers are located in isolated places where direct telephone lines do not reach.

Transmitters at the field station can telemeter data to satellites overhead that then relay the recorded seismic signals to receiving stations. Magnetic or gravity measuring satellites do reveal much about the Earth's interior which is of use in interpreting internal conditions related to plate motions that are factored in models used to analyze earthquake mechanisms.

There is one other use of satellites that bears on earthquake effects. This employs radar interferometry to examine displacements of the surface (SLR and altimeter profiling also can contribute to determining changes in surface position and elevations brought on by an earthquake). Color fringes produced from interferometric data indicate the distribution of changes in the surface (data from a pre-earthquake passage of the satellite are used to work out the differences in position).

An excellent case study which used this technology is the 7.1 magnitude Hector Mine earthquake of October 16, 1999 40 km (25 miles) northwest of Barstow, California (edge of Mojave Desert). This map shows the intensity distribution (relates to how people perceived the quake) around the epicenter.

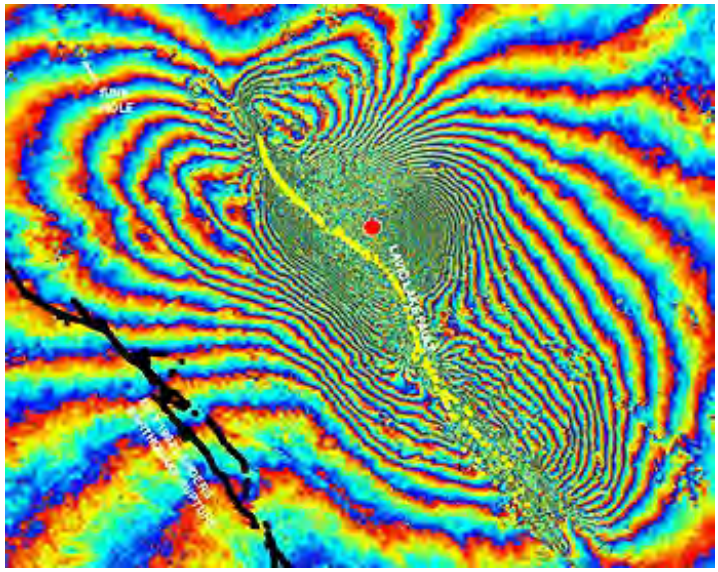


This quake occurred in what is known as the Mojave shear fault zone. Lateral displacement along this predominantly strike-slip fault ranged from 3.8 to 4.7 meters (almost 15 ft). Surface rupture extending for 41 km can be seen in this aerial photo taken near the mine; the offset road (right) reveals this displacement.

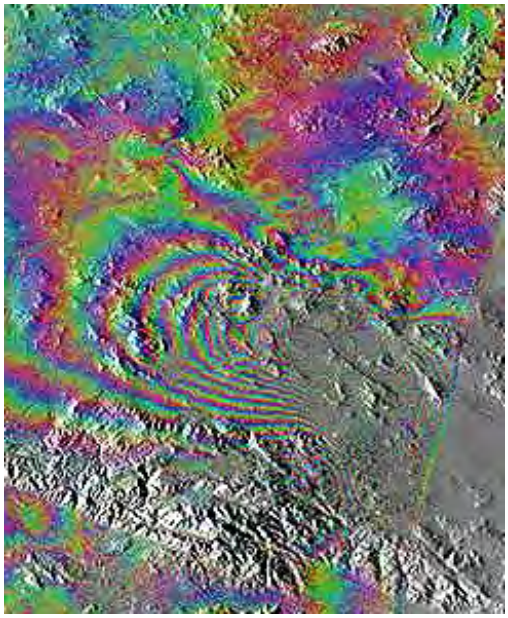
Photo courtesy of Paul "Kip" Otis-Diehl, USMC, 29 Palms

Radar interferometry from aircraft and spacecraft have provided useful information on vertical ground displacements (see page 11-10 that may be precursor hints of impending events (this also applies to volcanoes where eruptions are accompanied by ground swelling [Mt. St. Helens

case study; page 3-7]). Here is an interferogram (displayed by color fringes) showing displacement strain along the surface around the Lavic Lake fault as determined from analysis of SAR data from the ERS-1 satellite, soon after the day on which the Hector Mine earthquake was triggered:



A similar plot, superimposed on an image background, shows vertical displacement after an earthquake near Landers, California in the Mojave Desert. This occurred in 1992 within the same shear zone as the Hector Mine earthquake.



This subsection on geophysics as a component of remote sensing in its broadest meaning is, from what you have seen, a worthy diversion from the main theme. We move on to that theme - the more customary types of remote sensing - in the rest of this Introduction and the Sections that follow, first by investigating the role of the *photon* and its place in determining the characteristics of the *electromagnetic spectrum.*

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

The Introduction, replete with images and illustrations, is designed to cover the meaning embodied in the concept of "remote sensing", some of the underlying principles (mainly those associated with the physics of electromagnetic radiation [other related topics are deferred until Sections 8 and 9]), a survey of the chief satellite programs that have depended on remote sensors to gather information about the Earth, and some specialized topics . Emphasis is placed on the Landsat series of satellites that, starting in 1972, have provided a continuous record of the Earth's land (and some ocean) surfaces using the multispectral approach. In this Introduction, and most of the Sections that complete the Tutorial, as well as several of the Appendices, each page will be individually summarized at the top and all illustrations will have captions accessible by clicking at the lower right of each display.

The page you are now on once again defines the term "remote sensing", develops a brief discussion of implications, and places limits on its meaning. The meaning of the term "image" is discussed. The page also draws distinctions between what are the usual areas of application (confined to measurements at selected wavelengths in the electromagnetic spectrum) and what can more conventionally be called geophysical applications which measure particles and fields.

INTRODUCTION:

The Concept of Remote Sensing; Sensors

If you have heard the term "remote sensing" before you may have asked, "what does it mean?" It's a rather simple, familiar activity that we all do as a matter of daily life, but that gets complicated when we increase the scale at which we observe. As you view the screen of your computer monitor, you are actively engaged in remote sensing.

A physical quantity (light) emanates from that screen, whose imaging electronics provides a source of radiation. The radiated light passes over a distance, and thus is "remote" to some extent, until it encounters and is captured by a sensor (your eyes). Each eye sends a signal to a processor (your brain) which records the data and interprets this into information. Several of the human senses gather their awareness of the external world almost entirely by perceiving a

variety of signals, either emitted or reflected, actively or passively, from objects that transmit this information in waves or pulses. Thus, one hears disturbances in the atmosphere carried as sound waves, experiences sensations such as heat (either through direct contact or as radiant energy), reacts to chemical signals from food through taste and smell, is cognizant of certain material properties such as roughness through touch (not remote), and recognizes shapes, colors, and relative positions of exterior objects and classes of materials by means of seeing visible light issuing from them. In the previous sentence, all sensations that are not received through direct contact are remotely sensed.

I-1 In the illustration above, the man is using his personal visual remote sensing device to view the scene before him. Do you know how the human eye acts to form images? If not, check the answer. ANSWER

However, in practice we do not usually think of our bodily senses as engaged in remote sensing in the way most people employ that term technically. A formal and comprehensive definition of applied remote sensing * is:

Remote Sensing in the most generally accepted meaning refers to instrument-based techniques employed in the acquisition and measurement of spatially organized (for the Earth, most commonly geographically distributed) data/information on some property(ies) (spectral; spatial; physical) of an array of target points (pixels) within the sensed scene (anywhere in the Universe) that correspond to classes, features, objects, and materials, doing this by applying one or more recording devices not in physical, intimate contact with the item(s) under surveillance (thus at a finite distance from the observed target, in which the spatial arrangement is preserved); techniques involve amassing knowledge pertinent to the sensed scene (target) by utilizing electromagnetic radiation, force fields, or acoustic energy sensed by recording cameras, radiometers and scanners, lasers, radio frequency receivers, radar systems, sonar, thermal devices, sound detectors, seismographs, magnetometers, gravimeters, scintillometers, and other instruments.

I-2 To help remember the principal ideas within this definition, make a list of key words in it. ANSWER

This is a rather lengthy and all-inclusive definition. Perhaps two more simplified definitions are in order: The *first*, more general, is built on this idea: Remote Sensing involves gathering data and information about the physical "world" by detecting and measuring signals composed of radiation, particles, and fields emanating from objects located beyond the immediate vicinity of the sensor device(s). The *second* is more limited as to the nature of the carrier signal but is expressive of most of the subject matter (targets about which information is sought) of this Tutorial: In its common or normal usage, Remote Sensing is a technology for sampling ***electromagnetic radiation*** comprising a signal emanating from its source target that is used to acquire and interpret non-contiguous *geospatial data* from which to extract information about features, objects, and classes on the Earth's land surface, oceans, and atmosphere (and, where applicable, on the exteriors of other bodies in the Solar System, or, in the broadest framework, celestial bodies such as stars and galaxies).

I-3 What is the meaning of "geospatial"? Are there any differences in meaning of the terms "features", "objects", and "classes"? ANSWER

The term "remote sensing" is itself a relatively new addition to the technical lexicon. It was coined by Ms Evelyn Pruitt in the mid-1950's when she, a geographer/oceanographer, was with the U.S. Office of Naval Research (ONR) outside Washington, D.C. It seems to have been devised by Ms Pruitt to take into account the new views from space obtained by the early meteorological satellites which were obviously more "remote" from their targets than the airplanes that up until then provided mainly aerial photos as the medium for recording images of the Earth's surface. No specific publication or professional meeting where the first use of the term occurred is cited in literature consulted by the writer (NMS). Those "in the know" claim that it was verbally used openly by the time of several ONR-sponsored symposia in the 1950s at the University of Michigan.

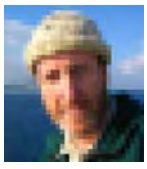
Applied remote sensing involves the detecting and measuring of electromagnetic energy (normally in the form of photons) emanating from (typically distant) objects made of various materials, by which means the user can identify and categorize these objects - usually presented/depicted as **images** - by class or type, substance, and spatial distribution. Generally, this conventional description of what the action of remote sensing produces points to the specific use of the term "images" - the most common and practical output. Dictionaries provide multiple definitions of "image'. The two at the top of the list are 1. a physical likeness or representation of a person, animal, or thing, photographed, painted, sculptured, or otherwise made visible, and 2. an optical counterpart or appearance of an object, as is produced by reflection from a mirror, refraction by a lens, or the passage of luminous rays through a small aperture and their reception on a surface.

For our purposes, an image is a pictorial representation of a scene (any collection of contiguous objects that give off or reflect visible light or other modes of electromagnetic radiation). An example of an image in a familiar format is the analog **photograph**, in which the scene is recorded in film (usually represented by an emulsion in which different shades of gray are correlated with varying amounts dark silver grains or, in color, with several dyes that have specific colors). A TV screen's picture is another example of an image. It is made up of a collection of tiny points of emitted light arranged in an X-Y two dimensional array. These points are called **pixels** - a mnemonic term for "picture elements".

A necessary digression now - just what is a pixel. We will try to explain this important concept that bears upon creating an image by the following discussion (more aspects of what is involved appear later in this Section). (You may also want to read the Wikipedia entry given as pixel.) Let us start by showing this illustration:

What you see is a familiar mode of artwork: a mosaic made of individual tile squares that vary in color. The composite array of tile squares is interpretable as an image - of a young woman. The tile squares are analogous to pixels.

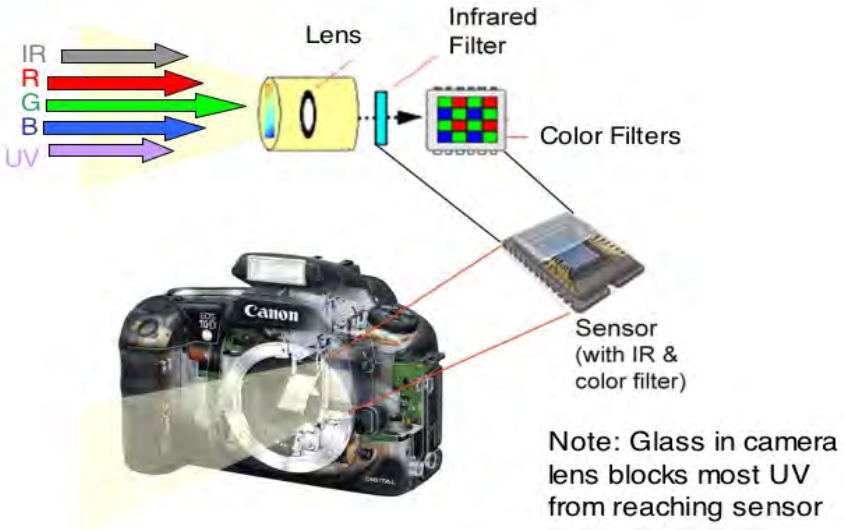
Now look at this image, again, an array of squares. The large version is almost uninterpretable because the squares are too big. But the small version beneath it is recognizable as a man's face. This recognition results when the individual squares are not noticeable, that is, in effect they blend together in one's minds eye so as to produce an intelligible object - a human face. This factor of "size" is related to the idea of "spatial resolution", a subject for later in this Section.



Those three illustrations should help to fathom what a pixel is and does in building up an image. Lets shift our thoughts to this exposition of how pixels are involved in remote sensing. To do this we will keep our sensor stationary; the digital camera so common today is the instrument that will produce the image (this camera is described later in this Section at the bottom of page 5a). A simplified diagram of a digital cameral appears below:

Inside the Digital Camera

How does it detect light?



The light detector in a digital camera (shown as the "sensor" in the above illustration) uses charge coupled devices (CCDs; check this Wikipedia website for details about them), contained within small silicon chip-based sensors that rely on the photoelectric effect (discovered by Einstein in 1905) in which photons of light expel electrons from certain metals that can be gathered into a current that is involved in creating the image. The analog signal thus formed can be digitized; the image is displayed immediated on a small screen (back of camera) or recreated later by a computer program into some other display format (a conventional photo). We will ignore at this point just how spectral information (different colors) is utilized in making a color image. A digital camera can contain millions of CCDs (the pixels; cameras are rated in units of "megapixels") arranged as an X-Y array in parallel columns and rows in what amounts to the image detector.

We now point the camera at a scene. Coming from that scene are huge numbers of photons of different energy (that is, they embody different wavelengths that discriminate the various colors). These photons are scattered in different directions. The camera is located along some direction relative to the scene; the scene may be illuminated from still a different direction (e.g., the Sun) or from the camera itself (flashbulb); this duality between angle of incoming illumination and angle of outgoing reflection to the camera is called bidirectional reflectance (page 3 of this Section). A tiny sample of the scene-emanating photons is passed into the camera using a lens which optically converges the light rays from all parts of the scene onto the CCD array. Each individual CCD corresponds to a small area in the scene being recorded; adjacent CCDs correspond to adjacent areas of the scene. Variations of light intensity and of spectral content (visible colors and/or radiation beyond the visible, e.g., infrared; ultraviolet or even other regions of the electromagnetic spectrum [introduced two pages hence]) when rendered into a display medium (a screen or a photograph) that has its X-Y geometry in correspondence with the scene X-Y geometry produces the image. The size of the individual CCDs is tied to the size (area) of the scene it represents; smaller CCDs record smaller scene areas (thus, the size controls the resolution or size of individual objects in the scene that can be resolved or distinguished); the larger the number of megapixels in a camera the sharper (better resolution) will be the image it makes. If the individual CCD is larger, it records a larger "piece of real estate", so that if that ground area contains several different features or classes, these may not be distinguishable (visually not separable), producing "mixed pixels" (see page 2 of Section 13).

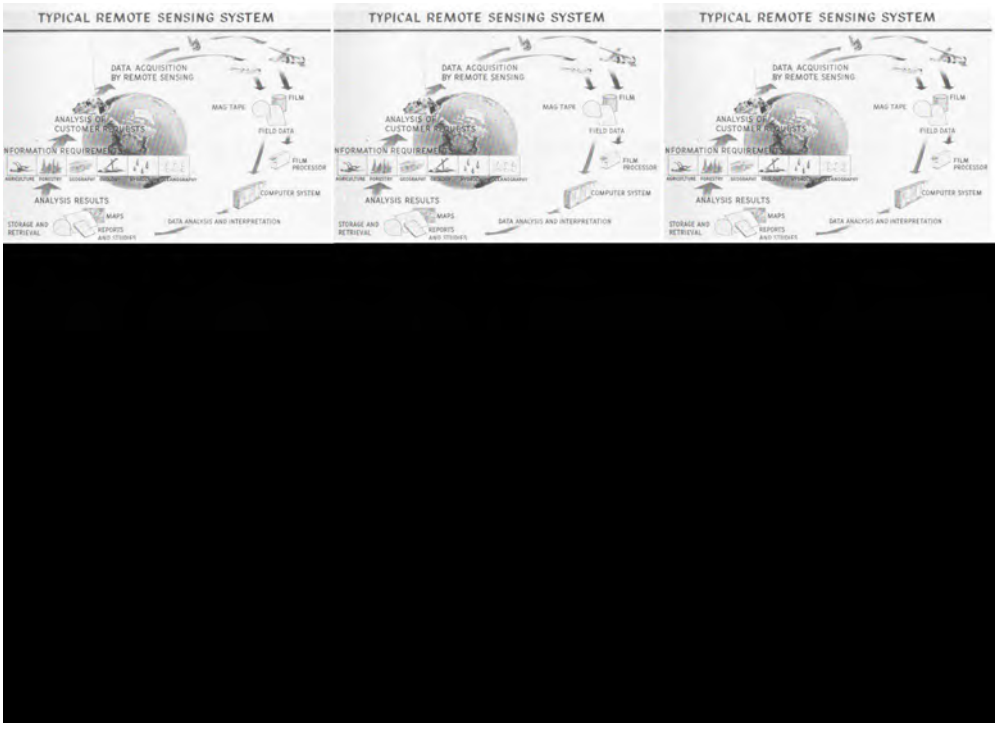
More on CCDs, considered from an electronics viewpoint and from their role in setting an image's resolution, is included on page 5a of this Introduction.

The digital camera is a sophisticated remote sensing instrument familiar to most of us. But

sensors that operate on aircraft and spacecraft have to operate under more challenging conditions that require additional capabilities. Unlike a digital camera that normally takes a picture and then awaits the next opportunity - moments to hours later - without the operator necessarily repositioning to a new location, sensors on moving platforms above the Earth have to continuously make observations of new segments of the scene below. This requires a means for constantly downloading the previously acquired segment of imagery signals, refreshing the data storage component, and sensing the next segment coming into view. The target below acts as though it is in steady motion! To function effectively, the sensor (which senses the target either by scanning to and fro or by using an array of CCDs) must continuously unload the data it acquires. This is feasible and effective because modern electronics are capable of facilitating very rapid discharge and refresh of their imaging device.

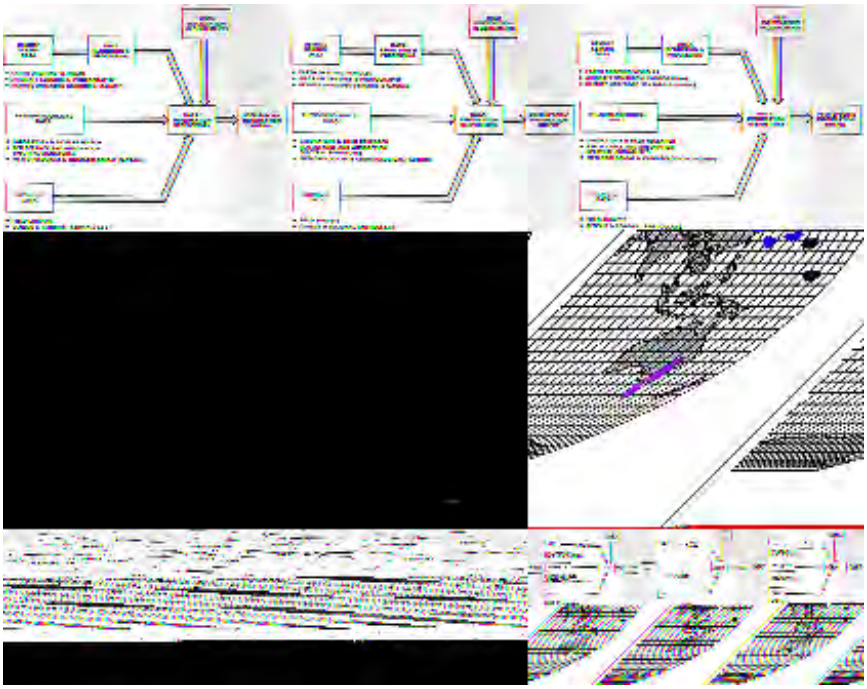
But there is more to remote sensing outputs than just making "pictures". In addition to images, the data often can also be shown as "**maps**" and "**graphs**", or to a lesser extent, as digital numbers that can be input to computer-based analysis, and in this regard are like the common data displays resulting from geophysical remote sensing. As applied to meteorological remote sensing, both images (e.g., clouds) and maps (e.g., temperature variations) can result; atmospheric studies (especially of the gases in the air, and their properties) can be claimed by both traditionalists and geophysicists. Note that in this paragraph, we restrict remote sensing to the production of images using electromagnetic radiation. However, a case can be made that in the broadest sense at least three other types of signals emanating from objects not in direct contact with the detecting sensor - namely sound waves, seismic waves, and magnetic force fields - should be included in the term "remote sensing". In this Tutorial, that argument is accepted but only a few pages are devoted to these special applications. (The aspects of remote sensing that are relevant to geophysical methods are discussed near the end of this Section, along with the special topic of medical "remote sensing".)

Suffice to say that remote sensing is a tool for gathering information, usually about what is at the surface of Earth or planets and their moons, but also about the atmosphere, and as indicated in Section 20, about celestial bodies. Remote sensing data are commonly combined with other kinds of data (typically, from field or "on-the-spot" studies [commonly called ground truth]) to act together as a system (for example, Geographic Information Systems, or GIS, as discussed in Section 15). This next diagram indicates the basic elements involved in an earth survey system:



A variant of this diagram emphasizes the idea that the ultimate goal in applied remote sensing is

determined and driven by user requirements. It also shows that in most cases remote sensing data are integrated with other types of data, such as phenomenological and archival.



As said earlier on this page, remote sensing is conducted by means of sensors - in effect, instruments that can pick up (sense) objects and features not in contact with the sensors, by detecting, coordinating, and recording electromagnetic radiation or acoustical (sound) waves. The best known (to the average individual) sensor is the human eye and the human ear. An artificial or manmade instrument familiar to all is the photographic camera (and in the last 80 years, the television camera). Since the beginning of the era of satellite remote sensing, more sophisticated instruments include multispectral scanners, radar, lidar, thermal imagers, and other systems treated later in this Section. Specialized sensors used in geophysics and in medical diagnosis are also considered in Section 1.

As the above comments suggest, some technical purists arbitrarily stretch the scope or sphere of remote sensing to include other measurements of physical properties from sources "at a distance" that are more properly included in the general term "Geophysics". (Geophysics has a scientific connotation: it is pertinent to the study of the physical properties of Earth and other planets. It likewise has an applied connotation: it is the technology often used to search for oil and gas and for mineral deposits.) This latter is especially conducted through such geophysical methods as seismic, magnetic, gravitational, acoustical, and nuclear decay radiation surveys. Magnetic and gravitational measurements respond to variations in force fields, so these can be carried out from satellites. Remote sensing, as defined in this context, would be a subset within the branch of science known as Geophysics. However, practitioners of remote sensing, in its narrower meaning, tend to exclude these other areas of Geophysics from their understanding of the meaning implicit in the term.

Still, space systems - mostly on satellites - have made enormous contributions to regional and global geophysical surveys. This is because it is very difficult and costly to conduct ground and aerial surveys over large areas and then to coordinate the individual surveys by joining them together. To obtain coherent gravity and magnetic data sets on a world scale, operating from the global perspective afforded by orbiting satellites is the only reasonable alternate way to provide total coverage. One could argue that Geophysics deserves a Section of its own but in the remainder of this Tutorial we choose to confine our attention almost entirely to those systems that produce data by measuring in the electromagnetic radiation (EMR) spectrum (principally in the Visible, Infrared, and Radio regions).

It may seem surprising to realize that going to the doctor can involve remote sensing. Most

obvious, on a miniature scale, is listening to a heartbeat using the stethoscope. But in the field of modern medical technology, powerful, often large, instruments such as CATscans and Magnetic Resonance Imaging, are now almost routinely used for non-invasive subskin investigation of human tissue and organs. This is indeed another major application of remote sensing that will be surveyed on pages I-26c through I-26e.

Another mode of remote sensing treated near the end of this Section of the Tutorial (on the page dealing with remote sensing of the ocean floor) is acoustic monitoring of sound waves in atmospheric and marine environments. For example, volcanic eruptions or nuclear (testing) explosions can be detected by sensitive sound detectors. Sonar is used to track submarines and surface ships in the oceans. Sound through water are also involved in listening to marine animals such as whales and porpoises and in locating schools of fish. (Acoustic Remote Sensing will not be exemplified anywhere else in the Tutorial.)

The traditional way to start consideration of what remote sensing is and means is to set forth its underlying principles in a chapter devoted to the Physics on which remote sensing is founded. This will be done in the next 6 pages, beginning with a page that considers platforms on which the sensors are mounted and their launch vehicles. These pages contain the "technical jargon" that remote sensing specialists like to banty about. With this caveat in mind, work through the pages, try to understand the esoteric, and commit to memory what seems useful. Then, progress through the subsection that deals with the history of remote sensing. This Section closes with treatment of some special applications of remote sensing; some of the topics will not be revisted anywhere else in the Tutorial.

At this point the writer (NMS) offers the option of clicking here to go to a brief page that describes how he (unexpectedly) entered the field of remote sensing that has been his dominant science/technology career activity since 1969. (In effect, it outlines his history that underwrites his qualifications for producing this Tutorial.)

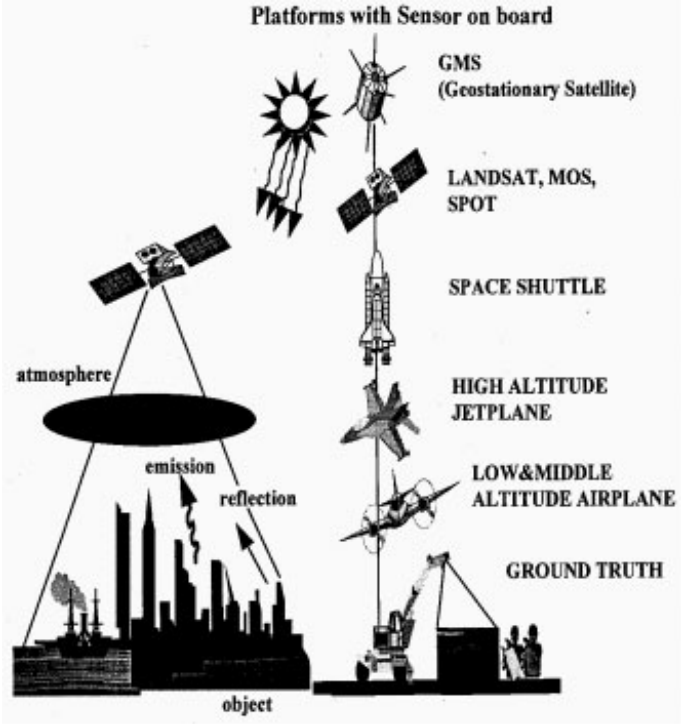
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Remote Sensing is conducted from a variety of platforms, such as satellites, space stations, aircraft, and planetary probes. These are discussed on this page. The use of rockets to get the platforms into orbit or to distant planets is then considered in some detail.

Platforms used by Remote Sensors: Aircraft; Balloons; Satellites; Spacecraft; Probes; Rovers;; Launch Vehicles.

Remote Sensing involves four basic inputs: 1. The Target; 2. The Platform; 3. The Sensor(s); and 4. The Signal (usually electromagnetic radiation or acoustical waves). The target is comprised of the features or materials being sensed. Sensors and signals (including the concept of signatures) are treated later in this Section. On this page we will concentrate on the topic of Platforms, the structures that house or support the sensors. In most instances, the platforms will be in motion; by moving they automatically proceed to new positions from whence they sense new targets. A satellite orbiting the Earth is a typical platform. Most of the platforms involved are pictured in this diagram:



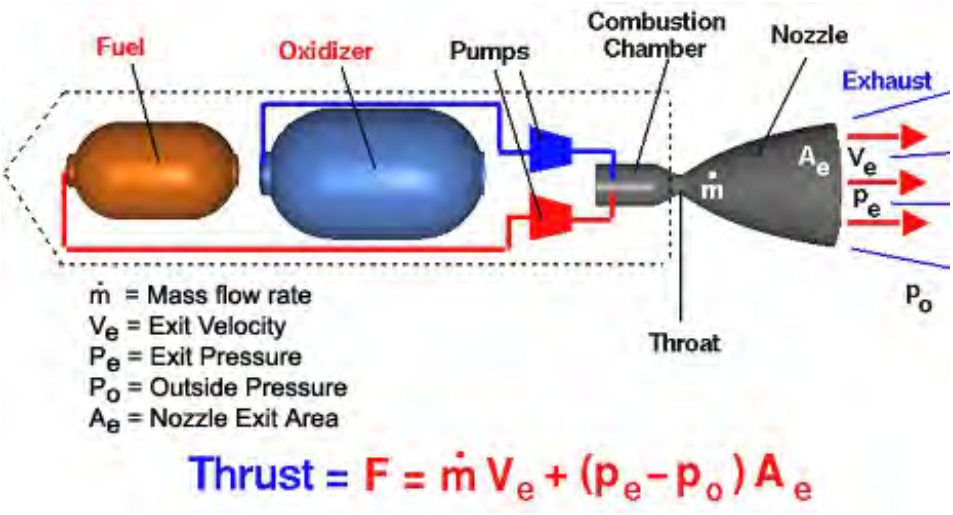
Those platforms above Earth's atmosphere and beyond include satellites, spacecraft, space stations, all in orbit around Earth, and planetary satellite orbiters, flyby satellites, probes, and landers/rovers that are used to explore planets, moons, and asteroids/comets.

We concentrate on those that function within the atmosphere, in orbit around Earth (including ones that look back at Earth and ones that peer outward at planets, stars, and galaxies), and from systems that are sent to explore other regions of the Solar System. We mention here only aircraft (planes and helicopters), blimps, and balloons, on which remote sensors have been mounted; the principles of operation are the same as those on platforms in Earth orbit and beyond.

The next few paragraphs will describe how rockets are used to get platforms into space. (It is assumed that the reader already understands how balloons are launched (they contain gases that are lighter in density than air, and thus rise, and they can be propelled by motors; and how airplanes fly aerodynamically by getting differential lift from air flowing above and below wings and forward motion from propellers or jets)

What we wish to consider briefly now is the physics and mechanics of launching platforms into space. The key ideas are found in these three websites: Orbital mechanics, How satellites are launched and Principles of rockets After reading through these Internet sources, take into account the next comments.

Satellites and probes are launched by rockets (as is the Space Shuttle, which uses rockets for initial boost [these are dropped off] and then its own rocket engine for further propulsion). Rockets are propelled by the burning (combustion) of either solid or liquid (or both) fuels whose gaseous end products are allowed to escape into air or space, providing a thrust. The thrust is a force that propels the rocket forward; the amount of thrust is calculated by this equation:



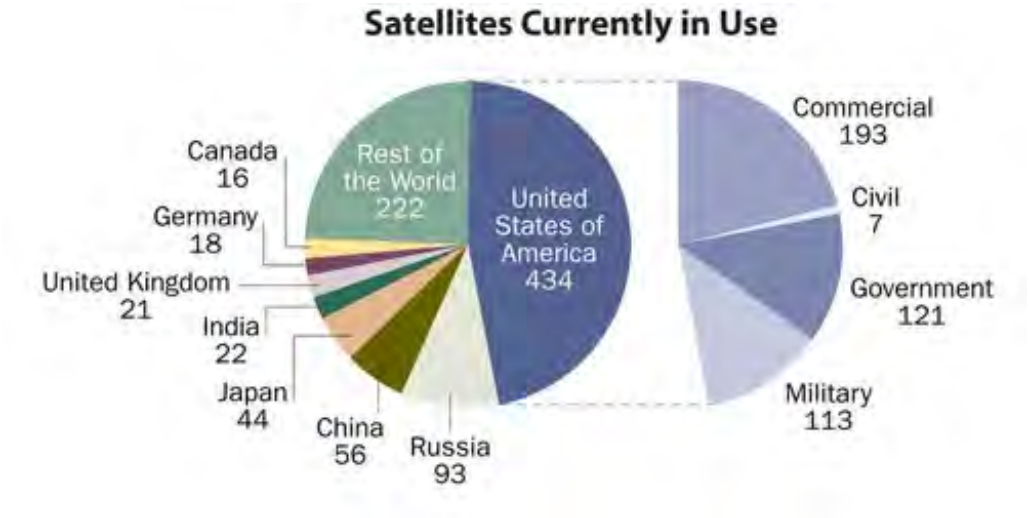
$$\text{Thrust} = F = \dot{m} V_e + (p_e - p_o) A_e$$

Two other factors are the amount of fuel and the time duration of the combustion. At any instant there is a specific velocity; the velocity increases systematically, causing the rocket vehicle to accelerate. At the time of engine combustion turnoff, some terminal velocity will have been achieved. Depending on that velocity, the rocket and its payload (such as a satellite) will either fall back to Earth, or go into Earth orbit, or eventually escape Earth's gravity and go into outer space.

To achieve a circular orbit, the payload must travel at 17,000 mph (27,359 kph). To leave Earth, the payload must accelerate to at least 25,039 mph (40,320 kph). It the desired orbit is to be geosynchronous, i.e., remain in orbit such that it seems to hover over the same local area on the Earth's surface, it must reach altitude of about 22,223 miles (~ 36000 km). The escape velocity from any body (such as a planet or moon) is calculated with this formula:

$$v_{escape} = \sqrt{\frac{2GM_{planet}}{R_{planet}}}$$

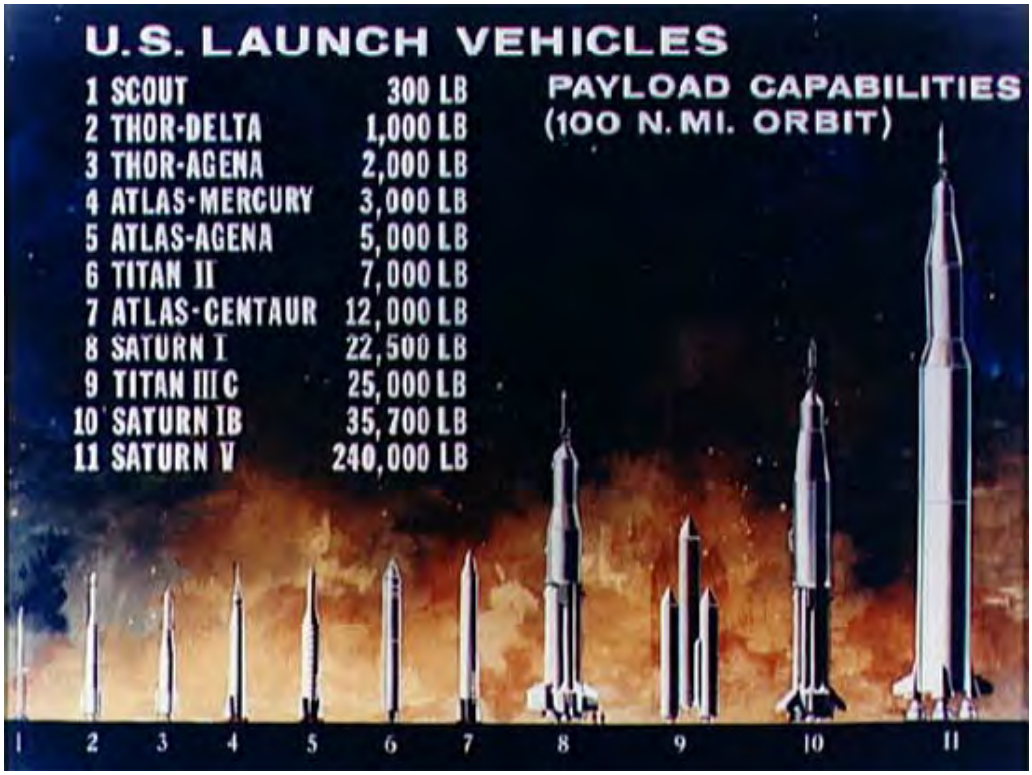
Since the opening of the Space Age in 1957, thousands of satellites have been orbited around the Earth. This is a general summary diagram:



Many of these satellites are described on this website: Remote Sensing satellites.. One trend in recent years has been to make satellites (as sensor platforms) smaller so as to cut the costs of launching. These microsatelites, as they are called, are discussed at this site:types of orbits

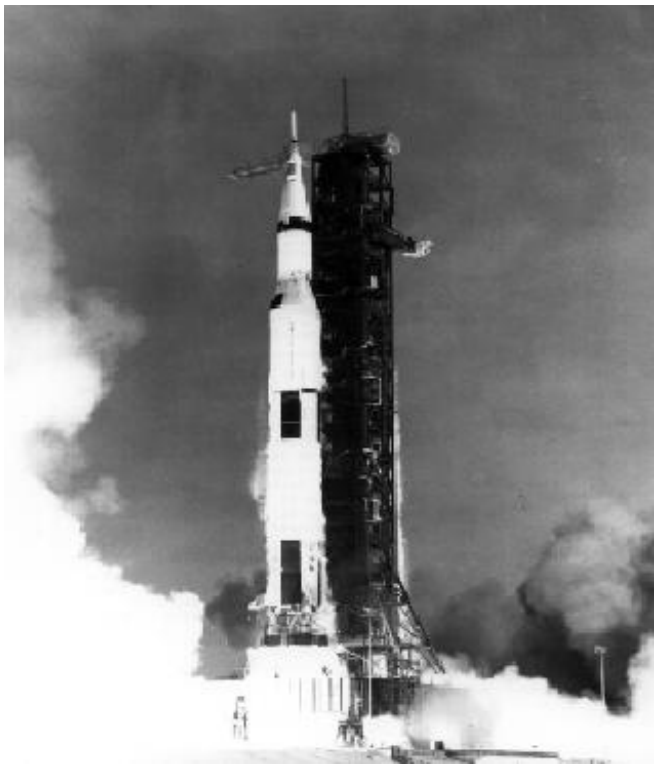
Next, we turn to the topic of "rockets" used to launch satellites and probes. There is a wide variety and range of rockets - some small, being used for suborbital flights (as sounders), others much larger, and some reserved for manned missions. The pictorial below shows many of the U.S. and International rockets (the title is misleading in that some U.S. rockets are included) that have been and/or are now being used to launch payloads or humans into space. Obviously, with this multitude, only a few will be described on this page.

As an example of the evolution of launch vehicles in the NASA program, note this diagram that depicts rockets developed during the early days of the space program:

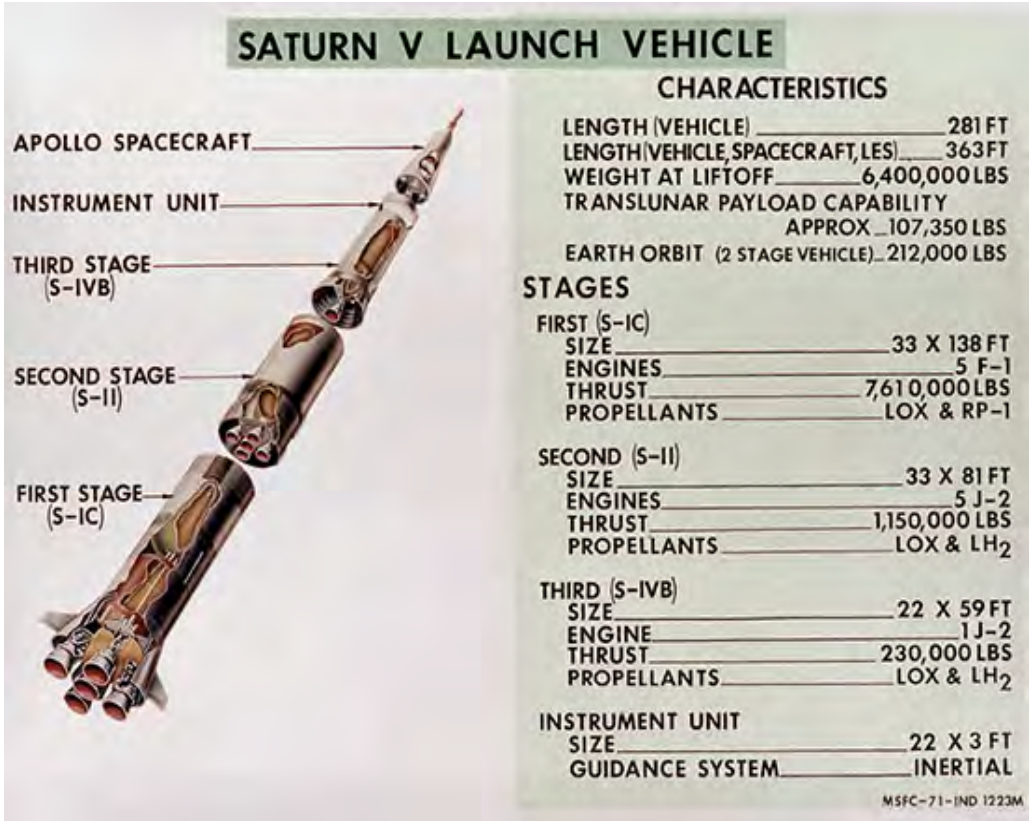


One of the earlier rockets, online in 1956, is Thor-Able, shown here:

We now call attention to six more modern launch vehicles, since there is no other page in this Tutorial that focuses primarily on this topic. The first is the mightiest rocket of them all - The Saturn V - which remains Wernher von Braun's greatest triumph. The top scene below shows this rocket (which traces its lineage back to the V-2) on its conveyor vehicle enroute to a launch pad at Cape Canaveral. The second photo portrays a "surplus" Saturn-V (never used after the Apollo program ended) vehicle on display at the Johnson Space Center in Houston; its massive size is evident from the people near it. The bottom image captures the famous moment when Apollo 11 lifted off for its historic journey to the Moon and back. Saturn V was 33 stories tall (113 m or 363 ft) and could deliver a thrust of 7.5 million pounds.



The specifications for the Saturn V are shown next:

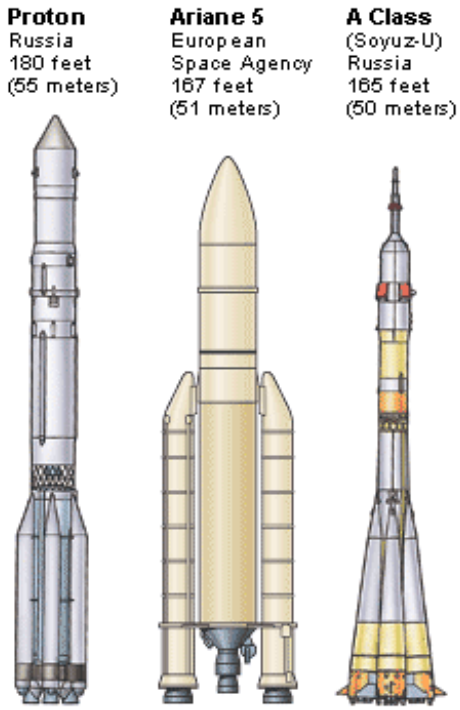


The two workhorse launch vehicles in the U.S. space program have been the Atlas and Delta rockets, each having been upgraded over the last 20 years. Both satellites orbiting Earth and spacecraft leaving for outer space have utilized these rockets.

The most recent of the Delta rockets built by Boeing is an advanced Delta 4, which is as tall as a 23 story building:

The U.S. Air Force's primary launch vehicle is the Titan.

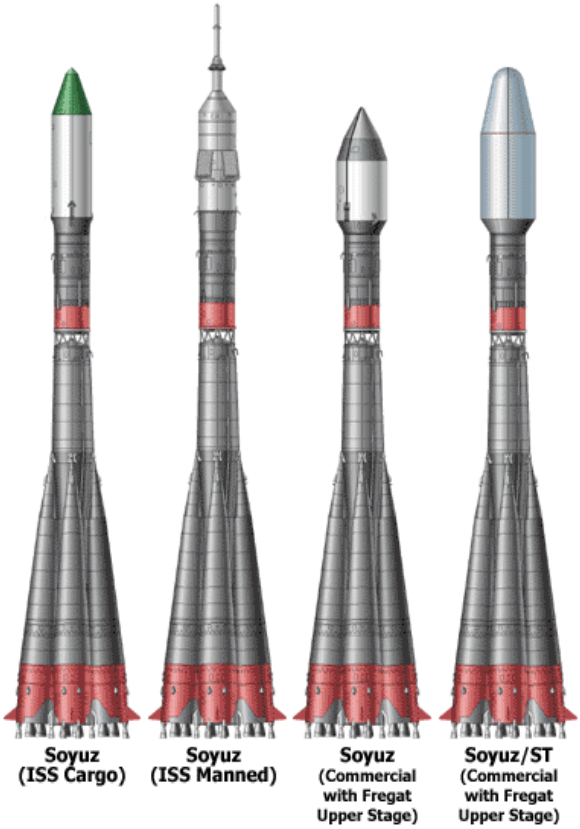
Other nations have built their own vehicles for varied purposes. The French Ariane rocket, operating out of Guiana in northern South America, is frequently chosen by companies launching commercial satellites. It is shown schematically below, along with two Russian rockets:



The Ariane 5 is the launch vehicle of choice for European Space Agency (ESA) launches. It is quite powerful, being able to put two large astronomical observatories (Herschel and Planck) into deep space orbit in May of 2009:

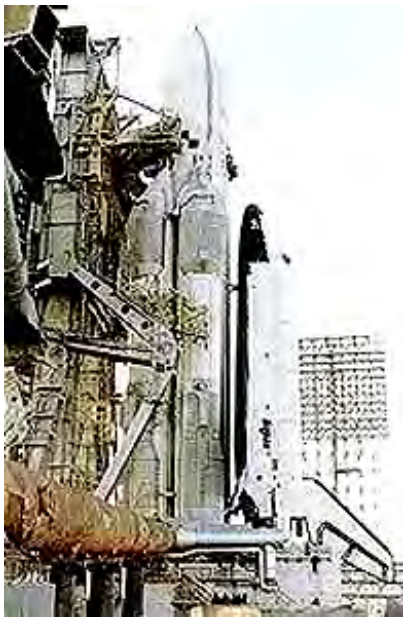
The Soviet/Russian rockets come in a wide variety of powerful vehicles. One group is the Soyuz

family:



The Chinese have come more recently into the space program. But they have developed a broad range of launch vehicles, as illustrated here:

The main launch vehicle for the NASA manned program is the Space Transport System (STS), more commonly known as the Space Shuttle. It uses two recoverable external fuel tanks plus its own central engine. The Soviets copied this vehicle (calling it Buran) but they use an Energia rocket (expendible) to launch it; very few flights have occurred, especially since it is now carried over into a Russian space program that has more limited funding. The Russians also use a Proton rocket to launch unmanned spacecraft.



As we move through the 21st century, the variety of launch vehicles proliferates. Launch sites include several in Asia, one in South America, and Cape Canaveral (Florida) and Vandenberg Air Force Base (Lompoc, California) as primary locations. A novel "portable" site, Sea Launch, is just starting operation. Based out of Long Beach, CA, and operated by a private consortium, this converted Norwegian drill platform can go almost anywhere in the open Pacific from which to launch. Then it returns to port to be refit with another rocket for the next event.

For those who may want to know more about space travel in general and launching in particular, we recommend this website: Rocket and Space Technology, which also includes other aspects of space history. More can be learned about launch sites, and launch vehicles themselves, by visiting this Space Today Internet site - it's loaded with information. A briefer review of launch procedures is presented at this Wikepedia site. A useful summary of the variety of rockets sent into space is found at this List of launch vehicles website maintained by Wikipedia.

PREVIOUS rst NEXT

Primary Author: Nicholas M. Short, Sr.

A new 7 spectral band scanner, called the Thematic Mapper, was/is the prime instrument on Landsats 4, 5, and 6. This improved sensor system includes a blue band so that quasi-natural color images can be made, two bands in the mid-IR, and a thermal band. System resolution is increased to 30 meters. Landsat 7 carried a single sensor, the ETM+, which, besides the 7 bands, also has a panchromatic band that achieves 15 m resolution. The ability of the TM to better identify materials and classes is demonstrated.

History of Remote Sensing: Landsat's Thematic Mapper (TM)

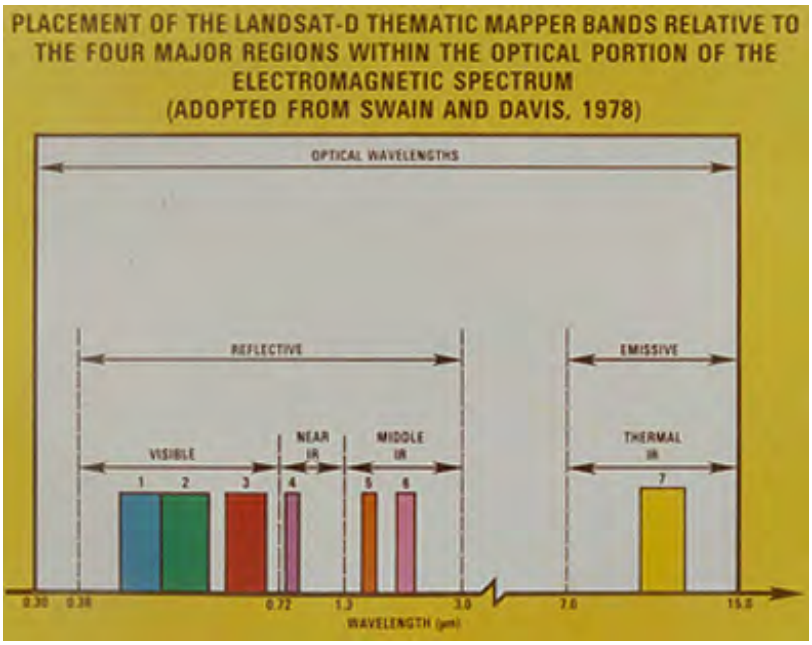
A more sophistical multispectral imaging sensor, named the Thematic Mapper (TM) has been added to Landsats 4 (1982), 5 (1984), and 6 (this last failed to attain orbit during launch and thus has never returned data) and a modified version to Landsat-7 (1999). These TMs flew on redesigned, more advanced platforms, the first of which, Landsat-4, is pictured below:

Although similar in operational modes to the MSS (which was also part of the Landsat 4 and 5 payloads, to maintain continuity), the TM consists of 7 bands that have these characteristics:

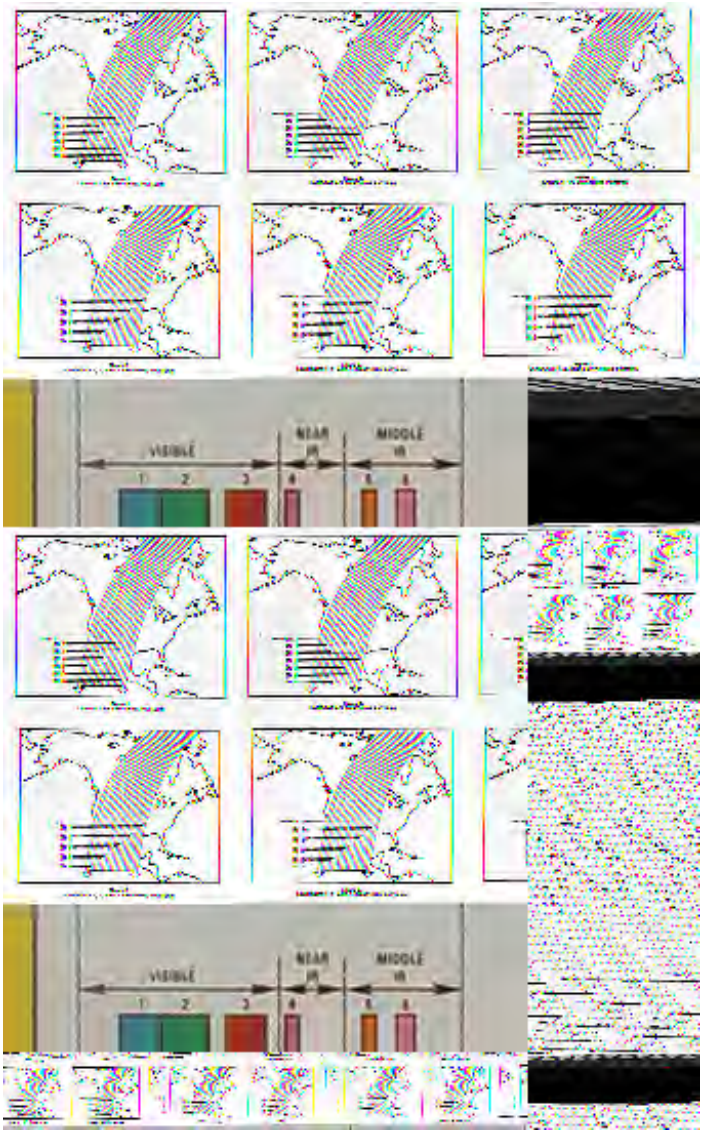
Band No.	Wavelength Interval (µm)	Spectral Response	Resolution (m)
1	0.45 - 0.52	Blue-Green	30
2	0.52 - 0.60	Green	30
3	0.63 - 0.69	Red	30
4	0.76 - 0.90	Near IR	30
5	1.55 - 1.75	Mid-IR	30
6	10.40 - 12.50	Thermal IR	120
7	2.08 - 2.35	Mid-IR	30

Six reflectance bands obtain their effective resolution at a nominal orbital altitude of 705 km (438 miles) through an IFOV of 0.043 mrad. The seventh band (but designated as Band 6) is the thermal channel, which has an IFOV of 0.172 mrad, which reduces resolution.

This diagram shows the placement of each band on a wavelength base:



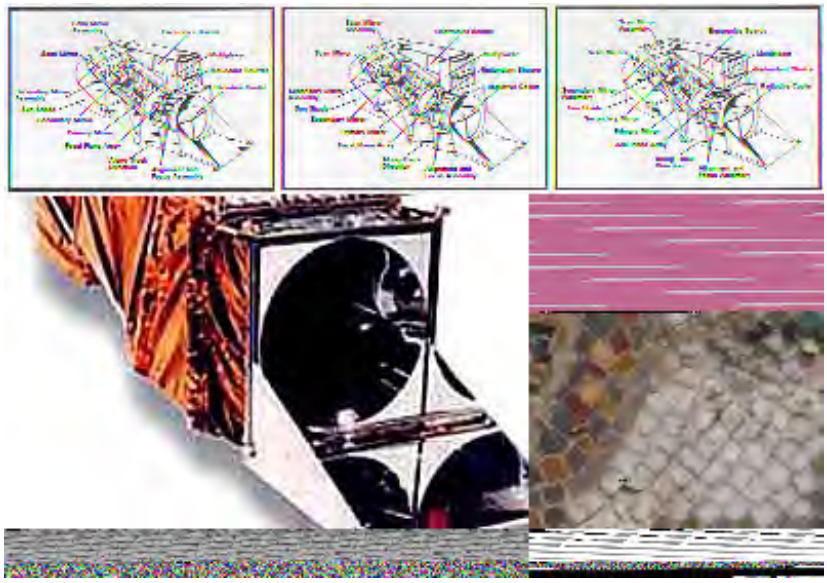
The TM spacecraft differs from the earlier MSS Landsats in the orbital repeat cycles, as shown in this diagram:



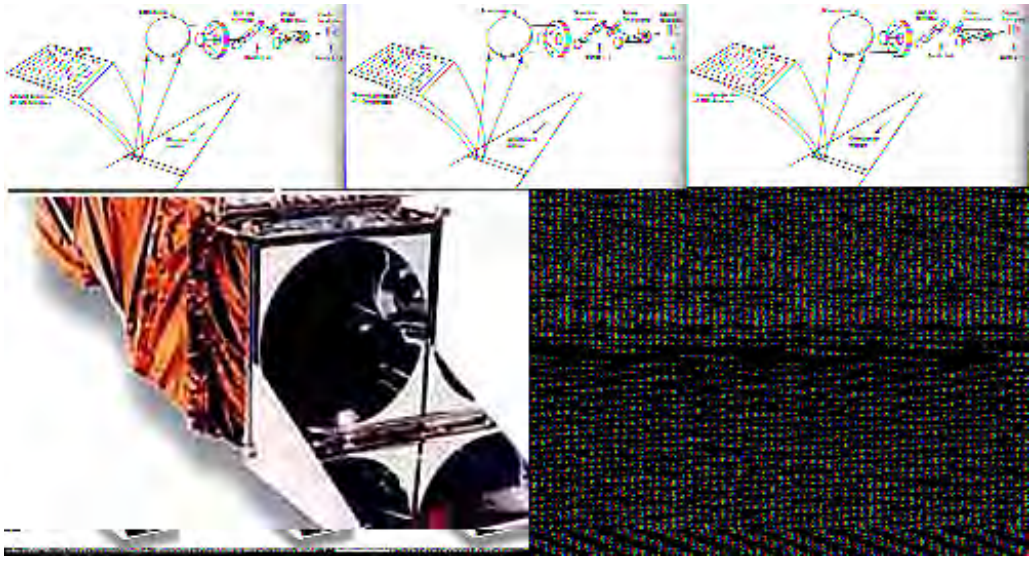
Note that for the first three Landsats there is an 18 day swath cycle in which the shift to the west (left) is systematic in that Path 1 is reoccupied after the 251 orbits have been occupied over an 18 day period. But keep in mind that for each day there are 14 circumglobal orbits (each orbit taking 103 minutes) such that as the Earth turns below each next path will have shifted less (160 km) than one frame (full size = 185 km) width to the west, allowing for some (variable) overlap between successive frames. During the 14 cycle single day history the total western shift produces a continuous combined swath width of 14 x 160 = 2240 km (1400 miles). In the full 18 day cycle this leads to a total composite coverage of (40900 km) 25200 miles. Landsats 4 and 5, at a lower altitude (which helps in improving resolution), accomplish 233 orbits in 16 days. The repeat pattern is notably different, as shown by the variable offset lines in the upper map.

Here is a photograph of the Thematic Mapper on the ground before it was mated to the spacecraft. Note the gold leaf that is used to shield the inner workings.

This cutaway diagram shows the major components of the TM system. Not shown are the interference filters used to separate the radiation into spectral bands:

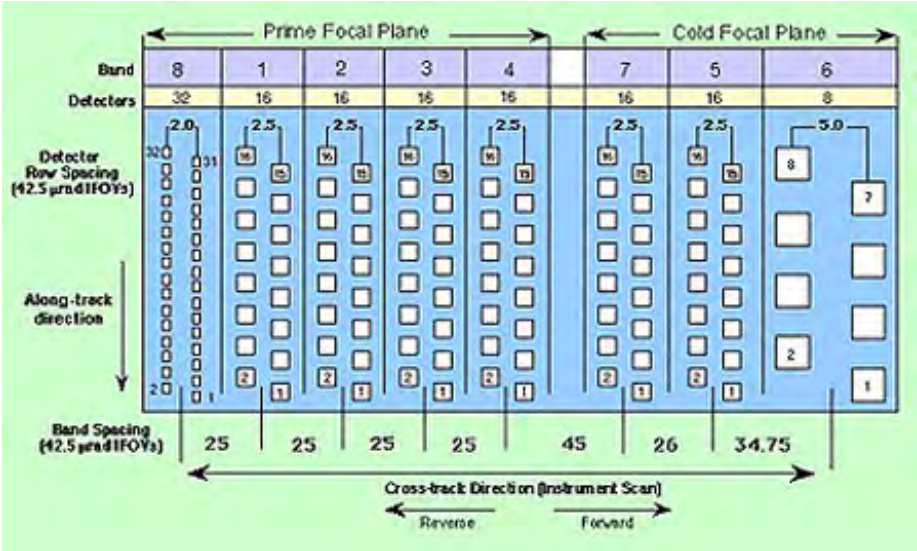


The sketch below shows some of the components in the optical train and detector layout of the TM.

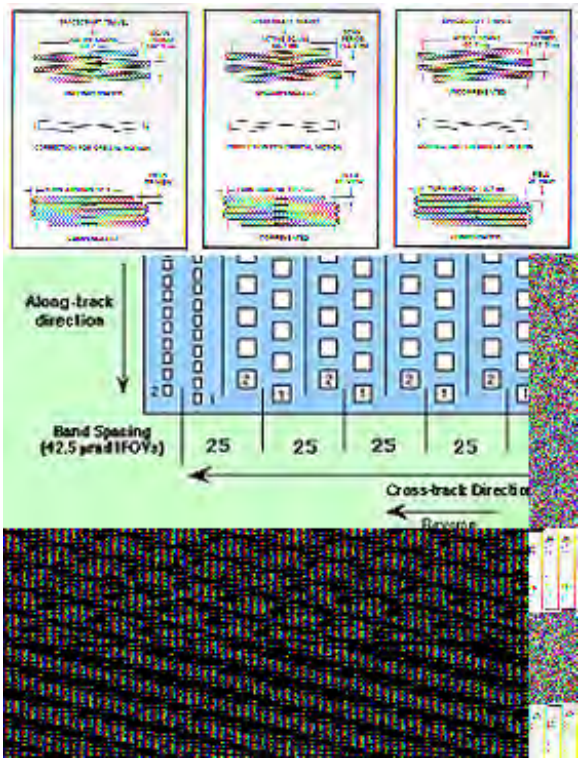


Instead of the 24 detectors on the MSS, the TM has a total of 96 for the reflective bands (16 each for a band; the mirror scan produces 16 lines at once) and 4 for the thermal band. For

Bands 1-4, silicon metal is the photoelectric detector; for 5 and 7 an indium antimony (InSb) alloy is used; for Band 6, the detector is a mercury cadmium telluride alloy (HgCdTe). At any given IFOV, each detector views a slightly different part of the ground that will be represented in the pixel being activated by radiation at any instant. The radiation collected for each band passes through a scan line corrector which compensates for the forward motion of the spacecraft. Radiation comprising Bands 1 through 4 is sensed by silicon detectors located in one focal plane; Bands 5 and 7, in the Mid IR (SWIR) use InSb detectors and Band 6 used a HgCdTe detector. Radiation from all three band sets are focused on a second focal plane at the detectors; the spectral filters are just before the detector plane. Bands 5, 6, and 7 are subjected to a radiative cooling system to improve sensitivity. This figure summarizes the detector array layout for the Landsat ETM+:

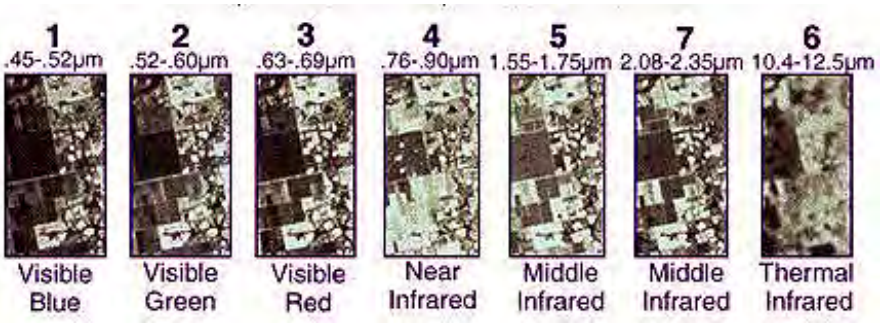


The TM's primary scan mirror takes imagery during both its left and its right swings (a full cycle; swing rate: there are 7 cycles/second). This results in a zig-zag pattern, in part because of the small but steady rotation of the Earth's surface below, as shown in the next figure. When the scan data are processed to produce an image, the lines are made to be parallel by using data acquired by two secondary mirrors (in parallel to one another; each rotating completely):

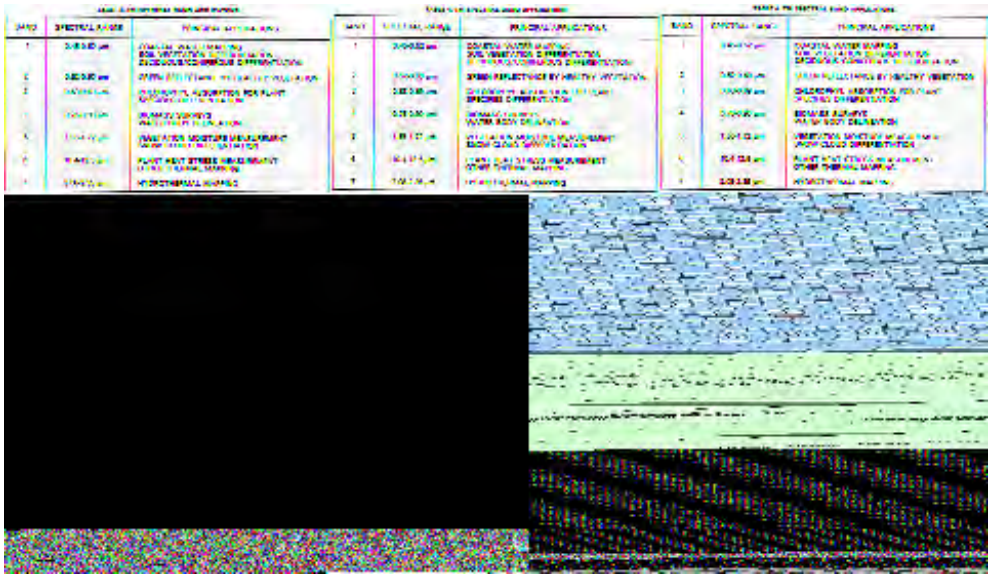


To introduce you to the differences in a TM image shown in all seven bands, we take a quick

look at a scene in Florida:



This table indicates the principal identification tasks that each band is especially adept at doing:



In more detail: Band 1 is superior to the MSS band 4 in detecting some features in water. It also allows us to form quasi-natural color composites. Band 5 is sensitive to variations in water content, both in leafy vegetation and as soil moisture. It also distinguishes between clouds (appearing dark) and bright snow (light). This band also responds to variations in ferric iron (Fe_2O_3) content in rocks and soils, which show higher reflectances as the iron content increases. Band 7 likewise reacts to moisture contents and is especially suited to detecting hydrous minerals (such as clays or certain alteration products) in geologic settings. Band 6 can distinguish a radiant temperature difference of about 0.6° C and is helpful in discriminating rock types whose thermal properties show differences in temperatures near their surface. It often can pick out changes in ground temperatures due to moisture variation and can single out vegetation due to its evaporative cooling effect. The higher resolution achieved in the reflective bands is a significant aid in picking out features and classes whose minimum dimension is usually on the order of 30 m (98 ft) . Thus, it can often discern houses and smaller buildings, which were unresolvable in MSS images.

The size and shape of full TM images from Landsats 4 and 5 are identical to the MSS images. At first glance, the quality and characteristics of these full scene TM images seem similar to those made by the MSS after optimal computer-based processing, but on closer inspection they do appear sharper. This apparent similarity is due to the need to resample the TM images for TV monitor displays (which are not high resolution systems capable of reproducing all TM pixels) by dropping some pixels. The influence of the better TM resolution (when un-resampled) becomes apparent whenever photographs of full scenes are enlarged (pictures more than a meter on a side can be produced with exceptional clarity) or subscenes are extracted and enlarged. Also, higher resolution improves scene feature classification since many objects on the Earth's surface are smaller than 30 meters; thus computer-based classification should result in higher accuracies for individual class identification and location.

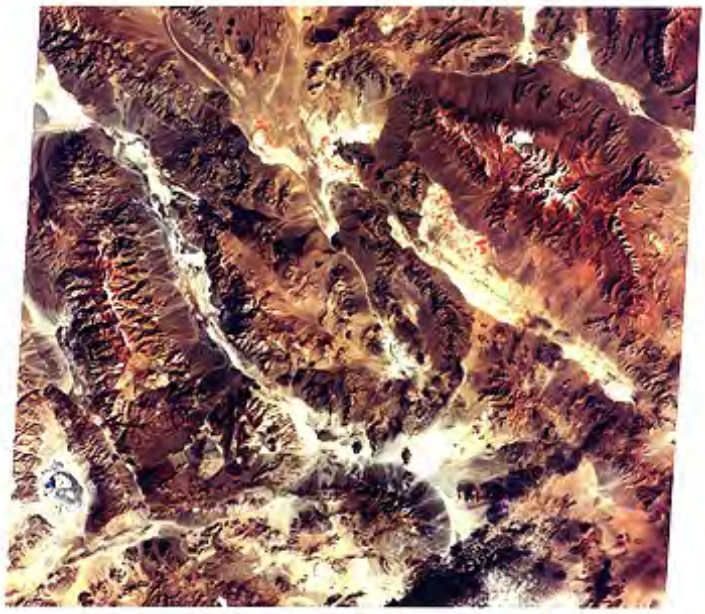
Primary Author: Nicholas M. Short, Sr.

This page is devoted primarily to displays of some Landsat TM images in natural and false color (and one comparison with an earlier MSS image); each is described. Presentation of black and white individual TM band images is deferred until the first part of Section 1 in this Tutorial.

History of Remote Sensing: Examples of TM Imagery

Below are three examples of Landsat-4 TM imagery in color. We refrain here from displaying any of the individual black and white TM bands because Section 1 presents and examines excellent examples of these from a subscene of Morro Bay, California.

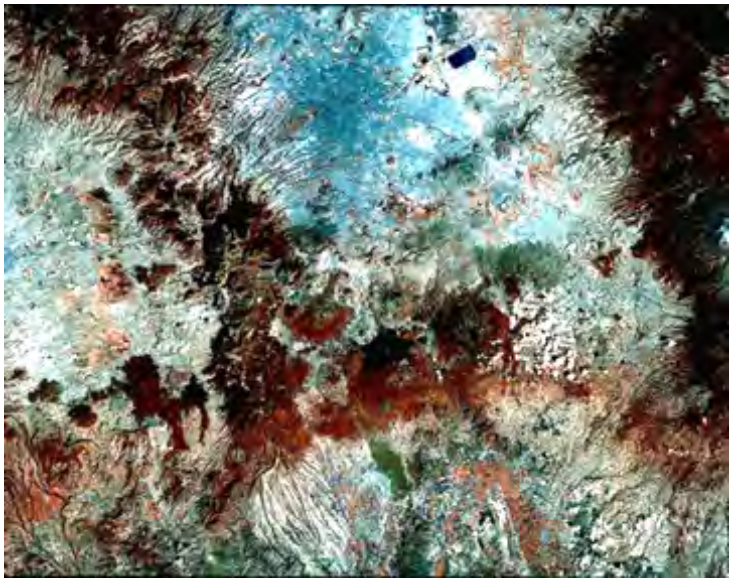
The first TM image is a late Fall, false color (TM bands 2, 3, 5 in blue, green, and red) rendition of mountain ranges in southeastern California and western Nevada. The large valley towards the left is Death Valley, with the Panamint Range to its left. The large range near the upper right is the Spring Mountains, whose reddish tones indicate vegetation (mixed evergreens and deciduous trees). The bottom of the image includes the north edge of the Mojave Desert.



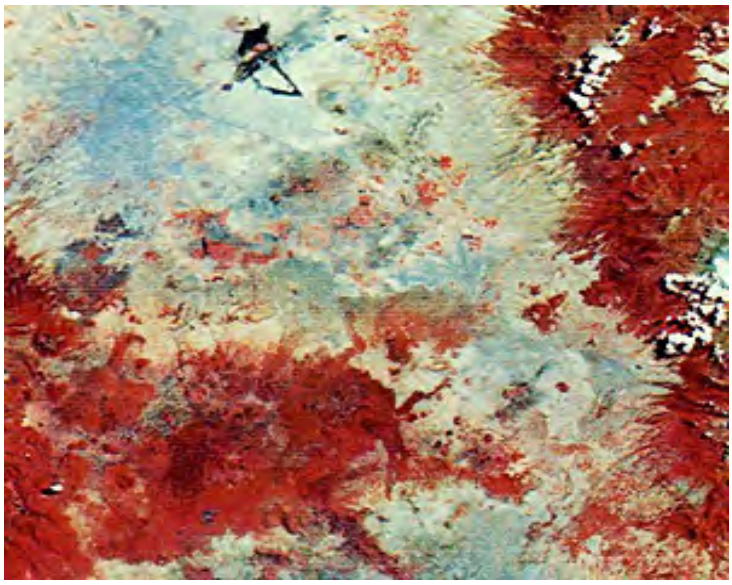
I-24: **Until you have gained familiarization by close inspection of TM full images, you may by inclined to think that they are not much better than corresponding MSS images. Ponder that; why? For what image situation(s) will the improved resolution really make a difference.** **ANSWER**

The second scene is an 80 km (50 mile) enlargement of part of a TM image covering the Sonoran Desert of northwest Mexico (a bit of the Gulf of California appears in the lower left), shown here in true color. Star and crescentic sand dunes dominate this subscene in this vast sand sea deposited over igneous lavas. The dark patches in the upper right are volcanic lavas but the dark mass to its southwest is the Sierra del Rosario, composed of granitic rocks.

The third TM image is also a subscene, about 70 km (43 miles) on a side, in west-central Mexico. Mexico City, with the largest urban population in the western hemisphere (about 30 million), appears in this false color version as the medium blue area in the upper left part of the image. Note that its area is much less than that of Los Angeles (one of the opening scenes in Section 4), indicating a high population density, i.e., crowding. The city lies at an average elevation of 2800 meters (9184 feet) astride the Neo-Volcanic Plateau, a zone that runs across Mexico and is seismically active. Just off the image to the right is a cluster of active volcanoes including the famed Popocatepetl which is over 5100 meters (almost 17000 feet).



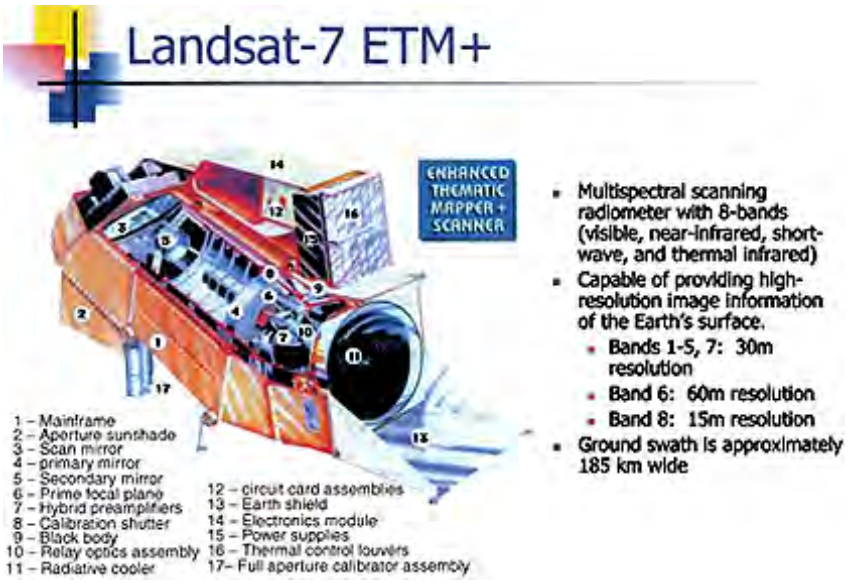
For comparison, we reproduce below a subscene imaged in 1973 by the first Landsat MSS. The area shown occupies about 2/3rds of the TM scene, pinned to the upper right corner. Draw your own conclusion about the relative details seen in the TM versus the MSS. Note that the size of Mexico City was not much smaller then even though its population was just over 7.5 million. The two images, when compared, illustrate the concept of change detection.



A Landsat-4 TM subscene shows the Cape Canaveral area of the east-central coast, where NASA's Kennedy Space Center (KSC) is located. Note the many individual launch sites. Compare this image with the RBV image of the same location shown near the bottom of page I-2-15.

A Landsat TM subscene in quasi-natural color that pinpoints many of the buildings and launch facilities at NASA’s Kennedy Space Center. The Banana (right) and Indian (left) inlets are visible as is part of Titusville on the mainland.

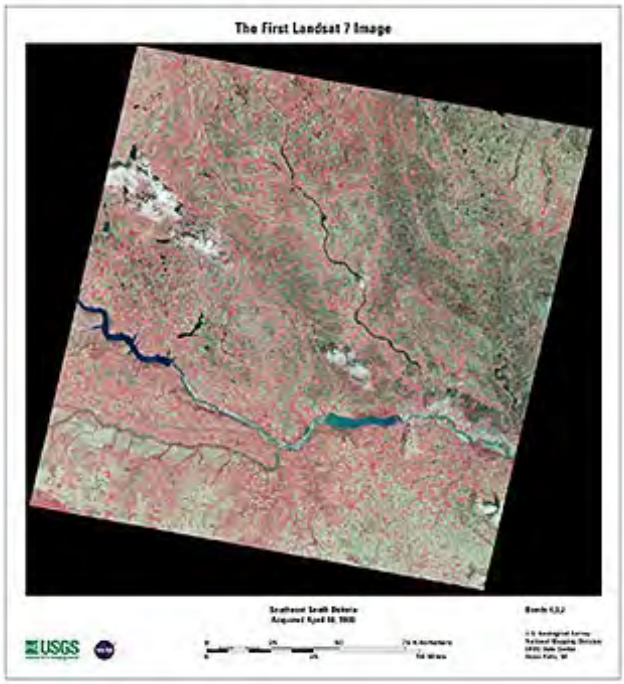
Landsat 7 has come on line in April of 1999 after its last working predecessor, Landsat 5 (remember, Landsat 6 failed to orbit), had continued to operate faithfully for 15 years (since 1984). Landsat 7 has only a single instrument (those who argued for including an MSS to maintain continuity lost the argument), called the Enhanced Thematic Mapper (ETM+). A cutaway diagram of this instrument appears below:



From *A History of Civil Land Imaging Satellites,* by Wm.Stoney
Encyclopedia of Space Science and Technology, J. Wiley & Sons.

The instrument consists of the same 6 bands in the Visible and Near Infrared as the TMs, again at 30 m resolution. The thermal band has an increase in spatial resolution by a factor of 2 - to 60 meters. The new component is a panchromatic (0.52 to 0.90 µm) black and white sensor (somewhat analogous to the RBV) which images at a 15 meter spatial resolution. The Landsat 7 program is operated jointed by NASA Goddard Space Flight Center and the U.S. Geological Survey. Here are some representative scenes.

The first scene acquired by Landsat 7 covers a part of southeastern South Dakota that includes the city of Sioux Falls. The U.S. Geological Survey's EROS Data Center (EDC), where Landsat imagery can be inspected and ordered, lies just off the image to the right.

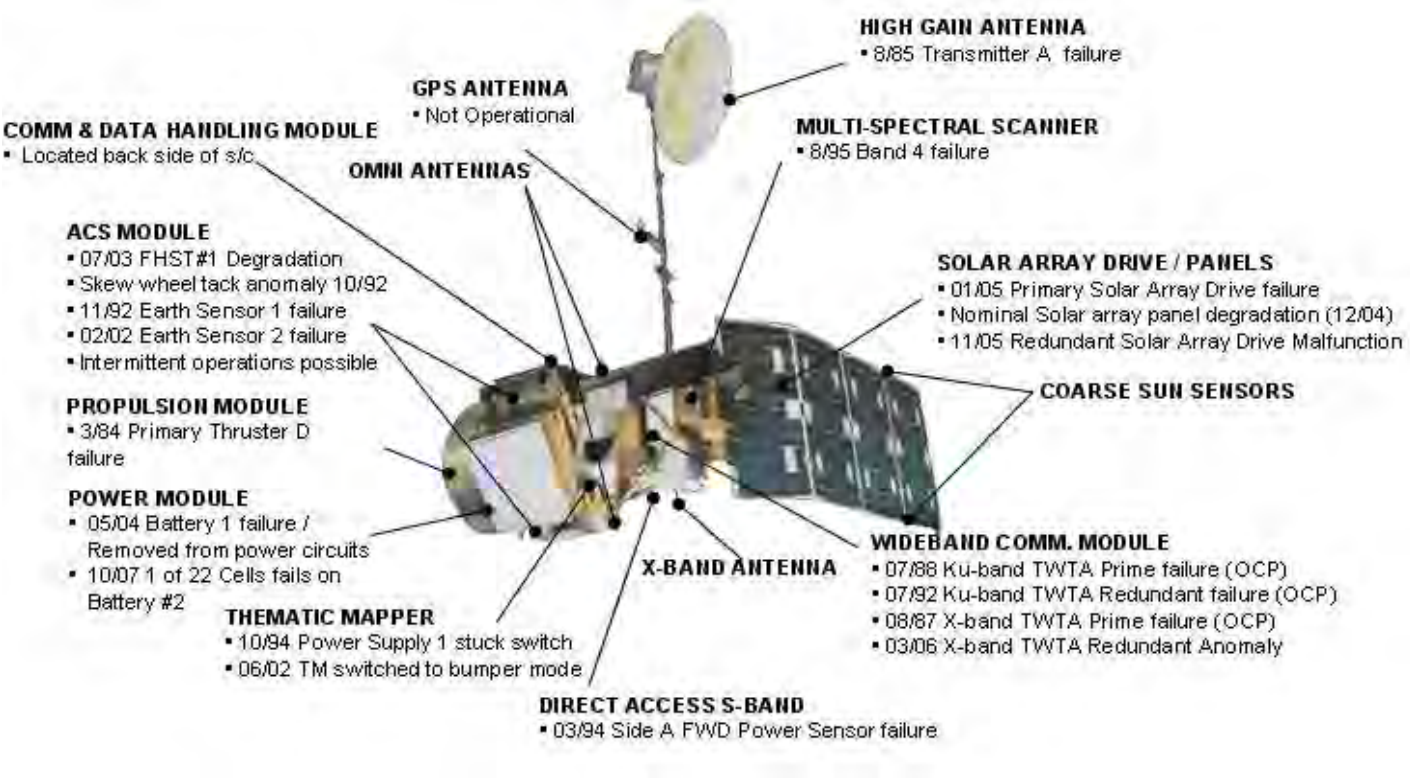


Part of the Landsat 7 panchromatic image of this same scene, showing a portion of Sioux Falls, with individual buildings now resolvable, is presented next.

Another urban area is seen in this quasi-natural color Landsat 7 subscene of the "Peninsula" area south of San Francisco. At the top is the San Mateo Bridge and Foster City (just beneath its west terminus), a residential area built on extensive fill into the San Francisco Bay, thus on newly created land. The Dumbarton Bridge near the bottom right leads into Palo Alto, home of Stanford University. Note the salt pans to the left of the bridge. The lake - actually the Upper Crystal Springs Reservoir - near the bottom left lies right on top of the infamous San Andreas Fault Zone. The nearby road (a 100 feet or so higher) is Interstate 280.

Goddard maintains a Landsat web site that now features a selected group of Landsat 7 images. Other information on Landsat 7 is included in the site established by the U..S. Geological Survey.

As of September 1, 2010, Landsats 5 and 7 are still operational (Landsat 6 failed after its launch). Both spacecraft have experienced threatening malfunctions, and some subsystems have failed leading to problems over the years, but clever engineers have found ways to maintain their imaging capability. This illustrations records the difficulties encountered by Landsat 5 as well as some of the solutions:



For those interested in the history of the Landsat program, click on this 11 page summary : http://geo.arc.nasa.gov/sge/landsat/lpchron.html. This Chronology was prepared for NASA at its Ames Research Center. They also maintain a general information bulletin board on Landsat (which includes the above historical perspective) that you can access at http://geo.arc.nasa.gov/sge/landsat/landsat.html. And for those curious about Landsat's future, read the paragraph that considers this on page 20-1, located just before the "Some Future Thoughts" subsection.

So, what's the future for the Landsat program? Users worldwide have gotten used to relying on Landsat for a wide range of applications. Despite the many newer satellites that perform similar functions, there is a strong clamor for maintaining continuity of data gathering and image acquisition. NASA and the U.S. Geological Survey have been charged by Congress with designing a workable program to continue Landsat after the current satellites lose functionality. This is referred to as the "Landsat Data Continuity Mission 2011". Occasional general information about the LDCM can be found at this USGS Landsat site.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

The first "competitor" to Landsat in providing high resolution multispectral images of the Earth is the SPOT series of satellites operated by a French-owned company. SPOT is briefly described here (in detail in Section 3). Its sensors use a different type of detector, the CCD, whose function is discussed in depth on this page.Two examples of SPOT images are displayed.

Other Remote Sensing Systems - MOMS and SPOT

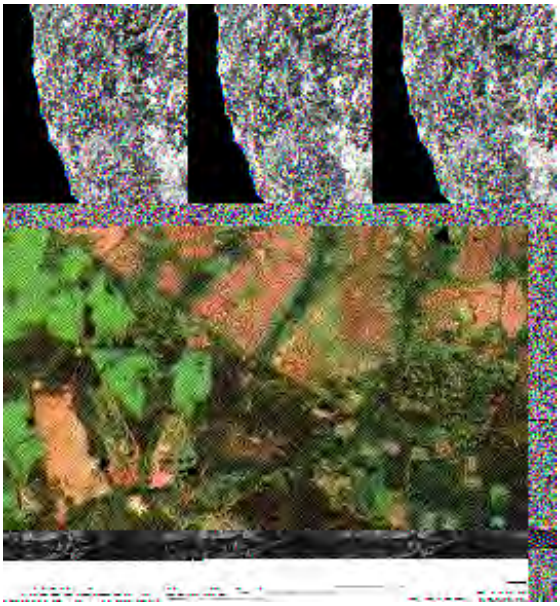
Scanners, such as those on the Landsats (MSS and TM) were the prime Earth-observing sensors during the 1970s into the 1980s. But these instruments contained moving parts, such as oscillating mirrors that were subject to wear and failure (although remarkably, the MSS on Landsat 5 continues to operate into 1999 after launch in March of 1984). Another approach to sensing EM radiation was developed in the interim, namely the Pushbroom Scanner, which uses charge-coupled devices (CCDs) as the detector.

CCD detectors are now in common use on air- and space-borne sensors (including the Hubble Space Telescope which captures astronomical scenes on a two-dimensional array, i.e., many parallel rows of detectors). The first airborne pushbroom scanner to be used operationally was the Multispectral Electro-optical Imaging Scanner (MEIS) built by Canada's CCRS. It images in 8 bands from 0.39 to 1.1 µm (using optical filters to produce the narrow band intervals) and uses a mirror to collect fore and aft views (along track) suitable as stereo imagery.

The German Aerospace Research Establishment (DFVLR) developed the first pushbroom scanner to be flown in space. The Modular Optico-electronic Multispectral Scanner (MOMS) was aboard Shuttle Mission STS-7 and STS-11 in 1983 and 1984. It uses two bands, at 0.575-0.625 and 0.825-0.975 µm, to produce 20 m resolution images. The MOMS image below is an area of farmland in Zimbabwe:

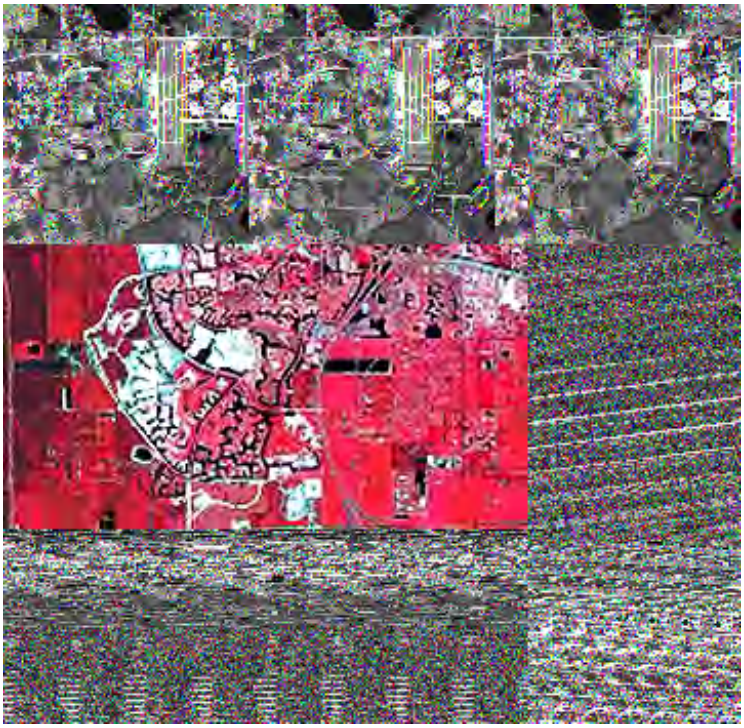
MOMS-2 was flown on STS-55 in May of 1993. It has four multispectral channels (13 m resolution) and a panchromatic band (4.3 m), and is in stereo mode. Here is a panchromatic image of a city on the western shore of the Persian Gulf (locus incertae; probably in Qatar):

The first use of CCD-based pushbroom scanners on an unmanned Earth-observing spacecraft was on the French SPOT-1 launched in 1986. (Page 3-2 describes the SPOT system, which is operated as a commercial program; 4 SPOTS have now been launched) An example of a SPOT image, from its high-resolution video (HRV) camera, covering a 60 km section (at 20 m. spatial resolution) of the coastal region in southwest Oregon, is the next image we show. Note that scan lines are absent, because each CCD element is, in effect, a tiny area analogous to a pixel.



False color composites are a standard product from SPOT. Here is part of a SPOT image taken in 1994 of developing farmland west of Miami, FL and near the edge of the Everglades, in which a new housing development has been put in since the mid-80s.

The third image is a panchromatic image (with 10 meters resolution) showing the edge of Orlando, Florida, including its airport.



Like Landsat, the SPOT satellites are widely used. The above only highlights the basic program. For more details, hit on the SPOT Image Internet home page.

Primary Author: Nicholas M. Short, Sr.

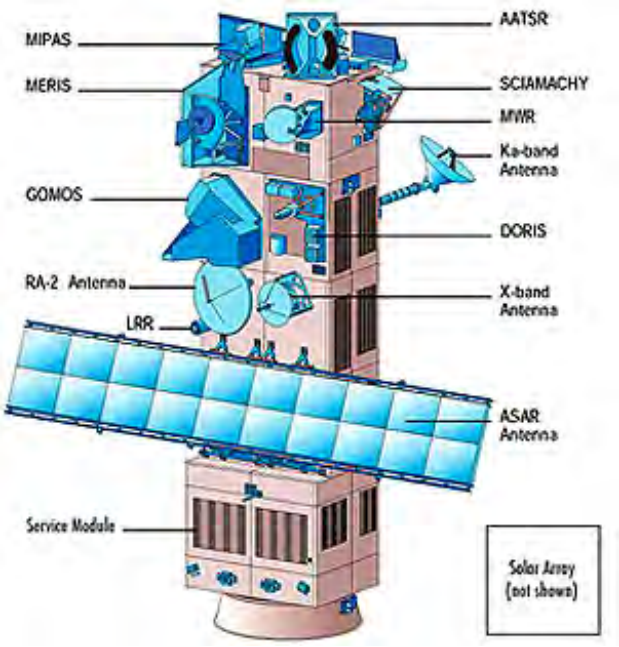
Other countries are now active in the satellite remote sensing "game". Canada was one of the first, with its Radarsat. The European Space Agency (ESA) has operated ERS-1 and ERS-2, and then Envisat, with both spectral imaging and radar. The Russians with their RESURS series now offer imagery similar to Landsat and some meteorological satellites. India has launched four satellites, the IRS series, each with multispectral sensors. Japan is operating two satellites, the JERS series, with sensors that cover much the same spectral regions as the Landsat TM, but at higher resolution and in 2006 the ALOS satellite. The Ukraine has launched its own OKEAN satellites. China and Brazil have jointly developed the CBERS series. A series of Microsats, destined to be used as disaster monitors, are being placed in orbit, the first now operational is AlSat-1.

Other Remote Sensing Systems - ERS, Envisat, RESURS, OKEAN, IRS and ResourceSat, JERS, ALOS, CBERS, MicroSat Series (AlSat-1)

Following Landsat and SPOT, other earth-observing satellite programs proliferated. Several are sponsoreded by international governments while others are now operated by private industry. Some spacecraft supported exclusively, or primarily, radar sensors. (These are considered again on page I-25 (two pages hence) and in more detail on page 8-7.)

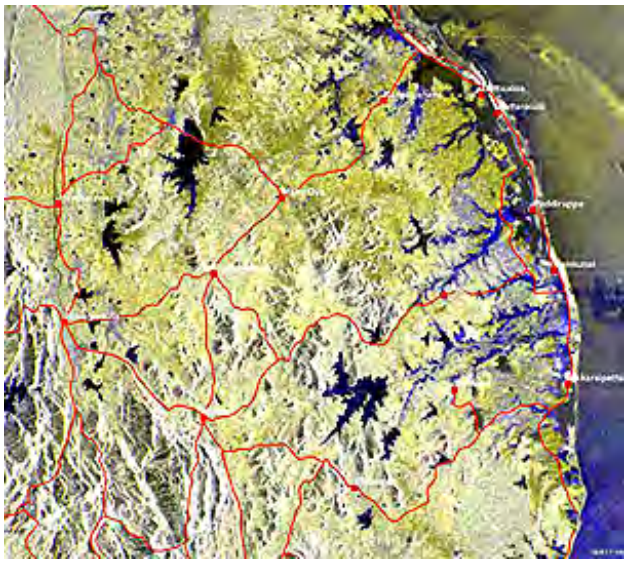
The European Space Agency (ESA) sent two missions, ERS-1 in March, 1991 and ERS-2 in April, 1995 with radar systems. Here is the ERS-1 spacecraft and a scene (southern coast of Finland) it acquired:

Envisat is an ESA satellite that has both wide and medium angle viewing sensors as well as radar. The satellite looks like this and has these instruments: Envisat home page



Here is its view of Europe as a whole, using the wide angle capability of MERIS:

The ASAR radar on Envisat imaged the east coast of Sri Lanka shortly after the massive tsunami struck in December, 2004 (see second page of Overview):



Starting in the mid 1980s, the Soviet Union (and now Russia) entered the world arena with an Earth-observing satellite program available on the open market. Prior to that, the USSR had the largest 'competitive' space program next to the U.S.'s. Most of its unmanned satellite programs had primarily military surveillance and support objectives. The first surveillance system was Zenit-2, launched in 1961. Some data were released but the bulk were kept secret, with the data used internally within the Soviet Union. The earth-observing programs evolved from some of these earlier satellite missions. A simplified chain of missions leading to these programs begins with the Meteor Series (meteorological) --> Okean Series (oceanographic) --> Salyut/Almaz series --> ResursF1 and F2 series --> Resurs-O --> Resurs-F (these are not progressive but overlap). We will only consider a few of these in the next paragraphs. To get more insight, log on to these websites:Soviet/Russian space systems and Soviet and Russian earth observation systems. We will not show any results from the Meteor, Okean and Almaz series here. Consult these sites for descriptions and a few images.

The Resurs (also presented as RESURS) program has a long history. We will simplify it, mentioning only a few aspects. The Resurs F series has operated for more than 40 years - most of this time for military purposes. These spacecraft were about 7 meters long, as shown here

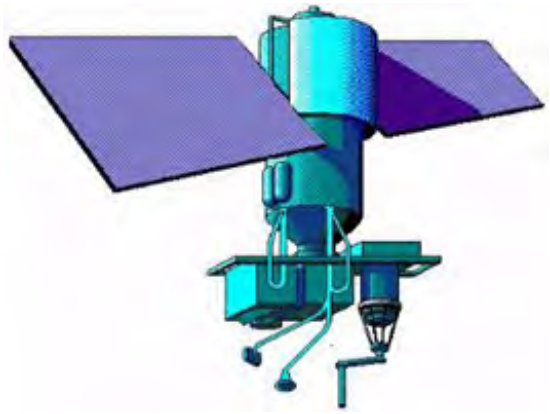
This group containing high resolution photo-cameras and after a few weeks released the camera package in a capsule for return to Earth by parachute (this retrieval method is discussed on page I-26e. Cameras used include those designated KATE-200, KVA-1000, and KVR-1200. Here is a view into one such capsule, showing the cameras in place:

This is a Resurs-F photo of the Vatican and adjacent Rome (you will see high resolution images of the Vatican again in Section 6):

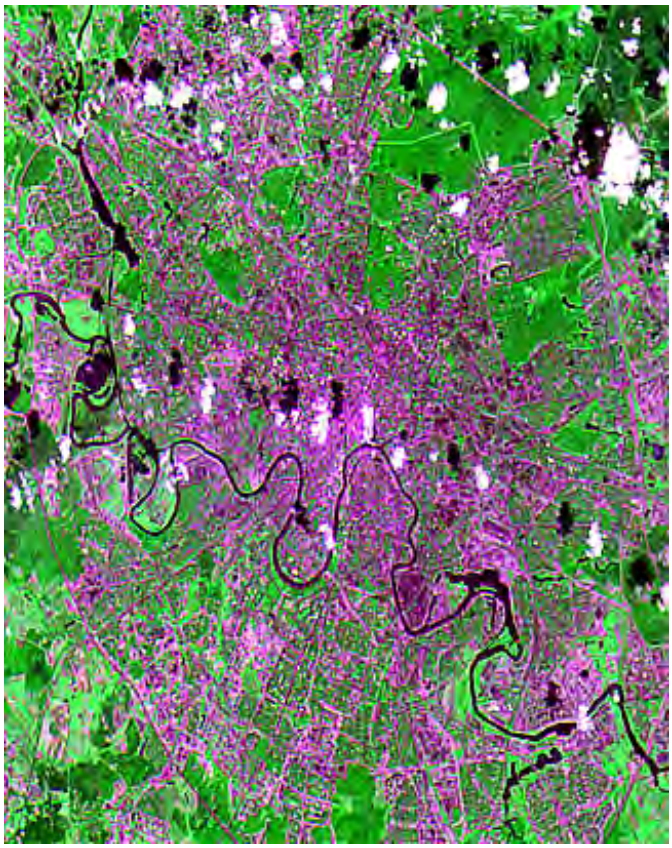
The camera retrieval program, which utilizes low cost equipment, continues with a 2006 launch of the first Resurs-DK spacecraft, shown below:

The Russian film retrieval technique is at the heart of its SPIN-2 series. Resolution of 2 meters is achieved in the film products. This quality enticed the TerraServer Corp. into partnership to distribute the imagery to a worldwide market. This scene of downtown Manhattan (pre-mid2001; can you tell how this is known?) shows the sharpness of this photography (compare this scene to the one shown on page 2 of the Overview and another on page 4-3 of Section 4).

The RESURS-01 series (4 so far) provided a multispectral system (3 Vis-NIR bands; 2 thermal) whose resolution (160 m, and 600 m for thermal) is intermediate between that of Landsat/SPOT and the AVHRR on meteorological satellites. Like Landsat RESURS are placed in near-polar, sun-synchronous orbits. A drawing of the Resurs01-3 spacecraft appears here:

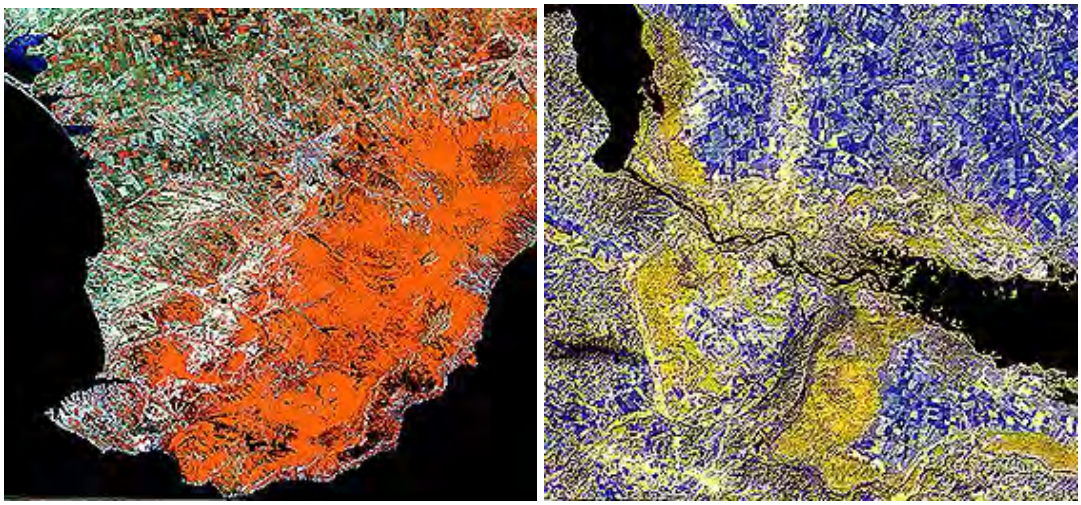


Two images from this RESURS system appear below: the first is a false color composite Moscow and its surroundings.

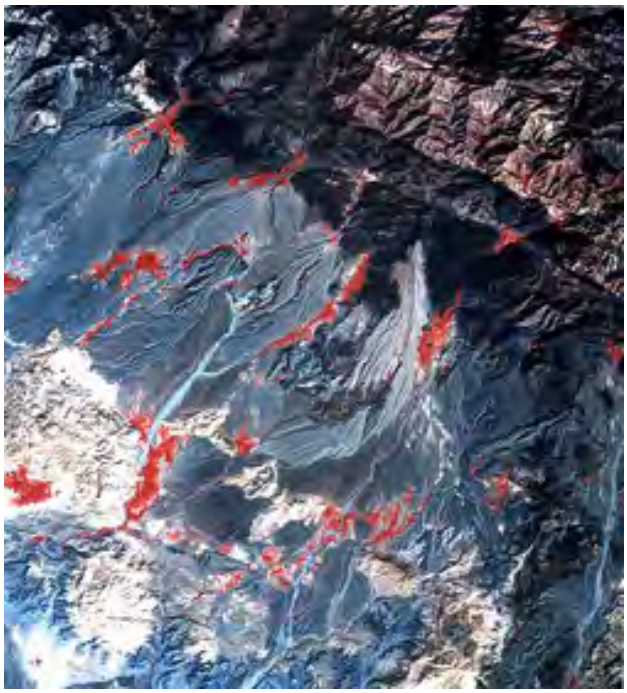


The second RESURS (Resources) image is part of a mosaic of Europe which here includes all of Norway, Sweden, and Denmark, and part of Finland and several Baltic nations.

The National Space Agency of the Ukraine begn its own program of space observations in 1992; it works in cooperation with the Russian Federation in using certain facilities. Its OKEAN series includes multispectral scanners, thermal sensors, and radar. Two MSU-V images (50 m resolution) show a standard false color composite (left) of the southern Crimea (Sebastapol in lower left) and a different color combination (right) of the Dnieper River in the Ukraine Lowlands, with Kiev just below the upper "lake" (caused by river damming).



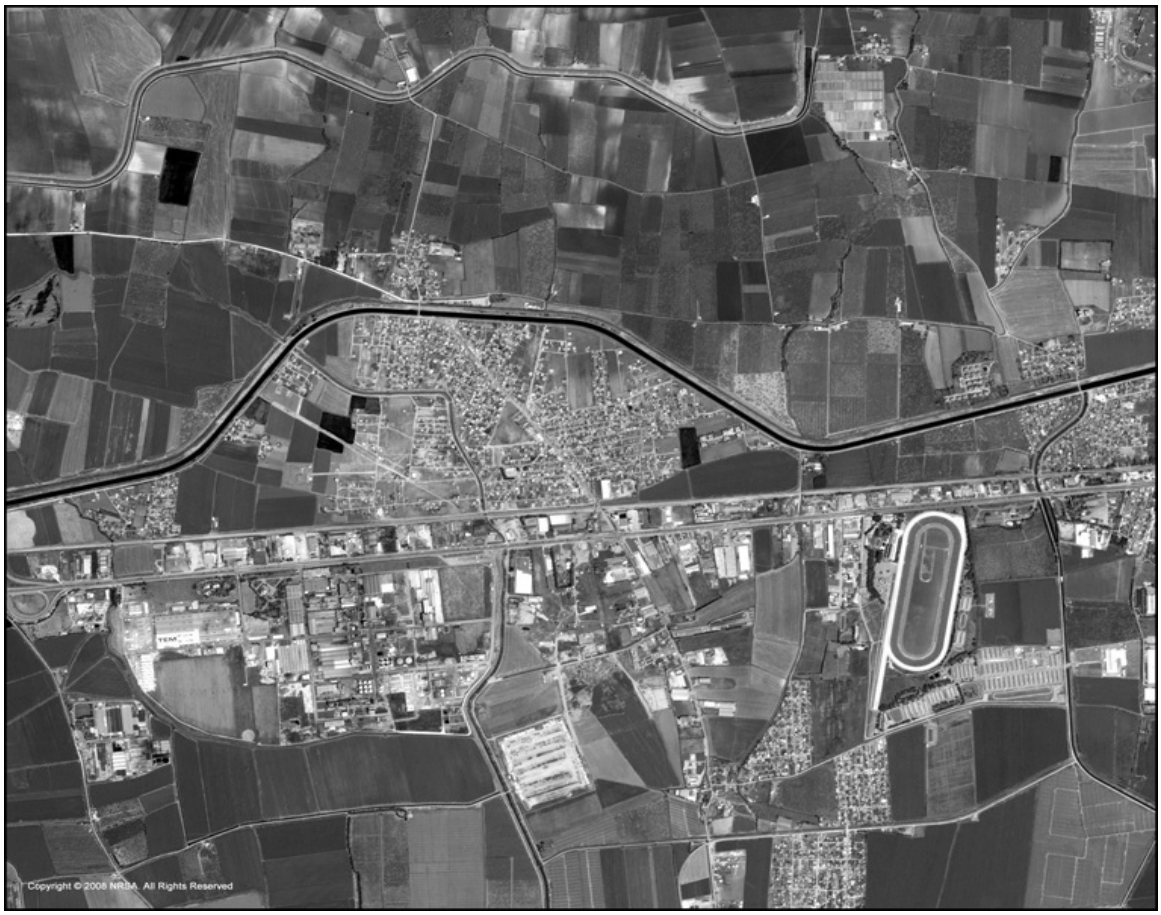
India successfully operates several Earth-resources satellites that gather data in the Visible and Near IR bands, beginning with IRS-1A in March of 1988. The latest in the series, IRS-1D, launched on September 29, 1997. Its LISS sensor captures radiation in the blue-green, green, red, and near IR bands at 23 m spatial resolution. The spacecraft also produces 5.8 m panchromatic images, as well as 188 m resolution wide-field (large area) WiFS multispectral imagery. Below are three recent images from this system, the one on the top (WiFS) showing the Grand Canyon of Arizona, in the middle a three-band color composite made by the 23 m LISS, showing mountainous terrain and pediments with alluvium fans in southern Iran, and at the bottom a 5.8 meter panchromatic view of part of the harbor at Tamil Nadu in India.



On December 4, 2003 the IRS program orbited the first in a new series, ResourceSat-1, whose multiple sensors produce multispectral images at 56, 23 and 5 meters. Details on its orbital parameters and instrument capabilities are found at this Internet Site Here are several examples of the imagery its produces; location and resolution is given in the captions (remember, click on lower right of the image).

MANASAROVAR
LAKE
RAKAS LAKE

India launched a new series of satellites with Cartosat-1 (placed in orbit on May 5, 2005). It has two panchromatic cameras, each producing a 30-km wide image having 2.5 meter resolution. They are arranged so as to produce a stereo pair. Here is a Cartosat-1 image of Adana, Turkey:

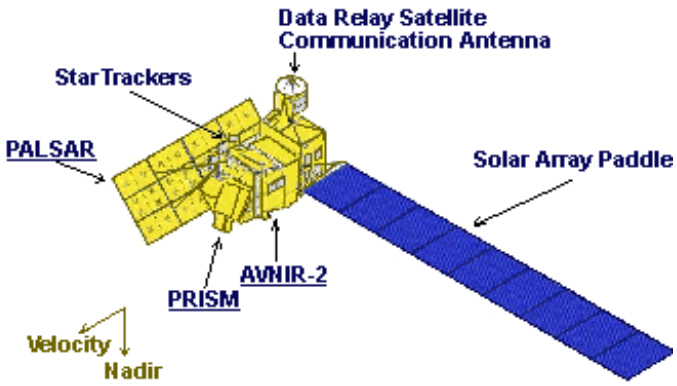


More information on the Indian remote sensing program is available on this Wikipedia website.

The Japanese, beginning in 1990, have flown JERS-1 and JERS-2 which include optical and radar sensors. Here is an artist's conception of JERS-1 in space:

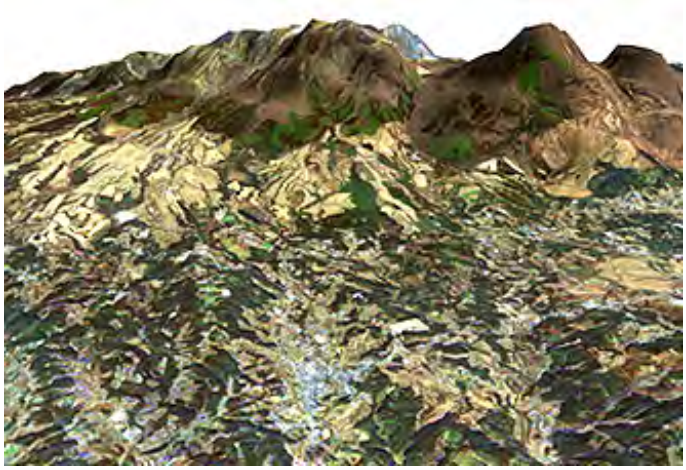
The optical system is a seven band scanner similar in coverage to the TM. The satellites are operated by the National Space Agency, JAXA Here is a false color JERS-1 image of Tokyo and Tokyo Bay:

NASDA's ALOS (Advanced Land Observation Satellite) was successfully launched on January 23, 2006. The ALOS has three remote-sensing instruments: the Panchromatic Remote-sensing Instrument for Stereo Mapping (PRISM) for digital elevation mapping, the Advanced Visible and Near Infrared Radiometer type 2 (AVNIR-2) for precise land coverage observation, and the Phased Array type L-band Synthetic Aperture Radar (PALSAR) for day-and-night and all-weather land observation. Here is the spacecraft:

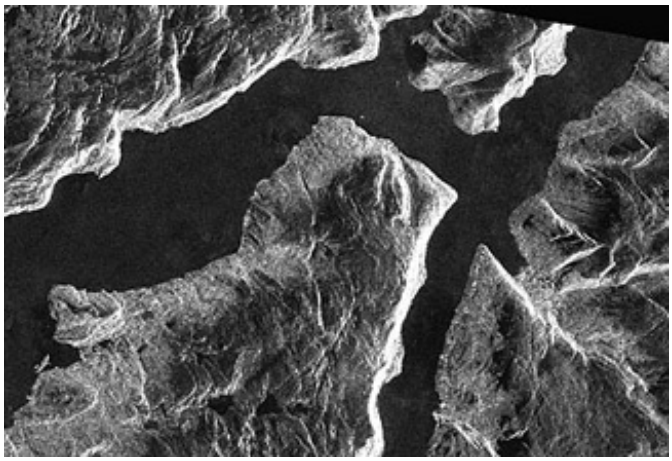


This scene of Naples, Italy and the Bay of Naples is typical of AVNIR inagery:

Using digitized land elevation data convolved with a corresponding AVNIR image, the ALOS can produce a perspective scene, such as this view of the Kujyu mountain range in Japan:



It is increasingly common now for earth-observing satellites to have a radar system working in conjunction with a visible imager. Here is a view of a fjord near Bergen, Norway imaged by ALOS PALSAR radar:

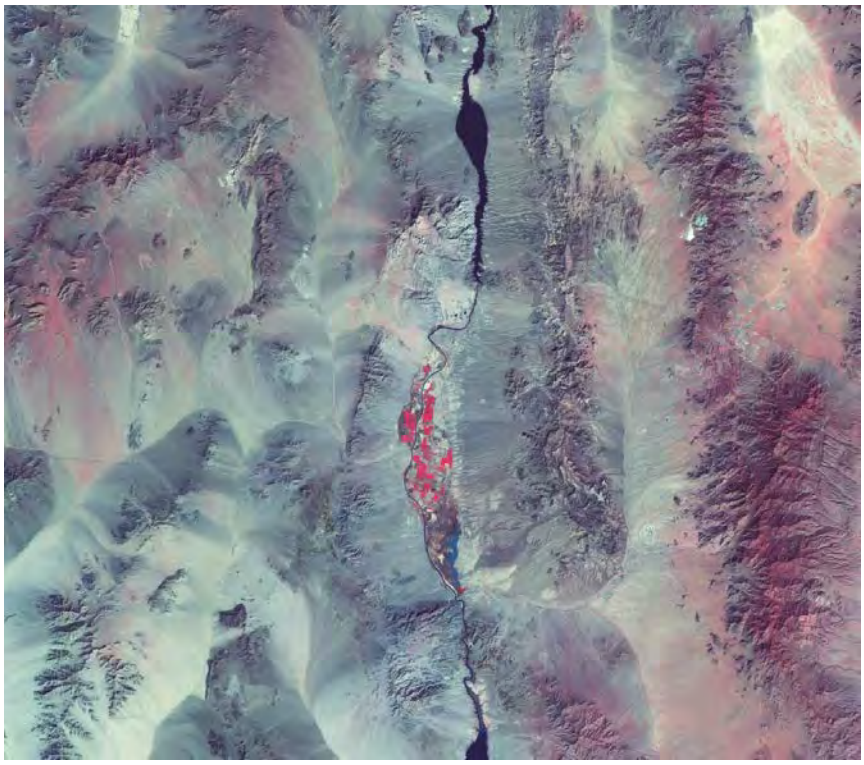


The Peoples Republic of China has joined forces with the Brazilian government to develop a series of earth-observing satellites launched by Long March rockets from China. Their program goes by the name of CBERS (China-Brazil Earth Resources Satellites); in China these satellites are called the Zujuan series. CBERS-1 was orbited on October 14, 1999. It includes three sensors: 1) WFI (300 km swath; 260 m resolution; 4 bands); 2) IR-MSS (20 km swath; 80 m

resolution; 4 bands including thermal); and 3) CCD (20 m resolution; 4 bands). The more than 280000 images received are concentrated mainly over Brazil and China and are not generally available to other nations. CBERS-2 (YZ-2) was launched on September 1, 2000; although reputed to be available for earth resources applications, western observers have concluded that its 3-meter resolution sensor is being used primarily for military reconnaissance. Here is a CBERS-1 CCD image of the area in Brazil that includes the capital, Brazilia.:

The British, in cooperation with the various national space agencies, have developed a program that is inserting into (700 km altitude) orbits a series of small (440 kilograms [200 lbs]) but versatile satellites (because of their size, referred to as *Microsatellites* consisting of a single sensor that can image in the green, red, and near IR wavelength band intervals. This is the DMC (Disaster Monitoring Consortium) program which will (as now planned by the end of 2003) have four similar satellites in staggered orbits that have spacings and repeat cycles which allow any location on Earth to be visited once during a 24 hour period. The sensor can cover an area 600 x 600 km (roughly, 400 miles on a side) in dimension at a resolution of 34 km. Higher resolution - smaller area - coverage is also an option. While capable of many scene monitoring and classification tasks, the prime utilization which is the "selling point" for this system of "formation flying" satellites (see page 16-11) is to obtain images that (if clouds are not a hindrance) can see damage or other consequences of a group of natural or manmade disasters detectable at the resolutions available. The first satellite (AlSat-1) was built at Surrey, England for operation by Algeria. This is AlSat-1 seen in its fabrication room.

An image encompassing part of the U.S. Southwest from ALSat-1 appeared at the bottom of page Overview-2. Below are two more ALSat-1 images, the first including the land adjacent to Colorado River between Arizona and California and the second a smaller area view of Las Vegas, Nevada.



Others in this series are BilSat (Turkey), UK-DMC (England), NigeriaSat (Nigeria), and ThakPaht (Thailand).

PREVIOUS rst NEXT

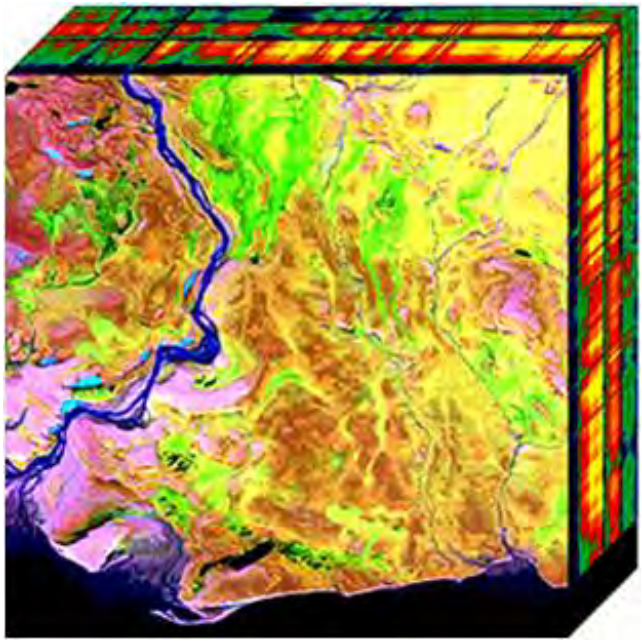
Primary Author: Nicholas M. Short, Sr.

Imaging spectroscopy, also known as hyperspectral remote sensing, allows a sensor on a moving platform to gather reflected radiation from a ground target such that a special detector system consisting of CCD devices can record up to 200+ spectral channels simultaneously over the range from 0.38 to 2.5 µm. With sampling thus at a 0.01 µm (10 nanometer) interval it is possible to plot the data as quasi-continuous narrow bands that approximate a spectral signature rather than histogram-like broader bands. With such detail, the ability to detect and identify individual materials or classes greatly improves. The AVIRIS instrument developed at JPL is described. Examples are shown that confirm that the hyperspectral approach is now the state-of-the-art cutting edge of remote sensing from air and space. Two hyperspectral satellites, Hyperion and Proba, now in orbit, demonstrate this capability.

Other Remote Sensing Systems - Hyperspectral Imaging

Another major advance, now coming into its own as a powerful and versatile means for continuous sampling of broad intervals of the spectrum, is *hyperspectral* imaging. This page expands on the treatment of imaging spectrometers, but the background to appreciate how they operate is reserved for the second half of Section 13, starting on page 13-5; consult those pages for more principles and details that could expand your understanding of the remainder of this page.

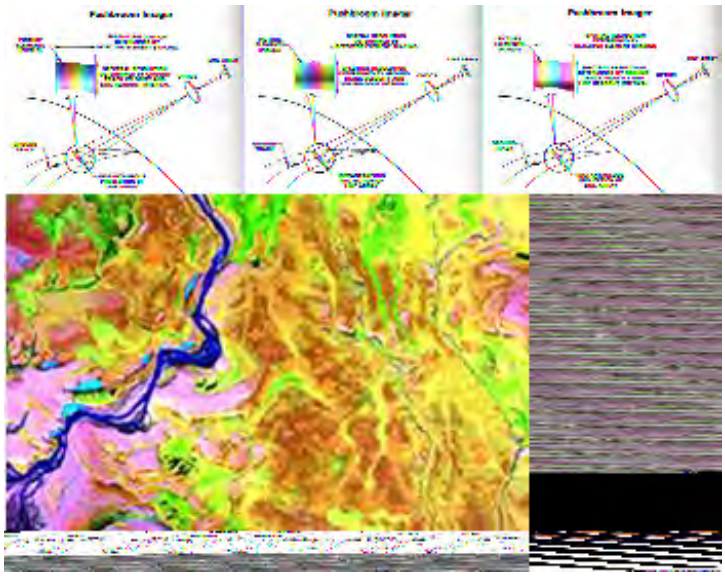
Multispectral remote sensors produce images of bands that have relatively broad spectral widths - typically about 100 nanometers. If that can be reduced to narrow widths, say 10 nanometers, then a plot of the intensity (for example, the percent reflectance) for each now narrow band yields a spectral curve that is similar to that produced by a laboratory spectrometer. If an instrument capable of 10 nanometer resolution can be flown on an aircraft or spacecraft, the resulting set of 10 nanometer images provides a far better and more discriminating package of imagery. If, as an example, the imaging spectrometer operates between 400 and 1100 nanometers, there would be 70 individual narrow bands used to construct either an image (using any three) or an approximating spectral curve for each target. The target, furthermore, can be small, such as a few meters and perhaps then homogeneous with respect to the class(es) it represents. The target is here synonomous with a pixel (the equivalent "piece of real estate" whose size is defined by the spatial resolution). The idea is further expressed in the next figure, the so-called hyperspectral cube, in which the full front is the scene itself, in X-Y coordinates, made up by the succession of pixels. The Z axis (the slant or 3-D part of the figure) is just the set of the many individual narrow band images.



Heretofore, because of the high speeds of air and space vehicle motion, insufficient time was available for a spectrometer to dwell on a small area of Earth's surface or an atmospheric target. Data are necessarily acquired for broad bands in which spectral radiation is integrated within the sampled areas to cover ranges, such as 0.1 µm (for instance Landsat bands). In hyperspectral data, that interval narrows to 10 nanometers (1 micrometer [µm] contains 1000 nanometers [1 nm = 10^{-9} m]). Thus, we can subdivide the interval between 0.38 and 2.55 µm into 217 intervals, each approximately 10 nanometers (0.01 µm) in width. These are, in effect, narrow bands.

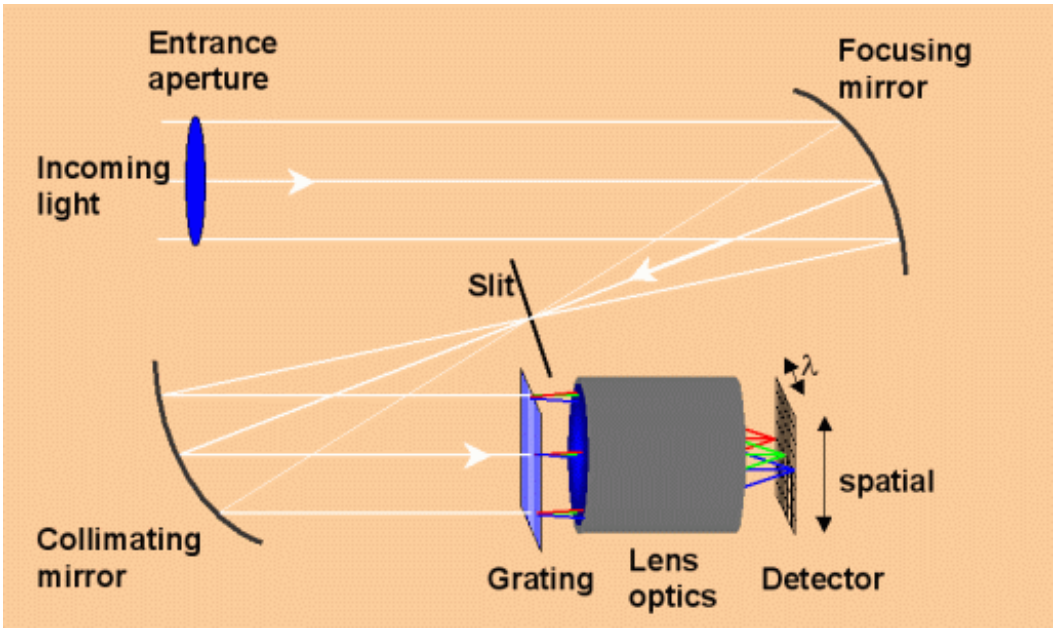
The problem that prevented development of hyperspectral sensors prior to 1985 (the date refers to work by Alex Goetz and others at JPL at that time) is one of being able to rapidly sample small intervals of the spectrum as these pertain to the spectral signatures (pure or mixed) associated with each resolution element or pixel. This necessarily fast sampling rate had to await development of appropriate electronics. The breakthrough that makes hyperspectral sensors possible was the technology involved in Charged Coupled Devices (CCDs)(see this Wikipedia website for a review of CCDs). The CCD detectors for VNIR intervals are silicon microchips, while those for the Short Wave InfraRed (SWIR, between 1.0 and 2.5 µm) intervals consist of an Indium-Antimony (In-Sb) alloy. If a radiance value is obtained for each such interval, and then plotted as intensity versus wavelength, the result is a sufficient number of points through which we can draw a meaningful spectral curve.

The use of CCDs permitted establishment of the so-called "pushbroom" technology, discussed in Section 13 and elsewhere in the RST (see page Intro-5a), which replaces the older, mirror-dependent "wiskbroom" scanners, such as used by Landsat's MSS (as the mirror oscillates it "sees" and samples one pixel at a time over its lateral extent [swath width]). (The terms draw upon the analogy: an actual wiskbroom, such as one used to brush one's clothes, swings back and forth - the mirror on a scanner rocks back and forth; an actual pushbroom is elongated and is pushed forward ahead of the person using it - on a scanner the line of CCDs is advanced forward in a similar manner.) The difference is visualized in this website.. The scanner used with your computer is normally of this type. This illustration is a simplified expression of the pushbroom or line all at once type scanner:



The entire line is instantaneously imaged all at once, and then very rapidly sampled (signal extracted from each pixel) sequentially. The sensor is advancing (in orbit or in the air) to the next linear target on the ground. As the previous line is sampled, it is then refreshed, so that this next line now is recorded by linear scan. This process is carried out in microseconds.

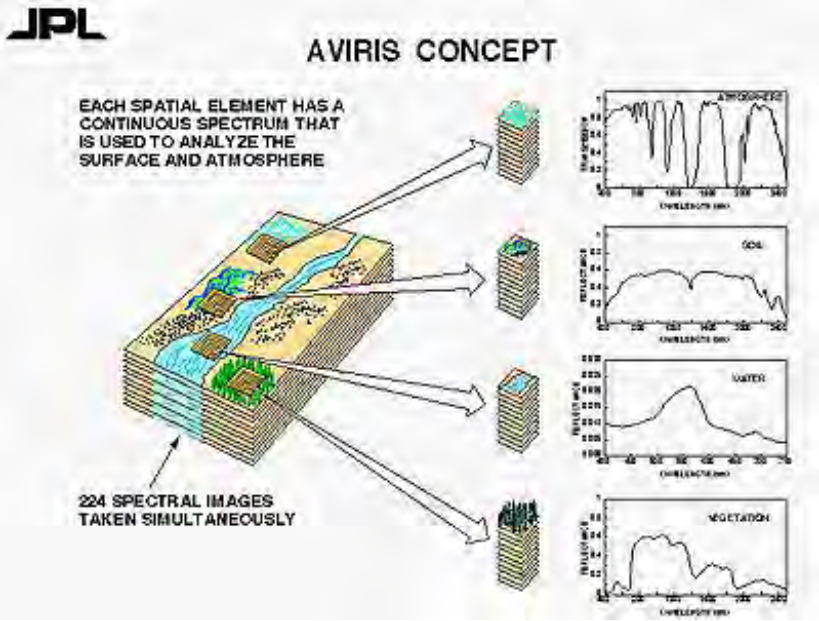
The design of most hyperspectral sensors is based on the pushbroom approach. The spectral sensing is accomplished usually by diffraction gratings (several may be included, to sample different parts of the spectrum). This is a typical schematic of such a sensor:



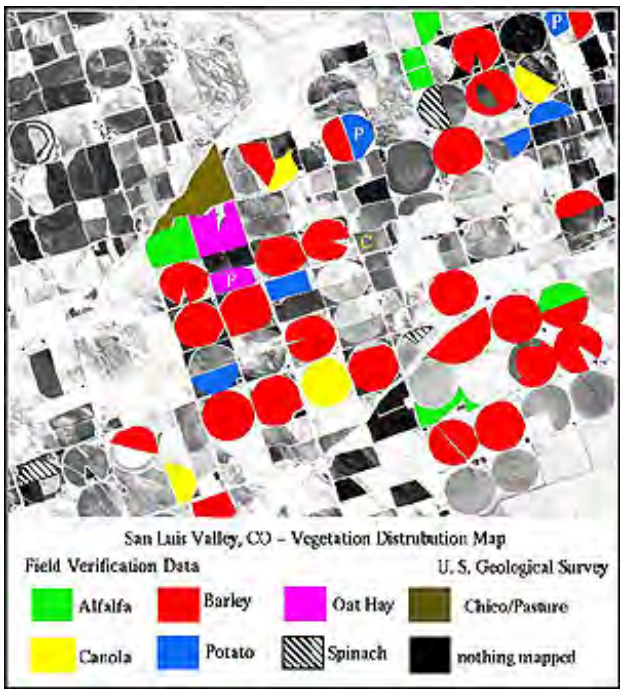
Hyperspectral sensors can be small enough to fit in one's hand. Some are used for signature acquisition of features examined on the ground. Here is a setup marketed by Photon, Inc.

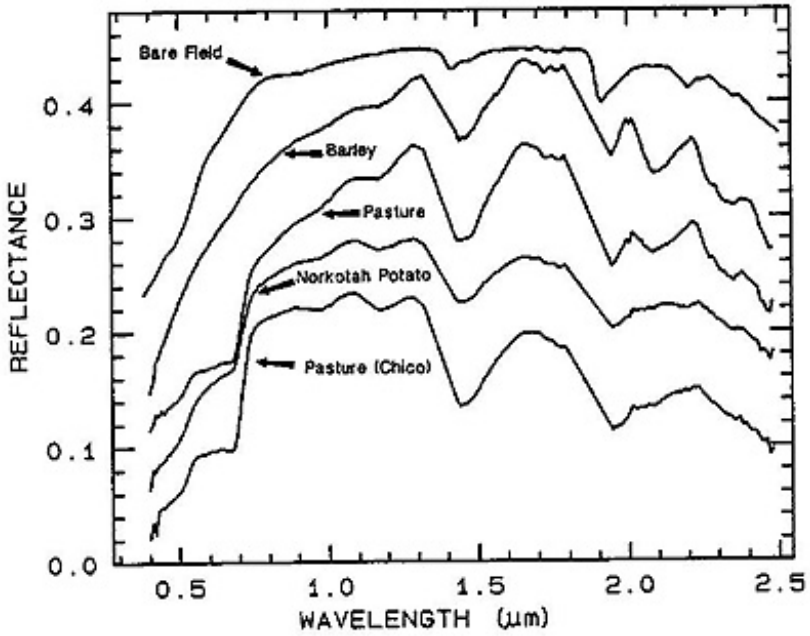
In ground operations, one normally remains still during signature acquisition. But rapid sampling permits flying hyperspectral sensors on aircraft. The Jet Propulsion Lab (JPL) has produced two hyperspectral sensors, one known as AIS (Airborne Imaging Spectrometer), first flown in 1982, and the other known as AVIRIS (Airborne Visible/InfraRed Imaging Spectrometer), which continues to operate since 1987. AVIRIS consists of four spectrometers with a total of 224 individual CCD detectors (channels), each with a spectral resolution of 10 nanometers and a spatial resolution of 20 meters. Dispersion of the spectrum against this detector array is accomplished with a diffraction grating. The total interval reaches from 380 to 2500 nanometers (about the same broad interval covered by the Landsat TM with just seven bands). It builds an image, pushbroom-like, by a succession of lines, each containing 664 pixels. From a high altitude aircraft platform such as NASA's ER-2 (a modified U-2), a typical swath width is 11 km.

From the data acquired, we can calculate a spectral curve for any pixel or for a group of pixels that may correspond to an extended ground feature. Depending on the size of the feature or class, the resulting plot will be either a definitive curve for a "pure" feature or a composite curve containing contributions from the several features present (the "mixed pixel" effect discussed in Section 13). In principle, the intensity variations for any 10 nm interval in the array extended along the flight line can be depicted in gray levels to construct an image. In practice, to obtain strong enough signals, data from several adjacent intervals are combined. Some of these ideas are elaborated in the block drawing shown here.

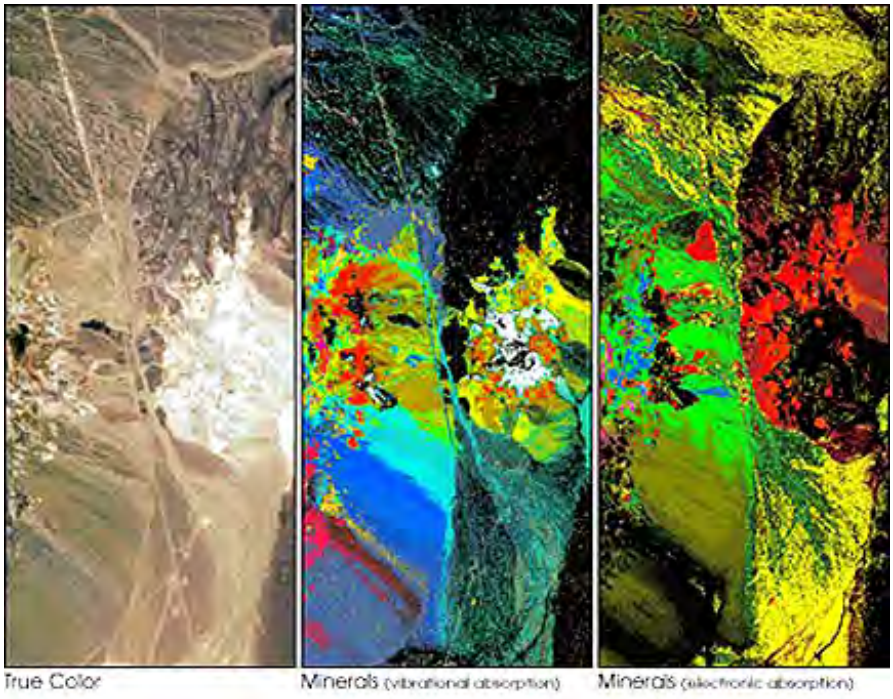


Below is a hyperspectral image of some circular fields (see Section 3) in the San Juan Valley of Colorado. The colored fields are identified as to vegetation or crop type as determined from ground data and from the spectral curves plotted beneath the image for the crops indicated (these curves were not obtained with a field spectrometer but from the AVIRIS data directly).





In Section 13 other AVIRIS images, used for mineral exploration near Cuprite, Nevada and other mining districts are displayed (see page 13-10) following an extended narrative on principles of spectroscopy and further consideration of the hyperspectral approach. A preview of the remarkable results achievable by this technology is given by this trio of images of the Cuprite district. The left image shows the area mapped as rendered in a near natural color version; the center image utilizes narrow bands that are at wavelengths in which certain minerals reflect energy related to vibrational absorption modes of excitation; in the right image, modes are electronic absorption (see page 13-7). Shown here without the mineral identification key, the reds, yellows, purples, greens, etc. all relate to specific minerals.



We know hyperspectral data are usually superior for most analyses to broader-band multispectral data, simply because such data provide so much more detail about the spectral properties of features to be identified. In essence, hyperspectral sensing yields continuous spectral signatures rather than the band histogram plots that result from systems like the Thematic Mapper which "lump" a range of varying wavelengths into single-value intervals. Plans are to fly hyperspectral sensors routinely on future spacecraft (see Section 21, page 21-1). The U.S. Navy is presently developing a more sophisticated sensor called HRST and industry is also designing and building airborne hyperspectral instruments such as ESSI's (Earth Search Science,

Inc., headquartered in Lakeside, MT) Probe 1.

I-25: **In your own words, using a single sentence, state the major advantage of hyperspectral sensors over broad band sensors. ANSWER**

A major goal in the space program has been to put several hyperspectral sensors on orbiting platforms. One such instrument, called Hyperion, is part of EO-1, the first satellite in NASA's New Millenium series, launched in December, 2000. It was inserted into an orbit that places it just about 50 km (30 miles) behind Landsat 7, which allows similar images acquired at almost the same time to be compared for performance evaluation. Operated by Goddard Space Flight Center, this satellite is a test bed for new ideas in instrumentation that can be made smaller and lighter, so that launch costs can be lowered. Here is an artist's illustration of EO-1, with its 3 main sensors:

Here is the actual hyperspectral sensor before it was mounted in EO-1:

The Hyperion consists of CCD detectors and other components that break the spectral range from 0.4 to 2.5 µm into 220 channels. Each resulting image is 7.5 by 100 km in ground coverage; resolution is 30 meters. This next image shows a color composite made with 3 narrow channels all in the visible in which the scene is of Maryland and Virginia along the Potomac River:

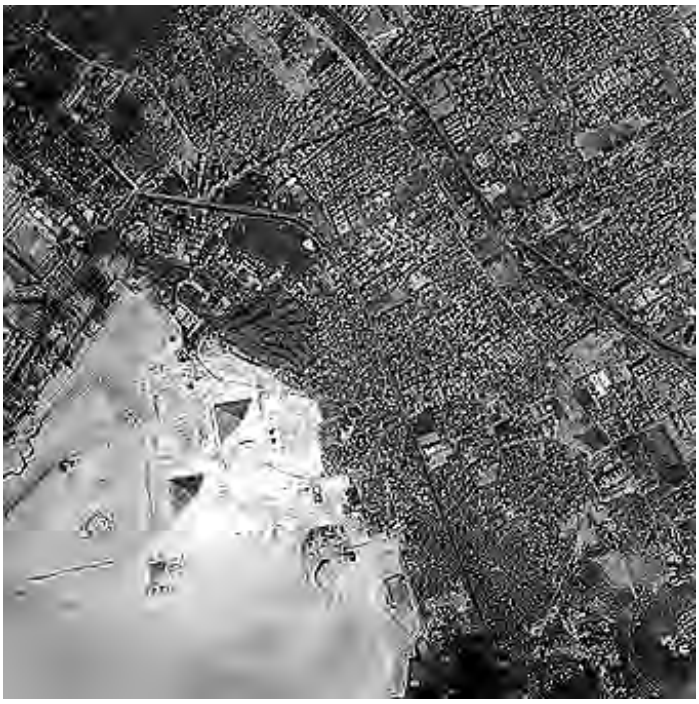
Hyperion images commonly are presented as long strips corresponding to down track scene acquisition. This image shows sedimentary rocks in a fold belt in the Mount Fritton area of the Flinders Range in South Australia:

The Atmospheric Corrector takes measurements that help to remove adverse effects from the atmosphere on image/data quality. A third sensor, the ALI (Advanced Land Imager) has 9 spectral bands and provides both multispectral images (30 m resolution) and panchromatic images (10 m). Here is an ALI image of the central part of Washington, D.C.

ESA has launched in 2001, as part of the MicroSat program, the satellite PROBA (Project for On-Board Autonomy). This small (40 x 60 x 80 cm; 94kg) boxlike system, with solar panel collectors on its surface, has remarkable image-making qualities. It is basically a hyperspectral system (200

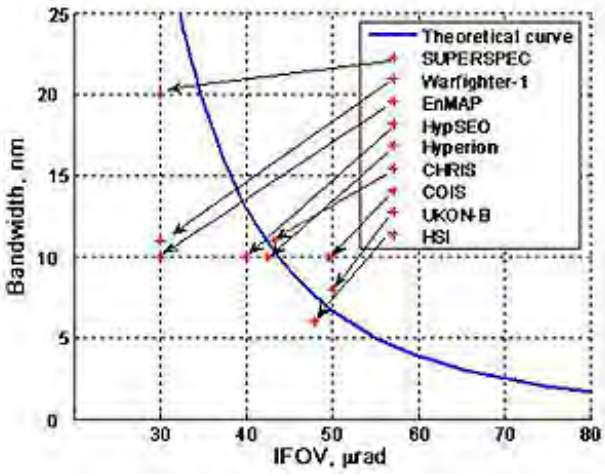
narrow bands) that image at 30 m, plus three in the visible that have 15 m resolution. Here is an artist's conception of Proba in space.

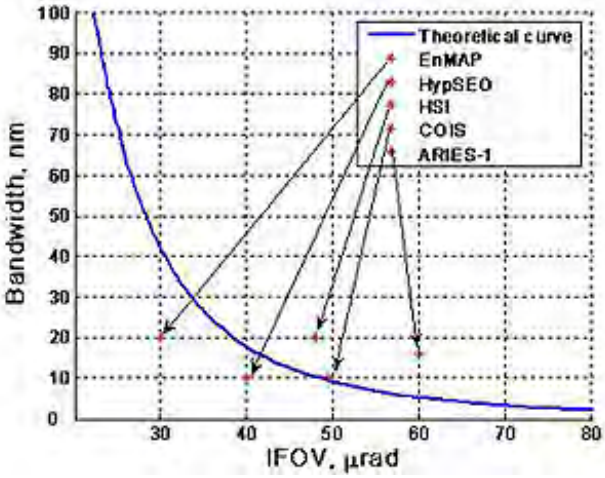
We show below 30 m color images of San Salvador Island in the Caribbean (where Christopher Columbus first landed in the New World in 1492) and Mauna Kea volcano in Hawaii, and a 15 m black and white image of the Great Pyramids of Egypt:



One is tempted to rename this satellite "Mighty Mite". But to be fully convinced of the power of hyperspectral remote sensing, you should read the pages related to it in Section 13.

We await the placement of other hyperspectral sensors into space orbit. The OrbView-4 commercial satellite had a 200 channel spectrometer on board, but this spacecraft failed during launch. EO-3 will have GIFTS (Geosynchronous Fourier Transform Spectrometer) to be used in hyperspectral data acquisition to analyze weather systems from a geosynchronous orbit. Some of the planned hyperspectral satellites, with data on channel bandwidth and on IFOV resolution, appear in this pair of plots:





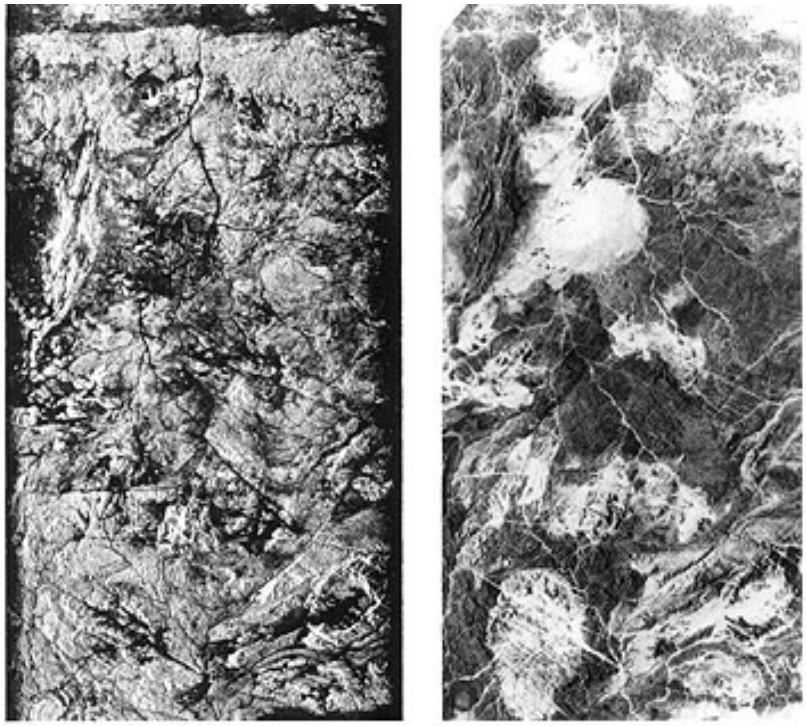
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Radar (an active microwave system) has been flown on both military and civilian spacecraft because of its ability (for certain wavelengths) to penetrate clouds. Seasat, the SIR series, and Radarsat are among the instruments used so far. ESA's ERS-1 and ERS-2 are also radar satellites. Thermal remote sensing, operating primarily in the 8-14 µm but also in the 3-5 µm wavelength region of the spectrum produces diagnostic data that can aid in identifying materials by their thermal properties. Some meteorological satellites have thermal sensors, as does the Landsat TM. HCCM was a mission dedicated entirely to thermal remote sensing. A multispectral airborne thermal scanner, TIMS, is described.

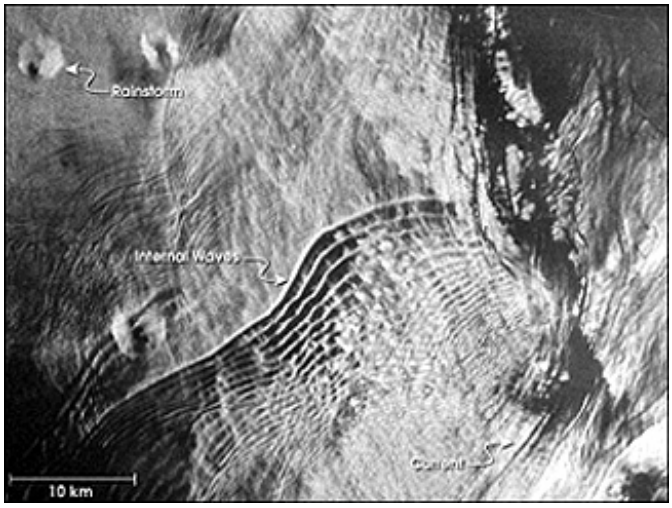
Other Remote Sensing Systems - Radar and Thermal Systems

Another class of satellite remote sensors now in space are radar systems (these are treated in detail in Section 8). Radar commonly provides a very different view of the same landscape compared with a visible image. This is obvious in this pair showing an ancient terrain in Egypt with fractures in a crystalline terrain evident in the left image (SIR-A radar) and plutons in the same scene in the Landsat image on the right.

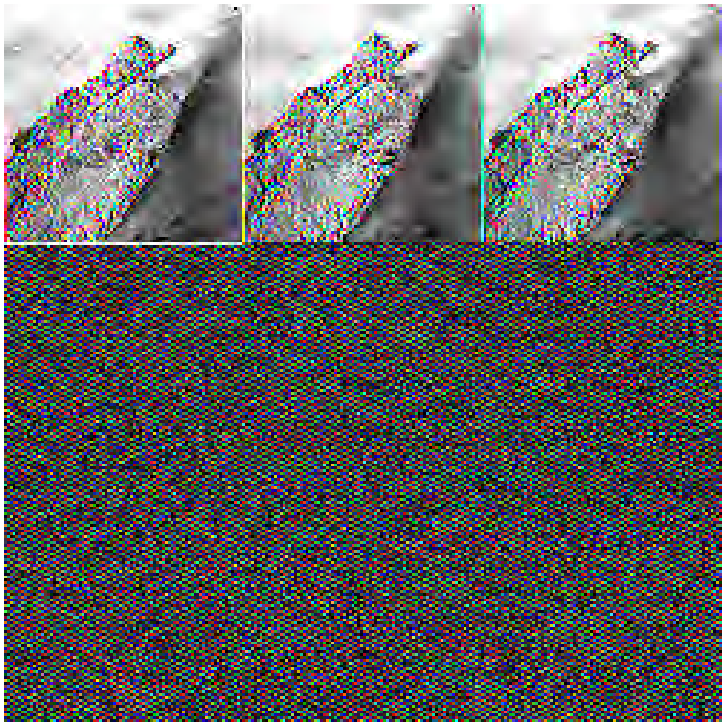


The first civilian radar system to operate from space was mounted in a Space Shuttle. The first independent spacecraft, Seasat, was an experimental L-Band radar whose primary mission was to measure ocean surfaces. It failed several months after launch in 1978 but did return many images that verified "proof of concept". Here is the spacecraft:

Seasat produced very informative images of the land and surfaces, including this scene that includes Death Valley, one of the prime test sites for determining the capabilities of various sensors. Below it are waves in the Gulf of Mexico:

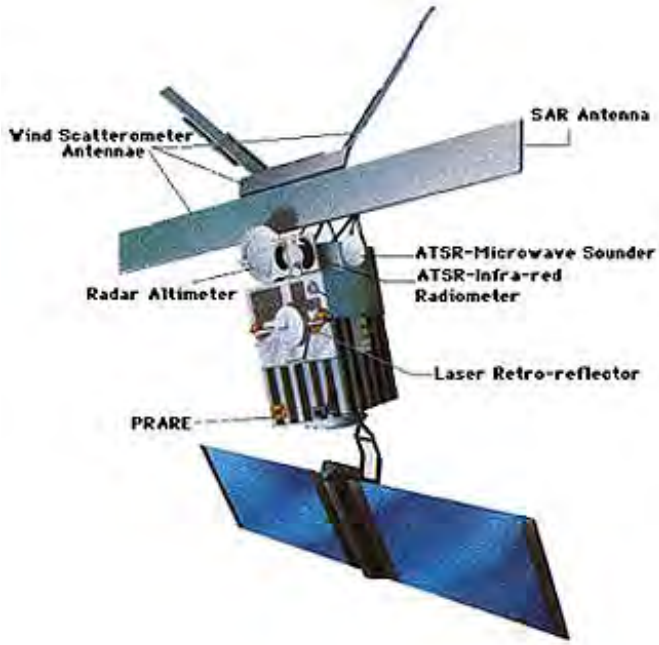


Among systems now operational are the Canadian Radarsat, ERS-1 and ERS-2 managed by the European Space Agency, and JERS-1 and JERS-2 under the aegis of the National Space Development Agency of Japan, NASDA. As an example, here is the first image acquired by Radarsat, showing part of Cape Breton in Nova Scotia, and the surrounding waters.

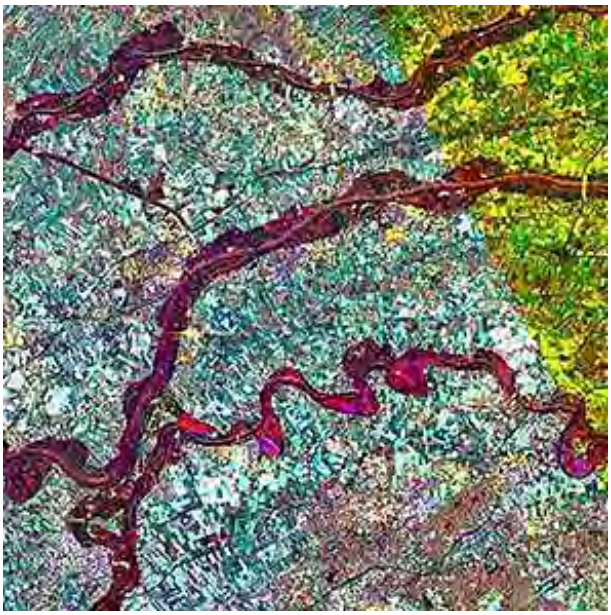


NASA, through its Jet Propulsion Laboratory (JPL) in Pasadena, California, has flown three radar missions on the Space Shuttle. The SIR (Shuttle Imaging Radar) series has used different wavebands and look conditions, with many excellent images over much of the globe having been acquired. Appearing below is a SIR-C image obtained on October 3, 1994 during a flight of the Space Shuttle. This is a false color composite made by assigning the L-Band HV, L-Band HH, and C-Band images to red, green, and blue respectively (see page 8-7). The area shown is that part of Israel containing disputed West Bank territory that includes Jerusalem (yellowish patterns on left) and the top of the Dead Sea.

The first major venture into the earth-observing satellite field by the European Space Agency were its ERS-1 and ERS-2 (ERS = European Remote Sensing) satellites. Each contains a SAR and other instruments. Here is a painting of ERS and a schematic of its instrumentation:



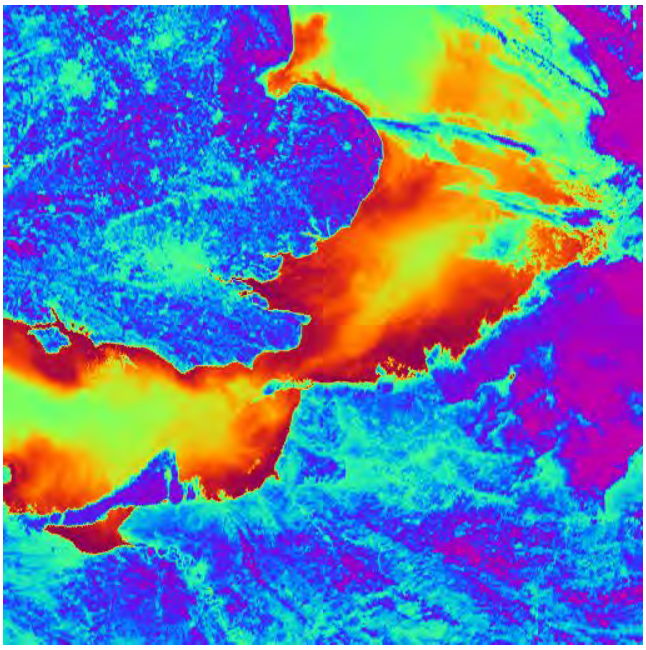
An example of the quality of radar imagery is this view of floods in the branches of the Rhine River around Nijmegen in Holland:



Here is an ERS-2 image in black and white showing the San Francisco, California, metropolitan area and the peninsula to its south, as well as Oakland, California, the East Bay, and beyond.

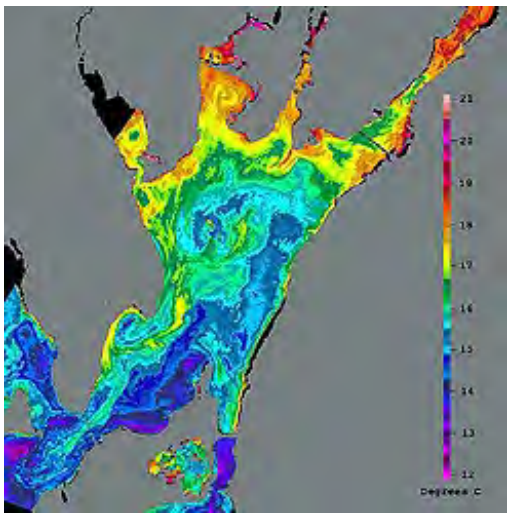
I-26: **Look at the above two radar images, especially the one showing San Francisco. State two characteristics of the radar images that seem to differ from those of Landsat. ANSWER**

The ERS satellites had (no longer operational) other sensors, as was indicated in the Overview. One in wide use is ATSR (Along Track Scanning Radiometer). Here is an image of the English Channel made by that instrument:

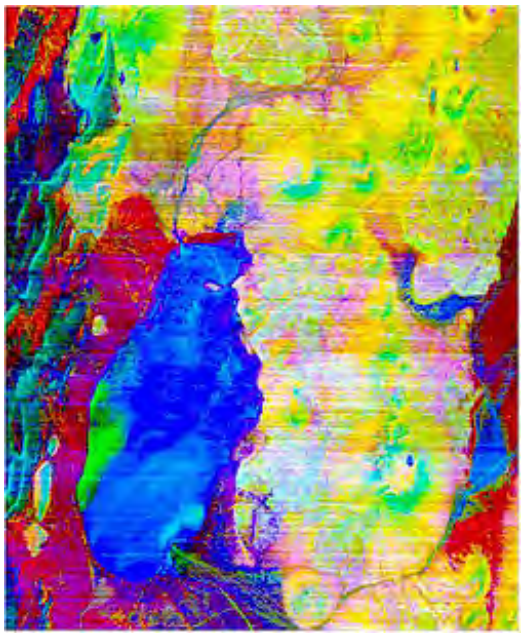


The principles behind thermal remote sensing are treated in some detail in Section 9. For now, let us look at representative samples of the types of thermal images that indicate the kinds of information resulting from operation of thermal sensors on moving platforms above the Earth's surface.

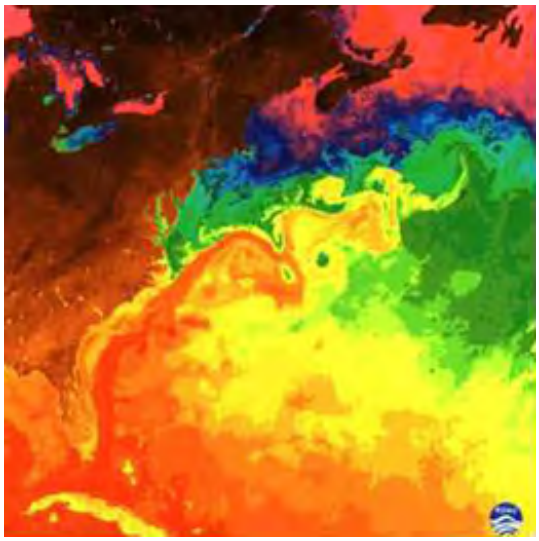
Remote sensors that cover two thermal intervals - the 3-5 µm and 8-14 µm broad bands (corresponding to two atmospheric windows) allowing sensing of thermal emissions from the land, water, ice and the atmosphere - have been flown on airplanes for several decades. Here are temperature variations in Mt. Hope Bay and part of Naragansett Bay, RI, made from an aircraft survey:



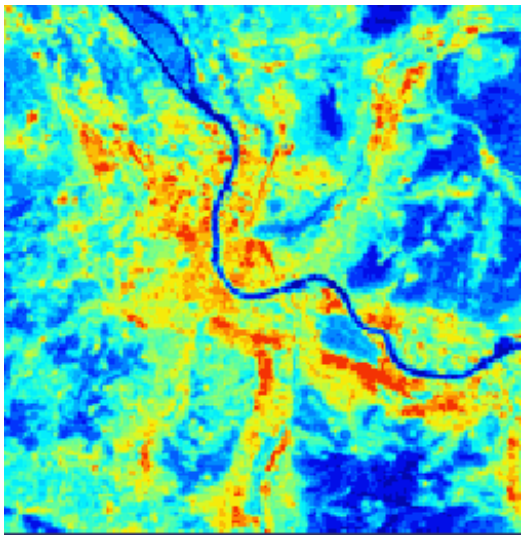
Thermal data, especially from the 8-14 µm region, become more valuable in singling out (classifying) different materials when this spectral interval is subdivided into bands, giving multispectral capability. NASA's JPL has developed an airborne multiband instrument called TIMS (Thermal IR Multispectral Scanner) that is a prototype for a system eventually to be placed in space. The images it produces are notably striking in their color richness, as evident in this scene that includes a desert landscape around Lunar Lake in eastern California.



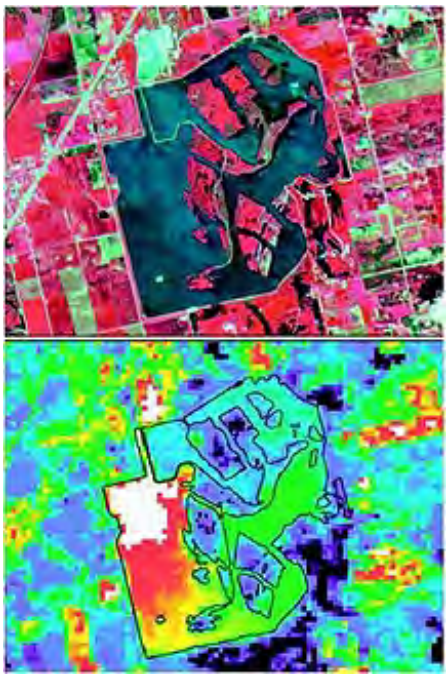
Thermal imaging has been done from various space systems. Many of the meteorological satellites (see next page) include at least one thermal channel. These include most geostationary satellites. This is a map of thermal variations off the east coast of the U.S. made from meteorological data:



A thermal band is included on the Landsat Thematic Mapper. This is a typical thermal map made from Band 6 data sensed the ETM+ on Landsat-7:



ASTER, on Terra, has a thermal band. Here are a false color image of fields and a large pond (strip mine waste) near Joliet, Illinois and the thermal band equivalent, which shows the thermal hotspot to be in the water.



This next image was made from a satellite dedicated to sensing one thermal property - thermal inertia (defined on page 9-3). The Heat Capacity Mapping Mission (HCMM) was launched in 1978 and is described on page 9-8. This image covers about 700 km (435 miles) on a side and was taken at night (daytime thermal images were also generated) on July 16, 1978 over southern Europe using a sensor that integrates thermal emissions within the wavelengths from 10.5 to 12.5 µm. The darker area in the upper left portrays lowlands in eastern France and southwestern Germany. The Alps form a broad arc crossing the image. The blackish pattern within the Alps corresponds to the cold higher elevations (with some snow). The lighter-toned land below the Alps is the Piedmont and western plains of Italy's Po Valley. The light tones near the image bottom are the waters of the Mediterranean Sea, which at night are warmer (heat sink) than most land surfaces.

While these color patterns make some sense when interpreted through geologic maps, aerial

photos, field visits (ground truth), etc., it is often hard to envision what they mean just from thermal images alone. Thermal images are best understood and utilized when they are combined with images covering other parts of the spectrum.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

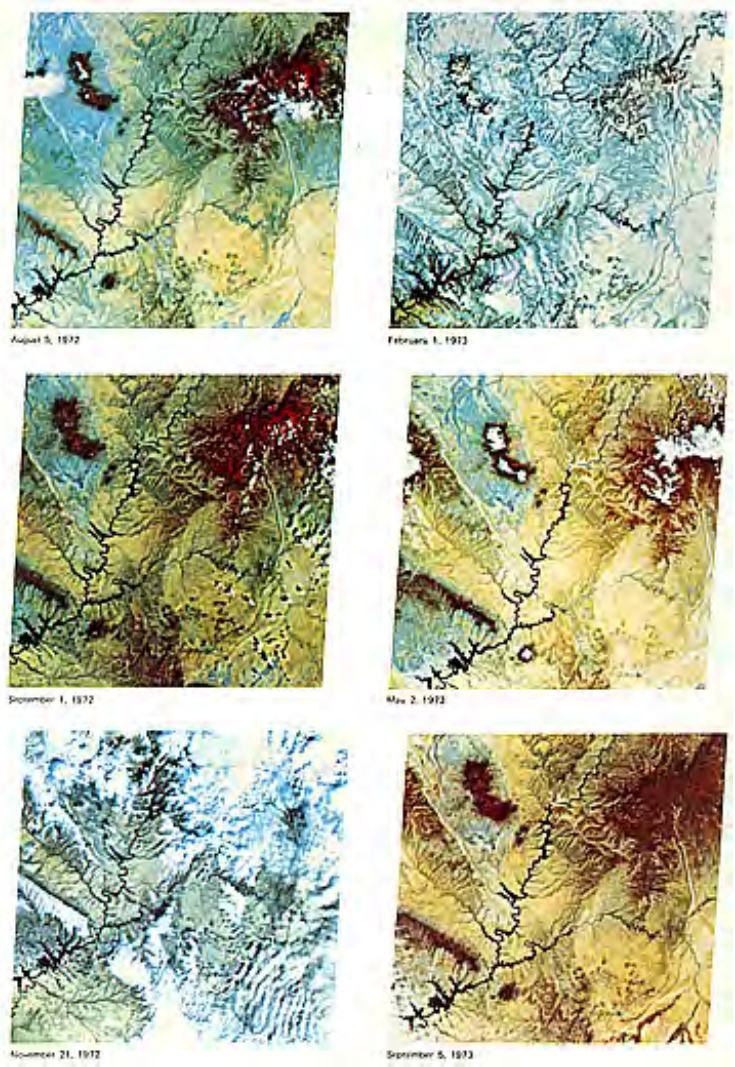
As may already be evident from earlier pages in this section, remote sensing data are of such nature and volume as to require the output to be compatible with processing and outputing by computers. They are the easiest, fastest, and most efficient way to produce images, extract data sets, and assist in decision making. One special function is to assist in manipulating other kinds of data about the spatial or locational aspects of areas in the world that are the subjects of interpretation and decision making. Today, the approach to analyzing a problem or determining a plan dealing with some aspect(s) of monitoring or managing these areas for specific uses or development is embodied in the concept of a Geographic Information System (GIS). This page previews this tool but Section 15 will be devoted to understanding its capabilities and applications.

The Systems (Multisource) Approach to Remote Sensing

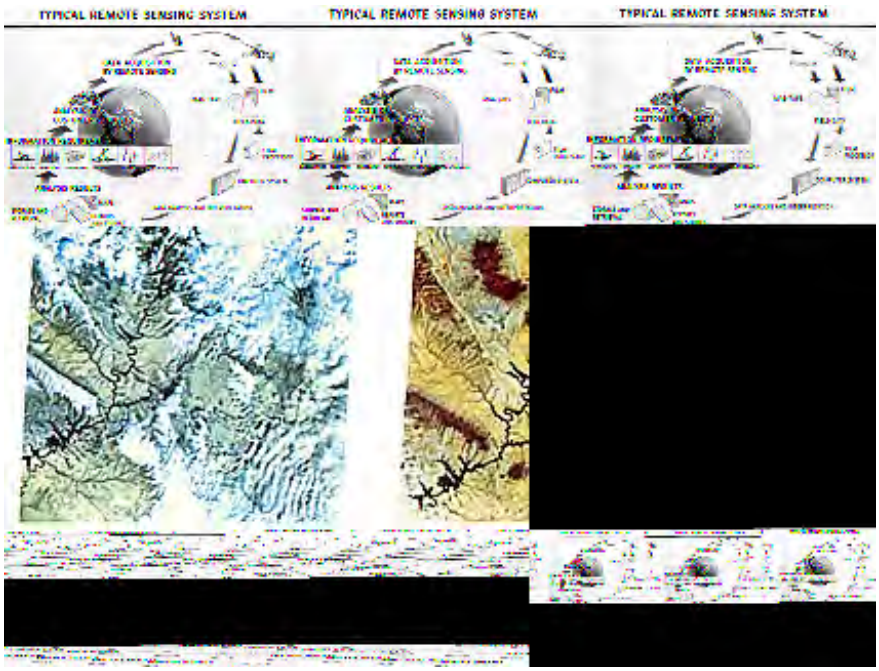
Since the early days of monitoring the Earth by orbiting spacecraft, the development of computer-aided techniques for reliably identifying many categories of surface features within a remotely sensed scene, either by photointerpretation of enhanced images or by classification, ranks in itself as an outstanding achievement. Numerous practical uses of such self-contained information are being made without strong dependence on other sources of complementary or supporting data. Thus, automated data processing assists in recognizing and mapping, as an example, major crop types, estimating their yields, and spotting early warning indicators of potential disease or loss of vigor. However, many applications, particularly those involving control of dynamic growth or change systems, or decision making in management of natural resources, or exploration for nonrenewable energy or mineral deposits, require a wide variety of input data of various kinds (from multisources) not intrinsic to acquisition by spaceborne sensors such as those on Landsat, SPOT, and others of similar purpose.

Some data are essentially fixed or time-independent - for example, slope aspect, rock types, drainage patterns, archaeological sites, etc. - in the normal span of human events. Other data come from measurements or inventories conducted by people on the ground or in the air - weather information, population censuses, traffic flow patterns, soil erodability, etc. - these can be gathered in time frames of months to years. However, many vital data are transient or ephemeral - crop growth, flood water extent, insect infestation, limits of snow cover, etc. - and must be collected in a timely sense. Pertinent remote sensing data play a key role in this last instance, and in fact satellite monitoring is often the only practical and cost-effective way to acquire data frequently over large regions.

A given scene imaged at different times of the year can show great variety. Changing Sun angles, atmospheric variations, seasonal differences in vegetation cover, presence of snow, and other variables will produce often pronounced contrasts in the spectral responses that determine "how an image looks". This is evident in this montage of 6 Landsat MSS images of an area in the desert of Utah, in what is called the Canyonlands region. The seasons involved are: Upper to Lower Left: August 5, September 1, November 21, 1972; Upper to lower right: February 1, May 2, November 6, 1973



One should always keep in mind that remote sensing is an integral part of a larger Information Management System. In fact, in many applications, the user community employs remote sensing inputs as a key component of a continuing cycle of decision making. Consider this diagram:



This chart shows a simple closed-loop cyclic process, unencumbered by the various feedback loops that no doubt exist. The starting point, and end point as well, is the set of panels labeled Information Requirements. This focuses on the ultimate driver of any information management system: the user and his/her recurring needs. Various disciplines concerned with Earth

observations and resources are represented (one not shown is Meteorology). The terrestrial globe in the background reminds us that the system should be worldwide in scope. Information requirements logically lead to user/customer demands. The best remote sensing system approach is the one most responsive to these demands.

There has by now been full realization that the best current and future uses of most Earth-observing data from satellites (or astronauts) stem from correlating and interleaving this type of data with various other types that together are essential inputs to decision making and applications models. This is embodied in the "Multi" concept, described in Section 13 (page 13-4a ff), but summarized here by these terms: Multistage; Multilevel; Multisensor; Multispectral; Multitemporal; Multisource. Remote sensing data constitute an integral element of a more general Earth Survey Information System, as exemplified here:

The bulk of the data in such systems have in common a geographical significance, that is, they are tied to definite locations on the Earth. In this sense, they are similar to or actually make up what has become a powerful tool in decision making and management: the Geographic Information System (GIS; also known as geobased or geocoded systems). Because vast amounts of spatial or geographically referenced data must be gathered, stored, analyzed in terms of their interrelations, and rapidly retrieved when required for day to day decisions; a GIS that accepts these data must itself be automated (computerized) to be efficiently utilized.

The importance of GIS as a unifying means of handling geospatial data, including often mandatory inputs from remote sensing, warrants an extended explanation of how its works and what it does. This is the subject of Section 15. In Section 1 you will learn how computers with appropriate software are an essential part in processing, manipulating, and integrating data such as is the output of Landsat and other systems. It is safe to say that today, without computers, remote sensing from space would be next to impossible.

PREVIOUS rst NEXT

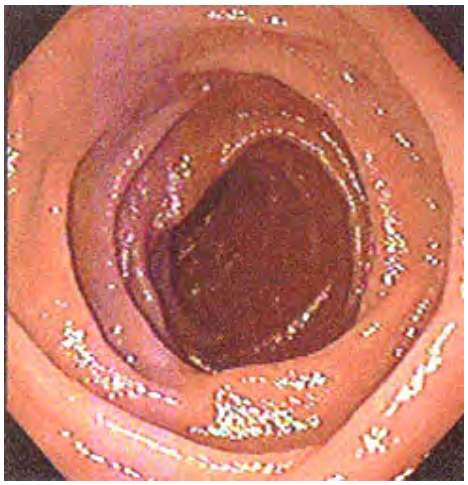
Primary Author: Nicholas M. Short, Sr.

The use of various instruments/machines as diagnostic tools in medical examinations falls within the broad definition of remote sensing, although the target or surface being analyzed is close to the sensor, which may be exterior to the body, placed on the body's surface, or inserted inside the body to examine internal organs. Electromagnetic radiation is the sensing medium in most analyses; sound waves are also applied. Both active and passive sensors are used. The usual end product is an image. Most medical remote sensing is designed to "see into" the body without having to be invasive (cutting it open). Some techniques produce only static images; others can actually display the features being examined in dynamic, real time images which show the functional movements of the organ(s) within the body. On this page, one of the earliest medical imaging devices - X-ray units that use radiography or fluorescence to produce the pictures - is described.

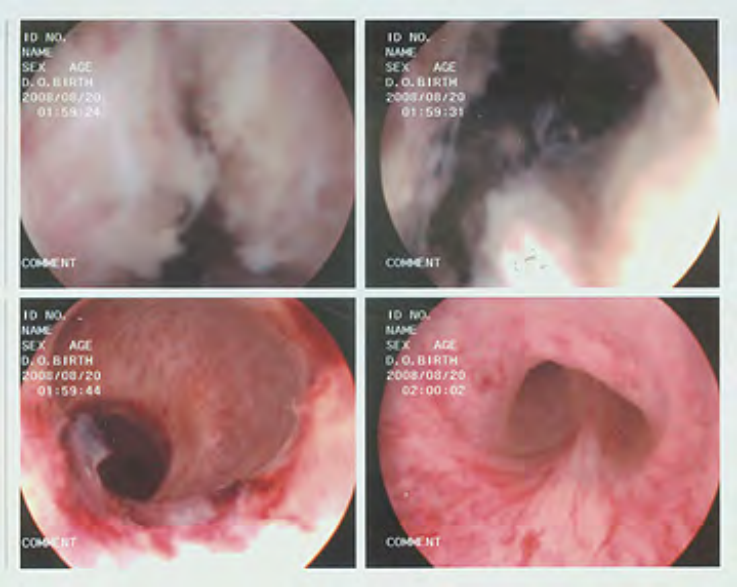
Medical Applications of Remote Sensing

The next three pages are concerned with the applications of remote sensing to the medical profession's need to examine humans and animals by sophisticated imaging instruments. Many of the techniques of using high-powered instruments to send electromagnetic and sonic waves into the human target (we shall assume that "animal" is understood implicitly to be included since at least some of these instruments have been used to examine dogs, cats, horses, etc.) fall within the broader definition of remote sensing - electromagnetic radiation photons (at different wavelengths) or sonic wave trains are generated and coupled to the body and then detected as transmitted, absorbed, or reflected signals to an external detector a short distance away. Most medical remote sensing is of the active mode, i.e., EM radiation or acoustical waves generated by the instrument are sent into or through the body. Some examinations of bodily functions depend on implanting (by injection or swallowing) a source of radiation, such as radioactive element(s) as tracers which can be sensed by appropriate detectors as they move about (as in the blood) or concentrate within organs - this approach is analogous to passive remote sensing.

Two common uses of imaging probes are an endoscopy and a colonoscopy. This again had elements of classical remote sensing (electromagnetic radiation associated with the scanning light) except that the signal from the moving target beyond direct reach passed through a flexible fiber optic cord rather than air or empty space. As introductory examples, below are two of the results of probing to into the human digestive system and bladder:

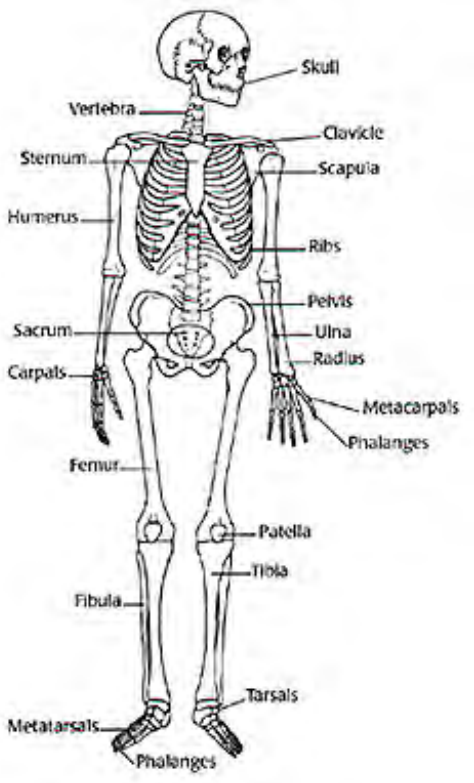


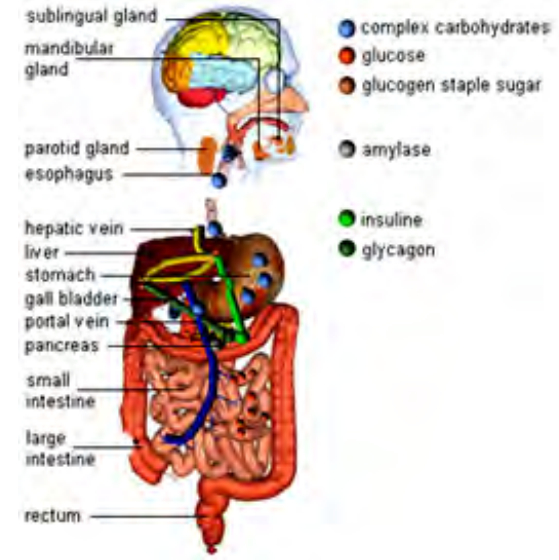
<p><div align=

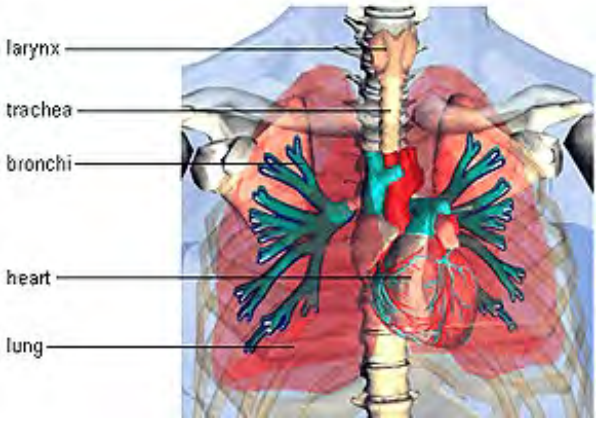


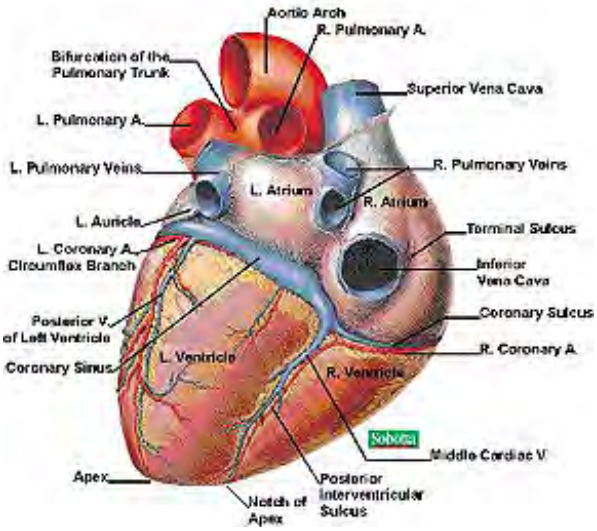
The subject of medical remote sensing (more commonly referred to as medical imaging) is now a major topic covered on the Internet. A good synopsis is online at this Wikepedia entry. Of general interest is this Imaginis site.

Before we begin, a word about the human body (for more details, consult the City University of New York website. The anatomy of the body is treated in several general categories, chief of which are the human skeleton, the internal organs, and the heart. These four illustrations are almost self-explanatory:









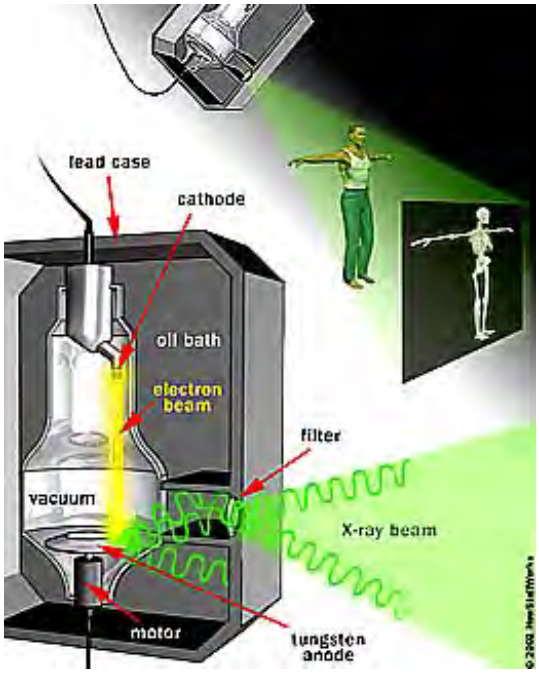
In the Tutorial, we will cover these methods: X-ray Radiography; X-ray Fluoroscopy; Computer Assisted Tomography (CAT scans); Magnetic Resonance Imaging (MRI); SPECT (Single Photo Emission-Computed Tomography; Positron Emission Tomography (PET); CATscan-SPECT combined; Infrared Imaging Thermography; Ultrasound; and Endoscopy. PET and SPECT also fall in the realm of nuclear medicine. These various methods can produce "static" images or can be viewed in real time to examine "movements" within the body. Also, some methods concentrate on skeletal parts (bone), others on soft-tissue internal organs (e.g., brain; heart; kidneys); others on circulation and other functions. Most methods are used to detect abnormalities such as malignant growths, bone breaks, and disease effects.

Modern medical imaging began with an almost accidental discovery in the lab of Professor Wilhelm Roentgen in Germany on a November day in 1895.

Roentgen was experimenting with a Crooke's Tube he had recently obtained from its inventor. This is a glass vessel from which air is withdrawn creating a near vacuum; at one end is an anode (positively charged) and at the other a cathode (negatively charged source of electrons); the tube is wired to be part of an electrical circuit. When a current is passed between these electrodes, the few particles within the tube are excited and fluoresce or glow (commonly blue or green); this results from the flow of high speed electrons (cathode rays) across the (voltage) potential difference imposed in the circuit. Roentgen had placed the Tube in a black box but to his amazement noted that a fluorescent screen nearby was glowing which he deduced to be excitement of its phosphors by radiation escaping the box. This unknown (X) radiation he simply labelled X-rays (they are also called Roentgen rays). As he studied their properties, he experimented by putting a hand on a fluorescent screen directly in the path of this radiation, getting this famous picture:

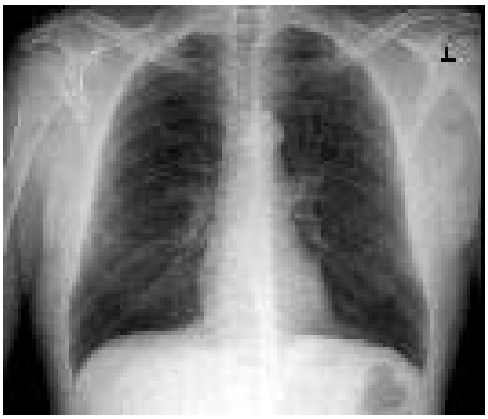
Soon others were experimenting with X-rays. The first medical uses of X-ray machines occurred within a year. Roentgen's achievement was recognized in 1901 when he received the first Nobel Prize in Physics. A fascinating account of his discovery is given at this Internet site.

(The physics principles appropriate to this paragraph are reviewed on page I-2a.) X-rays are produced when electrons are impelled against an anode metal target (tungsten; copper; iron; molybdenum; platinum; others) as they pass through a vacuum tube at high speeds driven by voltages from 10 to 1000 kilovolts (kV). When incoming electrons interact with inner electrons in the metal, these latter are driven momentarily to higher energy levels (these orbital electrons are pushed into outer orbitals); when these excited electrons drop back to their initial orbits (a transition from a higher to a lower energy level), the energy they acquired is given off as radiation, including X-rays (wavelengths from 0.03 to 3.0 nanometers). Some of the scattered X-rays are collimated into beams (typically at conical angles up to 35°) that are directed towards targets (such as the human body). Soft body tissue absorbs less X-rays, i.e., passes more of the radiation, whereas bone and other solids prevent most of the X-rays from transmitting through their mass. (X-rays have other uses, such as examining metals for flaws or determining crystal structure.) Here is a diagram of a typical X-ray machine setup:

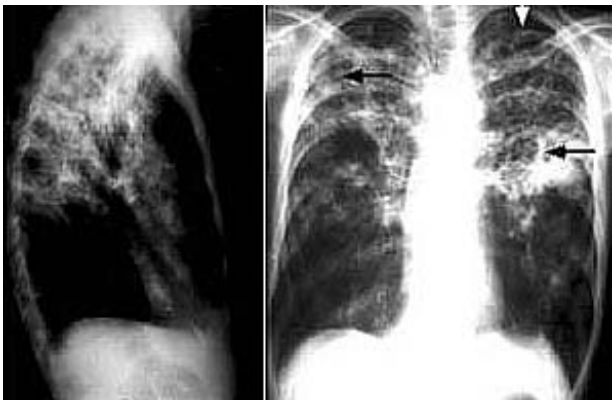


Two classes of detectors record the X-ray-generated image: 1)Photographic film, in which the difference in gray levels or tones relates to varying absorption of the radiation in the beam impinging on the target. The convention is to use the exposed film (X-rays act on the silver halide {see page I-12 of this Introduction} to reduce it to metal silver grains) as a negative. In its negative form, bone will appear nearly white (as said above, because bone absorbs efficiently, few X-rays strike the corresponding part of the film, leaving it largely unexposed; the soft tissue equivalents pass much more radiation and darken the film); 2) fluorescent screens, that include phospors (compounds that fluoresce or phosphoresce) coating a substrate; this occurs when electrons in the phosphors jump to higher level orbitals, with visible light given off either instantly when the electrons transition back to the lower state or with a time delay fractions of a second or seconds (afterglow), in a process similar to X-ray production; typical phosphors include Calcium tungstate or Barium lead sulphate; these screens in certain configurations allow realtime movements of the medical patient to be observed and the sreen images can be photographed or digitized.

X-ray radiology is still the most commonly used medical instrument technique. Here are a sequence of images that illustrate typical uses and results. The first is a chest X-ray:

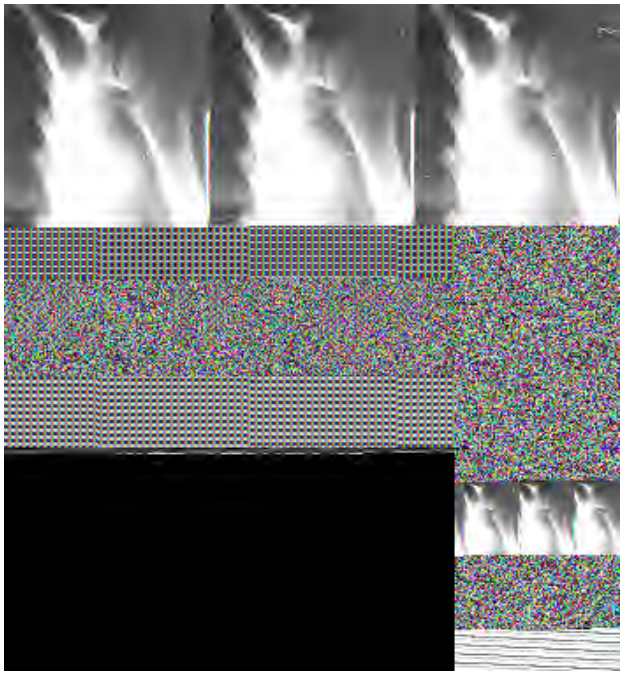


This next is a front and side view of the upper torso; the arrow points to a tuberculosis patch in the left lung:

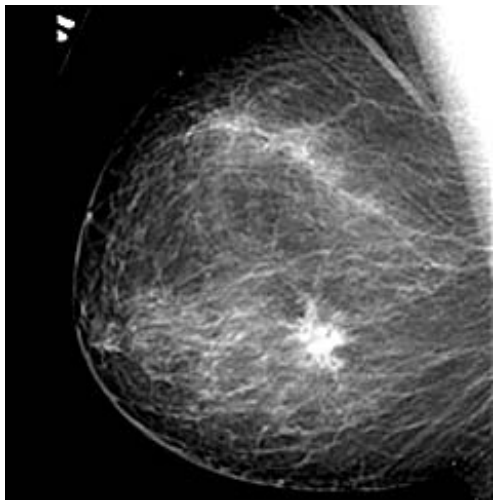


Here is a negative X-ray film image of the pelvic area:

This next X-ray image shows the upper arm and shoulder area. There is a slanted break across the arm (humerus bone) near the shoulder socket. Fractured in several places, this X-ray image is that



This next picture is a mammogram showing a growth in the female breast:



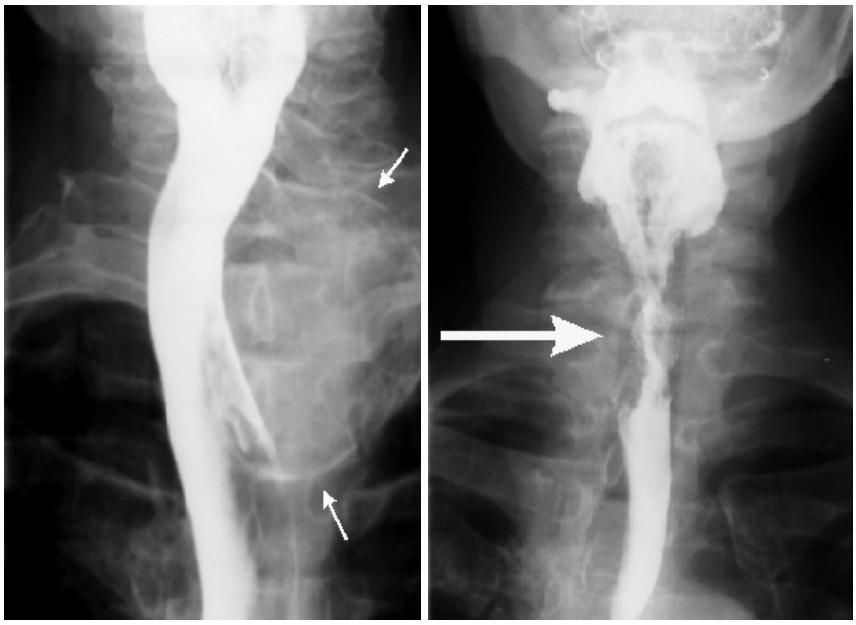
The human skull is x-rayed mainly to spot signs of fracture. But, sometimes indications of tumors are present, as shown by the darker gray patch in the cranium of this individual's skull:

The jaw and teeth are evident in this lateral view of the lower human skull:

Most of us gain our first experience and insight into X-rays diagnostics when we have a small film inserted into our mouth and then the X-ray machine is placed against that part of our jaw. Here is a typical X-ray image of teeth, in which the whitest part of the negative corresponds to metal fillings (great absorbers):

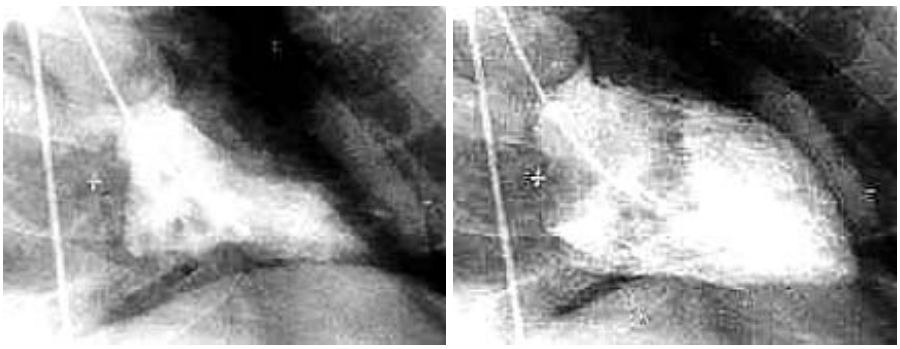
An important variation in X-ray radiography is *Fluoroscopy*. In this method, either chemicals that react with X-rays are swallowed or inserted as an enema or chemicals/dyes are injected into the blood stream. These tend to increase the contrast between soft tissue response in the parts of the body receiving these fluids and surrounding bone and tissue. This pictorially highlights abnormalities.

Barium sulphate is a good example. When swallowed (either at once or commonly in gulps), the "Barium Cocktail" is especially useful in examining the digestive track. In this image, an obstruction in the esophagus carrying food and liquids into the stomach is made evident:

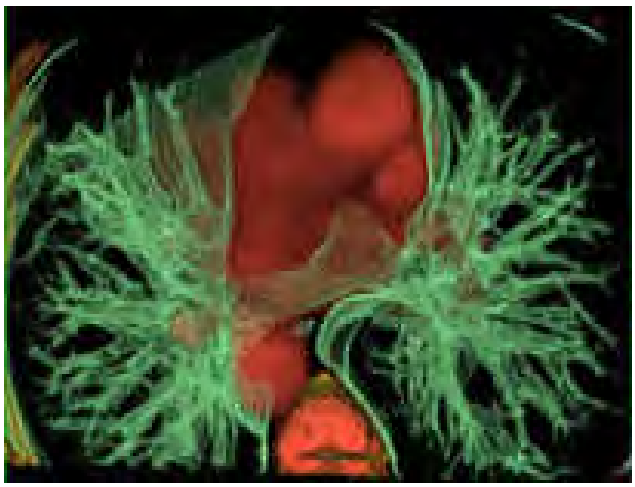


The large intestine or colon is strikingly emphasized in a patient who has just received a Barium enema:

Still another variant is the *Angiogram*. This involves insertion of a catheter into an artery, accompanied by a dye that reacts to X-rays. It is commonly used to explore the areas in and around the heart. Here is a pair of views of the left ventricle of the heart when it is pumping and squeezing blood and thus contracting (systolic phase) and then expanding as blood is returned (diastolic phase):



This next image is an angiogram that has been colored to show blood vessels including the great trunk artery or aorta around the heart:



Using special methods, angiogram-like images can be made for the blood vessels in the human head:

We move on now to a powerful new approach to medical imaging, based on the technique of *tomography*, which uses computers to assist in obtaining three-dimensional images or image slices when either X-rays or radioactive elements (nuclear medicine) are involved in producing radiation-based imagery.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

NOTE: THIS PAGE HAS BEEN PREPARED BY NICHOLAS M. SHORT; WHILE SOME OF ITS CONTENT IS RELATED TO COMPARABLE TOPICS PRESENTED IN APPENDIX A (MODERN HISTORY OF SPACE) CONTRIBUTED BY AIR FORCE ACADEMY PERSONNEL, THE INFORMATION GIVEN HERE HAS BEEN ENTIRELY DETERMINED BY THIS WRITER WHO ACCEPTS SOLE RESPONSIBILITY FOR ITS VERACITY AND PERTINENCE.

This page is devoted entirely to one of the prime uses of remote sensing that has driven the entire technology since the early days of the entry of satellites into space. Depending almost exclusively on imaging capabilities, "spy satellites" (more formally called "reconnaissance satellites") have been orbited by the hundreds (by several countries) to gather military intelligence or information about terrorist activities. Visible, Near-Infrared; Thermal Infrared, and Radar sensors are applied to gathering information about ground targets and activities of national security significance. Many of the military or intelligence satellites, up until recently, have had superior resolutions when compared with Space Agency systems.

Military Intelligence Satellites

Looking down and out (as from a mountain) to survey the battlefield for information useful to military leaders goes back to ancient times. In Napoleonic times, the French used observation balloons to scan their foes before and during battles. This technique was often a factor in the U.S. Civil War. By the First World War, airplanes and dirigibles were employed over enemy lines and their staging areas and cities as platforms from which aerial photography provided reconnaissance and intelligence pertinent to the content of battle. This approach was much expanded during the Second World War, as for example the follow-ups to a bombing raid to assess damage to the target.

With the advent of rockets and then satellites, observations of both military and political activities on the ground became possible, ushering in the so-called Age of Spy Satellites. Since the beginning of entry into space, hundreds of these satellites have been launched, first by the U.S. and the Soviet Union and then other nations. Besides surveillance of a wide variety of targets of interest to military intelligence units (in the United States, these include the Department of Defense, the CIA, the National Security Agency, and Homeland Defense), satellites can now assist in areas other than simply observing features on the ground - this includes communications, meteorology, oceanography, location (Global Position Systems [GPS]), and Early Warning Systems (none of these latter applications will be discussed on this page). In addition to satellites, manned aircraft continue to be platforms and in recent years UAV's (Unmanned Aerial Vehicles) such as drones have assumed some of the intelligence-gathering tasks.

As one would suspect, there is extensive coverage of "spy satellites" on the Internet, although much information remains classified and thus not released to the public. On this page, only the broadest outline of the history of military intelligence operations from the air and space in the last 50 years will be treated, along with some representative image examples that are now declassified into the public domain. The reader is offered these five Internet sites as among the best found by the writer. Two top the list: 1) An overview (click on that link) prepared by Federation of American Scientists, with an offshoot or link from the same organization that shows specific imagery; and 2) this site, with an information history supported by good imagery, produced by a group at George Washington University. Also worth a visit: SpySat. It's your

choice, either visit these first or read further on this page (most images come from these Web pages just cited) and then check out the above sites.

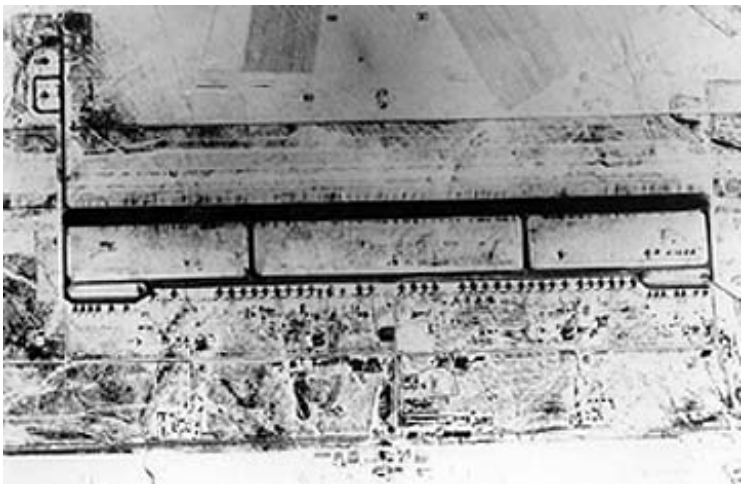
Before moving through this page, which will only synopsize part of the story, we suggest going to one more Web site that specifically includes the effects of spatial resolution in military satellite surveillance. This is again two of the links on the FAS Imint site (1)and (2). The single idea to draw from these illustrations is that military observations work best and reveal the desired intelligence when resolution is a few meters or better. Note what can be found and, more importantly, identified at each resolution level.

As you should expect, the military reconnaissance satellites are usually very sophisticated - advanced platform and advanced sensors. There aren't many pictures or sketches of spy satellites on the Internet. Here's one - as you might predict, it is not named; the black surroundings suggest it was actually photographed while in space (during a servicing mission from the Shuttle - yes, the Pentagon has allowed military activities on this so-called civilian system):

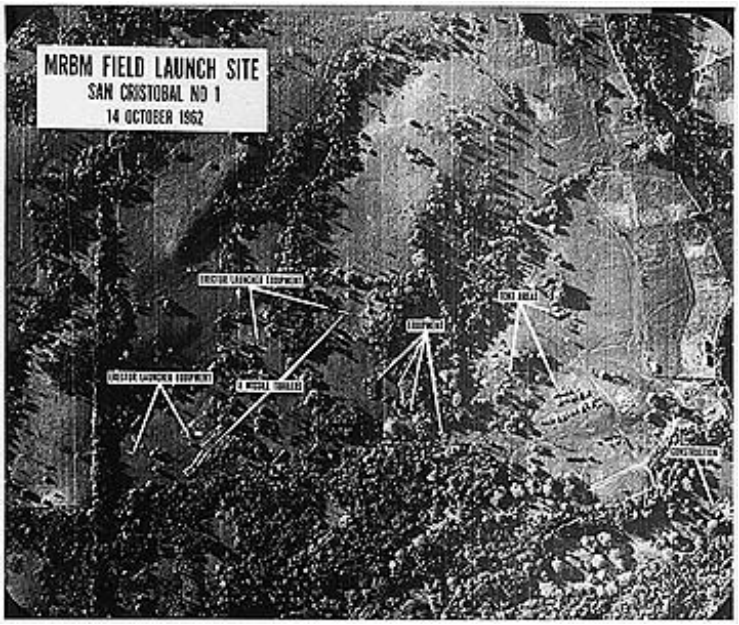
For nearly four decades the military had high resolution systems - this was their most salient advantage - that could not be matched by non-military earth observation systems (i.e., there were imposed limits of resolution below which (namely, towards greater/better resolution [picking out smaller objects]) no civilian space agency or private group were permitted to design into the sensors on the satellites they operated). All this changed in the 1990s when the Russians began to sell high resolution (~2 meters) imagery from their SPIN-2 on the open world market. After that the U.S. placed more than 800,000 of its earlier military space photos into the public domain.

Now to some specifics: When one thinks of any postwar spy from the sky incidents, the first and most famous case that many recall was the U-2 high altitude airplane that was shot down over the Soviet Union during President Eisenhower's second term in which the pilot, Gary Powers, was captured and held for many months.

Here is an example of a spy reconnaissance U-2 photo (from another mission) over a military air base.



The U-2 achieved even more fame during the October 1962 Cuban Missile Crisis, when Pres. John F. Kennedy went to the brink with Soviet Chairman Nikita Krushchev over the installation of Medium and Long Range nuclear rockets in parts of Cuba. The next two U-2 images show the facilities that were in place before the Soviet/Cuban block agreed to remove these weapons (probably avoiding World War III or at least partial annihilation):

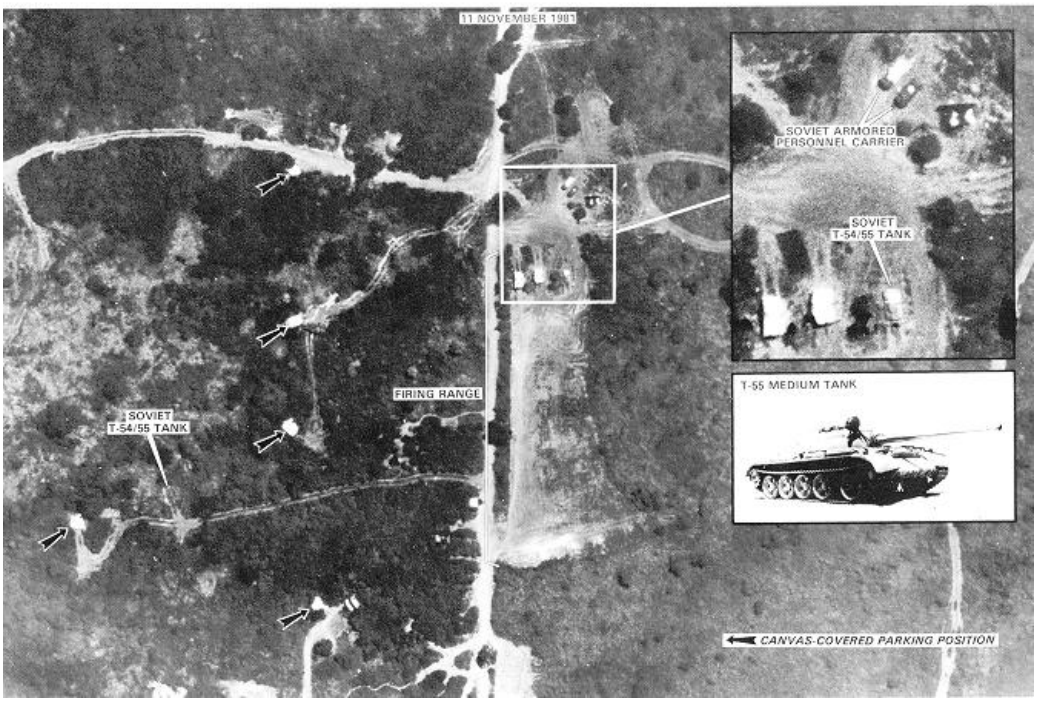


The U-2 is talked about again in Section 10 in connection with a mission flown for the writer

(NMS) in Wyoming as part of a Landsat geology study. The plane used is one of the two that are flying over San Francisco in this photo:

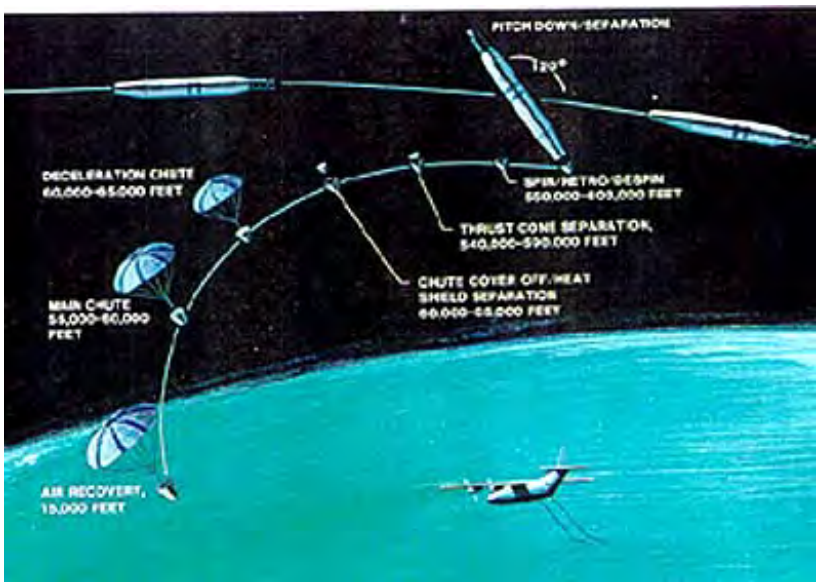
Added to this fleet of spy planes was the SR-71, known as the Blackbird. Here it is in flight; beneath is a photo shows a military storage base in Nicauragua (in 1987) :

TEMPORARY ARMOR STORAGE AREA
MANAGUA, NICARAGUA

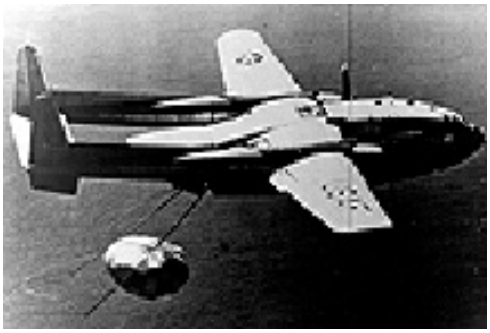


The first military satellites carried photographic cameras that had high resolution optics. This required well-constructed automatic photo systems with a reliable film advance. Here is a cutaway of the Corona nosecone showing the photographic setup:

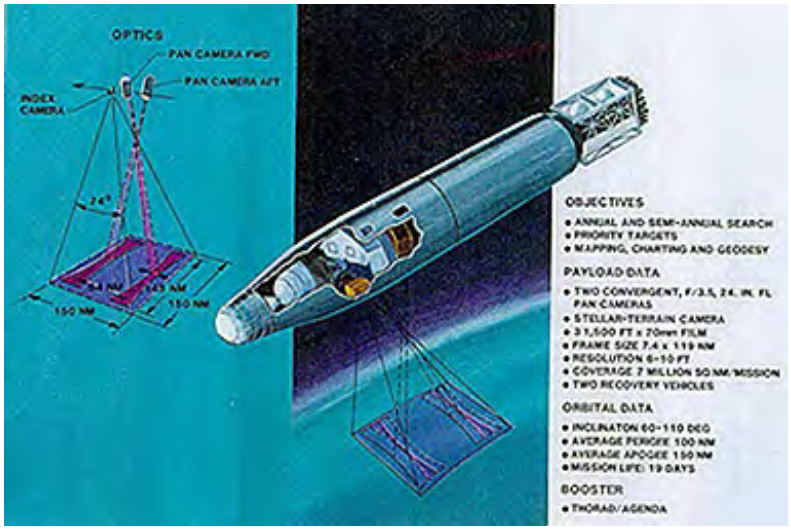
The problem, of course, was that of retrieving the film (no outer space drugstores). It could have been developed onboard like was done with the Lunar Orbiter pictures. Instead, the operators of this type of satellite chose to eject the (undeveloped) film towards Earth, with its container deploying on a parachute when the object entered the upper atmosphere. The film package was then "snatched" in mid-by an airplane sent to its expected point of entry. This seemingly difficult feat was remarkably successful, accomplished most of the time. This diagram shows the sequence:



And here is a C-119 aircraft just after it snagged the film dropped on a parachute:

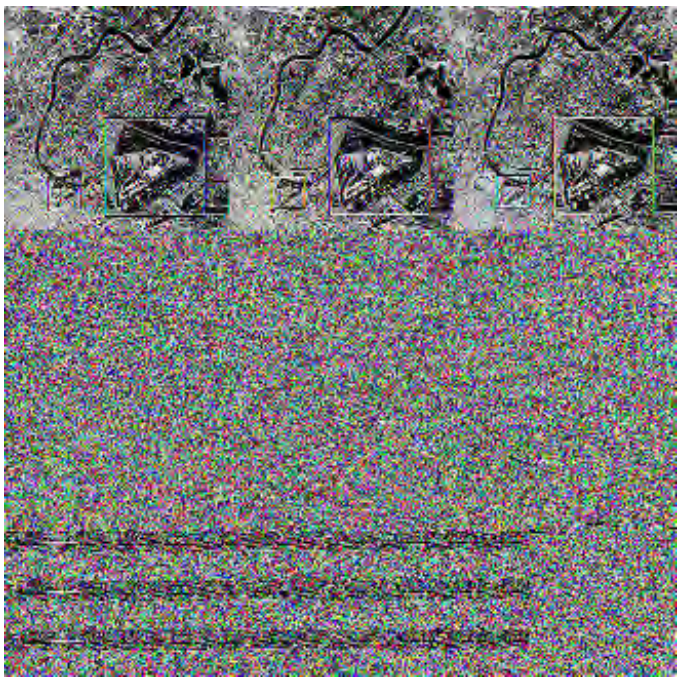


The primary group of U.S. military satellites used in reconnaissance and surveillance is known as the **KeyHole** series. The first four of these were given the code name CORONA. The first of these was placed in orbit in 1960 and used the parachute retrieval system until 1972. Each numbered KeyHole mission series consisted of multiple satellites. Different identifiers denoted specific satellite types that varied in launch vehicle, orbital characteristics and satellite instrumentation. For example, the KH-1 through KH-4 series also included the appellation **Corona**, KH-5 is called the **Argon** series, and KH-6 the **Lanyard**. Each involved various numbers of satellites. There were 105 successful Corona missions, which were operated by the U.S. Air Force, with CIA involvement. Spatial resolution, initially about 2 meters with the first KH-1 launch in 1960 to work properly, steadily improved with the higher KH numbers. Here is a diagram showing the main components of the entire Corona spacecraft:

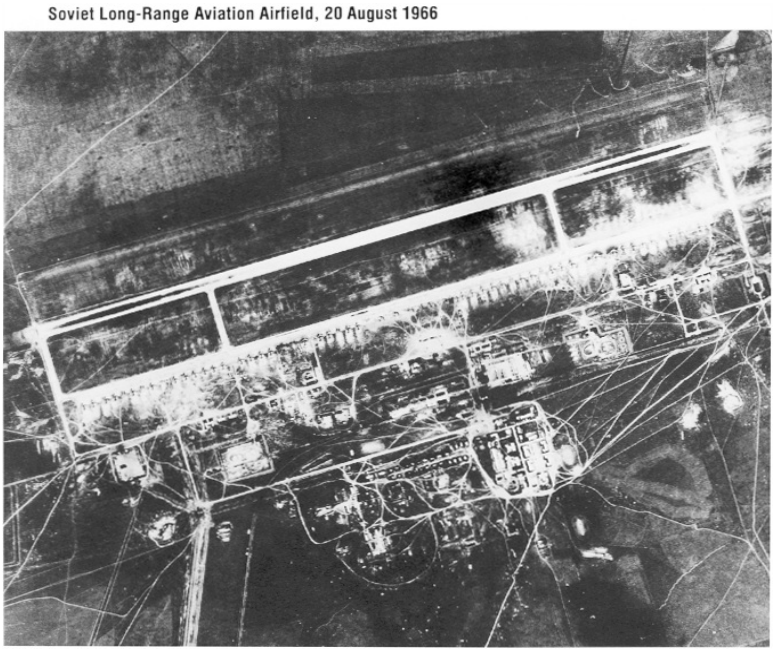


To illustrate the types of images obtained (as photos) from the Cornonas (that constitute the

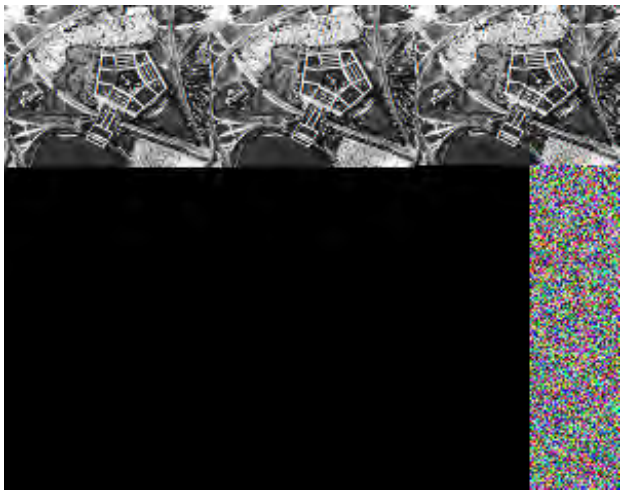
now declassified images released during the Clinton Administration), first examine this view of Moscow made by a KH-4 pass:



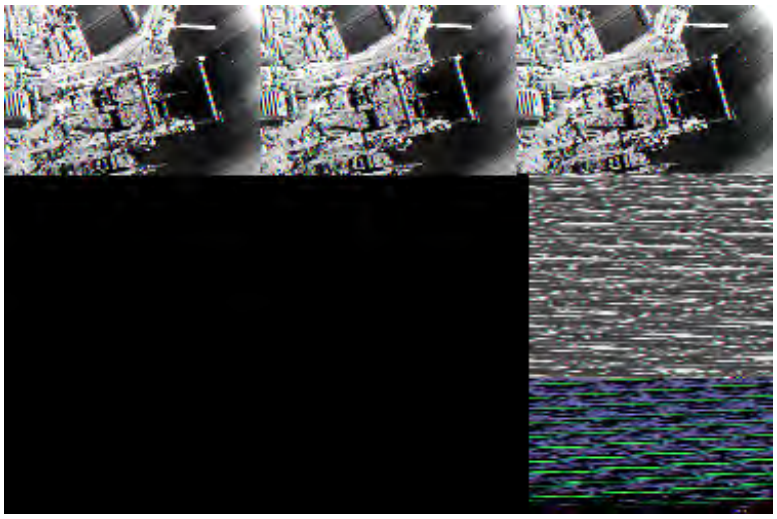
Next, examine this KH image (specific group unspecified) of a Soviet Airfield.



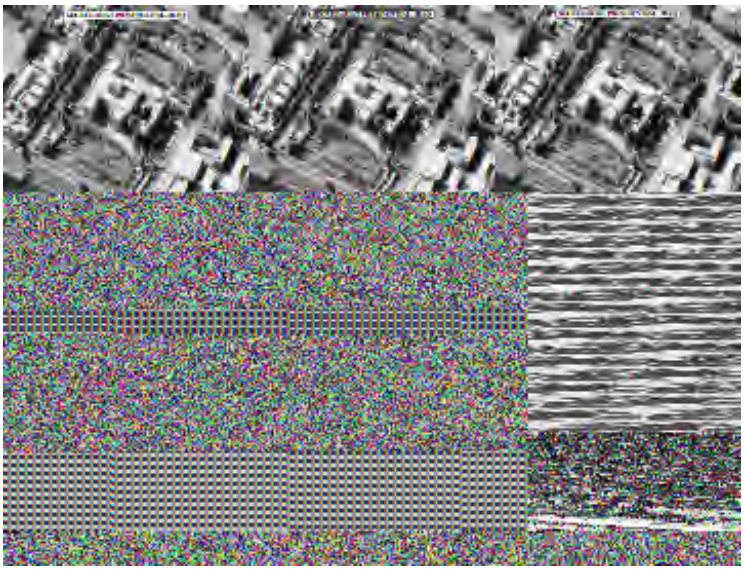
The U.S. has released a KH photo of the very center of its military complex - the Pentagon:



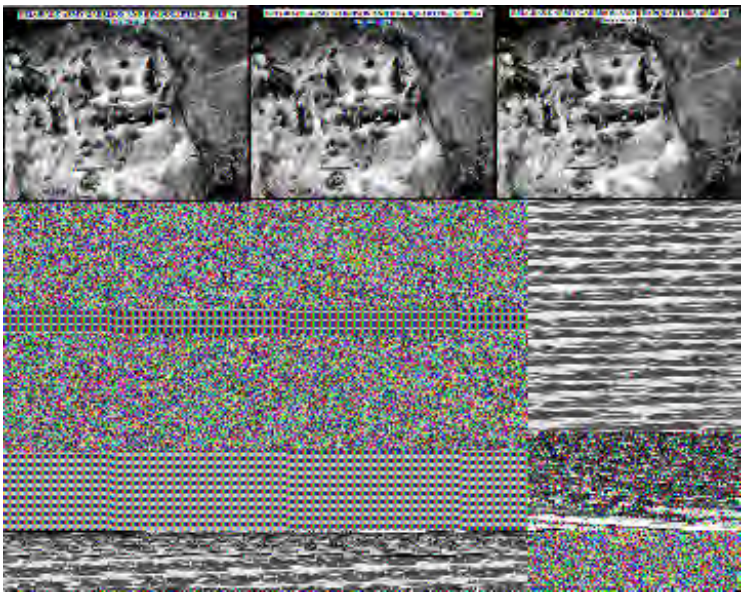
All of the KH satellites, of which more than 150 have been launched, consist of film cameras or electro-optical cameras that view the ground through telescopes. KH-7 and KH-9, the Gambit series, had resolutions of about 7 and 2.5 centimeters respectively. The KH-9 Hexagon satellites had 5 to 10 meter resolutions. It is not clear from any of the Internet sources consulted when 1) parachute drops ceased and 2) when electro-optical scanners replaced film. But, KH-11 is designated as an ELINT type (Electronic Intelligence) and did relay its imagery to receiving stations. The KH-11 Crystal satellites produce both SWIR and Thermal Infrared imagery, suggesting a non-photographic component. The KH-12 series, the last for which some specifications can be found, is reputed to achieve a resolution of 2+ cm, although images of this sharpness haven't been released. Below are several illustrations of declassified (but less than optimum resolution) KH-11 and KH-12 images:

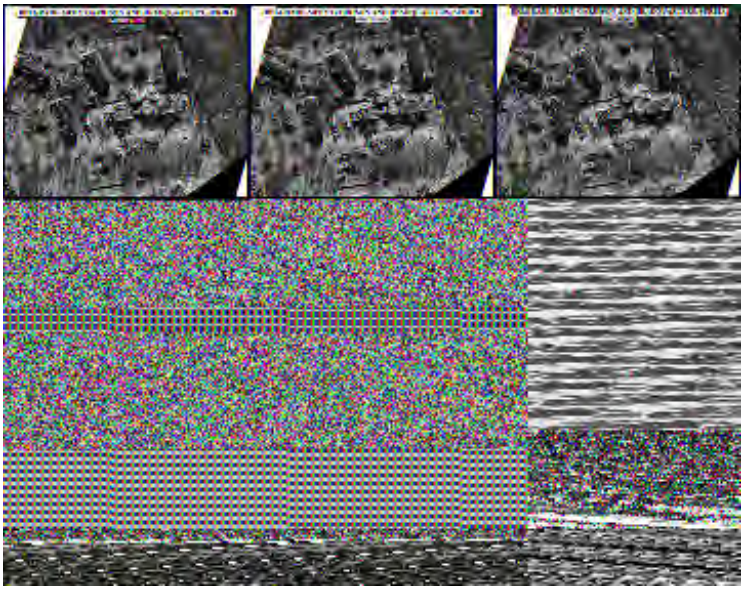


Here is one of Saddam Hussein's palaces, a potential target in any renewed Iraqi conflict.

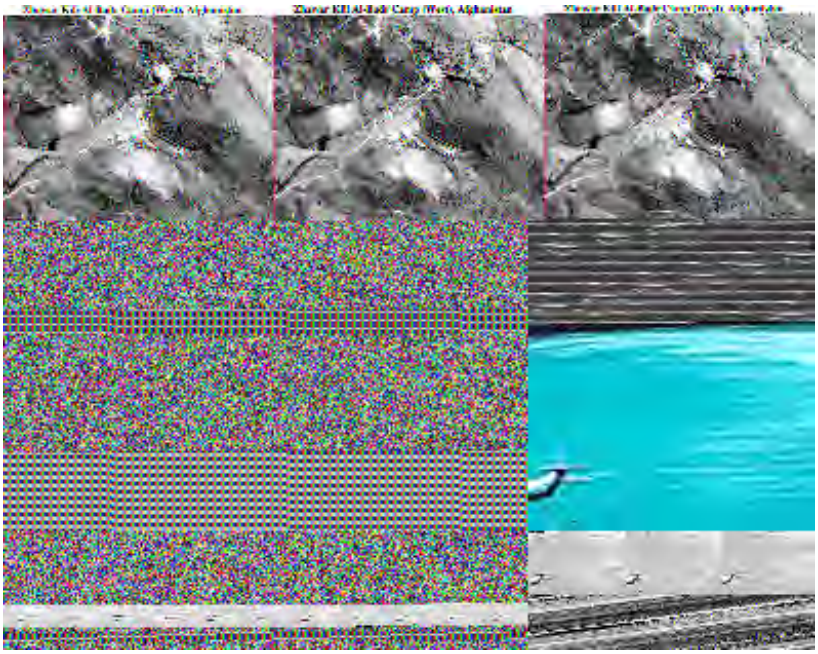


The next pair shows a military barracks in Serbia before and after an airstrike during the Bosnian conflict.



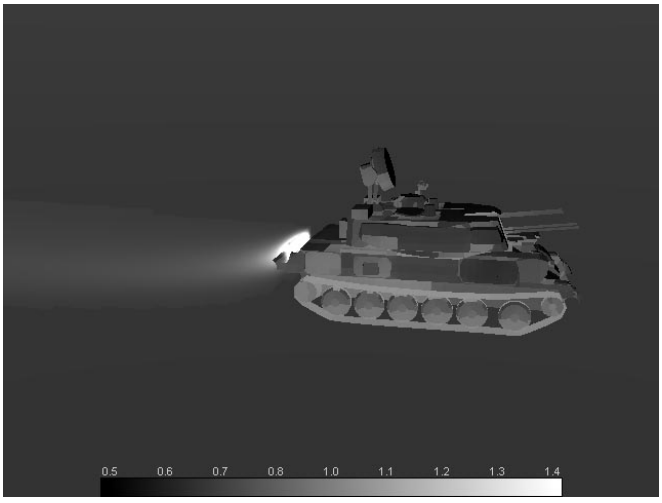


This next image was made by a KH-12 satellite using an improved Crystal sensor to image the Zawahr Kili Al-Bahr terrorist camp in western Afghanistan in 1998

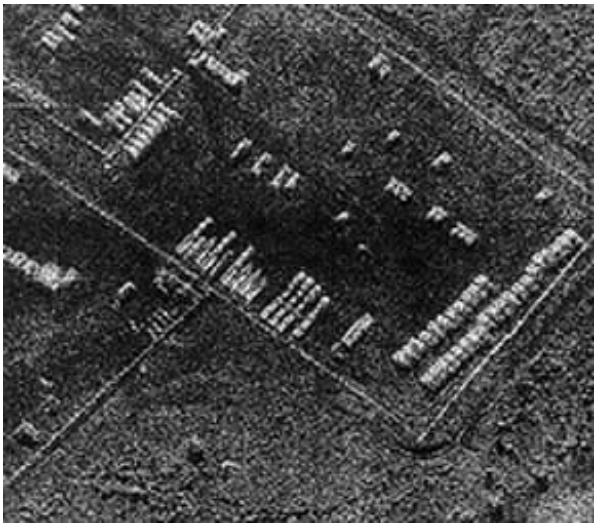


Examples of thermal infrared aerial and satellite imagery with specific military applications are scarce on the Internet. But here is one which appears to be a thermal IR image. It was taken during the Contra conflict in Nicauragua.

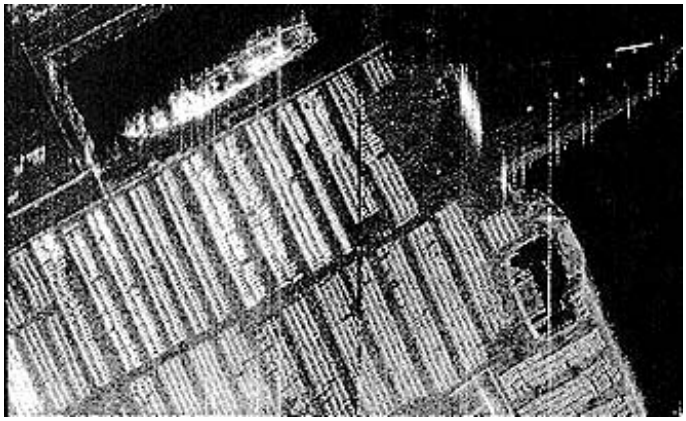
Thermal cameras used on the ground have numerous uses by the military. Look at these two thermal IR images, one of a tank, the other of a Hummer vehicle:



Radar has special military value because, using the right wavelengths, this active system can "see through" clouds and can operate at night. The U.S's Lacrosse series consists of a SAR sensor mounted on a very large (reputed to be schoolbus sized) platform that ties to extended solar arrays. No Lacross/Vega images were found from an Internet Search through more than 200 sites - this confirms the highly secretive and classified nature of this system. The first Lacrosse was orbited in 1988; Lacrosse-4 launched in 2000. The speculation is that this radar can achieve 1 meter or better resolution. The next image was made by an airborne SAR (TIER program) which simulates the Lacrosse products:

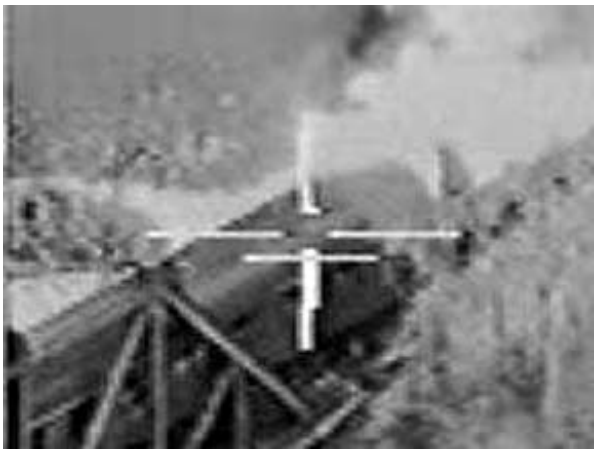


The TESAR (Tactical Enhanced SAR) program has produced these high resolution images: first of shipping crates in a military depot at a Baltimore, MD harbor; second, the most famed military building in the world - the Pentagon outside Washington. D.C.



Well before the civilian Quickbird and IKONOS satellites were made operational with high resolution sensors, military satellites had reached that stage and better. Claims are made of the ability to see people, read license plates, and even spot golf balls but no released images are that good (so the claims are speculative). But Hollywood has adopted this optimistic belief, with shows often pinpointing individuals in motion Example: The Fox network hit "24"; the Counterterrorism Unit seems able to call in a space sensor to dwell on buildings and cars and people in motion; maybe so, but since most satellites orbit at low altitudes and hence pass over a scene for only a few minute in about 120 minutes, this would suggest that the satellites are geostationary ones which, being far out, would have to have incredible optics to monitor the action. What is probably actually involved are unmanned drones, treated next.

Over these past 45 years, and well before, aerial reconnaissance and intelligence surveillance has relied heavily on aircraft (ranging from low flying slow single engine propeller planes to high flying fast jets). A U-2 and an SR-71 example were shown above. In recent years, the military has turned to another aerial method - the use of UAVs - Unmanned Aerial Vehicles - to conduct pre-programmed surveys of targets and personnel. Some vehicles are called "drones". One class of UAVs is the Predator series. Below are a photo of this large vehicle and then two examples of images obtained from so-called gun-mounts, where the camera or electronic sensor is placed at or near the front of the vehicle.



UAVs have become the workhorse for military reconnaissance. As such, they perform classic remote sensing functions. Their cameras operate in the visible and infrared (including thermal imaging). It is predicted that as of 2010 more UAVs will be purchased by the U.S. Defense Department than traditional aircraft. But the need for trained pilots will remain since these are usually the ones to operate the controls of the UAV during a mission.

UAVs are useful in non-military uses as well. Search and rescue comes to mind. A good summary of UAVs is offered at this Wikipedia web site.

The Soviet Union, as stated above, has operated its own network of spy satellites. Most of the images obtained have been declassified for sale on the open market. It has flown several different series, of which Kosmos (the USSR code name was Zenot) is the best known. Here is an example of once military-now civilian Soviet satellite photos/images showing the City Hall of New York:

The Soviet, and later Russian, spy satellites are, like the American one, quite advanced and loaded with powerful sensors. Here is a photo of the Kosmos 994 spacecraft, which obtains its internal electricity from a small nuclear power source:

Aside from the now declassified imagery discussed on page I-23 and elsewhere, much of the older photographs and scanner imagery obtained by the USSR/Russian reconnaissance satellite remains unavailable. Here is an exception: Part of the supersecret Area 51 outside of the Nevada Test Site (see page 6-8):

Other nations have followed the Superpowers to run their own surveillance programs. The Israelis are using their EROS 1 and 2 satellites parttime in military reconnaissance. Here is their image of Kandahar in Afghanistan:

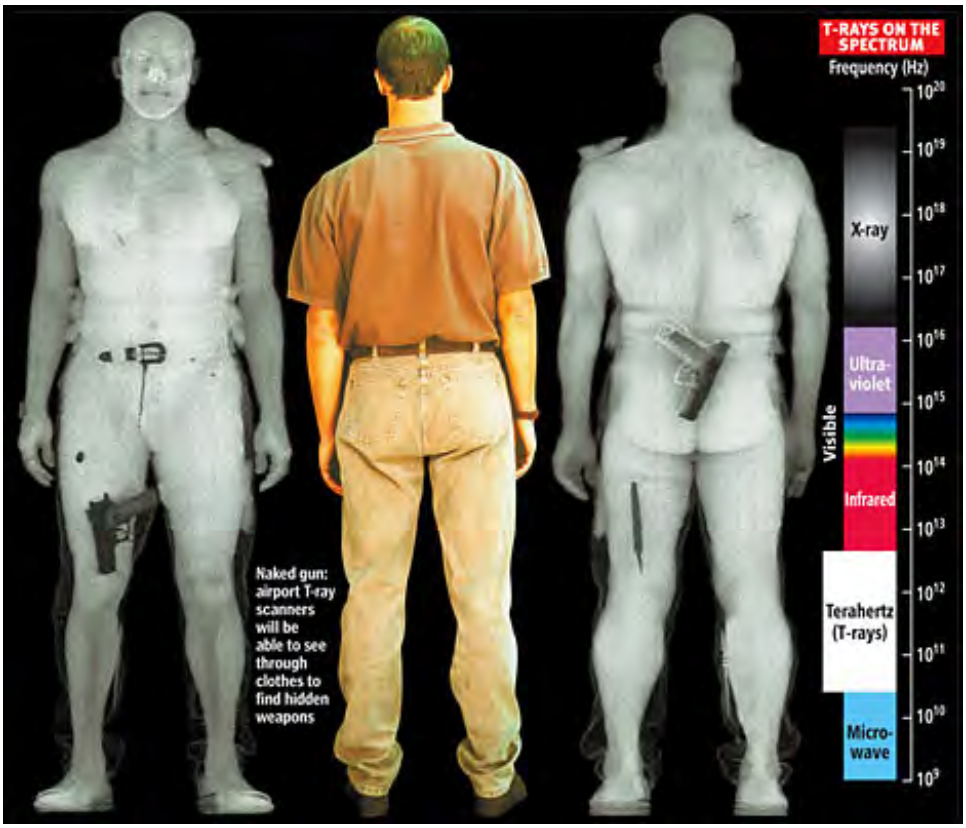
Now, intelligence imagery has come full circle. Among the first satellites launched were those used for military purposes. The experience with this technology was invaluable to NASA and other space agencies in developing their (until recently, lower resolution) earth-observers. But, with the Russian declassification followed by orbiting of civilian commercial satellites, e.g., IKONOS and Quickbird, high resolution imagery (1 to 4 meter range) has proved to have its military applications and is being purchased by many nations. The U.S. military and the CIA/NSA complex have contracted with these companies to obtain imagery, especially since the 9/11 terrorist attacks on the World Trade Center and the Pentagon. We illustrate with this one example of an Afghanistan Taliban Air Force base at Baghram, north of Kabul. This is an IKONOS product:

This availability of high resolution satellite images from unrestricted civilian sources has another important ramification: Images of possible military value can now be purchased by any country. Most nations cannot afford their own spy satellite programs but IKONOS, Quickbird and other private company products are affordable.

We turn next to a set of two examples of such imagery that have been gathered by Quickbird 2 (DigitalGlobe) in its now routine coverage of Baghdad, Iraq during 2002 and 2003 in anticipation of a possible invasion of that country by the U.S. military (and, probably other cooperating nations) as an initiative to topple the government of Saddam Hussein. Read the captions of each image for details.

Great concern for a while in June of 2006 was voiced by the U.S. and other major powers about a long-range (5000 km; 3000 miles) rocket on the launch pad of the Taepo-Dong facility in North Korea. Although a highly unlikely scenario, the launch could threaten U.S. territory, and the American military was alerted to be ready to intercept and destroy the missile if it seemed headed for a target deemed dangerous to U.S. interests. IKONOS obtained images of the launch facility:

Although it may seem "a bit of a stretch", the writer has chosen this page to mention 'full body scanning' - an application of remote sensing that many readers of this Tutorial have probably experienced. This is the scan done at airports, train stations, bus terminals, and other settings to search for bombs and hidden weapons that terrorists would choose to use to blow up the transportation carrier and its passengers - in a real sense a military action. The scanners are capable of seeing through clothes to image the naked body, and any metallic objects such as guns. Two scanner modes are now being used: one that uses X-rays and the other T-rays (in the near-microwave region, at Terahertz wavelengths). Here is an airport scanner and the kind of image it produces using T-rays:



So, the Question of the Day: Can anything on Earth really be private anymore?

Finally, for still more information on security-based surveillance from space, we suggest you check out the Global Security Organization web site.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The biggest advance in earth-monitoring remote sensing systems from the 1990s into the next Millenium has been the significant privatization or commercialization of the field as it continues to grow through both aerial and space observing systems. The principal new feature of these systems is high spatial resolution - comparable to earlier "spy" satellites. The germane point: remote sensing has become a business, not just an endeavor by governments to demonstrate the value of going into space to learn about the Earth.

The Commercialization of Space Remote Sensing

Because of the great costs involved, until the last decade most satellites that look at the Earth with remote sensors or make geophysical measurements have been conceived, designed, built, launched, and subsequently operated either by government agencies or have been supported (underwritten) in part by government funds. Since the early 1960s, of the approximately 2400 satellites in orbit - with many active (now operating), most are/were either military in origin or flown as a general service to a gradually developed user community by the U.S., France, Russia, India, Japan and other governments. Many satellites were directed towards scientific research or as monitoring systems that make environmental observations. In time some new satellites are/were developed by companies that have entered what can be called "Space Business".

Over the years, there have emerged industries that now operate in space (or plan to) to seek profits from a variety of activties besides remote sensing, e.g., communications (radio/television/phones), navigation (GPS, etc.), manufacturing (as on space stations), and even eventual tourism. Tourism, especially, has captured the imagination of the public. At present only a few millionaires have paid to take short trips into suborbital high altitudes. But in the lifetime of some now living it is plausible, even likely, that "ordinary citizens" will have the chance to reach orbit or perhaps venture to the Moon. These two illustrations embody these possibilities - no longer just the fiction of movies but a real likelihood:

On this page we will concentrate on what can be referred to as "the commercialization of space (imagery)". Remote sensing commercialization falls into two principal categories: 1) the sale of imagery or data from space satellites inaugurated by governments (national space programs, such as Landsat; this includes products made available to this industry from both U.S. and International space programs; see page I-23), and specialized services such as image interpretation or other processing (e.g., map making) related to the data; and 2) actual development of completely private satellites, their launching, data retrieval, and data (image) distribution, and their marketing. This second category is the subject of this page (thus, it omits discussion of any satellite program primarily sponsored by a government). The Earthsat Corporation is a good example of the first category; Space Imaging (GeoEye) of the second.

At the beginning commercial firms had been concerned principally with communications satellites. In that category are ComSat, IntelSat, PanAmSat, Inmarsat, and Eutelsat. For much of this earlier period, earth observations and resources satellites as well as most meteorological satellites - two groups relying on visual image gathering - have remained the province of governmental sponsorship and control. The shift from government programs for space platforms mounting remote sensors to programs devised and funded by private industry has been slow and only now, after 25 years since Landsat 1, has commercialization of earth-observing satellites begun to take off. This contrasts with one of the first privatized space programs developed first by NASA (aided by other federal agencies): namely, communications satellites, which with the incorporation of Comsat began in just a few years to assume a lead role in which development and capitalization from industry became the mainstay. Launch vehicles (i.e., rockets) also were built by commercial firms to put private satellites into space.

As we saw on page I-23, by the 1980s the success of the Landsats prompted other governments across the world to underwrite ventures into satellite remote sensing. Some of these branched into the private sector in time. The SPOT program developed by the French, with first launch in 1986, soon thereafter became the pervue of a commercial corporation, SPOT Image. The Indian government began its IRS program in 1988. The European Space Agency (ESA) actuated its ERS-1 and -2 operational satellites beginning in 1990 as did the Japanese Space Agency (NASDA), with JERS-1 the same year. The Canadians entered the picture with their Radarsat in 1991. In all of these the governments developed and launched the satellites but either created or arranged with private corporations for the operation of the data gathering and distribution system. Governments are still key players in recent and impending launches of an ever-increasing variety of space platforms and sensor systems.

In the early days, the principal users of civilian remote sensing data (aside from the military who have always been involved in reconnaissance systems) in the U.S. were government agencies, state agencies, and academia. With Landsat, industry came to recognize the basic value of the synoptic coverage afforded by space systems and started to purchase data in abundance. To handle dissemination to all kinds of users, the Dept. of the Interior through its U.S. Geological Survey established the EROS Data Center (EDC) in Sioux Falls, SD (see last page of this

Section). Its primary sales were a mix of Landsat data and aerial photography.

A number of companies offering value-added services, mainly to industry, emerged starting in the 1970s. One with whom the writer (NMS) has had repeated contacts is the Earthsat Corporation in Rockville, MD, which provides a wide range of capabilities including a full analysis of a given application problem. (On the Internet, you can find sites both under the Earthsat name, but also under MDA-Federal, since the company has been purchased by McDonald-Dettwiler Associates.) Other companies have concentrated on providing software for image processing and analysis to be performed at the customer's facility. ERDAS in Atlanta, GA. was one of the first in this field. Many other software programs have been developed since the 1970s, some by commercial firms and others by universities (e.g., IDRISI, used in this Tutorial; developed in the Geography Department at Clark University). As GIS grew into a major means of handling geospatial data and solving problems related to siting, land use, environmental monitoring, etc. a few companies, such as ESRI (Environmental Systems Research Institute) and Arc/Info sprang up to offer services and since then many more have joined in providing powerful programs that convolve GIS operations with image processing tools.

Some indication of the breadth of services now available from commercial organizations can be garnered from at the lists and ads in the Green Book: Who's Who in the GeoTechnologies, published each year by EOM, Aurora, CO (http://www.eomonline.com).

The first push to privatize remote sensing is traceable to some key historical dates in the Landsat program. (This is treated in some detail in the Landsat Chronology put together by Ed Sheffner at the Ames Research Center [NASA], at: http://geo.arc.nasa.gov/sge/landsat/lpchron.html.) In 1979, NASA relinquished operation of its Landsats (Landsat-3 had been launched by then) to NOAA. At the start of the Reagan administration, the drive for commercialization was accelerated. During 1984, the Dept. of Commerce set into motion procedures for privatizing remote sensing in those areas under conditions that did not conflict with military interests. Out of this, an agreement was reached in 1985 for the newly incorporated Eosat in Landover, MD, formed by Hughes and RCA, to operate the Landsat program for 10 years while EROS continued in its role as the National Archive for all land images. However, in 1992 a government decision was made to return the development and launching of Landsat-7 to the Dept. of Defense in cooperation with NASA; DoD opted out of the program by 1996 and Landsat-7 was dovetailed into the EOS/MPTE program. By 1996, however, Landsat-7 had reverted to NASA to operate, with its data to be distributed by an EOSDIS facility built at the EROS Data Center to handle the distribution of the multi-terrabytes of data that will ensue from the MPTE program. It is noteworthy that Landsat-7 will be the first satellite ever to be programmed to gather and archive complete global land scene sets several times a year. The commercial entities' policies are generally to acquire only scenes ordered by customers.

To assure that possibilities for commercialization of remote sensing satellites could prosper and continue on a parallel course with this and other government programs, an important workshop entitled The New Millenium Program was held at JPL in 1996 to foster coordination of government and private industry in those aspects of remote sensing that rely on expensive spacecraft to gather data.

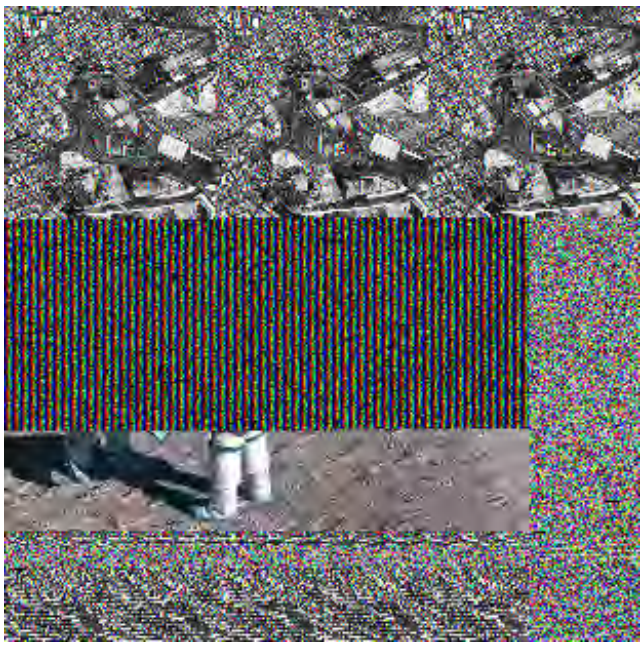
The inception of commercial satellite remote sensing probably began with the *Eosat* (now SpaceImaging) Corp.'s development in the 1980s as a business enterprise. Another early player in providing satellite remote sensing services was the *Earthsat* Corp., for years in Bethesda and now in Rockville, MD.

As commercial involvement took root, the demands for higher resolution imagery from the user community prompted these businesses to lobby for improvements in resolution - which at first conflicted with military policy. Heretofore, the only way to get such pictures was with aerial photography, which was restricted to regions in a country open to civilian access. (Landsat, SPOT, etc. had, however, opened up the entire globe to surveillance at lower resolutions.) In time, as we shall see, this improved resolution was permitted - allowing the industry to take off.

Today, high resolution imagery of the entire world (except in some systems the polar regions) is now available to all users; the tradeoff is that such images are generally limited to individual scenes that cover smaller areas (requiring more adjacent scenes to duplicate the regional coverage that Landsat attains, by piecing the images together as mosaics).

The big breakthrough in making satellite remote sensing products both highly desirable and readily available for general and commercial use was the result of the Russian program in the early 1990's to sell **high spatial resolution** data on the open market. Prior to that, U.S.-influenced military restrictions had placed a lower limit on allowable resolution so as to maintain the class of remote sensors that were known as "spy satellites". The Russians, in need of foreign capital, released many of the military camera photographs taken from space at resolutions of a few meters. In 1998, the Russian space program began to digitize their KR-1000 camera images, then placed these for sale to the entire world. This became known as the SPIN-2 program. Here is an example of a 2-meter image of central Bangkok, in Thailand:

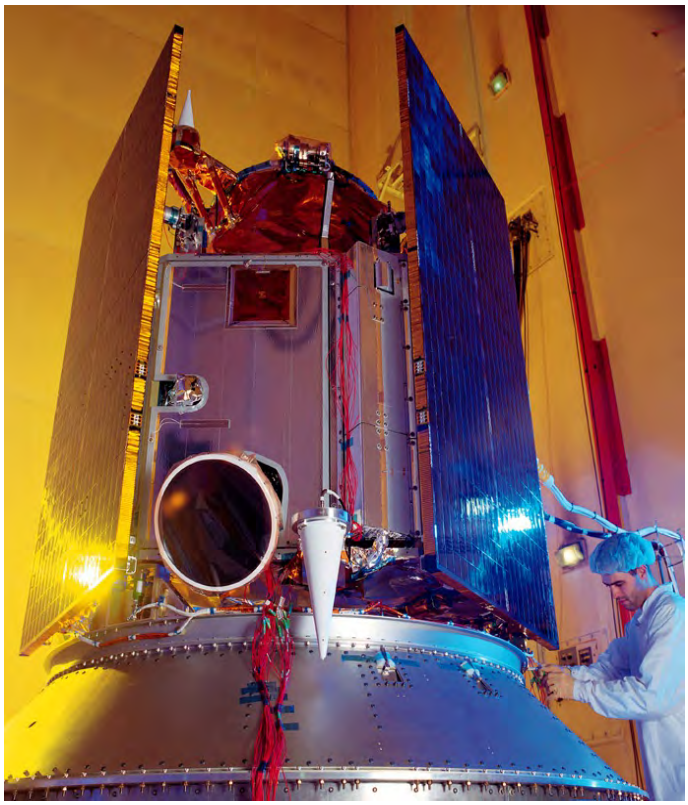
The Russian space program has remained the principal supplier of high resolution space photography, with pictures obtained onboard and returned in capsules, up to the present. (Its obvious disadvantage is the finite limits of film that can be packed onboard.) Among camera types used are the KVR-1000, KVR-3000, and KFA-1000. Two examples:



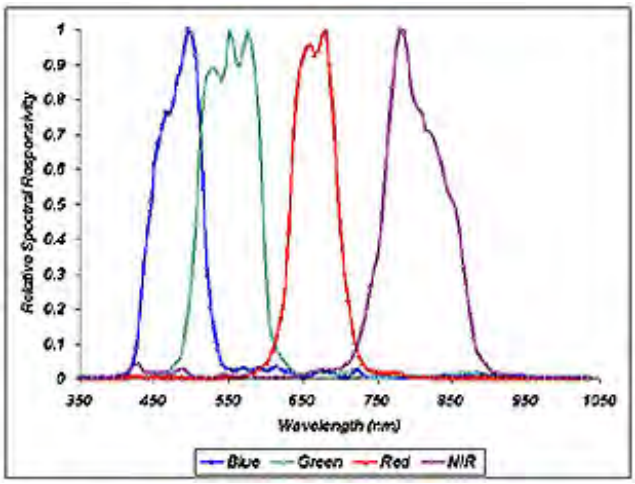
Soon after the Russian products appeared on the open market, encouraged by the anticipation that the United States would shortly waver its resolution restrictions various private corporations began to consider the possibilities in developing and operating high resolution satellites. These they would build but launch still required government-operated facilities - at first the U.S.(Vandenberg AFB in Calif.; Cape Canaveral in Florida), the French (in French Guiana along the northeast coast of S. America), and the Russians (in the northwestern Siberia and the southern steppes of central Asia). It was envisioned at this time that such satellites could compete with the aerial photography industry (although aerial photos can routinely provide pictures that were less than 0.3 meters in resolution). Thus, a combination of govenment-furnished space imagery and that supplied by private corporations - all with a range of resolutions and sensor systems - together moved earth observations into a new range of opportunities that have wide commercial implications.

Besides the U.S. Federal government, Canada, the Europeans, the Russians, the Indians, the Chinese, and the Japanese all have active remote sensing programs based on satellite observations; their products are available to anyone on a commercial basis. In the U.S., one finds mention of three companies that have launched remote sensing satellites, lately on commercial rockets: Space Imaging, Inc (company now named GeoEye), Digital Globe (formerly EarthWatch), Inc., and Orbital Sciences Corp. (before reorganization known as Orbital Imaging Corp. or just "Orbital"). GeoEye was formed in January 2006 when Virginia-based ORBIMAGE purchased the assets of Colorado-based Space Imaging; it is now the largest private commercial supplier of remote sensing products. GeoEye has more than 530 employees and is headquartered in Dulles, Virginia, with facilities in Denver (Thornton), Colorado; St. Louis, Missouri; Norman, Oklahoma; and Mission, Kansas. Its imagery sequence, in terms of satellites involved, has been IKONOS, OrbImage, GeoEye. All supply high resolution, multispectral products, as well as images taken with wide-angle viewing instruments that provide broad area coverage. We have seen several examples of these images in Sections of this Tutorial, including at the outset page 2 of the Overview. We now supplement these images with several examples that have been downloaded from the above Internet sites.

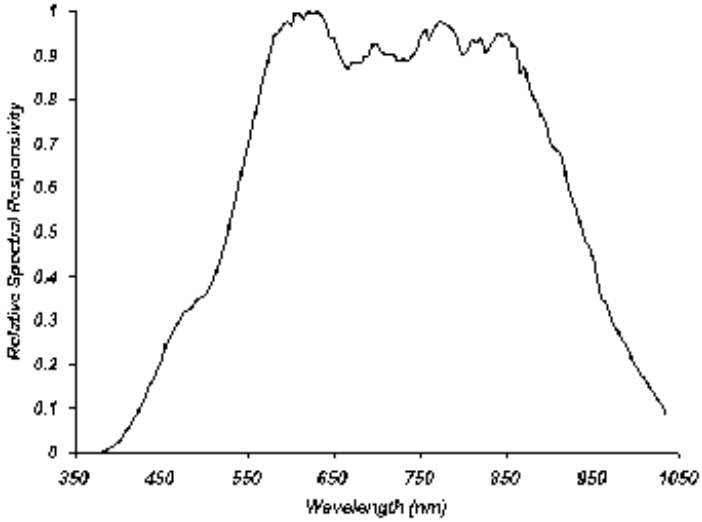
First, we look back at SpaceImaging's (headquartered in Thornton, Colorado) IKONOS-2; here is the satellite where it was built by Lockheed-Martin Missiles and Space Systems of Colorado:



The IKONOS satellites are multispectral, having approximately the same bands as the Landsat MSS plus a blue band, so that both natural and false color composites can be made:



These spectral band images provide 4 meter resolution. In addition, like SPOT, there is a very high resolution panchromatic sensor that produces a 1 meter black and white image. The spectral response curve for that sensor, and an example of its image product, are:



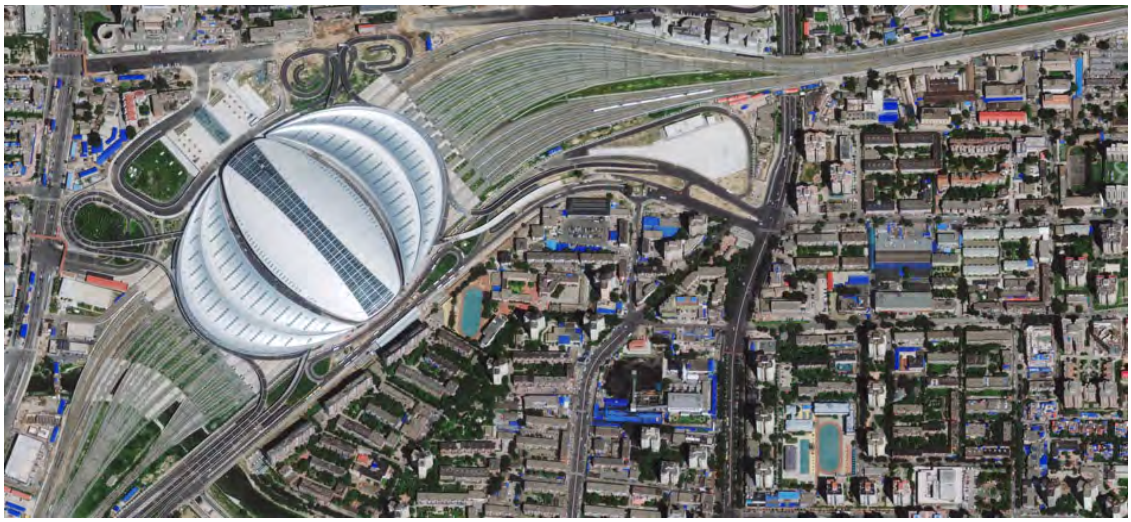
Here is an IKONOS high resolution panchromatic scene that shows the pyramids outside of Cairo, Egypt.

This example of high resolution IKONOS color composites shows a 1 meter image (involving registration of a pan image with the spectral band images) of Launch Pad 40, used mainly by the military at NASA's Cape Canaveral:

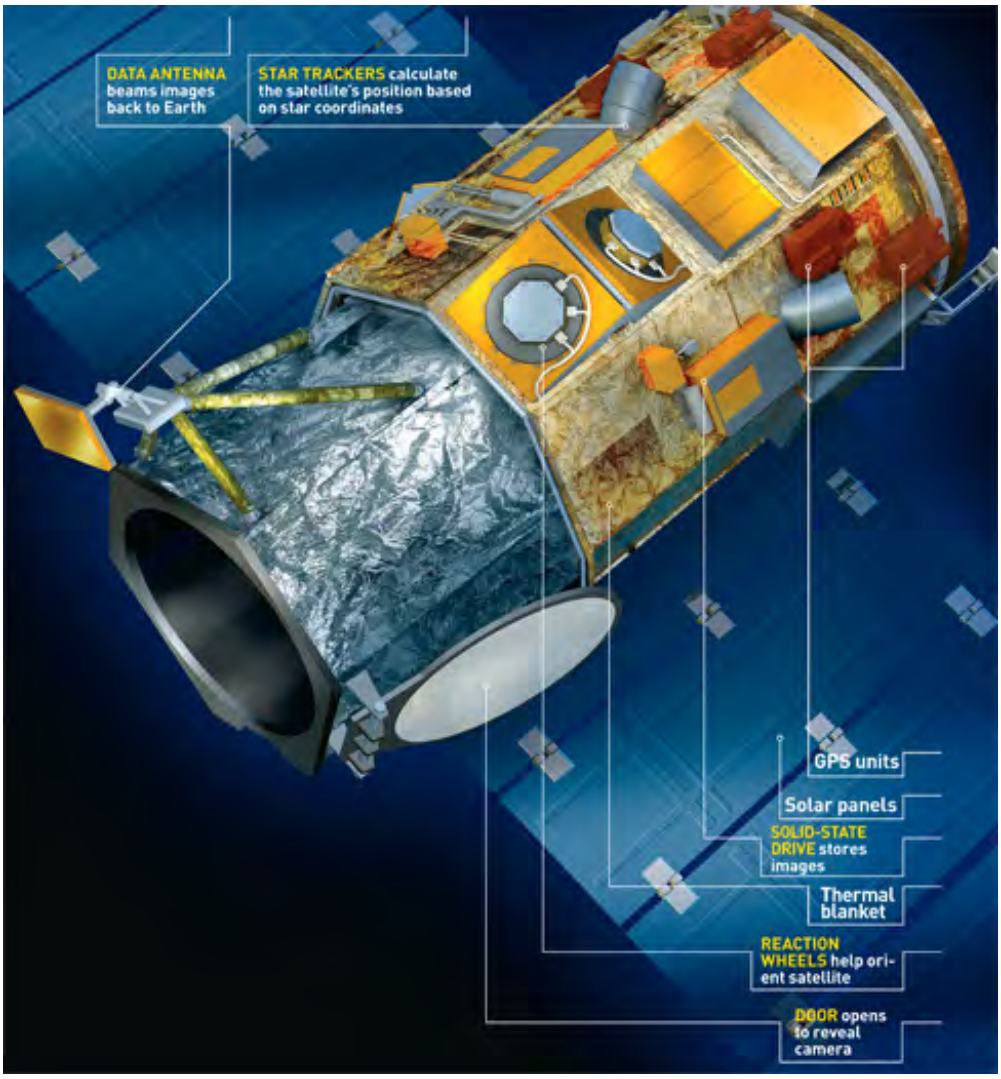
Here are two more, from Space Imaging (now GeoEye). At the top is another 1 meter color image showing a cruise ship at berth in the harbor of Rio de Janeiro:

The other image, at 4 meters, shows a dissected plateau near Kelan in the Shanxi Province of China; the peculiar curved markings were not identified but the writer thinks they may be some kind of agricultural plow patterns:

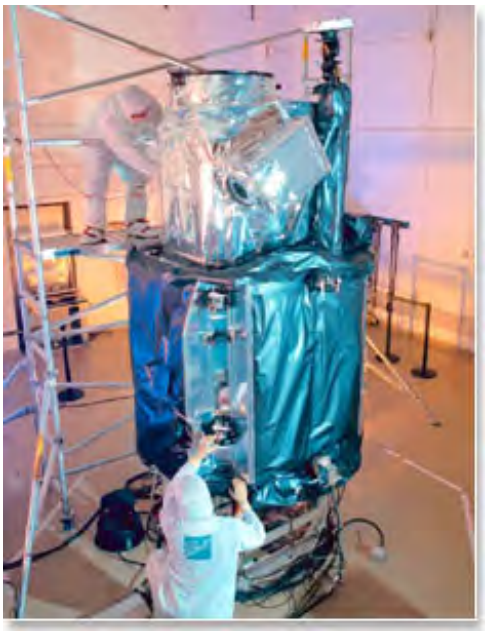
GeoEye images are distinctive - they rival good aerial photos. Here are two examples:



The GeoEye Corp. launched GeoEye-1 on Sept. 6, 2006. Its panchromatic camera produces 0.4 m imagery; its multispectral 4 band sensor produces 1.8 m images. Here is the GeoEye spacecraft and one of its images, namely of a Shuttle at the Kennedy Space Center:

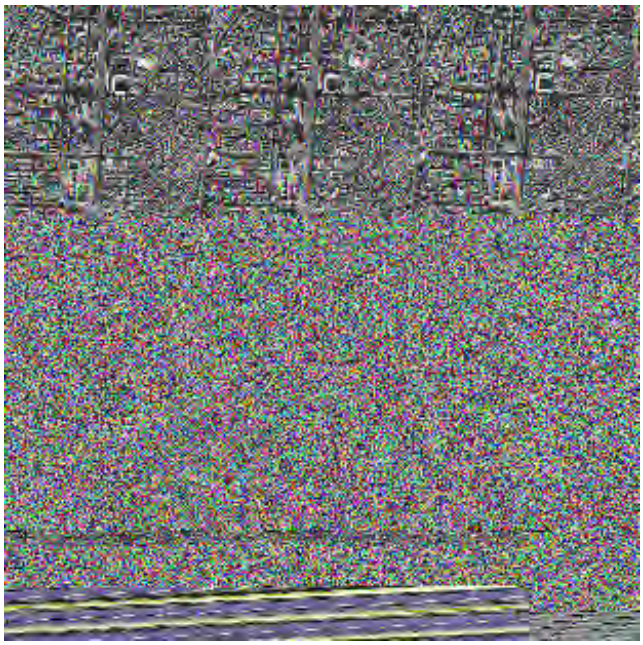


Quickbird-2 is the flagship of Digital Globe, a company located in Longmont, Colorado. Here is this spacecraft shown in its fabrication room at the builder, Ball Bros., also in Colorado:



The spacecraft orbits at 450 km. Its multiband sensors provide two levels of high resolution: 1) at 2.4 m, with red, green, blue and near IR bands; and 2) 0.64 m, with a panchromatic (wide wavelength) band. More details about these sensors are found on this Satellite Imaging, Inc. (a company allied with Digital Globe) site.

Here is a panchromatic black & white Quickbird image of Madrid, Spain:



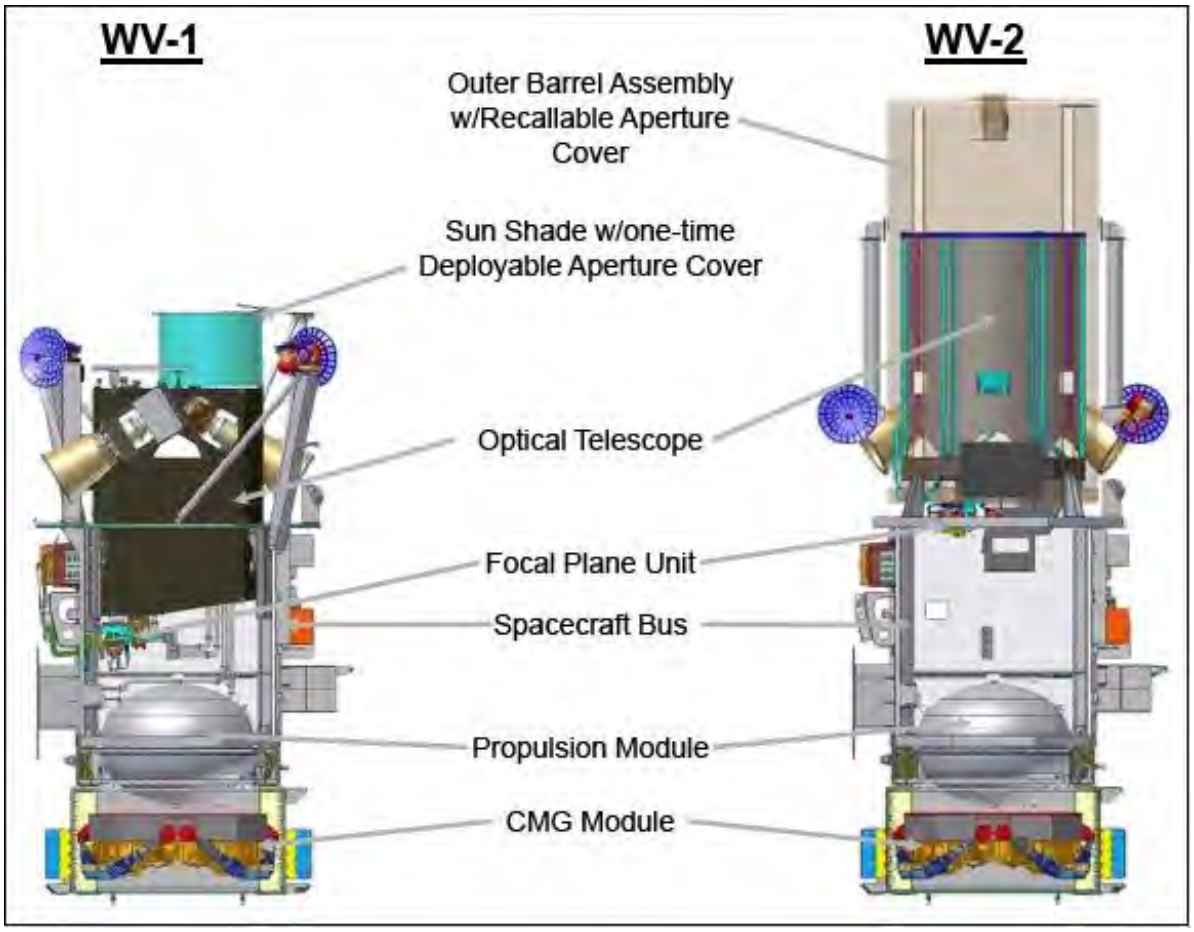
The next image shows part of Sunnyvale, California next to the Moffett Field airbase where NASA's Ames Research Center is located. This color image achieves a 0.64 m by using a registered panchromatic image.

The image illustrates one problem that can distort and image. Note that some buildings appear to be non-rectangular and one tall building seems to be leaning. For low altitude aerial photography this occurs as described in Section 10 but in this case the spacecraft sensor telescope was not pointed vertically downward, so that all buildings tend to be misshapen.

Here are three more Quickbird images, each looking at a noteworthy ground feature. The first pictures the largest community in the Antarctic, the McMurdo Sound Base managed by the United States. Below that is a view of the Pentagon (a year after it was damaged by a terrorist air assault) on the Virginia side of the Potomac River outside Washington, D.C. The bottom is a view of the Statue of Liberty in the New York Harbor at the mouth of the Hudson River, enlarged from the full Quickbird image.

Claims have been made (and appear to be true) that military satellites have high enough resolution of see a golf ball on a golf course. Commercial satellites don't yet do that well, but they show up the courses to advantage, as evident in this Quickbird image of a course near Kinsale, Ireland:

A follow-up to Quickbird is Digital Globe's WorldView satellites, which provide imagery at 0.46 m resolution. Here are details about the two spacecraft now in orbit:



This WorldView image shows part of Rome, Italy.

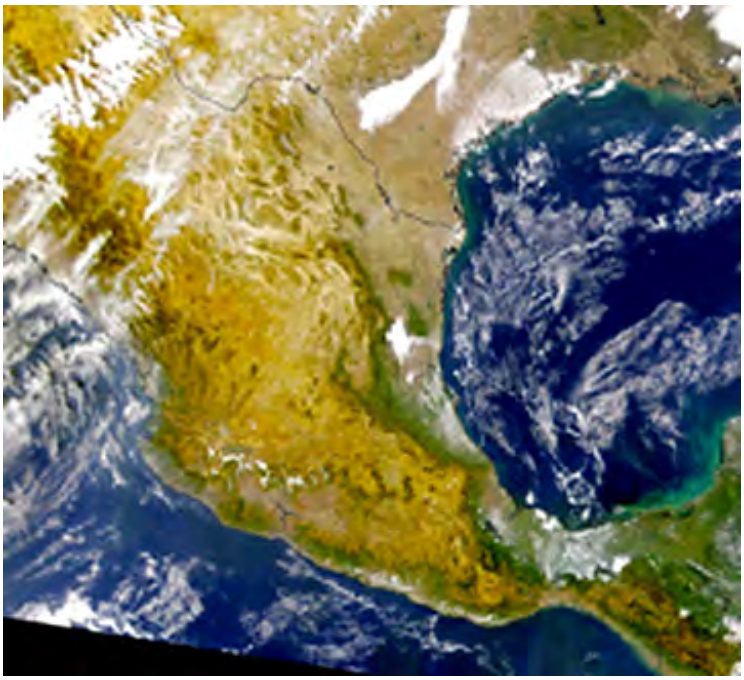
Both SpaceImaging (GeoEye) and DigitalGlobe really "came into their own" during the Iraqi war of 2003, since each provided exceptional images of all of Iraq both before and during the battles, as we have seen in the Overview. This provided exceptional publicity for the capabilities of their product.

There is also a market for synoptic imagery, i.e, at the other extreme, in which large tracts of land or sea, at subcontinental scales, are depicted in color but at low resolutions. OrbImage (its full name is Orbital Imaging Corp.), headquartered in Dulles, VA, is presently filling this need. This company claims to have been the first to launch a purely commercially-developed remote sensing satellite in April of 1995. Its OrbView-1, shown below, was designed to monitor lightning worldwide, and needed only a panchromatic sensor at low resolution:

OrbView-2, launched in August, 1997, operates from an altitude of 765 km, and produces images with 1 km resolution that extend over a swath width of up to 2800 km; revisit time is 1 day. Its

sensor has 8 channels ranging from 0.4 to 2.8 µm. Here is the "bird" before launch:

Two images exemplify the products sold by OrbImage. The first shows Mexico; the second the Himalayas and northern India.

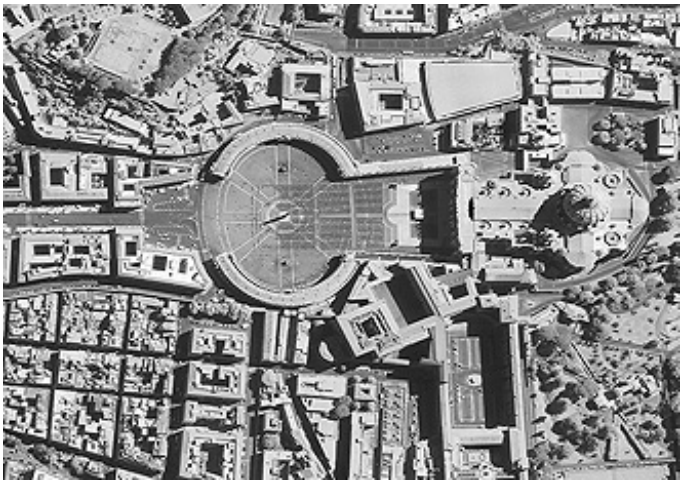


Reacting to the popularity of high resolution commercial imagery, OrbImage will launch OrbView-3 in the Spring of 2003 to a planned altitude of 470 km. Its 4-band (0.45 to 0.90 µm range) will provide images at 1 and 4 meter resolution, having dimensions of 8 km widths. OrbImage has also entered into partnership with MacDonald-Detwiler of Vancouver, BC to operate RadarSat-2 as a commercial venture providing radar imagery.

Israel decided it wished to enter the space imagery market in the 1980s. With some government "seed money", ImageSat International was incorporated. It launched its first satellite, EROS-1, on December 5, 2000:

The spacecraft carries a single sensor which produces a panchromatic (black & white) image at two resolutions, 3.3 meters (swath width of 12.5 km) and 1.8 m (swath width 6.5 km). Here is an EROS-1 image of central Stockholm, Sweden:

The idea behind this satellite was to provide high resolution black and white images. Because other satellites could provide multispectral images at almost the same resolution, ImageSat did not get their expected market share. A second EROS, slightly larger than the first, was put into orbit on April 25, 2006. Its panchromatic imagery has a spatial resolution of 0.7 m, making it one of the best systems available. Here is its view of the Vatican:



Since the market for space imagery keeps growing, other nations have entered the field, supplying timely high resolution images. These next two images (see their captions for identity) illustrate the types of products being supplied:

The future looks bright for commercial remote sensing (see Section 21 for additional

information). As more users purchase these products, the prices of individual scenes have dropped to under $2000. The ability to get the data as individual bands, so as to allow some image processing like that described in Section 1 and Appendix B, makes these scenes ever more attractive compared to the more conventional aerial photography that met the needs of wide areas of coverage at high resolutions. In time, more companies, from different nations, will emerge on their own (without government subsidies). The sensor systems will tend towards wider spectral range coverage, including thermal. In the foreseeable future, hyperspectral remote sensing - the "creme de la creme" of sensor capability - will become commonplace both from space and from aerial platforms.

The upshot of the review of government space program, military, and civilian (commercial) earth-observing remote sensors in space as examined on this and preceding pages is this: there are now so many systems actively looking down at the planet that it is likely the any spot on the surface can be imaged on demand on any given day - provided cloud cover is favorably low.

If you wish to learn more about the history and prospect for space commercialization of remote sensing, a tour of entries in your Search Engine will bring up 1000s of Web sites. Most are of limited value. Some tie commercialization with Space Law. The U.S. Congress has passed several enabling Acts. The Department of Commerce has part of the responsibility for satellite commercialization programs. Within NASA, the Stennis Space Center in Mississippi is managing many of the diverse space commercialization ventures. Some insight into how commercialization fits into satellite remote sensing can be gained from this Brief History of Operational Remote Sensing site. A review of difficulties as well as potentialities for commercial remote sensing is presented in this Rand Corp site (in .pdf format).

PREVIOUS rst NEXT

Primary Author: Nicholas M. Short, Sr.

Sensors mounted on satellites have proved to be the modern "workhorses" for gathering data about the Earth, the planets, and the stars. While it is usually cheaper and safer to send such unmanned systems into outer space, the role of humans - men and women - has often proved both cost-effective and often uniquely expedient when these folks can be put into orbiting capsules, transport vehicles, and ultimately space stations. There are many tasks that astronauts and cosmonauts can do quicker, better, and more efficiently than satellites. While the human brain can connect to the satellites via radio, and can be aided by computers, when on Earth, his/her performance in space is commonly the best use of brainpower for many tasks. Problems can be solved on the spot and if repairs of ailing but reachable satellites are needed these are often executed in the visiting space vehicle. The Americans in NASA programs were in the early days motivated to learn to operate in space by the challenge to go to the Moon. The Soviets, after dropping out of the Lunar "race", decided to concentrate their efforts on placing their cosmonauts in earth-orbiting space stations, using these both for military and for scientific applications. Since the 1970s the Russians have sent personnel up to the Salyut series of stations; this was followed by their Mir program. NASA in turn has put astronauts on Skylab and then in many Space Shuttle flights. With the new spirit of cooperation following the cold war, the U.S. and Russia, joined by other nations, have been putting intense efforts into building and manning the most advanced observational platform, the International Space Station, that will also serve as a space laboratory well into the 21st century.

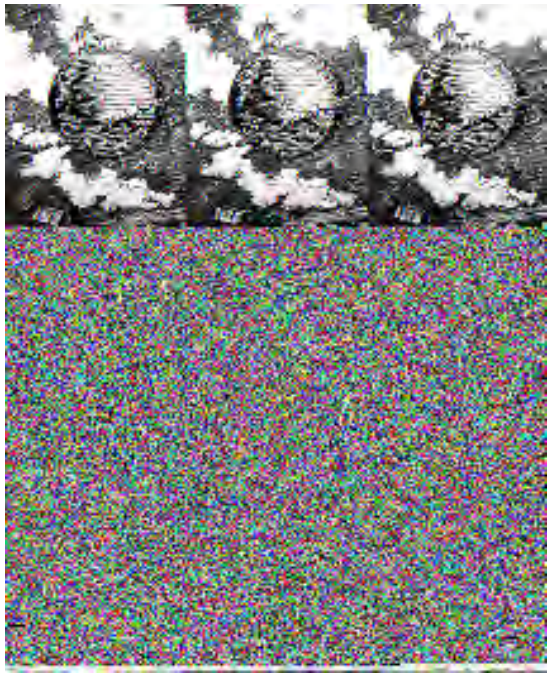
Humans in Space: Long Term Mobile and Fixed Stations.

So far, the emphasis in this Introductory Section has been on satellites that orbit the Earth (and in Sections 19 and 20, satellites that orbit or fly by other planets/moons or remain around the Earth as platform-mounted telescopes that peer beyond the Solar System). Yet experience has shown time and again that humans are often superior in performing certain remote sensing tasks (such as photography [Section 12]) at Earth from a variety of space platforms. Overall, humans have large and vital roles in conducting space experiments and other operations. The subject of "humans in space" or "manned space flight" is a vast one, but not of prime concern in this Tutorial. For those interested in its characteristics and history, consult this Wikipedia website. And for the curious, there is a Web site that lists all the flights of every kind of humans into space, visited at List of Human Space Flights.

This page summary will be constructed primarily from information from two Internet sites dealing with space stations: (1), (2). These numbers will be used below with individual images to tie them to their source site. You can log on to each site for more details.

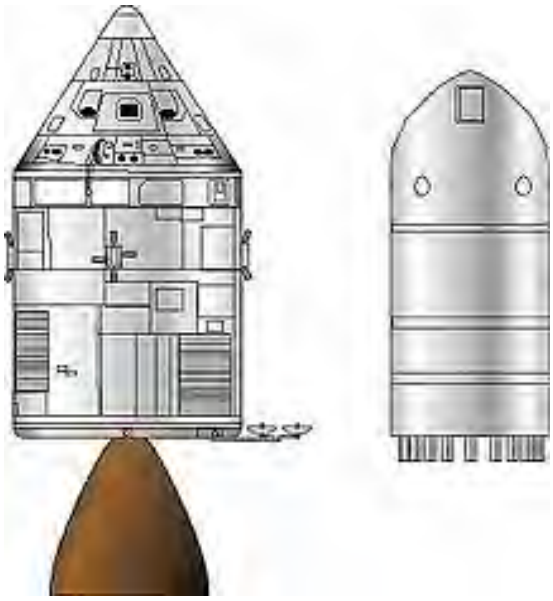
In the Overview we alluded to small space capsules or larger Space Transportation System (STS) vehicles (Shuttles) that are built to allow humans to spend short periods in orbit or in transit (as to the Moon) for days to a few weeks. Yet there is a distinct and valuable role for putting astronauts/cosmonauts into earth-orbiting space systems that are designed to stay in space for durations of months to years. These systems can be conveniently dubbed Space Stations - characterized by being larger, more versatile, and more fully equipped to harbor the human occupants safely for long periods, so that they can perform a variety of special tasks that take time. These include time-extended scientific experiments, manufacturing items in near-zero gravity, and repeated observations of Earth or astronomical phenomena. A Space Station can also be designed to serve as an intermediate stop and jumping-off point for space vehicles (manned or unmanned) that go to higher earth-orbits, the Moon, and the planets.

The notion of a manned orbiting platform around Earth actuallly occurs in a book published just two years after the Civil War. The famed writer, Edward Everett Hale (author of "The Man without a Country") conceived of a space station as a sphere made of bricks, shown below:



Even more imaginative is "From the Earth to the Moon" by Jules Verne (it appeared in French in 1865 and in English in 1873). The spaceship in the Verne book looked like this:

The engrossing story had three Frenchmen, one the builder of this bullet-shaped vehicle, fired from a 274 meter deep vertical circular shaft dug into a Florida site. The craft was launched by exploding guncotton. This propelled it to escape velocity (> 17,500 mph) to the Moon, which it orbited, and then returned (total trip time 243.5 hours) for a landing on Earth. A number of aspects of the trip were remarkably prescient (after all, Verne ranks as one of the top science fiction writers of all time), not least of which was the name Columbiad that he gave to the spacecraft. Compare his craft with the Apollo Columbia Lunar Service Vehicle that actually went to the Moon less than 110 years later:

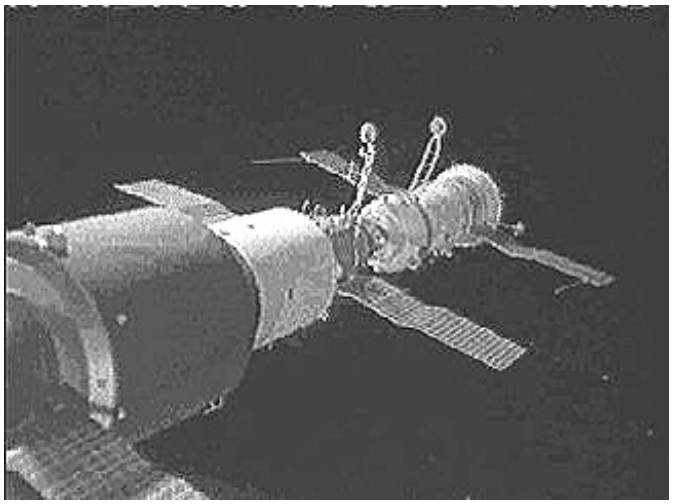


In 1928, the German Hermann Noordung first published his ideas on the feasibility of a manned platform in space. His colleague, the Romanian Hermann Oberth actually coined the term "space station" to descibe this concept

The famed rocketer, Wernher von Braun, may have been the first (in late 1945) to envision a believable configuration for a large space platform in space. It consisted of circular modules in a wheellike design and rotated to simulate gravity.

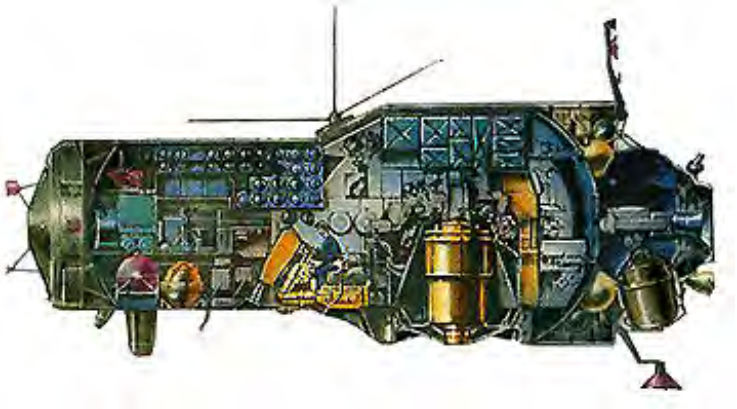
But the Space Age, which opened with Sputnik in 1958, had to get started before serious efforts were made to investigate the potentialities of establishing a station that would remain for long periods and be periodically to continually inhabited with humans performing useful tasks. In the 1960s, ideas were "floated" in both the Soviet and American space programs.

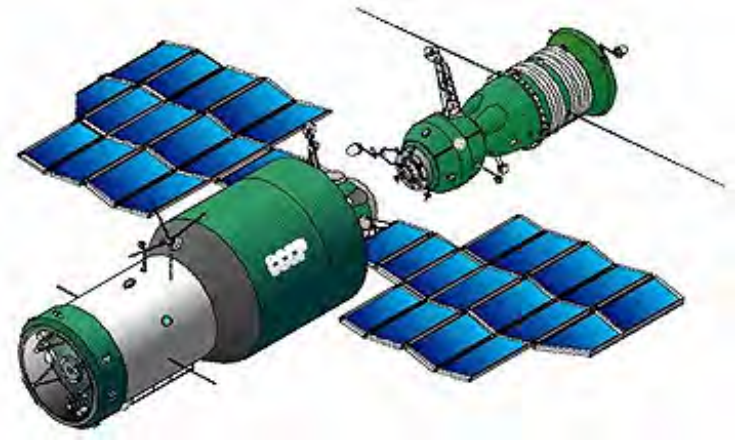
The Soviet space program focused early on establishing Space Stations. It began with the Salyut series. The first Station - Salyut 1 - was inserted into orbit on April 19, 1971. Here is a photo of this facility:



The first 3-man crew arrived a few days later after being launched by the workhorse Soyuz booster (shown below). But, something stuck and they couldn't get into it, requiring them to return to Earth. The second crew, with tools to open the Salyut, did enter and stayed for 16 days. But, tragically, upon return, a leak in their spacecraft sucked out all oxygen, so that they died enroute.

The Russians were developing another spacecraft to be used for military observations. This Almaz vehicle was a response to the U.S. Air Force's plans for manned observatories. The first Almaz was "disguised" as Salyut-2 to prevent other nations from awareness of the "spy" effort. This Salyut-2 was launched unmanned on April 3, 1973 but was never visited because a fire broke out onboard, so damaging the ship that it was uninhabitable and was decelerated into the atmosphere on May 28, 1973. The Almaz space station as it was being build appears below. Beneath it is a cutaway that shows its sophistication, and its deployment in space with its transfer vehicle, Soyuz, is at the bottom:



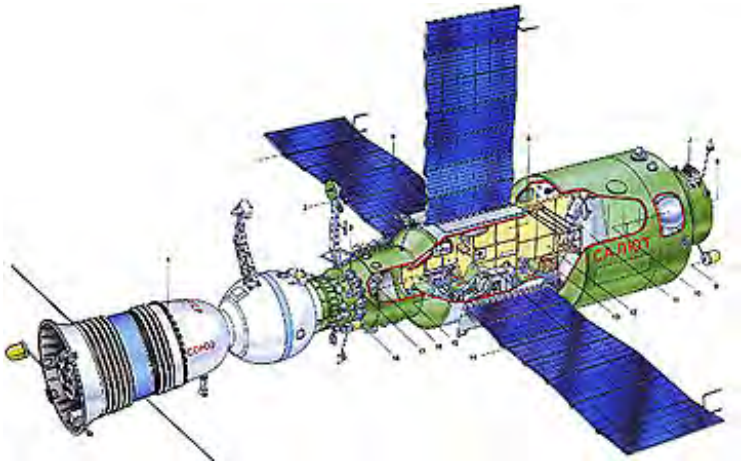


Another aborted military Space Station, Cosmos 557, launched on May 11, 1973, was never visited.

Although the Soviet space program had an early lead in operating Space Stations, the U.S. program made a "giant leap" with its Skylab station by putting up a facility that had a number of scientific instruments and experiments. Launched on May 14, 1973 on a Saturn 14-B rocket (page Intro-7), Skylab successfully achieved orbit but suffered two serious setbacks: 1) it lost one of its solar panels as that was deployed; and 2) part of its outer hull was ripped off. Nevertheless, the first crew (which arrived in a separate vehicle) of three astronauts was able to improvise a protective covering made of gold-surfaced sheeting. This Skylab-2 mission was followed by Skylab-3 and Skylab-4, leading to a combined occupancy time of 73 days by the three crews, each able to conduct most experiments. Several EVAs (Extra-Vehicular-Activity) or spacewalks occurred from Skylab. (The very first EVA was made by Aleksei Leonov on March 18, 1965 from a Voshkod-2 spacecraft; the first astronaut to EVA was Edward White who exited Gemini-4 on June 3, 1965.) Skylab was functional until February of 1977 after which it was decelerated to burn up over the Pacific Ocean. Here are views of Skylab in orbit and a cutaway

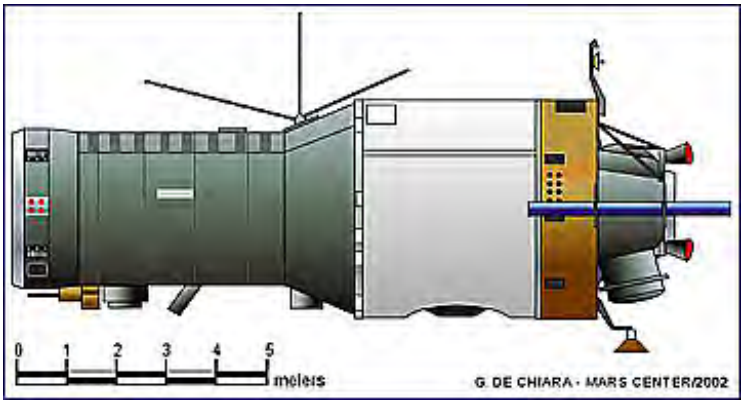
which shows some of the crew cabin:

Still in the competition, the Soviet program launched another military station, Salyut-4, on December 26, 1974. Six visits, totaling 93 days, continued until February, 1977.

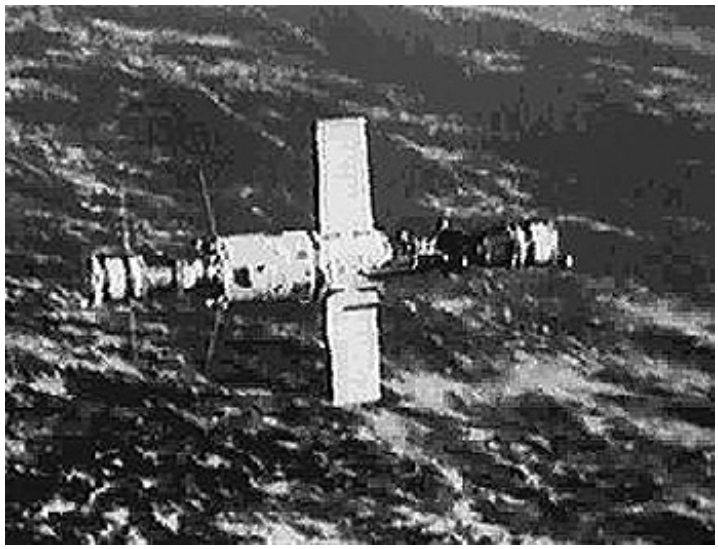


A Space Station of sorts was created for a short time when an American Apollo Command Module spacecraft (once destined to be the Apollo 18 mission vehicle before the Moon landing program was terminated) docked with a Russian Soyuz craft on July 17, 1975. Ostensibly, a show of "comradeship" in space, a practical purpose was to determine the ability (and possible problems) of a spacecraft from either of these countries to come to the aid of the other in case of emergency and need for rescue. Here is an artist's sketch of the two ships joined and beneath it are the five spacemen who shared their spacecraft for a brief time:

Salyut-5 (Almaz-3) was the last Soviet strictly military Space Station venture. It lasted from June 26, 1976 to August 8, 1977 and was visited by six crew missions.



With Salyut-6, the Station remained viable for 4 years, 306 days (Sept.'77-July'82) During that interval there were 35 trips (cumulative 1 year, 170 days) by 2-cosmonaut crews. The Station was capable of docking two spacecraft.

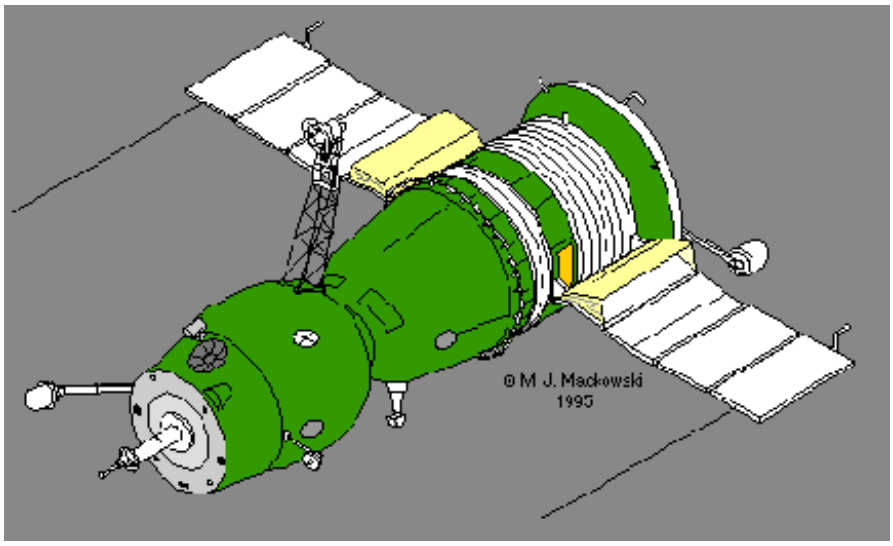


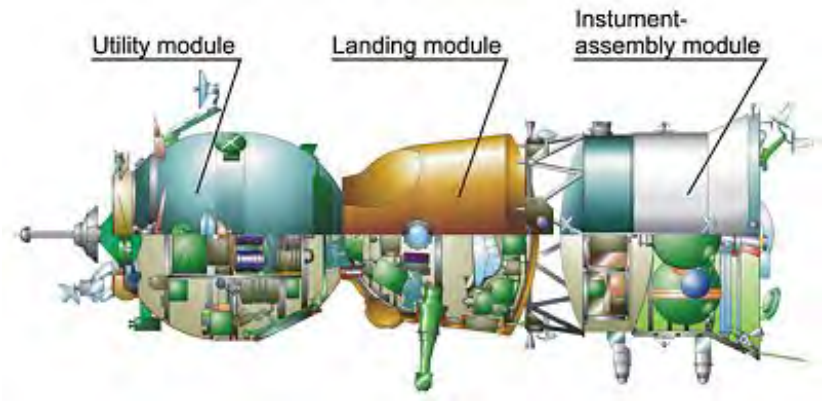
The last and longest (8 years, 268 days; April 19, 1982 to February 7, 1991) of the Salyut series was Salyut-7. It hosted 29 visits (although 2 crews failed to dock and enter) including one that reached it from the new MIR station (see next page).

The reader may be wondering by now just how the astronauts and cosmonauts traveled and and from the various Space Stations. Two space vehicles provided the vast majority of transportation systems (although two new ones are being used for the International Space Station ISS)(next page).

The longest transfer spacecraft still active is the Soviet's Soyuz vehicle, now operating in more advancing versions after its first launch in 1969. The Soyuz was first developed to compete with the Apollo system when in the '60s the Soviet space program still considered itself competively in the race to the Moon. Soyuz has proved itself so reliable and versatile that it has served many times as the transfer ship for visits to the Salyuts, MIR, and the ISS. "Soyuz" can refer both to the launch vehicle and the transfer module. Here is this pairing during a launch.

The first Soyuz, developed starting 1966 with launch in 1969, is shown below. This basic configuration has remained essentially as shown to the present. But there have been four variants, each improved over its predecessor and given new modifications. The sequence: Soyuz, Soyuz-T, Soyuz-TM, and Soyuz-TMA. A cutaway shows the general interior of the three component modules of the latest, the TMA.



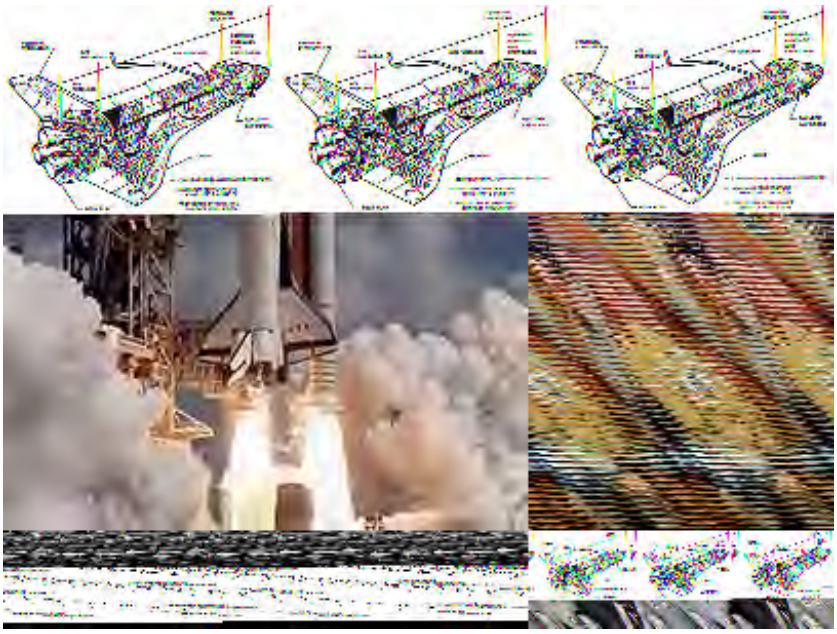


The last pair of photos show the Soyuz transfer vehicle approaching the ISS and docked with the ISS. The Soyuz, originally designed for two passengers, now holds 3 comfortably and four cramped inside if the Soyuz serves as an escape vehicle in an emergency. The Soyuz Landing Module returns to Earth as a capsule (separated from the other components) supported by parachutes that touches down on land at one of the space port facilities the Russians manage

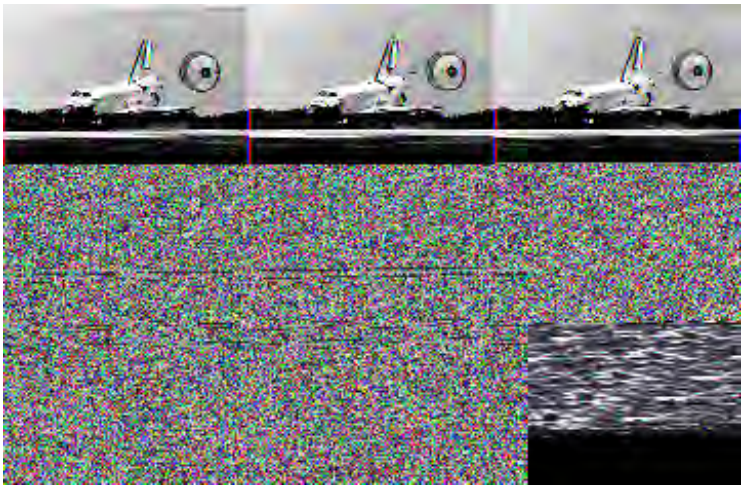
or "rent" (Baikinur, in Kazakhstan; its replacement, Svobodny; the Northern Cosmodrome; Yepatorlyas in the Ukraine).

For its first 20+ years, the U.S. space program depended on returning its astronauts to safe landings in the Pacific Ocean (or other marine waters if needed) by splashdown. By the 1970s, the need for a capability of touching down on land and being able to reuse the spacecraft led to design of a "space plane" known popularly as the Space Shuttle (technically, they are called Space Transport Systems [STS]). The first flight of the Shuttle was a launch on April 12, 1981 from Kennedy Space Center and a landing at Edwards Air Force Base in California.

The Space Shuttle is a complicated vehicle, and we will not treat it in detail here. Make note in the above photo of the large (orange-brown) liquid fuel tank and the two solid rocket boosters that are external to the Shuttle (which does have its own engines for use at the end of launch and for maneuvering in space and during the landing cycle. Here is a cutaway diagram that indicates the cabin area, the cargo bay, and tail assembly. Many Shuttles had a large crane for removing cargo from the bay.



While launch and insertion into orbit are usually the most precarious, but thrilling, phase of a Shuttle mission, landings have been almost routine. The prime site of choice is back at Kennedy Space Center. Here is a Shuttle about to land at KSC; a drove parachute helps to slow its landing speed which still exceeds 100 mph.



The next photo shows the Shuttle as it landed at Edwards Air Force Base (the first 9 missions used the natural playa surface in the Mojave Desert of California as a landing strip; most later Shuttles landed at Kennedy Space Center, but as of now 50 have used Edwards AFB, usually because of bad weather at Cape Canaveral; there is a third site - not used sofar - in New Mexico.

Since launch at Vandenberg AFB on the California coast was less practical than at Kennedy, those Shuttles that had to use Edwards AFB have to get back to Florida. This is accomplished by lifting the Shuttle onto fittings atop a 747 airliner. This photo gives a good impression of the size of the Shuttle:

There have been 119 flights of 5 Shuttles: Atlantis, Challenger, Columbia, Discovery, and Endeavor. (Enterprise was flown once as a test mission.) Over the years, two were destroyed while in flight - Challenger during launch in Florida; Columbia over Texas during final approach to landing. Information on individual flights appears as a List of Space Shuttle Missions on the Web.

The Shuttle has performed a variety of missions during its usual 6 to 14 day stay in space. One is to launch spacecraft into final orbit; the best known cargo was the Hubble Space Telescope. Also conducted are scientific experiments and medical investigations into the human response to weightlessness and other body hazards. EVAs have allowed servicing of satellites in reachable orbits. For much of the last 7 years the Shuttle has carried some of the structural components of the International Space Station (that effort has been suspended following the Columbia accident in 2003 but after major improvements in safety features the flights Discovery completed a resumption mission in August 2005 but again problems with debris have grounded the fleet until Spring of 2006.

The Soviet/Russian space program has its own Shuttle, Buran, which is essentially a copy of the American vehicle. It has flown only once (unmanned) in 1988 but was dropped from their space program because of signs of unreliability. Here is Buran in its landing mode.

Both the Shuttle and Soyuz have been key transport vehicles that make round trips to MIR and ISS - the two Space Stations discussed on the next page.

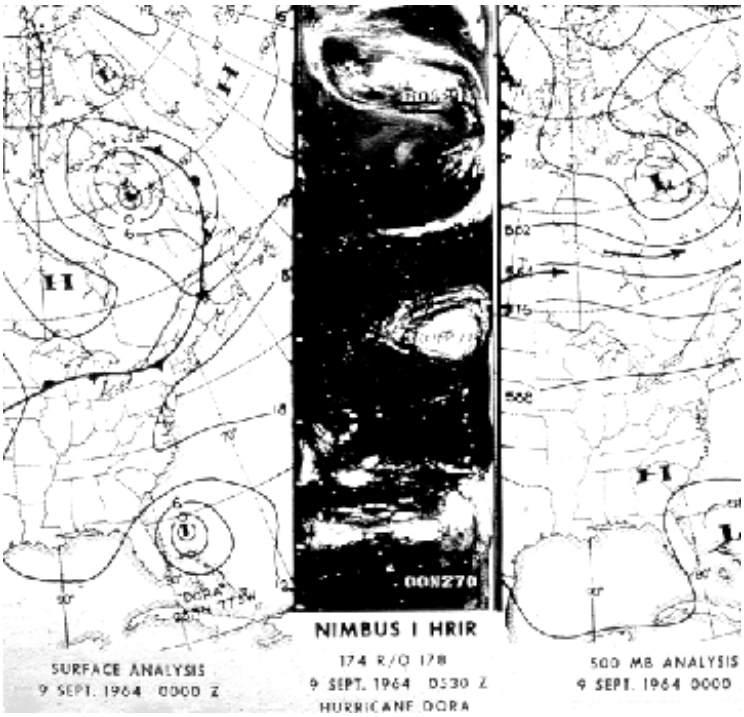
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Satellite-and Shuttle-based remote sensors are especially adept at providing image and physical property data regarding atmospheric and ocean surface conditions, usually scanned from meteorological/oceanographic platforms. Example from two of these (GOES and NOAA) are described on this page but the main treatment is reserved for Section 14. Metsats look mainly at cloud cover, water vapor, wind patterns, advancing fronts, air temperatures, and certain other atmospheric properties. SeaWiFS is a dedicated oceanographic satellite whose sensor measures in multibands (visible and near IR) that specify ocean color and chlorophyll content (in algae and plankton). To these specialized satellites should now be added a third group - a large array of satellites just launched or to be launched in the next 5 years that gather integrated information on the Earth System - land, marine, atmosphere, and biological aspects of the terrestrial environments. Although not the first directed towards this endeavor, Terra and Aqua are now operating and returning data that prove the value of this approach.

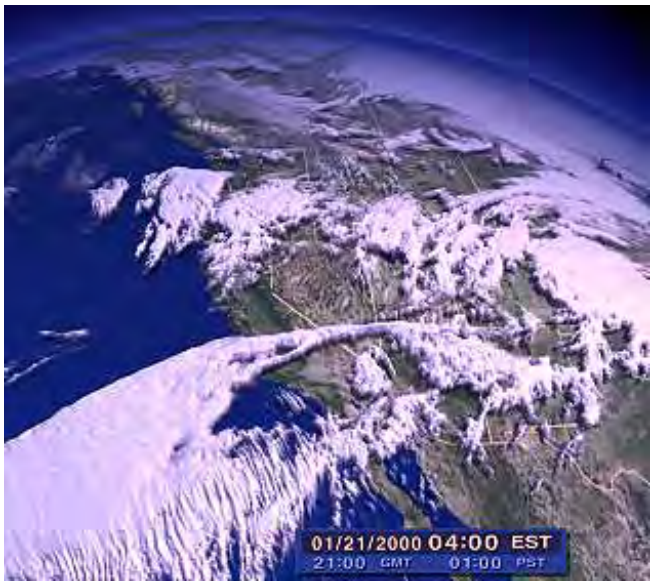
Other Remote Sensing Systems: Meteorological, Oceanographic and Earth System Satellites

As suggested earlier in this Introduction, among the first satellites in the U.S.'s entry into space were those designed to demonstrate that weather systems and climate variations could be monitored at regional or even continental scales, thus greatly improving the realtime surveillance of clouds, temperature variations, water vapor, and moving fronts (especially tornadic vortices and hurricanes). The TIROS and Nimbus series have already been mentioned (page I-7), but the example images there emphasize land features rather than meteorological conditions. Here are two 1960s images that are typical of the practical weather monitoring and forecasting that began to impact meteorology - and with eventual appearance on local TV, the very way in which people now draw upon satellite remote sensing to keep abreast of weather conditions for the moment and for ahead at least a week.

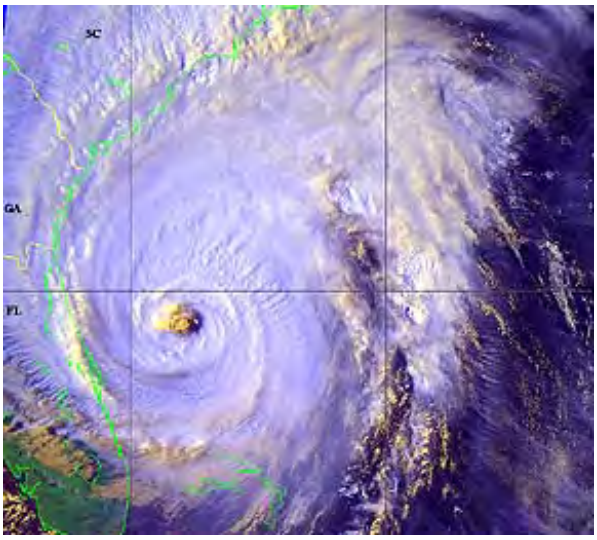


In Section 14, we will review the entire history of the "Meteorology from Space" programs that include satellites operated by several other countries. Suffice here to show three typical examples produced by more recent American metsats.

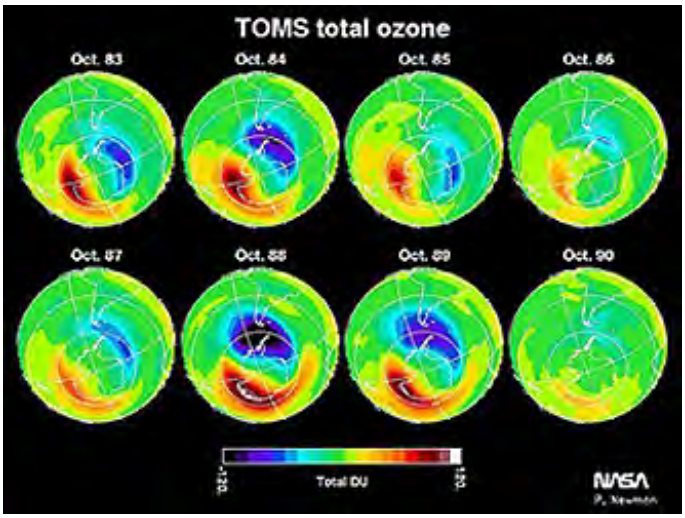
The first is a January 21, 2000 scene covering part of the western U.S. and adjoining Pacific Ocean as imaged many times each day by the GOES 10 (geostationary) satellite. The land tones are darkened so that the cloud patterns stand out.



Hurricane detection and monitoring as the storm progresses have been one of the triumphs of satellite remote sensing. The second scene was made by the NOAA 15 satellite that was the principal monitoring system that followed the westward progress of the powerful Hurricane Floyd which struck the U.S. mainland in mid-September of 1999. This color composite shows Floyd in the early morning of September 15 after it had passed over the northern Bahamas and was bearing down on the north Florida coast. The size of this hurricane can be grasped by noting that the bottom left of the image covers the Everglades (in green) whereas the top left includes all of the Georgia coast into South Carolina.

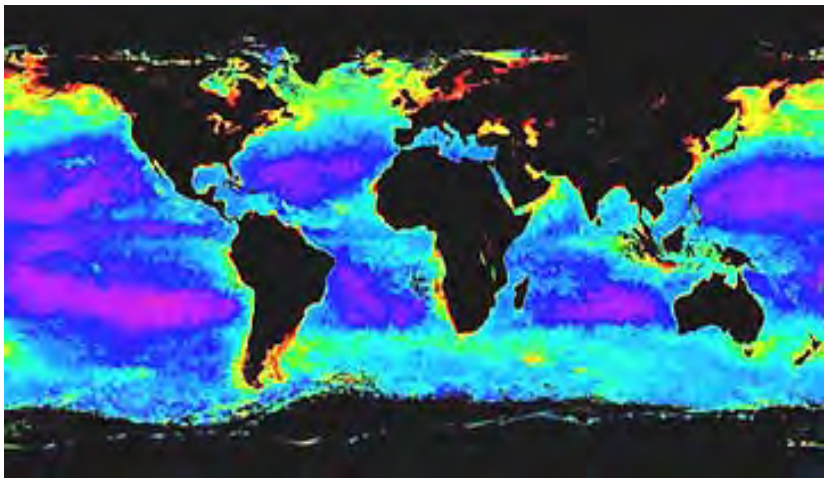


Aside from weather phenomena such as severe storms, a prime application of satellite observation has been to monitor over time deletorious effects of human activity in the atmosphere. One serious threat is from ozone accumulation (ozone - a form of oxygen - is released from aerosol cans and from combustion engines). Here are the TOMS (Total Ozone Monitoring Satellite) maps of the southern ozone hole (over the Antarctic) for a stretch of 8 years:



Many meteorological satellites are adept at picking out chracteristics of the oceans such as silt/sediment patterns, temperatures, wave trains, and current circulation. But several satellites have been flown primarily to sense these and other properties of the ocean surface (again, see Section 14). Among these are Seasat, Radarsat, the Coastal Zone Color Scanner (CZCS) on Nimbus 7, and SeaWiFS (now operating).

On SeaWiFS, several bands cover the blue, green, and red parts of the visible spectrum, and into the near infrared, yielding data that can be used to display variations in ocean color or, for particular bands, indications of the distribution and intensity of chlorophyll that resides mainly in surficial plankton. This SeaWiFS image maps the generalized ocean colors as well as chlorphyll concentrations (in red, yellow, and orange colors) on a near global scale during September, 1997.



Terra is the "flagship" satellite in the Earth Science Enterprise (ESE) that is the United States contribution to a continuing scientific effort often referred to as the International Geosphere-Biosphere Program. This, and Aqua (a second satellite in the program), are designed to support the growing field of Earth System Science. Mentioned near the end of the Overview, the IGBP and its spin-offs are of sufficient scope and merit to deserve its own Section (16) in this Tutorial.

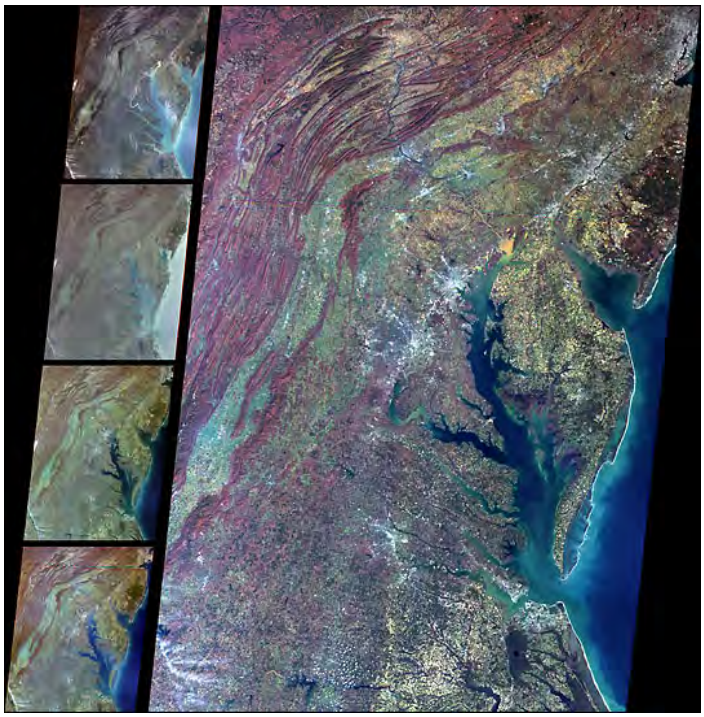
The five sensors on Terra are: **MODIS, MOPITT, MISR, ASTER, AND CERES**. For the moment, we show here representative images (others will appear in Sections prior to Section 16) made by the ASTER , MODIS, and MISR instruments on Terra. Examples of these will be encountered throughout the Tutorial, so some familiarization with the appearance of each sensor's images made by each is warranted.

ASTER is the acronym for Advanced Spaceborne Thermal Emission and Reflection Radiometer. Here is a false color rendition of a part of southern California that includes Palm Springs.

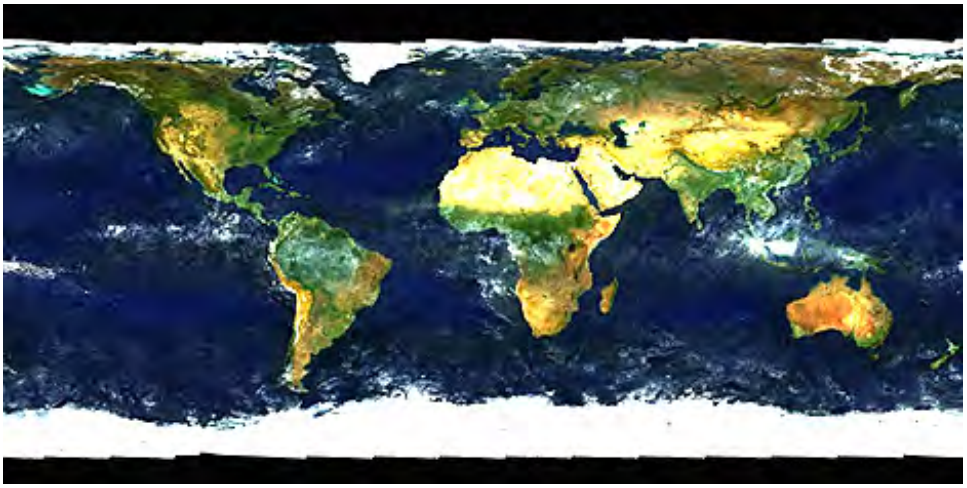
This next ASTER scene is of volcanoes in the Andes mountain chain of South America. Volcanoes are important components of the Earth System being studied by the ESE in that they affect the environment on regional to worldwide scales by expelling into the atmosphere gases and dust that can affect weather patterns.

MODIS, for Moderate (resolution) Imaging Spectrometer, covers usually large areas, such as the Black Sea shown here:

MISR stands for Moderate Imaging SpectroRadiometer. The instrument is capable of pointing off nadir at various angles as well as straight down. Here is Chesapeake Bay, with a vertical and side views; note the wide swath:



The last image is also constructed from a multispectral satellite sensor which produces color imagery. It is a natural color "portrait" of the entire globe, in which vegetation-rich areas are in green, vegetation-poor (including deserts) areas are in various shades of yellow and brown, and ice is in white. One thing brought out in this world view, is that a large part of the total land surface does not have extensive vegetation cover - this helps to visualize the possibly precarious state of those biomes that contain most of the living species, recycle oxygen to the atmosphere, and provide organic raw materials and foodstuffs for the health and survival of the human race and many members of the animal kingdom.



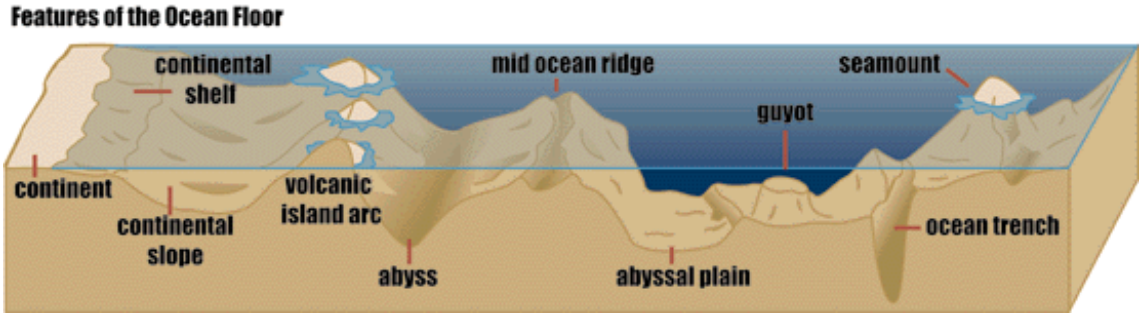
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

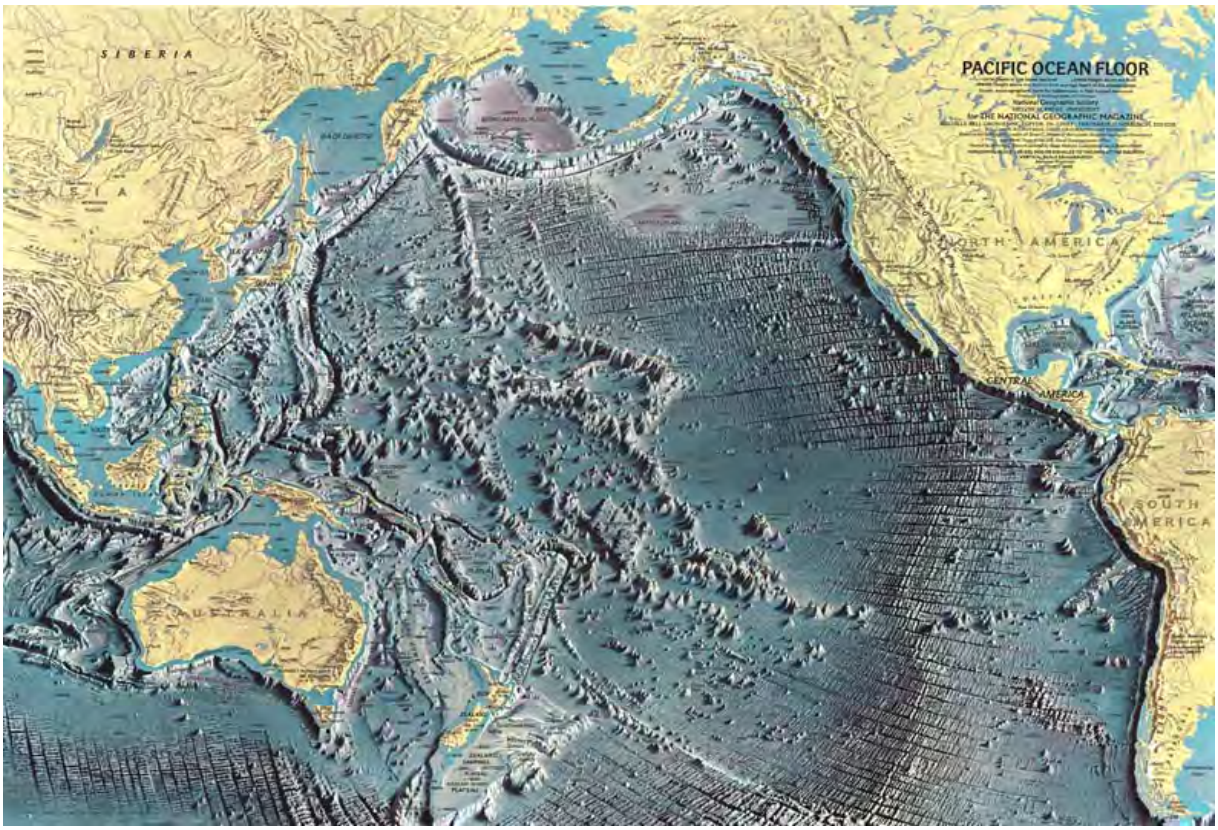
Remote sensing normally looks at surfaces. On Earth, most of the surface consists of water, mainly in the oceans. But beneath the marine surface at the bottom of the seas is a solid rock and sediment surface. Knowledge of that surface comes in part through remote sensing and in part through direct observation from diving vehicles. The principal remote sensing tool used for ocean floor studies is sound waves transmitted by Sonar instruments. Another mode of studies is photography or imaging cameras mounted on submersible platforms, either unmanned or manned. In recent years, oceanographic exploration has gone a long way towards exposing what is at the floor of the still largely unexplored ocean environment.

Remote Sensing of the Ocean Floor

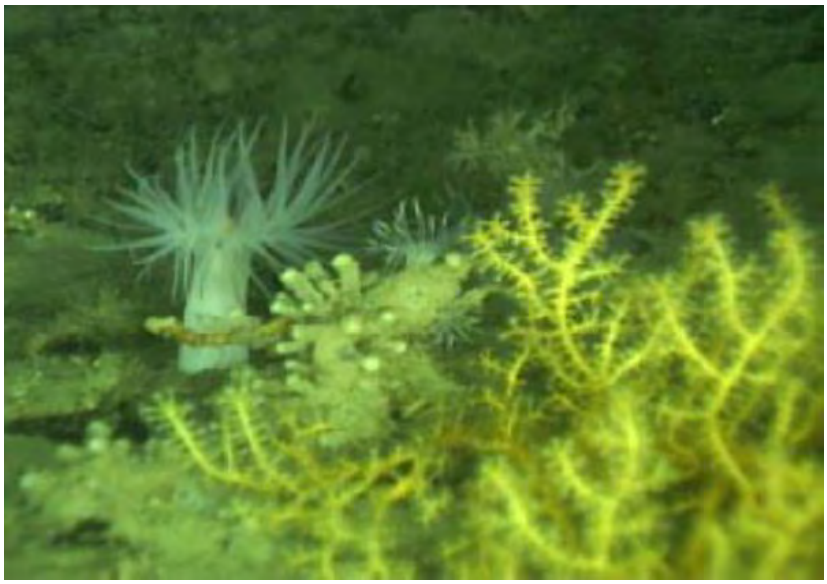
About 70% of the Earth's surface is covered by water - mostly marine (oceanic). The shallowest water is adjacent to continents and islands; the deepest (down to 12000 meters; 37000 feet) is in trenches near subduction zones. The first illustration shows the main features on ocean floors:



This is a topographic sketch map of the Pacific Ocean, the largest continuous expanse of submarine topographic which contains these features. Especially prominent are the trenches, the volcanic island chains, and the ridges cut by transverse faults:

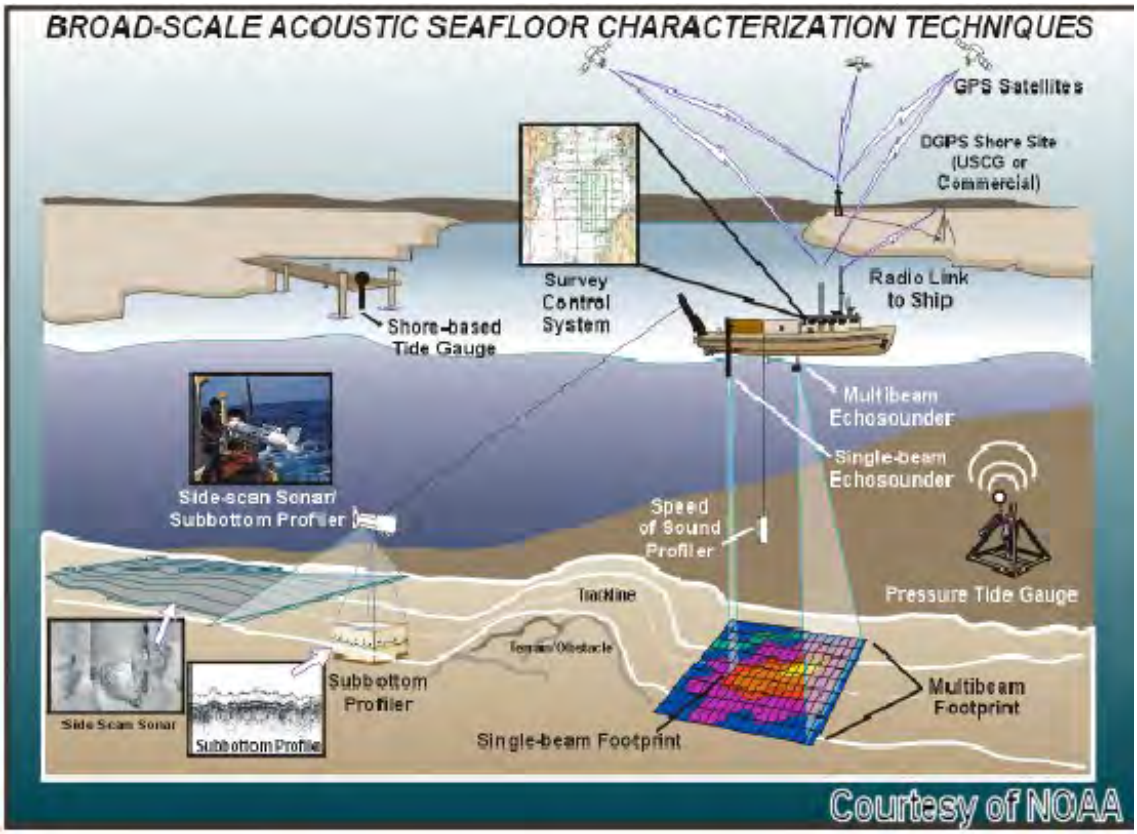


Even prior to modern times it was well known that the ocean floor near the continents was teeming with life. Here are two examples of life forms attached to the sea floor in shallow seas.



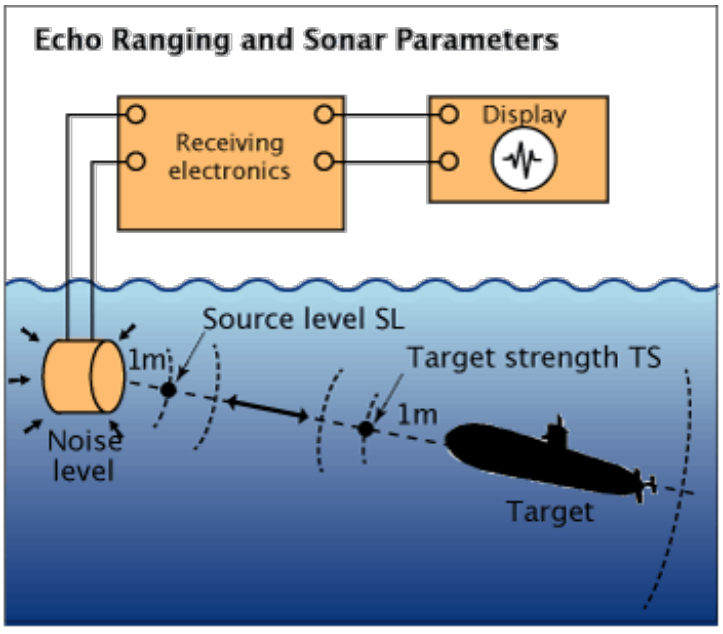
But, despite some knowledge of the ocean floor in offshore environments, surprisingly little detail is known about the vast majority of the ocean floors across the globe. Only about 3% has been explored to any extent. Exploration is twofold: instruments such as cameras have imaged small areas on the floor or much larger areas by sending probing waves from transmitters to the floor and measuring the return signals, and direct observation by humans either in shallow waters (scuba divers) or deep waters (in submersible vessels). Most subaerial remote sensing devices that peer at the Earth, as from airplanes or satellites, do not work in water. Sonar (described below) is the customary remote sensing device; cameras (photographic; digital; television) also are used. And, direct sampling of the floor is possible by drilling into ocean beds from platforms or from ships;

The principal way in which the ocean floor is mapped and studied is through the use of acoustic waves. This diagram depicts the main techniques for using sound waves to map the floor, locate objects on the floor, and search for manmade and animal objects within the ocean.

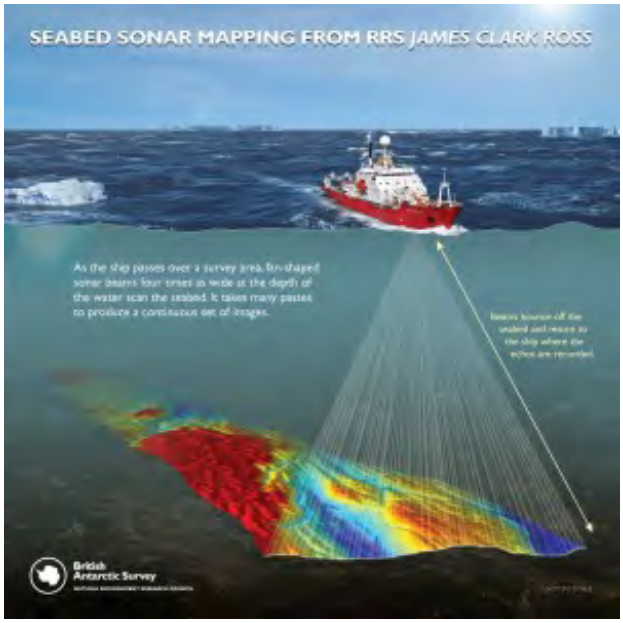


The most widely used technique for ocean floor mapping is Sonar (acronym for Sound Navigation and Ranging). For an overview of the principles behind Sonar, consult the Wikipedia and LEI websites:Sonar Overview How Sonar Works.

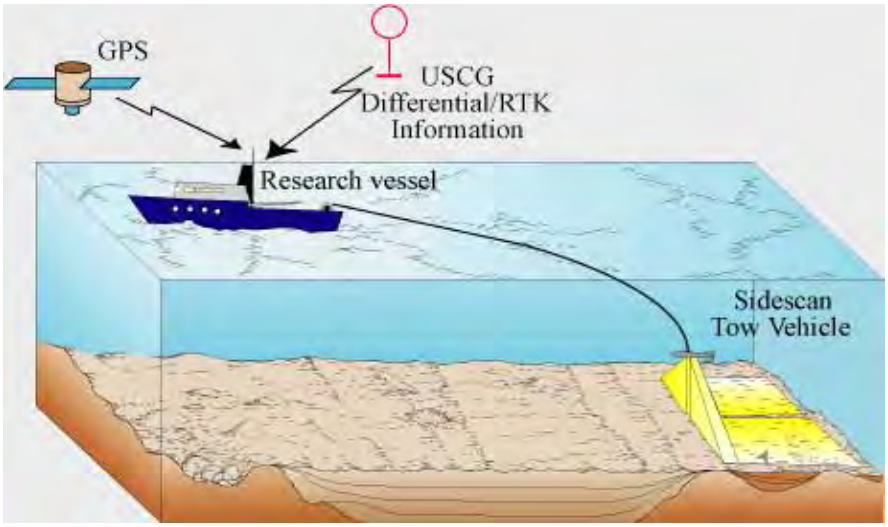
The Sonar technique dates back to the 1930s and was much advanced by naval military needs in World War II. The principle behind Sonar is similar to that involved in radar and in making sonograms of the human body's organs. A sound pulse (either single frequency or multi-frequencies) is sent by a transmitter in a narrow direction or more spread out. The rate at which the pulse travels is generally well known for water; it varies with density and with temperature. The pulse can hit an object - ocean floor, a boat, or a school of fish - and is reflected from same as an echo. The reflection can move off in various directions but one such direction goes directly back to the transmitter source and is picked out by a receiver. Since travel times out and back are known for the water medium, the delay between outward pulse and return pulse gives a time that leads to the distance to the reflection object. This diagram shows the concept for a signal sent from a submarine;



When pulses are sent in various directions at once and their return signals are likewise received, the small differences in travel time translate (in modern systems with the aid of a computer) into indications of variations in the shape and distances of surfaces that are usually varied in form, such as those characteristic of topography. This diagram depicts a general case:



In one mode of data gathering, a Sonar device is towed behind a ship, as shown here:



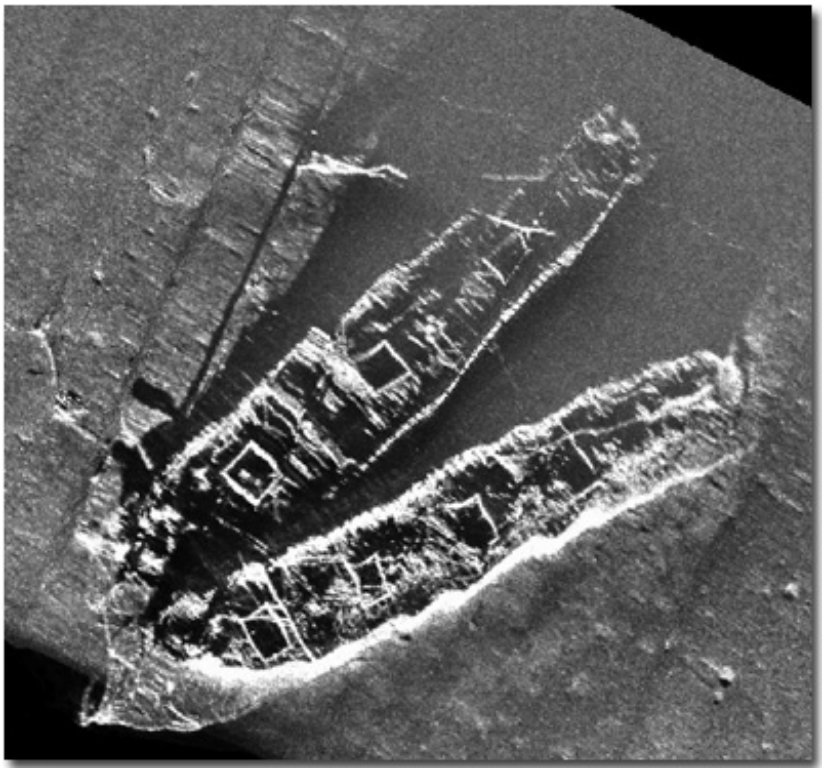
Here is a close up showing one of the smaller Side Scan Sonar units as it is being lowered on cable from its ship:

One of the most sophisticated Side Scan Sonar devices is GLORIA (GLORIA stands for Geological Long Range Inclined Asdic), mounted on the side of a ship:

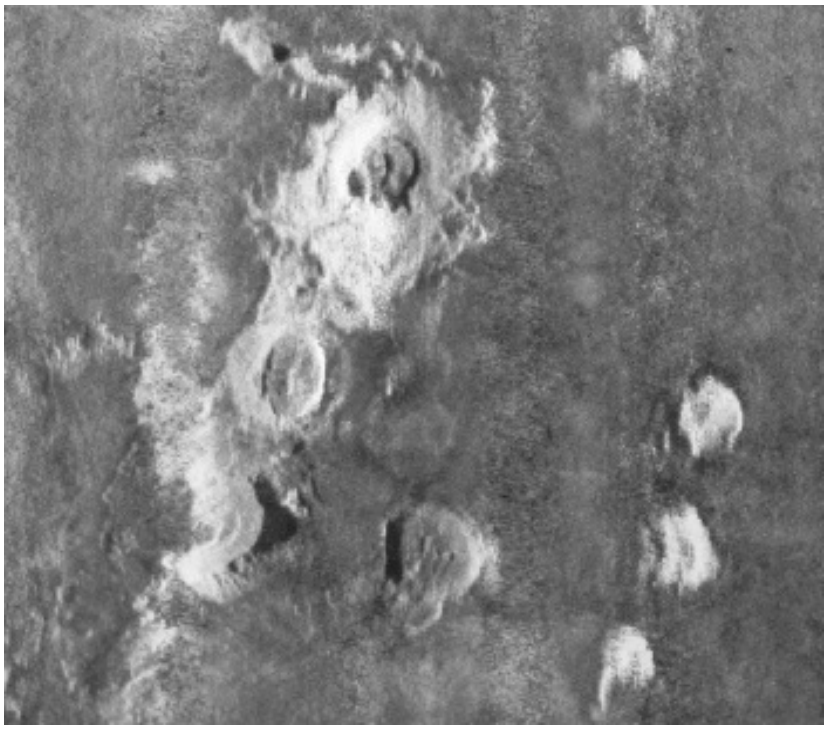
Sonar is frequently used by navies to search for, then examine, sunken ships. Here is a Sonar view of a submarine resting on the sea floor.

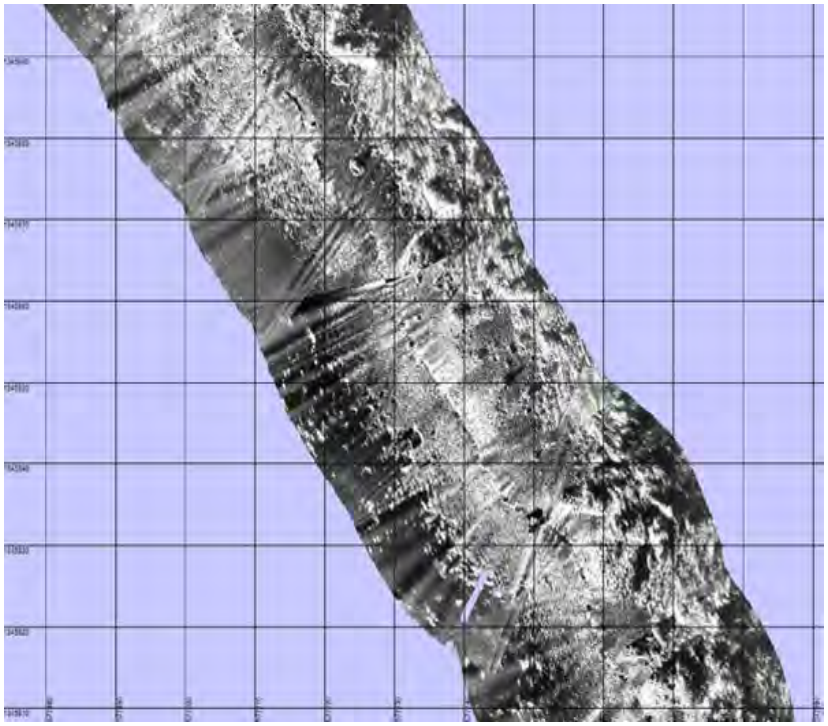
German U-boat, U-166
Side Scan Sonar Image from the HUGIN 3000 AUV

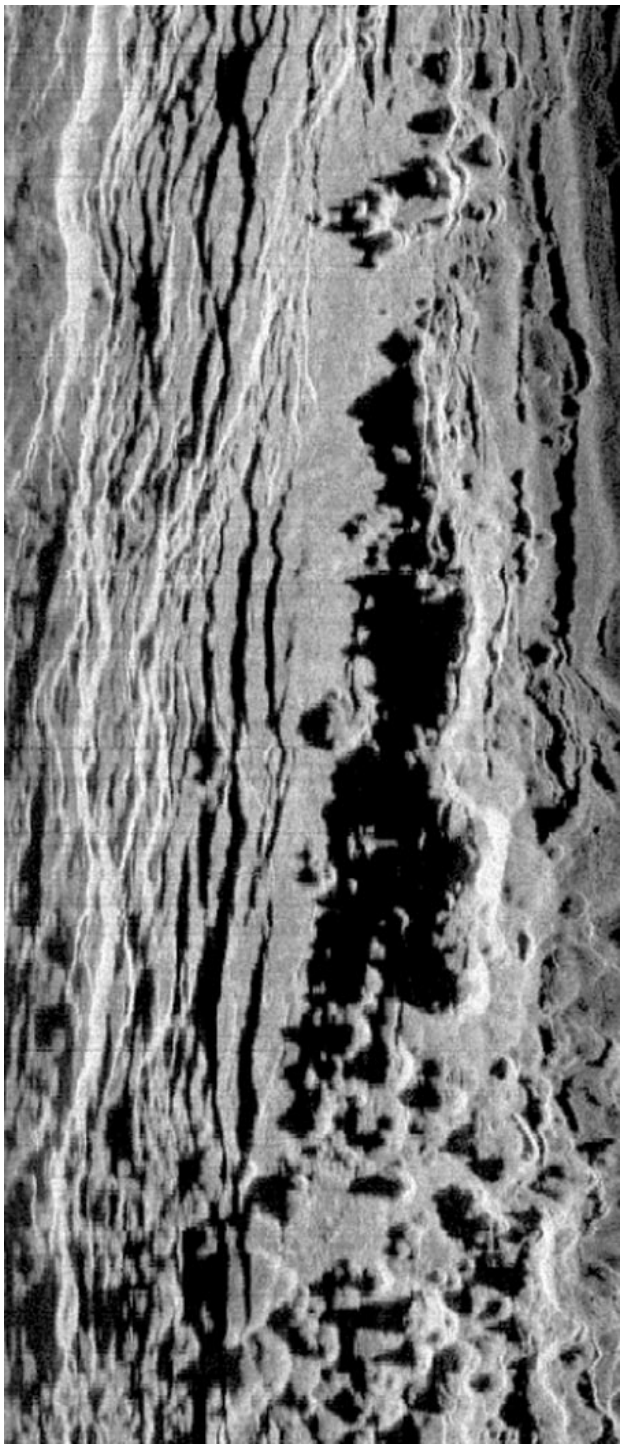
Sonar has been used to look at ships lost at sea. Two merchant vessels, the Frank Palmer and the Louise Craig, collided in December of 1962 and both sank. Here is a Sonar image of the two resting side by side.



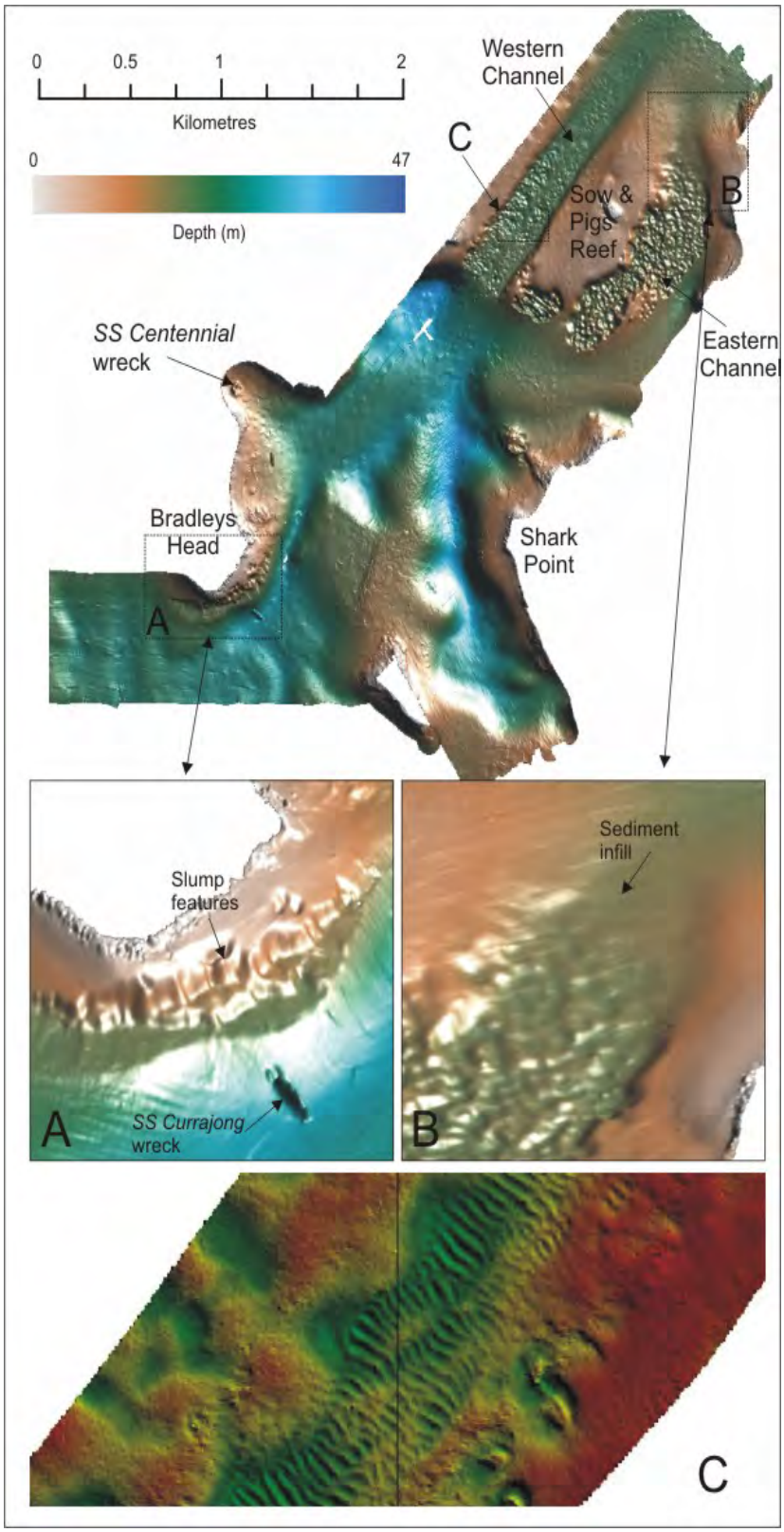
Parts of the sea floor have been mapped using Sonar to obtain images that display the topography. Three examples appear here; their captions provide explanatory information:

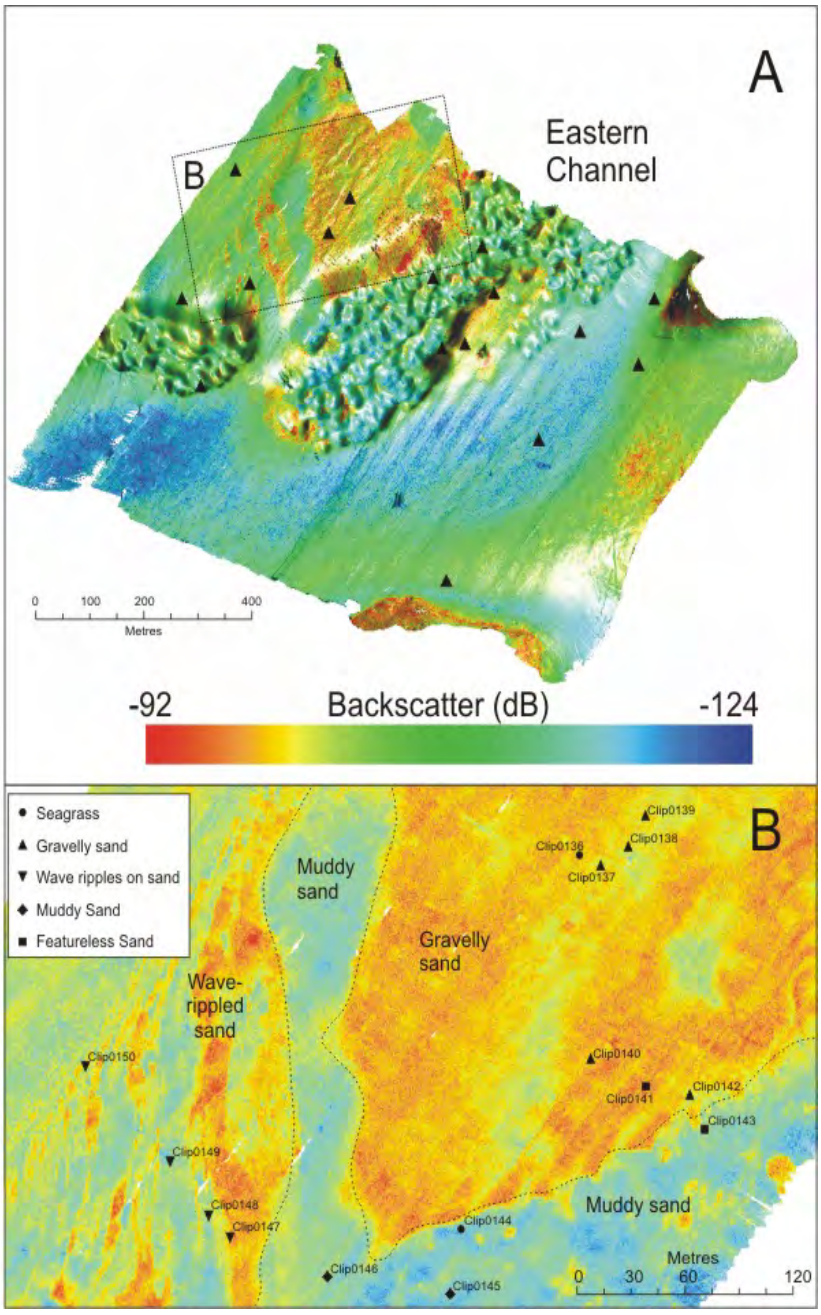




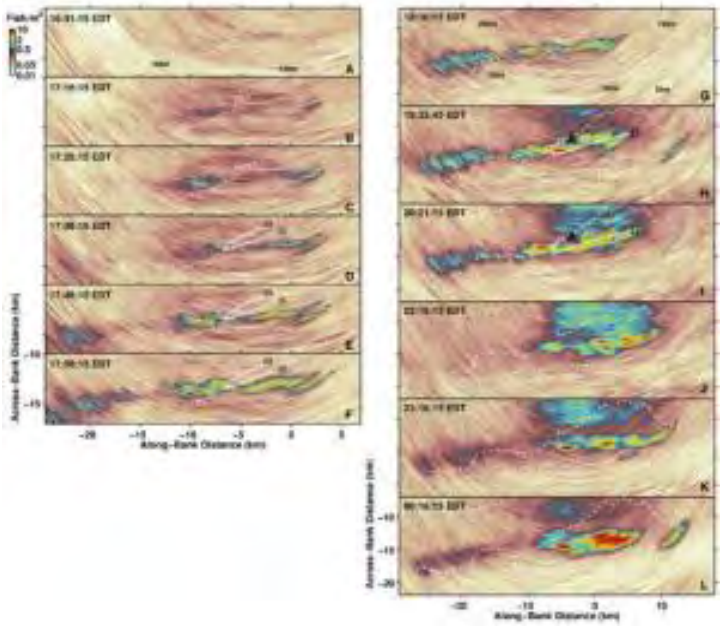


Here are two maps of the seafloor off the Australian coast that show the end product of a Sonar scan of the topography:


0
0.5
1
2
Kilometres
0
47
Depth (m)
Western Channel
C
Sow & Pigs Reef
B
SS Centennial wreck
Eastern Channel
Bradleys Head
A
Shark Point
Slump features
SS Currajong wreck
A
Sediment infill
B
C



As mentioned above, Sonar can detect objects in the water above the ocean floor. A common use is to locate schools of fish. Here is a sequence of images that show how one such school was monitored as it grew in numbers of individual herring:



The size of a Sonar unit has now diminished to one that can fit in a hand. These fishing Sonars sell for about $60 and thus are affordable for the average sports fisherman.

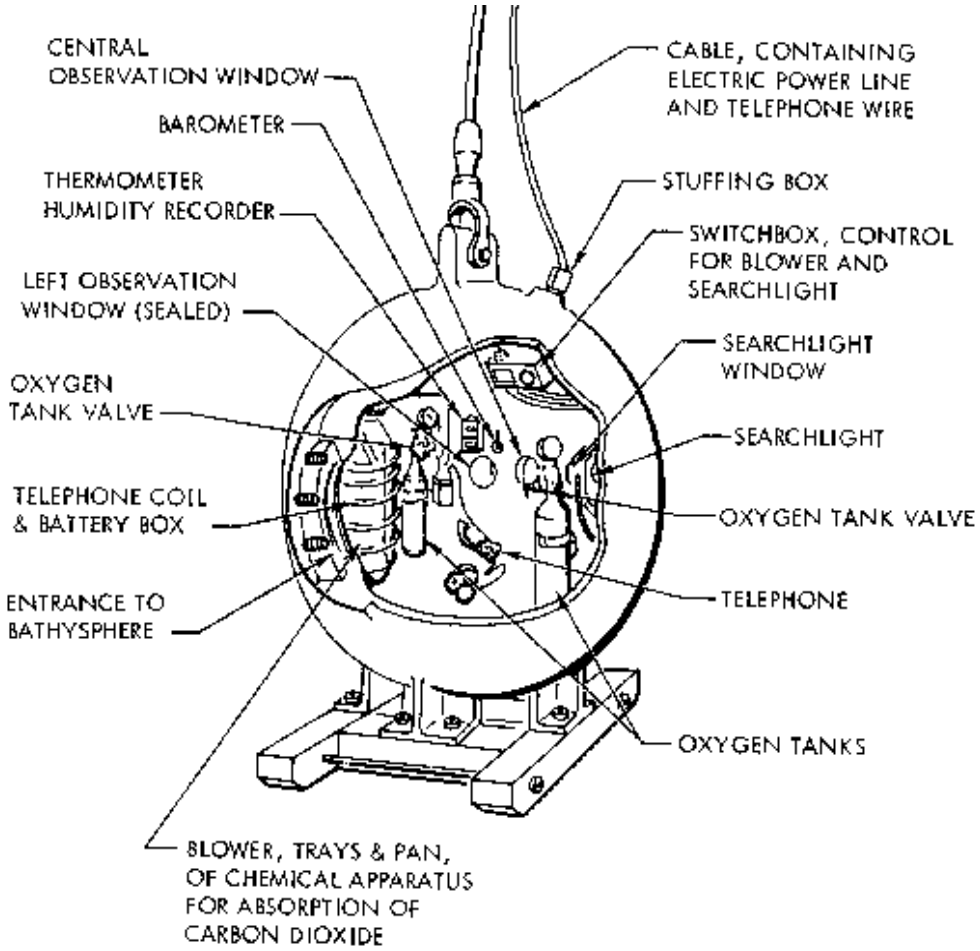
One might think that underwater cameras or televisions would be useful in obtaining images of the sea floor. They have been used but suffer from one severe limitation. Most of the ocean water below a few hundred meters is totally dark; illumination associated with the camera would be restricted to a small area. Thus most photography and image making is done from light provided by a submersible craft.

Submersibles, some unmanned and others with one to several humans, have been lowered into the ocean on cables but a few have been operated as "free" craft that control their movements including return to the mother ship. Submersibles have been employed for sea-oriented exploration since the end of World War II. They can be operated much like submarines, from whose technology they sprang. Most are small and have been likened to midget submarines.

Some of the submersibles mounted on cables are called bathyspheres (although some are

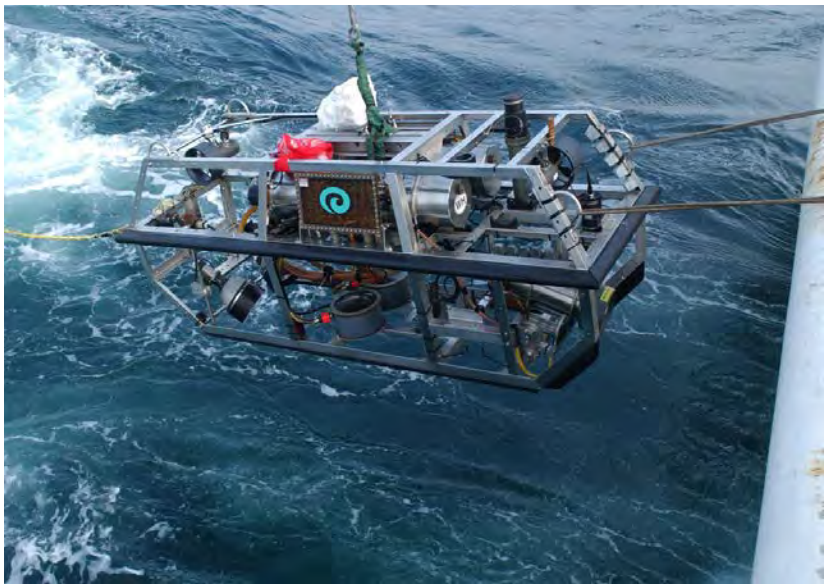
spherical, today they normally are elongate). The first dive made from a manned bathysphere was to a depth of about 300 meters off Bermuda. Its occupant was the famed oceanographer William Beebe.

Here is a cutaway schematic of a typical bathysphere:

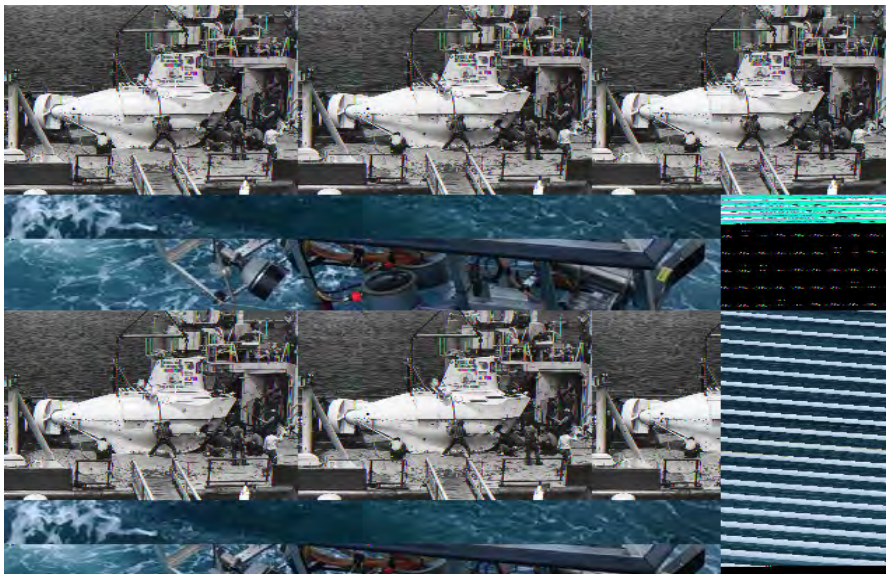


Two oceanographers associated with bathyspheres are the brothers Auguste and Jacques Picard.

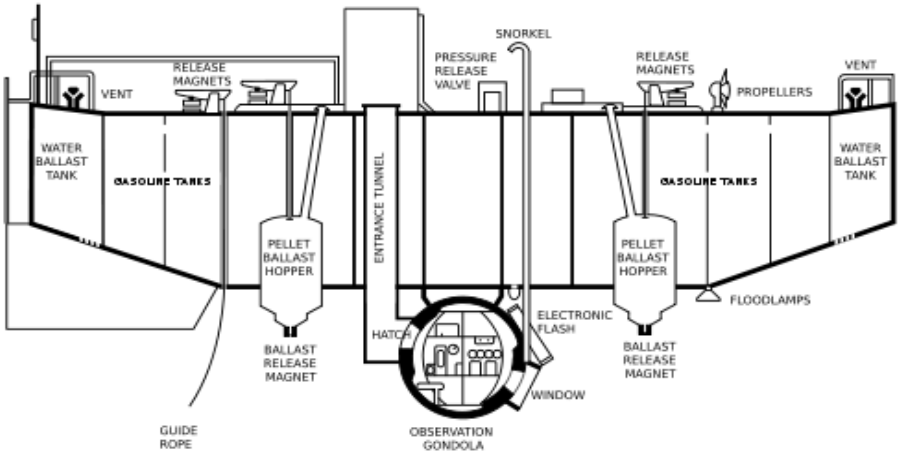
Some unmanned submersibles are simply open platform frames within which are the instruments that make the observations and measurements. These ROVs (Remotely Operational Vehicle) are lowered on cables to the sea floor to observe in place. These are then raised off the floor and moved to other locations. Examples of these ROVs are NOAA's Hercules and Argus, which sometimes were lowered together but performed different functions:



True midget subs used in oceanographic studies were first deployed in the 1960s. The best known of these is Alvin, first used in 1964. It takes its name from its designer, Allyn Vine.



Made of structural strong and re-enforced metals, these submersibles were usually employed to make truly deep dives. They became the main vehicles for exploring below 3000 meters. The deepest dive on record was to the bottom of the Marianas Trench in the western Pacific. A depth of 37000 feet was attained by the Trieste on January 23, 1960, with Jacques Picard onboard.

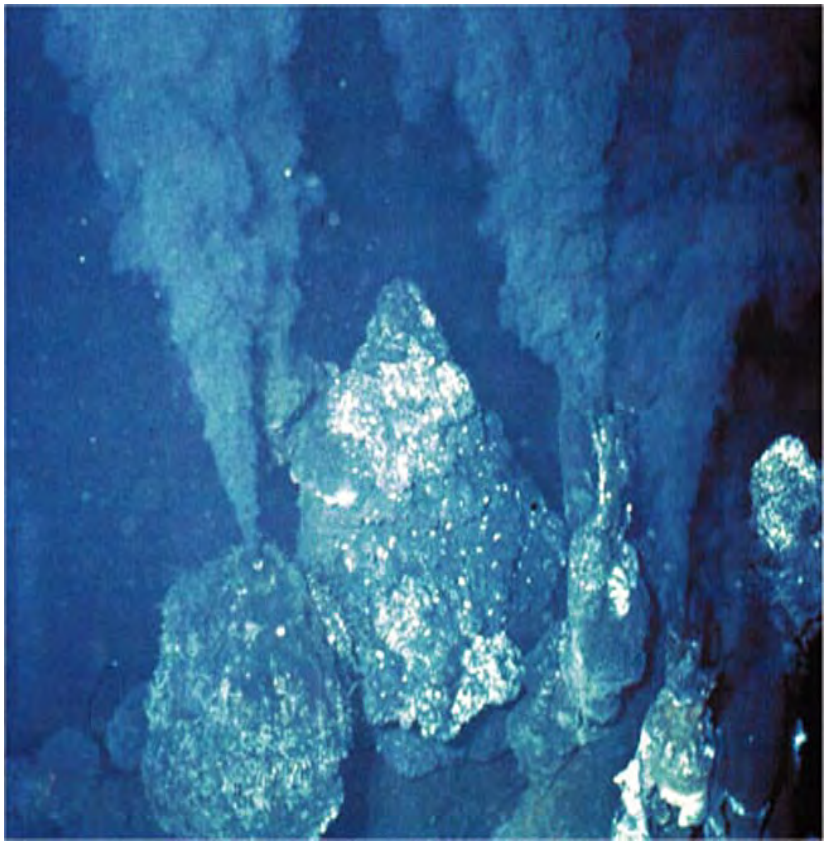


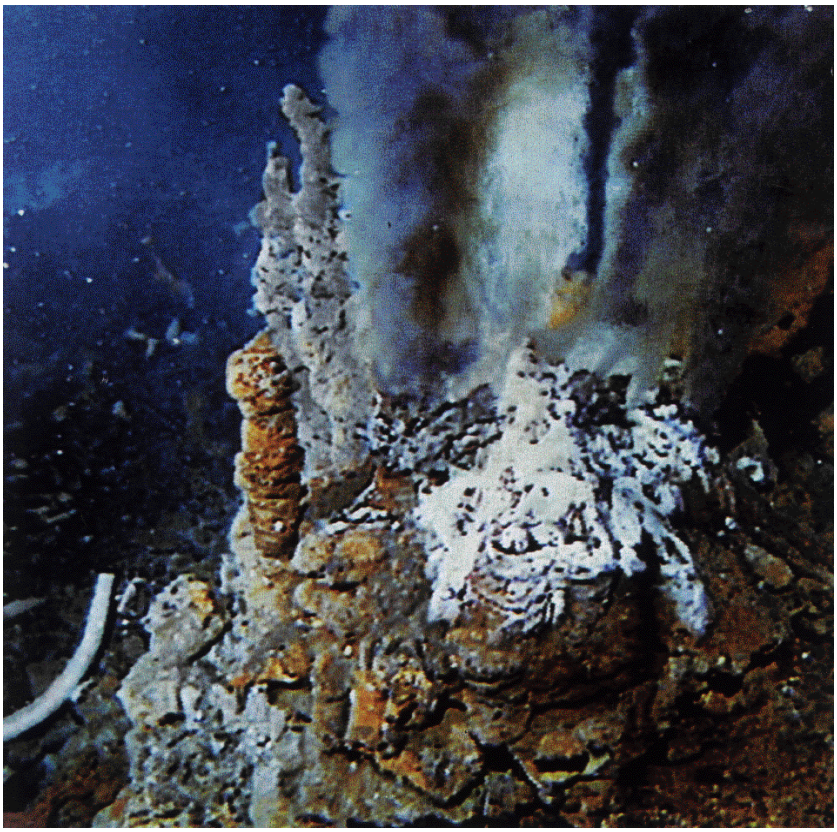
GENERAL ARRANGEMENT DRAWING OF TRIESTE, CA. 1959

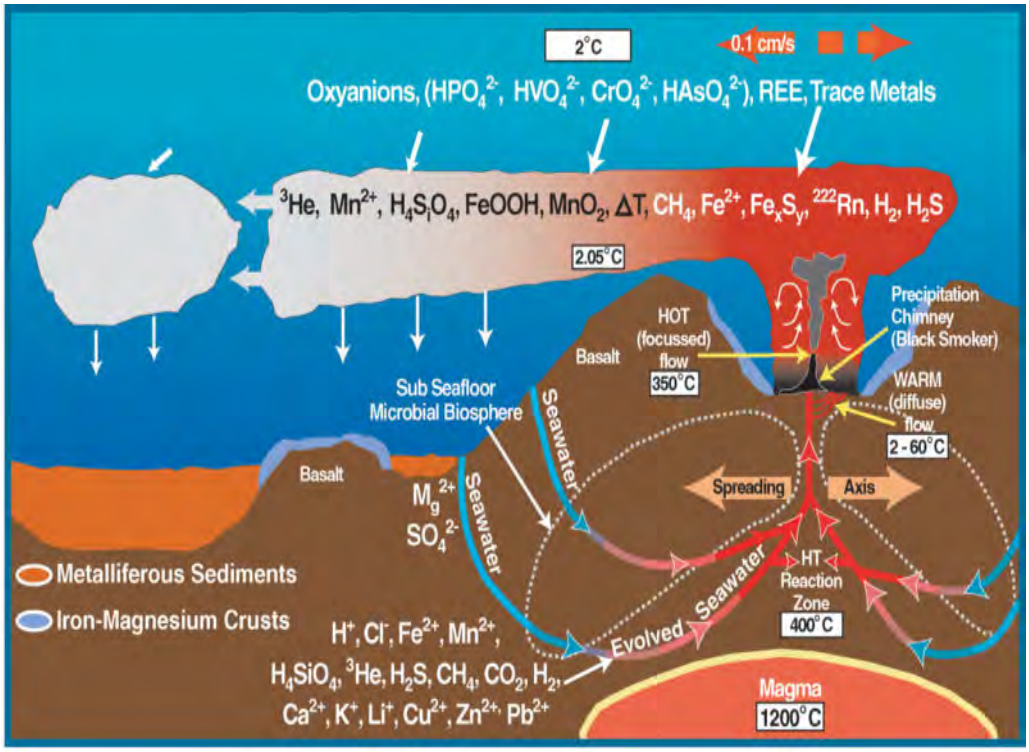
One of the most intrepid and versatile of the deep sea explorers is Dr. Robert Ballard. He has made many dives in manned submersibles and has made major discoveries with unmanned submersibles. His fame was firmed up when he discovered the wreckage of the Titanic, which sank on April 15, 1912 during its maiden voyage after striking an iceberg about 523 km east of Newfoundland. The ship lies at a depth of 3780 m; 12600 feet. Artifacts have since been recovered from the vessel. (This notorious maritime incident is immortalized in two great movies, both entitled "Titanic".)

Ballard has discovered many wrecks including cargo sailing ships hundreds of years old, even older boats lost in Roman times, and some modern vessels (the German battleship Bismark, for instance). This is a view of the U.S.S. Yorktown lost during the battle of Midway in 1942 in the Pacific in which 4 Japanese carriers were sunk in what many consider the greatest sea battle in history.

Ballard has financed his oceanographic studies in part through these spectacular underseas finds. But his main goals have always been scientific knowledge. He is one of the first to have discovered 'black smokers' - vent structures on mid-oceanic ridges in which heated waters build up pillars from precipitation of dissolved chemicals. Here are two view of black smokers, followed by a diagram that shows their nature in detail:







The black consists in part of manganese oxides, but many other metal elements are enriched in the particulates that make up the "smoke". Manganese and other elements also form away from the ridges as nodules produced by precipitation of dissolved elements in seawater. These nodules occur in huge clusters in ocean beds and represent a major source of valuable metal elements (Mn, Ni, Co, V, Cr, others). Schemes to mine these nodules by sucking them up have been proposed and demonstrations of feasibility implemented.

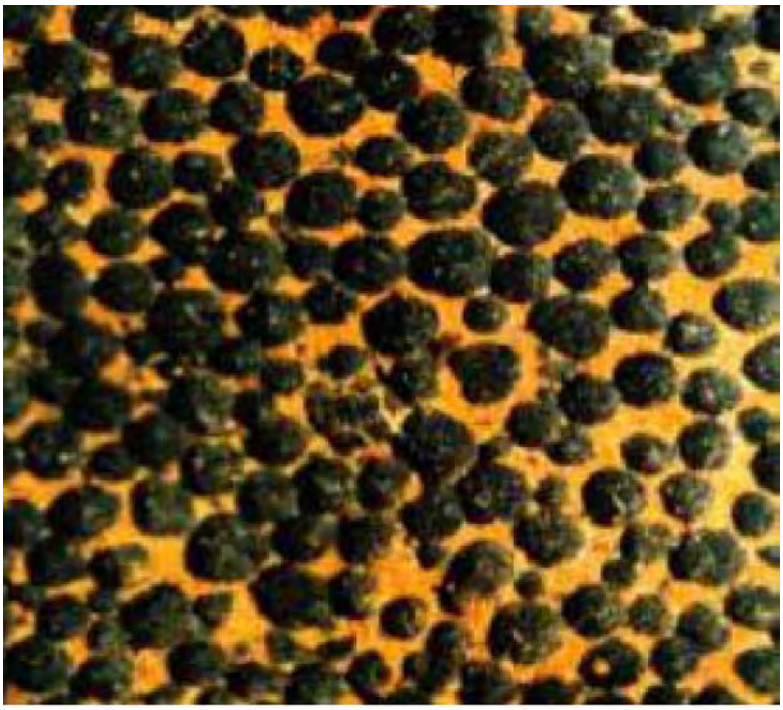


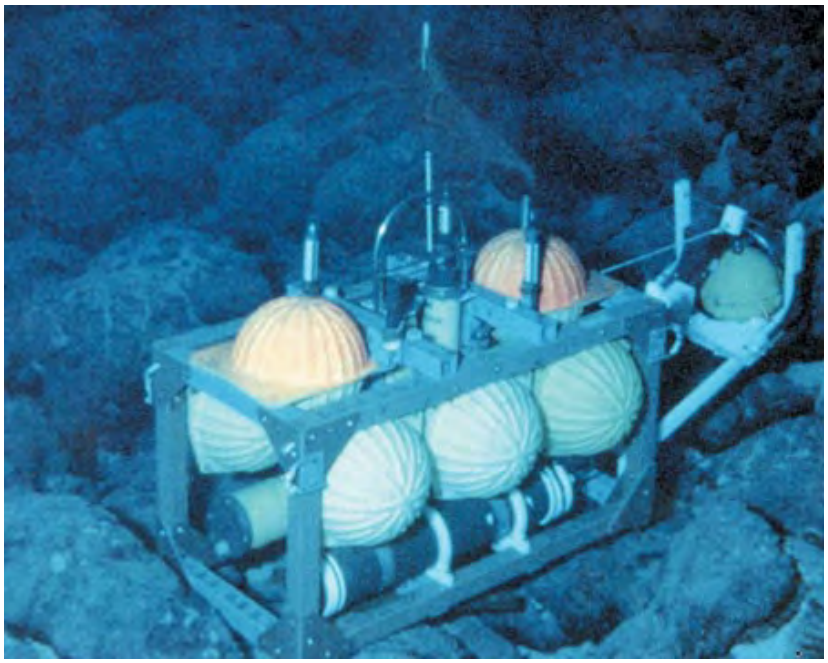
Fig. 39. Ferromanganese nodules on sea bed

But the discovery of black smokers led to an associated discovery that has major implications for the origin of life. Tube worms and other creatures are encrusted on and near the smokers:

These animals develop at great depths where sunlight never reaches. Unlike most animals they do not utilize sunlight directly or indirectly to maintain life. They derive their energy from the heat associated with the smoker fluids. One hypothesis, based on this observation, is that life may have originated in ocean deeps where thermal energy brought about its synthesis.

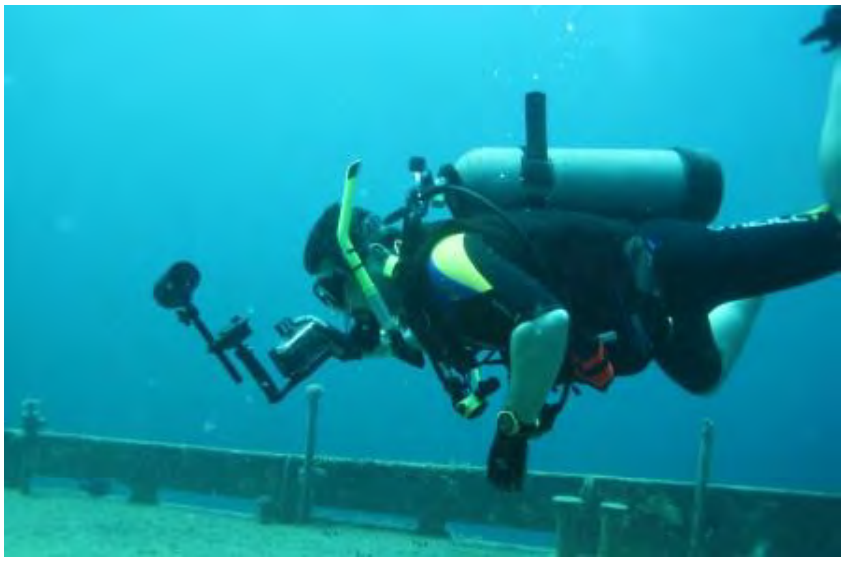
Oceanography as a science has matured in the 20th century and remains a major field of research as the still largely unexplored ocean floors are being mapped, sampled, and documented visually. The variety of methods involved in the study range from emplacement of instrument as platforms on the floors, such as the geophysical seismometers shown in the next illustration, to voyages of research vessels that have their own onboard instruments and can

lower submersibles.



One of the best known oceanographers who has popularized the study of the oceans is Jacque-Yves Costeau whose famous boat, the Calypso 2 (a converted World War II minesweeper), is shown here:

Nowadays individuals without any scientific background can enjoy exploring the ocean floor. In the last half century scuba diving has opened up the sea bottom in shallow waters off islands and continents. Submersible cameras allow photography to document sea life and other features:



We will encounter illustrations that deal with the surfaces of the oceans in several Sections of this Tutorial. Treatment of remote sensing of the marine surface is assigned it own pages in Section 14, pages 14-11 through 14-14.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

This Introduction closes by citing sources of data, information, and image hard copy such as continues to be supplied by the Eros Data Center (EDC) and other distributors of products from Earth-monitoring satellites; reference is made to the Idrisi image processing program.

Concluding Miscellany

A notable number of space missions were described, or at least mentioned, in this Introductory Section. We once again provide this NASA website (earlier in the Overview) that provides a list of most of the missions conducted by NASA and international space agencies.

A principal source for Landsat imagery and much of the astronaut space photography is the EROS Data Center (EDC) in Sioux Falls, S.D., operated by the U.S. Geological Survey (USGS).

EDC has assembled an Internet on-line collection of satellite images (mainly Landsat) designed to introduce the general viewer to scenes worldwide that focus on several environmental themes (e.g., cities, deserts, forests). We have linked their three-page Index for this *Earthshot* collection, which you can access at the Earthshots page.. To further stimulate interest in the practicality of this imagery, in addition to its beauty, we suggest taking time to sample scenes of interest and read the accompanying descriptions. We also recommend returning to this collection for repeat looks or to examine new examples. A recent Web site dedicated to showing a wide variety of Earth Science-related images from various satellites is at NASA's Visible Earth.

Based on sales by U.S. distribution centers, such as EROS and SpaceImaging, indicate that remote sensing is now truly a worldwide activity. The big trend in the 1990's and into the next century is the *commercialization of space*. Remote sensing is becoming a multi-billion dollar industry and new national organizations and private companies emerge each year to take advantage of its income-producing aspects, because they identify many applications (some covered in this tutorial) of great practical value. There is even a monthly magazine, EOM, dedicated to these increasing uses. We shall more fully touch upon trends and issues in *commercialization* of remote sensing as we review the outlook for the future of remote sensing in Section 21.

With this Introduction into basic principles and to characteristics of Landsat and other systems

completed, and, hopefully, digested and understood, you should move on to Section 1, with its protracted tutorial development of the "whys" and "hows" of image processing (largely computer-driven). There you should gain real insight into and practice in the efficacies of remote sensing from satellites and other types of air and space platforms.

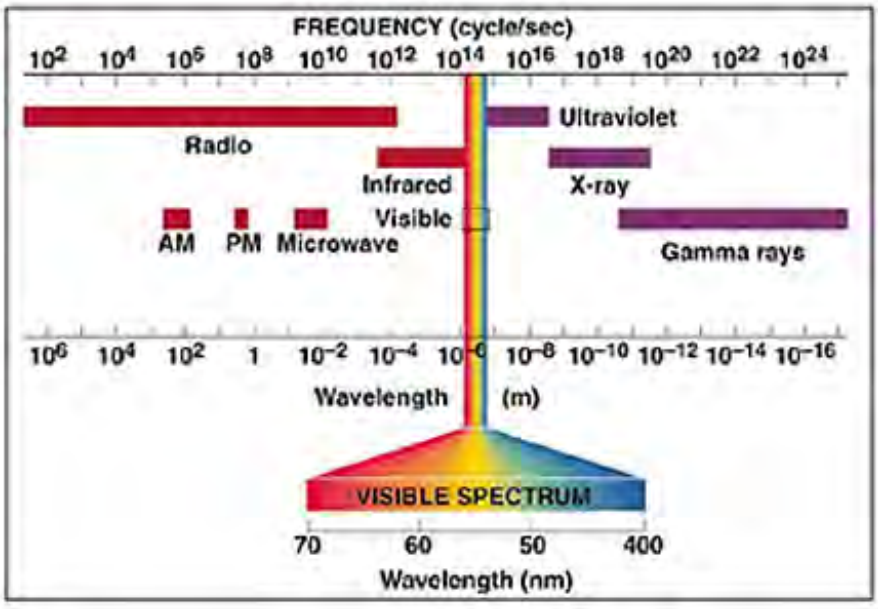
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

THE QUANTUM PHYSICS UNDERLYING REMOTE SENSING; USE OF SPECTROSCOPY IN DETERMINING QUANTUM LEVELS

This page is a supplement to your reading of page I-2. The page here was reconstructed (some changes and additions) from the relevant text in the Landsat Tutorial Workbook.

The physical phenomena most frequently sampled in remote sensing are photon energy levels associated with the Electromagnetic Spectrum. The Electromagnetic Spectrum is discussed at some length in this Introduction on page I-4. We reproduce an EM Spectrum Chart here in a rather unusual version in which the plot from left to right is "geared" to Frequency rather than the more conventional Wavelength spread (shown here at the bottom; both Frequency and Wavelength are treated below). The terms for the principal spectral regions are included in the diagram:

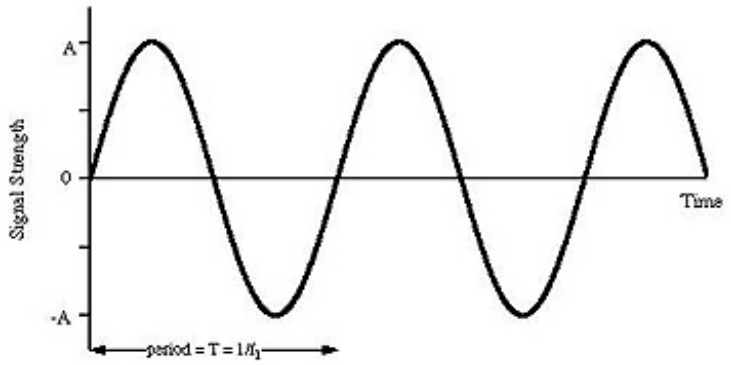


The Electromagnetic Spectrum is a plot of Electromagnetic Radiation (EMR) as distributed through a continuum of photon energies. Thus EMR is dynamic radiant energy made evident by its interaction with matter and is the consequence of changes in force fields. Specifically, EMR is energy consisting of linked electric and magnetic rields and transmitted (at the speed of light) in packets, or *quanta* in some form of wave motion. Quanta, or photons (the energy packets first identified by Einstein in 1905), are particles of pure energy having zero mass at rest. The demonstration by Max Planck in 1901, and more specifically by Einstein in 1905, that electromagnetic waves consist of individual packets of energy was in essence a revival of Isaac Newton's (in the 17th Century) proposed but then discarded corpuscular theory of light. Until the opening of the 20th Century, the question of whether radiation was merely a stream of particles or was dominantly wave motion was much debated.

The answer which emerged early in the 1900s is that light, and all other forms of EMR, behaves both as waves and as particles. This is the famous "wave-particle" duality enunciated by de Broglie, Heisenberg, Born, Schroedinger, and others mainly in the 1920s. They surmised that atomic particles, such as electrons, can display wave behavior, for example, diffraction, under certain conditions and can be treated mathematically as waves. Another aspect of the fundamental interrelation between waves and particles, discovered by Einstein between 1905-

07, is that energy is convertible to mass and that, conversely, mass is equivalent to energy as expressed by the famed equation $E = mc^2$. In one set of units, m is the mass in grams, c is the speed of EMR radiation in a vacuum in centimeters per second (just under 3×10^{10} cm/s), and E is the energy in ergs. (Other combinations of units are also used; thus, in high energy processes at high temperatures the electron-volt is the most convenient expression of E.)

EMR waves oscillate in harmonic patterns, of which one common form is the sinusoidal type (read the caption for more information); as seen below, this form is also a plot of the equation for a sine wave.



(a) Sine Wave

The sine wave is a function that occurs often in mathematics, signal processing, alternating-current power engineering, and other fields. Its most basic form is:

$$y = A \sin(\omega t - \varphi)$$

which describes a wavelike function that describes the progression of a signal with time (t); sin refers to the trigonometeric sine function; peak up/down deviation from center = A (amplitude); the angular frequency is given by ω (radians per second); initial phase (t = 0) = -φ (φ is also referred to as a phase shift, e.g., when the initial phase is negative, the entire waveform is shifted toward future time [i.e. delayed]); the amount of delay, in seconds, is φ/ω)

In the physics of wave movement, sinusoidal (also called periodic) types are known as transverse waves because particles within a physical medium are set into vibrational motion *normal* (at right angles) to the direction of propagation. EMR can also move through empty space (a vacuum) lacking particulates in the carrier medium, so that the EMR photons are the only physical entities. Each photon is surrounded by an electric field (*E*) and a magnetic field (*H*) expressed as vectors oriented at right angles to each other. The behavior of the *E* and *H* fields with respect to time is expressed by the Maxwell equations (4 needed to completely describe the behavior of the radiation). These equations include the terms μ_0 (permeability of the electric field in a vacuum; Ampere's Law) and ε_0 (permittivity of the magnetic field; Coulomb's Law). An important relationship between these terms and the speed of light is: $c = 1/(\mu_0\varepsilon_0)$ 1/2. The two fields oscillate simultaneously as described by covarying sine waves having the same wavelength λ (distance between two adjacent crest [trough] points on the waveform) and the same frequency ν (number of oscillations per unit time); ν is the reciprocal of λ, i.e., $\nu = 1/\lambda$. The equation relating λ and ν is c (speed of light) = λν since light speed is constant, as λ increases (decreases, ν must decrease (increase) the proper amount to maintain the constancy. When the electric field direction is made to line up and remain in one direction, the radiation is said to be plane polarized. The wave amplitudes of the two fields are also coincident in time and are a measure of radiation intensity (brightness).

Units for λ are usually specified in the metric system and are dependent on the particular point

or region of the EM Spectrum being considered. Familiar wavelength units include the nanometer; the micrometer (micron now obsolete); the meter; and the Angstrom (10^{-8} meters.

A fixed quantum of energy E (in units of ergs, joules, or electron volts) is characteristic of any photon transmitted at some discrete frequency, according to Planck's quantum equation: $E = h\nu = hc/\lambda$. From the Planck equation, it is evident that waves representing different photon energies will oscillate at different frequencies. It follows, too, that the shorter (longer) the wavelength, the greater (lesser) the energy of the photon(s) involved.

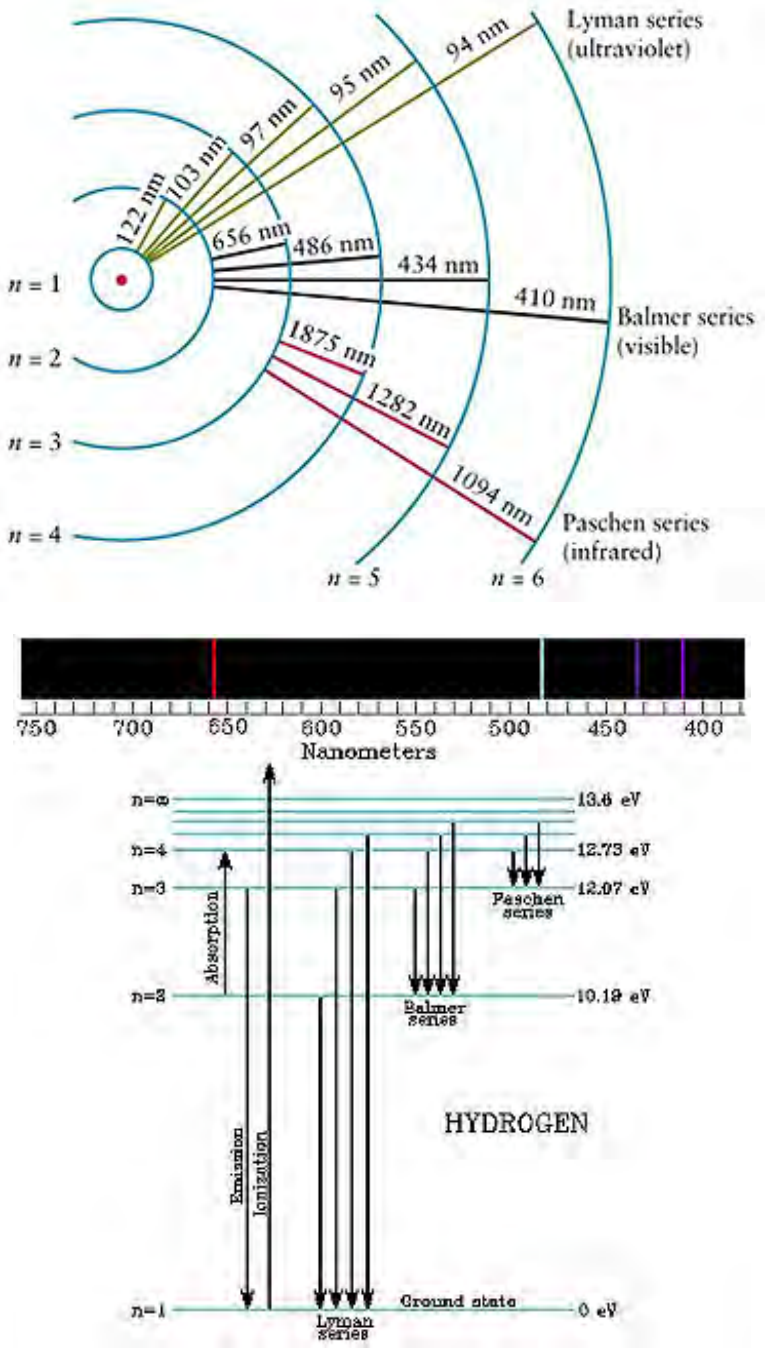
How is EMR produced? Essentially, EMR is generated when an electric charge is accelerated, or more generally, whenever the size and/or direction of the electric (E) or magnetic (H) field is varied with time at its source. A radio wave, for example, is produced by a rapidly oscillating electric current in a conductor (as an antenna). At the highest frequency (highest energy) end of the EM spectrum, gamma rays result from decay within the atomic nucleus. X-rays emanate from atoms within the source that are bombarded by high energy particles that cause the source electrons to move to an outer orbit and then revert to one further end (back to the ground state). (This process will be described later in this page.) In stars, high temperatures can bring about this electron transition, generating not only gamma rays and X-rays but radiation at longer wavelengths. Radiation of successively lower energy involves other atomic motions as follows: UV, Visible: transitions of outer electrons to a higher metastable energy level; Infrared: inter- or intra-molecular vibrations and rotations; Microwave: molecular rotations anf field fluctuations.

The interaction of EMR with matter can be treated from two perspectives or frameworks. The first, the *macroscopic* view, is governed by the laws of Optics, and is covered elsewhere in the Tutorial. More fundamental is the *microscopic* approach, which works at the atomic or molecular level. This is being considered on this page.

The basis for how atoms respond to EMR is found in the Schroedinger equation (not stated here because of its complexity and the need for background information that appears towards the bottom of the Preface in Section 20). This equation is the quantum analog to the Newtonian force equation (simplest form: F = ma) in classical mechanics. A key term in the Schroedinger equation is E, a characteristic energy value (also known as the "eigenvalue'). For many atomic species, there are a number of possible energy states (eigenstates) or levels within any speciies. For different species , the characteristic states are different and possess unique values (e.g., energies expressed in electron-volts) diagnostic of each particular element. There are, therefore, certain allowable levels for each atomic species whose discrete eigenvalues satisfy the Schroedinger equation. These levels are related to acceptable solutions for determination of the wave function associated with the Schroedinger equation.

Under the influence of external EMR (or thermal inputs), an atomic species can undergo a transition from one stationary state or energy level to another. This occurs whenever the oscillating EMR field (normally, but not necessarily, the electric field) disturbs the potential energy of the system by just the right amount to produce an allowable transition to a new eigenstate (new energy level in the quantum sense). The change in eigenvalue is given by $\Delta E = E_2 - E_1 = h\nu$, the expression for Planck's Law, in which h = the Planck constant (see Section 20-Preface) and ν is the wave frequency. This transition to the new energy state is evidenced by a reversion to a ground (starting) state or to an intermediate state (which may decay to the ground state) with a consequent output of radiation that oscillates sinusoidally at some specific frequency (wavelength) determined by the exact change in energy (ΔE) associated with the permissable transition. For most atomic species, a large amount of energy input can produce multiple transitions (to various quantum energy levels designated by n = 1, n = 2, ...) and thus multiple discrete energy outputs each characterized by a specific ν (λ).

A convenient way to appreciate this process is to examine several of the principal transitions of atomic hydrogen that has been excited by external energy. Consider these two illustrations:



In the upper diagram circles are used to represent the energy levels, the lowest being the ground state. In the Bohr atom model, these levels relate to orbits (orbitals) occupied by electrons (hydrogen has 1 electron in n = 1; helium has two electrons in n = 1; lithium has these two, plus 1 electron in the next orbit (or shell, but here called "Period", n = 2, and the next element in the Periodic Table, beryllium has 2 electrons in n = 2, then boron with 3 electrons in n =2 until that Period has a maximum number of 8 electrons (the element is helium, with a total of 10 electrons [and 10 protons as these two particles are always equal in the normal {unionized} state]. There are 7 Periods in all, with some containing up to 8 and others more (comprising series, built up from rules not addressed in this review). Currently, there are 103 elements (each with its specific number of protons), some of the higher numbered ones are known only or mainly from high energy physics "atom smashers".

The single hydrogen electron in n = 1 can be transferred into higher energy states within Periods 2 through 6. On decay (transition back to a lower energy level), from any of these levels back to n = 1, discrete energies are released for each level transition, according to the Planck equation. These are expressed here in nanometers but that unit can be converted to micrometers by multiplying its number by 10^{-3}. The first group of excited states, starting from n = 1, comprises the Lyman series (here, series is *not* used in the sense of the Period series mentioned in the

previous paragraph). The δE for each gives rise to spectra that fall within the Ultraviolet region of the spectrum. The electron may be placed in the n = 2 and n = 3 states and then jumped to higher states. The results are two more series, the Balmer (Visible Range) and the Paschen series (Infrared). Each transition shown on the diagram has a specific wavelength (frequency) representing the energy involved in the level changes.

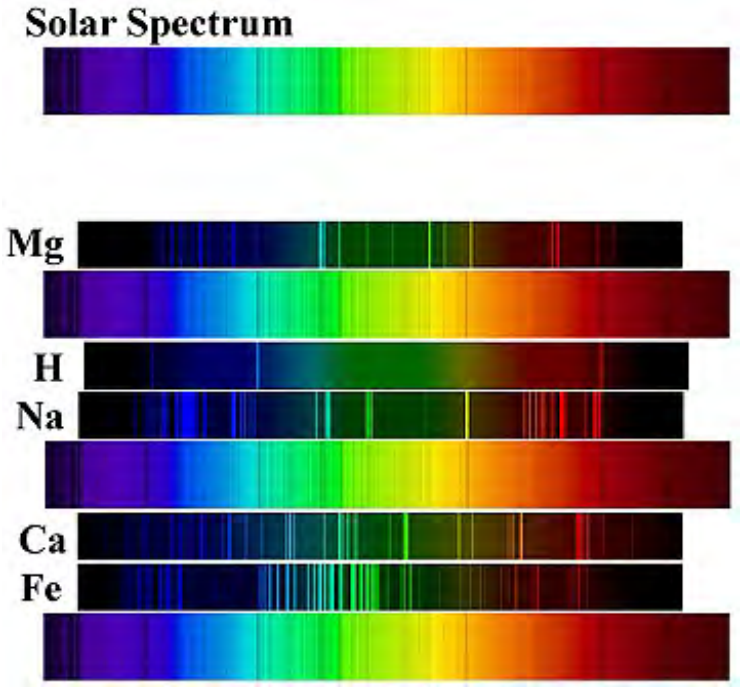
The lower diagram shows much the same information but with some different parameters. Thus, the energy level for each n is given as a specific value in electron-volts (eV). At the top of this diagram is a black band (part of the spectrum) representing a range of wavelengths, with thin colored lines fixing the location of each δE (as λ) for certain transitions.

Optional reading: As an aside, the writer (NMS) wants to relate how spectral information about atomic species is obtained: The following pertains to the 1950s version of the Optical Emission Spectroscope, the instrument I used during my Ph.D. research project. This type of instrument was responsible for acquiring data that helped in the fundamental understanding of EMR and the EM Spectrum.

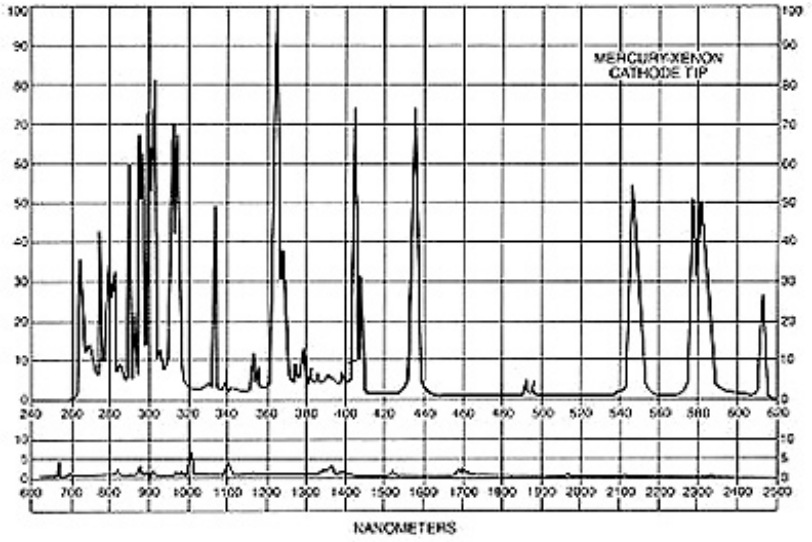
The spectroscope I used works this way. A single- or multi-element sample is placed usually in a carbon electrode hollowed to hold the material at one end; a second opposing pointed electrode helps to establish a small gap (arc) across which an electric current flows. The sample, subjected to a high voltage electric current, experiences high temperatures which vaporizes the material. The excited sample, which in effect "burns", has its material dissociated into constituent atoms, ionizing the elements involved. The excitation produces a series of emission spectra resulting from electron transitions to higher states and almost simultaneous reversions to lower states. There can be hundreds of diffracted spectra formed in the bright white light in the arc gap. Since the spectra relate to different elements, each species producing spectra of discrete, usually unique, wavelengths, this white light must be spread out into its constituent wavelengths. This occurs after part of the radiated light first passes through a narrow aperture or collimating slit and then onto the dispersing component (prism or grating) that accomplishes the spreading (next paragraph). The spread spectra are recorded as images of the slit; an image is reproduced for each wavelength involved. The resulting series of multiple images are extended one-dimensionally by the spreading onto unexposed photographic film (either in a long narrow strip or, in the thesis case, a 4 x 10 inch glass plate with a photo-emulsion coating on one side) which becomes exposed and then is developed to make visible the multiple images (in a film negative, appearing as thin black lines making up a spectrogram).

The spreading is accomplished by passing the light onto a glass (or quartz) prism (which bends the light because the glass refractive index varies with wavelength causing refractive separation) or (in my case) a diffraction grating (thousands of thin ruled parallel lines per linear centimeter on a glass or metal plate). Each given wavelength - representing some exitation state associated with the particular energy transition that an element attains under the burn conditions - is the main factor that determines how much the light is bent by refraction (prism) or diffraction (grating). The equation that describes the angle of spread (relative to the direction of the initial light beam reaching the dispersing prism) as a function of wavelength is: $n\lambda = b \sin\Theta$, where b is the width of the narrow collimating slit, Θ is the angle of spread, and n (0, 1, 2, 3,...) accounts for multiple orders of dispersal. For a diffraction grating b is replaced by d, the reciprocal of the number of diffraction lines (parallel to each other) per centimeter. Each diffracted line (an image of the narrow slit) can be indexed to a particular wavelength. The lines can be identified with multiple elements (different species) whose excited states from ionization give off characteristic ΔE's; this involves painstaking inspection of the recording film (often consuming hours). The appearance of the resulting spectrogram is one of a multitude of irregularly-spaced lines appearing dark (where the film was exposed) against a lighter background. Each line must be measured to obtain its Θ value which serves to determine the wavelength (and its causative element) in a lookup table. Below is a color version of spectrograms. The one at the top is for the spectrum of emitted solar radiation; others are for individual elements occurring in the Sun (the solar spectrum is repeated for reference). The brightness (or darkness, if a film negative is use) is

a function of line intensity that in turn depends on the amount of that element present.



A more modern strip chart record is shown below; the height of each line (which widens by dispersion) is a measure of the quantity of the element involved (in the above spectrogram, measurements of photographic density (relative brightness) accompishes this).



Present day emission spectrometers can record the data (the dispersed wavelengths and the intensity of the radiation associated with each) electronically (light sensors) and can digitize these discrete radiation values to facilitate automated species recognition and even amounts of each element (percent or parts per million [ppm]) present.

Returning to the Hydrogen diagrams, the situation for this element s modes of excitation is relatively "simple". For higher number atomic species, the transitions may be more complex and only approximate solutions to the Schroedinger equation (which can be used instead of the Planck equation to approach the energy changes differently) may result. Other factors enter into the determination of the energy level changes: the nature of the bonding, the coordination of the atoms in molecular or ionic compounds, the distribution of valency electrons in certain orbitals or conduction bands, et al. Without further pursuing these important considerations, we concentrate on this aspect: The transitions relate to three types of non-nuclear energy activity - electronic, vibrational, and rotational. Analysis of each type not only can identify the elements

present, but can reveal information about state of matter, crystal structure, atomic motions, etc.

Electronic energy transitions involve shifts of electrons from one quantum level to another according to precise rules. Any allowable transfer is determined from four quantities called the quantum numbers for that atomic system. These are (1) the principal quantum number n; (2) the angular momentum quantum number L; (3) the magnetic quantum number m; and (4), for polyelectronic atoms, the spin quantum number m_s. Electronic transitions occur within atoms existing in all familiar states of matter (solid; liquid; gas; plasma). Elements such as iron (Fe)(also Ni, Co, Cr, V and others) can undergo many transitions (yielding multiple wavelength lines in different parts of the spectrum), each influenced by valence state (degree of oxidation), location in the crystal structure, symmetry of atomic groups containing the element, and so forth. For compounds dominated by ionic bonding, the calculations of permissible transitions are influenced by crystal field theory. (Organic compounds, which usually are held together by covalent bonds, can be studied with a different approach.) Electronic transitions are especially prevalent in the UV, Visible, and Near IR parts of the spectrum.

Vibrational energy is associated with relative displacements between equilibrium center positions within diatomic and polyatomic molecules. These translational motions may be linear and unidirectional or more complex (vary within a 3-axis coordinate system). Specific transitions are produced by distortions of bonds between atoms, as described by such terms as stretching and bending modes. There is one fundamental energy level for a given vibrational transition, and a series of secondary vibrations or overtones at different, mathematically related frequencies (yielding the n orders mentioned above), as well as combination tones (composed of two or more superimposed fundamental or overtone frequencies). A tonal group of related frequencies comprises a *band*. Again, vibrational energy changes are characteristic of most states of matter (excluding the nucleus). Because these changes require less energy to initiate, the resulting ΔE's tend to occur at lower frequencies located at longer wavelengths in the Infrared and beyond (Microwave).

Rotational energy involves rotations of molecules, These take place only in molecules in the gaseous state and are described in terms of three orthogonal axes of rotation about a molecular center and by the moments of inertia determined by the atomic masses. This type of shift is relevant to the action of EMR on atomspheric gases. Being lower level energy transitions, the resulting emissions are at longer wavelengths. During excitation of gaseous molecules, both vibrational and rotational energy changes can occcur simultaneously.

The net energy detected as evidence of electronic, vibrational, and rotational transitions over a range of wavelengths (tied to one or multiple element species in the sample) is, in another sense, a function of the manner in which energy is partitioned between the EMR source and its interactions with the atoms in the material being analyzed and identified. The EMR (in remote sensing usually solar radiation) may be transmitted through any material experiencing this radiation, or absorbed within it, or reflected by atoms near the surface, or scattered by molecules or particulates composed of groups of atoms, or re-radiated through emission, or, as is common, by some combination of all these processes. There are three prevalent types of spectra associated with a material/object (m/o) being irradiated (as by sunlight): 1) absorption (or its inverse, transmission), 2) reflection, and 3) emission. For absorption spectra, the m/o lies between EMR source and the radiation detector; for reflection spectra, the source and detector are positioned outside of the reflecting surface of the m/o at an angle less than 180°; for emission spectra, the immediate source is located within the m/o although an external activating source (such as the temperature-changing electrical current in the carbon electrode spectrometer described above) is needed to initiate the emission.

Further information on Spectroscopy is presented on page 13-6 ff.

Primary Author: Nicholas M. Short, Sr.

The concepts of the electromagnetic spectrum and the role of the photon in transmitting energy are introduced; the variations in sine wave form in terms of frequency, and the basic equations relating EM energy to frequency are covered. Some commonly used radiometric terms and quantities are defined and cursorily explained. The technique of expressing powers of ten is explained in the footnote at the bottom of this page.

CAUTION: This page, and the special page I-2a accessed from it, are strongly theoretical and rather heavy going. Some background in Physics is helpful. But the subjects dealt with are important concepts and principles underlying the basis for remote sensing. Even if your knowledge of this kind of topics is limited, try to work through the text (re-reading is a good strategy) to build up familiarity with and understanding of the physical phenomena that describe what actually happens in remotely sensing an object or target.

Principles of Remote Sensing: The Photon and Radiometric Quantities

Most remote sensing texts begin by giving a survey of the main principles, to build a theoretical background, mainly in the physics of radiation. While it is important to have such a framework to pursue many aspects of remote sensing, we do not delve into this complex subject in much detail at this point. Instead, we offer on this and the next several pages an outline survey of the basics of relevant electromagnetic concepts. On this page, the nature of the photon is the prime topic. Photons of different energy values are distributed through what is called the **Electromagnetic Spectrum**. A full discussion of the electromagnetic spectrum (EMS) is deferred to page I-4.

Hereafter in this Introduction and in the Sections that follow, we limit the discussion and scenes examined to remote sensing products obtained almost exclusively by measurements within the Electromagnetic Spectrum (force field and acoustic remote sensing are briefly covered elsewhere in the Tutorial). Our emphasis is on pictures (photos) and images (either TV-like displays on screens or "photos" made from data initially acquired as electronic signals, rather than recorded directly on film). We concentrate mainly on images produced by sensors operating in the visible and near-IR segments of the electromagnetic spectrum (see the spectrum map on page I-4), but also inspect a fair number of images obtained by radar and thermal sensors.

The next several pages strive to summarize much of the underlying theory - mainly in terms of Physics - appropriate to Remote Sensing. The reader can gain most of the essential knowledge by working through those pages. In addition, optionally you can choose to read a reproduction of extracts from the Landsat Tutorial Workbook that deal with the theory by clicking onto page I-2a. Or, if you choose not to, read this next inserted paragraph which synopsizes key ideas from both the present and the I-2a pages:

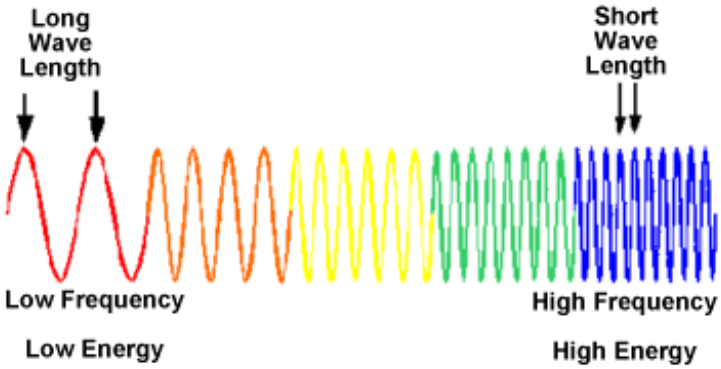
Synoptic Statement: The underlying basis for most remote sensing methods and systems is simply that of measuring the varying energy levels of a single entity, the fundamental unit in the electromagnetic (which may be abbreviated "EM") force field known as the *photon*. As you will see later on this page, variations in photon energies (expressed in Joules or ergs) are tied to the parameter *wavelength* or its inverse, *frequency*. EM radiation that varies from high to low energy levels comprises the *ElectroMagnetic spectrum* (EMS). Radiation from specific parts of the EM

spectrum contain photons of different wavelengths whose energy levels fall within a discrete range of values. When any target material is excited by internal processes or by interaction with incoming EM radiation, it will emit or reflect photons of varying wavelengths whose radiometric quantities differ at different wavelengths in a way diagnostic of the material. Photon energy received at detectors is commonly stated in power units such as Watts per square meter per wavelength unit. The plot of variation of power with wavelength gives rise to a specific pattern or curve that is the *spectral signature* for the substance or feature being sensed (discussed on page I-5).

Now, in more detail: The photon is the physical form of a quantum, the basic particle of energy studied in quantum mechanics (which deals with the physics of the very small, that is, particles and their behavior at atomic and subatomic levels). The photon is also described as the messenger particle for EM force or as the smallest bundle of light. This subatomic massless particle, which also does not carry an electric charge, comprises radiation *emitted* by matter when it is excited thermally, or by nuclear processes (fusion, fission), or by bombardment with other radiation (as well as by particle collisions). It also can become involved as *reflected* or *absorbed* radiation. Photons move at the speed of light: 299,792.46 km/sec (commonly rounded off to 300,000 km/sec or ~186,000 miles/sec).

Various aspects of the nature and behavior of photons are considered on this page but for a fuller treatment consult this Wikipedia website that deals with the nature of photons and the history of their discovery by Einstein as he first described them in a famous 1905 paper.

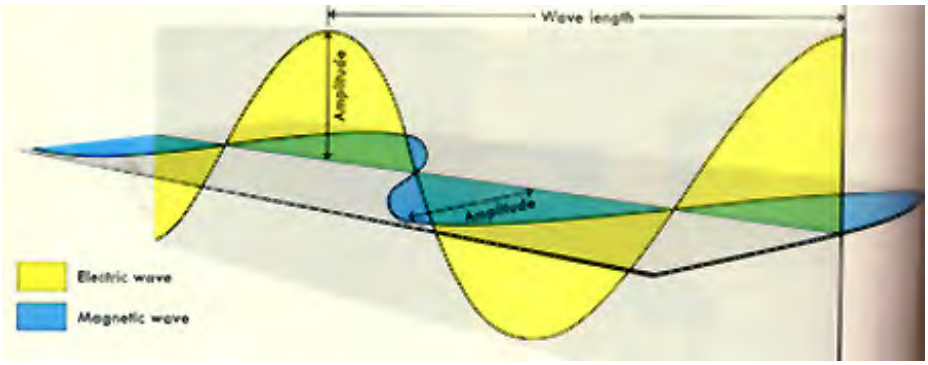
Photon particles also move as waves and hence, have a "dual" nature. These waves follow a pattern that can be described in terms of a sine (trigonometric) function, as shown in two dimensions in the figure below.



(NOTE: Frequency refers to number of crests of waves of same wavelength that pass by a point in one second.)

The distance between two adjacent peaks on a wave is its wavelength. The total number of peaks (top of the individual up-down curve) that pass by a reference lookpoint in a second is that wave's frequency (in units of cycles per second, whose SI version [SI stands for System International] is known as a Hertz [1 Hertz = $1/s^{-1}$]).

A photon travels as an EM wave having two components, oscillating as sine waves mutually at right angles, one consisting of the varying electric field, the other the varying magnetic field. Both have the same amplitudes (strengths) which reach their maxima-minima at the same time. Unlike other wave types which require a carrier (e.g., water waves), photon waves can transmit through a vacuum (such as in space). When photons pass from one medium to another, e.g., air to glass, their wave pathways are bent (follow new directions) and thus experience *refraction*.



A photon is said to be quantized, in that any given one possesses a certain quantity of energy. Some other photon can have a different energy value. Photons as quanta thus show a wide range of discrete energies. The amount of energy characterizing a photon is determined using Planck's general equation:

$$E = h\nu$$

where *h* is Planck's constant (6.6260... x 10^{-34} Joules-sec)* and *v* is the Greek letter, nu, representing frequency (the letter "*f*" is sometimes used instead of *v*). Photons traveling at higher frequencies are therefore more energetic. If a material under excitation experiences a change in energy level from a higher level E_2 to a lower level E_1, we restate the above formula as:

$$\Delta E = E_2 - E_1 = h\nu$$

where *v* has some discrete value determined by (v_2 - v_1). In other words, a particular energy change is characterized by producing emitted radiation (photons) at a specific frequency *v* and a corresponding wavelength at a value dependent on the magnitude of the change.**λ**.

I-4 Is there something wrong with the equation just above? ANSWER

Wavelength is the inverse of frequency (higher frequencies associate with shorter wavelengths; lower with longer), as given by the relationship:

$$c = \lambda\nu \text{ (transposable to } \nu = c/\lambda)$$

where *c is* the constant that expresses the speed of light, so that we can also write the Planck equation as

$$E = hc/\lambda$$

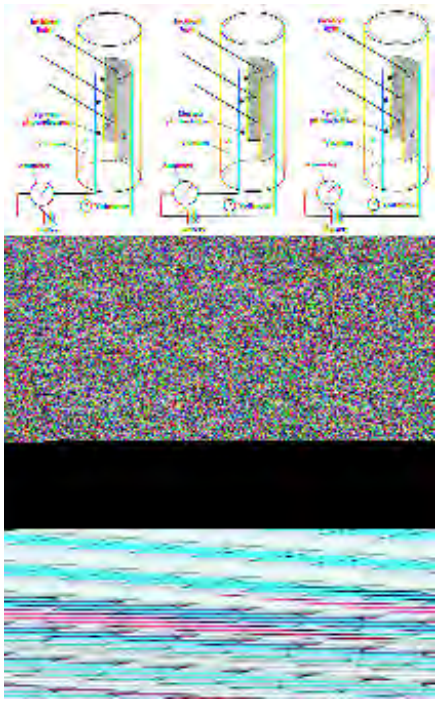
I-5 Come up with a very simple mnemonic phrase (one that helps your memory) for associating the energy level (amount of energy) with wavelength. ANSWER

I-6: Calculate the wavelength of a quantum of radiation whose photon energy is 2.10 x 10^{-19} Joules; use 3 x 10^8 m/sec as the speed of light c. ANSWER

I-7: **A radio station broadcasts at 120 MHz (megahertz or a million cycles/sec); what is the corresponding wavelength in meters (hint: convert MHz to units of Hertz). ANSWER**

A beam of radiation (such as from the Sun) is usually polychromatic (has photons of different energies); if only photons of one wavelength are involved the beam is monochromatic. The distribution of all photon energies over the range of observed frequencies is embodied in the term *spectrum* (a concept developed on the next page). A photon with some specific energy level occupies a position somewhere within this range, i.e., lies at some specific point in the spectrum

How are the photon energy levels in EM radiation quantified and measured? The answer lies in the discovery of the *photoelectric effect* made by Albert Einstein in 1905. Consider this diagram:



Einstein found that when light strikes a metal plate C, photoelectrons (negative charges) are ejected from its surface. In the vacuum those electrons will flow to a positively charged wire (unlike charges attract) that acts as a cathode. Their accumulation there produces an electric current which can be measured by an ammeter or voltmeter. The photoelectrons have kinetic energy whose maximum is determined by making the wire potential ever more negative (less positive) until at some value the current ceases. The magnitude of the current depends on the radiation frequencies involved and on the intensity of each frequency. From a quantum standpoint (Einstein's discovery helped to verify Planck's quantum hypothesis), the maximum kinetic energy is given by:

$$K.E. = hf + \varphi \text{ (the energy needed to free the electron)}$$

This equation indicates that the energy associated with the freed electron depends on the frequency (multiplied by the Planck constant h) of the photon that strikes the plate plus a threshold amount of energy required to release the electron (φ, the work function). By measuring the current, and if that work energy is known and other adjustments are made, the frequency of the photon can be determined. His experiments also revealed that regardless of the radiation intensity, photoelectrons are emitted only after a threshold frequency is exceeded. For frequencies below the threshold, no electron emission takes place; for those higher than the threshold value (exceeding the work function) the numbers of photoelectrons released re proportional to the number of incident photons (this number is given by the intensities involved.

When energies involved in processes from the molecular to subatomic level are involved (as in the photoelectric effect), these energies are measured in electron volt units ($1 \text{ eV} = 1.602 \times 10^{-19}$; this number relates to the charge on a single electron, as a fraction of the SI unit for charge quantity, the Coulomb [there are about 6.24×10^{18} electrons in one Coulomb).

Astute readers may have recognized the photoelectric effect as being involved in the operation of vacuum tubes in early radio sets. In remote sensing, the sensors used contain detectors that produce currents (and voltages, remember V = IR) whose quantities for any given frequency depend on the photoelectric effect.

From what's been covered so far on this page, let's modify the definition of remote sensing (previous page) to make it "quantum phenomenological". In this approach, *electromagnetic*

remote sensing involves the detection of photons of varying energies coming from the target (after the photons are generated by selective reflectance or by internal emittance from the target material(s)) by passing them through frequency (wavelength)- dependent dispersing devices (filters; prisms) onto metals or metallic compounds/alloys which undergo photoelectric responses to produce currents that become signals that can be analyzed in terms of energy-dependent parameters (frequency, intensity, etc.) whose variations are controlled by the atomic level composition of the targets. The spectral (wavelength) dependency of each material is diagnostic of its unique nature. When these photon variations are plotted in X-Y space, the shapes of the varied distributions of photon levels produce patterns that further aid in identifying each material (with likely regrouping into a *class* or *physical feature*.

There is much more to the above than the brief exposition and summary given. Read the next page for more elaboration. Consult a physics text for more information. Or, for those with some physics background, read the Chapter on *The Nature of Electromagnetic Radiation* in the **Manual of Remote Sensing**, 2nd Ed., published by the American Society of Photogrammetry and Remote Sensing (ASPRS). This last source contains a summary table that lists and defines what can be called basic radiometric quantities but the print is too small to reproduce on this page. The following is an alternate table which should be legible on most computer screens.

Quantity	Symbol	Unit	Definition
Radiant energy	Q	joule (J)	The total energy in a radiation field or the total energy delivered by such a radiation field
Radiant power, radiant flux	P	watt (W)	The rate at which radiant energy is transferred from one region to another by the radiation field. $P = dQ/dt$, where t is time
Radiant exposure	H	J m^{-2}	Energy per unit area received by a real or imaginary surface element. $H = dQ/dA$, where A is area
Radiant energy density	w	J m^{-3}	The radiant energy per unit volume of the radiation field. $w = dQ/dV$, where V is volume
Radiant energy fluence rate, spherical irradiance	ϕ	W m^{-2}	At a given point in space, the radiant power incident on a small sphere divided by the cross-sectional area of that sphere. $\phi = dP/dA$, where A is the cross section of the sphere
Radiant exitance	M	W m^{-2}	The flux per unit area leaving the surface of a source of radiation. $M = dP/dA$, where dA is the area of the source element
Irradiance	E	W m^{-2}	The flux per unit area received by a real or imaginary surface element. $E = dP/dA$, where dA is the area of the surface element
Radiant intensity	I	W sr^{-1}	The flux per unit solid angle emitted by a source in a given direction. $I = dP/d\Omega$, where Ω is the solid angle
Radiance	L	W m^{-2} sr^{-1}	The flux per unit projected area per unit solid angle leaving a source or reference surface. $L = d^2P/dA_{proj}d\Omega$. The projected area is given by $dA_{proj} = dA\cos\theta$, where θ is the angle between the outward surface normal of the area element dA and the direction of observation

We have extracted from Chapter 2 of the Manual of Remote Sensing the following useful information that explains some of the radiometric terminology and the concepts they represent as used by specialists in the remote sensing field:

Radiant energy (Q), transferred as photons moving in a radiation stream, is said to emanate in minutely short bursts (comprising a wave train) from a source in an excited state. This stream of photons moves along lines of flow (also called rays) as a *flux* (φ) which is defined as the time rate at which the energy Q passes a spatial reference (in calculus terms:φ = dQ/dt). The energy involved is capable of doing work. The SI units of work are Joules per second (alternately expressed in ergs, which equal 10^{-7} Joules). The flux concept is related to *power*, defined as the time rate of doing work or expending energy (1 J/sec = 1 Watt, the unit of power). The nature of the work can be one, or a combination, of these: changes in motion of particles acted upon by force fields; heating; physical or chemical change of state. Depending on circumstances, the energy spreading from a point source may be limited to a specific direction (a beam) or can disperse in all directions.

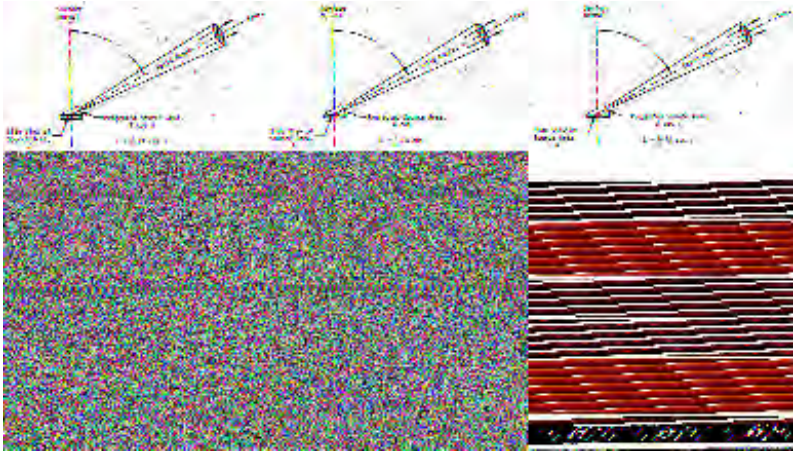
Radiant flux density is just the energy per unit volume (cubic meters or cubic centimeters). The

flux density is proportional to the squares of the amplitudes of the component waves. Flux density as applied to radiation coming from an external source to the surface of a body is referred to as irradiance (E); if the flux comes out of that body, it's nomenclature is exitance (M) (see below for a further description).

. The *intensity of radiation* is defined as the power P that flows through unit area A ($I = P/A$); power itself is given by $P = \Delta E/\Delta t$ (the rate at which radiant energy is generated or is received). The energy of an EM wave (sine wave) depends on the square of its amplitude (height of wave in the x direction; see wave illustration above); thus, doubling the amplitude increases the power by 4. Another formulation of radiant intensity is given by the radiant flux per unit of solid angle ω (in steradians - a cone angle in which the unit is a radian or 57 degrees, 17 minutes, 44 seconds); this diagram may help to visualize this:

Thus, for a surface at a distance R from a point source, the radiant intensity I is the flux Φ flowing through a cone of solid angle ω on to the circular area A at that distance, and is given by $I = \Phi/(A/R^2)$. Note that the radiation is moving in some direction or pathway relative to a reference line as defined by the angle θ.

From this is derived a fundamental EM radiation entity known as *radiance* (commonly noted as "L"). In the ASPRS Manual of Remote Sensing, "radiance is defined as the radiant flux per unit solid angle leaving an extended source (of area A) in a given direction per unit projected surface area in that direction." This diagram, from that Manual, visualizes the terms and angles involved:



As stated mathematically, L = Watt · m-2 · sr-1; where the Watt term is the radiant flux (power, or energy flowing through the reference surface area of the source [square meters] per unit time), and "sr" is a solid angle Ω given as 1 steradian. From this can be derived $L = \Phi/\Omega$ times $1/A\cos\theta$, where θ is the angle formed by a line normal to the surface A and the direction of

radiant flow. Or, restated with intensity specified, $L = I/A\cos\theta$. Radiance is loosely related to the concept of *brightness* as associated with luminous bodies. What really is measured by remote sensing detectors are radiances at different wavelengths leaving extended areas (which can "shrink" to point sources under certain conditions). When a specific wavelength, or continuum of wavelengths (range) is being considered, then the radiance term becomes L_λ.

In practical use, the radiance measured by a sensor operating above the ground is given by:

$$L_{tot} = \rho E T/\pi + L_p$$

where L_{tot} is the total spectral radiance (all wavelength) received by the sensor; ρ is the reflectance from the ground object being sensed; E is the irradiance (incoming radiant energy acting on the object); T is an atmospheric transmission function; and L_p is radiance from the atmospheric path itself.

Radiant fluxes that come out of sources (internal origin) are referred to as *radiant exitance* (M) or sometimes as "emittance" (now obsolete). Radiant fluxes that reach or "shine upon" any surface (external origin) are called *irradiance*. Thus, the Sun, a source, irradiates the Earth's atmosphere and surface.

The above radiometric quantities Q, φ, I, E, L, and M, apply to the entire EM spectrum. Most wave trains are polychromatic, meaning that they consist of numerous sinusoidal components waves of different frequencies. The bundle of varying frequencies (either continuous within the spectral range involved or a mix of discrete but discontinuous monochromatic frequencies [wavelengths]) constitutes a complex or composite wave. Any complex wave can be broken into its components by Fourier Analysis which extracts a series of simple harmonic sinusoidal waves each with a characteristic frequency, amplitude, and phase. The radiometric parameters listed in the first sentence can be specified for any given wavelength or wavelength interval (range); this spectral radiometric quantity (which has a value different from those of any total flux of which they are a part [unless the flux is monochromatic]) is recognized by the addition to the term of a subscript λ, as in L_λ and Q_λ. This subscript denotes the specificity of the radiation as at a particular wavelength. When the wavelengths being considered are confined to the visual range of human eyes (0.4 to 0.7 μm), the term "luminous" precedes the quantities and their symbols are presented with the subscript "v", as Φ_v for a luminous flux.

EM radiation can be *incoherent* or *coherent*. Waves whose amplitudes are irregular or randomly related are incoherent; polychromatic light fits this state. If two waves of different wavelengths can be combined so as to develop a regular, systematic relationship between their amplitudes, they are said to be coherent; monochromatic light generated in lasers meet this condition.

The above, rather abstract, sets of ideas and terminology is important to the theorist. We include this synopsis mainly to familiarize you with these radiometric quantities in the event you encounter them in other reading.

---

* The Powers of 10 Method of Handling Numbers: The numbers 10^{-34} (incredibly small) or 10^{12} (very large - a trillion), as examples, are a shorthand notation that conveniently expresses very large and very small numbers without writing all of the zeros involved. Using this notation allows one to simplify any number other than 10 or its multiples by breaking the number into two parts: the first part denotes the number in terms of the digits that are specified, as a decimal value, e.g., 5.396033 (through the range 1 to 10); the second part of the number consists of the base 10 raised to some power and tells the number of places to shift the decimal point to the right or the left. One multiplies the first part of the number by the power of ten in the second part of the number to get its value. Thus, if the second part is 10^7, then its stretched out value is 10,000,000 (7 zeros) and when 5.396033 is multiplied by that value, it becomes 53,960,330. Considering the second part of the number, values are assigned to the number 10^n where n can be any positive or negative whole integer. A +n indicates the number of zeros that follow the number 10, thus for n = 3, the value of 10^3 is 1 followed by three zeros, or 1000 (this is the same as the cube of 10). The number 10^6 is 1000000, i.e., a 1 followed by six zeros to its right (note: $10^0 = 1$). Thus, 10^{60} represents

1,000,000,000,000,000... (a total of 20 "000"s) out to 60 such zeros. Likewise, for the -n case, 10^{-3} (where n = -3) is equal to 0.001, equivalent to the fraction 1/1000, in which there are two zeros to the left of 1 (or three places to the right of 1 to encounter the decimal point). Here the rule is that there is always one less zero than the power number, as located between the decimal point and 1. Thus, 10^{-6} is evaluated as 0.000001 and the number of zeros is 5; (10^{-1} is 0.1 and has no zero between the . and 1). In this special case, 10^{0} is reserved for the number 1. Any number can be represented as the product of its decimal expression between 1 and 10 (e.g., 3.479) and the appropriate power of 10, (10^{n}). Thus, we restate 8345 as 8.345 x 10^{3}; the number 0.00469 is given as 4.69 x 10^{-3}.

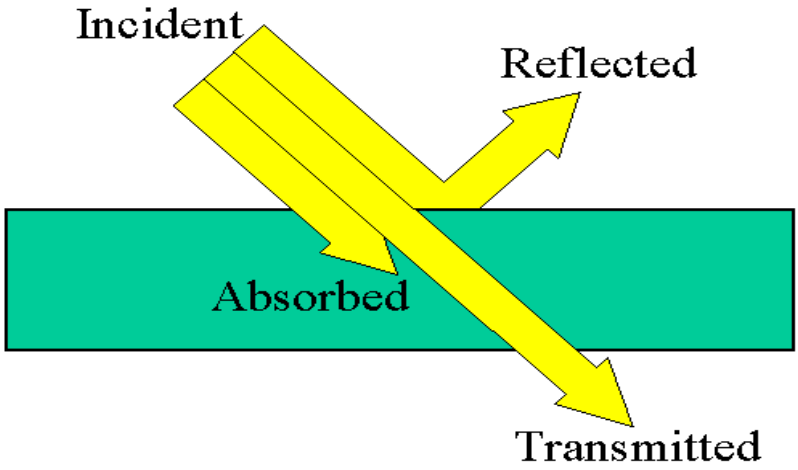
PREVIOUS rst NEXT

Primary Author: Nicholas M. Short, Sr.

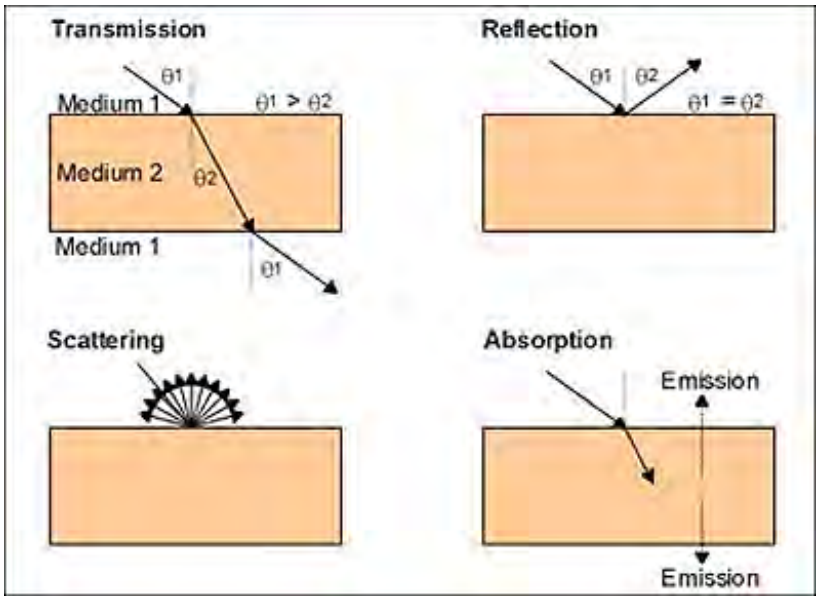
This page concentrates on what can happen to solar radiation as it passes through the atmosphere and hits objects/materials at the surface of Earth (or any planetary body); typically some of the irradiance is absorbed by the atmosphere and a fraction of the remaining radiation may be absorbed or reflected by surface features, or occasionally a large part is transmitted by penetrating the transparent/translucent features to varying depths (as occurs when light enters water). The effects of specular and diffuse surfaces on the radiometric values measured are examined.

Electromagnetic Spectrum: Transmittance, Absorptance, and Reflectance

When a beam of light impinges on a material, any one of three, or a combination thereof, changes or responses will occur:

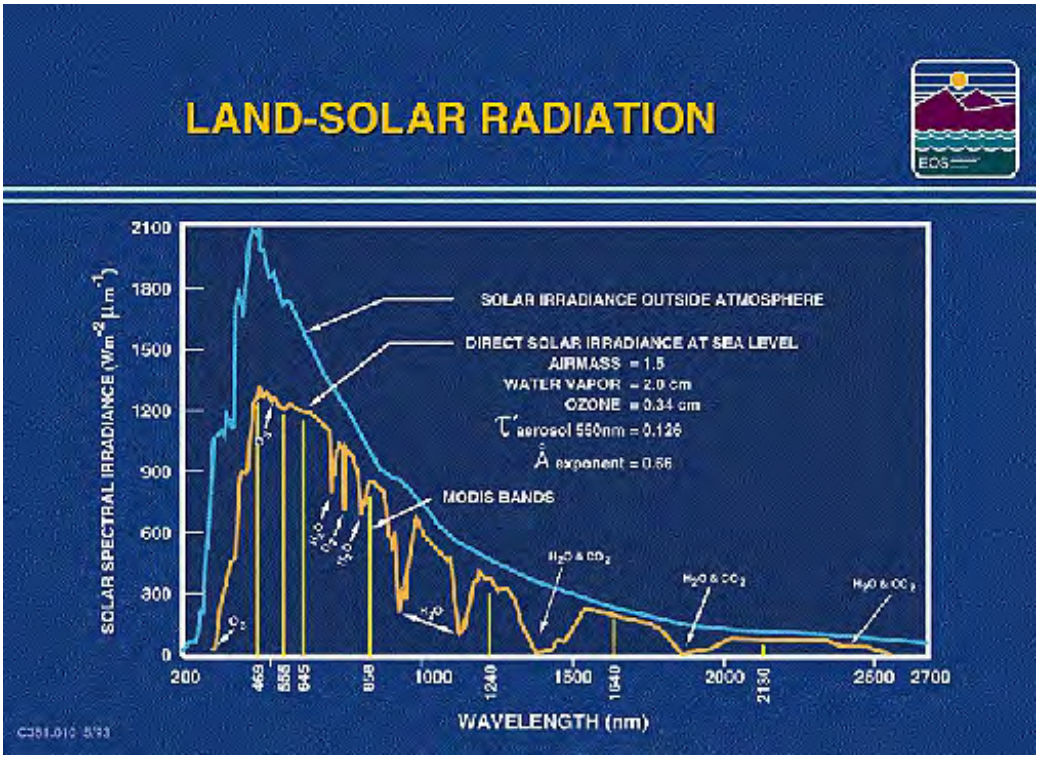


In more detail, these are the possible fates: Any beam of photons from some source passing through medium 1 (usually air) that impinges upon an object or target (medium 2) will experience one or more reactions that are summarized in this diagram:



Some objects are capable of transmitting the light through their bulk without significant diminution (note how the beam bends twice at the medium 1/medium 2 interface but emerges at the same angle as entry). Other materials cause the light energy to be absorbed (and in part emitted as longer wavelength radiation). Or, the light can be reflected at the same angle as it formed on approach. Commonly the nature of the object's surface (owing to microscopic roughness) causes it to be scattered in all directions.

The primary source of energy that illuminates natural targets is the Sun. Solar irradiation (also called insolation) arrives at Earth at wavelengths which are determined by the photospheric temperature of the sun (peaking near 5600 °C). The main wavelength interval is between 200 and 3400 nm (0.2 and 3.4 µm), with the maximum power input close to 480 nm (0.48 µm), which is in the visible green region. As solar rays arrive at the Earth, the atmosphere absorbs or backscatters a fraction of them and transmits the remainder.



Upon striking the land and ocean surface (and objects thereon), and atmospheric targets, such as air, moisture, and clouds, the incoming radiation (irradiance) partitions into three modes of energy-interaction response:

(1) Transmittance (τ) - some fraction (up to 100%) of the radiation penetrates into certain

surface materials such as water and if the material is transparent and thin in one dimension, normally passes through, generally with some diminution.

(2) Absorptance (α) - some radiation is absorbed through electron or molecular reactions within the medium ; a portion of this energy is then re-emitted, usually at longer wavelengths, and some of it remains and heats the target;

(3) Reflectance (ρ) - some radiation (commonly 100%) reflects (moves away from the target) at specific angles and/or scatters away from the target at various angles, depending on the surface roughness and the angle of incidence of the rays.

Because they involve ratios (to irradiance), these three parameters are dimensionless numbers (between 0 and 1), but are commonly expressed as percentages. Following the Law of Conservation of Energy: $\tau + \alpha + \rho = 1$.

A fourth situation, when the emitted radiation results from internal atomic/molecular excitation, usually related to the heat state of a body, is a thermal process. The theory underlying thermal remote sensing is treated in Section 9.

When a remote sensing instrument has a line-of-sight with an object that is reflecting solar energy, then the instrument collects that reflected energy and records the observation. *Most remote sensing systems are designed to collect reflected radiation*.

I-8 From the above graph, calculate (approximately) the percent decrease in surface irradiance (drop in power, in Watts/meter^2/μm) of maximum solar radiation (close to 500 nanometers) from the moment it reaches the outer atmosphere until it reaches the Earth's surface; assume a vertical rather than slant path through the atmosphere. ANSWER

There are two general types of reflecting surfaces that interact with EMR: specular (smooth) and diffuse (rough). These terms are defined geometrically, not physically. A surface may appear to be smooth in a physical sense, i.e., it appears and feels smooth, but at a scale on the order of wavelengths of light, many irregularities might occur throughout that surface. (A concrete roadway may appear smooth and flat from a distance but feels rough when a finger passes over it, owing to small grooves, pits, and protuberances.) Radiation impinging on a diffuse surface tends to be reflected in many directions (scattered). The Rayleigh criterion is used to determine surface roughness with respect to radiation:

$$h \text{ is less than or equal to wavelength } \lambda/8\cos\theta$$

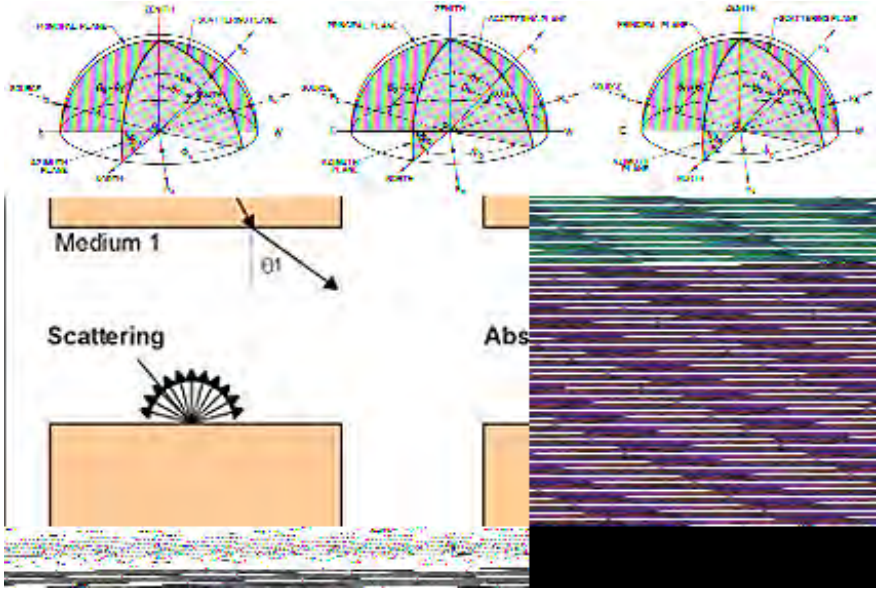
where h is the surface irregularity height (measured in Angstroms), λ is the wavelength (also in Angstroms) and θ is the angle of incidence (measured from the normal [perpendicular] to the surface). If λ is less than h, the surface acts as a diffuse reflector; if greater than h, the surface is specular.

A specular surface (for example, a glass mirror) reflects radiation according to Snell's Law which states that the angle of incidence θ_i is equal to the angle of reflectance θ_r (where the light ray moves in the principal plane that passes normal to the surface). Actual values (e.g., radiances) of specular reflected radiation depend on the type of material making up the specular surface. Specular reflectances within the visible wavelength range vary from as high as 0.99 (%) for a very good mirror to as low as 0.02-0.04 (%) for a very smooth water surface.

In general, natural surfaces are almost always diffuse and depart significantly from specular at shorter wavelengths (into the infrared) and may still be somewhat diffuse in the microwave region.

The behavior of a perfectly diffuse, or Lambertian, surface is described with the help of this

figure:

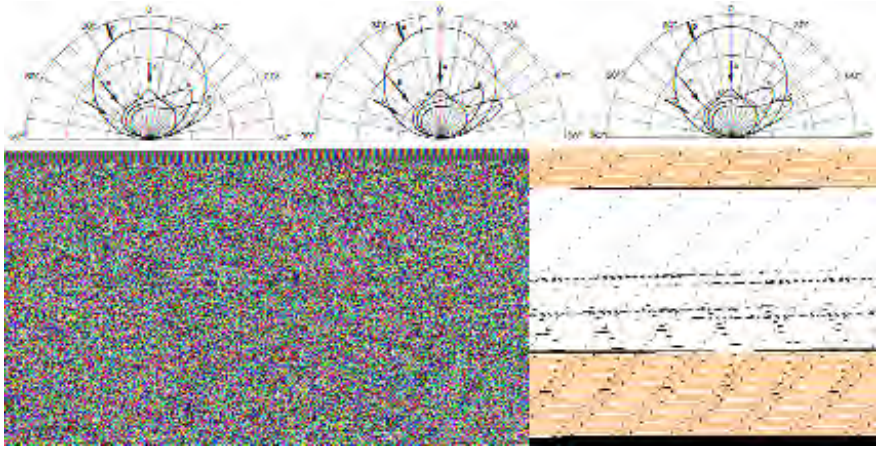


Consider a bundle of rays (making up the radiation flux) from a single distant source position at an incident angle θ_I (relative to the zenith direction) and an azimuth angle φ_0 (relative to north). Imagine the (arbitrarily) horizontal plane of a target being irradiated to be enclosed in a hemisphere (this simplifies some calculations because it allows polar coordinates to be used [which we will ignore here]). For the wavelength conditions we set, the surface is considered rough or diffuse and has irregularities where the surface departs from horizontal, at varying slopes. A given ray R_I located in the principal plane now strikes the surface at a point Q_1. It will be reflected according to the position of the minute surface at Q depending on its slope. If that surface is horizontal (slope = 0°), the ray moves away along path R_R in the principal plane as though this geometry is specular. But if the point surface Q_s is non-horizontal, i.e., has varying slopes defining the shape of the irregularity, the ray (mathematically treatable as a vector) will now move out along some direction R_D through its scattering plane whose position is defined by θ_D and φ_D. At other points (Q_n) on the surface, the direction of the outgoing R will differ according to the orientation of the slope at the immediate irregularity. Thus, a large number of incoming rays meeting the surface at other irregularities (most probably with randomly oriented slopes) will be redirected (diverge) in all possible directions extending through the hemisphere of reference.

The radiance in any one direction is, on average, the same as any other; in other words, radiance is constant at any viewing position on the hemishpere and is therefore independent of θ_0. However, the radiant intensity at any position will vary according to the relation $I_\theta = I_0 \cos\theta$. This states that as the angle of incident radiation I_θ is varied, the intensity of outgoing radiation also changes. For normal incidence (from the zenith), θ is 0 and cosθ is 1, so $I_\theta = I_0$. For all other angles cosθ is less than 1 and I_0 is reduced. Although a homogeneous, non-variant surface viewed from any position will seem to be uniformly illuminated (constant radiance), that surface will become less bright as the source is moved from a vertical (overhead) position towards the plane itself (horizon).

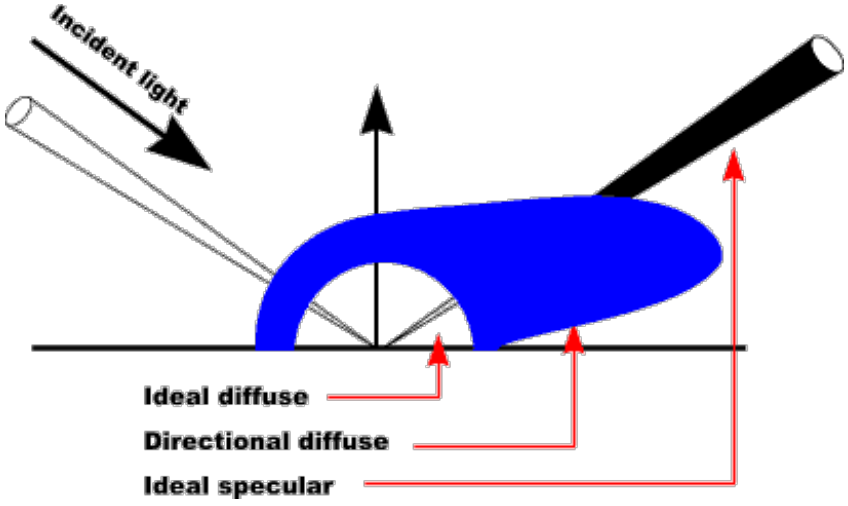
The term *bidirectional reflectance* describes the common observational condition in remote sensing in which the viewing angle φ differs from the angle θ of rays incident on a diffuse surface, and incoming/outgoing rays are not in the same principal plane (different azimuths). Thus, reflectances from the same target (type) change in value from various combinations of θ and φ: this is particularly important when the sensor operates off-nadir (looking sidewards) and the Sun angle and azimuth vary during the period of operation (as occurs when an aircraft moves back and forth along flight paths during an aerial photography mission). This is the

common state for most remotely sensed surfaces: they are essentially Lambertian rather than Specular and hence are sensitive to directions of illumination and of viewing. Consider this diagram (treating the two dimenionsal case; the behavior can be displayed in the third dimension using a hemisphere as reference):

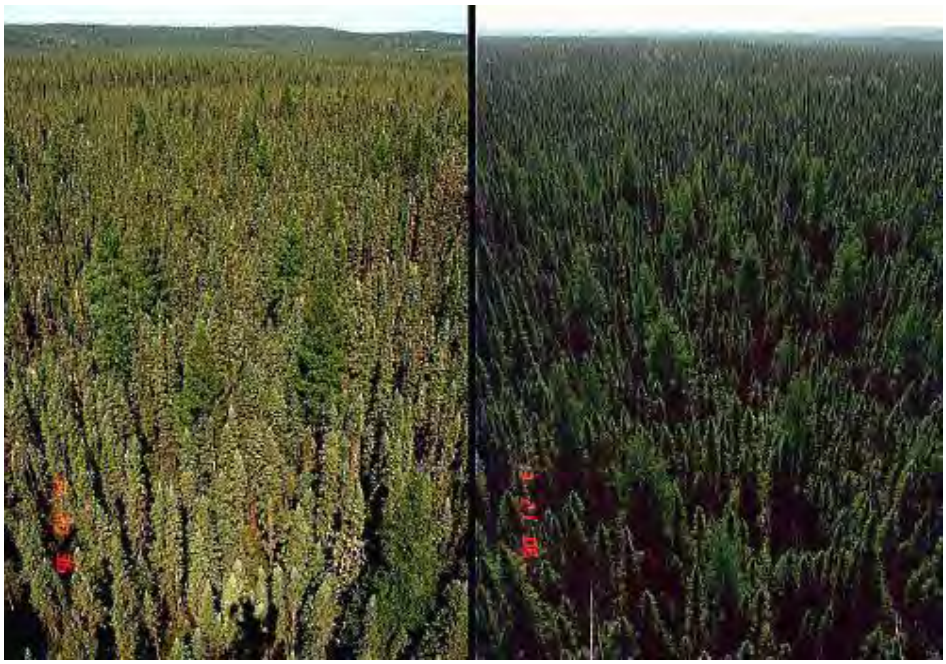


For an imperfectly diffuse reflector (having a specular component) and a viewer directly overhead (such as a sensor looking straight down normal to the Earth's surface), scattering will produce three-dimensional envelopes (shown here as a two-dimensional slice) of reflectances deriving from rays A, B, and C. These radiances vary in intensity in an asymmetric way (except for A). Thus, for B and C rays, there is general diffuse reflectance in all directions plus a "spike" or differential increase in directions spread around the angle of the specular component. Since reflectances also vary with wavelength, envelopes of bidirectional reflectance must be calculated for each wavelength in the range of λs considered; this is commonly not done in routine image analysis but is often necessary when a quantitative study involving field measurements is undertaken.

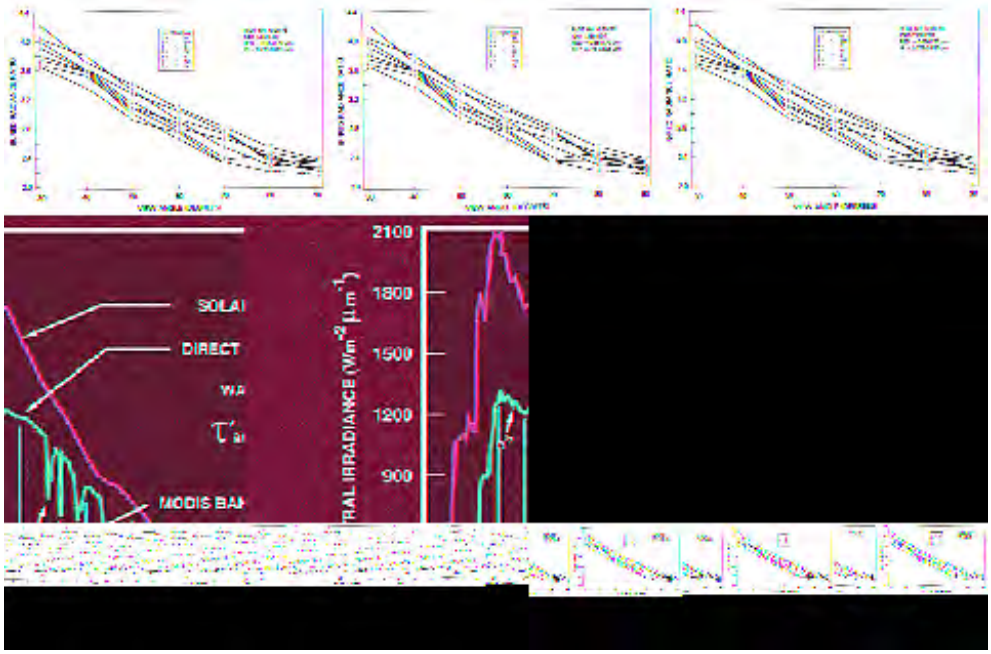
The next diagram is a variant of the one above. It shows the spread of incoming light for the ideal diffuse and specular cases, as well as the (bi-)directional case that describes most light scattering surfaces.



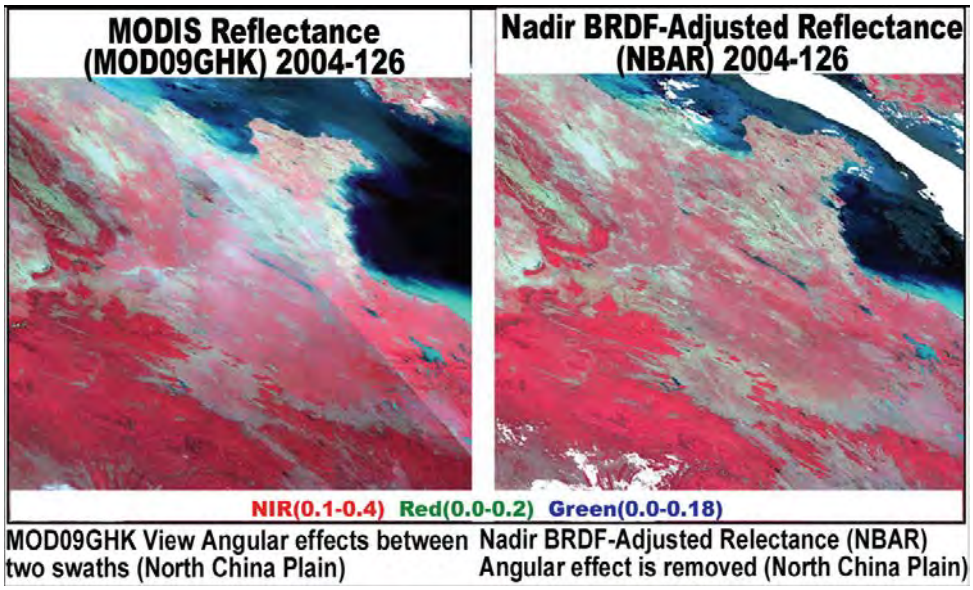
The reader no doubt has experienced bidirectional reflectance effects in which parts of a scene as viewed seem to get darker or even change color shades. The next figure illustrates an extreme example in which the same section of spruce forest is illuminated from behind the viewer (left) and from in front (right; Sun [not shown] is above the distant horizon.



The plots in the next chart show how reflectance (expressed as a ratio of near infrared to red radiances) vary with viewing angle (abscissa) and compass direction (family of plots). The conclusion: the effect of bidirectional reflectance can be significant (see page 13-4 for further discussion).



The influence of bidirectional reflectance is such that in a large area scene, in particular, its effects vary with distance out from the scene center. In principle, corrections should be made for the effect when a scene is classified since the radiances will change somewhat in different parts of the scene owing to angular shifts in viewing direction. In practice, such corrections are often not made, although correction programs can be written to account for the bidirectional variations. The writer searched the Internet looking for examples of scenes before and after correction and came up almost empty; the scene below is the only one found:



PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

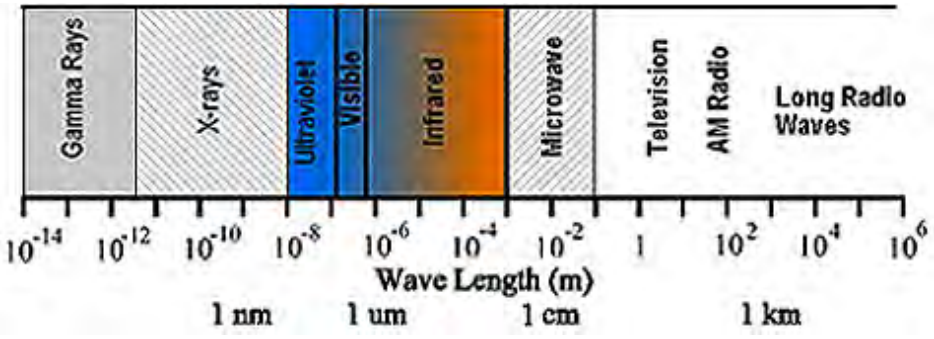
The specific regions or subdivisions of the electromagnetic spectrum are named and plotted. Mechanisms for generation of electromagnetic radiation are reviewed. The idea of incoming solar radiation or irradiance and its interaction (relative absorption) with the atmosphere is reviewed, and the notion of multispectral remote sensing over different spectral intervals is illustrated with a set of astronomical examples.

Electromagnetic Spectrum: Distribution of Radiant Energies

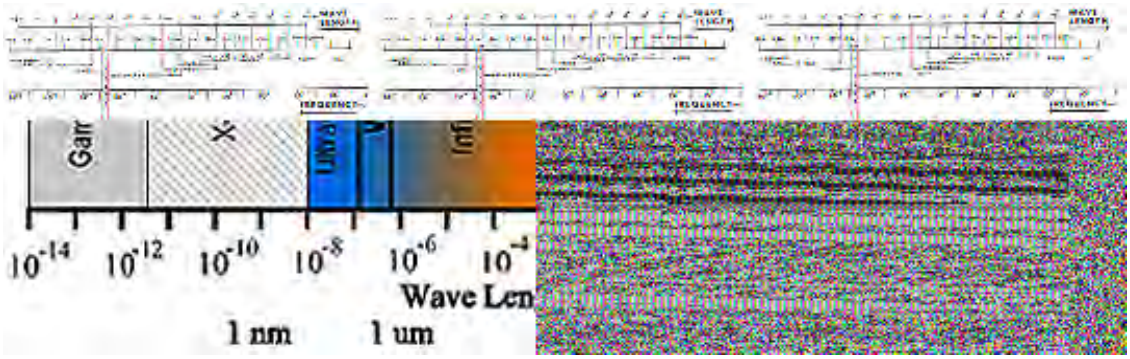
As noted on the previous page, electromagnetic radiation (EMR) extends over a wide range of energies and wavelengths (frequencies). A narrow range of EMR extending from 0.4 to 0.7 µm, the interval detected by the human eye, is known as the *visible region* (also referred to as *light* but physicists often use that term to include radiation beyond the visible). White light contains a mix of all wavelengths in the visible region. It was Sir Isaac Newton who first in 1666 carried out an experiment that showed visible light to be a continuous sequence of wavelengths that represented the different color the eye can see. He passed white light through a glass prism and got this result:

The principle supporting this result is that as radiation passes from one medium to another, it is bent according to a number called the index of refraction. This index is dependent on wavelength, so that the angle of bending varies systematically from red (longer wavelength; lower frequency) to blue (shorter wavelength; higher frequency). The process of separating the constituent colors in white light is known as *dispersion*. These phenomena also apply to radiation of wavelengths outside the visible (e.g., a crystal's atomic lattice serves as a diffraction device that bends x-rays in different directions).

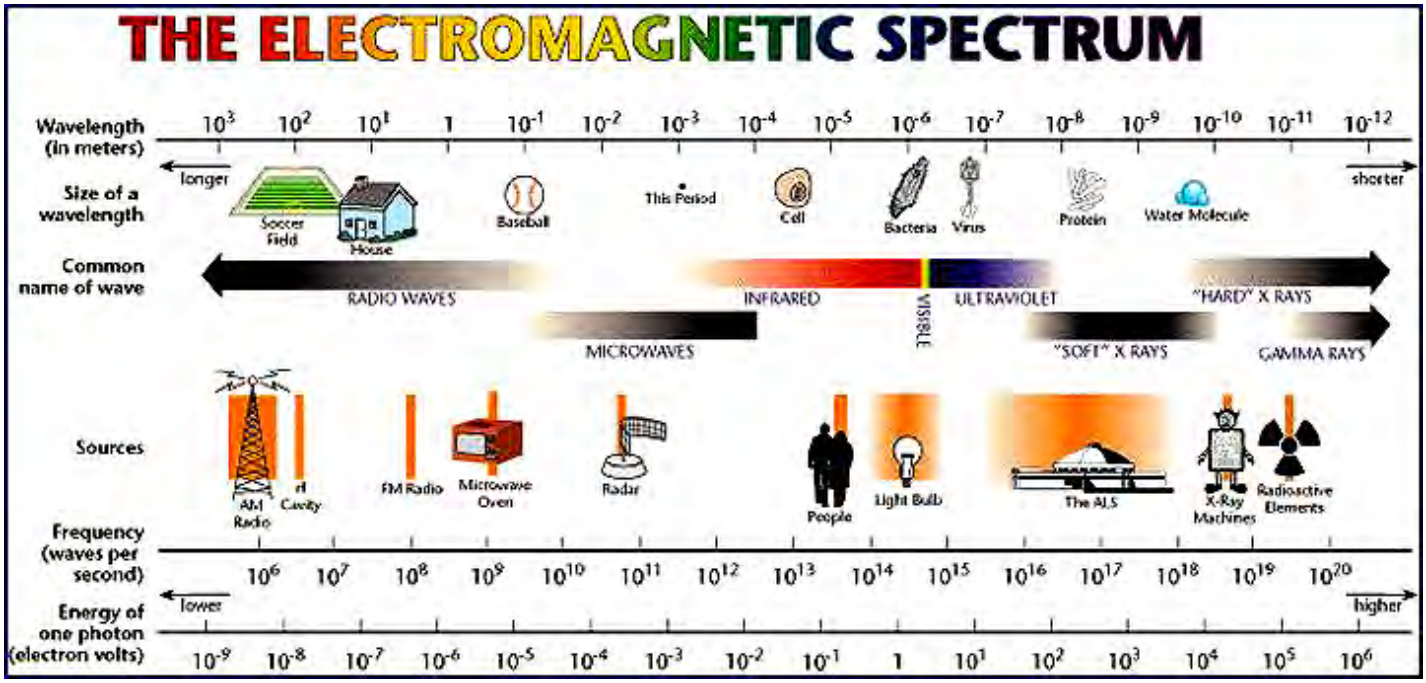
The distribution of the continuum of all radiant energies can be plotted either as a function of wavelength or of frequency in a chart known as the electromagnetic (EM) spectrum. Using spectroscopes and other radiation detection instruments, over the years scientists have arbitrarily divided the EM spectrum into regions or intervals and applied descriptive names to them.The EM spectrum, plotted here in terms of wavelengths, is shown here.



Beneath is a composite illustration taken from the Landsat Tutorial Workbook (credited there to Lintz and Simonett, *Remote Sensing of the Environment*, who identify it as a modification of an earlier diagram by Robt. Colwell) that shows in its upper diagram the named spectral regions in terms of wavelength and frequency and in the lower diagram the physical phenomena that give rise to these radiation types and the instruments (sensors) used to detect the radiation.

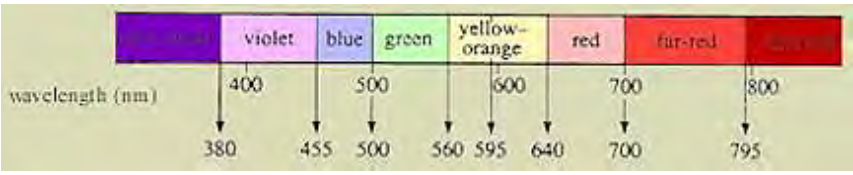


Although it is somewhat redundant, we reproduce here still another plot of the EM Spectrum, with added items that are self-explanatory:



Colors in visible light are familiar to most, but the wavelength limits for each major color are probably not known to most readers. Here is a diagram that specifies these limits (the purple on

the far left is in the non-visible ultraviolet; the deep red on the far right is the beginning of the infrared). The human eye is said to be able to distinguish thousands of slightly different colors (one estimate placed this at distinguishable 20000 color tints).



Different names for *(wave)length units* within intervals (those specified by types) that subdivide the EM spectrum, and based on the metric system, have been adopted by physicists as shown in this table:

Type of EM Wave	Typical Unit of Measure
radio	meter (m) centimeter (cm)=0.01 m
microwave (radar)	millimeter (mm)=0.001 m
infrared	micrometer (um)=10^{-6}m
visible	nanometer (nm)=10^{-9}m; 10^{-3}um
ultraviolet	angstrom (Å)=10^{-10}m

(Both in this Tutorial and in other texts, just which units are chosen can be somewhat arbitrary, i.e., the authors may elect to use micrometers or nanometers for a spectral location in the visible. Thus, as an example, 5000 Angstroms, 500 nanometers, and 0.5 micrometers all refer to the same specific wavelength; see next paragraph.)

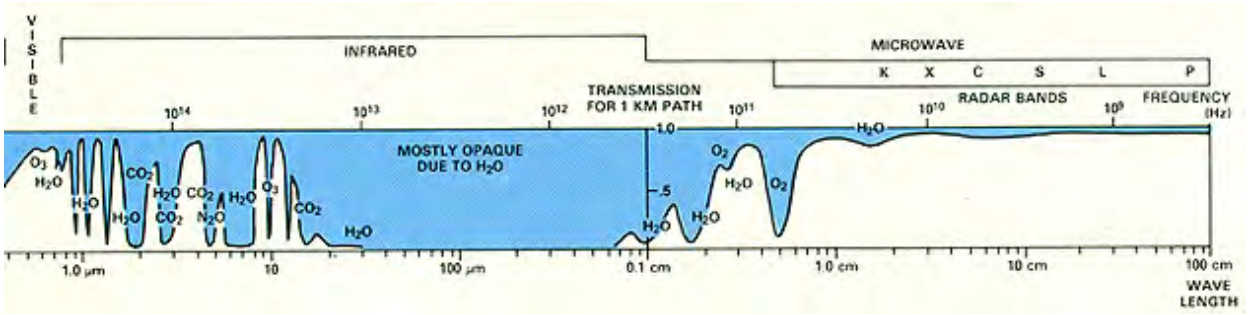
At the very energetic (high frequency and short wavelength) end are gamma rays and x-rays (whose wavelengths are normally measured in angstroms [Å], which in the metric scale are in units of 10^{-8} cm). Radiation in the ultraviolet extends from about 300 Å to about 4000 Å. It is convenient to measure the mid-regions of the spectrum in one of two units: micrometers (µm), which are multiples of 10^{-6} m or nanometers (nm), based on 10^{-9} m. The visible region occupies the range between 0.4 and 0.7 µm, or its equivalents of 4000 to 7000 Å or 400 to 700 NM The infrared region, spanning between 0.7 and 1000 µm (or 1 mm), has four subintervals of special interest: (1) reflected IR (0.7 - 3.0 µm), and (2) its film responsive subset, the photographic IR (0.7 - 0.9 µm); (3) and (4) thermal bands at (3 - 5 µm) and (8 - 14 µm). We measure longer wavelength intervals in units ranging from mm to cm. to meters. The microwave region spreads across 0.1 to 100 cm, which includes all of the interval used by radar systems. These systems generate their own active radiation and direct it towards targets of interest. The lowest frequency-longest wavelength region beyond 100 cm is the realm of radio bands, from VHF (very high frequency) to ELF (extremely low frequency); units applied to this region is often stated as frequencies in units of Hertz (1 Hz = 1 cycle per second; KHz, MHz and GHz are kilo-, mega-, and giga- Hertz respectively). Within any region, a collection of continuous wavelengths can be partioned into discrete intervals called *bands*.

I-9: **Given that 1 nanometer (NM) = 10^{-9} m, 1 micrometers = 10^{-6} m and 1 Angstrom (A) = 10^{-10} m, how many nanometers in a micrometer; how many Angstrom units in a micrometer? ANSWER**

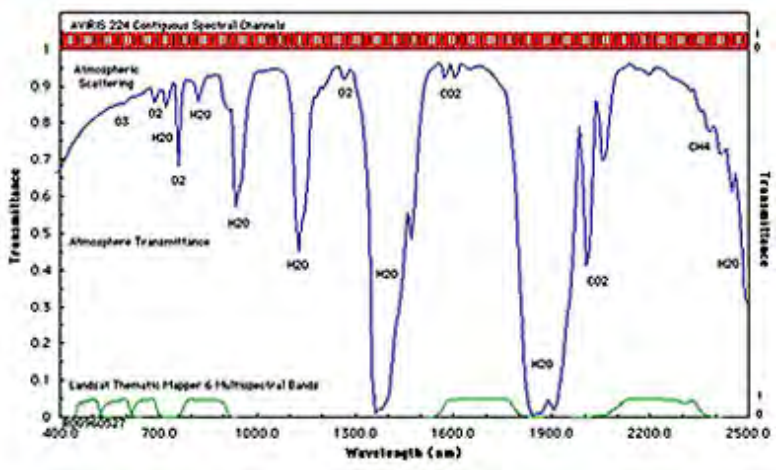
Referring to the Phenomenology diagram (fourth illustration above): That chart indicates many of the atomic or molecular mechanisms for forming these different types of radiation; it also depicts the spectral ranges covered by many of the detector systems in common use. This

diagram indicates that electromagnetic radiation is produced in a variety of ways. Most involve actions within the electronic structure of atoms or in movements of atoms within molecular structures (as affected by the type of bonding). One common mechanism is to excite an atom by heating or by electron bombardment which causes electrons in specific orbital shells to momentarily move to higher energy levels; upon dropping back to the original shell the energy gained is emitted as radiation of discrete wavelengths. At high energies even the atom itself can be dissociated, releasing photons of short wavelengths. And photons themselves, in an irradiation mode, are capable of causing atomic or molecular responses in target materials that generate emitted photons (in the reflected light process, the incoming photons that produce the response are not necessarily the same photons that leave the target).

Most remote sensing is conducted above the Earth either within or above the atmosphere. The gases in the atmosphere interact with solar irradiation and with radiation from the Earth's surface. The atmosphere itself is excited by EMR so as to become another source of released photons. Here is a generalized diagram showing relative atmospheric radiation transmission of different wavelengths.



Blue zones (absorption bands) mark minimal passage of incoming and/or outgoing radiation, whereas, white areas (transmission peaks) denote atmospheric windows, in which the radiation doesn't interact much with air molecules and hence, isn't absorbed. This next plot, made with the AVIRIS hyperspectral spectrometer (see page page 13-9), gives more a more detailed spectrum, made in the field looking up into the atmosphere, for the interval 0.4 to 2.5 μm (converted in the diagram to 400-2500 nanometers).



Most remote sensing instruments on air or space platforms operate in one or more of these windows by making their measurements with detectors tuned to specific frequencies (wavelengths) that pass through the atmosphere. However, some sensors, especially those on meteorological satellites, directly measure absorption phenomena, such as those associated with carbon dioxide, CO_2 and other gaseous molecules. Note in the second diagram above that the atmosphere is nearly opaque to EM radiation in part of the mid-IR and almost all of the far-IR region (20 to 1000 μm). In the microwave region, by contrast, most of this radiation moves through unimpeded, so radar waves reach the surface (although raindrops cause backscattering that allows them to be detected). Fortunately, absorption and other interactions occur over many of the shorter wavelength regions, so that only a fraction of the incoming radiation reaches the surface; thus harmful cosmic rays and ultraviolet (UV) radiation that could inhibit or

destroy certain life forms are largely prevented from hitting surface environments.

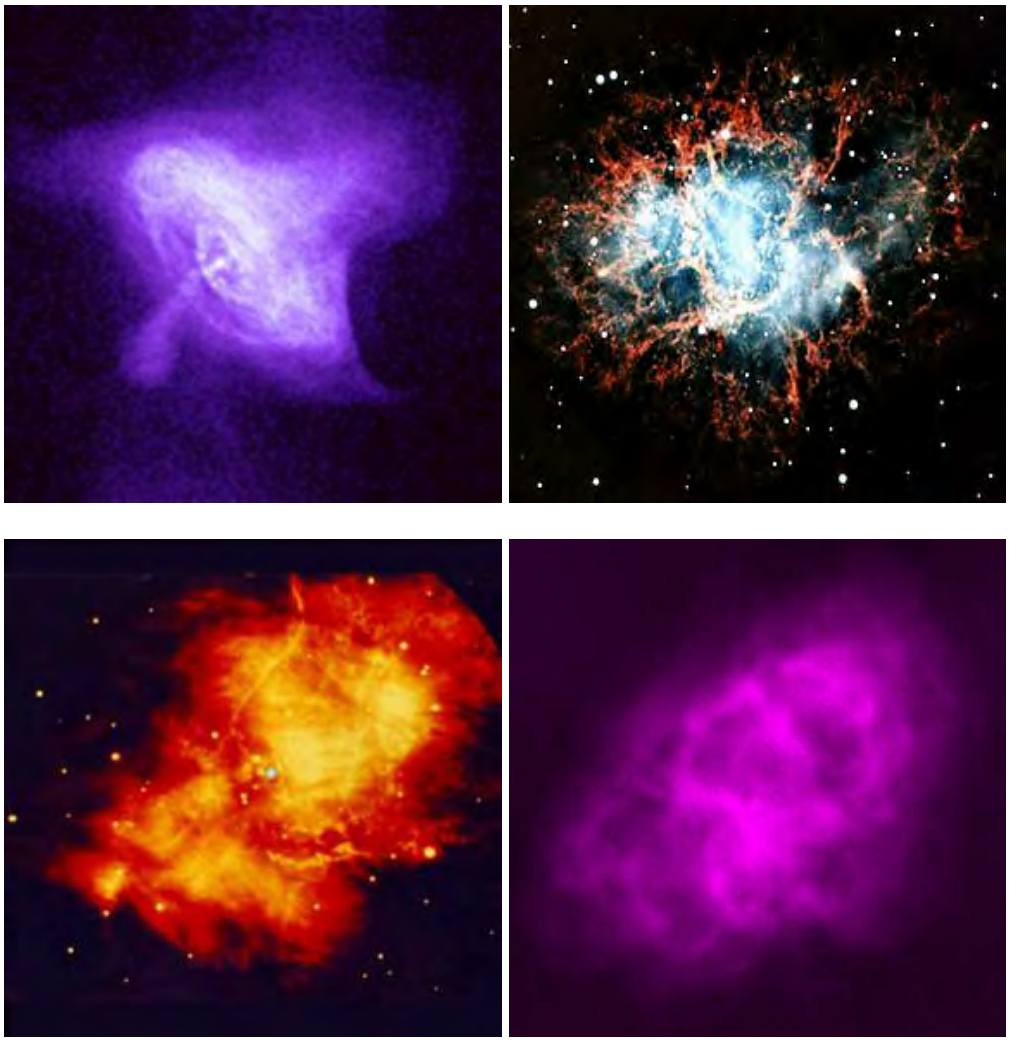
I-10: **From the first atmospheric absorption figure, list the four principal windows (by wavelength interval) open to effective remote sensing from above the atmosphere.** **ANSWER**

Backscattering (scattering of photons in all directions above the target in the hemisphere that lies on the source side) is a major phenomenon in the atmosphere. *Mie scattering* refers to reflection and refraction of radiation by atmospheric constituents (e.g., smoke) whose dimensions are of the order of the radiation wavelengths. *Rayleigh scattering* results from constituents (e.g., molecular gases [O_2, N_2 {and other nitrogen compounds}, and CO_2], and water vapor) that are much smaller than the radiation wavelengths. Rayleigh scattering increases with decreasing (shorter) wavelengths, causing the preferential scattering of blue light (blue sky effect); however, the red sky tones at sunset and sunrise result from significant absorption of shorter wavelength visible light owing to greater "depth" of the atmospheric path as the Sun is near the horizon. Particles much larger than the irradiation wavelengths give rise to *nonselective (wavelength-independent) scattering*. Atmospheric backscatter can, under certain conditions, account for 80 to 90% of the radiant flux observed by a spacecraft sensor.

Remote sensing of the Earth traditionally has used reflected energy in the visible and infrared and emitted energy in the thermal infrared and microwave regions to gather radiation that can be analyzed numerically or used to generate images whose tonal variations represent different intensities of photons associated with a range of wavelengths that are received at the sensor. This sampling of a (continuous or discontinuous) range(s) of wavelengths is the essence of what is usually termed multispectral remote sensing.

Images made from the varying wavelength/intensity signals coming from different parts of a scene will show variations in gray tones in black and white versions or colors (in terms of hue, saturation, and intensity in colored versions). Pictorial (image) representation of target objects and features in different spectral regions, usually using different sensors (commonly with bandpass filters) each tuned to accept and process the wave frequencies (wavelengths) that characterize a given region, will normally show significant differences in the distribution (patterns) of color or gray tones. It is this variation which gives rise to an image or picture. Each spectral band will produce an image which has a range of tones or colors characteristic of the spectral responses of the various objects in the scene; images made from different spectral bands show different tones or colors.

This point - that each spectral band image is unique and characteristic of its spectral makeup - can be dramatically illustrated with views of astronomical bodies viewed through telescopes (some on space platforms) equipped with different multispectral sensing devices. Below are four views of the nearby Crab Nebula, which is now in a state of chaotic expansion after a supernova explosion first sighted in 1054 A.D. by Chinese astronomers (see Section 20 - Cosmology - for other examples). The upper left illustration shows the Nebula as sensed in the high energy x-ray region; the upper right is a visual image; the lower left was acquired from the infrared region; and the lower right is a long wavelength radio telescope image.



By sampling the radiation coming from any material or class under observation over a range of continuous (or intermittent, in bands) spectral interval, and measuring the intensity of reflectance or emittance for the different wavelengths involve, a plot of this variation forms what is referred to as a **spectral signature**, the subject of the next page's discussion.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

In the first 5 pages of this Introduction, emphasis has been placed on the nature and properties of the electromagnetic radiation that is the information carrier about materials, objects and features which are the targets of interest in remote sensing. But to gather and process that information, devices called sensors are needed to detect and measure the radiation. This page looks at the basic principles involved in sensor design and development. A classification presented here indicates the variety of sensors available to "do the job". Discussion of film camera systems and radar is deferred to other Sections. This page concentrates on scanning spectroradiometers, a class of instruments that is the "workhorse" in this stable of remote sensors.

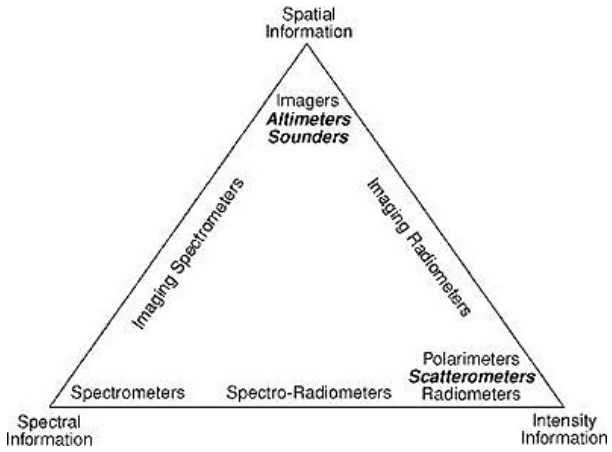
Sensor Technology; Types of Resolution

So far, we have considered mainly the nature and characteristics of EM radiation in terms of sources and behavior when interacting with materials and objects. It was stated that the bulk of the radiation sensed is either reflected or emitted from the target, generally through air until it is monitored by a sensor. The subject of what sensors consist of and how they perform (operate) is important and wide ranging. It is also far too involved to merit an extended treatment in this Tutorial. However, a synopsis of some of the basics is warranted on this page. Some useful links to reviews of sensors and their applications is included in this NASA site. A more specific treatment is given on the CNES (French) Remote Sensing website. (We point out here that many readers of this Tutorial are now using a sophisticated sensor that uses some of the technology described below: the Digital Camera; more is said about this everyday sensor near the bottom of the page.)

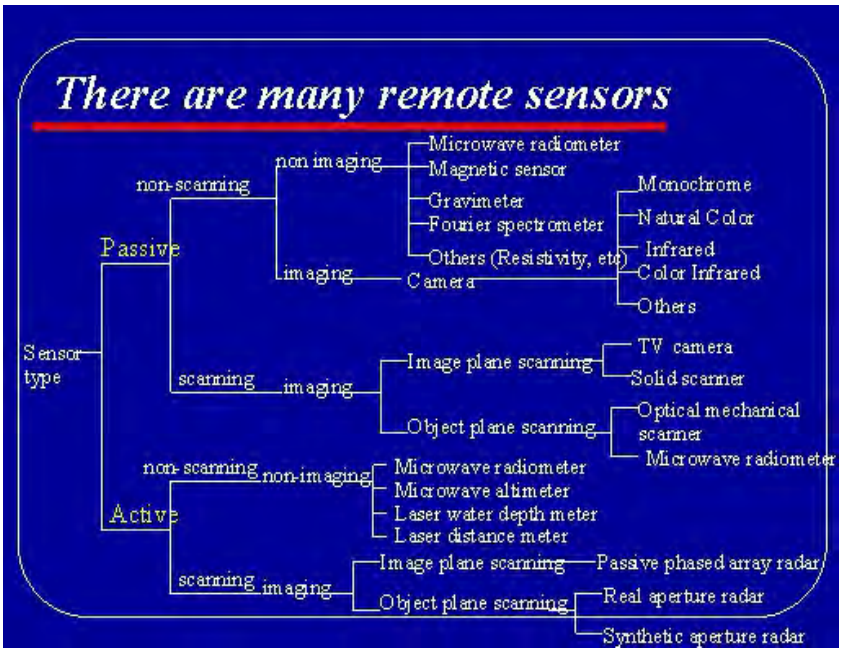
Most remote sensing instruments (sensors) are designed to measure photons. The fundamental principle underlying sensor operation centers on what happens in a critical component - the detector. This is the concept of the *photoelectric effect* (for which Albert Einstein, who first explained it in detail, won his Nobel Prize [not for Relativity which was a **much greater** achievement]; his discovery was, however, a key step in the development of quantum physics). This, simply stated, says that there will be an emission of negative particles (electrons) when a negatively charged plate of some appropriate light-sensitive material is subjected to a beam of photons. The electrons can then be made to flow as a current from the plate, are collected, and then counted as a signal. A key point: The magnitude of the electric current produced (number of photoelectrons per unit time) is directly proportional to the light intensity. Thus, changes in the electric current can be used to measure changes in the photons (numbers; intensity) that strike the plate (detector) during a given time interval. The kinetic energy of the released photoelectrons varies with frequency (or wavelength) of the impinging radiation. But, different materials undergo photoelectric effect release of electrons over different wavelength intervals; each has a threshold wavelength at which the phenomenon begins and a longer wavelength at which it ceases.

Now, with this principle established as the basis for the operation of most remote sensors, let us summarize several main ideas as to sensor types (classification) in these two diagrams:

The first is a functional treatment of several classes of sensors, plotted as a triangle diagram, in which the corner members are determined by the principal parameter measured: Spectral; Spatial; Intensity.



The second covers a wider array of sensor types:



From this imposing list, we shall concentrate the discussion on optical-mechanical-electronic radiometers and scanners, leaving the subjects of camera-film systems and active radar for consideration elsewhere in the Tutorial and holding the description of thermal systems to a minimum (see Section 9 for further treatment). The top group comprises mainly the geophysical sensors to be examined near the end of this Section.

The common components of a sensor system are shown in this table (not all need be present in a given sensor, but most are essential):

Major elements of electro-optical scanners

- Optical system: lenses, mirrors, apertures, modulators & dispersion devices
- Detectors: provide an electrical signal proportional to the irradiance on its active surface, generally some type of semiconductors
- Signal Processor: performing specified functions on the electrical signal to provide the desired output data

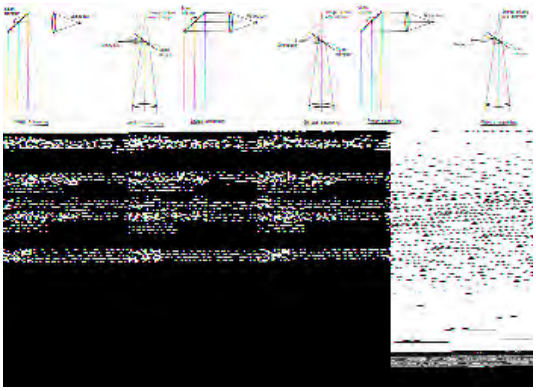
The two broadest classes of sensors are *Passive* (energy leading to radiation received comes from an external source, e.g., the Sun; the MSS is an example) and *Active* (energy generated from within the sensor system is beamed outward, and the fraction returned is measured; radar is an example). Sensors can be *non-imaging* (measures the radiation received from all points in the sensed target, integrates this, and reports the result as an electrical signal strength or some other quantitative attribute, such as radiance) or *imaging* (the electrons released are used to excite or ionize a substance like silver (Ag) in film or to drive an image producing device like a TV or computer monitor or a cathode ray tube or oscilloscope or a battery of electronic detectors (see further down this page for a discussion of detector types); since the radiation is related to specific points in the target, the end result is an image [picture] or a raster display [for example: the parallel horizontal lines on a TV screen]).

Radiometer is a general term for any instrument that quantitatively measures the EM radiation in some interval of the EM spectrum. When the radiation is light from the narrow spectral band including the visible, the term *photometer* can be substituted. If the sensor includes a component, such as a prism or diffraction grating, that can break radiation extending over a part of the spectrum into discrete wavelengths and disperse (or separate) them at different angles to an array of detectors, it is called a *spectrometer*. One type of spectrometer (used in the laboratory for chemical analysis) passes multiwavelength radiation through a slit onto a dispersing medium which reproduces the slit as lines at various spacings on a film plate (discussed on page I-2a). The term *spectroradiometer* is reserved for sensors that collect the dispersed radiation in bands rather than discrete wavelengths. Most air/space sensors are spectroradiometers.

Sensors that instantaneously measure radiation coming from the entire scene at once are called *framing systems*. The eye, a photo camera, and a TV vidicon belong to this group. The size of

the scene that is framed is determined by the apertures and optics in the system that define the *field of view, or FOV*. If the scene is sensed point by point (equivalent to small areas within the scene) along successive lines over a finite time, this mode of measurement makes up a *scanning system*. Most non-camera sensors operating from moving platforms image the scene by scanning.

Moving further down the classification tree, the optical setup for imaging sensors will be either an *image plane* or an *object plane* set up depending on where lens is before the photon rays are converged (focused), as shown in this illustration.



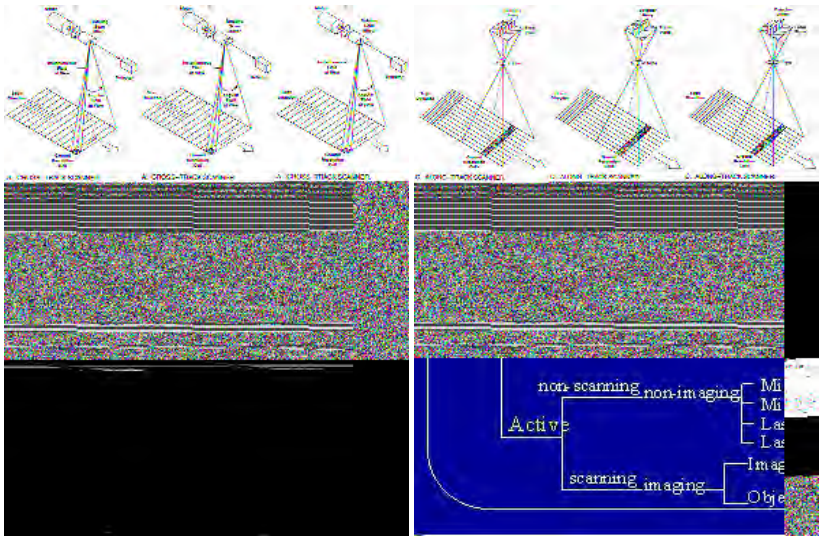
For the image plane arrangement, the lens receives parallel light rays after these are deflected to it by the scanner, with focusing at the end. For the object plane setup, the rays are focused at the front end (and have a virtual focal point in back of the initial optical train), and are intercepted by the scanner before coming to a full focus at a detector.

Another attribute in this classification is whether the sensor operates in a *non-scanning* or a *scanning mode*. This is a rather tricky pair of terms that can have several meanings in that scanning implies motion across the scene over a time interval and non-scanning refers to holding the sensor fixed on the scene or target of interest as it is sensed in a very brief moment. A film camera held rigidly in the hand is a non-scanning device that captures light almost instantaneously when the shutter is opened, then closed. But when the camera and/or the target moves, as with a movie camera, it in a sense is performing scanning as such. Conversely, the target can be static (not moving) but the sensor sweeps across the sensed scene, which can be scanning in that the sensor is designed for its detector(s) to move systematically in a progressive sweep even as they also advance across the target. This is the case for the scanner you may have tied into your computer; here its flatbed platform (the casing and glass surface on which a picture is placed) also stays put; scanning can also be carried out by put a picture or paper document on a rotating drum (two motions: circular and progressive shift in the direction of the drum's axis) in which the scanning illumination is a fixed beam.

Two other related examples: A TV (picture-taking) camera containing a vidicon in which light hitting that photon-sensitive surface produces electrons that are removed in succession (lines per inch is a measure of the TV's performance) can either stay fixed or can swivel to sweep over a scene (itself a spatial scanning operation) and can scan in time as it continues to monitor the scene. A digital camera contains an X-Y array of detectors that are discharged of their photon-induced electrons in a continuous succession that translate into a signal of varying voltage. The discharge occurs by scanning the detectors systematically. That camera itself can remain fixed or can move.

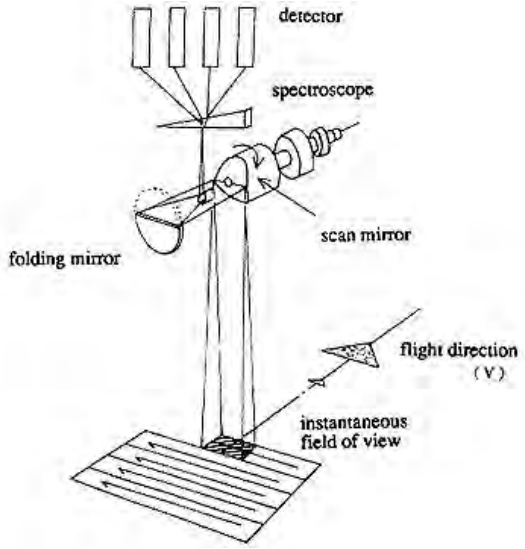
The gist of all this (to some extent obvious) is that the term **scanning** can be applied both to movement of the entire sensor and, in its more common meaning, to the process by which one or more components in the detection system either move the light gathering, scene viewing apparatus or the light or radiation detectors are read one by one to produce the signal. Two broad categories of most scanners are defined by the terms "optical-mechanical" and "optical-electronic", distinguished by the former containing an essential mechanical component (e.g., a moving mirror) that participates in scanning the scene and by the latter having the sensed radiation move directly through the optics onto a linear or two-dimensional array of detectors.

Another attribute of remote sensors, not shown in the classification, relates to the modes in which those that follow some forward-moving track (referred to as the orbit or flight *path*) gather their data. In doing so, they are said to monitor the path over an area out to the sides of the path; this is known as the *swath width*. The width is determined by that part of the scene encompassed by the telescope's full angular FOV which actually is sensed by a detector array - this is normally narrower than the entire scene's width from which light is admitted through the external aperture (usually, a telescope). The principal modes are diagrammed in these two figures:



From Sabins, Jr., F.F., Remote Sensing: Principles and Interpretation, 2nd Ed., W.H. Freeman

The *Cross Track* mode normally uses a rotating (spinning) or oscillating mirror (making the sensor an optical-mechanical device) to sweep the scene along a line traversing the ground that is very long (kilometers; miles) but also very narrow (meters; yards), or more commonly a series of adjacent lines. This is sometimes referred to as the Whiskbroom mode from the vision of sweeping a table side to side by a small handheld broom. A general scheme of a typical Cross-Track Scanner is shown below.



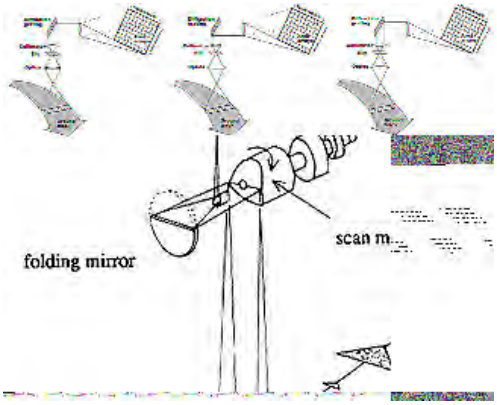
The essential components (most are shared with Along Track systems) of this instrument as flown in space are 1) a light gathering telescope that defines the scene dimensions at any moment (not shown); 2) appropriate optics (e.g., lens) within the light path train; 3) a mirror (on aircraft scanners this may completely rotate; on spacecraft scanners this usually oscillates over small angles); 4) a device (spectroscope; spectral diffraction grating; band filters) to break the incoming radiation into spectral intervals; 5) a means to direct the light so dispersed onto an array or bank of detectors; 6) an electronic means to sample the photo-electric effect at each detector and to then reset the detector to a base state to receive the next incoming light packet, resulting in a signal stream that relates to changes in light values coming from the ground targets as the sensor passes over the scene; and 7) a recording component that either reads the signal as an analog current that changes over time or converts the signal (usually onboard) to a succession of digital numbers, either being sent back to a ground station. A scanner can also have a *chopper* which is a moving slit or opening that as it rotates alternately allows the signal to pass to the detectors or interrupts the signal (area of no opening) and redirects it to a reference detector for calibration of the instrument response.

Each line is subdivided into a sequence of individual spatial elements that represent a corresponding square, rectangular, or circular area (ground resolution cell) on the scene surface being imaged (or in, if the target to be sensed is the 3-dimensional atmosphere). Thus, along any line is an array of contiguous cells from each of which emanates radiation. The cells are sensed one after another along the line. In the sensor, each cell is associated with a **pixel** (picture element) that is tied to a microelectronic detector; each pixel is characterized for a brief time by some single value of radiation (e.g., reflectance) converted by the photoelectric effect into electrons.

The areal coverage of the pixel (that is, the ground cell area it corresponds to) is determined by

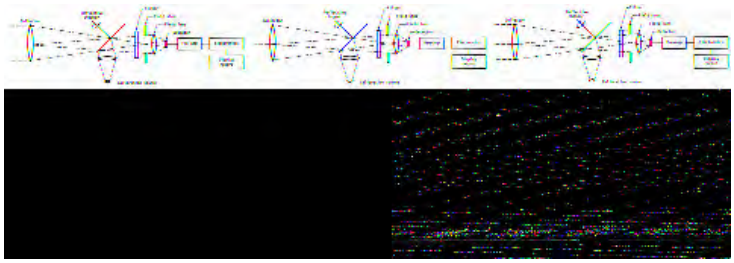
instantaneous field of view (IFOV) of the sensor system. The IFOV is defined as the solid angle extending from a detector to the area on the ground it measures at any instant (see above illustration). IFOV is a function of the optics of the sensor, the sampling rate of the signal, the dimensions of any optical guides (such as optical fibers), the size of the detector, and the altitude above the target or scene. The electrons are removed successively, pixel by pixel, to form the varying signal that defines the spatial variation of radiance from the progressively sampled scene. The image is then built up from these variations - each assigned to its pixel as a discrete value called the DN (a digital number, made by converting the analog signal to digital values of whole numbers over a finite range [for example, the Landsat system range is 2^8, which spreads from 0 to 255]). Using these DN values, a "picture" of the scene is recreated on film (photo) or on a monitor (image) by converting a two dimensional array of pixels, pixel by pixel and line by line along the direction of forward motion of the sensor (on a platform such as an aircraft or spacecraft) into gray levels in increments determined by the DN range.

The *Along Track* Scanner has a linear array of detectors oriented normal to flight path. The IFOV of each detector sweeps a path parallel with the flight direction. This type of scanning is also referred to as pushbroom scanning (from the mental image of cleaning a floor with a wide broom through successive forward sweeps). The scanner does not have a mirror looking off at varying angles. Instead there is a line of small sensitive detectors stacked side by side, each having some tiny dimension on its plate surface; these may number several thousand. Each detector is a charge-coupled device (CCD), as described in more detail below on this page. In this mode, the pixels that will eventually make up the image correspond to these individual detectors in the line array. Some of these ideas are evident in this image.

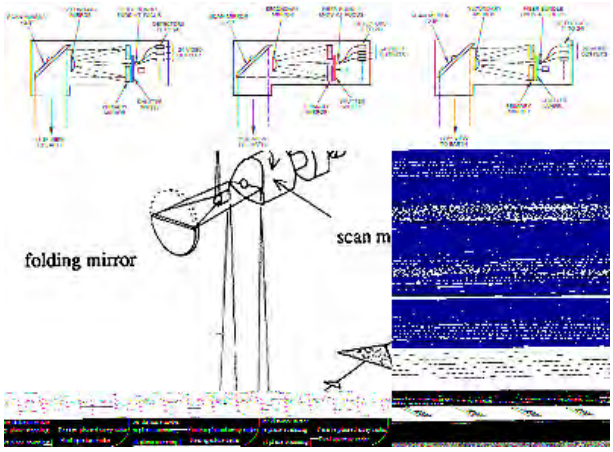


As the sensor-bearing platform advances along the track, at any given moment radiation from each ground cell area along the ground line is received simultaneously at the sensor and the collection of photons from every cell impinges in the proper geometric relation to its ground position on every individual detector in the linear array equivalent to that position. The signal is removed from each detector in succession from the array in a very short time (milliseconds), the detectors are reset to a null state, and are then exposed to new radiation from the next line on the ground that has been reached by the sensor's forward motion. The result is a build up of linear array data that forms a 2-dimension areal array. As signal sampling improves, the possibility of sets of continuous linear arrays all exposed simultaneously to radiation from the scene, leading to areal arrays, all being sampled at once will increase the equivalent area of ground coverage.

With this background, on to some more specific information. This next figure is a diagrammatic model of an electro-optical sensor that does not contain the means to break the incoming radiation into spectral components (essentially, this is a panchromatic system in which the *filter* admits a broad range of wavelengths). The diagram contains some of the elements found in the Return Beam Vidicon (TV-like) on the first two Landsats. Below it is a simplified cutaway diagram of the Landsat Multispectral Scanner (MSS) which through what is here called a *shutter wheel or mount*, containing filters each passing a limited range of wavelength, the spectral aspect to the image scanning system is added, i.e., produces discrete spectral bands:



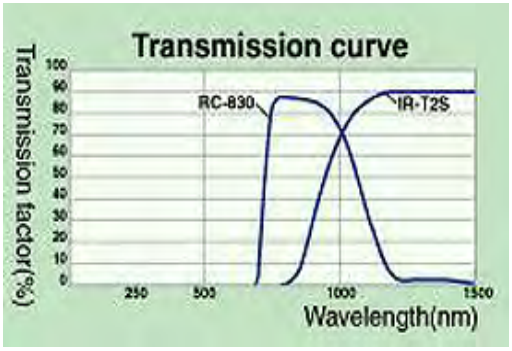
As this pertains to the Landsat Multispectral Scanner (the along-track type), check this cutaway diagram:



The front end of a sensor is normally a telescopic system (in the image denoted by the label 11.6°) to gather and direct the radiation onto a mirror or lens. The mirror rocks or oscillates back and forth rapidly over a limited angular range (the 2.9 ° to each side). In this setup, the scene is imaged only on one swing, here forward, and not scanned on the opposing or reverse swing (active scanning can occur on both swings, especially on slow-moving aircraft sensors). Some sensors allow the mirror to be pointed off to the side at specific fixed angles to capture scenes adjacent to the vertical mode ground track (SPOT is an example). In some scanners, a chopper may be in the optic train near the mirror. It is a mechanical device to interrupt the signal either to modulate or synchronize it or, commonly, to allow a very brief blockage of the incoming radiation while the system looks at an onboard reference source of radiation of steady, known wavelength(s) and intensity in order to calibrate the final signals tied to the target. Other mirrors or lenses may be placed in the train to further redirect or focus the radiation.

The radiation - normally visible and/or Near and Short Wave IR, and/or thermal emissive in nature - must then be broken into spectral intervals, i.e., into broad to narrow bands. The width in wavelength units of a band or channel is defined by the instrument's **spectral resolution** (see top of page 13-5). The spectral resolution achieved by a sensor depends on the number of bands, their bandwidths, and their locations within the EM spectrum. Prisms and diffraction gratings are one way to break selected parts of the EM spectrum into intervals; bandpass filters are another. In the above cutaway diagram of the MSS the filters are located on the shutter wheel. The filters select the radiation bands that are sensed and have detectors placed where each wavelength-dependent band is sampled. For the filter setup, the spectrally-sampled radiation is carried along optical fibers to dedicated detectors.

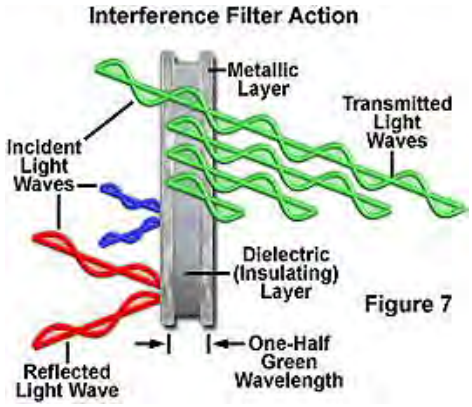
Spectral filters fall into two general types: Absorption and Interference. Absorption filters pass only a limited range of radiation wavelengths, absorbing radiation outside this range. Interference filters reflect radiation at wavelengths lower and higher than the interval they transmit. Each type may be either a broad or a narrow *bandpass* filters. This is a graph distinguishing the two types.



These filters can further be described as high bandpass (IRT2; selectively removes shorter wavelengths) or low bandpass (RC830; absorbs longer wavelengths) types.

Absorption filters are made of either glass or gelatin; they use organic dyes to selectively transmit certain wavelength intervals. These filters are the ones commonly used in photography (the various colors [wavelengths] transmitted are designated by Wratten numbers).

Interference filters work by using thin films that reflect unwanted wavelengths and transmit others through a specific interval, as shown in this illustration:

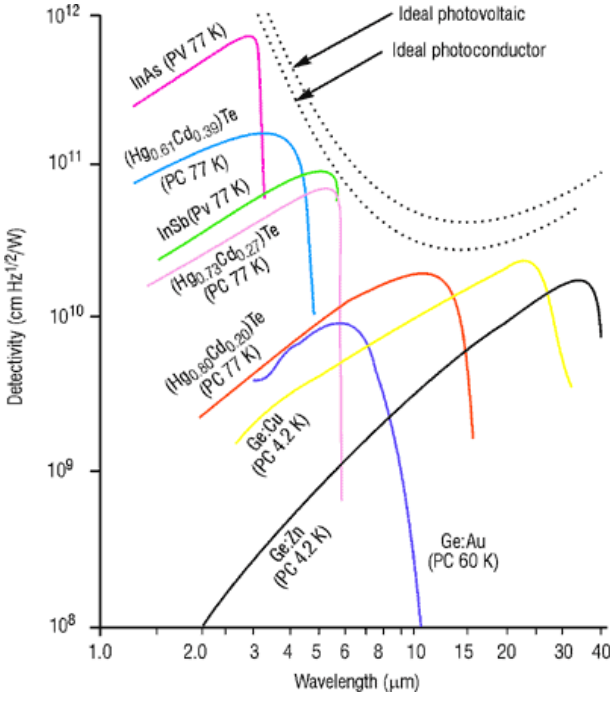


A common type of specialized filter used in general optics and on many scanning spectroradiometers is the *dichroic* filter.This uses an optical glass substrate over which are deposited (in a vacuum setup) from 20 to 50 thin (typically, 0.001 mm thick) layers of a special refractive index dielectric material (or materials in certain combinations) that selectively transmits a specific range or band of wavelengths. Absorption is nearly zero. These can be either additive or subtractive color filters when operating in the visible range (see page 10-2). Another type is the polarizing filter. A haze filter removes or absorbs much of the scattering effects of atmospheric moisture and other haze constituents.

The next step is to get the spectrally separated radiation to appropriate detectors. This can be done through lenses or by detector positioning or, in the case of the MSS and other sensors, by channeling radiation in specific ranges to fiber optics bundles that carry the focused radiation to an array of individual detectors. For the MSS, this involves 6 fiber optics leads for the six lines scanned simultaneously to 6 detectors for each of the four spectral bands, or a total of 24 detectors in all.

In the early days of remote sensing, photomultipliers served as detectors. Most detectors today are made of solid-state semiconductor metals or alloys. A semiconductor has a conductivity intermediate between a metal and an insulator. Under certain conditions, such as the interaction with photons, electrons in the semiconductor are excited and moved from a filled energy level (in the electron orbital configuration around an atomic nucleus) to another level called the conduction band which is deficient in electrons in the unexited state. The resistance to flow varies inversely with the number of incident photons. The process is best understood by quantum theory. Different materials respond to different wavelengths (actually, to photon energy levels) and are thus spectrally selective.

In the visible light range, silicon metal and PbO are common detector materials. Silicon photodiodes are used in this range. Photoconductor material in the Near-Ir includes PbS (lead sulphide) and InAs (indium-arsenic). In the Mid-IR (3-6 µm), InSb (indium-stibnium [an antimony compound]) is responsive.The most common detector material for the 8-14 µm range is Hg-Cd-Te (mercury-cadmium-tellurium); when operating it is necessary to cool the detectors to near zero Kelvin (using Dewars coolers) to optimize the efficiency of electron release. Other detector materials are also used and perform under specific conditions. This next diagram gives some idea of the variability of semiconductor detectivity over operating wavelength ranges.



Other detector systems, less commonly used in remote sensing function in different ways. The list includes photoemissive, photdiode, photovoltage, and thermal (absorption of radiation) detectors. The most important now are CCDs, or Charge-Coupled-Detectors (CCDs; the "D" is also substituted for by "Detector") , that are explained in the next paragraph. This approach to sensing EM radiation was developed in the 1970s, which led to the Pushbroom Scanner, which uses CCDs as the detecting sensor. The nature and operation of CCD's are reviewed in these two websites: CCD site 1 and CCD site 2. An individual CCD is an extremely small silicon (micro)detector, which is light-sensitive. Many individual detectors are placed on a chip side by side either in a single row as a linear array or in stacked rows of linear arrays in X-Y (two dimensional or areal) space. Here is a photograph of a CCD chip:

When photons strike a CCD detector, electronic charges develop whose magnitudes are proportional to the intensity of the impinging radiation during a short time interval (exposure time). From 3,000 to more than 10,000 detector elements (the CCDs) can occupy a linear space less than 15 cm in length. The number of elements per unit length, along with the optics, determine the spatial resolution of the instrument. Using integrated circuits each linear array is sampled very rapidly in sequence, producing an electrical signal that varies with the radiation striking the array. This changing signal goes through a processor to a recorder, and finally, is used to drive an electro-optical device to make a black and white image.

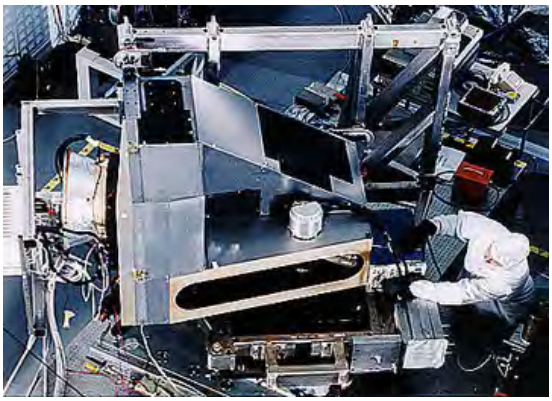
After the instrument samples the almost instantaneous signal, the array discharges electronically fast enough to allow the next incoming radiation to be detected independently. A linear (one-dimensional) array acting as the detecting sensor advances with the spacecraft's orbital motion, producing successive lines of image data (the pushbroom effect). Using filters to select wavelength intervals, each associated with a CCD array, leads to multiband sensing. The one disadvantage of current CCD systems is their limitation to visible and near IR (VNIR) intervals of the EM spectrum. (CCDs are also the basis for two-dimensional arrays - a series of linear CCDs stacked in parallel to extend over an area; these are used in the now popular digital cameras and are the sensor detectors commonly employed in telescopes of recent vintage.)

Each individual CCD corresponds to the "pixel" mentioned above. The size of the CCD is one factor in setting spatial resolution (smaller sizes represent smaller areas on the target surface); another factor is the height of the observing platform (satellite or aircraft); a third factor is tied to the use of a telescopic lens.

Once a scanner or CCD signal has been generated at the detector site, it needs to be carried through the electronic processing system. As stated above, one ultimate output is the computerized signal (commonly as DN [Digital Number] variations) used to make images or be analyzed by computer programs. Pre-amplification may be needed before the last stage. Onboard digitizing is commonly applied to the signal and to the reference radiation source used in calibration. The final output is then sent to a ground receiving station, either by direct readout (line of sight) from the spacecraft or through satellite relay systems like TDRSS (Tracking and Relay Satellite System) or other geosynchronous communications satellites). Another option is to record the signals on a tape recorder and play them back when the satellite's orbital position permits direct transmission to a receiving station (this was used on many of the earlier satellites, including Landsat [ERTS], now almost obsolete because of the much improved satellite communications network).

The subject of sensor performance is beyond the scope of this page. Three common measures are here mentioned: 1) S/N (signal to noise ratio; the noise can come from internal electronic components or the detectors themselves); 2) NEΔP and NEΔT, the Noise Equivalent Power (for reflectances) and 3) Noise Equivalent Temperature (for thermal emission detectors).

Sensors flown on unmanned spacecraft tend to be engineering marvels. Their components are of the highest quality. Before flight they are tested and retested to look for any functional weaknesses. With all this "loving care" it is no wonder that they can cost $millions to develop and fabricate. We show a photo of the MODIS sensor that now operating well on the Terra spacecraft launched in late 1999 and on Aqua two years later.

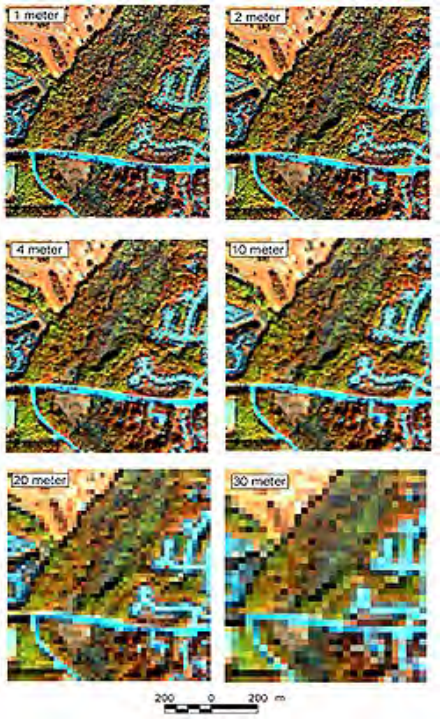


Finally, we need to consider one more vital aspect of sensor function and performance, namely the subject of *resolution*. There are three types: Spatial, Spectral, and Radiometric.

The spatial resolution concept is reviewed on page 10-3 as regards photographic systems and photogrammetry; check out that page at any time during these paragraphs. Here, we will attempt a generally non-technical overview of spatial resolution.

Most of us have a strong intuitive feeling for the meaning of **spatial resolution**. Think of this experiential example. Suppose you are looking at a forested hillside some considerable distance away. What you see is the presence of the continuous forest but at a great distance you do not see individual trees. As you go closer, eventually the trees, which may differ in size, shape, and species, become distinct as individuals. They have thus been resolved. As you draw much nearer, you start to see individual leaves. This means that the main components of an individual entity are now discernible and thus that category is being resolved. You can carry this ever further, through leaf macro-structure, then recognition of cells, and in principle with *higher resolutions* the individual constituent atoms and finally subatomic components. This last step is the highest resolution (related to the smallest sizes) achievable by instruments or sensors. All of these levels represent the "ability to recognize and separate features of specific sizes".

At this point in developing a feel for the meaning and importance of spatial resolution, consider this diagram:



At 1 and 2 meters, the scene consists of objects and features that, being resolvable, we can recognize and name. At 30 meters, the blocky pixels form a pattern but it is not very intelligible. However, when the image covers a wide area, as does a Landsat scene, the small area shown here becomes a small part of the total image, so that larger features appear sharp, discernible, and usually identifiable. As a general rule, high resolution images retain sharpness if the area covered is relatively small; IKONOS images (1-4 meter resolution) are a case in point.

The common sense definition of spatial resolution is often simply stated as the smallest size of an object that can be picked out from its surrounding objects or features. This separation from neighbors or background may or may not be sufficient to identify the object. Compare these ideas to the definition of three terms which have been extracted from the Glossary of Appendix D of this Tutorial:

resolution-Ability to separate closely spaced objects on an image or photograph. Resolution is commonly expressed as the most closely spaced line-pairs per unit distance that can be distinguished. Also called spatial resolution.
resolution target-Series of regularly spaced alternating light and dark bars used to evaluate the resolution of images or photographs.
resolving power-A measure of the ability of individual components. and of remote sensing systems, to separate closely spaced targets.

These three terms will be defined as they apply to photographic systems (page 10-3 again). But resolution-related terms are also appropriate to electro-optical systems, standard optical devices, and even the human eye. The subject of resolution is more extensive and complicated than suggested from the above statements. Lets explore the ideas in more detail. The first fundamental notion is to differentiate resolution from resolving power. The former refers to the elements, features or objects in the *target*, that is the scene being sensed from a distance; the latter concerns the ability of the sensor, be it electronic or film or the eye, to separate the smallest features in the target that are the objects being sensed.

To help in the visualization of effective (i.e., maximum achieved) spatial resolution, lets work with a target that contains the objects that will be listed and lets use the human eye as the sensor, making you part of the resolving process, since this is the easiest notion involved in the experience. (A suggestion: Review the description of the eye's functionality given in the answer to question I-1 [page I-1] in the Introductory Section.)

Start with a target that contains rows and columns of red squares each bounded by a thin black line place in contact with each other. The squares have some size determined by the black outlines. At a certain distance where you can see the whole target, it appears a uniform red. Walk closer and closer - at some distance point you begin to see the black contrasting lines. You have thus begun to resolve the target in that you can now state that there are squares of a certain size and these appear as individuals. Now decrease the black line spacing, making each square (or resolution cell) smaller. You must move closer to resolve the smaller squares. Or you must have improved eyesight (the rods and cones in the eye determine resolution; their sizes define the eye's resolution; if some are damaged that resolution decreases). The chief variable here is distance to the target.

Now modify the experiment by replacing every other square with a green version but keeping the squares in contact. At considerable distance neither the red nor green individuals can be specifically discerned as to color and shape. They blend, giving the eye (and its brain processor) the impression of "yellowness" of the target (the effects of color combinations are treated also in Section 10). But as you approach, you start to see the two colors of squares as individuals. This distance at which the color pairs start to resolve is greater than the case above in which thin black lines form the boundary. Thus, for a given size, color contrast (or tonal contrast, as between black and white or shades of gray squares) becomes important in determining the onset of effective resolution. Variations of our experiment would be to change the squares to circles in regular alignments and have the spaces between the packed circles consist of non-red background, or draw the squares apart opening up a different background. Again, for a given size the distances at which individuals can first be discerned vary with these changing conditions. One can talk now in terms of the smallest individual(s) in a collection of varying size/shape/color objects that become visibly separable - hence resolved.

Three variables control the achieved spatial resolution: 1) the nature of the target features as just specified, the most important being their size; 2) the distance between the target and the sensing device; and 3) some inherent properties of the sensor embodied in the term *resolving power*. For this last variable, in the eye the primary factor is the sizes and arrangements of the rods and cones in the retina; in photographic film this is determined in part by the size of the AgCl grains or specks of color chemicals in the emulsion formed after film exposure and subsequent, although other properties of the camera/film system enter in as well.

For the types of spaceborne sensors discussed on this page, there are several variables or factors that specify the maximum (highest) resolution obtainable. Obviously, first is the spatial and spectral characteristics of the target scene features being sensed, including the smallest objects who presence and identities are being sought. Next, of course, is the distance between target and sensor (orbital altitude). (For sensors in aircraft or spacecraft the interfering aspects of the atmosphere can degrade resolution). Last, the speed of the platform, be it a balloon, an aircraft, an unmanned satellite, or a human observer in a shuttle or space station, is relevant in that it determines the "dwell time" available to the sensor's detectors on the individual features from which the photo signals emanate.

Most targets have some kind of limiting area to be sensed that is determined by the geometric configuration of the sensor system being used. This is implied by the above-mentioned Field of View, outside of which nothing is "visible" at any moment. Commonly, this FOV is related to a telescope or circular tube (whose physical boundaries select or collimate the outer limits of the scene to be sensed) that admits only radiation from the target at any particular moment. The optics (mainly, the lens[es]) in the telescope are important to the resolving power of the instrument. *Magnification* is one factor, as a lens

system that increases magnifying capability also improves resolution. The spectral distribution of the incoming photons also plays a role. But, for sensors like those on Landsat or SPOT, mechanical and/or electronic functions of the signal collection and detector components become critical factors in obtaining improved resolution. This resolution is also equivalent to the pixel size. (The detectors themselves influence resolution by their inherent Signal to Noise (S/N) capability; this can vary in terms of spectral wavelengths). For an optical-mechanical scanner, the oscillation or rotation rate and arrangement of the scanning mirror motion will influence the Instantaneous Field of View IFOV); this is one of three factors that closely control the final resolution (pixel size) achieved. The second factor is related to the *Size (dimensions)* of an optical aperture that admits the light from an area on the mirror that corresponds to the segment of the target being sensed. For the Landsat MSS, as an example, the width of ground being scanned across-track at any instant is 480 m. The focusing optics and scanning rate serve to break the width into 6 parallel scan lines that are directed to 6 detectors thereby accounting for the 80 meters (480/6) of resolution associated with a pixel along each scan line.

The third factor, which controls the cross-track boundary dimension of each pixel, is *Sampling Rate* of the continuous signal beam of radiation (light) sent by the mirror (usually through filters) to the detector array. For the Landsat MSS, this requires that all radiation received during the cross-track sweep along each scan line that fits into the IFOV (sampling element or pixel) be collected after every 10 microseconds of mirror advance by sampling each pixel independently in a very short time by electronic discharge of the detector. This sampling is done sequentially for the succession of pixels in the line. In 10 microseconds, the mirror's advance is equivalent to sweeping 80 m forward on the ground; its cutoff to form the instantaneous signal contained in the pixel thus establishes the other two sides of the pixel (perpendicular to the sweep direction). In the first MSS (Landsat-1), the dimensions thus imposed are equivalent to 79 by 57 meters (79 is the actual value in meters, but 80 meters [rounded off] is often quoted), owing to the nature of the sampling process. Some of the ideas in the above paragraph, which may still seem tenuous to you at the moment, may be more comprehensible after reading page I-16 of this Introduction which describes in detail the operation of the Landsat Multispectral Scanner. That page further delves into this 79 x 57 pixel geometry.

The relevant resolution determinants (as pertains to pixel size) depend on the size of each fixed detector which in turn governs the sampling rate. That rate must be in "sync" with the ability to discharge each detector in sequence fast enough to produce an uninterrupted flow of photon-produced electrons. This is constrained by the motion of the sensor, such that each detector must be discharged ("refreshed") quickly enough to then record the next pixel that is the representative of the spatially contiguous part of the target next in line in the direction of platform motion. Other resolution factors apply to thermal remote sensing (such as the need to cool detectors to very low temperatures; see Section 9) and to radar (pulse directions, travel times, angular sweep, etc; see Section 8).

Since the pixel size is a prime factor in determining spatial resolution, one may well ask about objects that are smaller than the ground dimensions represented by the pixel. These give rise to the "mixed pixel" concept that is discussed on page 13-2. A resolution anomaly is that under circumstances of objects smaller than a pixel that have high contrasts to their surroundings in the ground space corresponding to the pixel dimensions sampled may actually so strongly affect the DN or radiance value of that pixel as to darken or lighten it relative to neighboring pixels that don't have the object(s). Thus, a 10 m wide light concrete road within a pixel's ground equivalent that has vegetation neighbors consisting of dark leaves, when contributing together, will reduce the averaged radiance of that pixel sufficient to produce a visual contrast such that the road is detectable along its linear trend in the image.

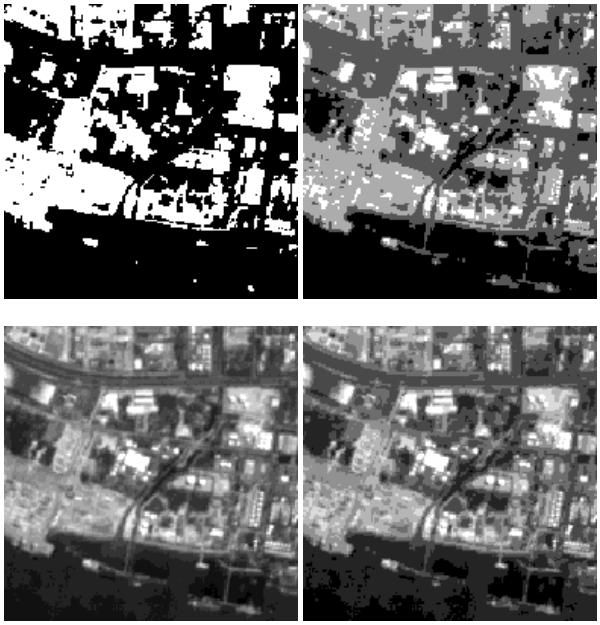
For sensed scenes that are displayed as photographic images, the optimum resolution is a combination of three actions: 1) the spatial resolution inherent to the sensor; 2) apparent improvements imposed during image processing; 3) the further improvement that may reside in the photographic process. To exemplify this idea: Landsat MSS images produced as pictures by the NASA Data Processing Facility, part of Goddard's Ground Data Handling System, were of notable quality; they had a certain degree of image manipulation imposed in reaching their end product pictures (later, this type of product was available through the EROS Data Center). But companies like the Earth Satellite Corp. took raw or corrected Landsat data and ran them through even more rigorous image processing algorthms which yielded superior pictures. These could be enlarged significantly without discernible loss of detail. The finest end product was then achieved by printing the images as generated electronically from the DN data on a type of film called Cibachrome, which maximizes the sharpness of the scene and enriches its colors, so that a viewer would rate the end result as of the highest quality.

Now, if you haven't already done so, go to page 10-3 to review that classic method by which spatial resolution is determined in photographs, but also applicable to electronically-generated images.

One goal for space-operated sensors in recent years has been improved spatial resolution (now, down to better than [smaller than] 1 meter) and greater spectral resolution (from band widths of about 10-30 nanometers [as pertains to the Landsat MSS] to 1 nanometer

or less, which carries capabilities into the hyperspectral mode discussed on page I-24 of this Section and again in Section 13. Spectral resolution has also been considered above on this page. Suffice to remind you that high spectral resolution allows spectral signatures to be plotted and these are superior to band signatures as aids to material and class identification. As with spatial resolution, a price is paid to get better spectral resolution: the number of detectors must be significantly increased; these must be physically placed to capture the wavelength-dependent radiation spread over part of the spectrum by a dispersing medium other than filters. This impacts signal handling onboard and data handling on the ground.

We have not yet defined **radiometric resolution** on this page. It is a rather esoteric concept that relates to levels of quantization that can be detected or be established to improve scene quality (such as tonal contrast). Consider, for example, a range of radiation intensities (brightness levels). This continuous range can be subdivided into a set of values of steadily increasing intensity. Each subdivision is a "level" that in a black and white rendition of a scene is represented by some degree of grayness. A two level rendition would consist of just black and white (all intermediate levels have been assigned to one or the other). A four level scene would include two intermediate gray levels). A 64 level image would have a range of distinguishable increasing (from black) gray tones up to the highest (white). Most sensors convert intercepted radiation into a digital form, which consists of a number that falls within some range of values. Radiometric resolution defines this range of values. A sensor with 8-bit resolution (e.g. Landsat TM) has a range of 256 levels, or 2^8, values (since 0 is one level, the range is 0-255). A 6-bit sensor (e.g. Landsat Multispectral Scanner (MSS) 1) has a range of 64, or 2^6, level values. To illustrate how this affects a scene representation, consider these panels (produced by Dr. S. Liew) that depict a scene at 2^1, 2^2, 2^3, and 2^4, that is, 2(upper left), 4, 8, and 16 (lower right) gray levels, or quantized radiometric values:



Note that the human eye is inefficient at distinguishing differences in gray levels much beyond the limit of 16. Thus, a 64-level image would look closely like a 16-level image. And so also for a 256-level (0 to 255) image which is just a little sharper than a 16 level image. Where higher radiometric resolution (range of intensity levels) becomes important is in image processing by computer. The broader range - smaller increments from one level to the next - provides a better measure of variations in radiation intensity which in turn prompts better spatial and spectral resolution. In classifying a scene, different classes are more precisely identified if radiometric precision is high.

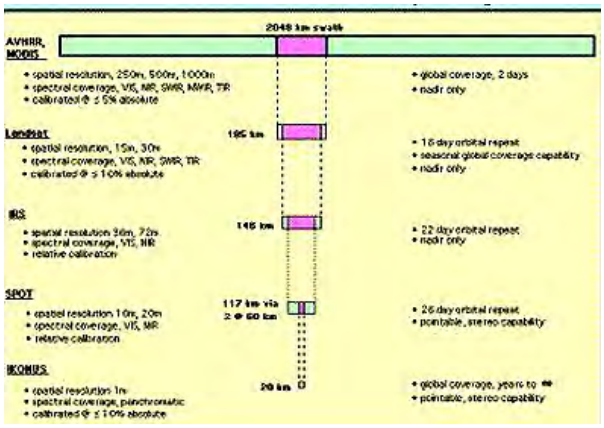
This is brought home by examining ancillary information in a Landsat scene as it was processed at NASA Goddard. At the bottom of an 8 by 10 inch print (the standard) is notation explained elsewhere in the Tutorial. Included is a 16-level gray scale or bar. When this 16 level bar was scanned to make the image below and then processed (by Photoshop), the resulting stretch of the full bar allowed only 11 levels to be distinguishable in this image (scroll right to see the full extent of the bar). In only, say, 6 levels were separable (as printed), the picture would be "flat" (less contrast). The goal of photo processing is to try to achieve the widest spread of levels which usually optimizes the tonal balance in a black and white image.

For more insight into radiometry, you are referred to this Web site prepared by CNES.

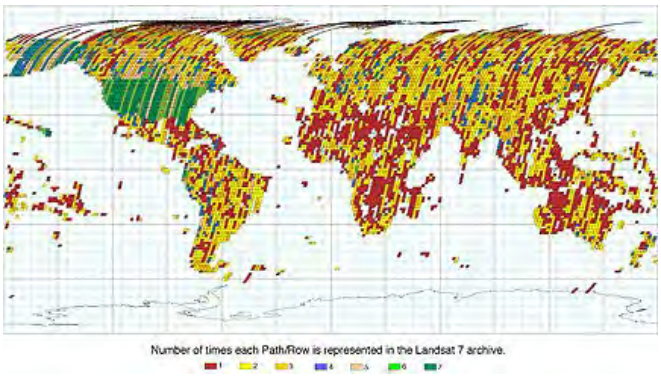
To sum up the above treatment of resolution, considerable effort and expense is applied

to designing and constructing sensors that maximize the resolution needed for their intended tasks. Greater spectral resolution (as now achievable in hyperspectral sensors) means that individual entities (classes; features) can be more accurately identified as the details of their spectral signatures are sensed. Superior spatial resolution will permit ever smaller targets to be seen as individuals; many of the items we live with while on the Earth's surface are just a few meters or less in size, so that if these can be resolved their recognition could lead to their identity and hence their better classification. Increased radiometric resolution gives rise to sharper images and to better discrimination of different target materials in terms of their relative brightness.

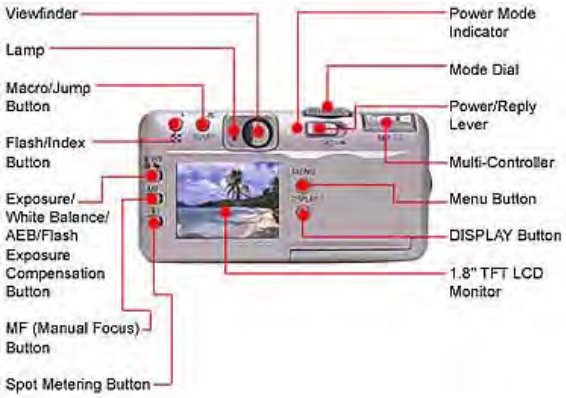
The trend over the last 30 years has been directed towards improved resolution of each type. This diagram shows this for five of the satellite systems in orbit; the swath width (across-track distance) is also shown. (However, a reminder: coarser space resolution has its place because it is what is available on satellites that deliberately seek wide fields of view so as to provide regional information at smaller scales, which can be optimal for certain applications.)



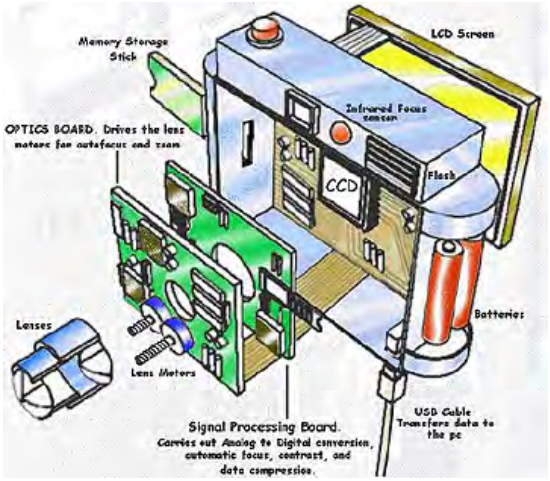
The field of view controls the swath width of a satellite image. That width, in turn, depends on the optics of the observing telescope, on electronic sampling limits inherent to the sensor, and on the altitude of the sensor. Normally, the higher the satellite's orbit, the wider the swath width and the lower the spatial resolution. Both altitude and swath width determine the "footprint" of the sensed scene, i.e., its across track dimensions and the frequency of repeat coverage. Taken together with cloud cover variations, the number of "good" scenes obtained along the total adjacent swaths occupied during a stretch of coverage will vary, as suggested by this map of scene frequency for Landsat-7 over a 123 day period (note that the greatest "successes" occur over the United States, largely because the principal market for the imagery is there [not fortuitous cloudfree opportunities] upping the number of "trys" for scene acquisition).



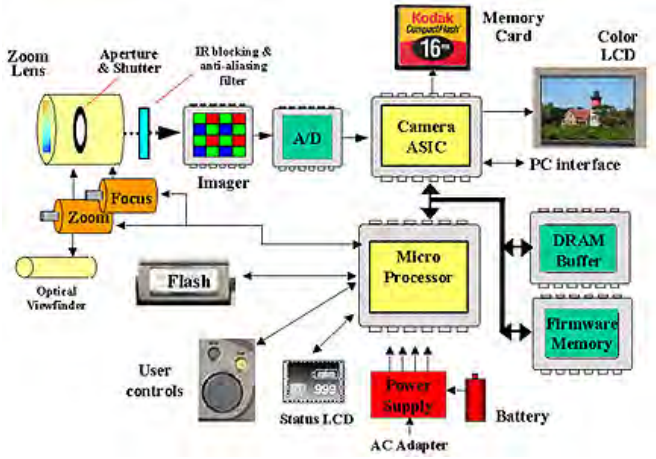
As a closing ancillary thought for this page, we mention that today's popular *Digital Cameras* have much in common with some of the sensors described above. Light is recorded on CCDs after passing through a lens system and a means of splitting the light into primary colors. The light levels in each spectral region are digitized, processed by a computer chip, and stored (some cameras use removable disks) for immediate display in the camera screen or for downloading to a computer for processing into an image that can be reproduced on a computer's printer. Alternatively,commercial processing facilities - including most larger drug stores - can take the camera or the disk and produce high quality photos from the input. Here are three websites that describe the basics of digital camera construction and operation: Site 1, Site`2, and Site 3. The next three diagrams show schematically the components of typical digital cameras:



Digital Camera Exterior



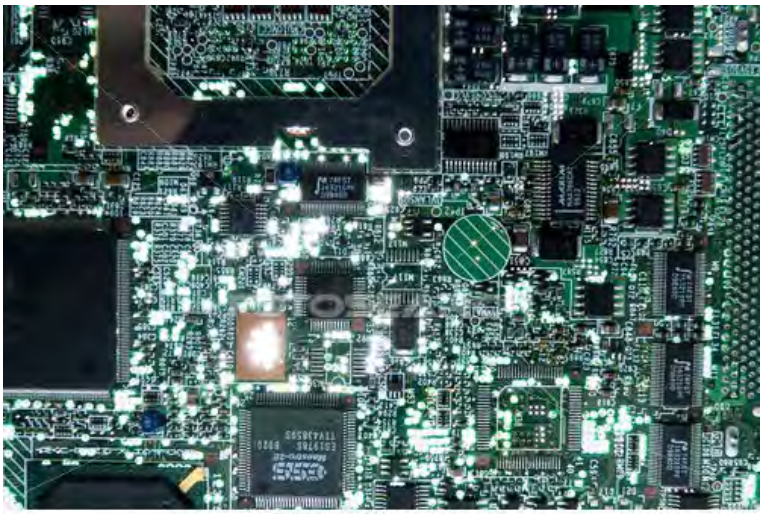
The Interior of a Digital Camera



How a Digital Camera Operates

The CCD unit, which detects the radiation (light) comprising the signal is housed within the ASIC (Application Specific Integrated Circuit):

The inside of a digital camera can look daunting and complication to anyone not familiar with electronic systems. Here is a camera circuit board:



The silicon CCD chip is at the bottom, left of center

Let's move on to get an inkling as to how remote sensing data are processed and classified (this subject is detailed in Section 1).

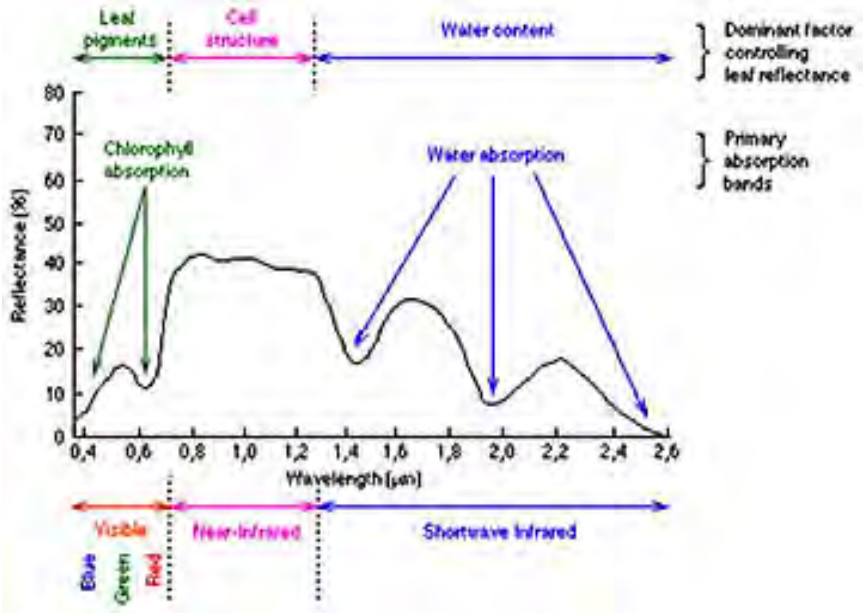
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

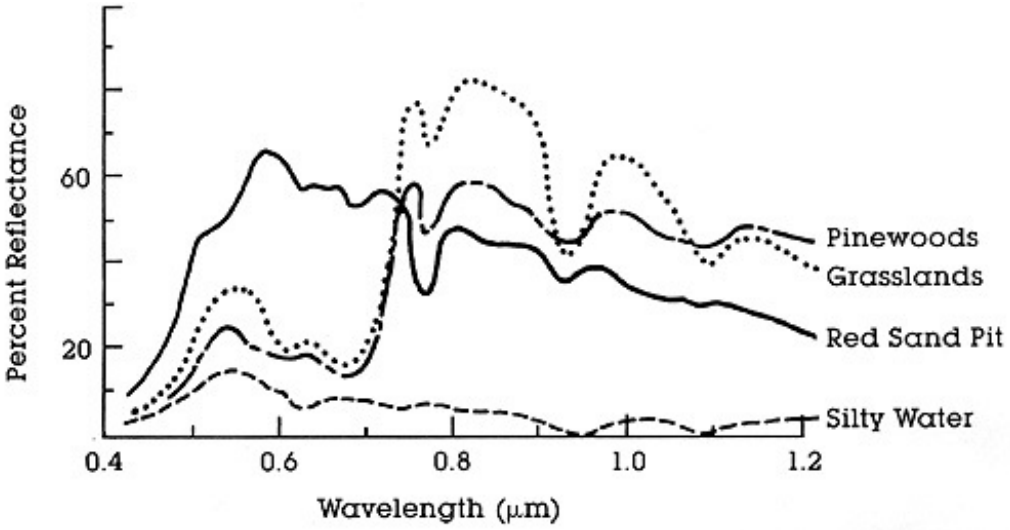
The concept of a "spectral signature", another name for a plot of the variations of reflected (or absorbed) EM radiation as function of wavelengths, gives rise to the widely used approach to identifying and separating different materials or objects using multispectral data obtained by remote sensors.

Electromagnetic Spectrum: Spectral Signatures

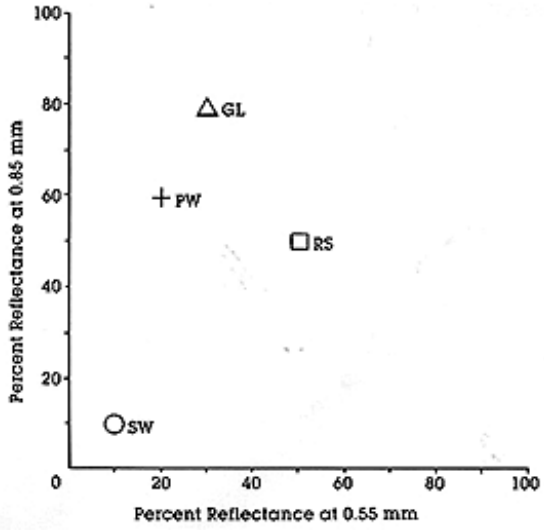
For any given material, the amount of solar radiation that it reflects, absorbs, transmits, or emits varies with wavelength. When that amount (usually intensity, as a percent of maximum) coming from the material is plotted over a range of wavelengths, the connected points produce a curve called the material's spectral signature (spectral response curve). Here is a general example of a reflectance plot for some (unspecified) vegetation type (bio-organic material), with the dominating factor influencing each interval of the curve so indicated; note the downturns of the curve that result from selective absorption:



This important property of matter makes it possible to identify different substances or classes and to separate them by their individual spectral signatures, as shown in the figure below. *



For example, at some wavelengths, sand reflects more energy than green vegetation but at other wavelengths it absorbs more (reflects less) than does the vegetation. In principle, we can recognize various kinds of surface materials and distinguish them from each other by these differences in reflectance. Of course, there must be some suitable method for measuring these differences as a function of wavelength and intensity (as a fraction [normally in percent] of the amount of irradiating radiation). Using reflectance differences, we may be able to distinguish the four common surface materials in the above signatures (GL = grasslands; PW = pinewoods; RS = red sand; SW = silty water) simply by plotting the reflectances of each material at two wavelengths, commonly a few tens (or more) of micrometers apart. Note the positions of points for each plot as a reflectance percentage for just two wavelengths:

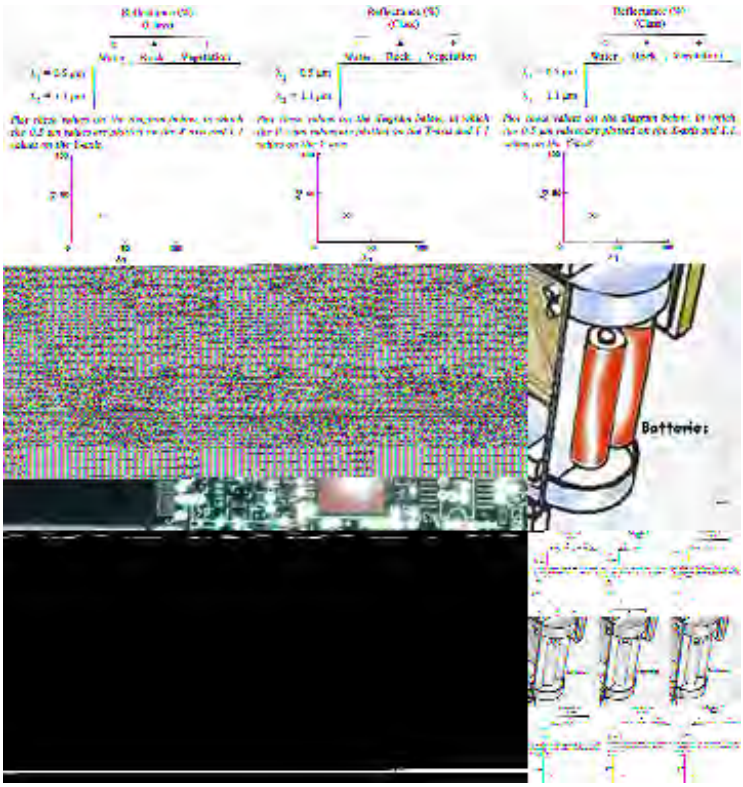


In this instance, the points are sufficiently separated to confirm that just these two wavelengths (properly selected) permit notably different materials to be distinguished by their spectral properties. When we use more than two wavelengths, the plots in multi-dimensional space (3 can be visualized; more than 3 best handled mathematically) tend to show more separability among the materials. This improved distinction among materials due to extra wavelengths is the basis for multispectral remote sensing (discussed on page I-6).

I-11: **Referring to the above spectral plots, which region of the spectrum (stated in wavelength interval) shows the greatest reflectance for a) grasslands; b) pinewoods; c) red sand; d) silty water. At 0.6 micrometers, are these four classes distinguishable? ANSWER**

I-12: **Which material in these plots is brightest at 0.6 micrometers; which at 1.2 micrometers? ANSWER**

I-13 Using these curves, estimate the approximate values of % Reflectance for rock (sand), water, and vegetation (choose grasslands) at two wavelengths: 0.5 and 1.1 micrometers, putting their values in the table provided below. Then plot them as instructed on the lower diagram. Which class is the point at X in this diagram most likely to belong? (Note: you may find it easier to make a copy of the diagram on tracing paper.) **ANSWER**



I-14: **Presume that two unknown surface features in an image or photo, which actually are a forest and a field crop with the plants close-spaced, are measured for their spectral values, and both display quite similar reflectances at three chosen wavelengths. How might these be separated and perhaps even identified? (Hint: think spatially.)** **ANSWER**

Spectral signatures for individual materials or classes can be determined best under laboratory conditions, where the sensor is placed very close to the target. This results in a "pure" spectral signature. But what happens if the sensor is well above the target, as when a satellite remote sensing device looks down at Earth. At such heights the telescope that examines the scene may cover a large surface area at any moment. Individual objects smaller than the field of view are not resolved (this is akin to spatial resolution limitations). Each object contributes its own spectral signature input. In other words, for lower resolution conditions several different materials/classes each send (unresolved) radiation back to the sensor. The resulting spectral signature is a compound of all components in the scene. Analytical techniques (e.g., Fourier analysis) can extract individual signatures under some circumstances. But the sampled area (corresponding to the pixel concept introduced on the next page) is usually assigned a label equivalent to its dominant class). This integration of several signatures is inherent to the "mixed pixel" concept examined in the bottom half of page 13-2.

---

* The principles of spectroscopy in general, as well as a survey of imaging spectroscopy and hyperspectral remote sensing, are explored in greater detail in Section 13 (pages 13-5 through 13-10). Also treated in that part of Section 13 is a brief review of the concept of "spectral resolution".

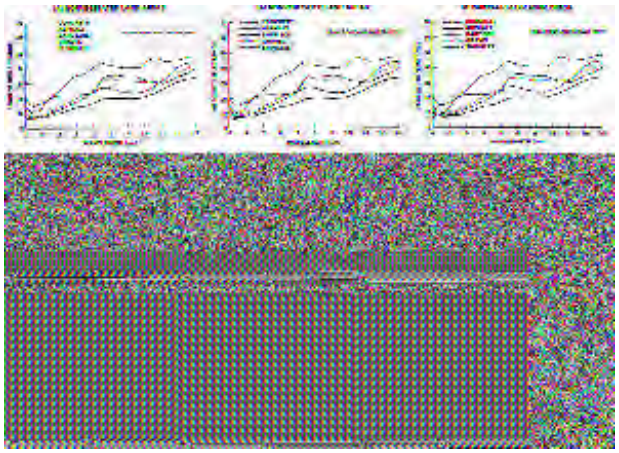
Primary Author: Nicholas M. Short, Sr.

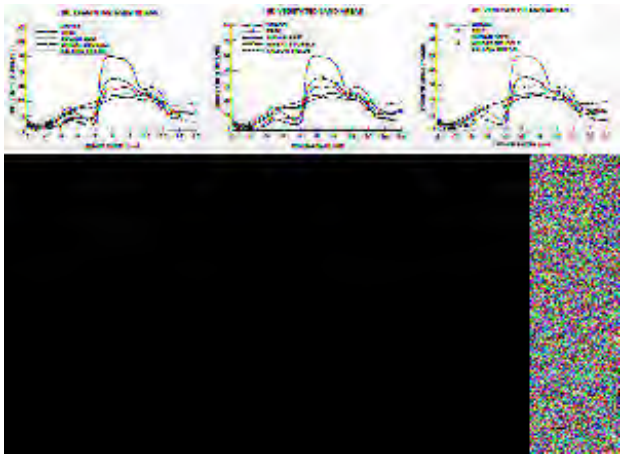
This page takes a quick look at methods involved in processing and interpreting remote sensing data that can be used to define and separate - thus classify - materials, objects, and features. The role of both spatial characteristics and distinctive colors in making classifications is discussed. The more commonly used image processing and display techniques are previewed; these include image enhancement and image classification. The two principal modes of classification - unsupervised and supervised - are described. The notion of ground truth is introduced, as is a quick reference to sensor types used to acquire the data to be analyzed. Finally, the relationship between remote sensing and pattern recognition is mentioned, along with suggestions on how to interpret the analysis results.

Processing and Classification of Remotely Sensed Data; Pattern Recognition; Approaches to Data/Image Interpretation

This part of the Introduction, which has centered on principles and theory underlying the practice of remote sensing, closes with several guidelines on how data are processed into a variety of images that can be classified. This page is a preview of a more extended treatment in Section 1; there is some redundancy relative to preceding pages for the sake of logical presentation. The ability to extract information from the data and interpret this will depend not only on the capabilities of the sensors used but on how those data are then handled to convert the raw values into image displays that improve the products for analysis and applications (and ultimately on the knowledge skills of the human interpreter). The key to a favorable outcome lies in the methods of image processing. The methods rely on obtaining good approximations of the aforementioned spectral response curves and tying these into the spatial context of objects and classes making up the scene.

In the sets of spectral curves shown below (made on site using a portable field spectrometer), it is clear that the spectral response for common inorganic materials is distinct from the several vegetation types. The first (left or top) spectral signatures indicate a gradual rise in reflectance with increasing wavelengths for the indicated materials. Concrete, being light-colored and bright, has a notably higher average than dark asphalt. The other materials fall in between. The shingles are probably bluish, in color as suggested by a rise in reflectance from about 0.4 to 0.5 µm and a flat response in the remainder of the visible (0.4 - 0.7 µm) light region. The second curves (on the right or bottom) indicate most vegetation types are very similar in response between 0.3 - 0.5 µm; show moderate variations in the 0.5 - 0.6 µm interval; and and rise abruptly to display maximum variability (hence optimum discrimination) in the 0.7 - 0.9 µm range, thereafter undergoing a gradual drop beyond about 1.1 µm.





I-15: **At what one wavelength does there appear to be maximum separability of the five Non-vegetated Classes; the five Vegetated Classes?** **ANSWER**

As we have seen, most sensors on spacecraft do not measure the spectral range(s) they monitor fast enough to produce a continuous spectral curve. Instead, they divide the spectral curves into intervals or bands. Each band contains wavelength-dependent input from what is shown in the continuous spectral curve, which varies in intensity (ordinate of the diagram plots), which is combined into a single value - the averaged intensity values over the spectral range present in the interval.

The spectral measurements represented in each band depend on the interactions between the incident radiation and the atomic and molecular structures of the material (pure pixel) or materials (mixed pixel) present on the ground. These interactions lead to a reflected (for wavelengths involving the visible and near-infrared) signal, which changes some as it returns through the atmosphere. The measurements also depend on the nature of the detector system's response in the sensor. Remote sensing experts can use spectral measurements to describe an object by its composition. This is accomplished either by reference to independently determined spectral signatures or, more frequently, by identifying the object and extracting its spectral response in each of the (broad to narrow) bands used (this is complicated somewhat by the degradation of the response by mixed pixels). In the second case, the signature is derived from the scene itself, provided one can recognize the class by some spatial attribute. The wave-dependent reflectances from limited sampling points (individual or small groups of pixels) become the standards to which other pixels are compared (if matched closely, it is assumed those additional pixels represent the same material or class); this approach is called the training site method.

In practice, objects and features on Earth's surface are described more as classes than as materials per se. Consider, for instance, the material concrete. It is used in roadways, parking lots, swimming pools, buildings, and other structural units, each of which might be treated as a

separate class. We can subdivide vegetation in a variety of ways: trees, crops, grasslands, lake bloom algae, etc. Finer subdivisions are permissible, by classifying trees as deciduous or evergreen, or deciduous trees into oak, maple, hickory, poplar, etc.

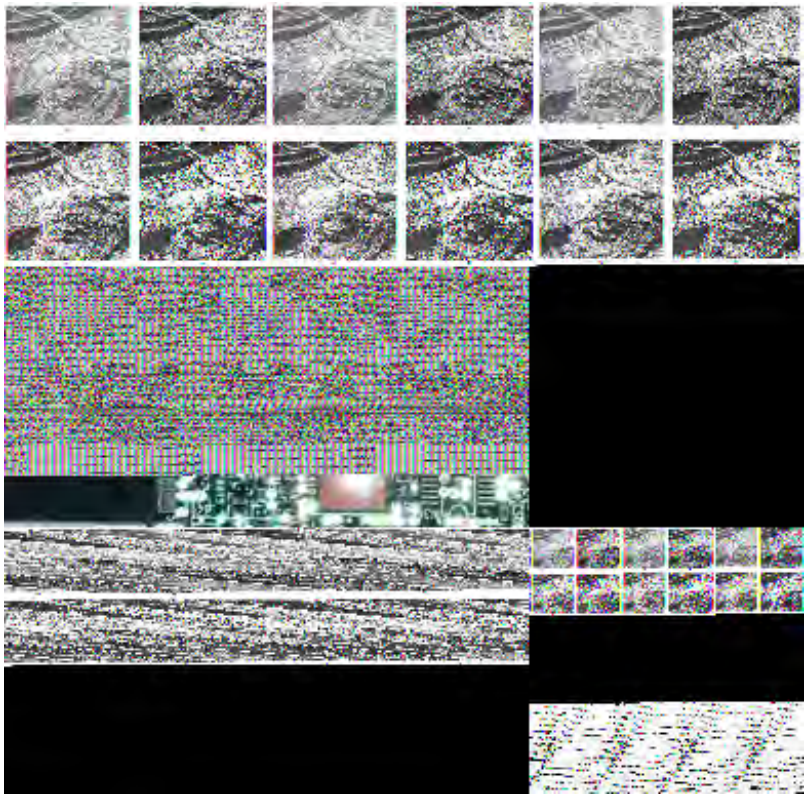
Two additional properties help to distinguish these various classes, some of which have the same materials; namely, shape (geometric patterns) and use or context (sometimes including geographical locations). Thus, we may assign a feature composed of concrete to the classes 'streets' and 'parking lots,' depending on whether its shape is long and narrow or more square or rectangular. Two features that have nearly identical spectral signatures for vegetation, could be assigned to the classes 'forest' and 'crops' depending on whether the area in the images has irregular or straight (often rectangular, the case for most farms) boundaries.

The task, then, of any remote sensing system is to detect radiation signals, determine their spectral character, derive appropriate signatures, and interrelate the spatial positions of the classes they represent. This ultimately leads to some type of interpretable display product, be it an image, a map, or a numerical data set, that mirrors the reality of the surface (affected by some atmospheric property[ies]) in terms of the nature and distribution of the features present in the field of view.

The determination of these classes requires that either hard copy, i.e., images, or numerical data sets be available and capable of visual or automated analysis. This is the function of image processing techniques, a subject that will be treated in considerable detail in Section 1 in which a single scene - Morro Bay on the California coast - is treated by various commonly used methods of display, enhancement, classification, and interpretation. On this page we will simply describe several of the principal operations that can be performed to show and improve an image. It should be worth your while to get a quick overview of image processing by accessing this good review on the Internet, at the Canadian Soonet site.

The starting point in scene analysis is to point out that radiances (from the ground and intervening atmosphere) measured by the sensors (from hand held digital cameras to distant orbiting satellites) vary in intensity. Thus reflected light at some wavelength, or span of wavelengths (spectral region), can range in its intensity from very low values (dark in an image) because few photons are received to very bright (light toned) because of the high reflectances representing much more photons. Each level of radiance can be assigned a quantitative value (commonly as a fraction of 1 or as a percentage of the total radiance that can be handled by the sensor's range). The values are restated as digital numbers (DNs) that consist of equal increments over a range (commonly from 0 to 255; from minimum to maximum measured radiance). As used to engender an image, a DN is assigned some level of "gray" (from all black to all white, and shades of gray in between). When the pixel array acquired by the sensor is processed to show each pixel in its proper relative position and then the DN for the pixel is given a gray tone, a standard black and white image results.

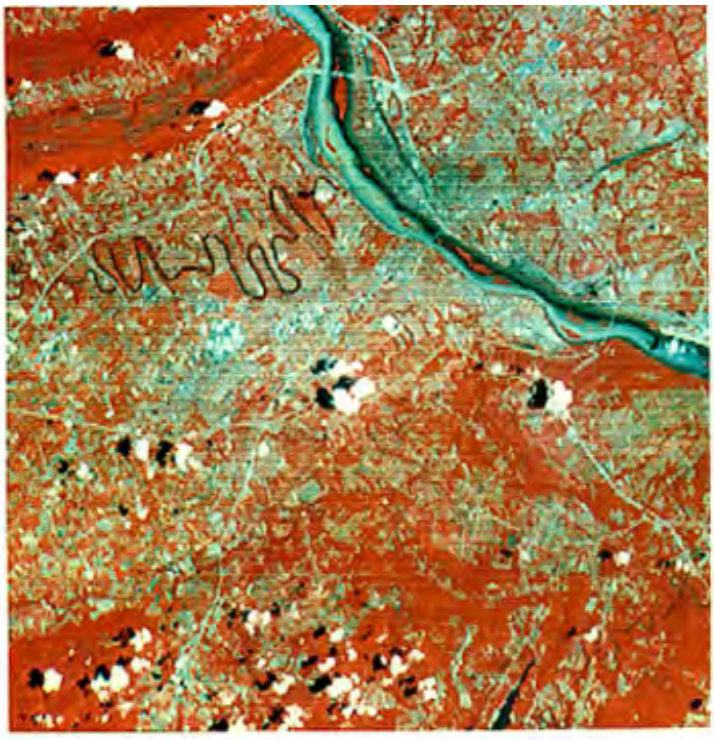
The simplest manipulation of these DNs is to increase or decrease the range of DNs present in the scene and assign this new range the gray levels available within the range limit. This is called contrast stretching. For example, if the range of the majority (say, 90%) of DN values is 40 to 110, this might be stretched out to 0 to 200, extending more into the darkest and the lighter tones in a black and white image. We show several examples here, using a Landsat subscene from the area around Harrisburg, PA that will be examined in detail in the Exam that closes Section 1.



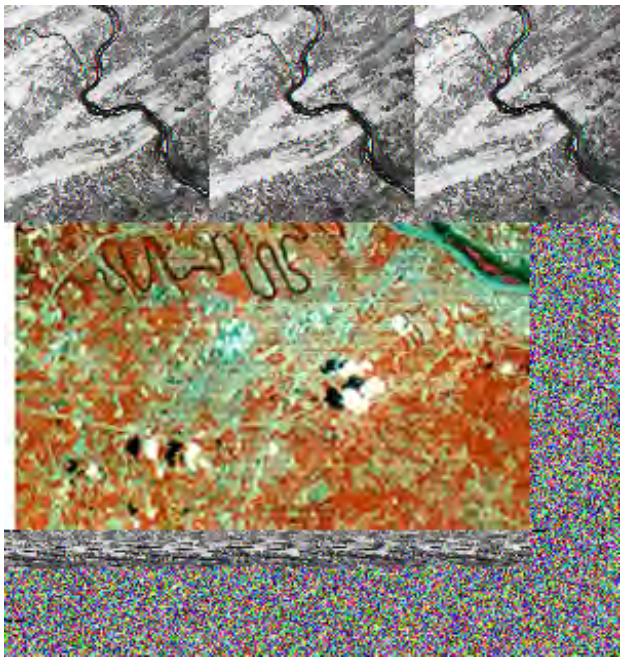
The upper left panel is a "raw" (unstretched) rendition of Landsat MSS band 5. A linear stretch appears in the upper right and a non-linear (departure from a straight line plot of increasing DN values) in the lower left. A special stretch known as Piecewise Linear is shown in the lower right.

Another essential ingredient in most remote sensing images is color. While variations in black and white imagery can be very informative, and were the norm in the earlier aerial photographs (color was often too expensive), the number of different gray tones that the eye can separate is limited to about 20-30 steps (out of a maximum of ~250) on a contrast scale. On the other hand, the eye can distinguish 20,000 or more color tints, so we can discern small but often important variations within the target materials or classes can be discerned. Liberal use of color in the illustrations found throughout this Tutorial takes advantage of this capability; unlike most textbooks, in which color is restricted owing to costs. For a comprehensive review of how the human eye functions to perceive gray and color levels, consult Chapter 2 in Drury, S.A., *Image Interpretation in Geology*, 1987, Allen & Unwin.

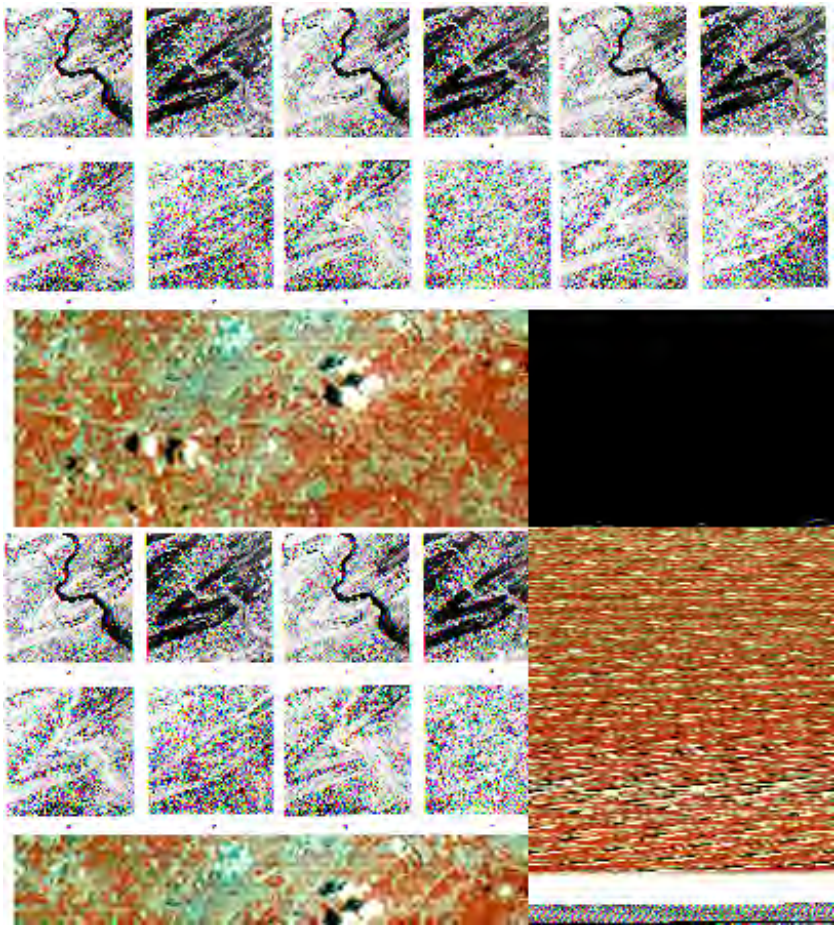
Any three bands (each covering a spectral range or interval) from a multispectral set, either unstretched or stretched, can be combined using optical display devices, photographic methods, or computer-generated imagery to produce a color composite (simple color version in natural [same as reality] colors, or quasi-natural [choice of colors approximates reality but departs somewhat from actual tones], or false color). Here is the Harrisburg scene in conventional false color (vegetation will appear red because the band used displays vegetation in light [bright] tones is projected through a red filter as the color composite is generated):



New kinds of images can be produced by making special data sets using computer processing programs. For example, one can divide the DNs of one band by those of another at each corresponding pixel site. This produces a band ratio image. Shown here is Landsat MSS Band 7 divided by Band 4, giving a new set of numbers that cluster around 1; these numbers are then expanded by a stretching program and assigned gray levels. In this scene, growing vegetation is shown in light gray tones.

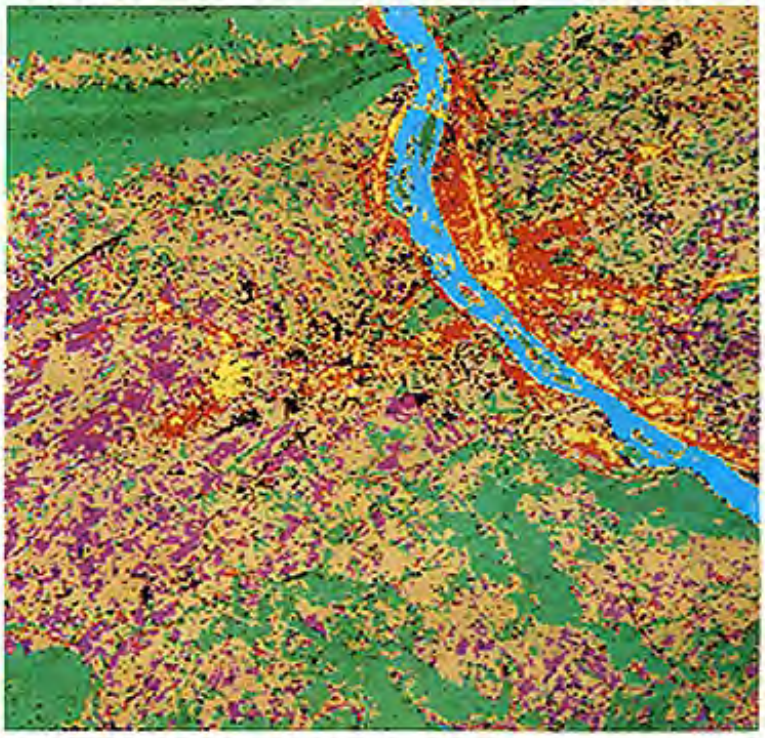


Data from the several bands that are set up from spectral data in the visible and Near-IR tend to be varyingly correlated for some classes. This correlation can be minimized by a reprocessing technique known as Principal Components Analysis. New PCA bands are produced, each containing some information not found in the others. This image shows the first 4 components of a PCA product for Harrisburg; the upper left (Component 1) contains much more decorrelated information than the last image at the lower right (Component 4).



New color images can be made from sets of three band ratios or three Principal Components. The color patterns will be different from natural or false color versions. Interpretation can be conducted either by visual means, using the viewer's experience, and/or aided by automated interpretation programs, such as the many available in a computer-based Pattern Recognition procedure (see below on this page).

A chief use of remote sensing data is in classifying the myriad of features in a scene (usually presented as an image) into meaningful categories or classes. The image then becomes a **thematic** map (the theme is selectable, e.g., land use; geology; vegetation types; rainfall). In Section 1 of the Tutorial we explain how to interpret an image using an aerial or space image to derive a thematic map. This is done by creating an **unsupervised** classification when features are separated solely on their spectral properties and a **supervised** classification when we use some prior or acquired knowledge of the classes in a scene in setting up **training sites** to estimate and identify the spectral characteristics of each class. A supervised classification of the Harrisburg subscene shows the distribution of the named (identified) classes, as these were established by the investigator who knew their nature from field observations. In conducting the classification, representative pixels of each class were lumped into one or more training sites that were manipulated statistically to compare unknown class pixels to these site references:



	AREA [%]
URBAN: INDUSTRIAL/COMMERCIAL	5.8
URBAN: RESIDENTIAL/MIXED	8.1
VEGETATED CROPLAND/PASTURELAND	36.5
BARREN/OPEN LAND	14.2
DECIDUOUS FORESTLAND (SUNNY)	16.1
DECIDUOUS FORESTLAND (SHADED SLOPES)	15.0
WATER (STREAMS AND LAKES)	2.8
UNCLASSIFIED (FINAL ITERATION)	1.9

We mention another topic that is integral to effective interpretation and classification. This is often cited as *reference* or *ancillary* data but is more commonly known as *ground truth*. Under this heading are various categories: maps and databases, test sites, field and laboratory measurements, and most importantly actual onsite visits to the areas being studied by remote sensing. This last has two main facets: 1) to identify what is there in terms of classes or materials so as to set up *training sites* for classification, and 2) to revisit parts of a classified image area to verify the accuracy of identification in places not visited. We will go into ground truth in more detail in the first half of Section 13; for a quick insight switch now to page 13-1.

Another important topic - Pattern Recognition (PR)- will be looked at briefly on this page. Pattern Recognition is closely related (allied) to Remote Sensing and warrants a few comments here. Strangely, a search on the Internet has failed to find any really adequate site that effectively provides an overview or any definitive explanatory illustrations. Two sites that offer some insight are 1)Wikipedia, which presents a brief overview, and 2) McGill University, which contains an impressive list of the scope of topics one should be familiar with to understand and practically use Pattern Recognition methodologies. Elsewhere on the Net this very general definition of PR was found:

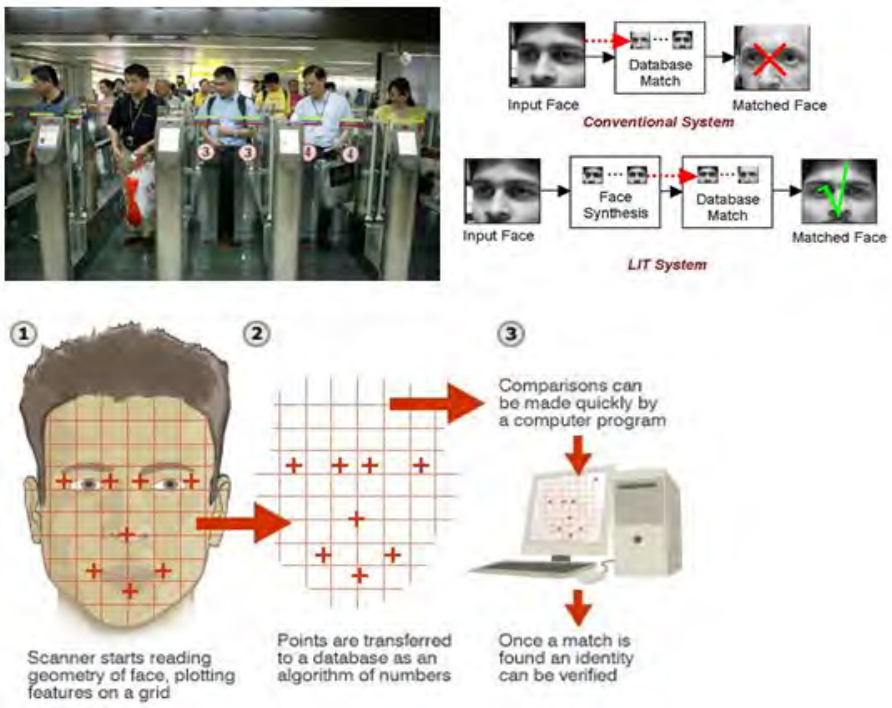
Pattern Recognition: The term refers to techniques for classifying a set of objects into a number of distinct classes by considering similarities of objects belonging to the same class and the dissimilarities of objects belonging to different classes. As an application in computer science, pattern recognition involves the imposition of identity on input data, such as speech, images, or a stream of text, by the recognition and delineation of patterns

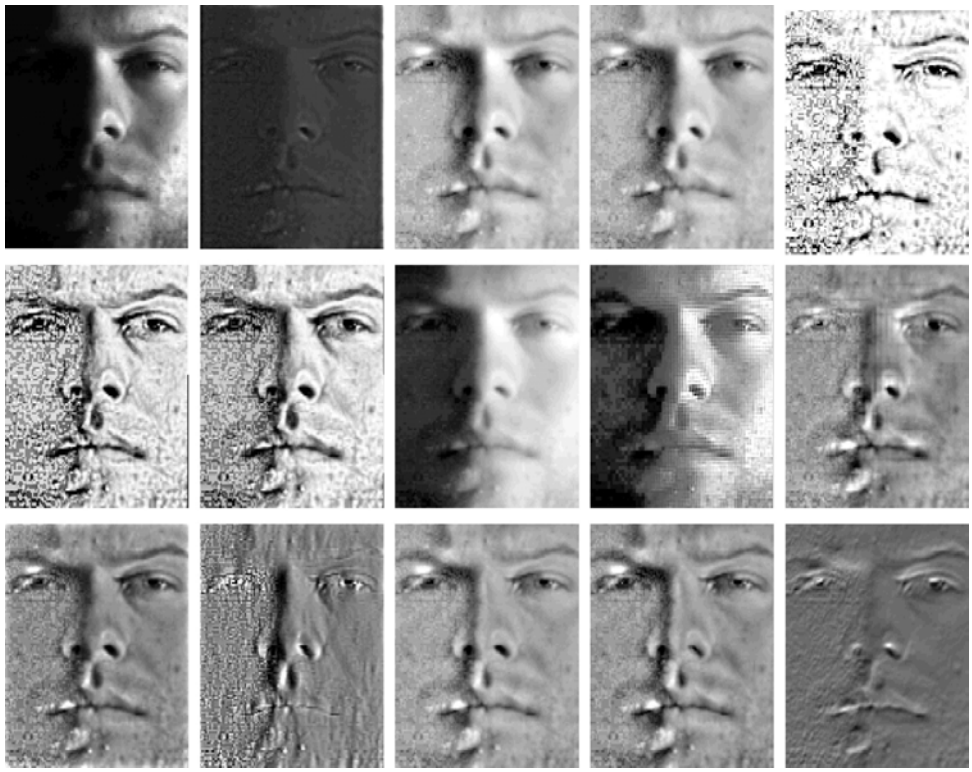
it contains and their relationships. Stages in pattern recognition may depend on measurement of the object to identify distinguishing attributes, extraction of features for the defining attributes, and comparison with known patterns to determine a match or mismatch. Pattern recognition has extensive application in astronomy, medicine, robotics, and remote sensing by satellites.

Pattern Recognition (sometimes referred alternately as "Machine Learning" or "Data Mining") uses spectral, spatial, contextual, or acoustic inputs to extract specific information from visual or sonic data sets. You are probably most familiar with the Optical Character Recognition (OCR) technique that reads a pattern of straight lines of different thicknesses called the bar code:

An optical scanner reads the set of lines and searches a data base for this exact pattern. A computer program compares patterns, locates this one, and ties it into a database that contain information relevant to this specific pattern (in a grocery store, for example, this would be the current price of the product on which the bar code has been included on, say, the package).

One application of Pattern Recognition familiar to many readers of this Tutorial is described by the term "facial recognition". The technique depends on selecting key metrics of features on the human face that can be specified by size, shape, color, spacing, etc. This information is digitized and entered into a data bank. When a new face is encountered, its metrics are compared to the reference set of faces to determine the extent of match. A summary of face recognition is given at the HowStuffWorks Internet site. Aspects of the concept appear in these illustrations:





In the 21st century era of terrorism and crime, surveillance of cities, airports, and other modes of transportation has led to emplacement of TV cameras that scan streets and other open areas. While most of these simply record what is happening in real time, which must be interpreted by human monitors, some are equipped with facial recognition and other types of pattern recognition software that provide rapid analysis of what/who is in the scene.

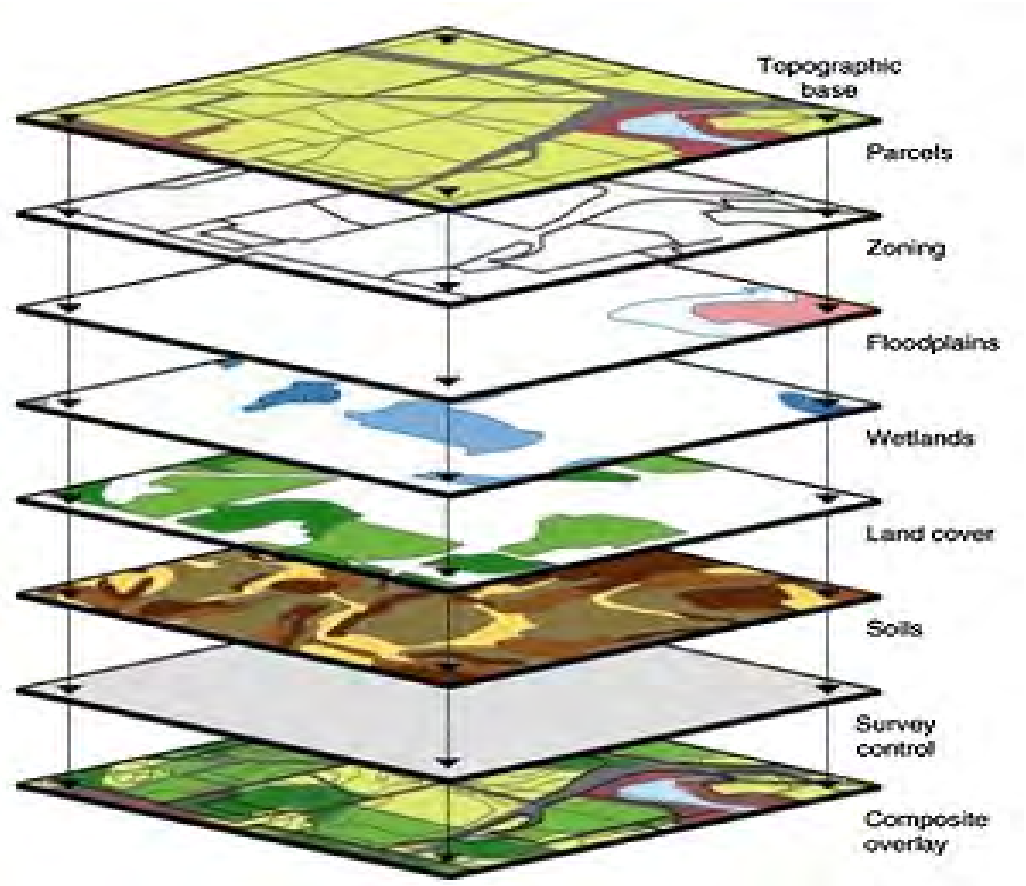
Other examples of PR encountered in today's technological environment include 1) security control relying on identification of an individual by recognizing a finger or hand print, or by matching a scan of the eye to a database that includes only those previously scanned and added to the base; 2) voice recognition used to perform tasks on an automated telephone call routing (alternative: push telephone button number to reach a department or service); 3) sophisticated military techniques that allow targets to be sought out and recognized by an onboard processor on a missile or "smart bomb"; 4) handwriting analysis and cryptography; 5) a feature recognition program that facilitates identification of fossils in rocks by analyzing shape and size and comparing these parameters to a data bank containing a collection of fossil images of known geometric properties; 6) classifying features, objects, and distribution patterns in a photo or equivalent image, as discussed above.

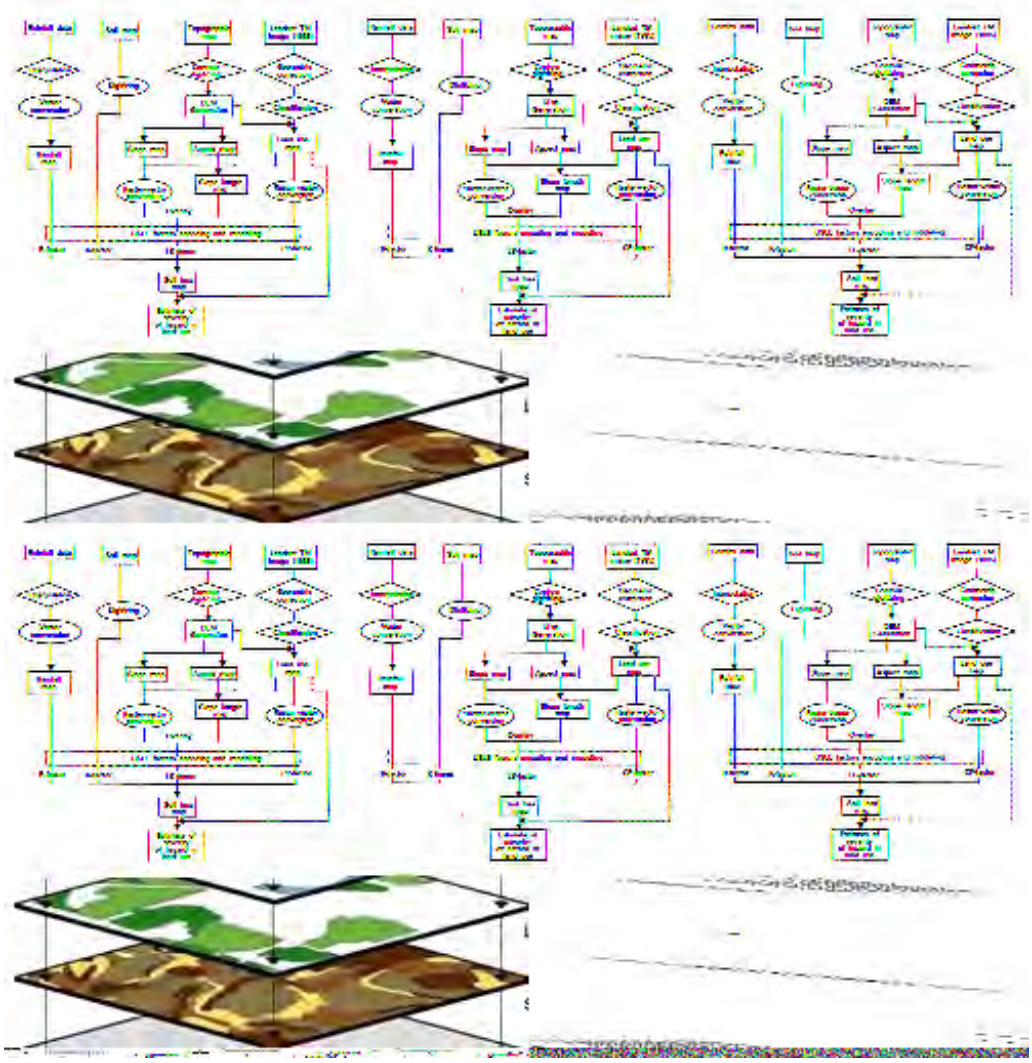
In the Remote Sensing Tutorial pattern recognition is an implicit component of the image processing techniques of Unsupervised and Supervised Classification of geographic and other spatial images acquired by remote sensing satellites, as discussed in Section 2.

Pattern Recognition is a major application field for various aspects of Artificial Intelligence and Expert Systems, Neural Networks, and Information Processing/Signal Processing (all outside the scope of coverage in this Tutorial) as well as statistical programs for decision-making (e.g., Bayesian Decision Theory). It has a definite place in remote sensing, particularly because of its effectiveness in geospatial analysis; however, it is ignored (insofar as the term Pattern Recognition *per se* is concerned) in most textbooks on Remote Sensing. Establishing a mutual bond between RS and PR can facilitate some modes of *Classification*. Pattern Recognition also plays an important role in **Geographic Information System**s (GIS) (the topic is reviewed in detail in Section 15).

GIS is a fast-growing, now widely used approach to analyzing spatial (in the geographic sense) data of many kinds for the purpose of managing systems (such as forest cutting, city planning, crop planting, etc.) that require frequent, ongoing *decisions*. Different types of systems can

have, in common, spatial references, such as geographic coordinates. Each type can be displayed as a map or layer. The different layers can be overlaid using the coordinates as a common means of registration. This is now done almost entirely on computers using software (such as marketed by the ESRI Co.) that facilitates analysis and decision making. Some hint of how this is done (described in detail in Section 15) is afforded by these two diagrams:





All these processing and classifying activities are done to lead to some sort of end results or "bottom lines". The purpose is to gain new information, derive applications, and make action decisions. For example, a Geographic Information System program will utilize a variety of data that may be gathered and processed simply to answer a question like: "Where is the best place in a region of interest to locate (site) a new electric power plant?" Both machine (usually computers) and humans are customarily involved in seeking the answer. Remote sensing data are commonly an integral part and input into GIS.

It is almost self-evident that the primary interpreter(s) will be one person or a group of humans. These must have a suitable knowledge base and adequate experience in evaluating data, solving problems, and making decisions. Where remote sensing and pattern recognition are among the "tools" used in the process, the interpreter must also be familiar with the principles and procedures underlying these technologies and some solid expertise in selecting the right data inputs, processing the data, and deriving understandable outputs in order to reach satisfactory interpretations and consequent decisions. But, with the computer age it has also become possible to have software and display programs that conduct some - perhaps the majority - of the interpretations. Yet these automated end results must ultimately be evaluated by qualified people. As the field of Artificial Intelligence develops and decision rules become more sophisticated, a greater proportion of the interpretation and evaluation can be carried out by the computer programs chosen to yield the desired information. But at some stage the human mind must interact directly.

Throughout the Tutorial, and especially in Section 1 we will encounter examples of appropriate and productive interpretation and decision making.

We will now move on to the second phase of the Introduction: the history of remote sensing with emphasis on satellite systems that employ radiance monitoring sensors.

Primary Author: Nicholas M. Short, Sr.

The chronological development or history of the use of remote sensing from platforms (synonym for spacecraft) that fly or orbit above the Earth's surface is introduced on this page. Brief mention is given to the history of aerial photography. Examples of image products photographed from sounding rockets, from one of the first satellite-mounted remote sensors, and taken by astronauts are presented. Links are provided to three Tables, each outlining some aspect of the history of remote sensing. Although not themselves directly associated with remote sensing, launch vehicles are needed to get the sensors into position. We mention and describe some of the best known: the V-2 modified from World War II weaponry; the Viking and Aerobee; the Delta, Atlas, and Titan rockets; the Apollo program's Saturn V; Russia's Proton and Energia; France's Ariane; the Space Shuttle and its Russian counterpart, Buran.

History of Remote Sensing: In the Beginning

Having now covered some of the principles behind the nature and use of remote sensing data and methodologies, including sensors and image processing, we switch to a survey of the era of satellite remote sensing (and some mention of aircraft remote sensing and space photography by astronauts/cosmonauts) introduced from an historical framework. Special topics near the end will be multiplatform systems, military surveillance, and remote sensing as it applies to medical imaging systems.

Remote sensing as a technology started with the first photographs in the early nineteenth century (see first page of Overview). To learn about the milestones in remote sensing up to the first Landsat, look over the next three pages that cover these three areas - Photographic Methods, Non-Photographic Sensor Systems, Space Imaging Systems (extracted from the Landsat Tutorial Workbook) covering events through 1979. You can also find more on the general history of U.S. and foreign space programs in Appendix A and this Wikipedia online Web site. It is also informative to look at Jack Estes' website that surveys the history of aerial photography. He later revised that with a website that includes the History of remote sensing. Another review of aerial photography is found on the Professional Aerial Photographers Association (PAPA) website, from which several of the illustrations on this page have been downloaded.

We present major highlights subsequent to 1979 both within this Introduction and throughout the Tutorial. Some of these highlights include short summaries of major space-based programs such as launching several other satellite/sensor systems similar to Landsat; inserting radar systems into space; proliferating weather satellites; orbiting a series of specialized satellites to monitor the environment using, among others, thermal and passive microwave sensors; developing sophisticated hyperspectral sensors; and deploying a variety of sensors to gather imagery and other data on the planets and astronomical bodies.

The photographic camera has served as a prime remote sensor for more than 150 years. It captures an image of targets exterior to it by concentrating electromagnetic (EM) radiation (normally, visible light) through a lens onto a recording medium. The Daguerrotype plate was the first of this kind. A key advance in photography occurred in 1871 when Dr. Henry Maddox, a Brit, announced development of a photographic negative made by enclosing silver halide suspended in an emulsion mounted on a glass plate (later supplanted by flexible film that is advanced to allow many exposures). Silver halide film remains the prime recording medium today. The film displays the target objects in their relative positions by variations in their brightness of gray

levels (black and white) or color tones (using dyes, as discussed in Section 10).

Although the first, rather primitive photographs were taken as "stills" on the ground, the idea of photographing the Earth's surface from above, yielding the so-called **aerial photo**, emerged in the 1860s with pictures from balloons. The first success - now lost - is a photo of a French Valley made by Felix Tournachon. One of the oldest such photos, of Boston, appears on the first page of the Overview. The first free flight photo mission was carried out by Monsieur Triboulet in 1879. The next illustration was taken from the PAPA website.

Meanwhile, an alternate approach, mounting cameras on kites, became popular in the last two decades of the 19th Century. E. Archibald of England began this method in 1882. Here is a montage of kite pictures taken from the PAPA website:

G.R. Lawrence took several famous kite photos of the devastation in San Francisco, California right after the infamous 1906 earthquake that, together with fire, destroyed most of the city. The best example was shown in the Overview, first page near the top:

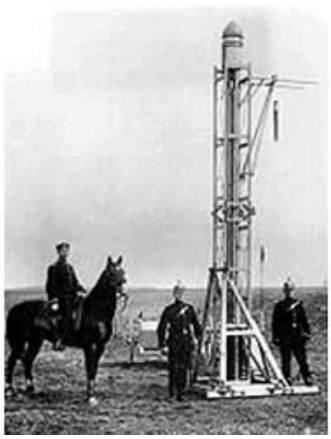
It appears that Wilbur Wright - the co-developer of the first aeroplane to leave the ground in free flight - himself was the first to take pictures from an airplane, in France (LeMans) in 1908 and Italy (Centocelli) in 1909.

By the first World War, cameras mounted on airplanes, or more commonly handheld by aviators, provided aerial views of fairly large surface areas that were invaluable for military reconnaissance. This is documented in these two photos:

From then until the early 1960s, the aerial photograph remained the single standard tool for depicting the surface from a vertical or oblique perspective. More on aerial photography is reviewed on page 10-1.

Historically, the first photos taken from a small rocket, from a height of about 100 meters, were imaged from a rocket designed by Alfred Nobel (of Prize fame) and launched in 1897 over a Swedish landscape; to see this photo, click on the Overview first page.

A camera succeeded in photographing the landscape at a height of 600 meters (2000 ft) reached by Alfred Maul's rocket during a 1904 launch:

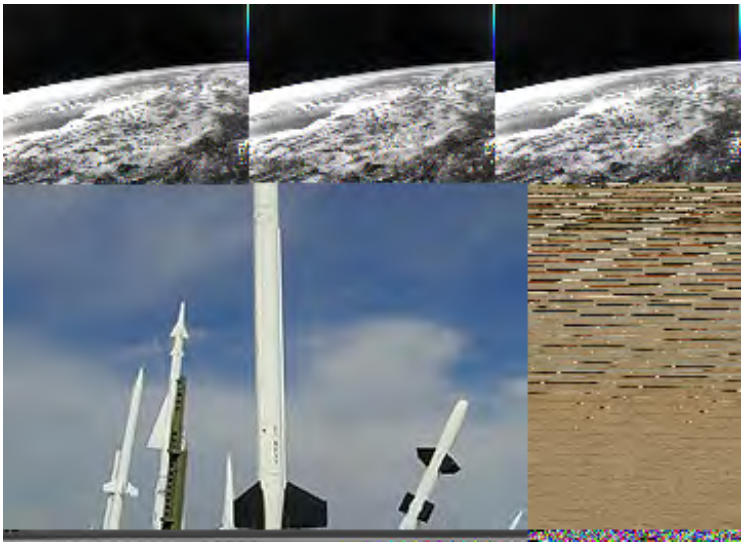


Remote sensing above the atmosphere originated at the dawn of the Space Age (both Russian and American programs). The power and capability of launch vehicles was a big factor in determining what remote sensors could be placed as part (or all) of the payload. By 1946 some V-2 rockets, acquired from Germany after World War II, were launched by the U.S. Army from White Sands Proving Grounds, New Mexico, to high altitudes (70 to 100 miles). The first V-2 launch in the U.S. took place in April 17 of 1946. Launch 13 in the V-2 series on October 24, 1946, contained a motion picture camera in its nose cone, which acquired a series of views of the Earth's surface as it proceeded to a 134 km (83 miles) altitude. We show here a photo of one of the White Sands launches of the V-2 (more V-2 pictures are included in an October 1950 article "Seeing the Earth from Space" in the National Geographic).

Here is a motion picture frame from a later V-2 launch that shows the quality of detail in a scene in nearby New Mexico just north of the White Sands launching site:

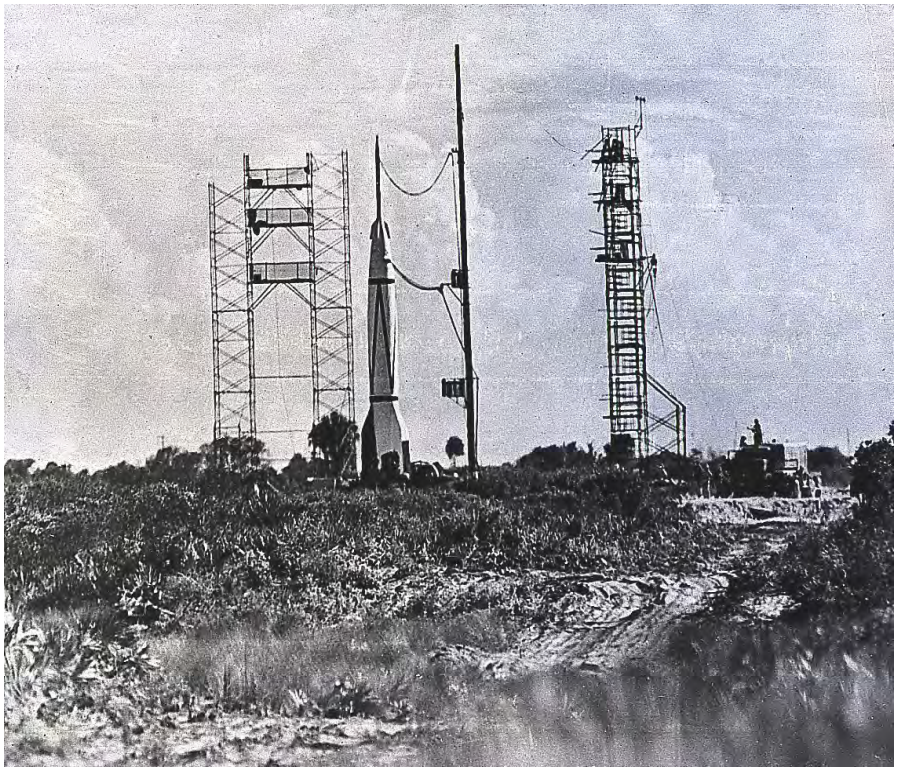
Smaller sounding rockets, such as the Wac Corporal, and the Viking and Aerobee series, were developed and launched by the military in the late '40s and '50s. These rockets, while not attaining orbit, contained automated still or movie cameras that took pictures as the vehicle ascended. In these early days there were many variants of sounding rockets, along with those being groomed for eventual insertion of objects into orbit. An outdoor display of these at the White Sands Museum is impressive:

Here is an example of a typical oblique picture made during a Viking Flight in 1950, looking across Arizona and the Gulf of California to the curving Earth horizon (this photo is shown again in Section 12).

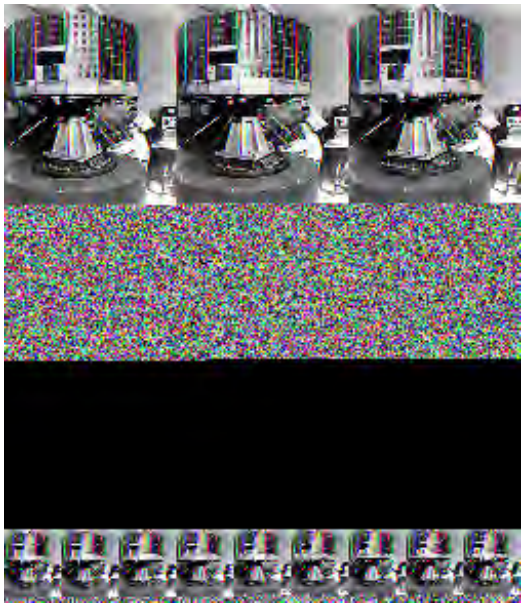


The White Sands Proving Grounds entered the writer's (NMS) life briefly in 1947. At the time I was in the U.S. Army at Fort Bliss. One of my duties was as a reporter on the Fort Bliss

newspaper. I was assigned to cover a V-2 launch at White Sands. There I met Dr. Wernher von Braun and interviewed him (met him again about 40 years later and he remembered that occasion). We watched the launch from a block house and then ran outdoors to see it disappear into the sky. This photo reminds me of the setting at the site:



Now, let us return to satellites that do remote sensing to consider some of the early vehicles. The first non-photo sensors were television cameras mounted on unmanned spacecraft and were devoted mainly to looking at clouds. The first U.S. meteorological satellite, TIROS-1, launched by an Atlas rocket into orbit on April 1, 1960, looked similar to this later TIROS vehicle.

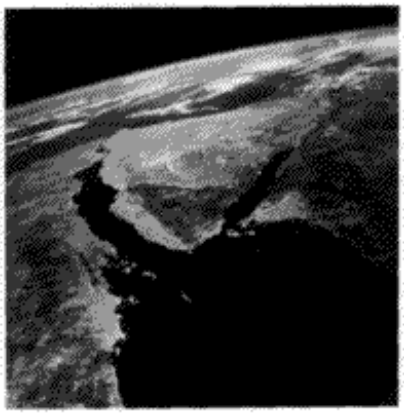


TIROS, for Television Infrared Observation Satellite, used vidicon cameras to scan wide areas at a time. The first television picture ever from a satellite was this TIROS view:

THIS IMAGE CAN BE NOMINATED AS THE OFFICIAL STARTING MOMENT OF REMOTE SENSING FROM SATELLITES.

Images quickly improved. The image below from May 9, 1960 was also returned by a TIROS spacecraft. Ten satellites in this series were flown, followed by the TOS and ITOS spacecraft, along with Nimbus, NOAA, GOES and others (see Section 14). Superimposed on the cloud patterns is a generalized weather map for the region; this kind of data display soon started to appear in television news broadcasts.

Then, in the 1960s as man entered space, cosmonauts and astronauts in space capsules took photos out the window. In time, the space photographers had specific targets and a schedule, although they also have some freedom to snap pictures at targets of opportunity. (As an aside, Scott Carpenter, the second American to orbit the Earth [after John Glenn], told the writer [NMS] that he became so excited by the views from space outside the capsule window, which he furiously photographed, that he forgot for a moment to flip a switch related to re-entry; as a result, the spacecraft reached the Pacific ocean about 50 miles beyond the planned touch down point. His tourist's enthusiasm earned him a "chewing out".) Here are three early astronaut photos:

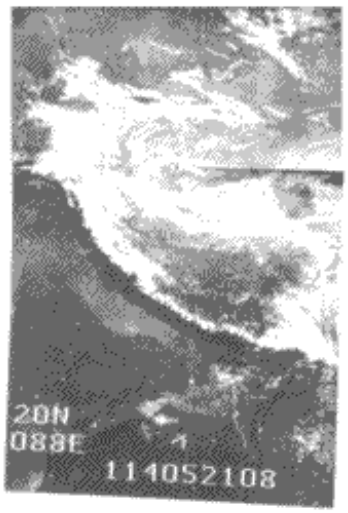


Sinai Peninsula and Red Sea

Gulf of California & Southern California

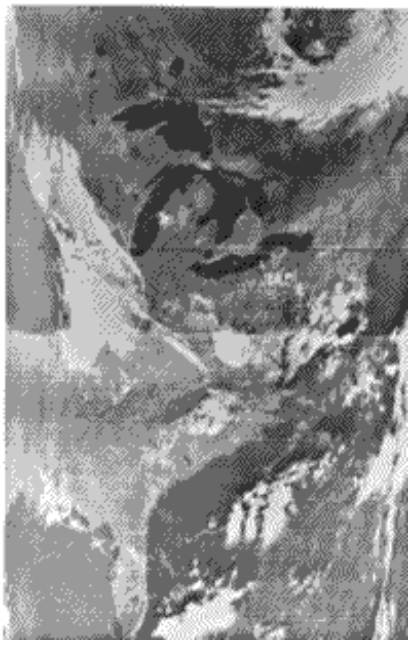
Note: these images, and others in the Tutorial, have a thin blue border; this means that you can click inside it and it will automatically enlarge to fill most of your screen. To close, try pressing escape, then close, using the X button in the upper right, or on some browsers, just try X first.

During the '60s, the first sophisticated imaging sensors were incorporated in orbiting satellites. At first, these sensors were basic TV cameras that imaged crude, low resolution (little detail) black and white pictures of clouds and Earth's surface, where clear. Resolution is the size of the smallest contrasting object pairs that can be sharply distinguished. Below, we show three examples from the Nimbus satellite's sensors to give an idea of how good the early photos were.



Eastern India, Bangladesh, Himalayas

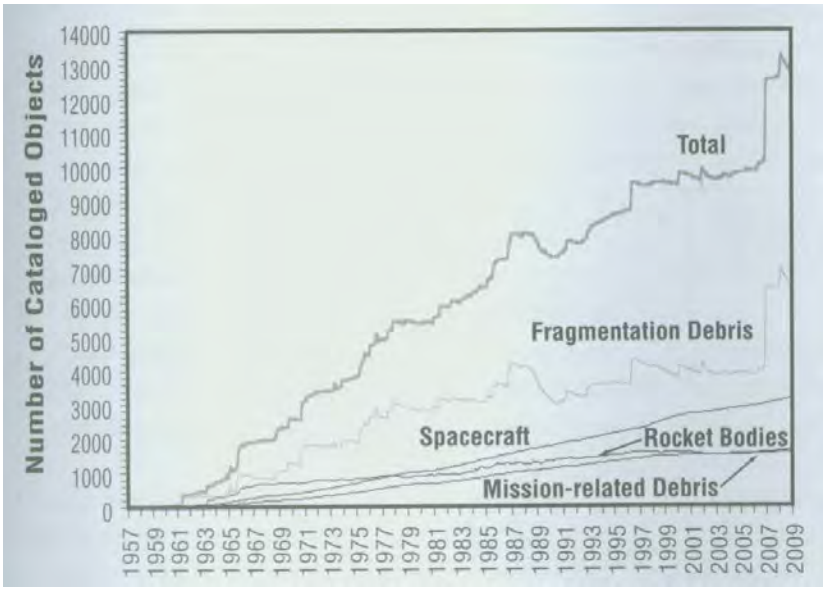
Saudi Arabia

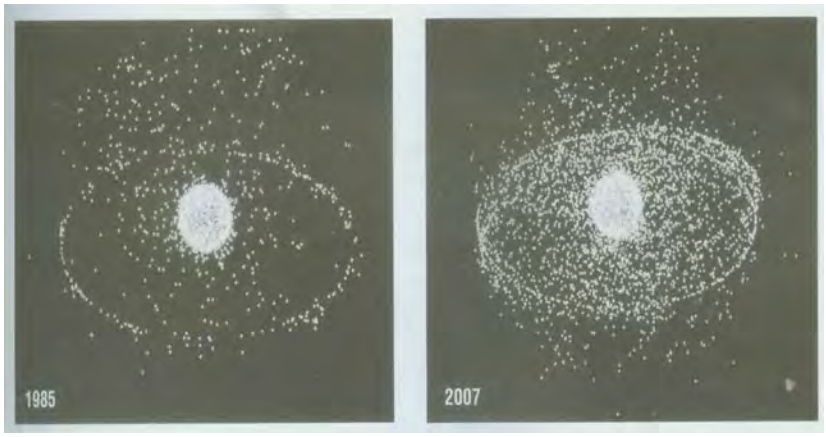


Eastern North America

Early on, other types of sensors were developed that " took images using the EM spectrum beyond the visible, into the near and thermal infrared regions. The field of view (FOV) was broad, usually 100s of kilometers on a side. Such synoptic areas of regional coverage were of great value to the meteorological community, so that many of these early satellites were metsats, dedicated to gathering information on clouds, air temperatures, wind patterns, etc.

We close this page with a subject now in the news rather often, one that is especially a concern during manned missions: the rapid increase in "space junk" which can damage satellites and threaten the International Space Station, the Shuttle and other vehicles carrying astronauts/cosmonautes. These two illustrations zero in on the problem:





The objects shown in the plots are those greater than 10 cm in minimum dimension. The sudden spike in 2007 was due to two events: 1) the destruction in 2007 of a Feng Yun 1C weather satellite by a Chinese missile as an anti-satellite weapons test; and 2) the collision between an active Iridium 33 satellite and an inactive COSMOS 2251 payload. These destructive occurrences validate the growing threats that "space junk" increasingly poses. Since the start of the space program, more than 4600 satellites have been inserted into earth-orbit; about 800 of these are presently still operating. The problem is partially alleviated by deliberately slowing down a satellite so that it falls towards Earth and burns up in the lower atmosphere. A nice summary of the hazards involved can be found at this Wikipedia website.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

II. Non-Photographic Sensor Systems

1800	Discovery of the IR spectral region by Sir William Herschel.
1879	Use of the bolometer by Langley to make temperature measurements of electrical objects.
1889	Hertz demonstrated reflection of radio waves from solid objects.
1916	Aircraft tracked in flight by Hoffman using thermopiles to detect heat effects.
1930	Both British and Germans work on systems to locate airplanes from their thermal patterns at night.
1940	Development of incoherent radar systems by the British and United States to detect and track aircraft and ships during W.W.II.
1950's	Extensive studies of IR systems at University of Michigan and elsewhere.
1951	First concepts of a moving coherent radar system.
1953	Flight of an X-band coherent radar.
1954	Formulation of synthetic aperture concept (SAR) in radar.
1950's	Research development of SLAR and SAR systems by Motorola, Philco, Goodyear, Raytheon, and others.
1956	Kozyrev originated Frauenhofer Line Discrimination concept.
1960's	Development of various detectors which allowed building of imaging and non-imaging radiometers, scanners, spectrometers and polarimeters.
1968	Description of UV nitrogen gas laser system to simulate luminescence.

Primary Author: Nicholas M. Short, Sr.

EXAM 1

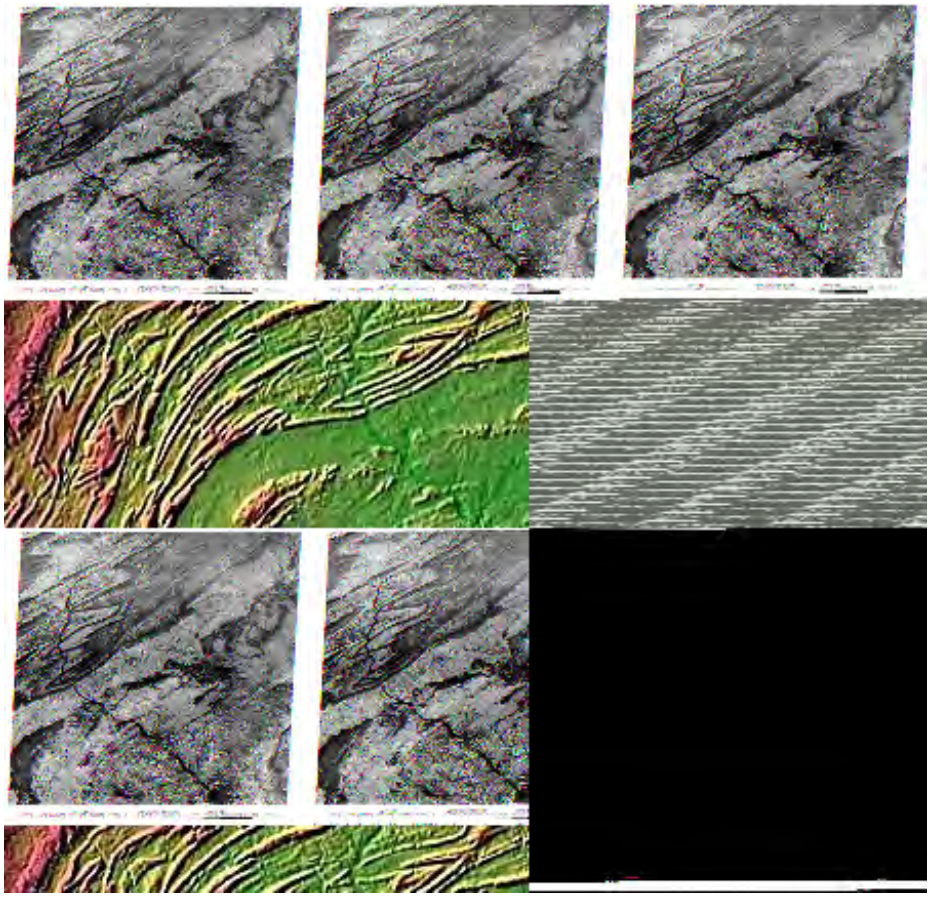
This is your first of two major interpretive Exams in this Tutorial, although subsequent Sections will have a fair number of questions of their own. What is different from earlier questions is that most here will be associated with images and maps included in this special subsection, which was drawn from the writer's (NMS) 1982 *Landsat Tutorial Workbook*. They all relate to a single regional scene that includes much of south-central Pennsylvania and in several images small parts of Delaware and Maryland. Most images are made by the Landsat MSS (remember, Band 4 = green part of the spectrum; Band 5 = red part; and 6 and 7 = Near Infrared). Two thermal and two radar images, made by other sensor systems, are used as well. Some of the image products are computer-derived outputs such as spatial filtered images and classifications.

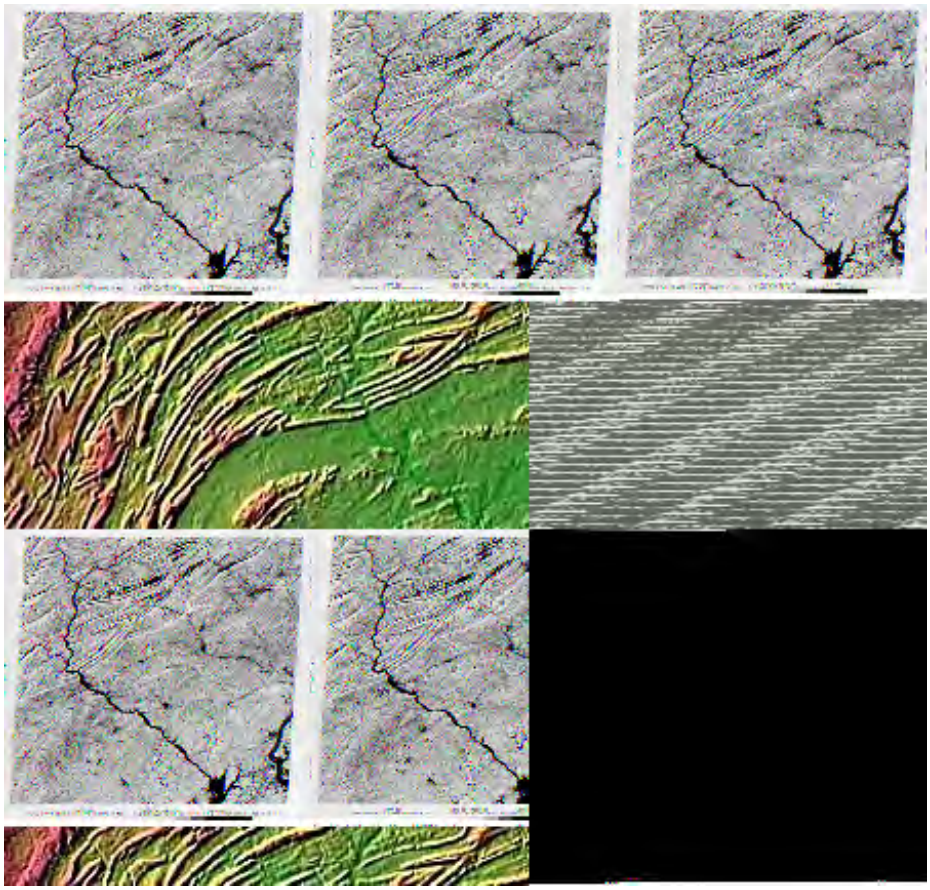
We begin by showing first a satellite view of the entire state, made by the AVHRR sensor on one of the meteorological satellites. For the fun of it, locate Pittsburgh and Philadelphia (they appear in a lighter blue tone).

Then, again to set the region we will be working in into context, we will look at much of the eastern half of the state in terms of its topography, using a color rendition of elevations (red is higher) made from the national DEM (Digital Elevation Model) data. Fit subsequent scenes into this as a reference; also fit it into the above AVHRR image.

Now, we start with an MSS 5 image, reproduced on your screen at a nearly 1:1,000,000 scale - the standard size in images distributed by NASA and other agencies and sold commercially.

Below it is the corresponding MSS 7 image. And beneath that is a false color composite of the same scene. All were acquired on October 11, 1972, a day after a major storm had hit much of the northeast as far south as southern Virginia. After the sky cleared, the air was especially crisp and dry, cutting down the scene-degrading effects of moisture.





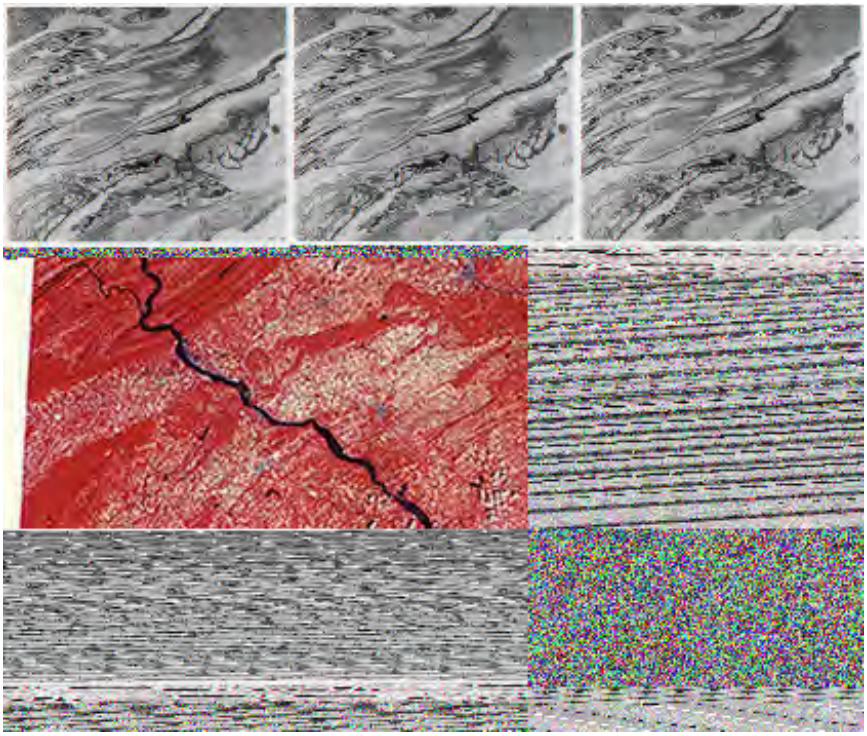
1-1: Using all three images, but especially the Band 7 image, with the help of an atlas or AAA Road Map, locate the following towns and other features: Harrisburg, York, Gettysburg, Hanover, Lancaster, Hershey, Lebanon, Reading, Allentown, Easton, Wilmington, DE, the Susquehanna River, the Delaware River; the Juniata River. Also, look hard for the small (15000+) town, Bloomsburg, PA ,which is the current home of the writer (NMS), where this question is being written. ANSWER

1-2: What other towns, rivers, or natural features and points of interest in this scene can you list?ANSWER

1-3: What band(s) seems best for picking out the towns and rivers in each of the images? What is MSS 5 especially good at bringing to your attention. ANSWER

To give you a feel for the typical appearance of the mountain ridges portion of the image, glance at this ground photo taken about 15 km (10 miles) north of the Blue Mountain - the first major ridge encountered going north from the Lebanon Valley. The area photographed, on a February 1999 day when the trees are bare, is along Interstate 81 and lies near the center of the Landsat image.

Now, look at the black and white print of the geological map below that covers most of this scene, plus some surrounding areas.



1-4: Fit the major rock unit patterns to the Band 5 image. Comment on the fit. ANSWER

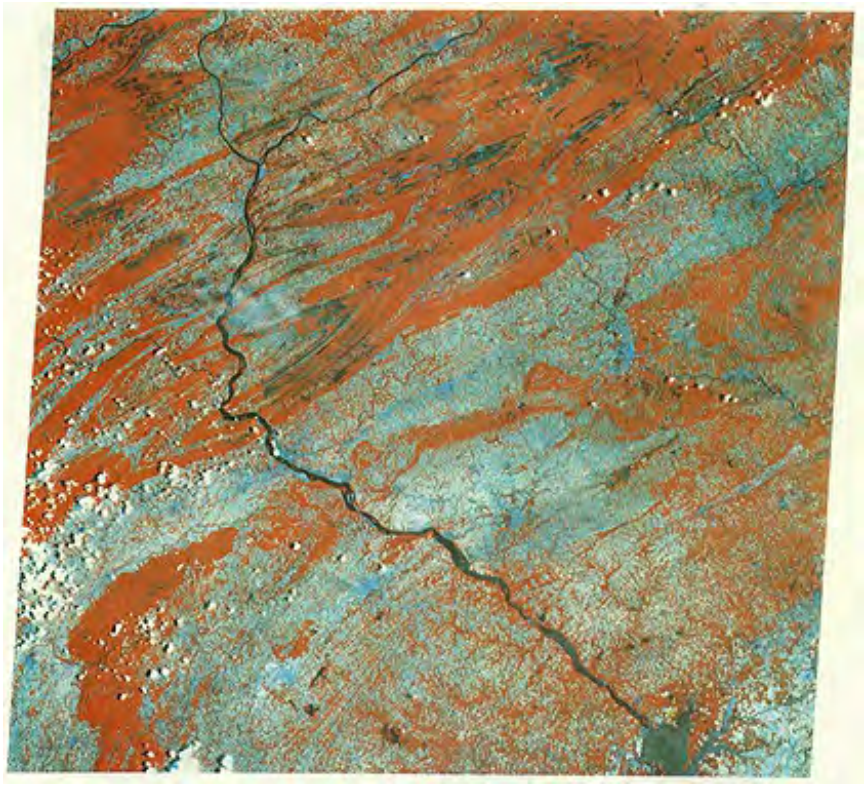
1-5: Comment on what may be unusual about (the color in) the false color composite. ANSWER

1-6: You may have noted that there is a strip of notation at the bottom of both the individual band images and the false color composite. This is on nearly all renditions of Landsat images regardless of the printed size. We will reproduce this strip, enlarged a bit, just below. Make an educated guess as to what the different groupings of letters/numbers mean. ANSWER

Next, familiarize yourself with this October 23, 1975 false color composite of the same area.

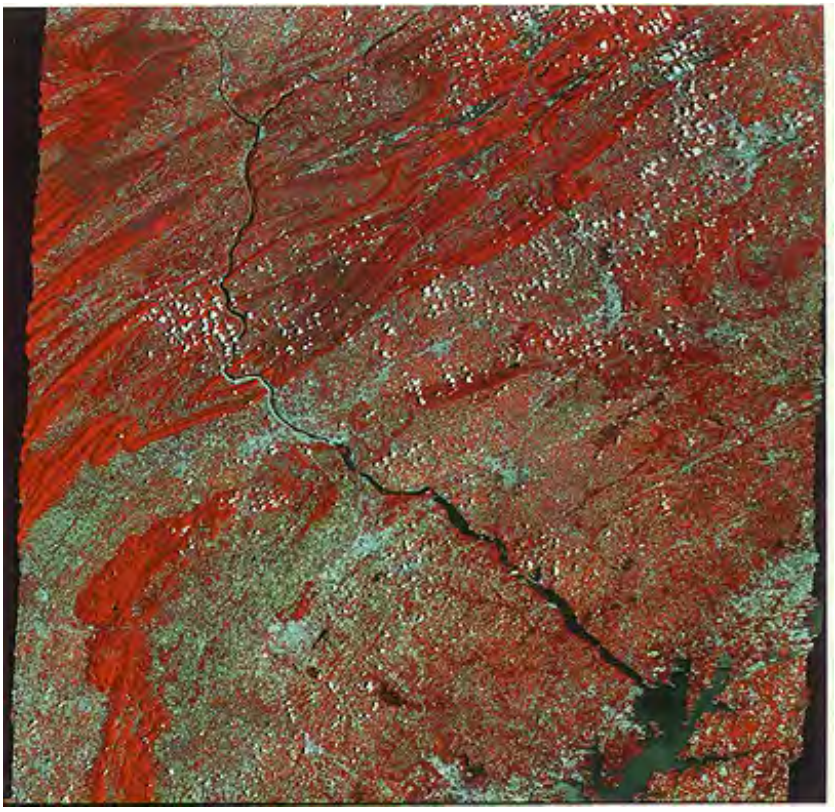
1-7: What overall is different and why? Account for the blue in the Susquehanna River. ANSWER

1-8: Now examine the full MSS false color image of this region as seen on June 8, 1977. Account for the large areas in blue. Take note of the dark grayish-black patches on some of the ridges; there are also some blackish areas in the valleys between ridges, primarily in the upper third of the image. We will take note of these special features later in this exam. Note the course of the Susquehanna River in the lower third of the image. At several places the river widens noticeably. Can you explain this. ANSWER

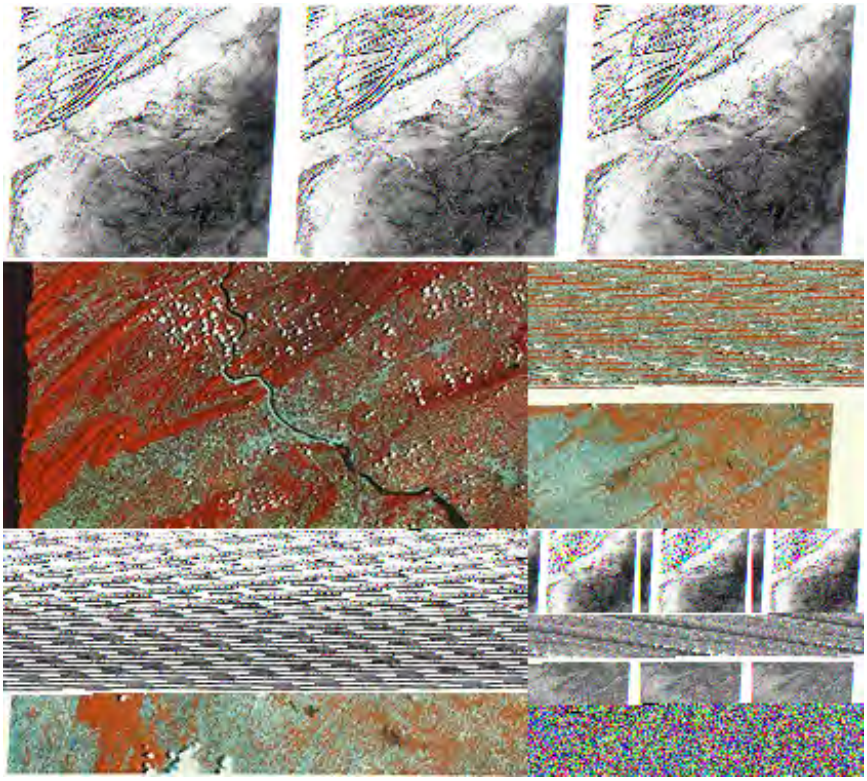


1-9: Below is a scene taken on July 14, 1977. Describe the principal difference between this and the June 8th image. Look also in the vicinity of Harrisburg at the Susquehanna

River. It is marked by a blue color on its left side; why? ANSWER



1-10: A Landsat 2 MSS Band 5 scene, taken on January 23, 1979, is reproduced below. Comment on differences you note in this image from those we've examined before. How might one distinguish snow from clouds? ANSWER

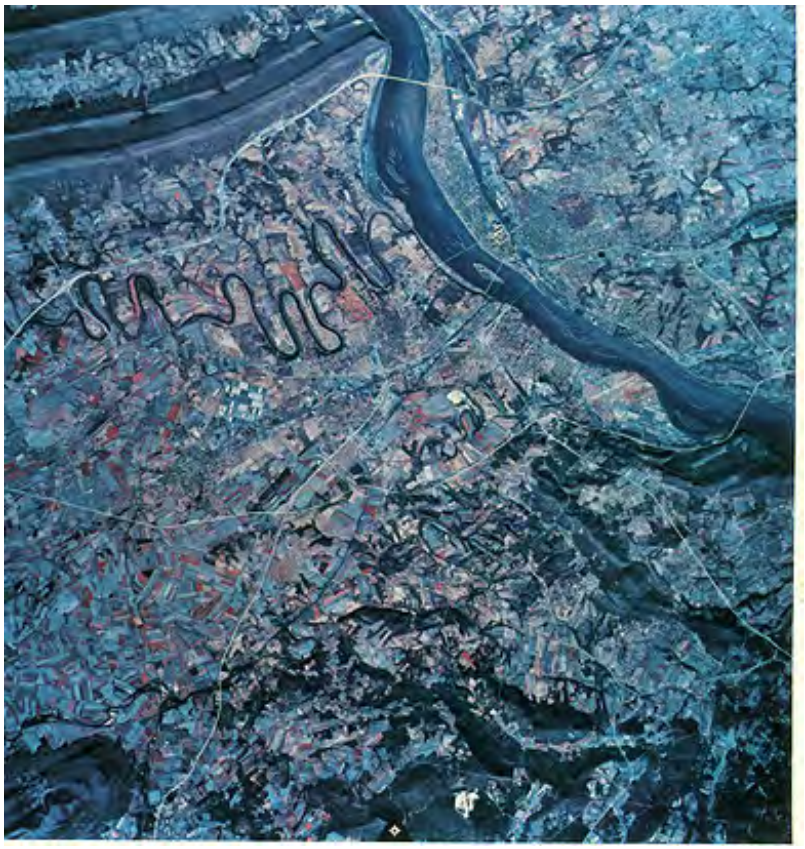


At this point, we will zero in on a single localized region, around Harrisburg, to get a better feel for the detail that can be picked out in Landsat MSS imagery (would be much sharper for Landsat TM images with better than 2 times higher resolution but a TM image was not available to the writer).

But first, to get up close and personal, look at this ground scene taken from the I-83 bridge

looking northeast at downtown Harrisburg. The bridge carries a railroad track across the Susquehanna River. At the far left skyline is a part of Blue Mountain and the water gap through which the river flows. Find the dome of the State Capitol building.

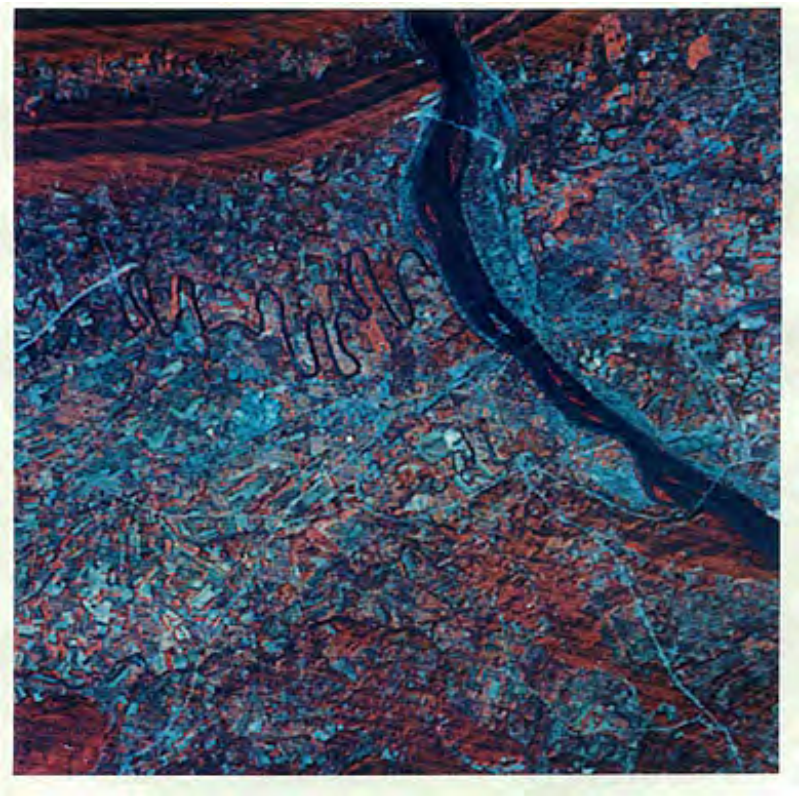
We start our survey from above with a false color aerial photo (scale 1:142,000) taken from an altitude of 50000 feet by a camera mounted on an RB-57's lower bay, on February 5, 1974.



1-11: Peruse this image. Locate Harrisburg first, then using an Atlas pinpoint other towns, major roadways, and the meandering (curving) river. Mention anything else you find interesting. ANSWER

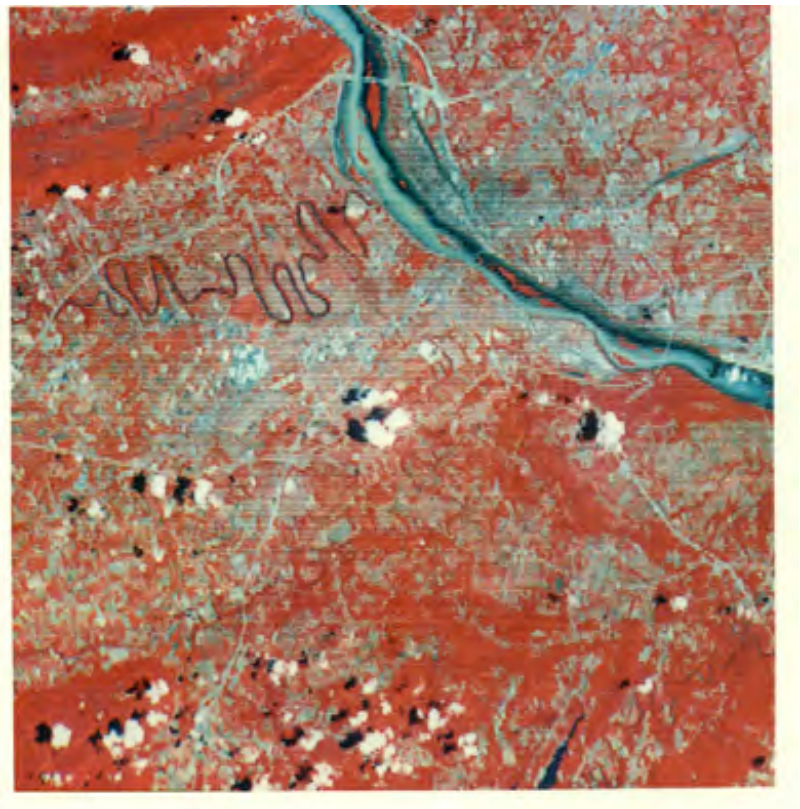
1-12:Knowing that the standard Landsat full scene (7.25 inches wide at its base, representing 115 miles across) has a scale of 1:1,000,000, calculate the width in miles of this aerial photo scene. ANSWER

Just below is a subscene from the October 11, 1972 Landsat MSS false color composite which was enlarged to about the same scale as the RB-57 scene above. It was made more blue to look more like the aerial photo.



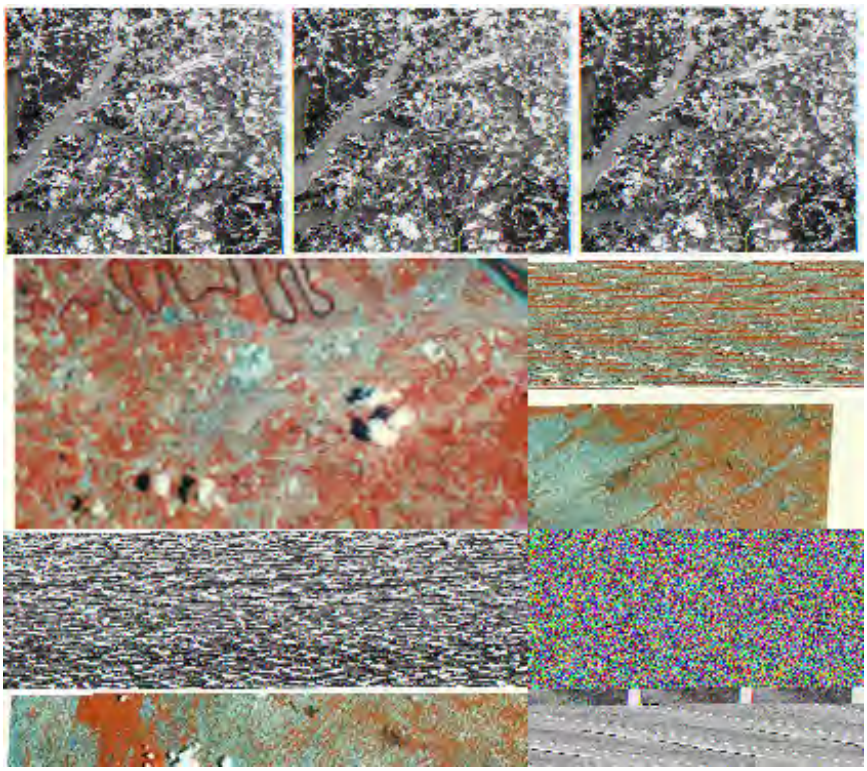
1-13: Compare this subscene with the RB-57 photo, noting what can be found in both and what is not "seeable" in the Landsat version. ANSWER

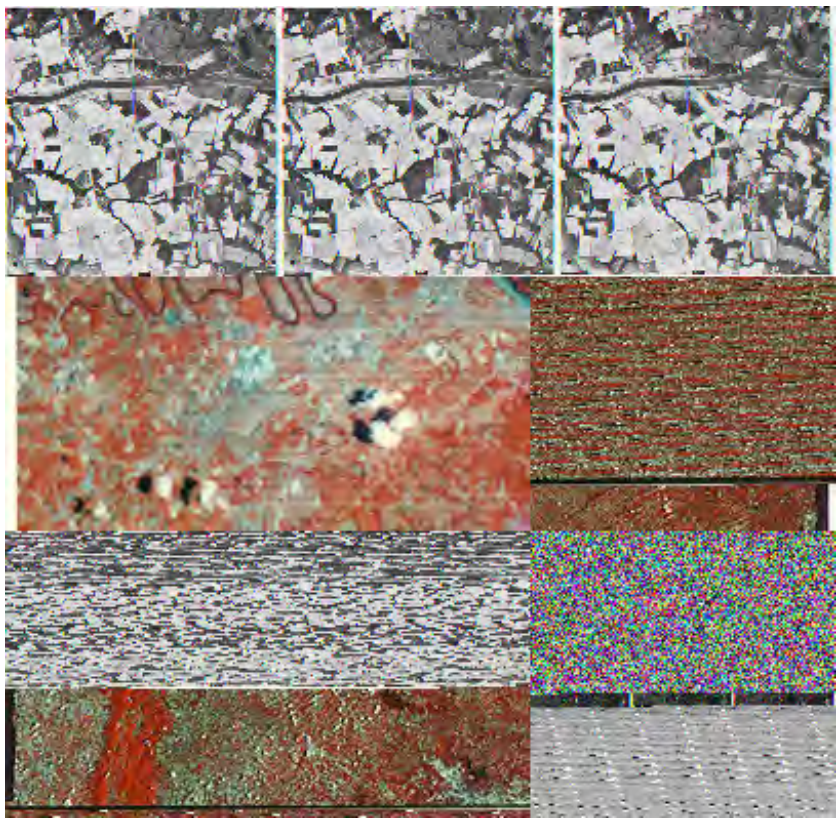
Still another Harrisburg subscene is this July 1977 false color composite.



1-14: Comment on anything unusual or different in the above image. ANSWER

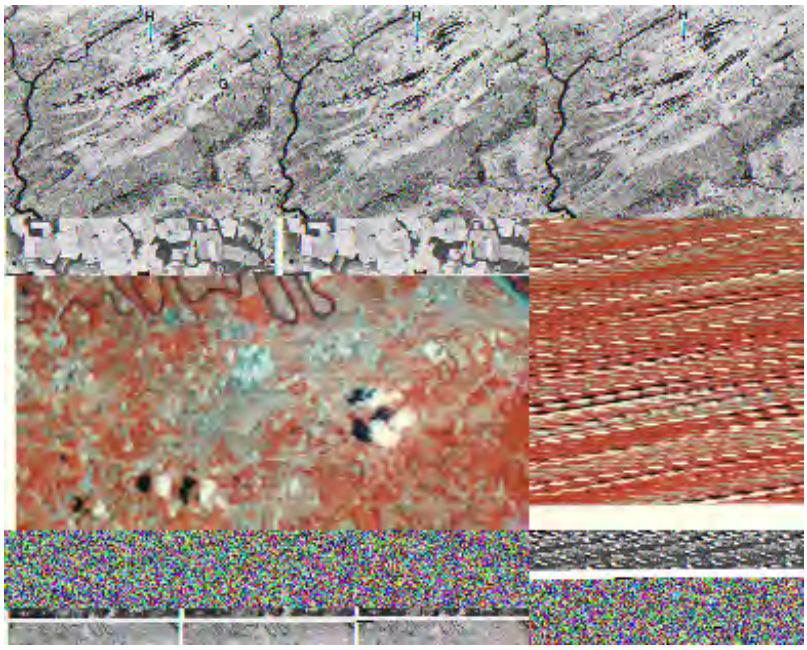
We will now look closer at details present in enlargements, concentrating on the area around the Chesapeake & Delaware Canal, south of Elkton, MD. (This is at the lower right corner of the October 1972 scene.) The first of two images below is a 512 x 512 pixel subscene from Band 5 of the July 1977 image set. Beneath it is an aerial photo, at a scale of 1:60,000 that contains farmlands included in the Landsat image; this was taken in May of the same year.





1-15: Can you locate the aerial photo in the Landsat subscene? Watch out - the field patterns have changed between the two dates for an obvious reason. ANSWER

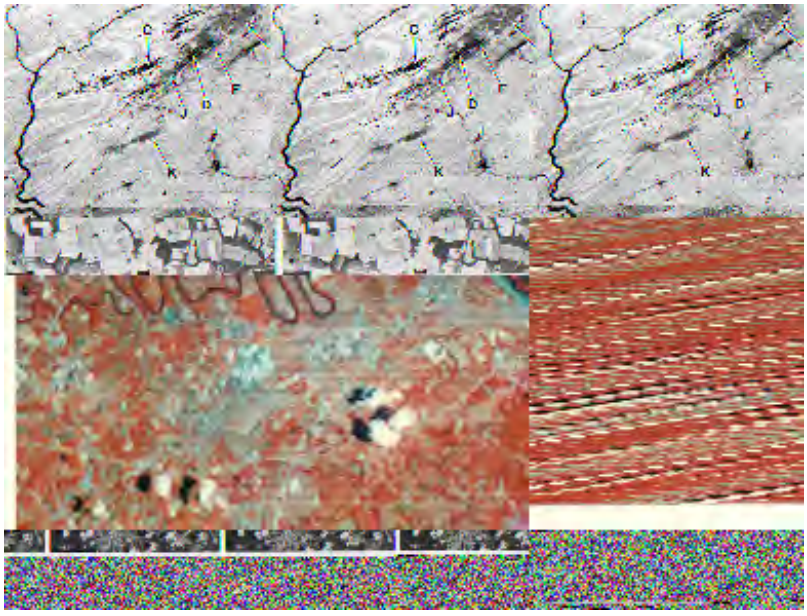
We are ready to change directions to concentrate on a prime use of imagery of photo-like nature: change detection. We are going to utilize several subscenes that have been cut from the upper half of the full images we have already encountered. Consider first this June 8, 1977 Band 7 image.



1-16: One thing that is obvious: widespread patches of almost black features, found almost entirely in valleys between ridges. Do you have any idea what these represent? Hint: they are related to a geological material that has immense commercial value; this material is in beds in rocks of Pennsylvanian age. ANSWER

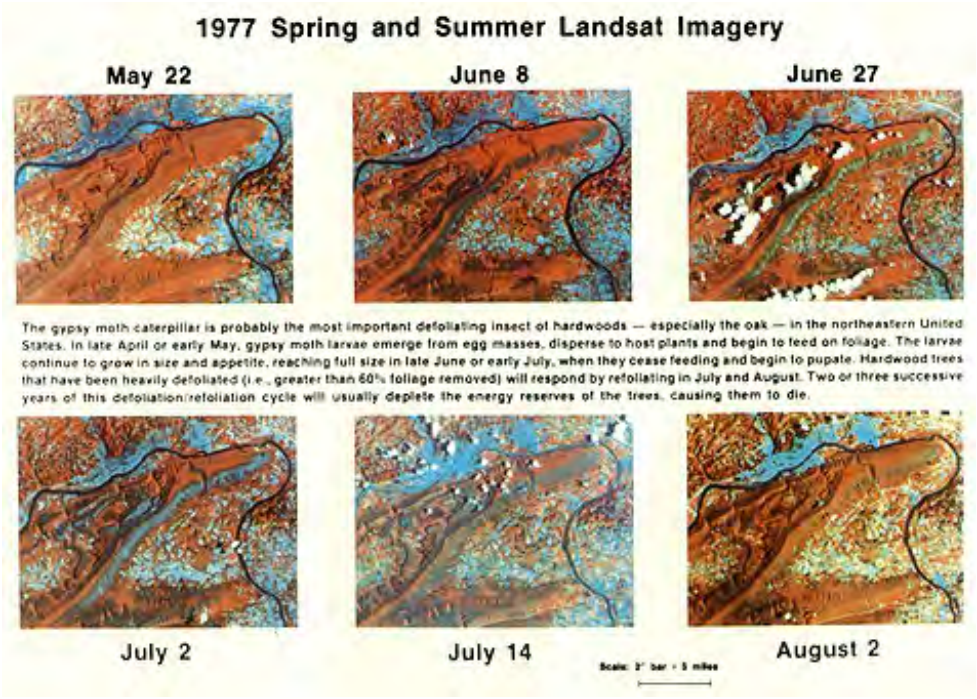
1-17: There are dark patterns, one elongate, the other as three spots, near the tips of arrows G and H. What do you think these are? Can you find them in corresponding local areas in the October 1972 MSS Band 7 image? What's going on? ANSWER

Now, inspect the scene below which is a July 8, 1973 MSS 7 subscene of the same area just looked at.



1-18: What's new and different? Identify the features at arrow tips of the letters indicated (the rightmost arrow is labelled "I"). Features such as at "K" are a new phenomenon, in which the ground is responsible for the reflectance that gives rise to a medium gray tone (that's a clue!) ANSWER

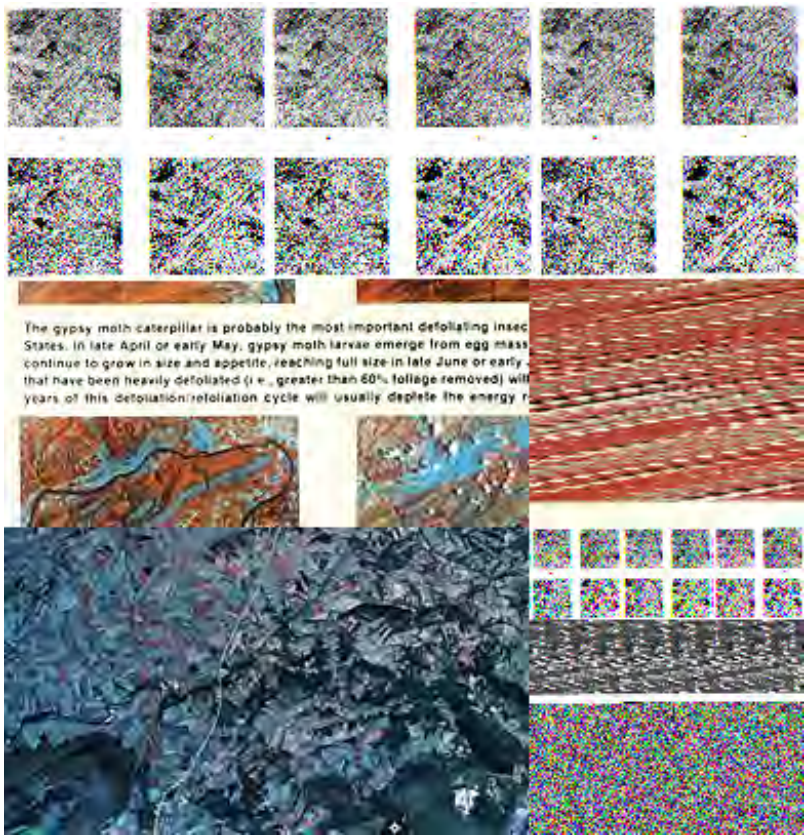
Pennsylvania was hard hit by the phenomenon at "K" again in 1977. The next illustration is a set of six subscene panels taken from an area north of Williamsport, PA (just outside the full scenes we have been looking at; here that town is the stretched out blue pattern near the top). For now, don't read the black print between the two panel rows.



1-19: Since you now know what "K" is from the previous question, comment on the progression of the feature that appears dark brownish-gray in these panels. ANSWER

We will now transition into a second phase of questioning in which you will be asked to analyze and comment on some computer-generated imagery produced in this study of south-central Pennsylvania.

First, is a panel of four subscenes of an area that includes Hanover, PA (at the upper left edge, west of the reservoir [thin black pattern]) about 20 miles south-southwest of York, PA. Passing diagonally across each subscene is a series of nearly parallel straight lines. These are also visible in most of the full scenes. They are low ridges carved from more resistant units in the Precambrian metamorphic complex that makes up this part of the Piedmont. Here is the scene:

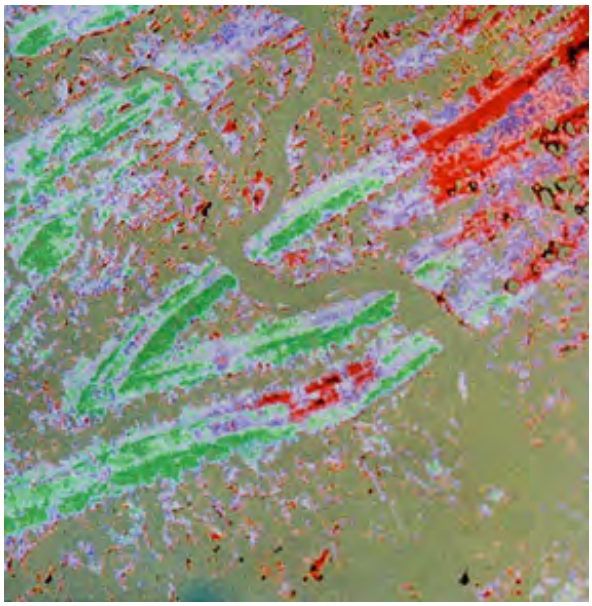


The upper left image is a MSS Band 7 rendition (stretched) of a June 1978 scene; the lower right is a linear stretch of a Band 7 of a February 1978 scene in which the ground was covered with snow. The upper right consists of a high bandpass filtering by using a Fast Fourier transform algorithm applied to the June subscene whereas the lower left is also a high bandpass filter version of Band 5, with Cubic Convolution resampling.

1-20: Which of the four subscenes seems best/worst for picking out linear features? Compare the Fast Fourier and Cubic Convolution processing for enhancing linear features and sharpening details. Make a guess as to the identity (nature) of the irregular feature (in dark gray to black) that lies just south of the reservoir lake.ANSWER

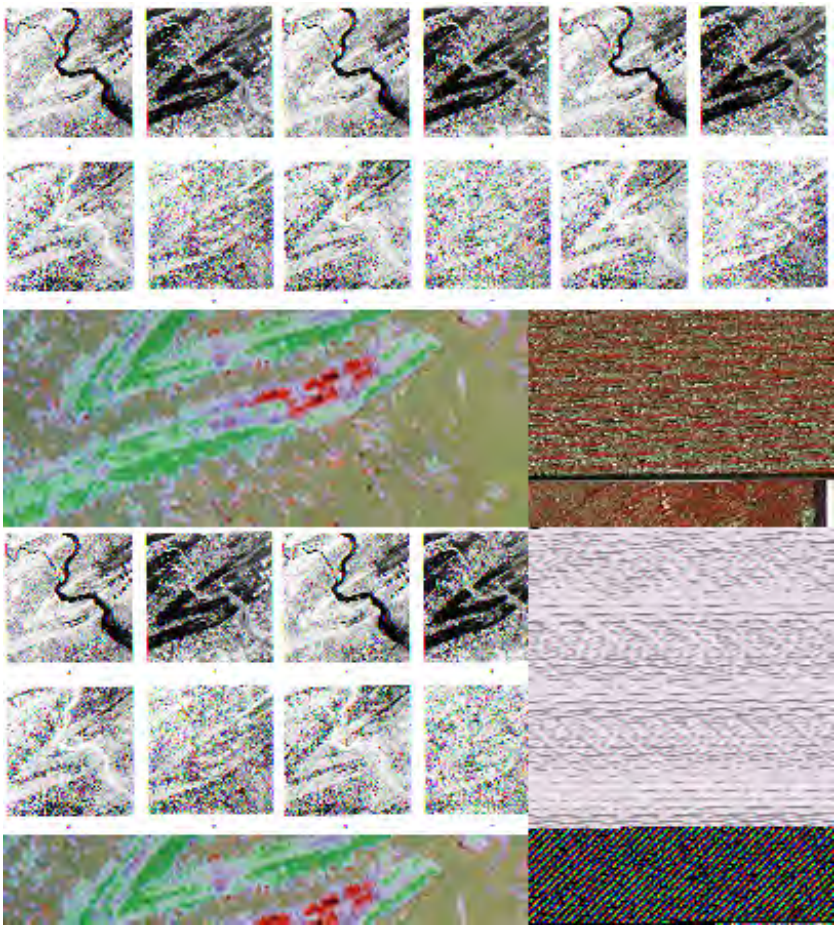
Now, back to Harrisburg and Gypsy Moth defoliation. Look at this June 27, 1977 subscene. The defoliation appears as very dark gray patches on the ridges.

There are several ways to focus on this defoliation while ignoring other features/classes in the image. One is to calculate the Vegetation Index for the image. At least six different formulae have been proposed for this VI. We will use the simplest, namely, a ratio of MSS7/MSS5. When this is done there will be a range of DNs that could be shown as gray levels. But, another way is to produce a color mask image (a variant of density slicing). All non-vegetated features are set to a single DN value (masking them out) and vegetated areas with signatures characteristic of trees and forest are assigned different colors. This is the result, with masked out areas shown in yellow-gray, healthy forest in green, defoliated forest in red, and stressed vegetation in purple.



1-21: What is the advantage of this VI image? Where may confusion occur? ANSWER

This same scene (and date) has been subjected to Principal Components Analysis. The first four component images are shown below (but, in an error in layout, the first component is B and the second, A, with C and D the third and fourth components respectively.

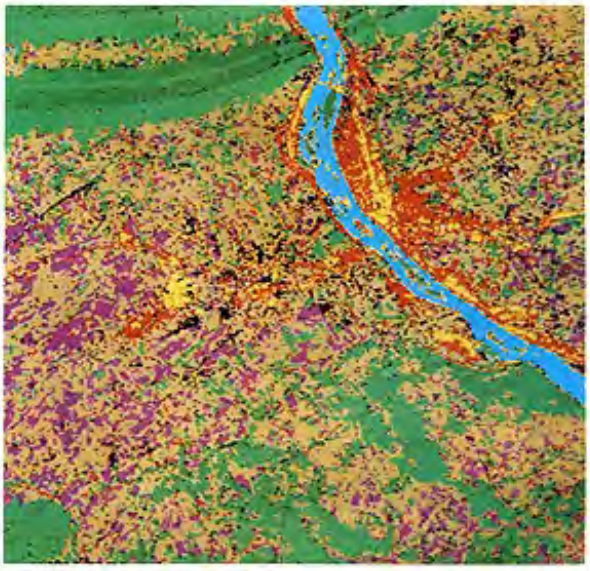


1-22: Describe the main gray patterns in each component image. ANSWER

A color composite can be made from any three of these images. Generally, the fourth component image is bland and blurry and is not used. Here below is a false color image made using PC 2 = red; PC 1 = green; PC 3 = blue.

1-23: What does this PC image show that is less obvious in the standard false color composite shown above?ANSWER

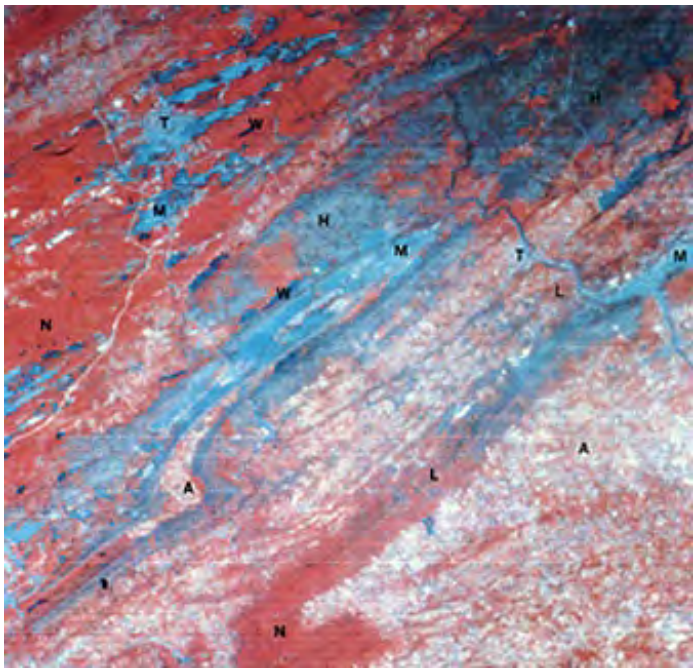
We are ready now to embark on Supervised Classifications of 4 areas within the original October, 1972 full MSS scene that we have examined at the beginning of the test. We'll start with this 8 class map of the Harrisburg subscene. This, and subsequent classifications were all done on the IDIMS computer system at NASA Goddard.



	AREA [%]
URBAN: INDUSTRIAL/COMMERCIAL	5.8
URBAN: RESIDENTIAL/MIXED	8.1
VEGETATED CROPLAND/PASTURELAND	36.5
BARREN/OPEN LAND	14.2
DECIDUOUS FORESTLAND (SUNNY)	16.1
DECIDUOUS FORESTLAND (SHADED SLOPES)	15.0
WATER (STREAMS AND LAKES)	2.8
UNCLASSIFIED (FINAL ITERATION)	1.9

1-24: Evaluate this classification. ANSWER

Onward to the anthracite coal belt. We present first a standard false color composite of the subscene that shows folded mountain ridges from Hazleton (T) southeast from Blue Mountain (lowest N) out into the agricultural lands (A) of the Great Valley itself. Most of the brighter blue is mine waste. Defoliation was moderate in this July 8, 1973 scene, being represented by the blue-gray tones.



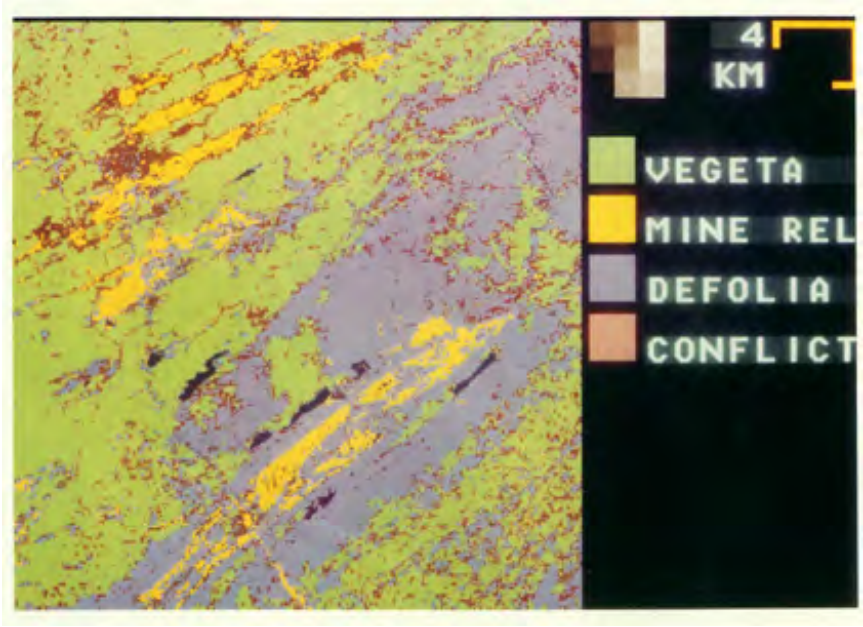
A 1:78,000 scale black and white aerial photograph covers part of this Landsat subscene.

1-25:Try to fit this aerial photo into the Landsat scene. Name the towns (use an Atlas) you find. Is the mine refuse well displayed in the photo. ANSWER

Let's look ahead to Section 16, in which products from the spacecraft Terra are introduced. One sensor, MISR, has yielded this near-natural color image of the Pennsylvania scene plus the New York-New Jersey areas to the east.

1-26:Is there anything special about the areas in the coal country we have been looking at? Why is this speciality ambiguous?ANSWER

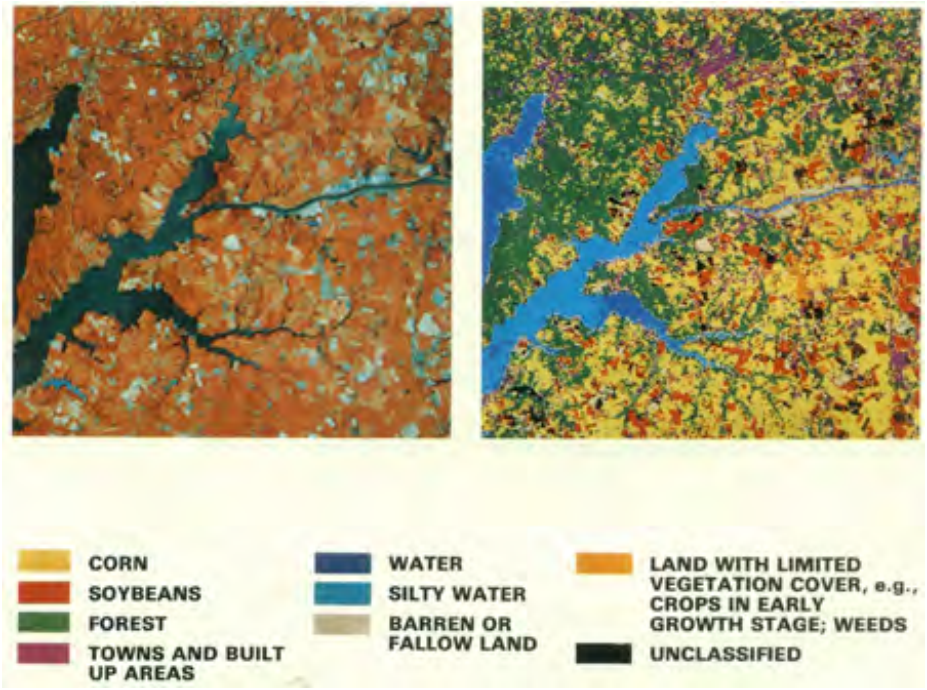
An Unsupervised Classification of the July 8th 1973 subscene near Hazleton was made on a General Electric Image-100 interactive processing system. Only four cluster classes were specified. After they were blocked out as patterns, they were then named, using the writer's general knowledge of the region. Here it is:



1-27: Evaluate this classification, compared with your expectations from the fcc image for this date. Use the lakes (shown in black) to relate the class map to the image. ANSWER

The next two classifications treat the area around the Chesapeake and Delaware Canal and Chesapeake Bay headwaters that occupy the lower right corner of the October, 1972 image.

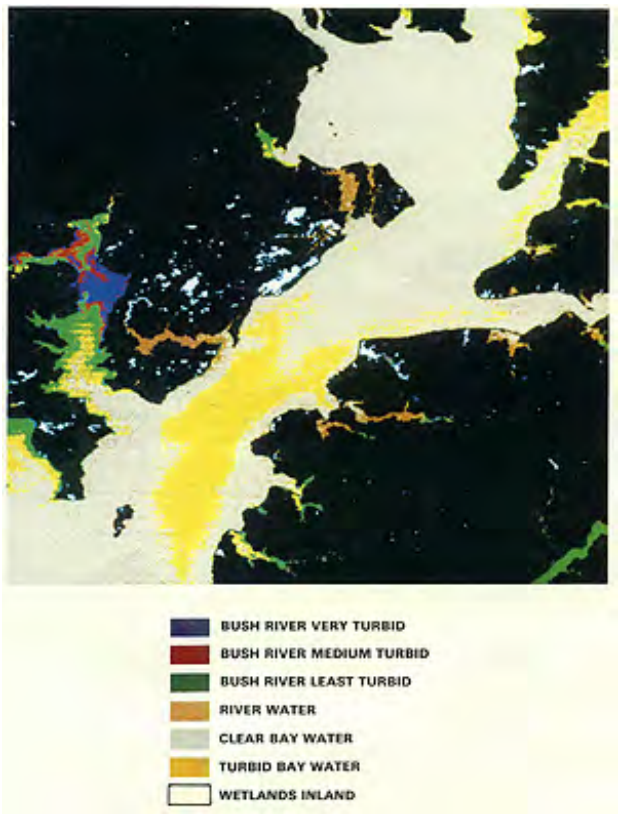
The first of these is a Supervised Classification (right) of the subscene extracted from the July 1977 full scene shown below on the left as a false color image. The classes were chosen by the writer following a field trip to the area and discussions with the local Soil Conservation Service representative for New Castle County, DE.



1-28: Which seems more abundant, corn or soybeans? If you consult a map of the area, you might conclude that most of the larger lavender patterns associate with towns and villages. But what about smaller clusters of lavender; what might they be? Give a reason why so many of the fields do not have straight boundaries, as is typical, nor are rectangular. ANSWER

The last classification shows categories of water purity in the Chesapeake Bay headwaters in

northern Maryland around Havre de Gras and Aberdeen. The date of image acquisition was May 28, 1978. This is the result:



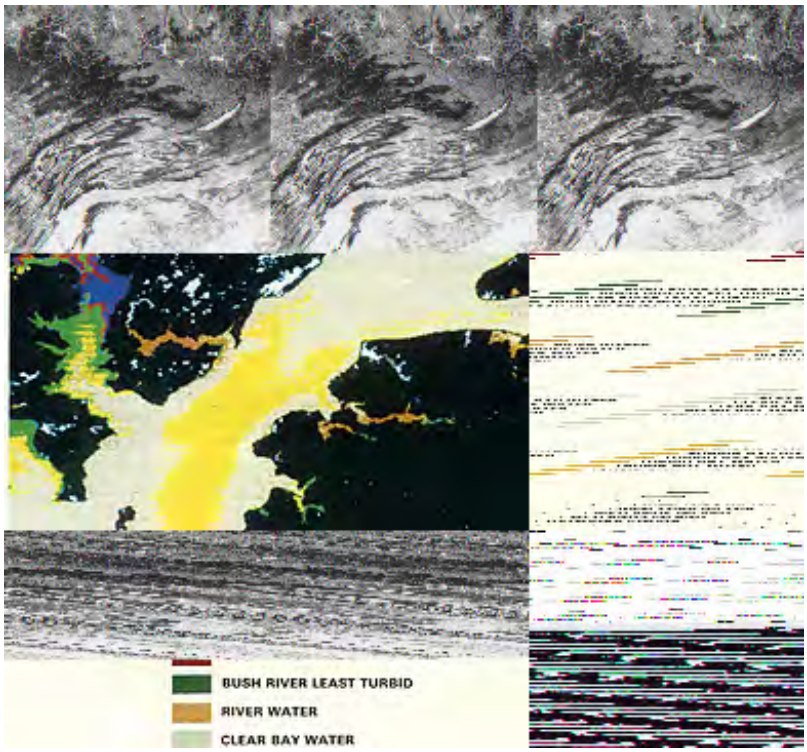
The mouth of the Susquehanna River is at center top. The Chesapeake Bay itself is the drowned segment of the lower Susquehanna which once emptied in Virginia near Norfolk but has been inundated by rising sealevel after the last glacial retreat. The land here is rendered in one color (except for white wetlands) by masking, which involves assigning all land pixels to a single value shown as black.

1-29: How do you think the turbidity in the Bush River was determined (assessed)? ANSWER

We are nearing the end of this "exam". Now we turn our attention to two other types of remote sensors - those that produce active microwave - radar - images and those that measure thermal properties . These sensor systems are treated in Sections 8 and 9 respectively. The following is just a preview.

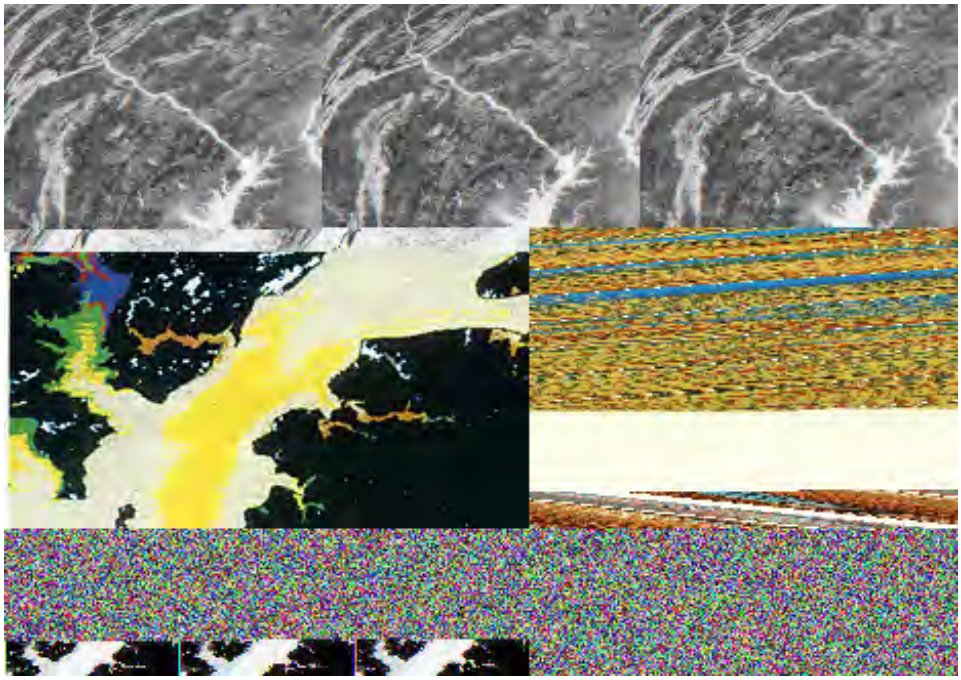
For thermal imagery covering the Pennsylvania region and some neighboring states, we will turn to the satellite known as HCMM (Heat Capacity Mapping Mission; see Section 9). This operated during 1978 and produced scenes 700 km wide in three modes: Day Visible; Day (thermal) IR; and Night IR.

Look first at a Day IR image (September 26, 1978) in which the bottom half includes much of the region around Harrisburg that we have been studying. Dark tones correspond to "cool"; light to "warm".



1-30:First locate yourself within this image relative to the landmarks you have come to know. Why are the ridges cool (dark) and the valleys warm (light-toned)? There is a long, canoe-shaped white area in the right center; why does it appear to be so hot? While it is not in the region you have been studying, speculate on the very dark patterns in the upper center. ANSWER

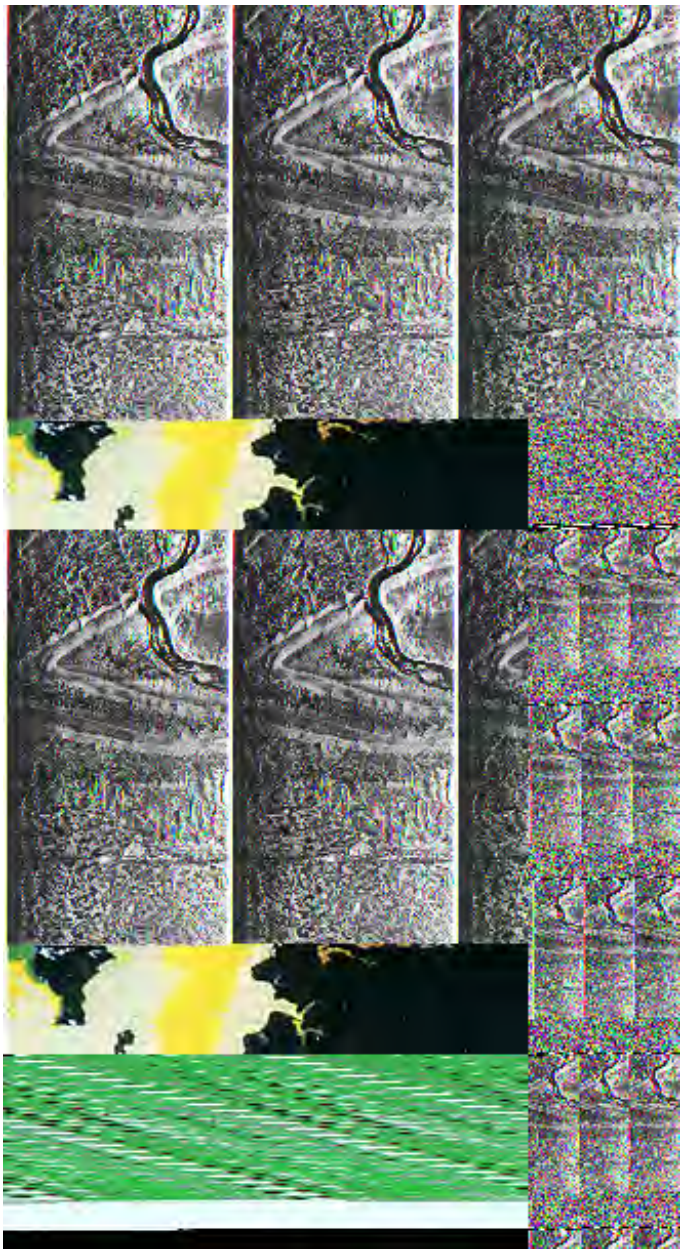
Then, look at this Night IR image, obtained on June 11, 1978.

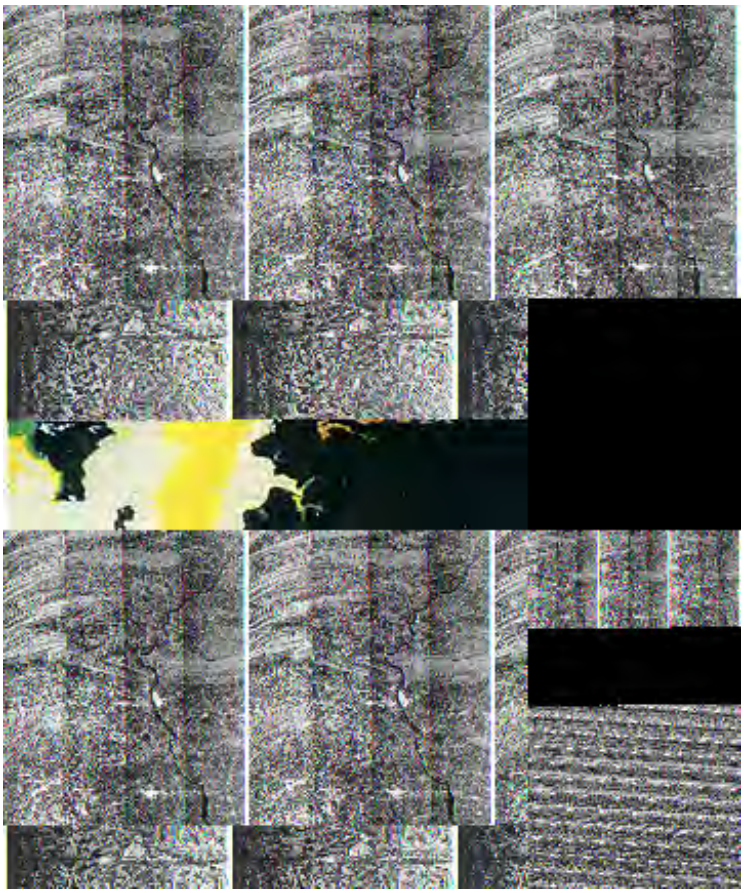


1-31: Why do the ridges and rivers and other water bodies appear to be warmer (lighter toned) than the valleys? Can you find the Harrisburg area? ANSWER

Finally, we take just a momentary peak at two radar images that include the Harrisburg area. Because interpretation of radar imagery requires considerable understanding of how the sensor operates and the images are formed (which we will learn about in detail in Section 8), we will only show you the two examples. We suggest you locate these images in the framework of the Landsat scenes and pick out any special features and landmarks of which you are already familiar. The first image is made from an aircraft-mounted SLAR system (K-Band), as it flew over

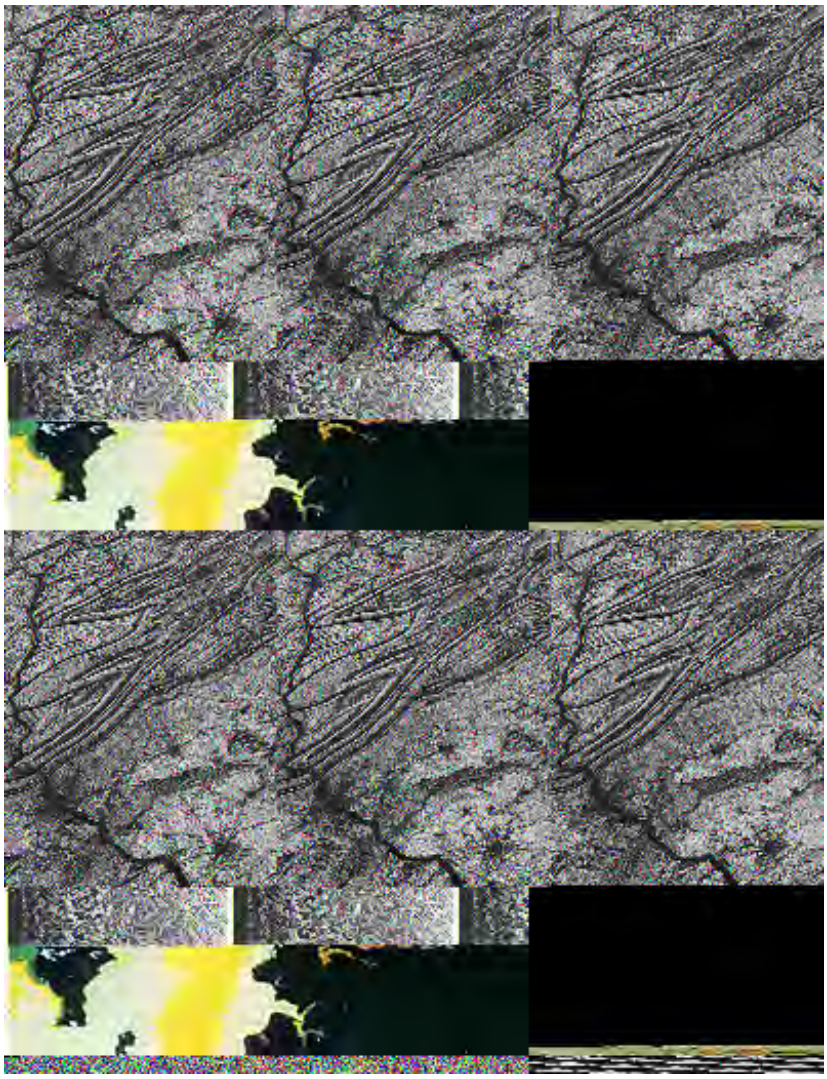
the terrain on July 22, 1966. The second was taken from space by the Seasat SAR (L-Band) on September 28, 1978. In that image, the location of Harrisburg should be obvious (its buildings are strong reflectors).

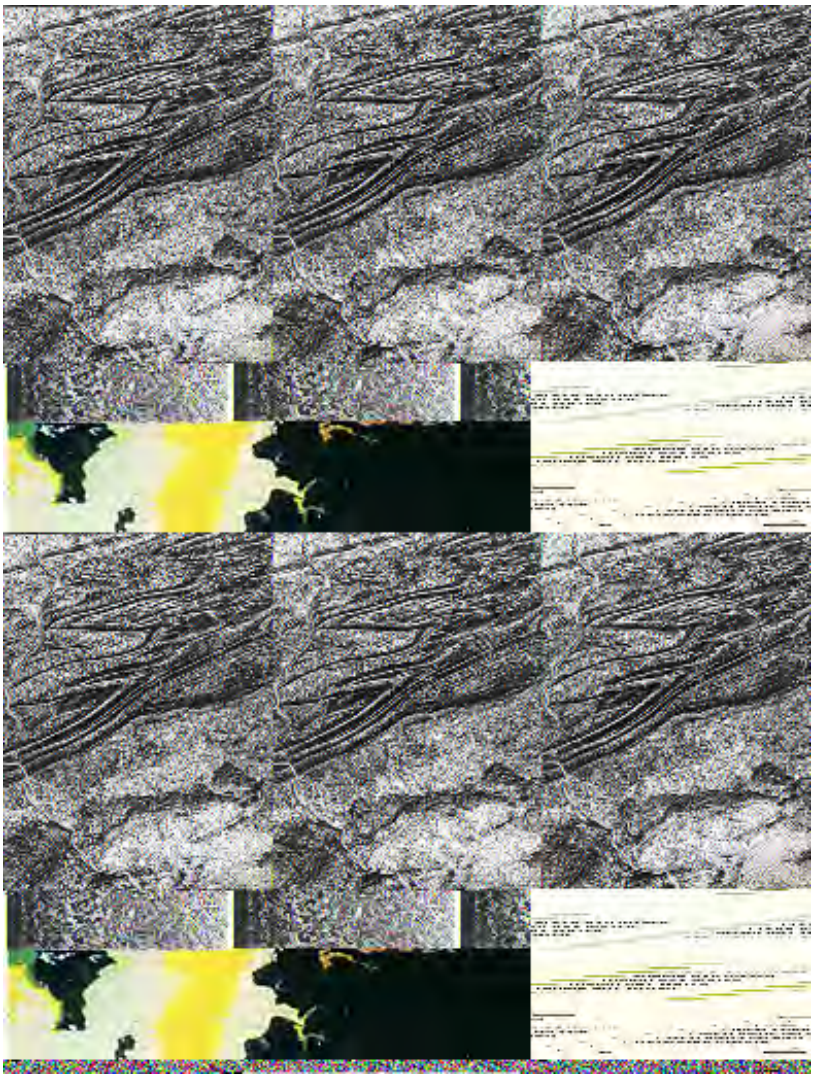




1-32: Compare the aircraft SLAR image with the Seasat image. Make note of any obvious differences between the two and suggest reason(s) why. ANSWER

Now, we will close by asking you to look at these two images that contain Harrisburg.





1-33: We challenge you to determine what evidence you note that indicates the time of the year when each of these Harrisburg subscenes was imaged. ANSWER

Well, if you tried the "exam" and have clicked back to the bottom, we expect you found it rather challenging, but hopefully not tedious. Grade yourself, but by merely working through it, you deserve an "A". In the rest of the Tutorial we will not again face such an extensive sequence of questions until the Final Exam (Section 21) - but in nearly all Sections in between there will be just enough queries to help you become better acquainted with the important concepts and features associated with remote sensing interpretations.

Click here to get back to the last page of Section 1.

Primary Author: Nicholas M. Short, Sr. email: nmshort@TowerMicro.net

THE MYSTERY FEATURE IS AN 18-HOLE GOLF COURSE!!

By projecting TM Bands 1, 2, and 3 through blue, green, and red filters respectively, a quasi-true color image of a scene can be generated. One is shown on this page, but if you are familiar with color aerial photos, you will likely protest that these are not the real colors you would see in such photos. That is correct, but at the least, vegetation appears green yet the ocean water is not the expected blue. Very careful processing of the Landsat data can improve the quality of a true color rendition (the Idrisi program only approximates this). Still, note the similarity between the Landsat expression of the marshland, and the general color seen in a ground photo. The identity of the mystery feature is finally exposed on this page, along with a supporting ground photo.

True Color View

The next scene displays (sometimes the verb "map(s)" is used instead) the Morro Bay region in quasi-natural color, achieved using the IDRISI processing system by assigning TM Band 1 to the blue, Band 2 to the green, and Band 3 to the red color source.

In practice, we use various color mapping algorithms to facilitate visual interpretation of an image (some types of analysis by computer work by manipulating the original DN [digital number] values of the pixels without necessarily producing direct imagery). The original DN values contain all of the information in the scene and though their range of values may make it necessary to remap them (such as by the contrast stretching described later in this Section) to create a good display, that doesn't always add much new information. In fact, although visual interpretation is easier with the remapped image, remapping potentially can lose and distort information; thus, for some kinds of analytical work, we may prefer to use the original DN values or DN values adjusted to calibrated radiances.

With this mapping, we see a pleasing and satisfying image because it depicts the world in the general color ranges with which we are naturally familiar. We can imagine how this scene would appear if we were flying over it at a high altitude.

We might comment that the ocean appears too dark in the IDRISI TM composite. This darkness is due in part to its general darkness in all of the bands, the consequence of water absorbing light from most of the visible range. (To a lesser degree, the algorithm that IDRISI uses to combine three 8-bit bands into a single 8-bit composite image also may compromise the image color display). When actually next to or above the ocean, we see blue tints partly because of some angular reflection of blue light and because of reflection of blue sky color when clouds are sparse; the ocean can in reality appear in some other colors to the eye (as for example a darkish jade green in the Caribbean). The notable exception is, of course, related to the breakers (here, with little internal structure) that are highly efficient at scattering light; there is much whitish tones in waves when standing near them. The sediment patterns as we noted vary through bands 1 to 3. Areas with higher sediment content are clearly separable (rather than gradational) in this color rendition, because areas of greater load density have some green band input (Band 3 [red] shows only the densest sediment areas). Note especially the dark brownish-green tones by the delta where the coastal marsh develops, interpreted as a heavy load dumped at the mouth of the small river (**p**) flowing over it.

Probably the most unrealistic color in the rendering of the IDRISI version of this "true color" scene is the bright hill slopes. In this rendition, the base color is tan with some purplish overtones, but there are small areas and patches marked by a darker brown tone which, if inspected at ground level would have some green tints. The color composite shows these surfaces in a somewhat misleading way. From an airplane, we might describe these surfaces as a light golden-brown (from the dry grass, one reason California is often called the "Golden State"), similar to the terrain shown below which matches the above true color mapping more closely.

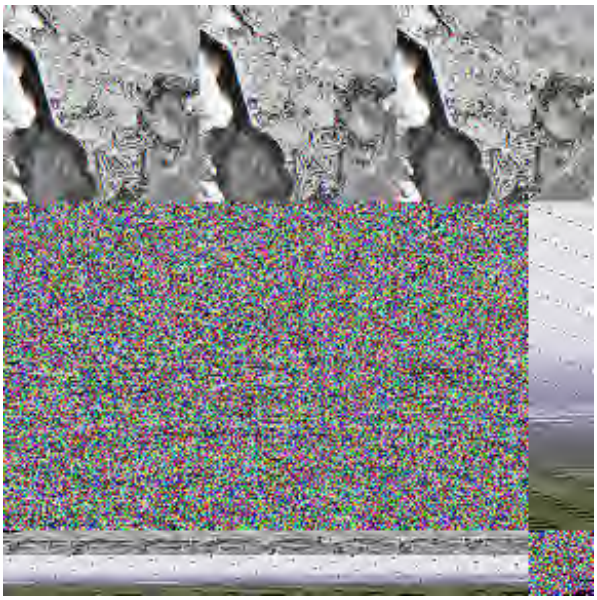
The reason for the red-brown-purple blotches in the natural color composite is that light from surfaces at certain orientations contributes considerable red from Band 3, some blue (hence purple) from Band 1, and little green from Band 2.

The areas we presumed to be active grasslands (**v**) appear in "true color" as light tones that are slightly reddish brown. Areas correlated with fields and meadows are a medium brown. Some areas that were associated with medium to dark reds in the false color composite are medium green in the natural color version. This is particularly valid for the National Forest growth around (**u**), the stream-lining vegetation (**i**), and the wooded patches (**l**) scattered throughout the scene. However, many false color red areas do not show as green in this natural color version. The diminution of green color is explainable as follows: most of this vegetation is either California oaks, which have dark green leaves, or perhaps some evergreens which tend to be less reflective than deciduous plants, thus the dark brown color. These leaves, or needles, have generally lower reflectance in Bands 1-3 but still are somewhat proficient reflectors (from cell walls) in Band 4. But, since the natural color composite only uses Bands 1, 2, and 3, there is little color contribution from any of these in the image above.

The coastal marsh (**o**) appears as a dark brownish-red in this true color rendering (as it seems to do in the aerial oblique photo). At least part of this effect may link to types of vegetation which

were envisaged in the photo shown on page 1-2 that indicates reddish tones in many of the plants. Another view of the marsh, with its outlet on the bay, is presented here; note the rather brownish tones over much of its expanse:

Now, lets return our attention momentarily to the mystery feature. It appears again in this high altitude aerial photo of Morro Bay (town and water) distributed by the U.S. Geological Survey:



In the natural color Landsat subscene at the top of this page, the mystery feature displays as a small patch of bright green. We saw earlier in the black and white enlargement that the feature is arranged roughly in elongated strips. These strips do not stand out in the natural color image, partly because at the size on the screen, the resolution is insufficient to bring out the strip pattern. From the top of the page, you now know (or confirmed) that this feature is an 18-hole golf course. The aerial photo shows these public links at its bottom center. Hole Number 18 is pictured in this ground view:

Let's repeat the Landsat enlargement in false color and then show a map of the greens as printed on the official scorecard picked up at the local Pro Shop. It is rotated 90° clockwise to align with the Landsat subscene.

We have also extracted the portion of the digitized aerial oblique photo (page 1-1) that contains this tract, known as the Morro Bay Golf Course on State Park Road. This photo shows the lighter green fairways amidst darker green trees in close-up, mainly in the center. One patch of trees (many being the Eucalyptus tree, imported from Australia, that has spread so widely that it is almost a commonplace pest), hides luxurious Condominiums built since the aerial photo was

taken. The two bright patches within the photo are sand traps. They are not visible in the TM subscene. This is surprising, in that even if they are actually smaller than the 30 m resolution of a TM pixel containing one of them, the great brightness of each should have swamped the reflectance response of other classes in that (mixed) pixel, making it appear overly light. This same phenomenon was observed early in the Landsat program where small [less than 8 m {26 ft} wide] light-toned dirt roads passing through dark sage brush or grasslands could be resolved because the pixels containing these pathways were substantially brightened.)

There are several parts of the natural color image that contain distinctive orange-brown tones (as at **x**). This coloration is apparently due to soils that have a rich brown color. We also see this brown tone in the valley north and east of Los Osos and in part of that town. This valley appears to have its own small stream, which is not easily resolvable, or perhaps it once had a larger stream, which is now gone. Such a valley is probably underlain by alluvial deposits that typically are browner (enhanced by hydrous iron oxides similar to rust).

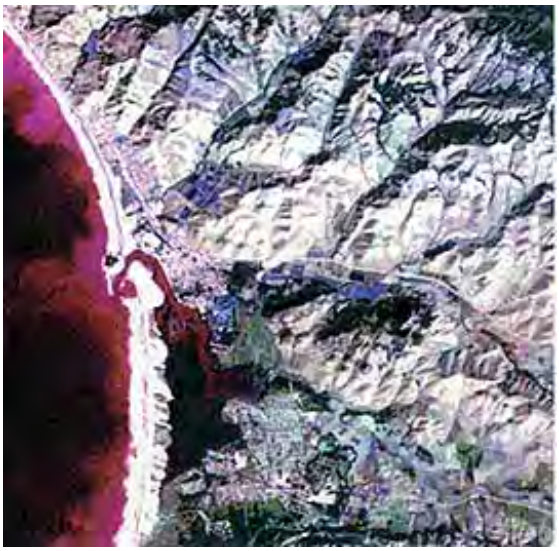
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Other combinations of bands and color filters (or computer assignments) produce not only colorful new renditions but in some instances bring out or call attention to individual scene features that, although usually present in more subtle expressions in the more conventional combinations, now are easier to spot and interpret. This page illustrates several examples.

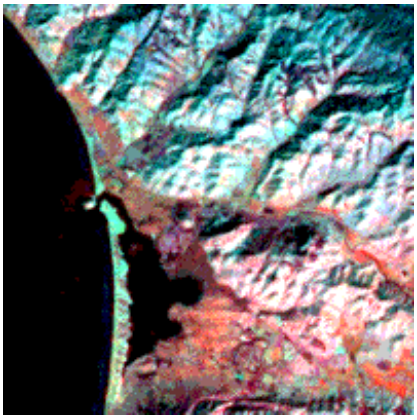
Other Color Combinations

Reluctantly, we leave these attractive false and true color composites to examine several others made by different three band/filter combinations. The first we will construct renders Band 1 = red, Band 3 = green, Band 2 = blue:



Features of special interest in the resulting image are, first, the intensification of the expression of sediment in the ocean, standing out sharply in red, and two, areas of more active vegetative growth (which would have lighter gray tones in the green band) appearing as distinct blues.

Let us now experiment further by using Bands 5, 6, and 7 in some combinations. The first casts 5 as blue, 6 as red, and 7 as green.



- TM Band 6 = red
- TM Band 7 = green
- TM Band 5 = blue

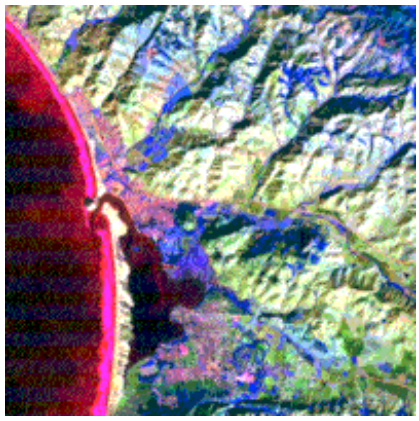
This combines the thermal band, sensitive to emitted radiation, (with its lower resolution 120 m.

pixels), with two infrared bands, 5 and 7 (with their 30 m. pixels). In this color composite generated with IDRISI, the thermal band image (check TM 6 again) dominates. In the IDRISI mapping, reds and blues greatly outweigh any greens contributed by light tones in Band 7. The reds represent the warmer surfaces, and the blues the cooler surfaces, as recorded in Band 6. Note that the blues extend over a broader fraction of the hill slopes than we might expect from the shadow effects seen in other bands. Some areas of blue that lie on the sunlit foreslopes are covered by the uplands vegetation at **g** and the grasslands at **v,** which offset the direct heating effect through cooling by evapotranspiration. Other areas supporting vegetation (as at **y**), and also along streams have dark red tones, not because they are hotter but because their tonal expressions in 5 and 7 are darker (almost no blue or green inputs).

Both urban communities show several shades of red, with the street patterns showing through because of their sharper definition in bands 5 and 7. This is, in part, due to the "urban island" thermal effect, which is the tendency of populated areas to be warmer because of heat-absorbing materials (darker streets, tar-covered roofs, etc.), reduced surface areas with vegetation, and heat emitted from furnaces, air-conditioners, and other human activities. Fully perceptible is the thermal plume (whitish patch) emanating from near the power plant (**t**), but the ocean sediments introduce no noticeable effect. Note that the breakers now completely disappear (are black) since they provide no reflectance input in these bands.

Some areas we see as light greenish-blue in this composite include the beach bar and several of the extraction pits (**u**). These surfaces reflect highly in most bands, hence, their light tones in Bands 5 and 7 combine the blues and greens assigned to these bands whereas their darkness in Band 6 (reflective materials are cooler) leaves red out.

Let's now experiment briefly with combining one of the reflective longer-wave IR bands (7, assigned to green) with the vegetation band, TM 4 (blue), and the ocean water/sediment band, TM 1 (expressed in red), as produced by the IDRISI Composite function. Remember it converts three 8-bit bands into a single 8-bit image that mimics the results of using the three bands to make a true 24-bit color image. The result is a colorful rendition that, in some respects, shows certain features best.



- TM Band 7 = green
- TM Band 4 = blue
- TM Band 1 = red

Again, blues and reds are the prevailing colors, whereas greens are subordinate. The blues correlate closely with vegetation, which stands out in sharp contrast to most areas that either lack this cover or now support dormant grasses, etc. (note the active grasslands in lighter blue). The reds tie into three principal surface classes: the waves (note the bluish-purple tones within them, representing the Band 4 contribution), which now sharply separate from the beach sands; the sediments; and the towns, with the red streets speckled by blues from local vegetation. The greens show some specific features that are relatively brighter in Band 7 and darker in Bands 1 and 4, including areas in the hills around Los Osos and some uncultivated areas (**m**) in valleys. Note scattered greens along the otherwise bright slopes.

In this last image, you no doubt noticed that there are some obvious similarities between this Band 7-4-1 combination (when the bands are stated this way, they refer implicitly to the color

sequence BGR or blue-green-red; sometimes they are also written 741) and the Band 2-3-1 composite at the top of the page. But several distinctions occur: In the 2-3-1 version the breakers are white whereas they are reddish in the 7-4-1 version. This is because the breakers have about the same gray levels in Bands 1, 2, and 3 but are darker gray in 5 and almost black in 7. The blues in this last version are more intense and widespread since Band 4, which shows vegetation as very light tones, is rendered in blue and thus effectively emphasizes the locations of active growth. Also, note several areas of red in the 7-4-1 composite that are found within parts of the built-up or urban class. The color patterns are not the same in the 2-3-1 rendition, suggesting that equivalent areas are suppressed tonally in Band 7 and/or 4. If curious, retrace to page 1-3, which has all 7 bands together, to try to deduce whether it is one or the other or both.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Almost without exception, the image will be significantly improved if one or more of the functions called Enhancement are applied. Most common of these is contrast stretching. This systematically expands the range of DN values to the full limits determined by byte size in the digital data. For Landsat this is determined by the eight bit mode (28) or 0 to 255 DNs. Examples of types of stretches and the resulting images are shown. Density slicing is also examined.

Contrast Stretching and Density Slicing

We move now to two of the most common image processing routines for improving scene quality. These routines fall into the descriptive category of Image Enhancement or Transformation. We used the first image enhancer, *contrast stretching*, on most of the Morro Bay TM images we have looked at so far in the previous 12 pages to enhance their pictorial quality . Different stretching options are described next, followed by a brief look at *density slicing*. We will then evaluate the other routine, *filtering*, shortly.

As we said on page 1-4, contrast stretching, which involves altering the distribution and range of DN values, is usually the first and commonly a vital step applied to image enhancement. Both casual viewers and experts normally conclude from direct observation that modifying the range of light and dark tones (gray levels) in a photo or a computer display is often the singlemost informative and revealing operation performed on the scene. When carried out in a photo darkroom during negative developing and positive printing, the process involves shifting the gamma (slope) or film transfer function of the plot of density versus exposure (H-D curve). This is done by changing one or more variables in the photographic process, such as the type of recording film, paper contrast, developer conditions, etc. Frequently the result is a sharper, more pleasing picture, but certain information may be lost through trade-offs, because some of the gray levels are "overdriven" into states that are too light or too dark.

Contrast stretching by computer processing of digital data (DNs) is a common operation, although we need some user skill in selecting specific techniques and parameters (range limits). The reassignment of DN values is based on the particular stretch algorithm chosen (see below). Values are accessed through a Look-Up Table (LUT).

The fundamental concepts that underlie how and why contrast stretching is carried out are summarized in this diagram:

From Lillesand & Kiefer, *Remote Sensing and Image Interpretation*, 4th Ed., 1999

In the top plot (a), the DN values range from 60 to 158 (out of the limit available of 0 to 255). But below 108 there are few pixels, so the effective range is 108-158. When displayed without any expansion (stretch), as shown in plot b, the range of gray levels is mostly confined to 40 DN values, and the resulting image is of low contrast - rather flat.

In plot c, a linear stretch involves moving the 60 value to 0 and the 158 DN to 255; all intermediate values are moved (stretched) proportionately. This is the standard linear stretch. But no accounting of the pixel frequency distribution, shown in the histogram, is made in this stretch, so that much of the gray level variation is applied to the scarce to absent pixels with low and high DNs, with the resulting image often not having the best contrast rendition. In d, pixel frequency is considered in assigning stretch values. The 108-158 DN range is given a broad stretch to 38 to 255 while the values from DN 107 to 60 are spread differently - this is the histogram-equalization stretch. In the bottom example, e (labeled 'special stretch'), some specific range, such as the infrequent values between 60 and 92, is independently stretched to bring out contrasty gray levels in those image areas that were not specially enhanced in the other stretch types.

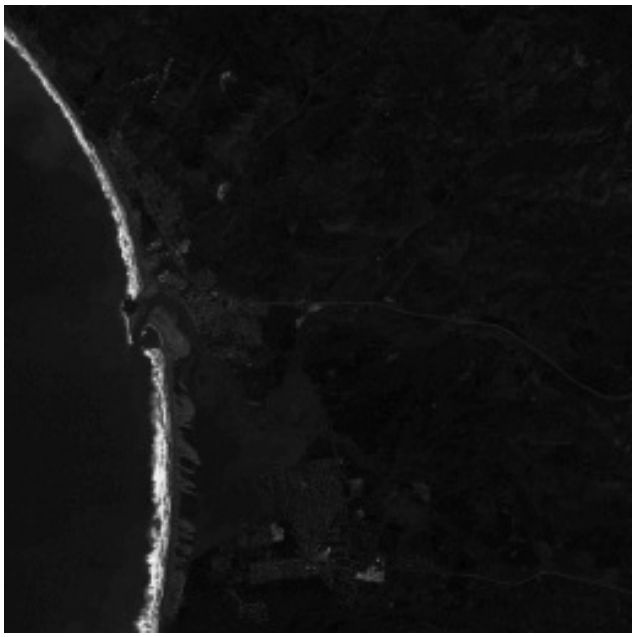
Commonly, the distribution of DNs (gray levels) can be unimodal and may be Gaussian (distributed normally), although skewing is usual. Multimodal distributions (most frequently, bimodal but also polymodal) result if a scene contains two or more dominant classes with distinctly different (often narrow) ranges of reflectance. Upper and lower limits of brightness values typically lie within only a part (30 to 60%) of the total available range. The (few) values falling outside 1 or 2 standard deviations may usually be discarded (histogram trimming) without serious loss of prime data. This trimming allows the new, narrower limits to undergo expansion to the full scale (0-255 for Landsat data).

Linear expansion of DN's into the full scale (0-255) is a common option. Other stretching functions are available for special purposes. These are mostly nonlinear functions that affect the precise distribution of densities (on film) or gray levels (in a monitor image) in different ways, so that some experimentation may be required to optimize results. Commonly used special stretches include:1) Piecewise Linear, 2) Linear with Saturation 3) Logarithmic, 4) Exponential 5) Ramp Cumulative Distribution Function, 6) Probability Distribution Function, and 7) Sinusoidal Linear with Saturation.

Now, to apply these ideas to the Morro Bay TM bands. For Landsat data, the DN range for each

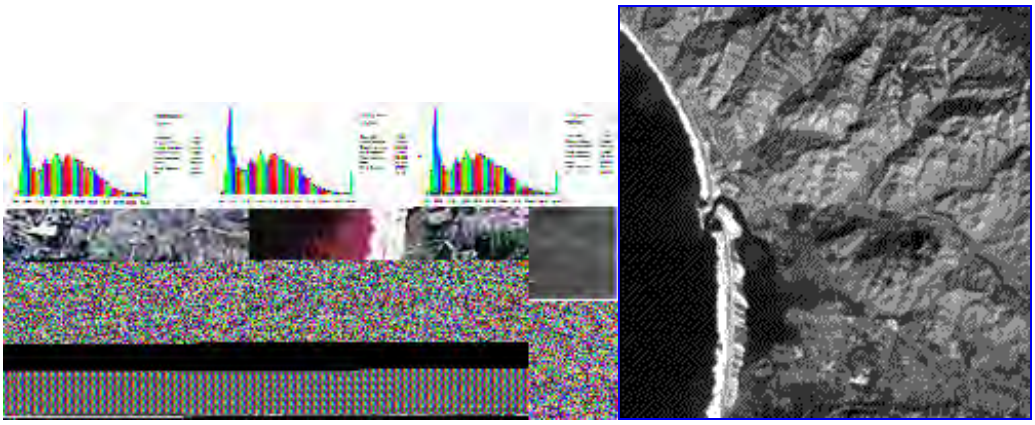
band, in the entire scene or a large enough subscene, is calculated and displayed as a histogram (recall the histogram for TM Band 3 of Morro Bay on page 1-1; there is a frequency peak at about DN = 10 and most of the pixels fall between DN 0 and DN 60).

To illustrate contrast stretching (also called autoscaling) we utilize several IDRISI STRETCH functions to enhance Band 3. Recall that the histogram of raw TM values shows a narrow distribution that peaks at low DN values. We might predict from this that we have a dark, flat image. This is indeed the case:

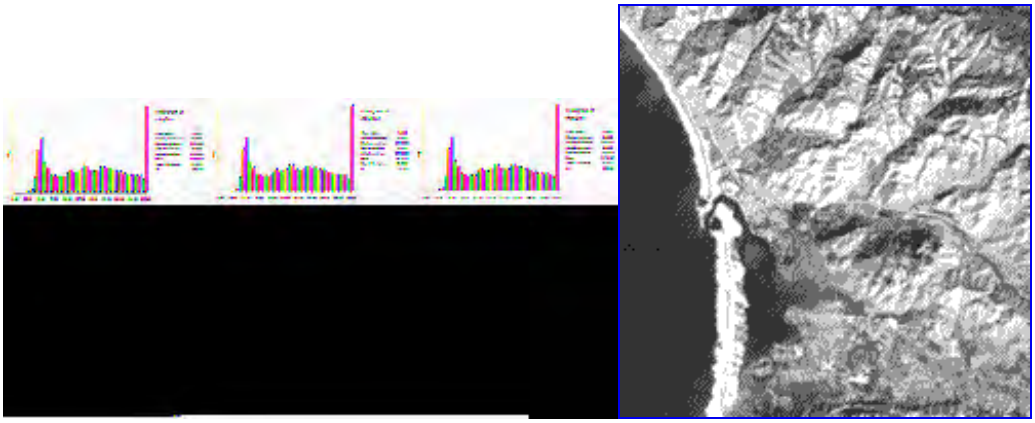


When IDRISI's standard linear stretch program was applied, this somewhat improved image resulted.

Most of the DN values, however, lie between 9 and 65 (there are values up to 255 in the original scene but few of them). We can perform a selective linear stretch so that 9 goes to 5 and 65 goes to 255, with all values in between getting stretched proportionately. We show this expanded histogram, and the resulting new image below it.

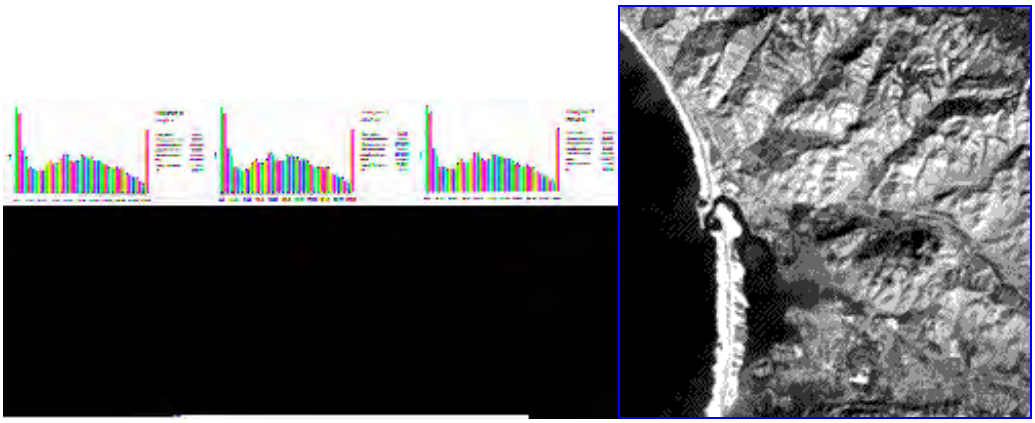


Now, most of the scene features are discriminable. But the image still is rather dark. Let's try instead to choose new limits in which we take DNs between 5 and 45 and expand these to 0 to 255. The procedure is like (special) case e in the diagram near the top of this page. This stretch results in the following histogram and image:



The histogram for this image is polymodal, with a lower limit near 25 and a large number of DN pixels at or near 255. This accounts for the greater scene brightness (light tones).

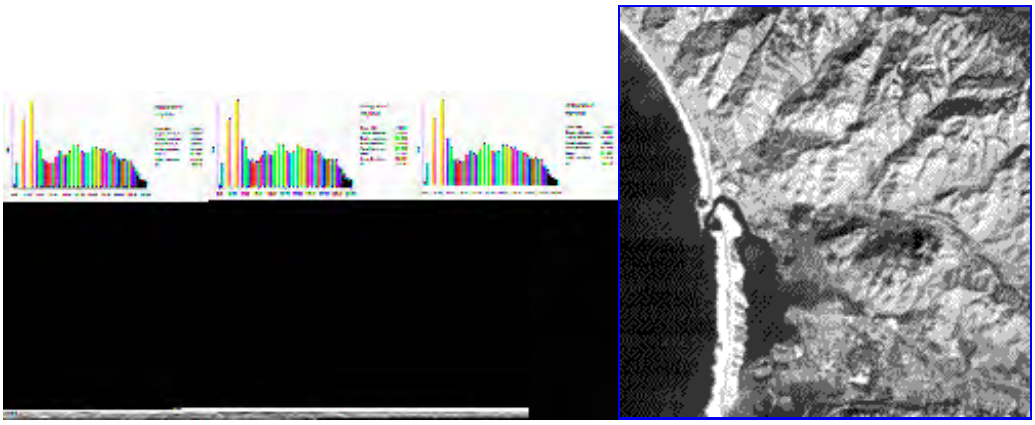
Next we try a Linear-with-Saturation stretch. Here we assign the 5% of pixels at each end (tail) of the histogram to single values. The consequent histogram and image are:



The image appears as a normal and pleasing one, not much different from the others. But, comparing this one with either of the linear-stretched versions shows that real and informative differences did ensue.

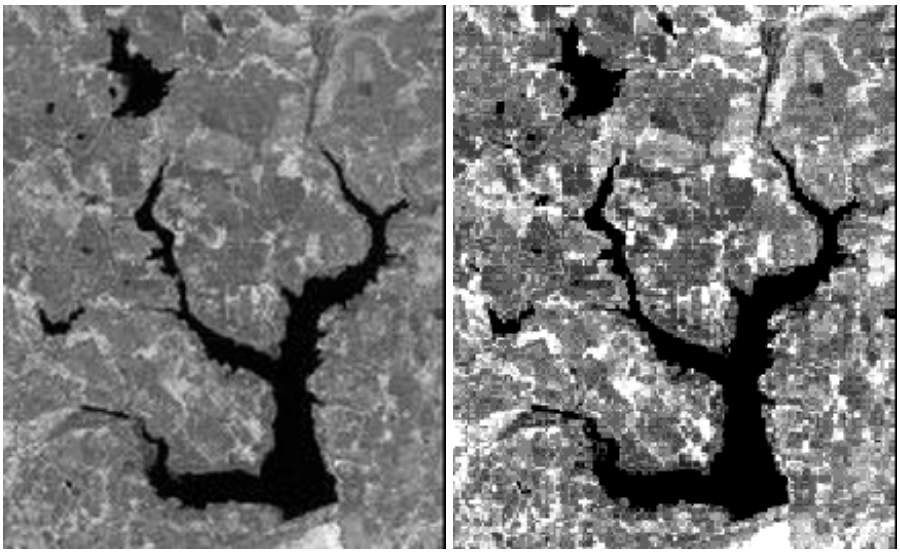
Histogram Equalization is a stretch that favorably expands some parts of the DN range at the expense of others by dividing the histogram into classes containing equal numbers of pixels. For instance, if most of the radiance variation occurs in the lower range of brightness, those DN

values may be selectively extended in greater proportion to higher (brighter) values. Here, we carry out a Histogram Equalization stretch, with these results:

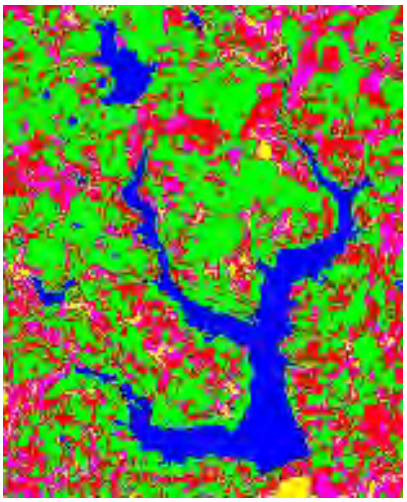


The image is similar to the Saturation version. Note that pixel frequencies are spread apart at low DN levels and (closely-spaced) at high intervals.

In the Introduction it was stated that the human eye can see and distinguish between thousands of differing colors whereas only about 20 or so successive levels of gray are discernible. Applying colors to stretched images can be a big aid in picking up differences that may relate to actual features or classes in the scene. Consider these three images (of an area not in the Morro Bay scene):



The left image is an unstretched Band 3 image; the right is a simple stretch. The stretch version is obviously easier for picking out differences in gray levels that seem to some degree representative of different ground classes. Now look at the scene when levels have been assigned different colors (red, green, blue, yellow):



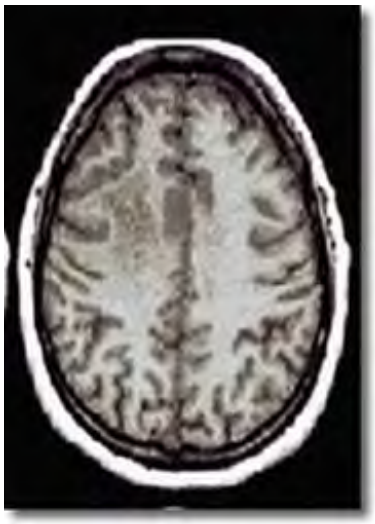
Let us reiterate. Probably no other image processing procedure or function can yield as much new information or aid the eye in visual interpretation as effectively as stretching. It is the first step, and most useful function, to apply to raw or preprocessed data.

1-12: Comments have been made above about the relative quality and information content of each of the stretches displayed. In your opinion, which seems best and is most pleasing to the eye. ANSWER

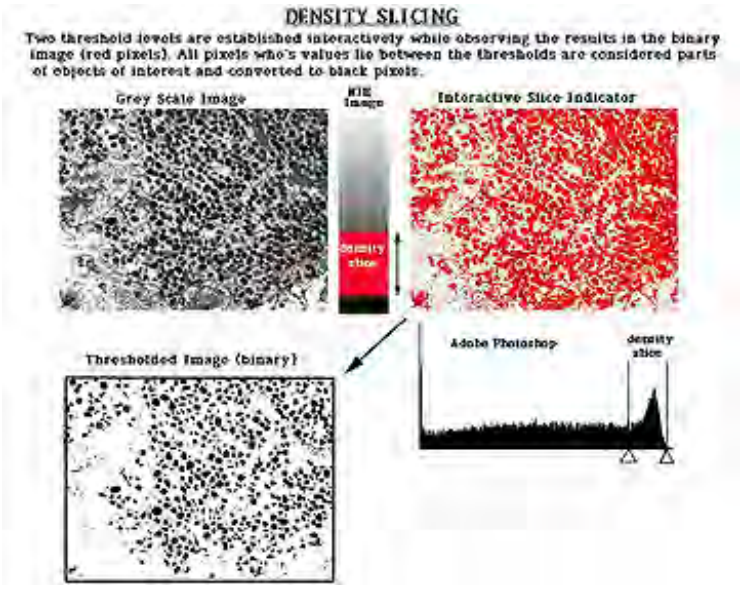
The last (colorized) image is related to another enhancement procedure. This is based on combining ("lumping together") DNs of different values within a specified range or interval into a single value. This density slice (also called "level slice") method works best on single band images. It is especially useful when a given surface feature has a unique and generally narrow set of DN values. The new single value is assigned to some gray level (density) for display in a photo or on the computer monitor (or in an alphanumeric character printout). However, for an image with a range of DNs, these will be subdivided into a series of single values or gray level slices. To illustrate: suppose the range of the predominant DNs is 30 to 150; every 5 successive DNs are combined into a single value; this produces 150 - 30 = 120 divided by 5 = 24 single value increments. All other DNs can be assigned another level, usually black. This yields a simple map of the distribution of the combined DNs. The new sets of slices are commonly assigned different colors in a photo or display.

Version 2.0 of IDRISI FOR WINDOWS unfortunately does not have a simple density slicing routine, so the Morro Bay scene cannot be used here to illustrate this type of enhancement. We will instead take examples from different (Internet) sources.

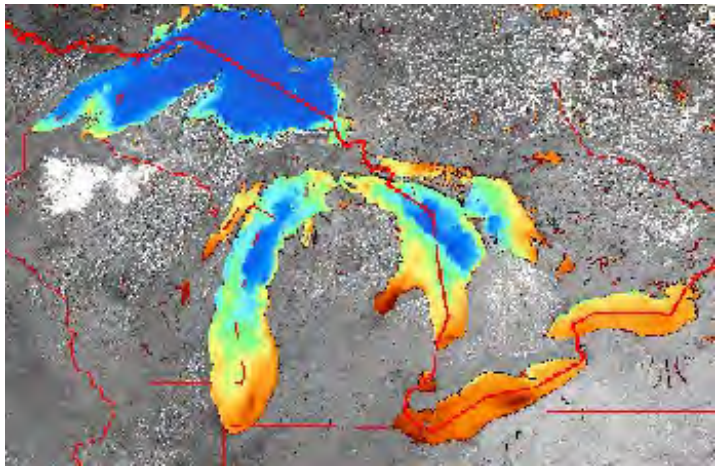
We saw several black and white level slices in the subsection of the Introduction covering Medical Remote Sensing. As a reminder, here is a density slice image made from a Magnetic Resonance Imaging scan of a human head, showing primarily variations in brain tissue:



Color-coded density slicing is commonplace in medical studies. One example would be selective color staining of organic tissue. Or, an image made by a scanner can be processed to bring out certain features recognizable as having distinctive tones, to which one or more colors are assigned:

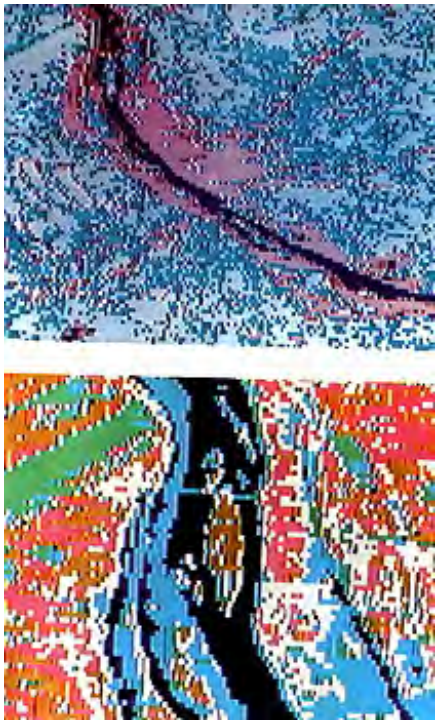


Examples of density- sliced and color-assigned images from satellites are scattered throughout the Tutorial. Here are gray level slices representing temperature variations that are colored to indicate thermal differences in the U.S. Great Lakes (AVHRR data set):



If classes or features within an image have gray tones that are narrow and distinct (not associated with other classes), the density sliced rendition can serve as a "quick and dirty"

classification. These sliced subscenes are from a MSS Band 6 (vegetation bright) image acquired in July 1975 over Harrisburg, PA (the subject of the Exam found at the end of this Section):



The upper map has four levels or slices. The lavender tends to demarcate a gray level (DN 43 to 48) that associates with urban areas. The lower map (pixels enlarged) covers part of the northern reaches of Harrisburg (the bridge is Interstate 81). Six gray levels (each representing a DN range) have been colorized as follows: Black = (DN) 0-19; Blue = 20-34; Red = 35-44; White = 45-54; Brown = 55-69; Green = 70+. The black pattern is almost entirely tied to water; the blue denotes heavily built up areas; the green marks vegetation; the other colors indicate varying degrees of suburbanization and probably some open areas.

However, one must be wary of accepting a density-sliced image as a bonafide classification map. In any band, the DNs and their corresponding image gray levels can include contributions from several classes to a limited range. But, if one class with a particular level dominates, the density slice rendition is a reasonable "first cut" at specifying its distribution.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

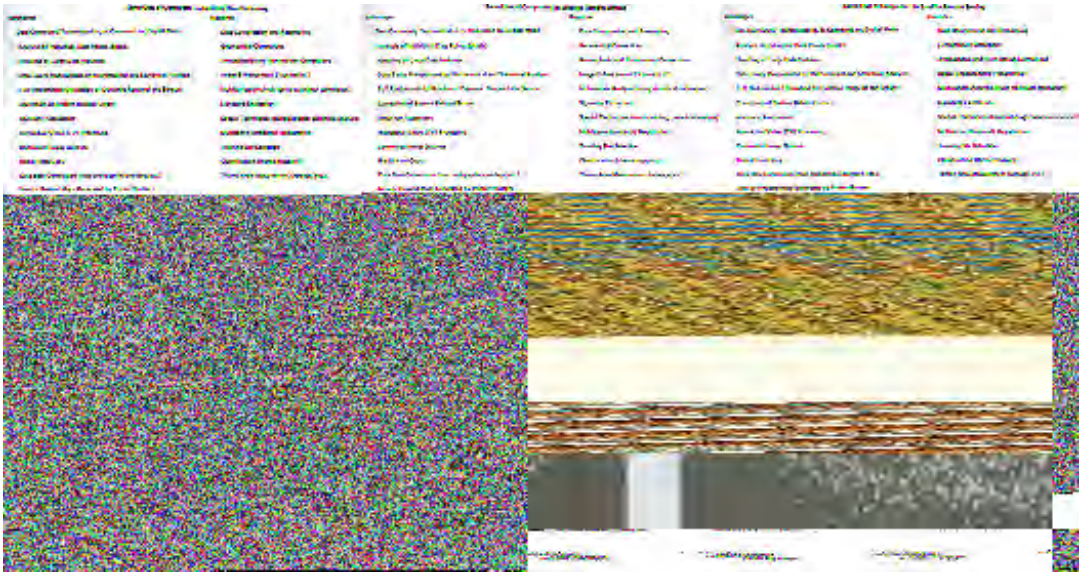
At this point we turn to a more intense examination of computer-based procedures for working with the digital data acquired by remote sensors. The three general types of image processing are introduced. The programs that convert the raw DN values from individual bands into photographs or computer displays often yield visually poor products, especially if there is a narrow range of DN's and/or an even narrower range dominates as a single histogram peak. This is improved, often greatly, by applying standard or specialized processing procedures. On this page, the first step in optimizing the data, especially for pictorial purposes, in called Preprocessing.

Some Basic Image Processing Procedures

The roots of remote sensing reach back into ground and aerial photography. But modern remote sensing really took off as two major technologies evolved more or less simultaneously: 1) the development of sophisticated electro-optical sensors (see page I-5a) that operate from air and space platforms and 2) the digitizing of data that were then in the right formats for processing and analysis by versatile computer-based programs. Today, analysts of remote sensing data spend much of their time at computer stations, as shown below, but nevertheless still also use actual imagery (in photo form) that has been computer-processed.

Now that you have seen the individual Landsat Thematic Mapper (TM) bands and color composites that have introduced you to our study image, we are ready to investigate the power of computer-based processing procedures in highlighting and extracting information about scene content. This involves the recognition, appearance, and identification of materials, objects, features, and classes (these general terms all refer to the specific spatial and spectral entities in a scene). A very helpful summary of the main ideas involved in digital image processing is provided in Volume 3 in the Remote Sensing Core Curriculum first cited in the Overview of this Tutorial. Jan Husdal has produced a good Internet site worth visiting for more on image processing (but it may "stick" on screen after you try to leave via the "Back" button). Another useful overview tutorial has been put on line by the Chesapeake Bay Institute at Towson University in Maryland. Chapter 7 on Image Processing, from Floyd Sabins, Jr.'s excellent text on remote sensing, has been reproduced online at this site. There are also some helpful supplementary ideas on basic processing in the first five pages of Appendix B of this Tutorial.

The advantages of computer processing and the common routines or methods are summarized in this Table:



Processing procedures fall into three broad categories: *Image Restoration (Preprocessing); Image Enhancement; and Classification and Information Extraction*. The first of these is considered on this page. We will discuss on the next two pages contrast stretching, densing slicing, and spatial filtering; producing stereo pairs, perspective views, and mosaics are considered in other Sections of the Tutorial. Under Information Extraction, ratioing and principal components analysis have elements of Enhancement but lead to images that can be interpreted directly for recognition and identification of classes and features. Also included in the third category but treated outside this Section is Change Detection. Pattern recognition is often associated with this category.

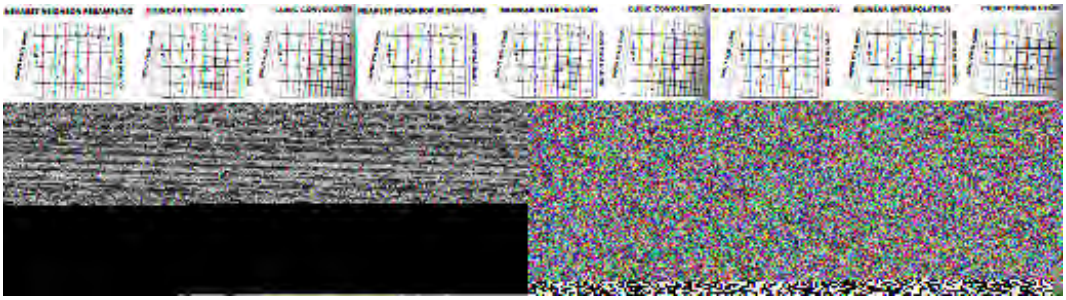
A word about a parameter we have mentioned before (e.g, first on page I-5a and at the beginning of this Section), namely the meaning of the term *Digital Number* or DN. We have said that the radiances, such as reflectances and emittances, which vary through a continuous range of values, are digitized onboard Landsat and other spacecraft after initially being measured by the sensor(s) in use. Ground instrument data can also be digitized at the time of collection. Or, imagery obtained by conventional photography is capable of digitization. A DN is simply one of a set of numbers based on powers of 2, such as 2^6 or 64. This range of numbers is coupled with the range of radiances. Instrument-wise these radiances can be, for example, recorded as varying voltages. The lower and upper limits of the sensor's response capability form the end members of the DN range selected. The voltages are divided into equal whole number units based on the digitizing range selected. Thus, a Landsat TM band can have its voltage values - the maximum and minimum that can be measured - subdivided into 2^8 or 256 equal units. These are arbitrarily set at 0 for the lowest value, so the range is then 0 to 255.

Preprocessing

Preprocessing is an important and diverse set of image preparation programs that act to offset problems with the band data and recalculate DN values that minimize these problems. Among the programs that optimize these values are atmospheric correction (affecting the DNs of surface materials [making them higher or lower] because of radiance from the atmosphere itself, involving attenuation and scattering); sun illumination geometry; surface-induced geometric distortions; spacecraft velocity and attitude variations (roll, pitch, and yaw); effects of Earth rotation, elevation, curvature (including skew effects), abnormalities of instrument performance (irregularities of detector response and scan mode such as variations in mirror oscillations); loss of specific scan lines (requires destriping), and others. Once performed on the raw data, these adjustments may require appropriate radiometric and geometric corrections.

Resampling is one approach commonly used to produce better estimates of the DN values for individual pixels that have been reshifted to better fit their real spatial distributions. Thus, after the various geometric corrections and translations have been applied, the net effect is that the resulting redistribution of pixels involves their spatial displacements to new, more accurate relative positions. However, the radiometric values of the displaced pixels no longer represent the real world values that would be obtained if this new pixel array could be resensed by the scanner. The particular mixture of surface objects or materials in the original pixel has changed somewhat (depending on pixel size, number of classes and their proportions falling within the pixel, and extent of continuation of these features in neighboring pixels [a pond may fall within one or just a few pixels; a forest can spread over many contiguous pixels]).

In effect, for any shifted pixel from a geometric location X to a new, more correct location Y, the calculated radiometric corrections lead to creation of a new DN number for that pixel. To further clarify, at the time of sampling during the sensor overpass the geometrically incorrect pixel covers ground X, with its set of DN values, but after being shifted occupies a position that, if not adjusted, has X's DN values. However after being properly relocated the pixel should have somewhat different DN values that represent ground Y. One way to find this new Y DN value set is to examine the values in pixels surrounding the X pixel and average them out as an estimate of what inputs from them to Y would have been (if correctly sampled in the first place). This estimate of the new brightness value (as a DN) that is closer to the Y condition is made by some mathematical resampling technique. Three sampling algorithms are commonly used:



In the Nearest Neighbor technique, the transformed pixel A (equals Y) takes the value of the closest pixel in the pre-shifted array. In the Bilinear Interpolation approach, the average of the DNs for the 4 pixels surrounding the transformed output pixel is used. The Cubic Convolution technique averages the 16 closest input pixels; this usually leads to the sharpest image.

Because preprocessing is an expansive topic that requires development of a broad background, we will omit further discussion here. Instead, we refer you to any of the textbooks listed in the Overview that treat this subject in more detail.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Just as contrast stretching strives to broaden the image expression of differences in spectral reflectance by manipulating DN values, so spatial filtering is concerned with expanding contrasts locally in the spatial domain. Thus, if in the real world there are boundaries between features on either side of which reflectances (or emissions) are quite different (notable as sharp or abrupt changes in DN value), these boundaries can be emphasized by any one of several computer algorithms (or analog optical filters). The resulting images often are quite distinctive in appearance. In particular, linear features such as geologic faults can be made to stand out. The type of filter used, high- or low-pass, depends on the spatial frequency distribution of DN values and on what the user wishes to accentuate.

Spatial Filtering

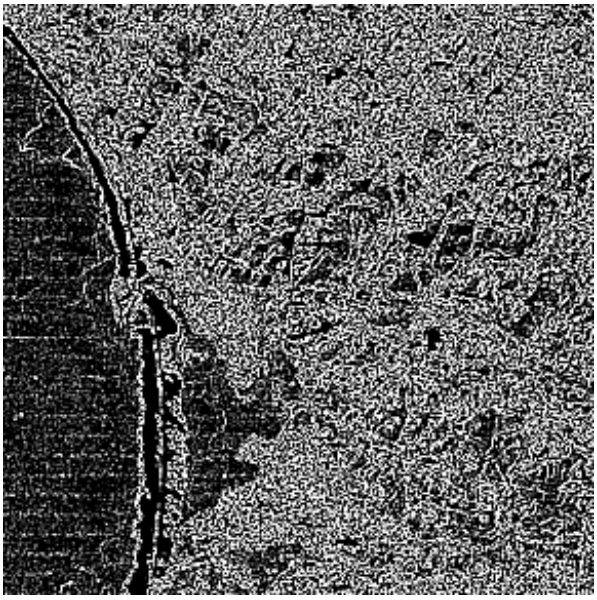
Another processing procedure falling into the enhancement category that often divulges valuable information of a different nature is *spatial filtering*. Although less commonly performed, this technique explores the distribution of pixels of varying brightness over an image and is especially sensitive to detecting and sharpening boundary discontinuities between adjacent pixel sets with notable differences in DN values. Patterns of sharp changes in scene illumination, which are typically abrupt rather than gradual, produce a relation that we express quantitatively as "spatial frequencies". The spatial frequency is defined as the number of cycles of change in image DN values per unit distance (e.g., 10 cycles/mm) applicable to repeating tonal discontinuities along a particular direction in the image. An image with only one spatial frequency consists of equally spaced stripes (in a TV monitor, raster lines). For instance, a blank TV screen with the set turned on, can have horizontal stripes. This situation corresponds to zero frequency in the horizontal direction and a measurable spatial frequency in the vertical, evidenced by line boundary repeats.

In general, images of practical interest consist of several dominant spatial frequencies. Fine detail in an image involves a larger number of changes per unit distance than the gross image features. The mathematical technique for separating an image into its various spatial frequency components is called Fourier Analysis. After an image is separated into its components (done as a "Fourier Transform"), it is possible to emphasize certain groups (or "bands") of frequencies relative to others and recombine the spatial frequencies into an enhanced image. Algorithms for this purpose are called "filters" because they suppress (de-emphasize) certain frequencies and pass (emphasize) others. Filters that pass high frequencies and, hence, emphasize fine detail and edges, are called highpass filters. Lowpass filters, which suppress high frequencies, are useful in smoothing an image, and may reduce or eliminate "salt and pepper" noise.

Convolution filtering is a common mathematical method of implementing spatial filters. In this, each pixel value is replaced by the average over a square area centered on that pixel. Square sizes typically are 3 x 3, 5 x 5, or 9 x 9 pixels but other values are acceptable. This square array is called a moving window, or kernel, since in the computer manipulation of an image the array can be envisioned as moving systematically from one pixel to the next. As applied in lowpass filtering, this tends to reduce deviations from local averages and thus smoothes the image. The difference between the input image and the lowpass image is the highpass-filtered output. Generally, spatially filtered images must be contrast stretched to use the full range of image display. Nevertheless, filtered images tend to appear flat unless strong frequency repeat patterns are brought out.

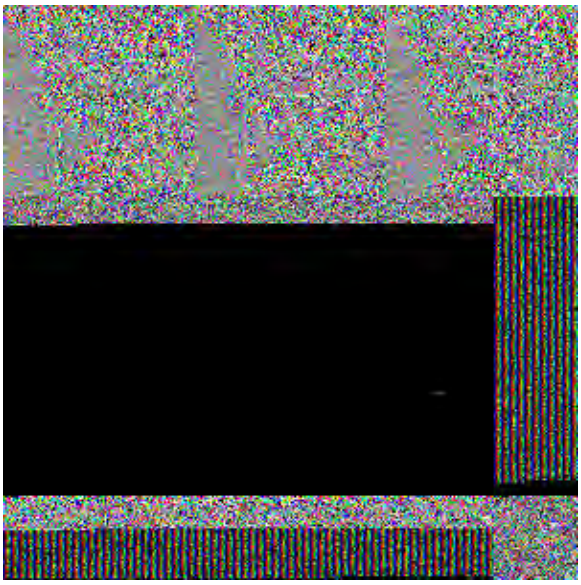
Next, we will apply three types of filters to TM Band 2 from Morro Bay. The first that we display is a lowpass (mean) filter product, which tends to generalize the image:

An edge enhancement filter highlights abrupt discontinuities, such as rock joints and faults, field boundaries, and street patterns:



In this example, the scene has been radically modified. The Sobel Edge Enhancement algorithm finds an overabundance of discontinuities but we have chosen this program (using Idrisi) to emphasize the sharp boundaries that can result from applying this mode of enhancement.

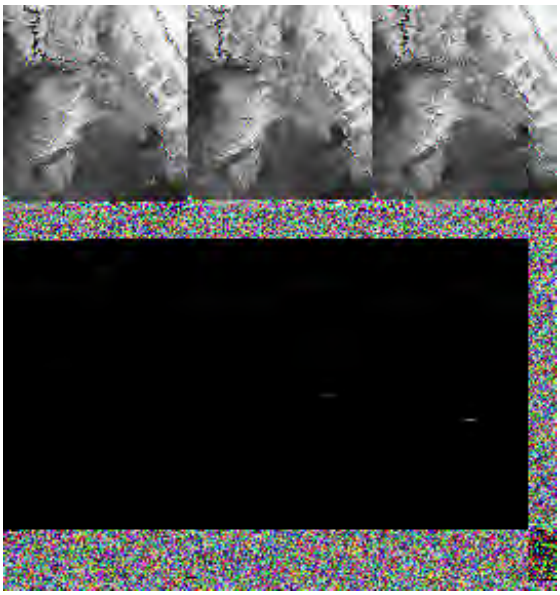
The highpass filter image for Morro Bay also brings out boundaries, but more subdued than above:



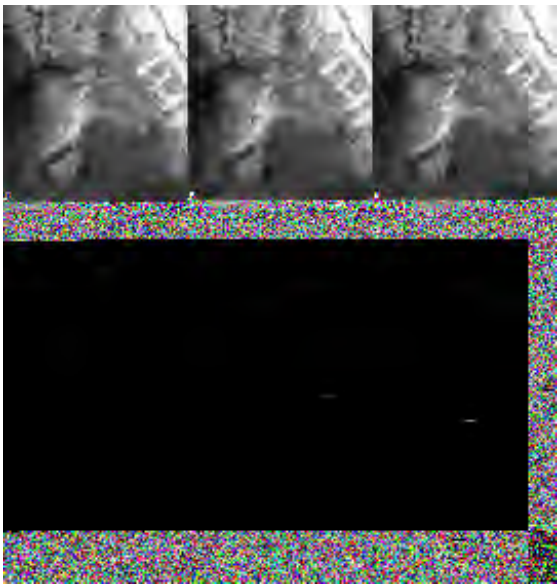
Here, streets and highways, and some streams and ridges, are greatly emphasized. The trademark of a highpass filter image is that linear features commonly appear as bright lines with a dark border. Details in the water are mostly lost. Much of the image is flat.

1-13: Comment further (evaluate) the three filter images shown above in terms of what information you extract visually. Include detrimental aspects. ANSWER

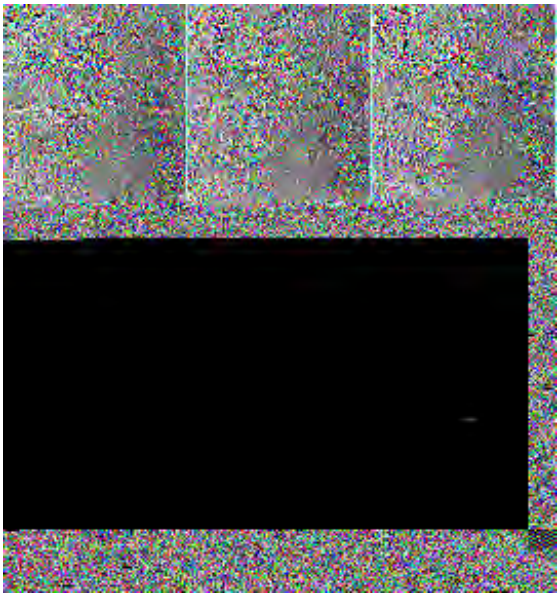
You will note in the answer to this question that the IDRISI filter products leave something to be desired. They can't really convince one that spatial filtering brings about helpfully different imagery. (Many of the linear patterns are spurious, i.e., they are artifacts of processing rather than real features.) To assuage that criticism, we are adding five images extracted from the textbook by Avery and Berlin (Macmillan Publ.; see Overview for reference). The Landsat TM 5 subscene shows a plateau landscape north of Flagstaff, AZ, with both sedimentary rocks and volcanic flows. The first image is just a normal contrast-stretched version.



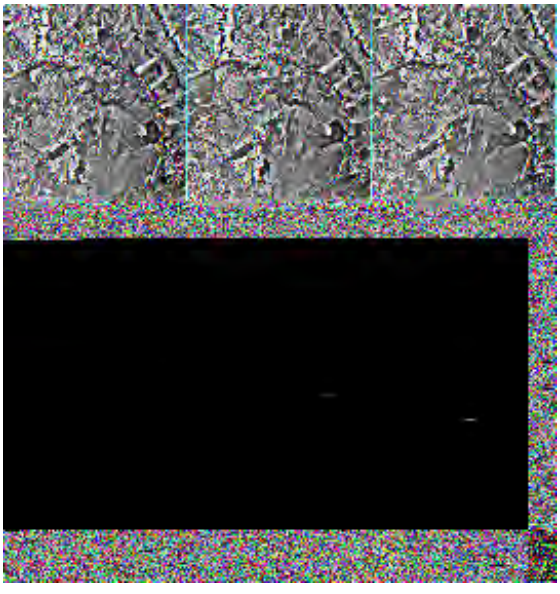
The next view shows this scene as a low bandpass image.



This is followed by a high pass filter image in which the convolution matrix is 11 x 11 pixels. This is made by subtracting the low pass filter image from the original (unstretched) data set. In this instance, the image is a form of edge enhancement.



As the number of pixels in the convolution window is increased, the high frequency components are more sharply defined. That is evident in this image which uses a 51 x 51 pixel matrix.



The last image in this set illustrates directional first differencing, analogous to determining the first derivative (a Calculus term) which establishes a gradient of spatial change as the filter, with its special algorithm, moves (computer calculational-wise) across the pixel array making up the image. Directional filtering from different directions tends to enhance or disclose linear features that lie preferentially near the perpendicular to the traverse direction. In the image below, the array was moved diagonally across the scene. Note the similarity to the third image above.

PREVIOUS NEXT

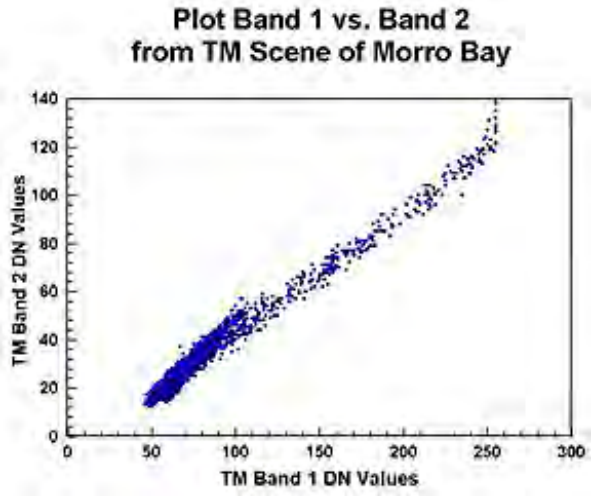
Primary Author: Nicholas M. Short, Sr.

There is a tendency for multiband data sets/images to be somewhat redundant wherever bands are adjacent to each other in the (multi-)spectral range. Thus, such bands are said to be correlated (relatively small variations in DNs for some features). A statistically-based program, called Principal Components Analysis, decorrelates the data by transforming DN distributions around sets of new multi-spaced axes. The underlying basis of PCA is described in a link. Color composites made from images representing individual components often show information not evident in other enhancement products. Canonical Analysis and Decorrelation Stretching are also mentioned.

Principal Components Analysis

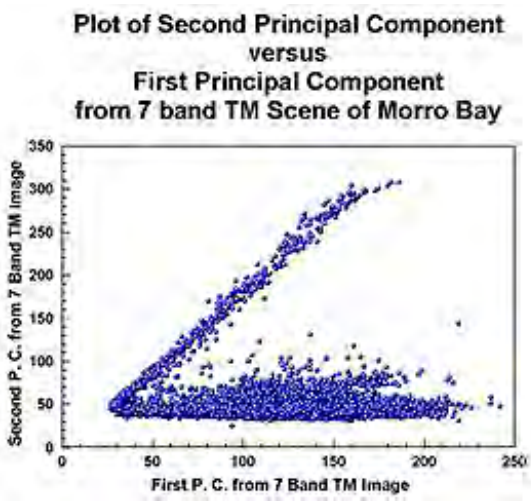
We are now ready to overview the last two types of image enhancement discussed in this Tutorial. Both are also suited to Information Extraction and Interpretation, but are treated separately from Classification (considered later in the Section). We will embark first on a quick run-through of images of Morro Bay produced by Principal Components Analysis (PCA), with explanation and commentary. But you have the option of going directly at this time to Appendix C which treats the theory of PCA; the theory is "tough", so take a look at the link and then decide. But first, a digression to discuss the possible effects of interband correlations.

You may have noticed in this Section that Band 1 and Band 2 images look almost alike. Bands 5 and 7 images are similar. Bands 1 and 4 appear dissimilar. The main reason for this is that the spectral responses for, say, Bands 1 and 2 for some given object or class are close (in a spectral signature plot, the intensities for each band, despite the moderate spectral separation, are often about the same). If the reflectance levels (as DNs) for all pixels in an image of, for example, TM Bands 1 and 2 are plotted, this results:

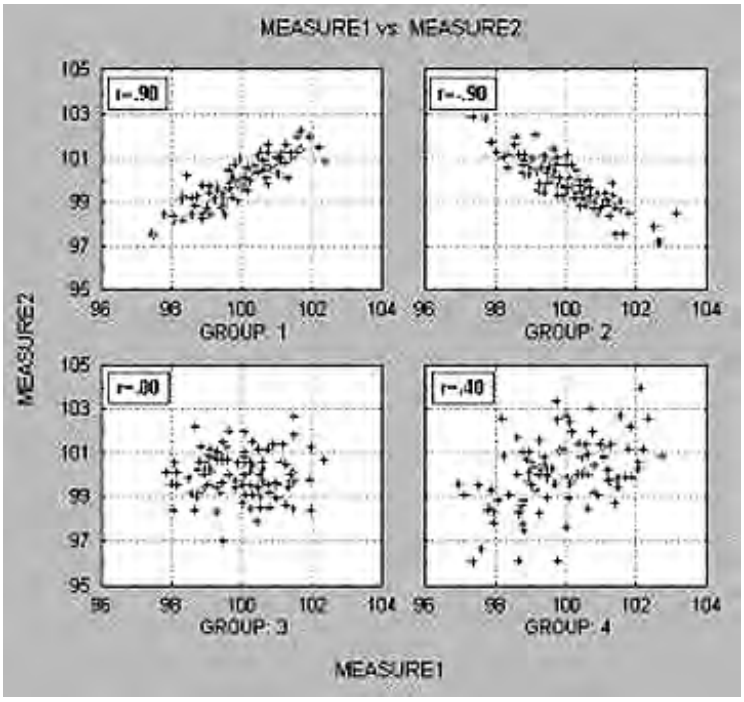


The slightly scattered data form a narrow plot that is almost a straight line. The two bands are said to be *correlated*, that is, as Band 1 varies, so does Band 2, and either could be used in place of the other. But, there are 7 bands in the TM data set, and some are sufficiently different from others as to behave as though uncorrelated. Principal Components Analysis is a mathematic technique that uses statistical methods to decorrelate the data and reduce redundancy. All available bands participate in the calculations and a new set of data values - the components - results. Here is a plot of the resulting first two components, showing the now emphatic

dissimilarities (again, for details you should work through Appendix C); note that the distribution of plotted values is spread apart, i.e., the data set is bimodal:

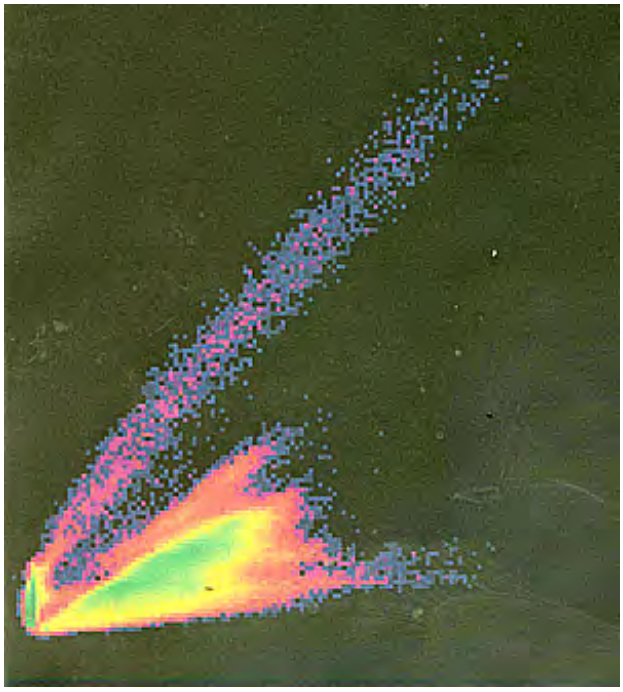


Having seen these two plots, we will digress for a moment to consider this general statistical property of data sets. Two measurements of the same variable, say reflectance, using two different ranges (for sensor data, two wavelength intervals or bands, here expressed as X and Y) will lead to several patterns in an X-Y two dimensional plot:



In the upper left diagram, the two measured variables X and Y plot for a large number of measurements in a Scatter Diagram along a central (unshown) straight line, on either side of which the points vary a little. This case denotes a high positive correlation - over the entire range, any measure of X (Measure 1) can predict the corresponding value of Y (2), and as X values increase, so do Y values. In the upper right, diagram a high negative correlation is indicated; as X values increase, their corresponding Y values decrease. The bottom two diagrams show a wide spread or scatter of values, such that the X and Y values are said to be poorly correlated, such that predictions of Y values from any X values are hampered since a wide range of Y values is possible. These diagrams refer to low correlations in which probabilities of getting specific values are low. As this applies to measurements of TM bands, 1 vs 2 produces a high correlation (so that their ability to discriminate is poor) whereas 2 vs 7 may give a low correlation (meaning that they may together yield independent assessments).

Let's check TM bivariate scatter plots for the Morro Bay bands. Various combinations were tried. One interesting one is Band 2 (abscissa) vs Band 4 (ordinate):



This plot shows a bimodal distribution of data points. The upper (blue & purple dots) plot shows strong correlation (spread around a mean line is small) between the two bands. This plot is for all the water in the scene - the DNs for water extend over a wide range of values but value changes in one band are matched by similar change increments in the other band. The second plot (orange/yellow/green) is for all other classes in the image. There is strong correlation when DN values are low but as these increase for both bands the plot widens. This means that for much of the DN value range, the two bands are less correlated and should serve increasingly well as discriminators in any classification.

Now, back to PCA: Lets assume for this page that you now know that the PCA output is a new set of DN pixels for each derived component. That DN set can be made to appear as an image that resembles to some extent any of the individual TM bands. We will now look at each of these components as images, keeping in mind that many of the tonal patterns in individual components do not seem to spatially match specific features or classes identified in the TM bands and represent linear combinations of the original values instead. We make only limited comments on the nature of those patterns that lend themselves to some interpretation.

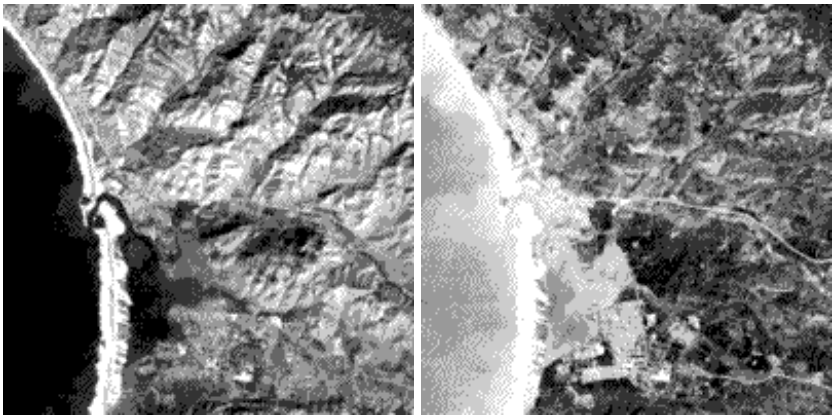
1-14: After reading through the special review of PCA accessed by link, plus the above paragraph, see if you can come up with a single key word (or perhaps a key idea in several words) that describes the main benefit from using Principal Components Analysis. ANSWER

The first Principal Component contains the maximum amount of variation in the 7-dimensional space defined by the seven Thematic Mapper bands. The image produced from PC 1 data commonly resembles an actual aerial photograph.

In fact, this is the normal character of the first component, in that it broadly simulates standard black and white photography and it contains most of the pertinent information inherent to a scene. The hills appear more realistic because the sharp light-dark contrast in most TM bands is subdued. Note the internal structure of the waves and the absence of any indication of sediment load in the sea. The histogram of the first PC shows two peaks. The first, on the left,constitutes the ocean pixels and the second one, to the right, the land pixels.

1-15: Describe this image relative to, say, the histogram-equalization stretched image seen on the previous page. ANSWER

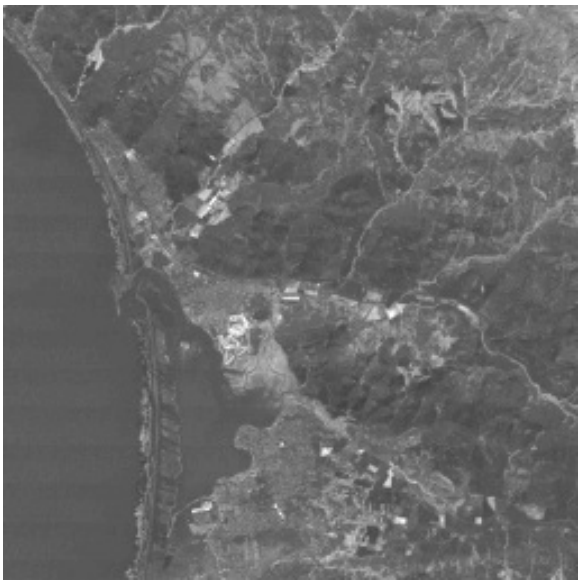
When we look at the histogram of the second PC (PC2), we see that even though the total range (maximum value - minimum value) is greater than for the first PC, most of the pixels fall in a small range around the mean of 49. Thus as is the convention the second PC has a smaller variance (variance is standard deviation squared) than the first PC. Since the bulk of the pixels falls in such a narrow range, the image does not seem to have as much interpretable patterns relatable to the classes as does any TM image (Band 3 appears in the next left image).



In order to make the PCA2 image (above right) viewable, we had to expand (extend the DN range numerically) its raw data set as a histogram equalization stretch. This procedure (histogram equalization) produces a histogram where the space between the most frequent values is increased and the less frequent values are combined and compressed. If we had not done this transformation, the image would appear tonally flat, with only two gray levels defining most of the land surfaces and one gray level defining the ocean. Some distinctions in the TM image that were previously small are now singled out and easier to see on the computer display. The breaker waves are uniquely singled out as very bright.

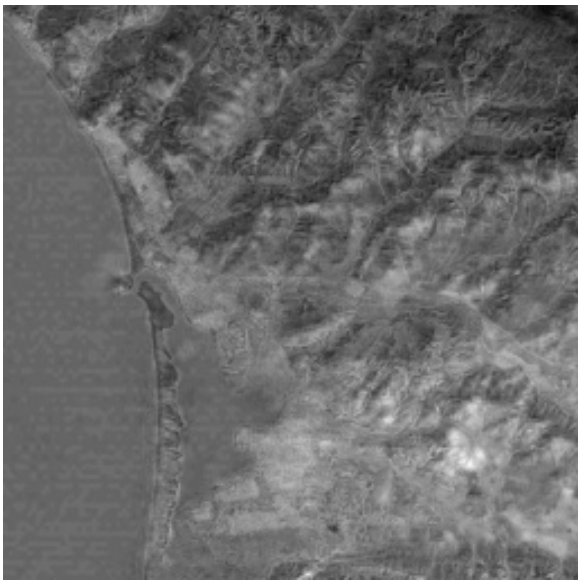
1-16: Make some general observations on how the tonal patterns in PC 2 differ from patterns observed in, for instance, Band TM 3. ANSWER

Some of the gray patterns in the PC3 image below can be broadly correlated with two combined classes of vegetation:



The brighter tones come from the fairways in the golf course and many of the agricultural fields. Moderately darker tones coincide with some of the grasslands, forest or tree areas, and coastal marshland. Note that both the beach and waves almost disappear as patterns.

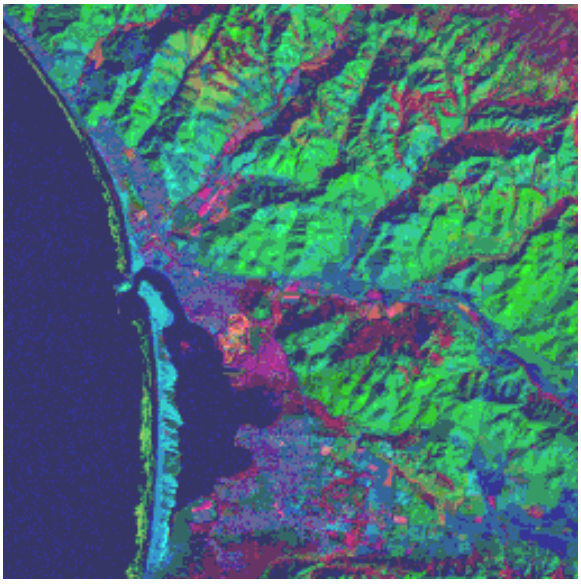
The breakers completely disappear in the PC4 image below while the rest of the scene is rather flat and mostly dark but with several patterns set forth in medium grays.



1-17: Anything unusual about PC 4 that might be meaningful? ANSWER

You may be wondering what the remaining PCs (through PC7) look like, and if they show any useful information. The response, after examining, for example, PC6, is that the features we are familiar with do appear but probably offer little new in interpretation. Note that the waves in the image below now are black - interesting but perhaps meaningless; the golf course pattern is also black.

The information available in the PCA images can be revealed better by combining them visually as registered overlays. Any three of these four PC images can be made into color composites with various assignments of blue, green, and red. In all, 24 different combinations are possible. Of those made experimentally for this review, this next image composed of PC 4 = blue, PC 1 = green, and PC 3 = red has proved the most interesting. In this rendition, the golf course has a singular color signature (orange-red) and a unique internal structure. Most other vegetation shows as red to purple-red tones, but the grasslands (**v**) has an unusual color, describable as greenish-orange. The brighter slopes of the hills and mountains appear as medium green, while some areas in shadow, are bluish. The urban areas also have a deep blue color. The beach bar now appears as turquoise and the adjacent breakers are olive-green.



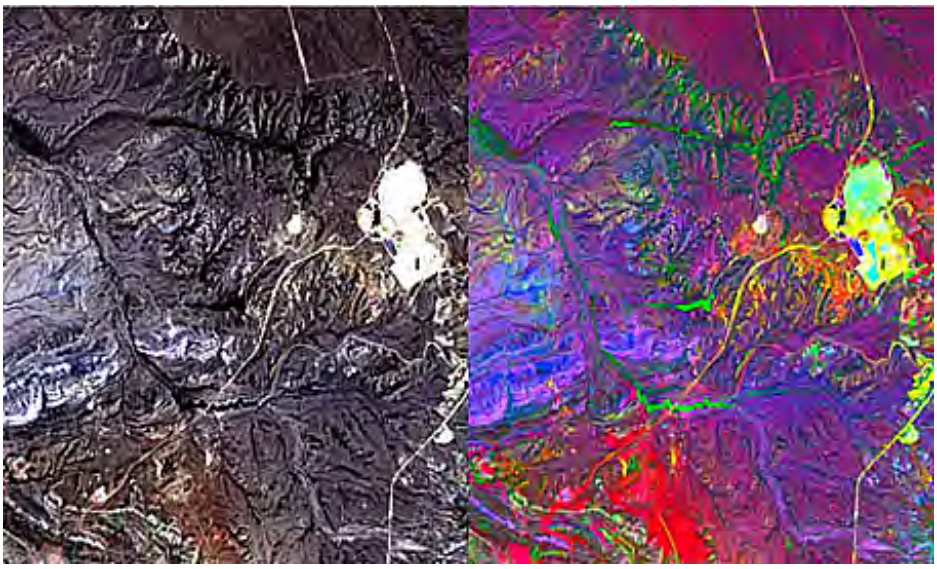
A very instructive example of a practical use of PCA is given in Section 5, page 3.

A variant of PCA is known as *Canonical Analysis (CA)*. Whereas PCA uses all pixels regardless of identity or class to derive the components, in CA one limits the pixels involved to those associated with pre-identified features/classes. This requires that those features can be recognized (by photointerpretation) in an image display (single band or color composite) in one to several areas within the scene. These pixels are "blocked out" as training sites much as you will see done in the Classification discussion beginning on page 1-16. Their multiband values (within the site areas) are then processed in the manner of PCA. This selective approach is designed to optimize recognition and location of the same features elsewhere in the scene.

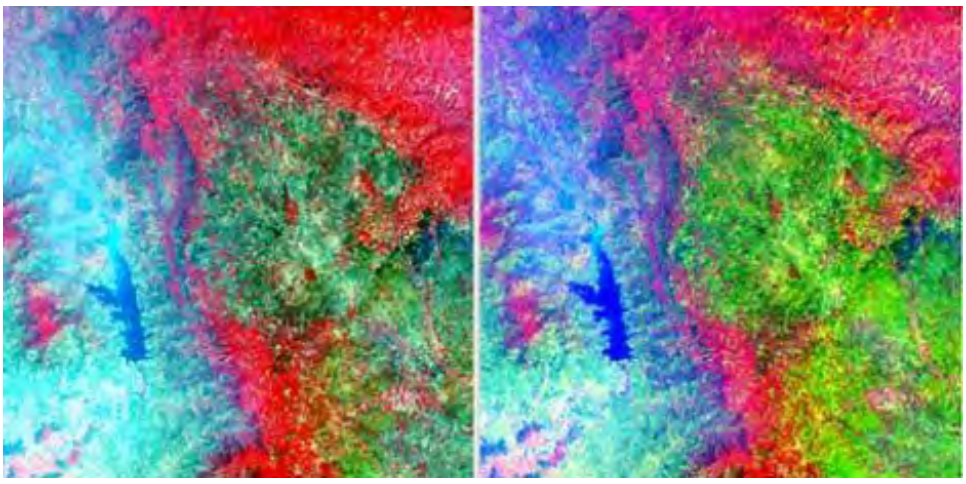
Another use of PCA, mainly as a means to improve image enhancement, is known as a *Decorrelation Stretch* (DS). The DS optimizes the assignment of colors that bring out subtle

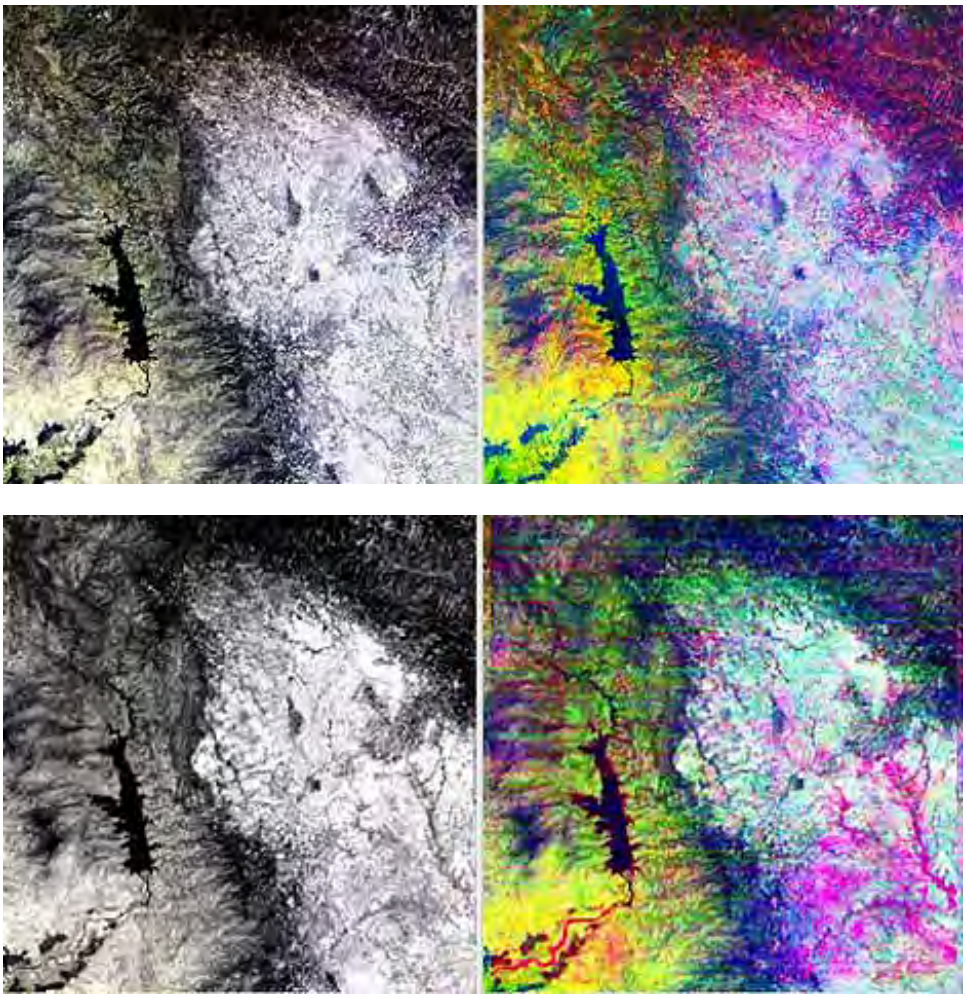
differences not readily distinguished in natural and false color composites. This reduction in interband correlations emphasizes small but often diagnostic reflectance or emittance variations owing to topography and temperature. The first step is to transform band data into at least the first three PCs. Each component is rescaled by normalizing the variance of the PC vectors. Then each PC image is stretched, usually following the Gaussian mode. The stretched PC data are then projected back into the original channels which are enhanced to maximize spectral sensitivity.

The IDRISI Windows version used to produce the various Morro Bay PCA images does not contain the last step in the DS process. However, here are two examples gleaned from the Internet. The first shows a Landsat subscene of an unidentified area: on the left is a standard false color composite; on the right a DS image - this illustrates the ability to extract and emphasize the tonal differences not apparent in the left image:



Users of ASTER data have found Decorrelation Stretching to be particularly effective. The stretches are variously informative depending whether the bands used are in the Visible, the SWIR, or the thermal IR interval. These three ASTER scenes (again, of an unidentified area) use inputs from 12 bands to show the effects of a DS; read the captions for the bands used.





The difference between the PC color composites and the DS color composites is generally not large, but extra statistic data manipulation in the latter often leads to a better product.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

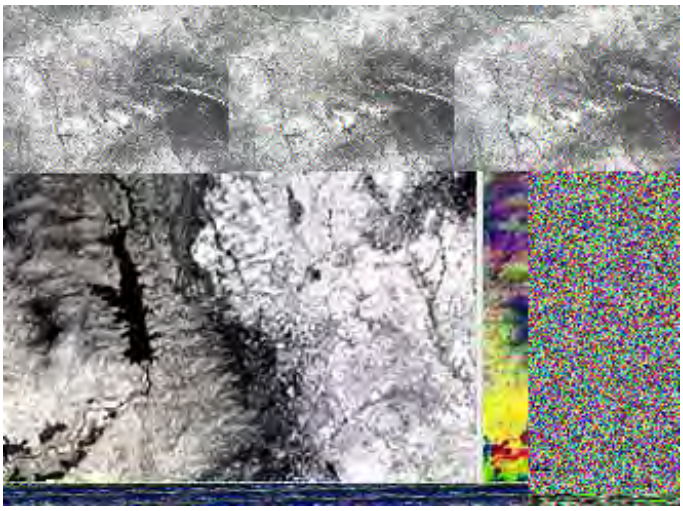
Ratioing is an enhancement process in which the DN value of one band is divided by that of any other band in the sensor array. If both values are similar, the resulting quotient is a number close to 1. If the numerator number is low and denominator high, the quotient approaches zero. If this is reversed (high numerator; low denominator) the number is well above 1. These new numbers can be stretched or expanded to produce images with considerable contrast variation in a black and white rendition. Certain features or materials can produce distinctive gray tones in certain ratios; TM Band 3 (red) divided by Band 1 tends to emphasize red- or orange-colored features or materials, such as natural hydrated iron oxide, as light tones. Three band ratio images can be combined as color composites which highlight certain features in distinctive colors. Ratio images also reduce or eliminate the effects of shadowing.

Ratioing

Another image manipulation technique is ratioing. For each pixel, we divide the DN value of any one band by the value of another band. This quotient yields a new set of numbers that may range from zero (0/1) to 255 (255/1) but the majority are fractional (decimal) values between 0 and typically 2 - 3 (e.g., 82/51 = 1.6078...; 114/177 = 0.6440...). We can rescale these by stretching to provide a gray-tone image of each Band X/Y ratio, in which each can fill out 16 or 256 levels, depending on the computer display limits.

One effect of ratioing is to eliminate dark shadows, because these have values near zero in all bands, which tends to produce a "truer" picture of hilly topography in the sense that the shaded areas are now expressed in tones similar to the sunlight sides. To illustrate this, suppose that an inclined surface of some material facing the Sun has a brightness value of 70 in one band (A) and 50 (B) in another. For the same material in shadow on a slope away from the Sun, the BV values may be 35 and 25 for A and B respectively. A divided by B in the Sun-facing case is 24; for the shadow case A/B = 1.4. So, in the ratio image, both front and back slope areas would have the same value. The problem, of course, is that for another material, with BVs of 51 (sunlit) and 37 (shadow), the A/B would also B 1.4, so that material is not differentiated from the first.

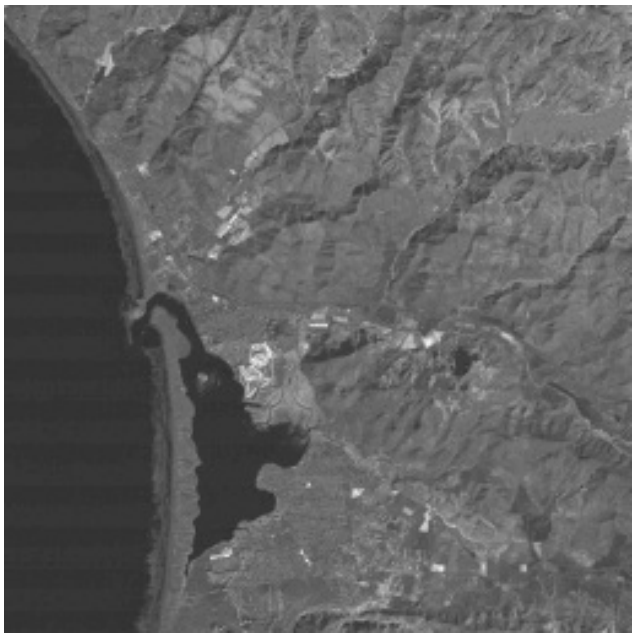
Still, while class-related tonal contrasts may smooth out, certain ratios can point to certain tonal anomalies that are diagnostic of special conditions. This first ratio image (in Wyoming), made from an MSS Band 5/MSS Band 4 data set, exemplifies this. A red surface would give a high DN value in Band 5 (MSS red band) and a low number in Band 4 (little green light), so that the ratio would be a large number that would produce a light tone in an image. In the image below, the bright tonal patch just below the center is an iron-rich anomaly (reddish on the ground) over a small oil field that has leaked petroleum which leached out iron oxide to form a tell-tale stain (the linear white pattern in the middle right is a red-colored geologic outcrop):



For a Landsat MSS data set, a total of six ratios (Bands 4/5. 4/6, 4/7, 5/6, 5/7, 6/7) and six reciprocals (5/4, 6/4....7/6) are possible. Three pairs of ratio images can be co-registered (aligned) and projected as color composites. In individual ratio images and in these composites, certain ground features tend to be highlighted, based on unusual or anomalous ratio values. For example, an ore deposit may be weathered or altered so that a diagnostic surface staining, called gossan, develops. This stain consists of hydrated iron oxide (rust) that is normally yellow-brown. In Band 3, this material reflects strongly in the red but it is apt to be dark in Band 4. The ratio quotient values for this situation tend, therefore, to exceed 2-3, giving rise to a bright spot pattern in a 5/4 image.

1-18: Although this is a rare natural situation, rocks containing native copper could develop a distinct greenish tone (think on occasions where you saw weathered copper roofing or a penny that had been outdoors a long time). What two TM bands might you ratio to distinguish this surface indicator of copper from natural vegetation. ANSWER

IDRISI generates ratio images through a module called OVERLAY. These may need to be rescaled and converted to byte format for display. However, the new images are difficult to combine in composites. We show just one ratio image made by dividing Band 1 DN values into Band 4 DNs, to illustrate a characteristic product.

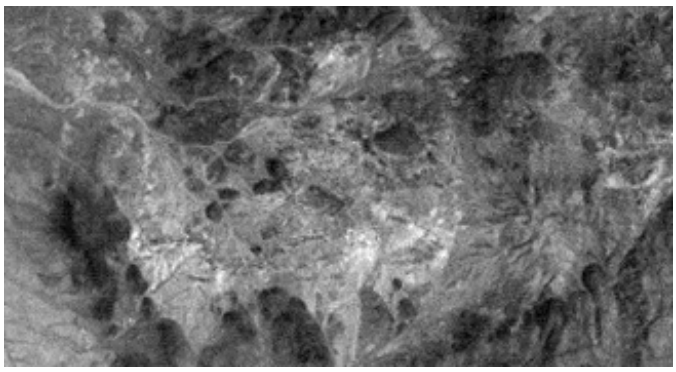


In this unstretched version, most of the image is moderately dark without much variation in gray levels. This implies similarities in DN values for equivalent pixels in the two bands. However, the tonally bright (high DNs) golf course fairways at (**s**) contrast sharply with their surroundings. Fields also stand out in this way. The grassy area at (**v**) has well-defined boundaries and is ligher

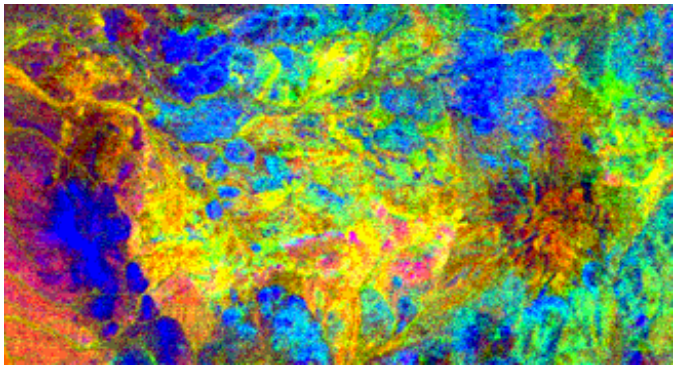
toned.

1-19: Which Principal Component image does this ratio image most resemble? ANSWER

(Note: the above ratio 1/4 image was produced on a DOS version of IDRISI. The writer (NMS) tried to create new ratio images on IDRISIW, the Windows version, with no luck. But excellent examples of ratio images, and ratio color composites, developed on the IDIMS system at NASA Goddard while the writer was still there, are shown on page 5-3.) To preview this, look at a TM Band 3/Band 1 ratio image of a mineralized area in southwestern Utah, and then a color composite made from 1/7 = blue; 4/2 = green; 3/1 = red.



The lightest tones coincide with iron enriched surface material (reds and yellows in the field).



The yellows and reds in this composite denote areas of rock alteration and mineralization.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

*This and the next four pages deal with the process of classifying multispectral images into patterns of varying gray tones or assigned colors that represent either clusters of statistically different sets of multiband data (radiances expressed by their DN values), some of which can be correlated with separable classes/features/materials (**Unsupervised Classification**), or numerical discriminators composed of these sets of data that have been grouped and specified by associating each with a particular class, etc. whose identity is known independently and which has representative areas (training sites) within the image where that class is located (**Supervised Classification**). The principles involved in classification, mentioned briefly in the Introduction, are explored in more detail. This page also describes the approach to unsupervised classification and gives examples; it is pointed out that many of the areas classified in the image by their cluster values may or may not relate to real classes (misclassification is a common problem).*

Classification

Now, at last we approach the finale of this Tutorial section during which we demonstrate two of the common methods for identifying and classifying features in images: *Unsupervised* and *Supervised Classification*. Closely related to Classification is the approach called *Pattern Recognition*. You may wish to read at the outset, the helpful Internet site that reviews Classification Procedures that is included in the Remote Sensing Core Curriculum online tutorial we have cited before.

Before starting, it is well to review several basic principles, considered earlier in the Introduction, with the aid of this diagram:

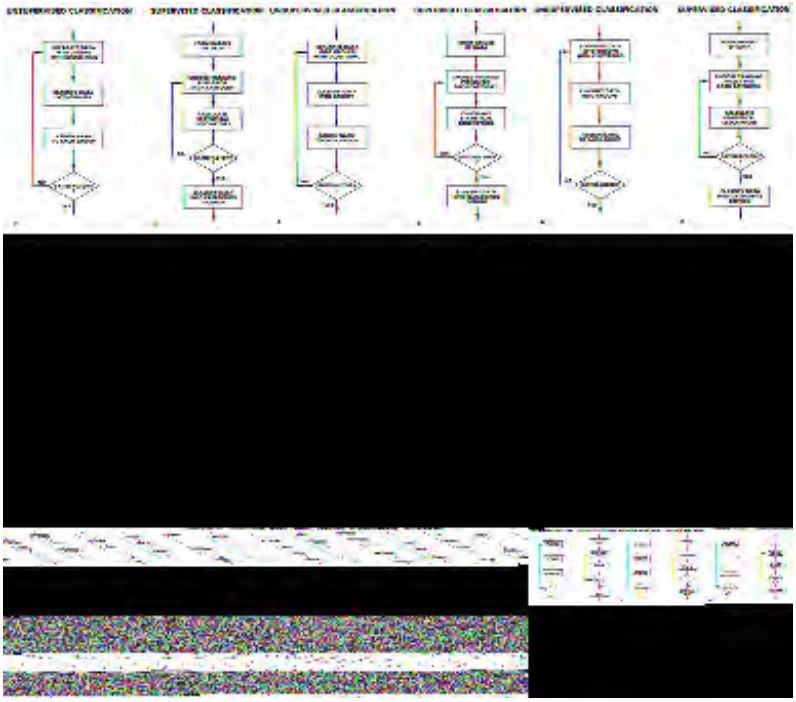
In the upper left are plotted spectral signatures for three general classes: Vegetation; Soil; Water. The relative spectral responses (reflectances in this spectral interval), in terms of some intensity unit, e.g., reflected energy (as a ratio of reflected to incident radiation) in appropriate units or as a percent, the unit times 100), have been sampled at three wavelengths. (The

response values are normally converted [either at the time of acquisition on the ground or onboard aircraft or spacecraft] to a digital format - the DNs or Digital Numbers cited before, commonly subdivided into units from 0 to 255 [2^8]).

For this specific signature set, the values at any two of these wavelengths are plotted on the upper right. It is evident that there is considerable separation of the resulting value points in this two-dimensional diagram. In reality, when each class is considered in terms of geographic distribution and/or specific individual types (such as soybeans versus wheat in the Vegetation category), as well as other factors, there will be usually some variation of a particular set of its characteristic DNs in one or both chosen wavelengths being sampled. The result is a spread of points in the two-dimensional diagram (known as a *scatter diagram)*, as seen in the lower left. For any two classes this scattering of value points may or may not overlap. In the case shown, which treats three types of vegetation (crops), they don't. The collection of plotted values (points) associated with each class is known as a *cluster*. It is possible, using statistics that calculate means, standard deviations, and certain probability functions, to draw boundaries between clusters, such that arbitrarily every point plotted in the spectral response space on each side of a boundary will automatically belong the class or type within that space. This is shown in the lower right diagram, along with a single point "w" which is an unknown object or pixel (at some specific location) whose identity is being sought. In this example, 'w' plots just in the soybean space.

Thus, the principle of classification (by computer image-processing) boils down to this: Any individual pixel or spatially grouped sets of pixels representing some feature, class, or material is characterized by a (generally small) range of DNs for each band monitored by the remote sensor. The DN values (determined by the radiance averaged over each spectral interval) are considered to be clustered sets of data in 2-, 3-, and higher dimensional plotting space. These are analyzed statistically to determine their degree of uniqueness in this spectral response space and some mathematical function(s) is/are chosen to discriminate the resulting clusters.

Two methods of classification are commonly used: Unsupervised and Supervised. The logic or steps involved can be grasped from these flow diagrams:



In *unsupervised classification* any individual pixel is compared to each discrete cluster to see which one it is closest to. A map of all pixels in the image, classified as to which cluster each pixel is most likely to belong, is produced (in black and white or more commonly in colors assigned to each cluster). This then must be interpreted by the user as to what the color patterns may mean in terms of classes, etc. that are actually present in the real world scene; this requires

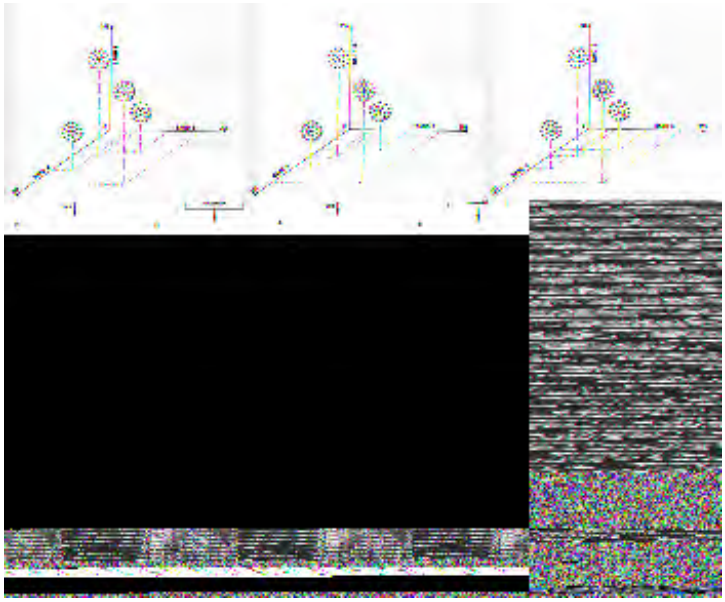
some knowledge of the scene's feature/class/material content from general experience or personal familiarity with the area imaged.

In a *supervised classification* the interpreter knows beforehand what classes, etc. are present and where each is in one to perhaps many locations within the scene. These are located on the image, areas containing examples of the class are circumscribed (making them training sites), and the statistical analysis is performed on the multiband data for each such class. Instead of clusters then, one has class groupings with appropriate discriminant functions that distinguish each (it is possible that more than one class will have similar spectral values but that is unlikely when more than 3 bands are used because different classes/materials seldom have similar responses over a wide range of wavelengths). All pixels in the image lying outside training sites are then compared with the class discriminants derived from the training sites, with each being assigned to the class it is closest to - this makes a map of established classes (with a few pixels usually remaining unknown) which can be reasonably accurate (but some classes present may not have been set up; or some pixels are misclassified.

Both modes of classification will be considered in more detail and examples given here and on the next 4 pages.

Unsupervised Classification

In an unsupervised classification, the objective is to group multiband spectral response patterns into clusters that are statistically separable. Thus, a small range of digital numbers (DNs) for, say 3 bands, can establish one cluster that is set apart from a specified range combination for another cluster (and so forth). Separation will depend on the parameters we choose to differentiate. We can visualize this process with the aid of this diagram, taken from Sabins, "Remote Sensing: Principles and Interpretation." 2nd Edition, for four classes: A = Agriculture; D= Desert; M = Mountains; W = Water.



From F.F. Sabins, Jr., "Remote Sensing: Principles and Interpretation." 2nd Ed., © 1987. Reproduced by permission of W.H. Freeman & Co., New York City.

We can modify these clusters, so that their total number can vary arbitrarily when four or more bands are involved. (This multidimensional situation is, of course, not readily displayed in a 3-D plot.) When we do the separations through a computer program, each pixel in an image is assigned to one of the clusters as being most similar to it in DN combination value. Generally, in an area within an image, multiple pixels in the same cluster correspond to some (initially unknown) ground feature or class so that patterns of gray levels result in a new image depicting the spatial distribution of the clusters. These levels can then be assigned colors to produce a

cluster map. The trick then becomes one of trying to relate the different clusters to meaningful ground categories. We do this by either being adequately familiar with the major classes expected in the scene, or, where feasible, by visiting the scene (**ground truthing**) and visually correlating map patterns to their ground counterparts. Since the classes are not selected beforehand, this latter method is called Unsupervised Classification.

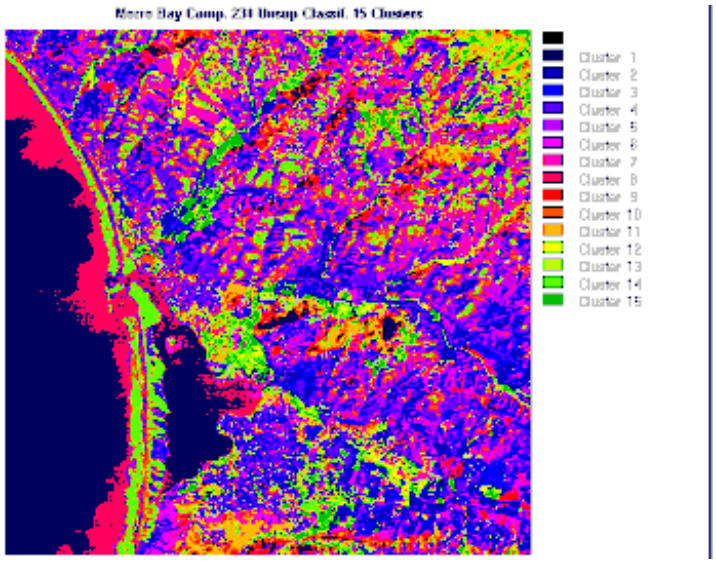
The IDRISI image processing program employs a simplified approach to Unsupervised Classification. Input data consist of the DN values of the registered pixels for the 3 bands used to make any of the color composites. Algorithms calculate the cluster values from these bands. The program automatically determines the maximum number of clusters by the parameters selected in the processing. This process typically has the effect of producing so many clusters that the resulting classified image becomes too cluttered and, thus, more difficult to interpret in terms of assigned classes. To improve the interpretability, we first tested a simplified output and thereafter limited the number of classes displayed to 15 (reduced from 28 in the final cluster tabulation).

1-20: What's to be done if one uses more than three bands to make an Unsupervised Classification? ANSWER

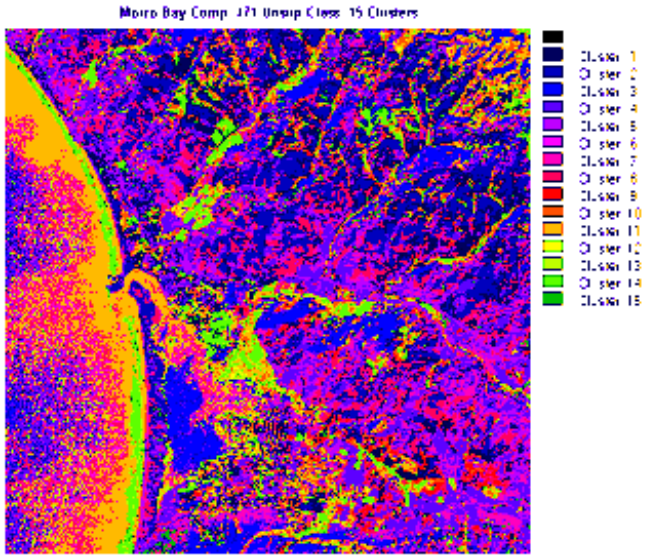
The first Unsupervised Classification operates on the color composite made from Bands 2, 3, and 4. Examine the resulting image when just 6 clusters are specified.

The light buff colors associate with the marine waters but are also found in the mountains where shadows are evident in the individual band and color composite images. Red occurs where there is some heavy vegetation. Dark olive is found almost exclusively in the ocean against the beach. The orange, green, and blue colors have less discrete associations.

We next display a more sophisticated version, again using Bands 2, 3, and 4, in which 15 clusters are set up; a different color scheme is chosen.



Try to make some sense of the color patterns as indicators of the ground classes you know from previous paragraphs. A conclusion that you may reach is that some of the patterns do well in singling out some of the features in parts of the Morro Bay subscene. But, many individual areas represented by clusters do not appear to correlate well with what you thought was there. Unfortunately, what is happening is a rather artificial subdivision of spectral responses from small segments of the surface. In some instances, we see simply the effect of slight variations in surface orientation that changes the reflectances or perhaps the influence of what we said in the Overview was "mixed pixels". When we try another combination, Bands 4, 7, and 1, the new resulting classification has most of the same problems as the first, although sediment variation in the ocean is better discriminated. One reason why both 15 cluster classifications don't grab one's attention is that the colors automatically assigned to each cluster are not as distinctly different (instead, some similar shades) as might be optimum.



1-21: Critique these two Unsupervised Classifications. What is shown well; poorly? Do you find it very helpful in pinpointing potential classes to be identified and then used in carrying out a supervised classification? ANSWER

The subject of "mixed pixels" is intimately embedded in any discussion of classification methodology, and particularly of accuracy of a classified map. For reasons of context, discussion of the mixed pixel concept is deferred until page 13-2; after reading it but you should understand its relevance to accuracy assessment (that page also expands upon the idea behind training sites).

There is another aspect of unsupervised classification that bespeaks of their limited value. Arbitrarily changing classification parameters can lead to images that don't closely resemble

each other and that are hard to judge as to the validity of the patterns as representative of actual ground classes or features. They are thus subjective! If 20 or 10 spectral clusters had been selected for the above examples instead of 15, the color display would have looked quite different. Part may be due to color assignments, but the patterns vary considerably in size and shape. Some features, such as the ocean remain constant but others such as those which associate with the hills can be notably changed. Again, if 6 reflected light bands were used instead of three, still different patterns emerge. The question of how many bands and how many clusters are assigned to the classification is not readily answered. One way to use unsupervised classifications is often impractical to execute, namely, to visit the scene with several different classifications and try to judge which seems to best describe the observables.

The writer's personal bias is that Unsupervised Classification is too much of a generalization and that the clusters only roughly match some of the actual classes. Its value is mainly as a guide to the spectral content of a scene to aid in making a preliminary interpretation prior to conducting the much more powerful supervised classification procedures.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The principles behind Supervised Classification are considered in more detail. The fact that the pixel DNs for a specified number of bands are selected from areas in the scene that are a priori of known identity, i.e., can be named as classes of real features, materials, etc., allows establishment of training sites that become the basis of setting up the statistical parameters used to classify pixels outside these sites. The process by which training sites were selected for the Morro Bay scene, together with a plot of spectral signatures and a table of statistical means and site sample sizes, illustrate the preparatory phase involved in this mode of classification.

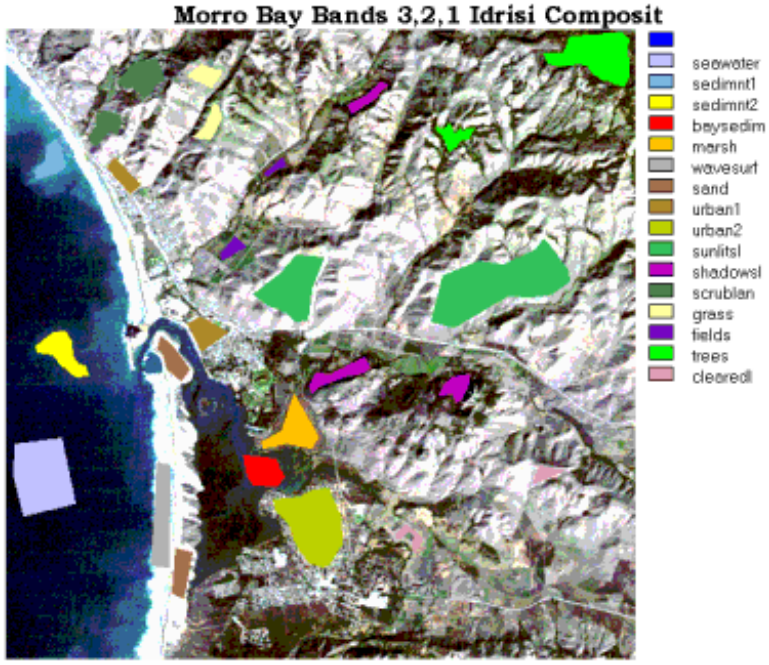
Supervised Classification

Supervised classification is much more accurate for mapping classes, but depends heavily on the cognition and skills of the image specialist. The strategy is simple: the specialist must recognize conventional classes (real and familiar) or meaningful (but somewhat artificial) classes in a scene from prior knowledge, such as personal experience with what's present in the scene, or more generally, the region it's located in, by experience with thematic maps, or by on-site visits. This familiarity allows the individual(s) making the classification to choose and set up discrete classes (thus supervising the selection) and then, assign them category names. As a rule, the classifying person also locates specific training sites on the image - either a print or a monitor display - to identify the classes. The resulting **Training sites** are areas representing each known land cover category that appear fairly homogeneous on the image (as determined by similarity in tone or color within shapes delineating the category). In the computer display one must locate these sites and circumscribe them with polygonal boundaries drawn using the computer mouse. For each class thus outlined, mean values and variances of the DNs for each band used to classify them are calculated from all the pixels enclosed in each site. More than one polygon is usually drawn for any class. The classification program then acts to cluster the data representing each class. When the DNs for a class are plotted as a function of the band sequence (increasing with wavelength), the result is a **spectral signature** or spectral response curve for that class. The multiple spectral signatures so obtained are for all of the materials within the site that interact with the incoming radiation. Classification now proceeds by statistical processing in which every pixel is compared with the various signatures and assigned to the class whose signature (usually as a data set within the computer rather than a plot) comes closest. A few pixels in a scene do not match and remain unclassified, because these may belong to a class not recognized or defined.

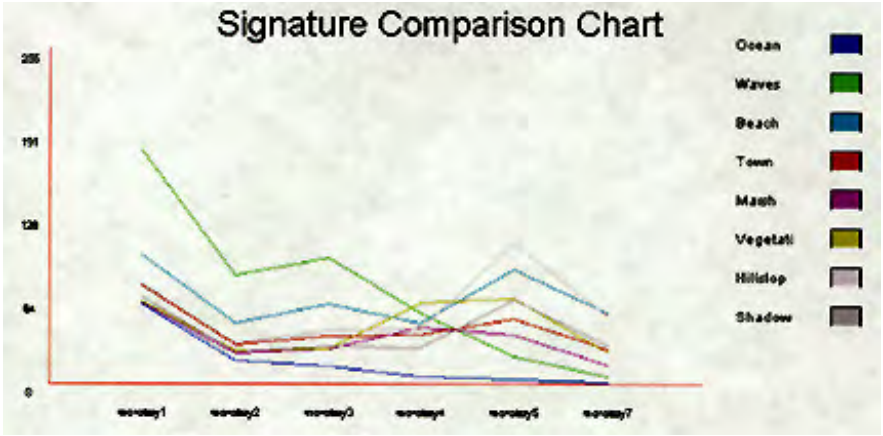
Many of the classes for the Morro Bay scene are almost self-evident: ocean water, waves, beach, marsh, shadows. In practice, we could further sequester several such classes. For example, we might distinguish between ocean and bay waters, but their gross similarities in spectral properties would probably make separation difficult. Other classes that are likely variants of one another, such as, slopes that faced the morning sun as Landsat flew over versus slopes that face away, might be warranted. Some classes are broad-based, representing two or more related surface materials that might be separable at high resolution but are inexactly expressed in the TM image. In this category we can include trees, forests, and heavily vegetated areas (the golf course or cultivated farm fields). For high spatial resolution and hyperspectral data, the trees can often be subdivided to the species level.

For the first attempt at a Supervised Classification, the writer set up 17 discretional classes. These were picked largely from my on site experience (as indicated on page 1-1, I went to Morro Bay on vacation in California even before I attempted the classification). These sites are shown

as color polygons traced on the true color (Bands 1,2,3) composite, as shown next. (Note that their site colors are assigned here for display convenience and do not correspond to their class equivalent colors in the maps shown on the next page).



Note that IDRISI does not name them during the stage when the signatures are made. Instead, IDRISI numbers them and names are assigned later. Several classes gain their data from more than one training site. IDRISI has a module, SIGCOMP, that plots the signature of each class. Here we show Signature plots for 8 general classes (not all the same as the training sites shown above; this is done to simplify the first product). These are Ocean = Deep Blue; Waves = Green; Beach = Lighter Blue; Town = Maroon; Marsh = Purple; Vegetation = Olive; Hillslope = Light Gray; Shadow = Darker Gray. All TM bands (listed on the abscissa as Morobay1..., Morobay2, etc.) except 6 are involved. The ordinate DN values range from 0 to 255.



The signature positions at each band are somewhat artificial in that the raw data were not adjusted for relative gains; this affects Bands 1 and 5 in particular. However, it is clear from these plots that most of the signatures are different from one another over the 6 bands used, even though some are close-spaced (nearly coincident) over intervals consisting of several adjacent bands, e.g, all but Waves and Beach between TM1 and TM2. The greatest separability among all classes occurs in Band 5.

IDRISI also has a program that presents pixel information for each signature, recording the number of pixels contributing to the data, and the mean, maximum, minimum, and standard deviation of DN values for each signature. To help you get a deeper feel for the numerical inputs involved in these calculations, we have reproduced a simplified version of these data in the following table:

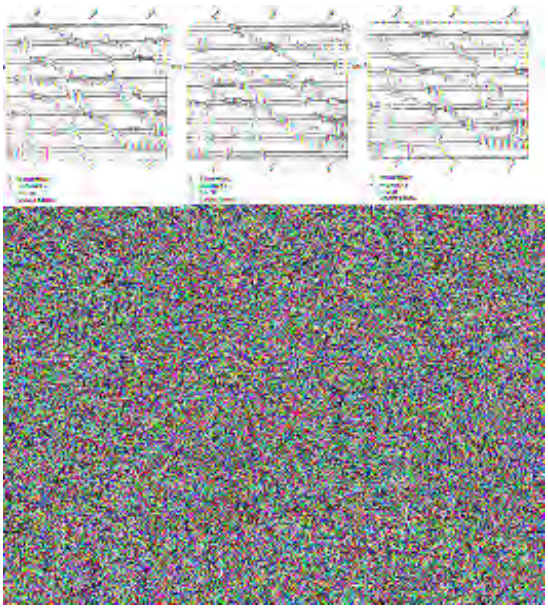
Table of Band Means and Sample Size for Each Class Training Set

BAND:	1	2	3	4	5	6 (TH)	7	No. of Pixels
Class								
1. Seawater	57.4	16.0	12.0	5.6	3.4	112.0	1.5	2433
2. Sediments1	62.2	19.6	13.5	5.6	3.5	112.2	1.6	681
3. Sediments2	69.8	25.3	18.8	6.3	3.5	112.2	1.5	405
4. Bay Sediment	59.6	20.2	16.9	6.0	3.4	111.9	1.6	598
5. Marsh	61.6	22.8	27.2	42.0	37.3	117.9	14.9	861
6. Waves Surf	189.5	88.0	100.9	56.3	22.3	111.9	6.4	1001
7. Sand	90.6	41.8	54.2	43.9	86.3	121.3	52.8	812
8. Urban1	77.9	32.3	39.3	37.5	53.9	123.5	29.6	747
9. Urban2	68.0	27.0	32.7	36.3	52.9	125.7	27.7	2256
10. Sun Slope	75.9	31.7	40.8	43.5	107.2	126.5	51.4	5476
11. Shade Slope	51.8	15.6	13.8	15.6	14.0	109.8	5.6	976
12. Scrublands	66.0	24.8	29.0	27.5	58.4	114.3	29.4	1085
13. Grass	67.9	27.6	32.0	49.9	89.2	117.4	39.3	590
14. Fields	59.9	22.7	22.6	54.5	46.6	115.8	18.3	259
15. Trees	55.8	19.6	20.2	35.7	42.0	108.8	16.6	2048
16. Cleared	73.7	30.5	39.2	37.1	88.4	127.9	45.2	309

1-22: Examine the signature plots and the table. What can you say about the plots in terms of similarities and differences? Based on numbers in the table, would you predict any notable differences in the signatures for towns; marsh; sunlit hillslopes and shadows? ANSWER

We can deduce from this table that most of the signatures have combinations of DN values that allow us to distinguish one from another, depending on the actual standard deviations (not shown). Two classes, Urban 1 and Cleared (Ground), are quite similar in the first four bands but apparently are different enough in Bands 5 and 7 to suppose that they are separable. The range of variations in the thermal Band 6 is much smaller than in other bands, suggesting its limitation as an efficient separator. However, as we will see next, its addition to the Maximum Likelihood Classification increases the spatial homogeneity of some classifications.

As another aside: In some older computer-based classification systems, one (sometimes the only) hard copy output was a printout that showed the different classes in alphanumeric symbols. Here is an example for a scene in central Pennsylvania:



Since the mid-1980sm most final products are color images that can be printed as photographs.

There are scores of different classifiers.We will limit our examination of specific Classification Methods to just three, the first two being widely used in determining classes in space imagery. Among those omitted we cite: Parallelepiped; Piecewise Linear; Binary Decision Tree; Density Based Clustering; others.

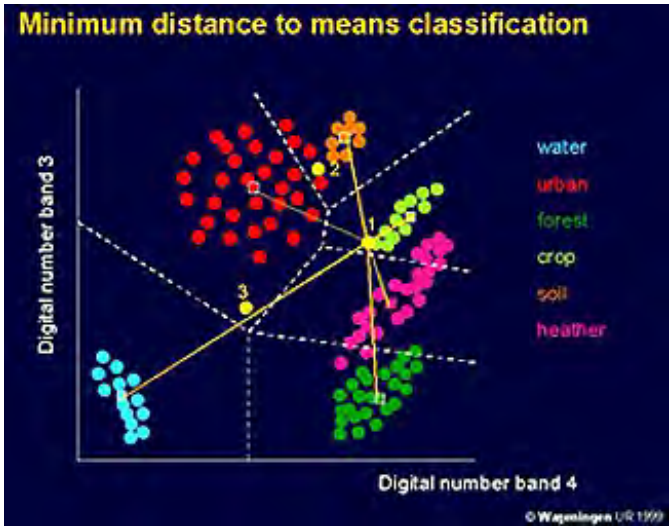
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

One of the simplest supervised classifiers is the parallelopiped method. But on this page we employ a (usually) somewhat better approach (in terms of greater accuracy) known as the Minimum Distance classifier. This sets up clusters in multidimensional space, each defining a distinct (named) class. Any pixel is then assigned to that class it is closest to (shortest vector distance).

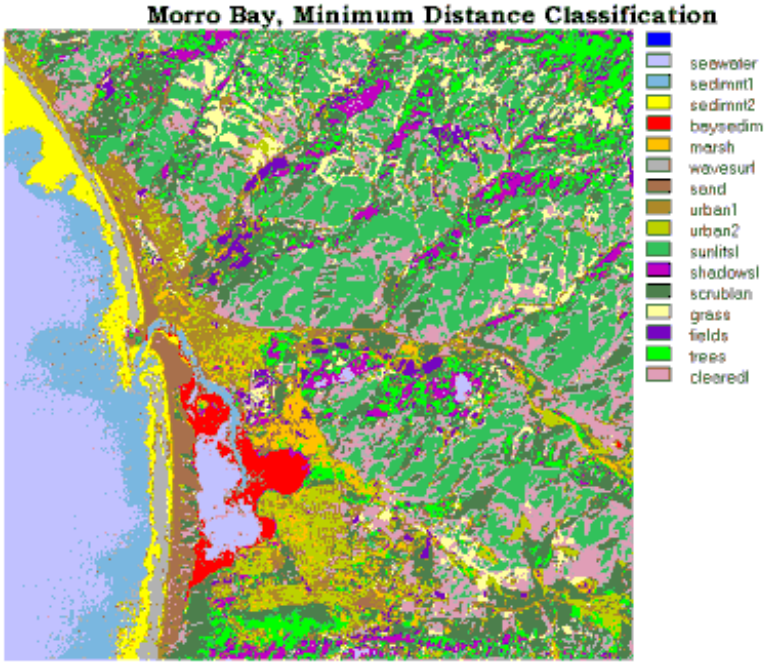
Minimum Distance Classification

We initiate our exemplification of Supervised Classification by producing one using the Minimum Distance routine. This involves fairly simple calculational procedures. Look at this diagram (for simplification, just two bands are used):



The data points for DNs from two bands are dots; the mean for each clustered data set are square. For point 1, an unknown, the shortest straight-line distance to the several means is to the class "heather". Point 1, then, is assigned to this category. Point 2 is slightly closer to the "soil" category but lies within the edge of the "urban" spread. Here, the classification seems ambiguous. By the minimum distance rule, it would go to "soil" but this may be erroneous ("urban" would have been a greater likelihood). Point 3 is not near any of the class DN clusters, but is about equidistance between "urban", "water", "forest", and "heather". If one plays the odds, "urban" is just a tad closer to 3; but this situation indicates how misclassification might occur.

The IDRISI program acts on DNs in multidimensional band space to organize the pixels into the classes we choose. Each unknown pixel is then placed in the class *closest* to its mean vector in this band space. For Morro Bay, the resulting classification image consists of 16 gray levels, each representing a class, to which we can then assign any color on the computer. We can elect to combine classes to have either color themes (similar colors for related classes) and/or to set apart spatially adjacent classes by using disparate colors. Examine this Minimum Distance classification below, in which we use all seven TM bands, including the thermal. Study it in relation to your acquired knowledge of this scene from the preceding pages in this section and compare it with the classifications we show on the next page.



A variant of this classifier is known as the Nearest Neighbor Classifier.

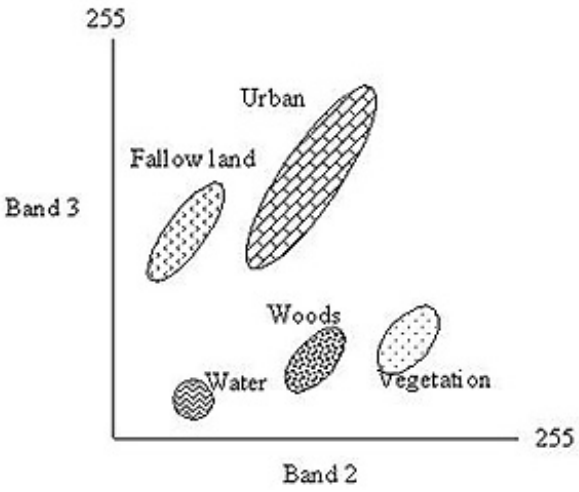
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

One of the most powerful classifier in common use is that of Maximum Likelihood. Based on statistics (mean; variance/covariance), a (Bayesian) Probability Function is calculated from the inputs for classes established from training sites. Each pixel is then judged as to the class to which it most probably belong. This is done with the Morro Bay TM data, using first the six reflected radiation bands and then with the longer wavelength thermal band added. The result is a pair of quite believable classification maps whose patterns (the classes) seem to closely depict reality but keep in mind that several classes are not normal components of the actual ground scene, e.g., shadows. A later classification (made using IDRISI for Windows) using a smaller number of classes give a somewhat different end product.

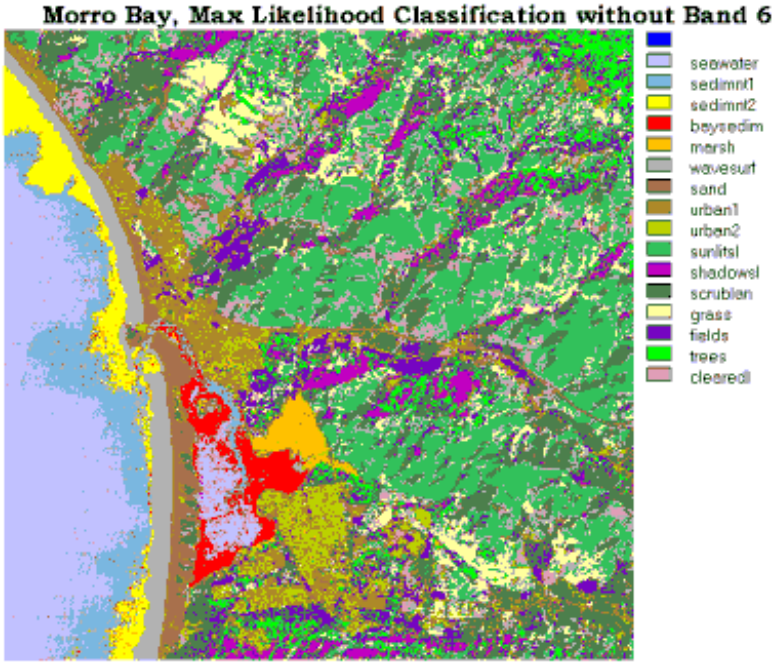
Maximum Likelihood Classification

In the computation of DN distributions, again determining the means and covariances for the classes that fall within the training sites, a statistical (Bayesian) Probability Function is also calculated for the class data. The class DN distribution, again for a two-dimensional case, gives rise to elliptical boundaries which define the equiprobability envelope for each class. Here is a scatter plot that shows the outer envelope bounding each class:



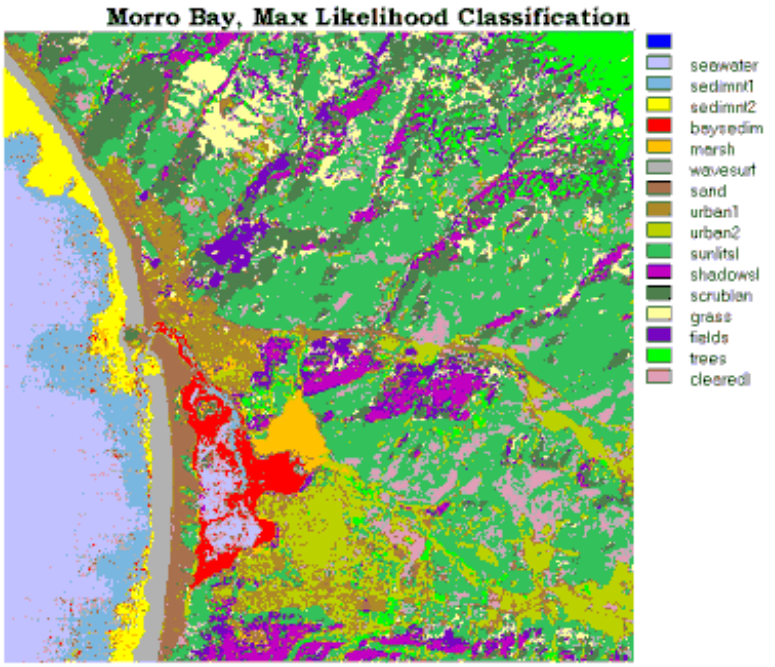
Not shown is the fact that inside each ellipse are contours that indicate the degree of probability. Associated with each ellipse is a separate plot that expresses a statistical surface (bell-shaped in three dimensions) called probability density functions. Using these functions, which relate to the contours, a likelihood that any unknown point U is most probably associated with some one ellipse is determined. A Bayesian Classifier is a special case in which the likely occurrence of each class (common to rare) is assessed and integrated into the decision making.

We made this Supervised Classification using the Maximum Likelihood classifier acting on all seven bands. Again, multiband classes are derived statistically and each unknown pixel is assigned to a class using the maximum likelihood method.



1-23: While the minimum distance and this first Maximum Likelihood classifications are, for the most part, quite similar, there are some differences. Mention several. ANSWER

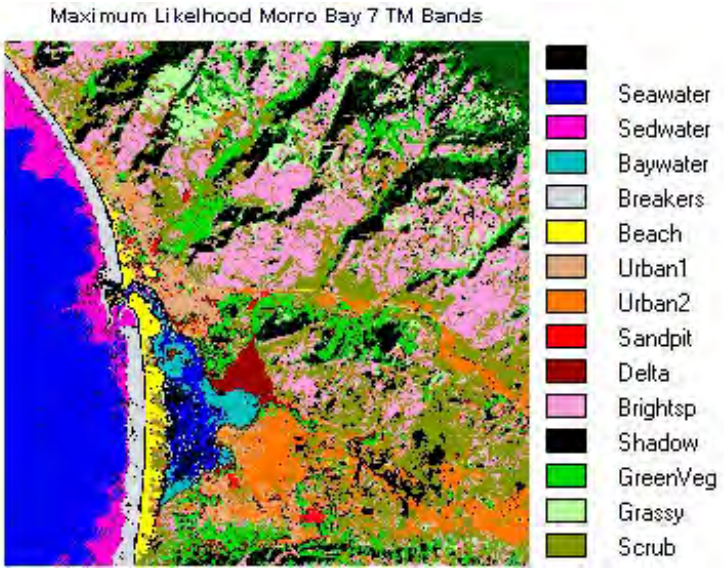
In this image we omit thermal Band 6 and define 16 classes (this is the maximum allowable in the IDRISI program). These classes are identical to the previous ones recorded in the Minimum Distance image. In both instances, the *Sediment* class has been subdivided into three levels (I and II in the ocean and a third in the Bay) and two *Urban* classes (I = Morro Bay; II = Los Osos) are attempted, to account for visual differences between them (mainly street patterns). Look at this image classification and decide how believable it is. Compare it with the minimum distance image as well. To assist you in comparing similar classes, we used the same color assignments. Next, look at a Supervised Classification that uses Band 6 and again specifies 16 classes. Notice how each urban area becomes more homogeneous. There is a similar increase in spatial homogeneity of vegetation and slopes in general with Band 6 added, but overall adding Band 6 didn't show much differences.



Each 16-class Maximum Likelihood version is a fairly dazzling image, with many classes "right on". Both *Breakers* and sand bar (*Beach*) seem uniformly classified. The sediment load distribution is credible. There are enough color tone differences between Morro Bay and Los Osos to justify the decision to make them two *Urban* classes (Los Osos differs in its street

patterns and in the presence of the orange-brown soil, seen in the composite of Bands 1,2, and 3). However, color elements of one *Urban* class are mixed with the other, in differing proportions, as one would expect. The bright orange given to the coastal *Marsh* area occupies a slightly larger area than its equivalent does in the Minimum Distance classification and is also distributed in small patches around the Los Osos coastline, and again along the river. Thus it is probably a true condition, in that, we expect such vegetation to be more widespread. No doubt the most uncertain group of classes is spread over the hills. The categories *SunLit Slope* and *Shadow Slope* are somewhat artificial, in that they refer mostly to an illumination condition. Whereas the grass and trees classes may be a mix of lighting effects and a lighter or darker surface. The class *Cleared Land* is, again, a depiction of land surfaces that may support, not only thin natural vegetation or even be partially barren but also may in some places have a shadowing effect. The *Grasslands* is properly placed in this image but appears to spread over wider areas than indicated in several other images, so it is doubtless a valid case. The *Green Vegetation* category proxies well for the actual distribution of reflective organic material (in Band 4) but in this choice of class assignments, several types of growing ground cover are not singled out. Thus, elements of the golf course and the mountain crest forest are shown as "like" and are not distinguished from field crops, etc. We could tell them apart to some degree of correctness, if we had given each its own class and selected training sites.

Nearly two years after the above Supervised Classifications were generated, an occasion arose to redo the same scene using new IDRISI software that operates from Windows, Version 1, rather than DOS 3.1. In performing this Supervised Classification, we used the same Maximum Likelihood classifier with all seven TM bands and 15 classes. But, as an experiment, we decided to drop several class categories and select new ones instead. Also, we established some slightly different training site polygons for each class. In effect, we achieved an independent classification without peeking at the results, shown above, for guidance. And, instead of using the natural color scene from which to pick training sites, we used the false color image. This is the result:



(ERROR: For some reason, the Windows IDRISI does not show the 15th class in the legend. This class should be "Trees", in dark green, present mainly in the upper right corner of the class map. Also, the first legend box (black here; blue in the two other classifications above) has no label; it is not a named class but refers to the color used outside the map image as background.)

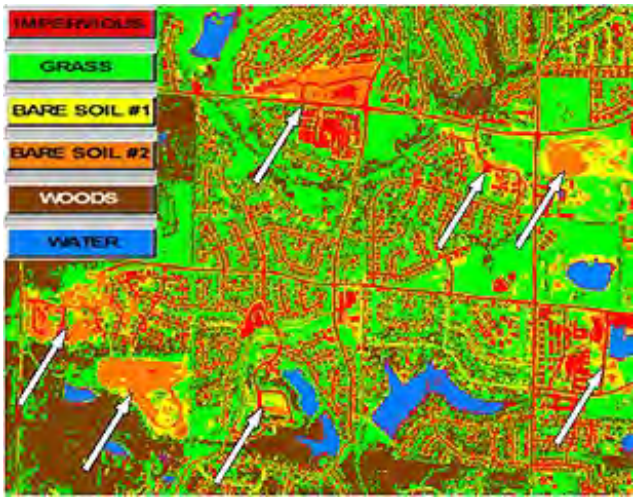
Note that for some of the classes, we assigned different colors than used in the first two maxlike classifications, which makes it rather difficult to compare the results with the earlier classifications. Nevertheless, scrolling between this and the 7-band, Supervised Classification just above, reveals differences and similarities.

In the Windows version, the two *Sediment* classes are combined. Also, the class, called *Fields* in

the DOS 3.1 version, is here renamed *GreenVeg*, and includes fields with crops and also some natural vegetation (probably local woodlands). Both show bright red in the false color rendition. The distribution of the class *Trees* is similar in both classifications but is a bit more widespread in the Windows version (but harder to see because dark green and black shadows do not show contrast well). The classes *Scrubland* and *Cleared* in the DOS 3.1 version are partially represented by *Scrub* in the Windows version. In DOS 3.1, *Urban II* (focused on the Los Osos street pattern) is olive and is orange in the Windows version. In both cases, the distribution of the Urban II class pattern is much more extensive than is the real situation. Town structures or clusters of buildings do not exist in the long orange strip near the highway, nor in the lower right part of the image. Apparently, some natural surfaces, as interpreted from the true and false color composite images, give rise to signatures that resemble this urban class. In the Windows version, several very bright areas, mainly around Los Osos, have been named *Sandpit*. This is a guess, because they may be excavated ground or inland remnants of beach sand (although they classify as distinct from the *Sand* Class); only an on-site visit could ascertain a correct identity.

The point in running and comparing these two classifications is probably obvious: the precise end result is sensitive to the variables involved and the choices we made - mainly in extrapolating classes from their training sites to the identities and distribution of the selected classes, i.e., the overall appearance and accuracy of the classifications. Interpretations differ depending on the colors and other factors present in the training image, by which we choose separable classes and block efficient training sites. The number of classes, the validity (purity) of the enclosed space in the training sites (and the number of pixels in the polygons assigned to each class), the nature of a class (the Urban division is somewhat artificial and Scrub may be rather dissimilar classes or features in the real world), the colors assigned to the final map, and other considerations all contribute to differences. Once again, we emphasize the argument that field work, if logistically possible, before and after computer-based classification of an image, is the key to selecting and then checking class locations. Thus it is the best insurance for achieving a quality product. But, if an on-site visit is not feasible, a skilled interpreter can develop a fairly reasonable classification based mainly on his/her abilities in recognizing obvious ground features in the scene. The writer (NMS) has achieved believable classifications of many parts of the world without any field work, but just from his knowledge of the appearance of the common components of a landscape or land-use categories. You, the reader have the same option of scene interpretation based on your general experience; surprisingly, this often works well.

Several of the commercial satellites described in the Introduction have supplied 4-band multispectral imagery and a higher resolution panchromatic band. A check of websites on the Internet indicates that use of the highest resolution band set increases the accuracy of class/feature identification, in some cases to the 90+% level. Here is an IKONOS image of part of the town of Columbia, Missouri (where the writer lived in 1954 while teaching at the University of Missouri), and beneath it a supervised classification involving classes pertinent to evaluation of neighborhood construction; the result had a 93% accuracy level.



PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

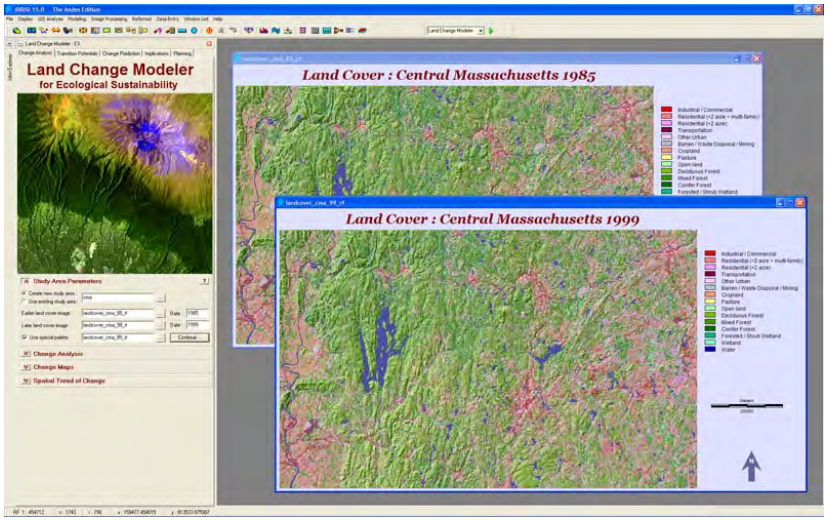
Section 1 is a key or pivotal "chapter" in this Tutorial, in that you will be introduced to two prime approaches in the use of remote sensing: 1) standard photo-interpretation of scene content, and 2) use of digital image processing and classification techniques that are generally the mainstay of practical applications of information extracted from sensor data sets. To accomplish this, we will utilize just one Landsat TM subscene that covers the Morro Bay area on the south-central coast of California. On this first page, a full Landsat TM image and one TM band (3) subscene image, both showing Morro Bay, are displayed. The concept of plotting its gray scale values or digital numbers (DN) is exemplified, and a brief description of the geographic setting of MB is given.

IMAGE PROCESSING AND INTERPRETATION:

MORRO BAY, CALIFORNIA

Assuming that you have read the preceding Introduction to remote sensing, you should be ready to work with the basic techniques that image interpreters utilize in examining and analyzing an image. In this Section we concentrate on computer processing of and information extraction from a single Landsat image - a subscene (from the digitized data set that covers a much larger area in California) of Morro Bay, a town on the California coast. But at times we will use images from other parts of the country either because the Morro Bay scene does not favorably illustrate the processing product or the computer program we will use does not have the particular function we are demonstrating.

We will apply and evaluate most of the routine or specialized analysis techniques used in computer processing of satellite remote sensing data (see end of page 1-20 for a brief list of books on the principles of image processing). For the Morro Bay subscene the various images shown in this Section were created using the IDRISI software processing program (it's worthwhile to check their ClarkLabs website [Clark University in Worchester, Mass.]). The IDRISI program is especially user-friendly to students wishing to gain experience in image processing. A typical product (2008-vintage) looks like this;



At the outset, we want to give you an option, namely to run through a Website that is brief but

comprehensive in its treatment of the basics of image processing of remotely sensed data. This preview should help you as we apply these ideas to the Morro Bay image. To choose this path, just click on this Canadian Soonet link.

Morro Bay in Context

The chosen subscene lies along the central California coast about half way between San Francisco and Los Angeles, in the county of San Luis Obispo. Before looking at the subscene, it is appropriate to locate it in a full Landsat scene. Below is a "true color" Landsat image acquired on December 8, 1982 that shows the California Coast, the Coast Ranges, and a small segment of the San Joaquin Valley. At the top left is part of Monterrey Bay. At the very bottom, on the coast, is Morro Bay itself. As seen here, almost the entire area included in the subscene (from a different date) is present but appears so tiny in this small scale full scene that it may prove hard for you to make it out.

TM Natural Color Image of Morro Bay, California

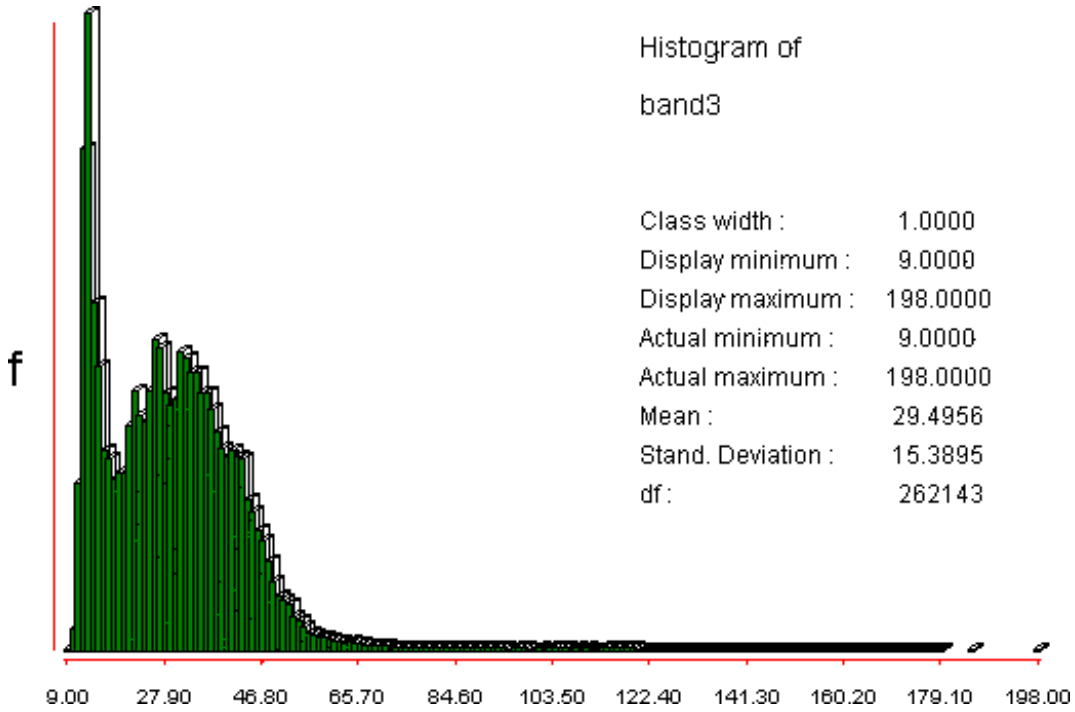
We extracted the subscene we will use from Landsat 5 Thematic Mapper scene 5026-31810 (Path 043; Row 035: Worldwide Reference System [WRS]) acquired on November 19, 1984. The actual data set is part of the Education sampler offered to users by the Space Imaging Corp. (the company has merged with another called Eosat that is located in Lanham, MD; Eosat, which produced a collection of TM imagery for educational uses, including the Morro Bay subscene, no longer supplies this set). To acquaint yourself with this subscene, look at the black and white TM Band 3 image which has been enlarged to a convenient size to allow you to look for details and specific features singled out later in the Section.

TM Band 3

1-1: After finding this subscene in the full image above, locate Morro Bay on the California map in a U.S. Atlas. Familiarize yourself with the area that includes the subscene and with the surrounding region (up to about 162 km [100 miles]). ANSWER

The subscene consists of a 512 x 512 pixel array (because each pixel represents 30 m [98ft] each side of the image is just over 16 kilometers [10 miles] long). Below is a histogram giving a frequency distribution of the brightness values in Band 3. The abscissa plots the spread of brightness values, specified by their Digital Numbers (see page 1-12) or DNs that, in a black and white Landsat TM image, range from 0 (black) to 255 (white) with numbers from 1 to 254 representing gray levels in progressively increasing lightness.

1-2: Analyze this histogram. What do its peaks represent? If the DNs shown along the abscissa were the ones used to produce a black and white image, what would it look like? ANSWER



The largest town in the image is Morro Bay, a resort community of about 10,000 permanent residents, some of whom are still actively involved in commercial fishing. The town is located about 21 km (13 miles) NW of the city of San Luis Obispo, along the conspicuous road evident in the scene (the famed coastal Highway 1 which wanders along much of the California coastline but in this scene extends inland to the east). The second major road, California Highway 41, is visible as it passes through a valley between hills enroute to Atascadero 27.5 km (17 miles) to the NE of Morro Bay. Cayucos is a small residential town along the coast just north of Morro Bay. Near the bottom of the image is another large settlement, the town of Los Osos (and its northern neighbor, Baywood Park), which is popular now as a retirement community, especially the areas along hillsides with commanding views of the ocean.

Assuming you didn't locate Morro Bay on the California State Map (first question above), this map version extracts the immediate area and should help to get your bearings relative to the above description.

Primary Author: Nicholas M. Short, Sr.

This page is a gallery of mostly ground photos taken by the writer (NMS) during two visits to Morro Bay, in 1994 and again in 1999. These provide a means for the viewer to gain a feel for what the towns and natural features in this area actually look like, so as to better correlate the TM subscene's contents with their counterparts as these appear on the surface. An aerial oblique photo covers much of the same area and should serve as the reference in comparing scene characteristics between the TM vertical view and the more familiar near surface manifestations.

Ground and Aerial Photographs of the Morro Bay Area

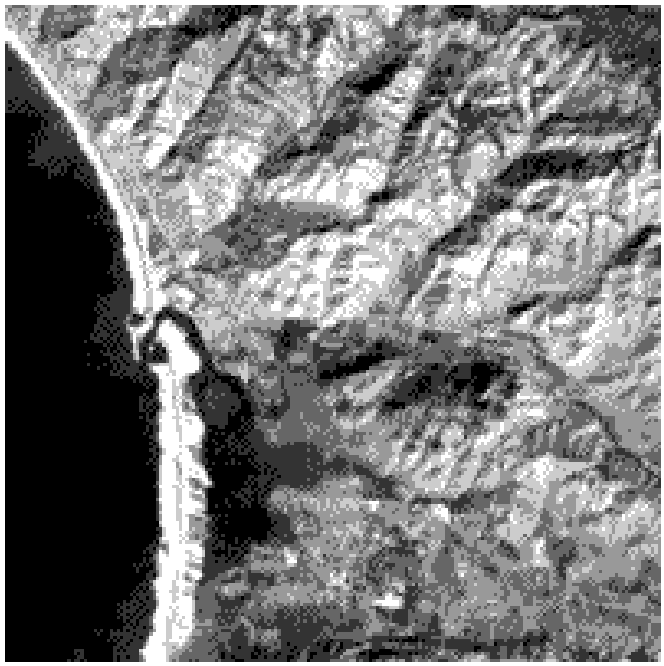
You can gain a feel for the terrain and the town layout by looking at the next group of views. The first is an aerial oblique photograph taken in mid-afternoon (note shadows) on January 25, 1988 (when, in winter, the grasses covering the hills and mountains are green from the rainy season) looking east at the northern part of Morro Bay towards the hills in the background. The town of Morro Bay is situated in the center of the photo. Morro Rock is near the bottom right.

(Credit: Golden State Aerial Surveys, San Luis Obispo, California)

Here is much the same area seen in a vertical aerial photo:

The next view was taken from a hill slope within the town of Los Osos (Spanish for "The Bears"), about 8 km (5 miles) south of Morro Bay town, on the ground looking north-northwest across the body of water named Morro Bay, with that town in the distance.

1-3: Glancing back and forth between these oblique photos and the Band 3 image repeated below, try to locate in the Landsat scene whatever ground features you note/identify in the aerial oblique photo. Also, see if you can find in the Landsat image the approximate place where the above land photo was taken. ANSWER



TM Band 3

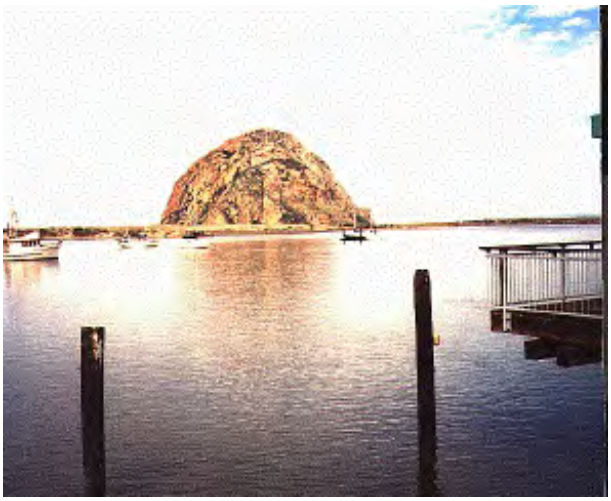
Lets look at some of these features more closely, as photographed by the writer during two visits, in 1994 and 1999. The town of Morro Bay is a popular tourist attraction streaming with visitors much of the year. The Main street area is typical of many smaller California towns located along the coast.

Its waterfront, called the Embarcadero, is lined with shops, restaurants, and boat moorings, as seen here:

Note the three smokestacks of the powerhouse, notable in the aerial oblique photo but hard to

pick out in the Landsat image, in the background. Behind this plant are five oil storage tanks, conspicuous in both the aerial photo and the Landsat imagery; as seen from the ground:

One surface feature stands out in both the aerial photo and the panoramic ground scene: Morro Rock, a great erosional monolith made of silicic volcanic rock (part of a chain known as the Seven Sisters; other volcanic necks [central vents] also appear in this scene) that reaches a height of 175 m. (574 ft) above the Pacific.



Technically, this is a sea stack, an outlier of rock that became detached from the mainland as waves cut away at the shoreline. Here, we see a narrow deposit of sand, caused by deposition in shallow water. This sand bridge connects the stack with the mainland, producing what is known as a "spit".

Just to its north is the public beach against which waves break in a pounding surf; this picture was taken as the ubiquitous summer fog was beginning to clear for the day.

Morro Bay itself is formed from a long barrier island tied to the south end and open as an inlet near Morro Rock. Exposed sand dunes occupy much of the surface but patches of low saltwater vegetation are scattered over it:

Extending into the bay near its southeast end is a delta formed by a small river. This delta supports estuarine and riverine vegetation as depicted in this east-looking view:

The southern side of the hills running eastward and perpendicular to the ocean, with several conspicuous volcanic peaks (other Seven Sisters), but vegetation-covered, appears in this view taken from a point SE of Morro Bay town.

On the north side of these hills, along Highway 1, the flat valley is given to agriculture. Seen in this photo are fields of the main crop in the area - snow peas:

More widespread farming takes place in the valley traversed by Highway 41. The brown area noted in the aerial oblique photo is largely made up of Avocado trees, as seen here:

Other cultivated acreage includes snow peas and grapes (wine-producing), evident in this photo which also shows the low grass-covered hills that make up much of the rolling topography inland from the coastline:

Denser forests extend from the higher elevations (about 260 m or 850 ft, in the Landsat subscene) into the lowlands along streams. The next view is a typical example.

The high coastal mountains to the east of the Morro Bay-Baywood-Los Osos developments make up the Santa Lucia Range, along which we find California Live Oaks and other trees and foliage that comprise the northernmost extension of Los Padres National Forest in south-central California.The hallmark of much of the hillsides in both the Coast Range and Sierra Nevada foothills is the natural grasses covering which reach their full greenery during the rainy season from October through April. By May, these grasslands assume a pleasing golden yellow-brown color that persists through Summer into early Fall. This brownish background contrasts with the deep greens of the oaks and evergreens exemplified in the next photograph.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

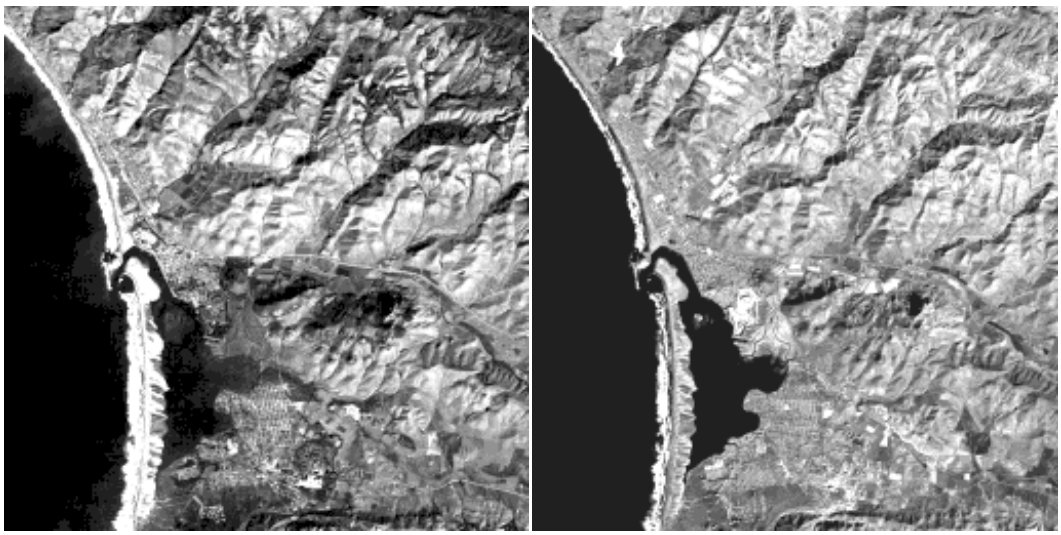
On this page are, side by side for easy comparison, all 7 TM bands that cover the Morro Bay subscene acquired in November, 1984. A DN histogram for each is also given. There is also a feature map, with letters in color (against a subdued Band 3 image as the base) referring to some of the main ground features referred to in Section 1.

Thematic Mapper Bands

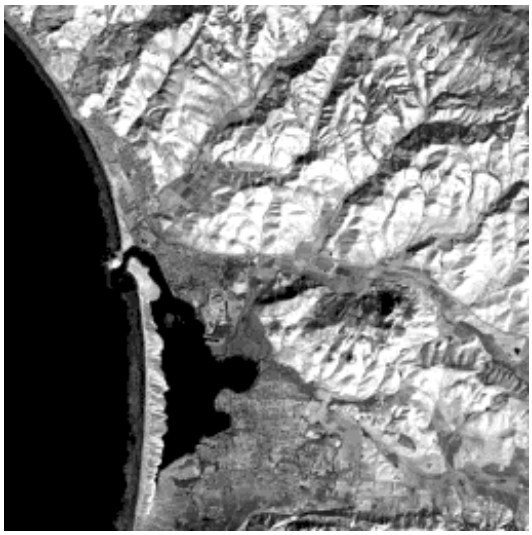
After that quick look at the scene setting, we are set to start our exercise in image interpretation. As this could be your first effort at this process, we suggest you now become familiar with this Landsat subscene as we rendered in all seven TM bands.

Morro Bay as Recorded In Different TM bands

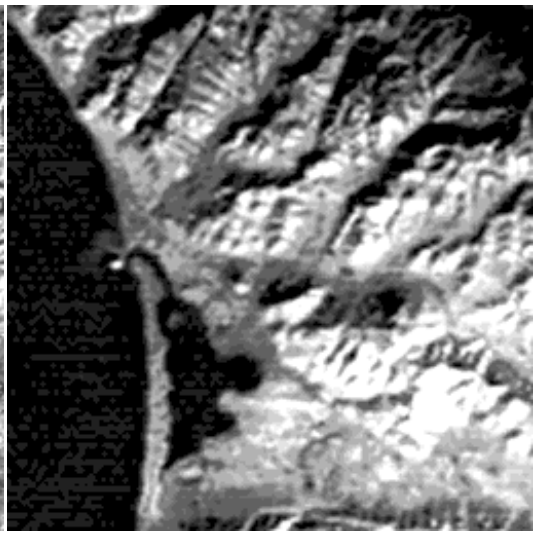
TM 1 **TM 2**



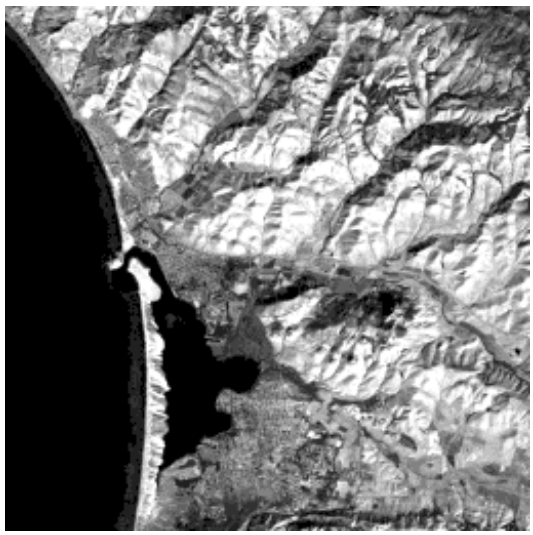
TM 3 **TM 4**



TM 5



TM 6



TM 7

Note that TM Band 6 appears blurred or fuzzy. This is the thermal band which has a spatial resolution of 120 m compared with the 30 m of the other bands.

Each image has its own tonal characteristics, along with distinctive shapes defined by tonal boundaries and patterns, that help to identify ground features. You can view the histograms for each band clicking on each of the following links: histogram 1, histogram 2, histogram 3, histogram 4, histogram 5, histogram 6, histogram 7. Note that each histogram is bimodal, having one peak at very low values (water) and the second at various mid-DN values.

Feature Identification Map

In the treatise that follows, to focus attention, we identify specific features and patterns by letters on an overlay of the background image in which we flattened the gray tones. Take a look at the map now. You can link to this map any time by clicking on the underlined and colored letter. (To remove it once displayed, click the X button at the upper right corner of the screen.) To make a convenient reference, we suggest you print this map and keep it handy if you do not wish to download the map each time it is referred to on later pages. The background image is made from TM 1 in which the gray levels were deliberately contracted to show only small contrasts.

1-4: As a preliminary to the more detailed feature descriptions that follow in this section, make a simple table with each of the seven TM bands running horizontally along rows and the following categories making up the vertical columns: silt in ocean

water; wave breakers; beach sand on a bar; the town of Los Osos; the marshy delta; sun-facing slopes (lower right surfaces in the hilly parts of the image); shadows. In this table' boxes (draw them or imagine them) put down some representative gray level ranging from very light tone; light gray; medium gray; dark gray; blackish (your choice will be somewhat subjective). ANSWER

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

This page introduces one approach to image interpretation: In general, a novice photointerpreter may not have the experience to look at an image and identify many of the features and landmarks that comprise this image. An expert often can recognize these features by their shapes or spatial patterns. The writer falls into that category from thousands of hours examining Landsat imagery and aerial photos. You probably do not. It helps very much to know the area personally or go visit it. This is what I did (twice now) for the Morro Bay scene. What you should do is to gain a general insight into the interpretive process by associating features/patterns that are labelled by letters on a reference image of the same scene with their appearance in this TM Band 1 image. Later in this section, color versions and special computer processing products will help you further in the interpretation. But, ultimately, you will need to develop your own skills in learning how to recognize/identify the information features in an image - one useful aid is to practice this identification procedure whenever you fly - ask for a window seat away from the wing and spend many moments looking out at the Earth's surface. The letters in blue (or purple if chosen earlier) refer to specific locales on the reference image which will automatically be called up any time you click on a letter.

Analysis of the Morro Bay Scene

Lets start our feature analysis with TM Band 1.

Since we will be referring often to band-related tonal gray levels, it would aid you in following the interpretation if you print out page I-19, which has a two-band guide to tonal patterns in MSS imagery (works just as well for TM), and refer to it when appropriate.

This and the other black and white individual band images shown in these first pages have all been contrast stretched, thus modifying the raw DN data. This is probably the most common and informative of the various image processing routines applied to digitized satellite images. Contrast stretching is discussed in detail on page 1-12a. As your option, you can preview that page now or wait until you have gained more familiarity with the Morro Bay images.

In Band 1, bluish and greenish objects and areas in the scene generally show in the Band 1 image as lighter gray tones than those representing other colors. In your first look at the image, you may notice the ocean on the left; hilly topography, especially in the upper right two-thirds of the scene; several areas with distinctive crisscross designs that are street patterns in the several towns; and one prominent linear feature and several similar features that are obviously a major highway and secondary roads. You may also recognize some other surface features, natural and manmade, that we will shortly explain.

The dominant water feature in the scene is the ocean. The tone for the open sea is surprisingly dark, inasmuch as we usually think of sea water as being bluish-green. In all bands, open water usually shows as low gray tones, since it absorbs most of the incident solar irradiation. Reflectance from water is greater in the blue and green bands but nevertheless is not high. But, water reflectance increases somewhat if silt and other sediments are present. In this image (Band 1), sediment is widespread, in the offshore water and in the bay (the sediment tones in Band 1are the lightest of any of the bands), but deviations within the tonal pattern such as at point **a**, suggest variations in sediment load. Broad changes in these patterns, from the coastline to offshore locations, represent tongues of sediment (barely discernible in the aerial oblique photo) being distributed by wave action. This displays better in Bands 2 and 3, where these tongues stand out in sharper contrast. Note the variable distribution of sediment within the more sheltered Morro Bay, especially at **b**, which is just beyond the entry point of a small river.

Right at the coastline, two very similar, but distinct, long strips of very bright (whitish) tones extend back to back from nearly the top to the bottom of the subscene. As you examine spectral bands at longer wavelengths, the strip on the left (to the west) will change character (becoming a mottled darker gray) whereas the one on the right stays about the same. If you are familiar with the West Coast shorelines, you probably surmised that the left strip corresponds to strongly breaking waves. These breakers are whitish to the eye when viewed from the ground, owing to the ability of water droplets to strongly scatter and reflect light, particularly at shorter wavelengths, because the water surface is no longer smooth when it churns into a foam. You may now suspect that the right strip relates to highly reflective sand that has deposited as a sandy beach bar. This coastal feature commonly results from longshore currents that cut across shallow water (here, Morro Bay) and deposit their sand load as their velocity is decreased. Two long jetties (see aerial oblique photo) protect the inlet from these currents. Note that these narrow structures are not visible in the image. The darker patches (as at **c**) that interrupt this bar are coastal vegetation areas that tend to have low reflectances in all bands, even for Band 4 which usually renders many other types of vegetation as much brighter.

Hills and low mountains in the scene appear as a distinctive light-dark pattern that gives the impression of an irregular surface, one showing relief. Map makers create this effect by using gray-tone shading that tricks the mind into perceiving the impression of undulating topography. In this image, as in aerial photographs, the same effect is produced by variations in surface illumination. Mountain slopes (as at **d**) whose faces incline toward the Sun, i.e., have more surface area facing the solar rays, reflect a high percentage of light towards the spacecraft's sensor (or camera on an airplane). These surfaces appear brighter (lighter tones). Slopes (**e**) that face away from the Sun, i.e., have less surface area facing the incoming radiation, reflect most of that light away from the sensor, and thus, appear variably darker, that is, they are more or less in shadow. These slopes generally appear in medium gray patterns, within which, other patterns, representing specific features or objects, may be discernible because of tonal contrasts (look at **f**). If the slopes are steep, and/or the Sun angle is low, the slopes tend to appear progressively darker (approaching black tones in the imagery).

1-5: What do you think would be the effect on this subscene of having been imaged at a different time of the day; at a different time of the year? ANSWER

In this subscene, most slopes in this band and the others are very light toned. This implies a scarcity of growing (green) vegetation and thus, a predominance of bare rock surfaces and soils. In this Fall image, the hills are in their late stage brown phase and the low chlorophyll levels

reduce the reflectance from vegetation. At higher elevations vegetation is greener and denser (as at **g**), because more moisture is available, due to condensation of fog and heavier precipitation owing to local topography. This denser vegetation appears in darker tones in Band 1 and Band 3 because chlorophyll absorbs blue and red wavelengths. In Band 2 the green vegetation is only slightly darker than the dry grass in adjacent areas because chlorophyll reflects green most strongly in that part of the visible spectrum. As was evident in the oblique air photo and the ground-based photo, granite bedrock is exposed in places within the scene. The best example (at **h**), is Morro Rock, whose light-toned volcanic rock is displayed in this image as a bright (Sun-facing surface) spot just below a dark spot (its shadow).

Two patterns correlate with actively-growing vegetation (which appear as somewhat darker tones in the first three TM bands [especially 1 and 3]). At (**i**), long thin lines of dark tones correspond to vegetation, mostly trees and bushes, that cluster along the banks of narrow, probably intermittent, streams in the hills. Similar tones (as at **j**) in more irregular patterns are usually associated with copses and irregular stands of trees, some of which are noticeable in the aerial oblique photograph. At (**k**) and elsewhere, primarily in the valleys and lowlands, rectangular to somewhat irregular, elongated dark-toned patterns indicate agricultural field plots containing actively growing crops (inspect the aerial oblique photograph again). Other irregular patterns in these settings (as at **l**), may be fields of brush or scrub (see the foreground in this photo) or even wetlands. Still other darker-toned areas, as at (**m**) and (**n**), that spread over wider expanses either are open uncultivated, or now barren, fields or are lowlands sporadically covered with natural vegetation.

The large, roughly triangular area on which (**o**) is centered lies at the end of the small river (**p**) that winds through the valley in which Highway 1 passes. At least part of this feature is a delta that has built up into Morro Bay. The live vegetation it supports, rendered in moderately dark tones in this band, is characteristic of wetlands that develop into saltwater marshlands along a coast.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Identification of select features continues on this page. The appearance of typical residential areas in Morro Bay is presented in a photograph, along with a view of the power plant which can just be made out in the imagery. A Band 4 TM image is shown, within which you should try to find the same features introduced to you in Band 1. Of special note is a mystery feature (in an enlarged subset) characterized in Band 4 by a series of long, connected light toned patterns. Depending on your interests and outdoor activities, you may recognize this feature immediately - or it for now may remain a complete "puzzlement". If it is still unknown to you, keep trying to determine what it is before we reveal its nature in a later page.

We can easily identify the towns in this subscene because of their prominent rectangular street patterns (towns labeled in yellow). The light tones emanating from the streets indicate that most are concrete rather than asphalt. Typical residences in downtown Morro Bay appear below.

At the north end of the bay is a major electric power-generating plant, as seen here.

Taking a close look in the image (at **t**), we see a row of two light dots and then three more to the north. A glance at the aerial oblique photo quickly reveals their identity. The stacks are not clearly resolvable in the image but the large central power plant just behind them does show as a dark-toned rectangle (it has a flat-topped roof coated with a material such as asphalt).

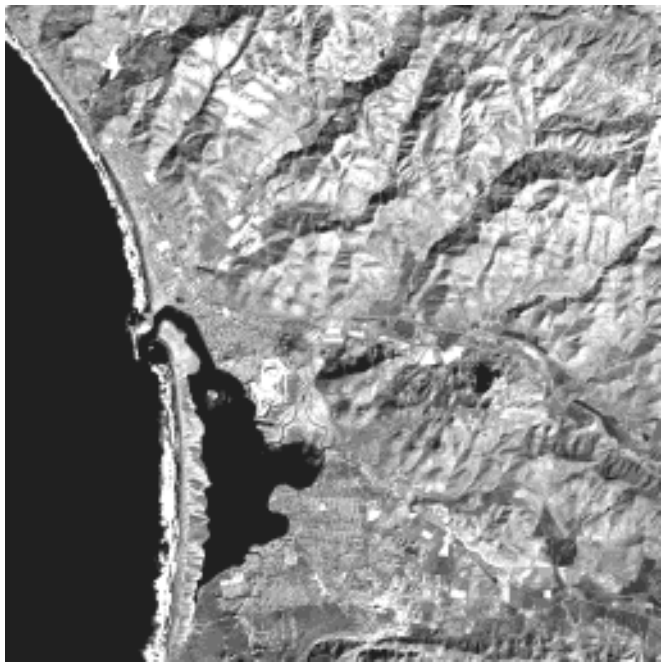
Several small, very bright-toned patches appear in Los Osos (as at **u**) and elsewhere. Some of these are borrow pits where sand and gravel deposited during earlier ocean floods is extracted

for construction and other purposes; others are outcrops of whitish sandstone.

1-6: Look for the five oil storage tanks in each of the seven TM images. Can you find them. Explain. ANSWER

At this point we will end our first look at the Morro Bay subscene. We will now hunt for the expression of the labeled features described above in other bands and call attention to several additional features that stand out better in one or more of those bands. Rather than proceed sequentially to the next band, TM 2, we investigate the subscene presented in Band 4, which is optimal for highlighting actively growing vegetation as bright tones.

Here is TM Band 4:

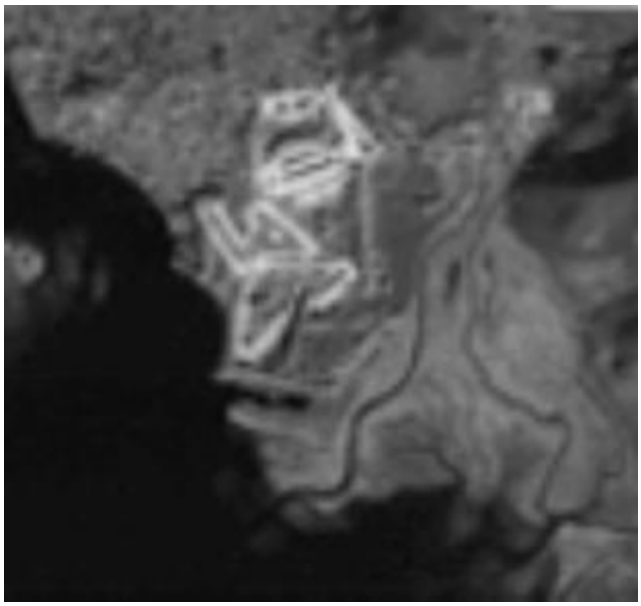


TM Band 4

Your first reaction may be that much of the land now displays in brighter tones and the ocean is almost uniformly dark (except where heavy silt still shows at the mouth of the river crossing the wetlands delta). Of course, much of this brightness comes from vegetation. Some of the dark fields (**k**) are now quite bright. Similar rectangles in Morro Bay and Los Osos are city parks and meadows.

Mystery Feature

There is a mysterious area at **s** that we challenge you to identify as we work through this Section. It's very bright in Band 4 but not in the overlay image. Its identity may be more obvious in the enlarged version below.



1-7: Take your first guess now. Clues: the patterns and tonal contrast. ANSWER

Band Information Characteristics

Looking again at the Band 4 image, the coastal marsh (**o**) is bright also, except near the water where it darkens (more mud, less vegetation). Another area within the mountains (at **v**), which is somewhat darker in Band 1 even though it's on a slope facing the sun, is uniformly bright. As a hypothesis, it may be a zone in which the grass cover is thicker and more active (perhaps new growth of grass from the first Fall rains).

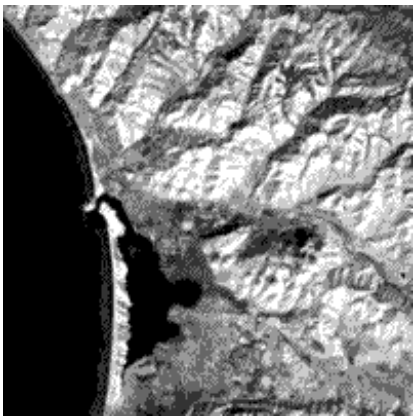
The shadowy sides of the mountains are, in most places, less dark than in Band 1 and reveal more detailed structure representing features on these slopes. This lightening is due to the presence of vegetation. However, very dark shadows do persist on the north slopes of several hills east of Morro Bay town.

The beach, quite bright in Band 1, is now darker in the sandy areas to a medium tone. To its west, the breakers that are so bright in 1 now appear in a mottled, streaky pattern that probably correlates with the different parts of the waves. Some parts face towards the sun (more reflectance into the TM sensor) and other parts of these transient curved surfaces face away (and appear darker).

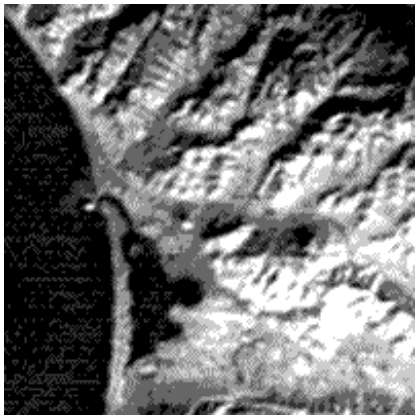
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

The two longer wavelength Landsat TM bands are imaged on this page. Band 7 depicts reflected infrared radiation beyond the narrow spectral interval where vegetation is highly reflective. For some materials, the amount of reflected radiation (and hence its tonal representation) is quite distinctive in these intervals. Band 6 produces an emitted radiation image in the Thermal IR; in general, light tones mean higher temperatures (surface has become warmer owing usually to greater radiative absorption) and darker tones denote a cooler surface. During the daytime, water is almost always cooler than most land surfaces.



Now look at TM Band 7 which is centered near 2.2 µm in the reflected IR band. The image has some similarities to bands 1 and 4 but there are several obvious differences. All water is now completely dark (no signs of reflecting sediments). Only faint traces of the wave breakers remain. The beach sand is somewhat brighter than in Band 4. The sun-facing slopes in the hills seem to be moderately brighter than in the other two bands. A bit more detail emerges in the partially shadowed slopes. Some areas, such as crop fields, that were quite bright in 4, are again dark as in 1. The mystery area near **s** shows some internal pattern or structure, including elongate strips of lighter and darker, that provides another clue to its identity. The towns have a more diffuse street pattern and are generally darker overall. The extraction pits so prominent in Band 1 are almost impossible to pick out in Band 7.



Next, we will examine Band 6 - the emitted thermal band (10.5 - 12.5 µm). Remember that the spatial resolution for this band is 4 times coarser - 120 m compared with 30 m - than the other bands. The effect of this reduced resolution is immediately evident in viewing the subscene in that many details discernible in the other bands tend to be "smeared" out in this one. Yet the main elements you have come to know from bands already scanned are still readily recognized.

In fact, the hilly terrain is even more like a 3-D rendition because of the sharp contrast between the hotter sun-facing slopes and the shadowed slopes that are cooler and appear now as fairly uniform, dark, shadow-patterns. This contrast gives a superior sense of topographic relief to this part of the scene and also highlights the valley regions.

Some medium gray tones, as at **v**, represent the cooling effects of grass ground cover. An extremely bright, but isolated, spot at **h** coincides with the sun-facing side of Morro Rock which, having a bare rock surface with steep slopes, absorbs solar radiation and heats up so that it emits more thermal energy than other parts of this sea stack. The towns, labeled in yellow, have moderate gray levels, indicating they are somewhat cooler than the natural hillsides, but there are several lighter tones within the downtown areas which are customarily warmer than residential areas (that have lawns and more trees).

1-8 Speculate on why the upper right corner area of TM 6 is notably darker. ANSWER

Water, which has a notably higher heat capacity (amount of heat needed to raise its internal temperature by 1° C) than most dry materials on land, tends to appear cooler in daytime thermal IR images than do most land surfaces despite both land and water receiving about the same direct sunlight. (Section 9 deals with thermal imagery and reviews the principles governing the tonal displays of features at different radiant temperatures.) If this image had been taken at night, the water would be expressed in much lighter tones than the neighboring land, because it retained much of its internal heat. The land, on the other hand, would appear darker, because it radiated its absorbed solar energy, thereby cooling to radiant temperatures lower than the water.

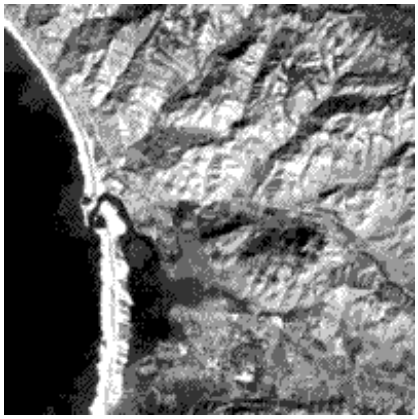
There is one pattern in Band 6 that is anomalous but meaningful (but may not express well in the relevant Tutorial image). Namely, in the water just north and west of Morro Rock (just above the bright dot) is an irregular pattern with a slightly lighter tone that jets out like a plume of sediment entering the ocean from the shore. This is precisely what we would expect from the nearby power plant (at **t**) which dumps its heated waste water into a nearby body of standing water.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

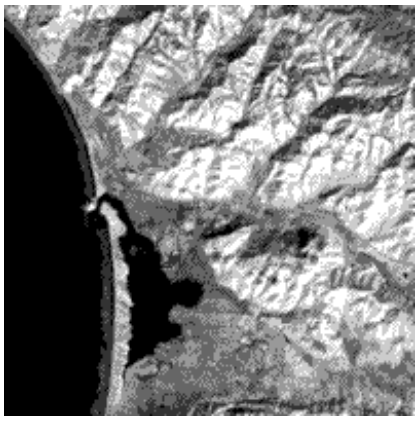
On page 1-5 Bands 1 and 4 were compared, with notable tonal differences between them. On this page, Bands 2 and 3, which are close together in the spectral intervals sensed by the TM, are shown side by side. At first glance, they appear almost the same in tonal balance and intensity. But small differences can be discerned, which become important when color composites are made and/or special computer processing (described later in this Section) is applied. In Band 5, is shown for the first time, you should note any differences in its tones compared to Bands 2 and 3 and to Band 7. The page ends with an "aside", namely, a brief discussion of what has been called "linears" (these will be discussed further in Section 2).

Returning for a moment to the two other visible bands, Landsat TM 2 and 3, which monitor reflected light broadly distributed in the green and red bands, respectively. First look at Band 2 (left) and then Band 3



(right) and, if you wish, compare each of them to Bands 1 and 4. On first glance, there seems to be little difference in tonal variations for features imaged in Bands 1 through 3. The most perceptible change is the extent to which the ocean sediment tonal patterns diminish in going from Band 2 to 3. The sun-facing slopes in the hill country seem a bit brighter in 3 than in the other two bands, perhaps influenced by the red component in light reflected from a surface underlain by brown soils and also by the brown color of the grass.

Because of the apparent similarities in tone for major features seen in the first 3 bands, we might wonder how they combine so well into a color composite, which we will inspect shortly. The answer to this question is subtle and rather complex but here is the essential reason: when images for the individual bands are generated, they have appropriate contrasts. The images have similar tonal variations for the ground classes, because most of the features have approximately the same relative differences in reflectance in each band. But, their actual reflectance values (partially embodied in their DN's or digital numbers [see page I-5a for a brief synopsis of what DNs are]) vary more from one band to the next. Thus, a green reflector may have a higher average DN value in Band 2 than in 1 or 3 and that will contribute more effective brightness to the green component of a color composite.



Now, we will finish our look at the black and white renditions of the individual bands by displaying TM Band 5. This band, centered around 1.65 µm, most closely resembles Band 7 but by flipping back and forth into 7 and 5 you should be able to pinpoint several moderate differences. In the beach bar (**c**), Band 5 shows a moderately darker tonal variation, with some structure in the patterns within the sand other than that caused by vegetation. As with Band 7, the breakers have lost most of their reflectance output and appear therefore as mottled medium-dark grays. The mysterious feature (**s**) has good internal contrast highlighting the distinctive patterns that give away its identity (to the imaginative). The urban areas' structure, defined by the crisscrossing street patterns, (within towns labeled in yellow on the overlay map), is rather washed out in the Band 5 version. The two rows of bright dots representing storage tanks near the power plant (**t**) almost disappear in 5 (and 7).

1-9: There is something in Bands 5 and 7 that is almost absent from view by its lack of contrast; but, this feature stands out in Bands 2 and 3. What is it? (Hint: it's in the lower half of the image.) Which band seems lighter overall, 5 or 7? ANSWER

Before leaving this seven-band review, we comment on a peripheral item but one that illustrates a typical danger and misuse of space imagery. Look at any of the images (particularly Bands 5 and 7) at two points marked **w**. You should see two thin dark features that are almost straight. These "linears," as some call them, are common phenomena, observed in satellite images and in aerial- or astronaut-camera photographs. Peruse the rest of the image you chose and you should find more linears (although some are actually sharp boundaries between two features or classes). In the early days of Landsat applications, many geologists reported that the MSS (and, later, TM) sensor was especially adept at highlighting such linear features. They chose to identify a significant fraction of these features as geological in nature, such faults or fracture systems. Maps showing numerous linears of presumed faults or fractures were produced and often published but too frequently without appropriate field checking. When several exacting studies discredited this interpretation of many such features (although some, and sometimes a majority, were verified), this use of Landsat led to widespread skepticism and negative criticism. Today, we know to be careful, and properly use field inspectionsto verify our work. Incidentally, we haven't established the nature of the two linears at **w** but one is probably vegetation lining a narrow gully. The fact that the two line up may be a coincidence and does not prove any commonality.

1-10: Make a list of other features besides faults and fractures that have linear expressions. ANSWER

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Now, we will see the first example of a color composite, made by combining (either photographically or with a computer-processing program) any three registered (superimposed) bands of images projected through some choice of color filters, usually blue, green, and red. The single image on this page is that of the customary false color composite made by projecting a green band image through a blue filter, a red band through green, and the photographic infrared image (in MSS, Bands 6 or 7; in TM, Band 4) through a red filter. Several characteristic color signatures are discussed.

False Color Rendition

We are now in a position to apply image processing to generating our first color composite combination. Review page I-13, if need be, to recall the discussion of how color composites are made.

In practice, to make a color composite one needs to project on to color film an illuminated black and white image of each of three different spectral bands through a set of color filter or if computer-generated using color beam produced by three electron guns on to a monitor*. Let's start by producing photographically the now-conventional standard false color composite, made by assigning the TM Band 2 (green) image to the blue filter , Band 3 (red) to the green filter, and Band 4 (the near or photographic reflective IR) to the red filter.

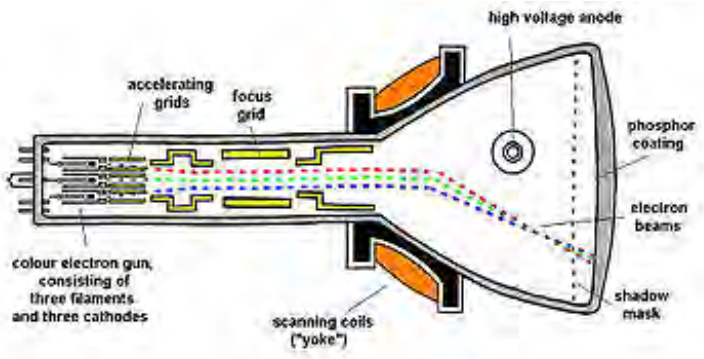
Two color patterns dominate the land classes: reds, depicting vegetation, and medium grayish-browns, found mainly along the bright sun-facing slopes. The ocean and the bay are evinced in deep blues that, near the shore (**a**), become a bit lighter where thicker sediments add reflectance. The breakers are presented in mottled blue and white patterns.

We can place the various expressions of vegetation in several categories based on their specific red tints and in most cases also from the spatial patterns they occupy. The continuous and rather deep red at **g** represents a segment of the forested areas in the Los Padres National Forest near the crest of the Santa Lucia Mountains. Elsewhere, as at **i**, we can attribute thin strings of red or irregular red patches (**j**) to trees and/or scrub vegetation (as at **l**), lining stream channels or scattered as copses and patches along the slopes. In the valleys, bright red areas (at **k** and other points), some rectangular and others more uneven, are primarily examples of field crops, hay meadows, or other types of cultivation. Areas believed to be barren to varying degrees (as at **m** and **o**), have darker gray-brown tones, but may have faint pink overtones implying limited vegetation cover. Where vegetation is sparse and scattered on the hills, particularly where well-illuminated by the sun, the prevailing tan to grayish brown colors imply joint contributions from underlying soils, combined with reflectances from the brown grasses (with much diminished Band 4 input). However, we temper this last statement by the fact that the image produced for this same scene by EOSAT (not reproduced here but examined by the writer) shows more pinks over most of the grassland areas than does this image. At the time of this scene acquisition (mid-November), if winter rains had started early, and had been only moderate to this date, the hills would still be relatively brownish. However, the EOSAT rendition suggests some greening had started.

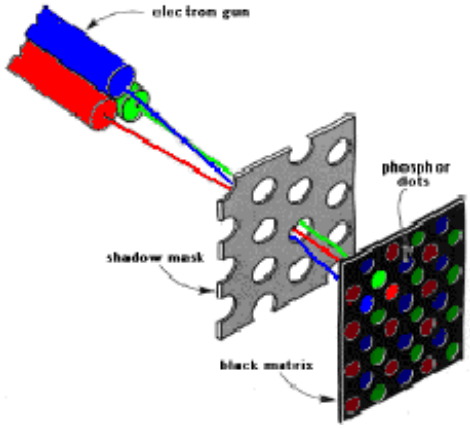
1-11: In the above false color composite, consider a different combination of colors and bands. Make Band 4 = blue; Band 3 = red; Band 2 = green. Describe the general appearance of this new false color composite. ANSWER

Once again, identify the urban areas from the street patterns. The streets, as well has Highway 1 and other major roadways, have bluish tones, expected because they are especially light-toned in Bands 1 and 2 (assigned here to blue). The mixed tones in Los Osos, with some yellow-brown scattered about, give it a subtly different color signature when compared with Morro Bay. We'll explain more about this difference when we examine the next color composite. Both town areas contain segments that have red blotches, which correspond to residential sections, park land, or other places where trees or grass grow. The extraction pits (**u**) are very bright, tinged with blue.

***** A brief discussion of how Color TVs (and with variations Color Monitors such as your computer screen) is appended here. A color vidicon produces an encoded signal which combines primary colors into a video version that can be transmitted. The signal upon reception at the receiver passes through a tuner and then electronic circuitry that separates audio from visual. The visual signal then goes through a decoder to extract three primary color signals that are each assigned to an electron gun that produces a color beam. Now, look at this figure:



As they enter the picture tube, these three electron beams are passed from the guns through an accelerator, than a focusing device, and then through a scanner which moves the beams through parallel lines (a high resolution picture tube displays in succession 700 horizontal lines or more on its screen face) that sweeps very rapidly across the picture tube's front surface. This diagram shows that surface assemblage:



The shadow masks serves as a collimator. In one arrangement color phosphors of different compositions, each sensitive to one of the primary colors, are activated by whichever electron beam yields a response to that color. If one of a color phosphor triad (RGB) is so excited, it will glow in that color as seen by an external viewer (such as you). Combinations of 2 or 3 beam color/phosphor responses, in varying intensitys, produce other colors than the primaries. In another arrangement, the phosphors are thin linear bands, red, green, and blue in parallel, that yield the same result. The phosphors are individually so small (a full screen can have more than 300000 dots), the eye does not see them as discrete, only the color they impart to individual pixels.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

This page is devoted to a closer look at the mystery feature, now shown in glorious living technicolor.

Areas that may support more actively growing grass (as at **v**), have a definite red-orange tone. The vegetation within the coastal marsh (**o**) takes on a distinctive orange undertone, implying that some of the plants have a reddish coloration (remember, orange can result from a higher input of green-filtered light [assigned to the red band in making the composite]). The mystery area **s** next to the marsh, and shown enlarged below, has a bright red expression alluding to the presence of highly IR-reflective vegetation such as might be found where grasses are luxurious (if you haven't guessed this feature's identity yet, this might be a decisive clue).

We show another ground photo which includes the mystery feature. Look at this:

In the foreground are homes in Los Osos that are on a curved strip of land that juts into the south end of Morro Bay. The marshlands can be seen in center right on the bay. To the left is a large cluster of trees that are the main part of Morro Bay State Park. The mystery feature is the thin green strip behind the trees. One of the rounded hills made of volcanic rocks like Morro Rock

itself is beyond and in a gap one can see many miles to the high tree-covered mountains of the Coast Ranges.

So, after all this preamble, just what is the mystery feature: **FORE** is the last hint. The answer will be on top of the next page

PREVIOUS NEXT

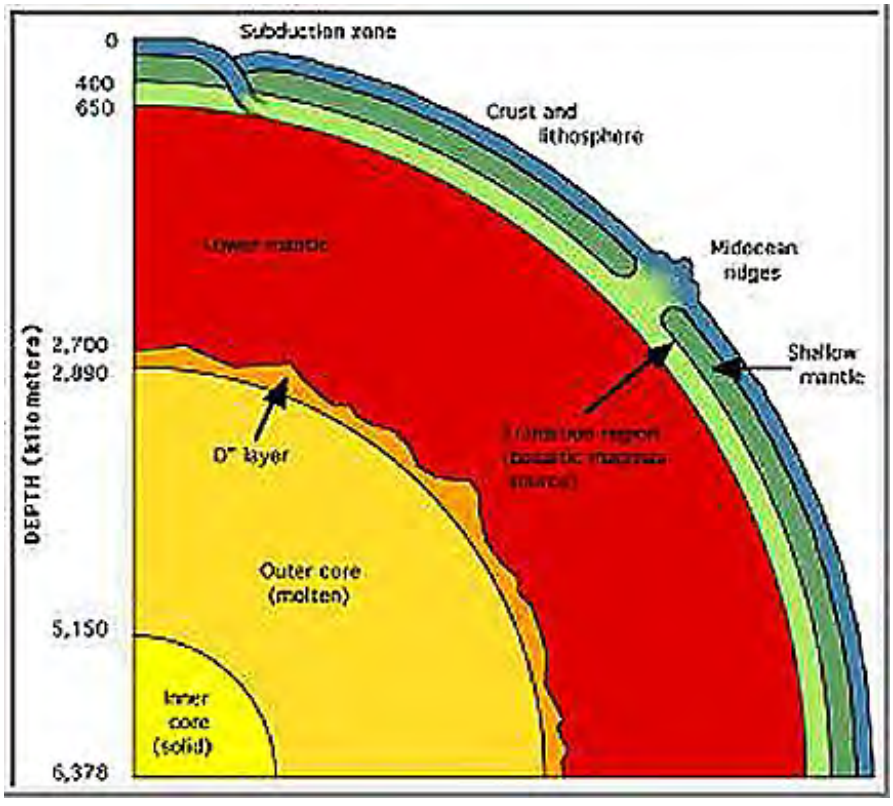
Primary Author: Nicholas M. Short, Sr.

SOME BASIC CONCEPTS UNDERLYING THE SCIENCE OF GEOLOGY

GEOLOGY - THE SCIENCE OF THE SOLID EARTH AND ITS INTERNAL PROCESSES, INFLUENCED ALSO BY THE ATMOSPHERE, OCEANS/RIVERS, AND BIOLOGICAL ACTIVITIES - COMES FROM THE GREEK WORDS *GEOS* (EARTH) AND *LOGOS* (STUDY). PLANETOLOGY INVOLVES ON-SITE AND SENSOR INVESTIGATIONS OF THE OTHER SOLAR (AND NOW EXTRASOLAR) PLANETARY BODIES, USING MAINLY COMPARISONS TO TERRESTRIAL GEOLOGICAL FEATURES AND PROCESSES AND APPLYING METHODS OF ANALYSIS PROVEN EFFECTIVE FOR EARTH STUDIES.

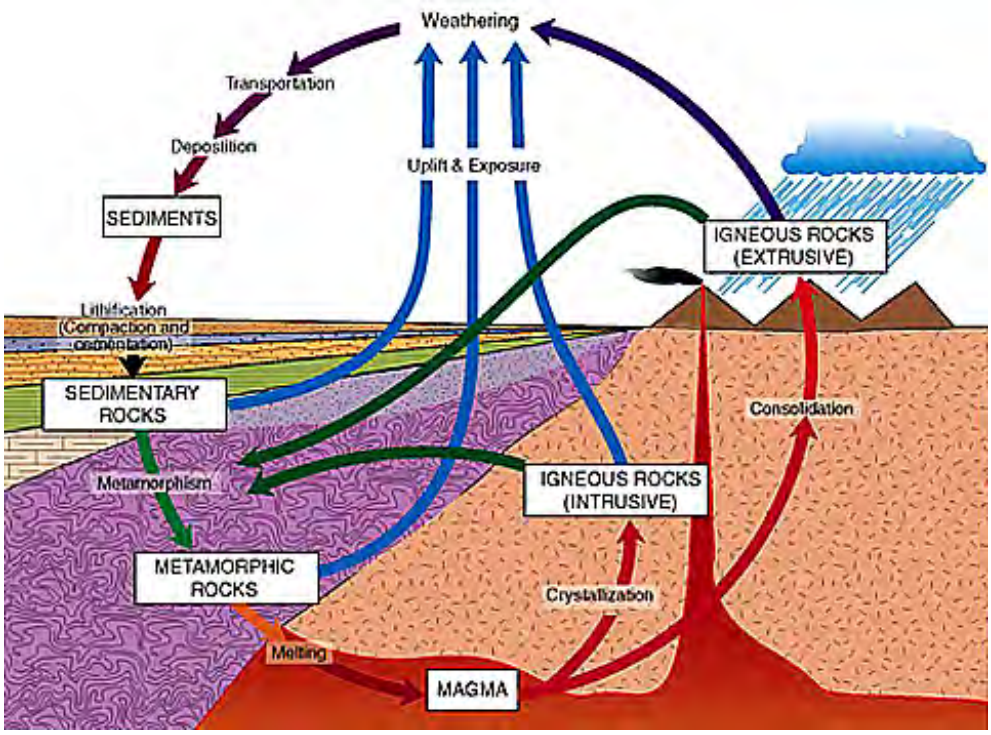
Six Fundamental Concepts about the Earth's Geology

1. The Earth formed about 4.6 billion years ago, along with the other solar planets and the Sun itself. The planets built up by accretion of rocky and gaseous debris (asteroidal, planetesimal [meteoritic] materials and comets) through collision of orbiting bodies. Aided by gravitational attraction which helped to compact these material, early on the assembling Earth underwent partial to complete melting, separation of different materials into an inner and outer core (Iron-Nickel), an extensive interior mantle, (Iron/Magnesium/Calcium-rich silicates), and a thin crust (enriched in Silica, Sodium/Potassium/Aluminum), all (except the outer core) solidifying by cooling over the first few hundred million years; escaping gases produced an atmosphere (principaly H, CO_2, N, CH_4) were held above the solid Earth by gravity owing to its large mass; in time (about 4 billion years ago), the Earth's exterior cooled sufficiently to allow vast volumes of water vapor to condense, forming in lower areas great concentrations of water collected into depressions (oceanic basins).



2. The Earth's materials are diverse and variable. Most variation occurs in the outermost 200 kilometers, in the lithosphere. **Igneous** rocks form directly by crystallization of hot melts made up

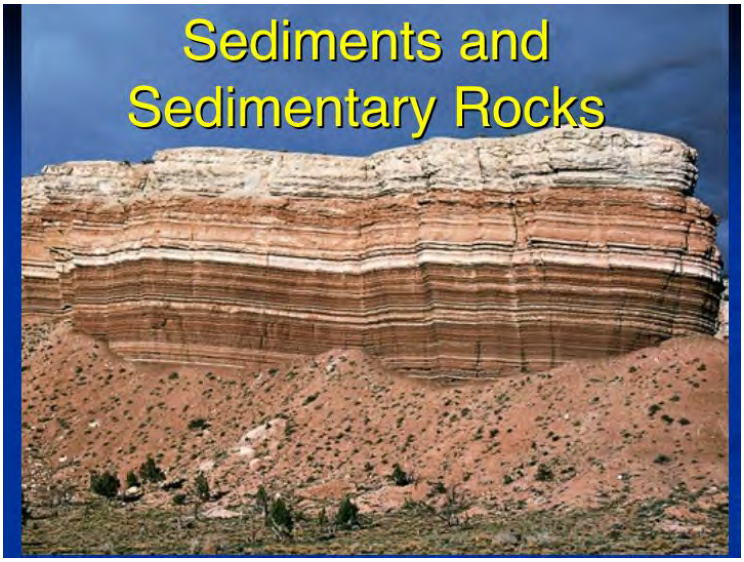
of silicates (Si_mO_n) combined with Fe, Mg, Ca, Al, Na, K, Ti, H_2O). Minerals formed from these make up nearly all the mantle and crust. Rocks at the surface decompose/disintegrate by reaction with the atmosphere/hydrosphere to produce solid debris and soluble chemicals that are transported/deposited to form sediments, that upon burial are converted to **Sedimentary** rocks (usually layered; strata). Previously formed rocks that are heated and pressurized when buried to shallow to moderate depths (5 to 70 km) of the crust recrystallize as solids under increased temperatures and pressures to form **Metamorphic rocks** (some may melt to become igneous again). The above processes comprise the *Rock Cycle*, shown below, and discussed in more detail near the bottom of this page.



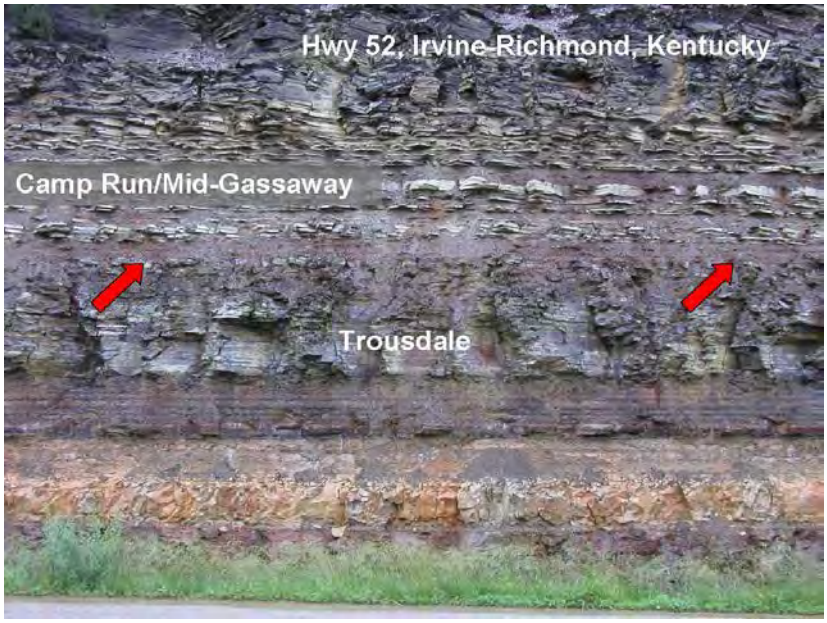
These rocks are usually distinctive in the field (out-of-doors). Igneous rocks are made from crystals - minerals that crystalized from once hot melts. Intrusive igneous rocks are commonly without notable layering - they can be described as massive, and result from cooling and solidification well beneath the Earth's surface. **Granite** is the most familiar intrusive rock type. Extrusive igneous rocks often occur in layers, which may vary in thickness, formed either by outflows of lava or accumulation of debris tossed into the air by erupting volcanoes. **Basalt** is the prevalent extrusive rock. Field examples of each are:

Sedimentary rocks are nearly always in layers. These are beds (also called *strata*) made up of deposits from the ocean or other water bodies that consist of clastic fragments (e.g., sand grains), or chemical or biochemical precipitates. Most frequently, the layers are initially laid down

as horizontal units. The most common types of sedimentary rocks are shales, mudstones, sandstones, limestones, and evaporites (such as salt beds). If the strata are inclined (dipping), this usually means that the rocks have subsequently been pushed by mountain-building forces that have tilted or folded the units. Here are examples:



Shales are the most common sedimentary rock type, being made up of fine particles (mud is composed of such small particles). Shales are usually characterized by their tendency to break along thin layers called parting. **Mudstones** lack this feature. Here is a typical outcrop of shale beds:



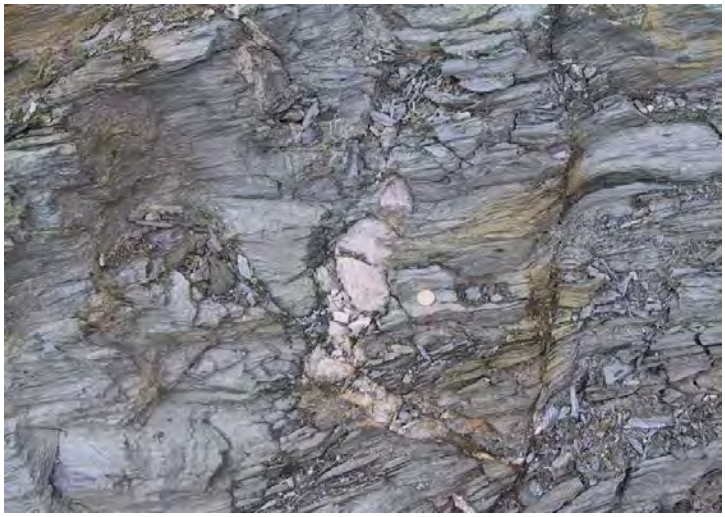
Sandstones are made up of sand-sized grains. Most common is quartz, the principal constituent of the thick sandstone beds seen below (graywackes are composed of sand fragments of eroded basaltic rocks; arkoses consist of fragments from granitic rocks)

Limestones are the principal member of carbonate rocks. Typically, the limestone is a light whitish-gray, as seen in this outcrop:

The bedding of sedimentary rocks is often conspicuous even when viewed from space, as shown by these limestone and sandstone units in Namibia.

Metamorphic rocks result from the action of heat and pressure on pre-existing rocks (usually sedimentary) that are brought to depths of a few to tens of kilometers below the surface. New minerals are produced by this metamorphism. Shales (mudstones) can be hardened into **slates**

and may be further recrystallized into mica-rich rocks called **schists**. As such rocks are heated to temperatures below but not far from those that would melt the rocks, they become soft, recrystallize further, and can be deformed into crenulated light and dark units that resemble layers, forming rocks called **gneisses**. Examples of slate, schist and gneiss are:

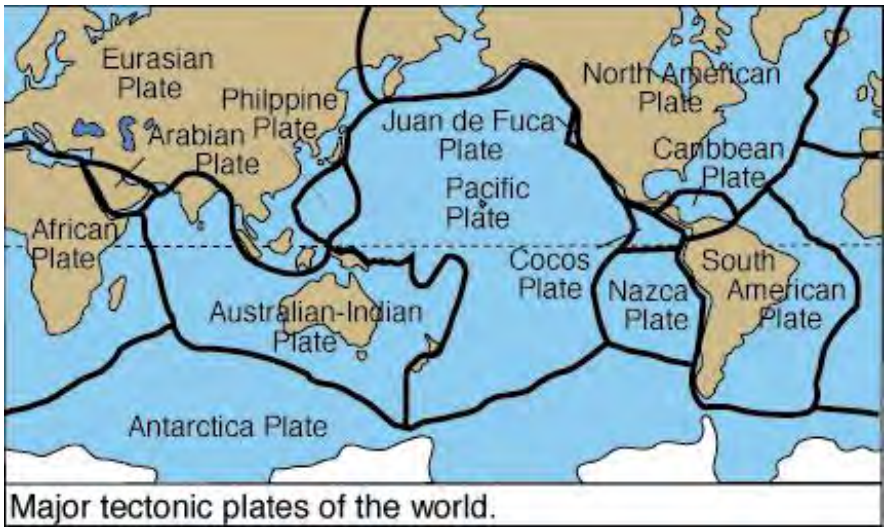


Metamorphosed sandstones are turned into **quartzites** which resemble their precursors but are harder.

Limestones when subjected to metamorphism become **marbles**

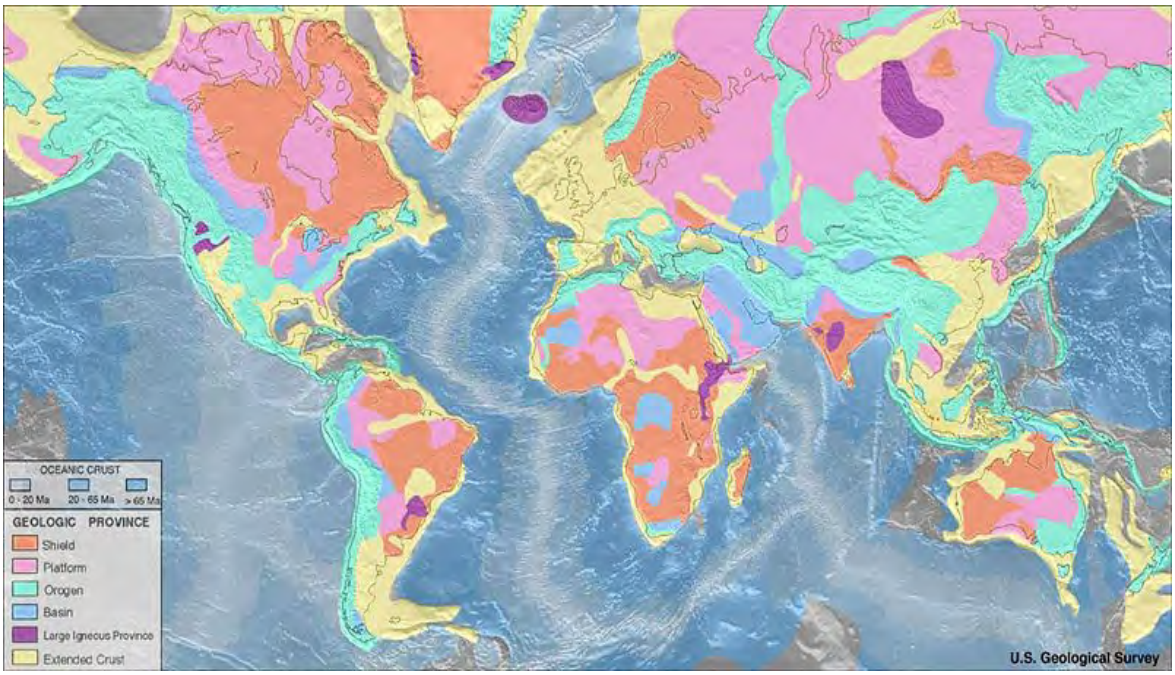
3. The Earth's outer shells (crust and upper mantle = lithosphere), about 150-200 kilometers thick under the continents (less so under the oceans), are subjected to dynamic forces that cause segments of the shells and materials at the top, to break up into plates and deposits on them that move laterally, bringing about deformation of their constituent rocks (mainly in and on

the crust) by bending, folding, flowing, fracturing, movement of blocks along faults, and melting. The branch of geology that studies these deformational effects is known as Structural Geology (a closely related term is "Tectonics", which refers to the processes that emplace and deform the structures assumed by the rocks involved). The dynamic processes, driven mainly by heat (much supplied by radioactive element decay) and gravity and resultant convection within and below the lithosphere (in the mantle), move plate units (segmented units of the Earth's spherical crust and upper mantle, that are in slow, essentially horizontal motion in some general direction) either away from each other or against each other (both situations can affect a plate); this general motion is called **plate tectonics**. Plates diverge from ridges rising from within oceanic basins (lower areas underlain by basaltic crust) and converge against boundaries of other plates (whose outer rocks are either oceanic or continental in nature and composition), causing melting, volcanism, metamorphism, mountain building, rise/fall of crustal blocks, continental growth and splitting.



Major tectonic plates of the world.

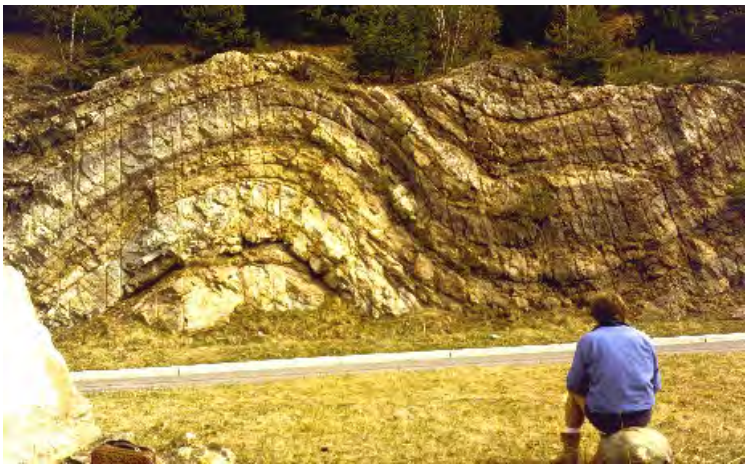
This next illustration has some of the inherent aspects of the plates within it. It shows however subdivisions of the continents in terms of structural settings that include 1) Shield (old igneous/metamorphic complexes; also called the craton); 2) Platform (supracrustal flat-lying sedimentary rocks); 3) Orogen (mountain belts; deformed rocks); 4) Basin (regions where sedimentary rocks have accumulated in geologically more recent times); 5) Large Igneous Province (areas of considerable cover by basaltic rocks); 6) Extended crust (parts of the continental crust now covered by marine waters).



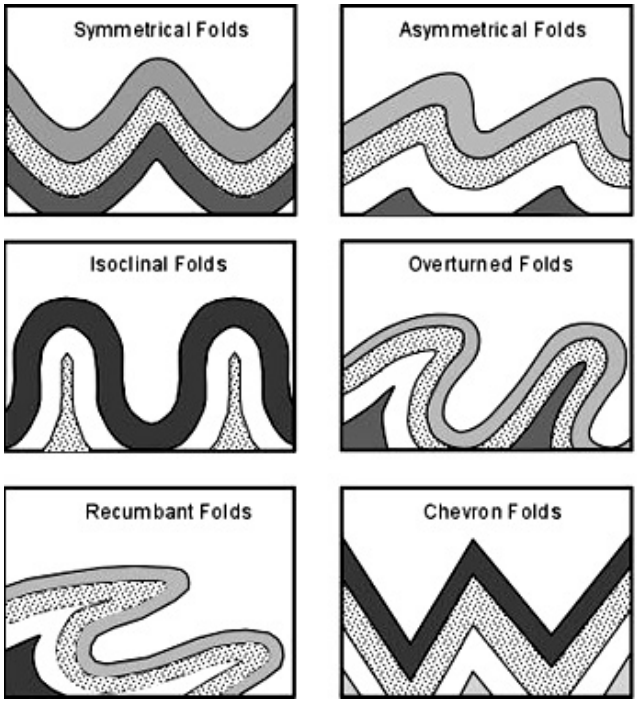
The previous two maps show large units of the Earth continental and oceanic crust. At much

smaller scales, rocks within the plates are subjected to pressures that cause them to bend or break. This ground photo shows folds and a fault (where rock is broken and displaced).

The two principal types of bends (folds) appear in this photo - the upward arch (left) is called an *anticline* and the downward bend is a *syncline*. Folds are the result of compressive stresses (the rocks are "pushed" by external forces causing them to buckle or "wrinkle").



The geometry of folds gives rise to different descriptive terms that relate to the attitude of the limbs (either side) of the fold with respect to the horizontal (the inclination of linear or tabular features such as layers or strata is referred to as "dip"). These are the terms.



Here is a ground photo of tight (isoclinal) folding. The upfold on the left is an *anticline*; on the right is a downfold or *syncline*:

This is a recumbent fold ("lying on its side").

Folds tend to die out at either end as one looks down on them. This effect is called "plunging" and is illustrated here:

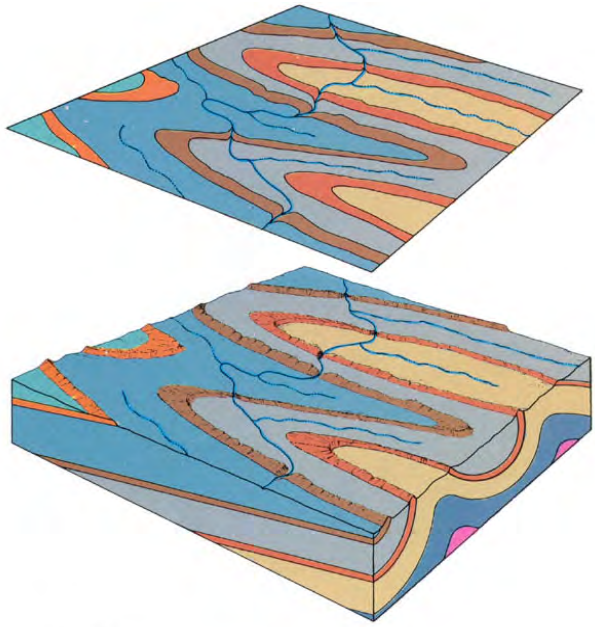
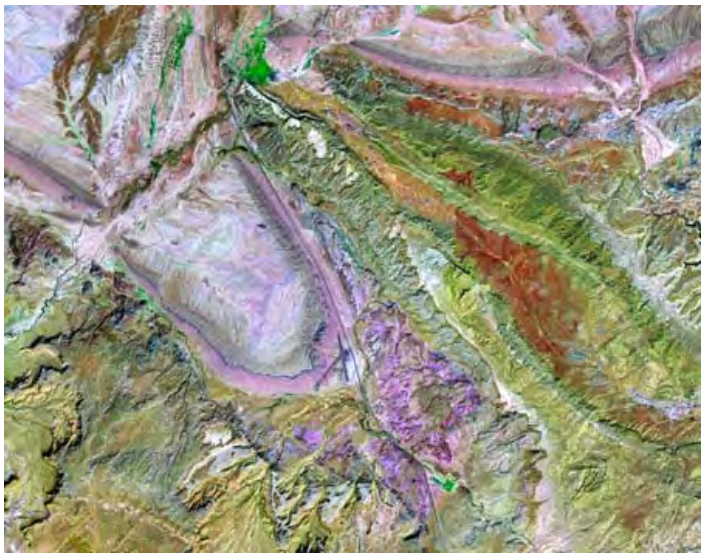


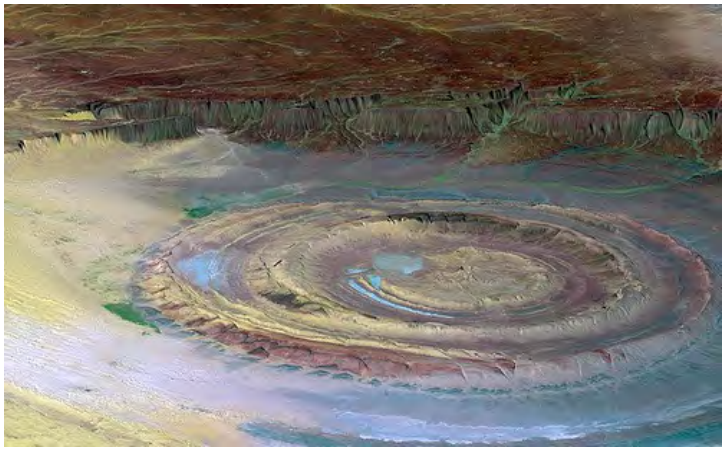
FIGURE 16.7
Outcrop Patterns of Plunging Folds

Note something else in this diagram. There is a general rule, the "Law of Superposition", that says younger layers of sedimentary rocks are laid down successively on top of older layers below. Note the sequence of strata in the diagram. In the anticline, the erosional surface (the horizontal plane) has produced a pattern in which the older rocks are interior to those on either side. The reverse is true for the syncline so that the younger rocks are interior. This hold for folds in general.

Here is an aerial view of a plunging anticline in Wyoming; below it is the surface of folded rocks with pronounced plunging as seen in a satellite image (the pattern can become complex):

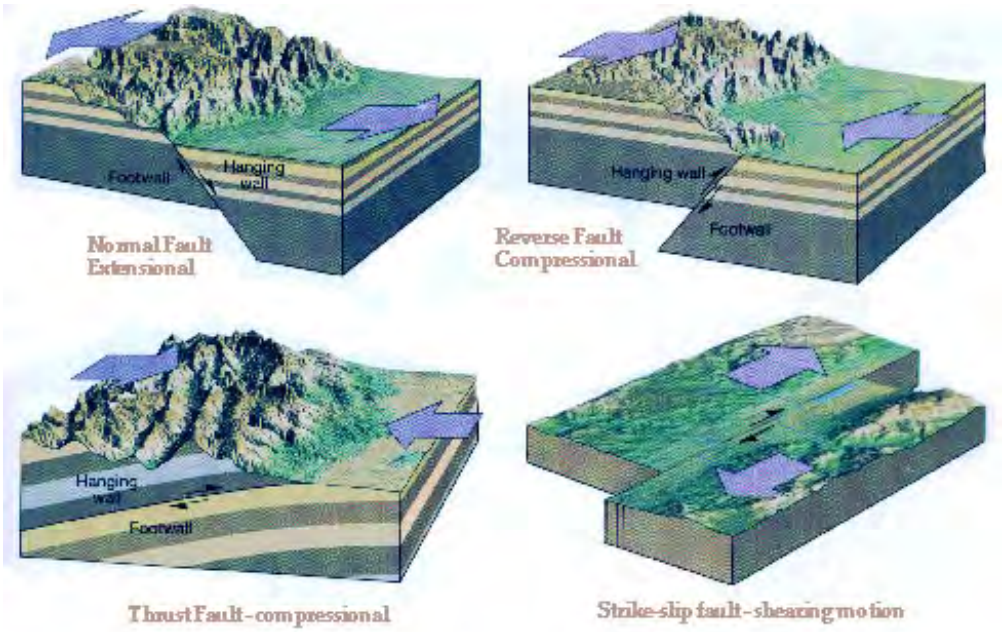


As seen from the air, a circular to elongate anticlinal fold comprises a *dome*; the older strata are exposed in the center:



At a small scale (outcrop-sized), contorted folds are found in metamorphic rocks such as gneisses. The rocks actually soften (see below) as they are heated during deep burial. This type of folding, called ptygmatic, is illustrated here:

Faults result from rocks that are stressed (usually resulting in folding) that then break with rocks on either side of the fault plane being displaced (shifted) so as to create a discontinuity. Faults are given names that indicate the mode of stress and geometric nature of the displacement. The extensional (when the rock units are subject to tension stresses) fault is commonly known as a *normal fault*, the compressional type is called a *reverse fault* if the fault plane is high angle and a *thrust fault* if low angle; the transform fault is one type of *wrench* or *rift faults* that is associated with oceanic ridges.



A small fault is seen in this ground outcrop. The fault - a plane - appears as a line marking a discontinuity in the once continuous sedimentary layers. Rocks on the left side are displaced upwards:

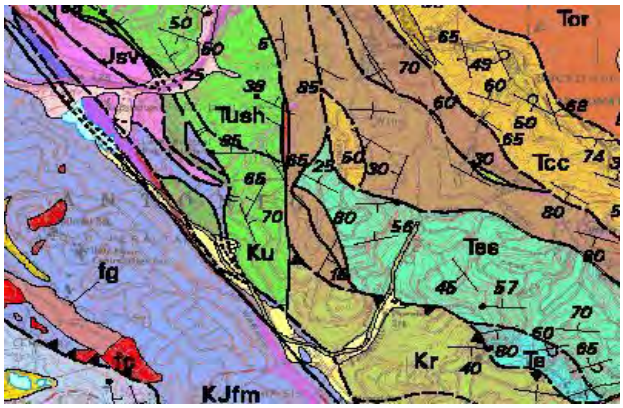
Both a normal fault and a reverse fault are exposed at this outcrop.

This fault can be recognized from the air as a line with dissimilar surface features on either side because the crustal blocks have shifted horizontally relative to each other; the movement is mainly horizontal making this a strike-slip fault.:

The most famous fault in North America is the San Andreas fault of California - another strike-slip fault. This reknown photo from the air shows offset of orange grove trees as the west side of the fault moved northward.

A thrust fault usually develops when an overturned fold breaks. Rocks are shoved up and outward such that older rocks are carried on top of younger (reversing the Law of Superposition), as shown here:

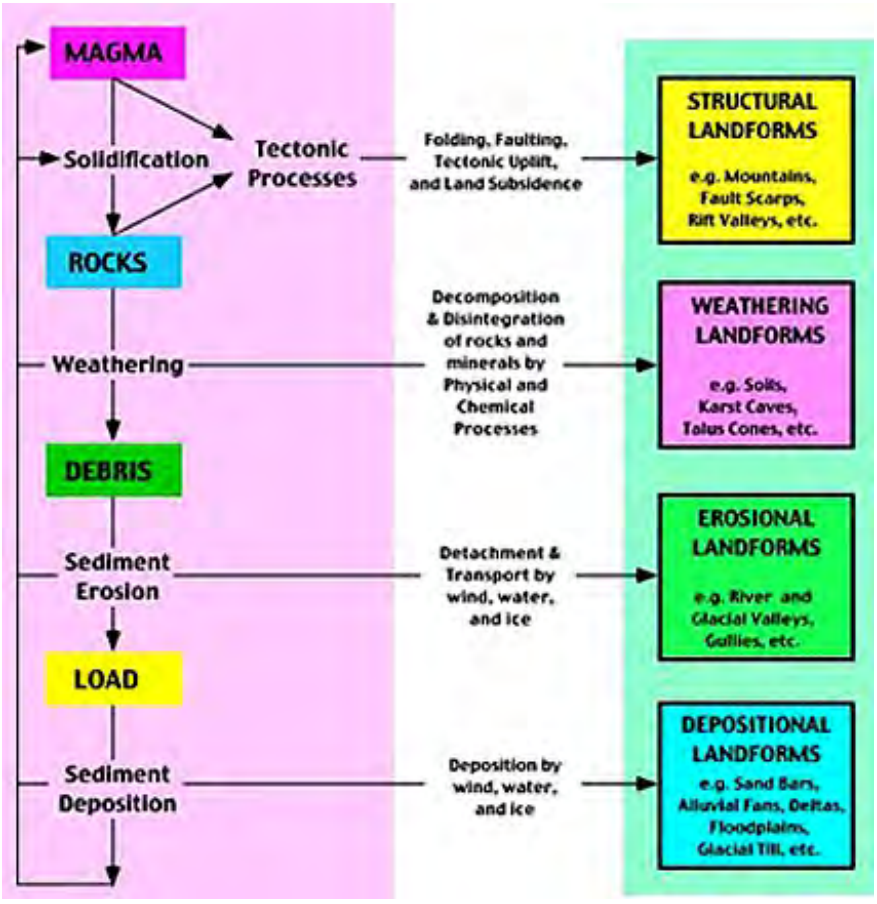
On a geologic map rock units are represented by different colors. In an area in which faulting has occurred, the fault trace at the surface is rendered as a black line; rock units on either side will show abrupt color mismatches:



Rocks subjected to stress also can break without any displacement. These planar breaks (in effect, fractures) are called *joints*. Here are three examples:

A question may have crossed the reader's mind, to which we will now try to respond. Rocks when held in the hand or examined in the field that are struck with a hammer usually break into chunks, that is, they are brittle. How then can great masses of rocks, particularly those that are layered, bend and fold without breaking into bits? Several factors make up an answer. These are: confining pressure, heat, time, slow rate of deformation. All but the surficial rocks involved in mountain-building deformation are buried. Any individual part of the rock assemblage (imagine a cube of material) is confined by its neighboring rock masses. The rocks are subjected to heat from the Earth's interior and other sources. They are pressed upon by external forces (see next page). The pressures are exerted over long periods of time. Under these conditions rocks behave as though they were "soft" or ductile rather than brittle. (Metamorphic rocks that contort erratically are actually more like taffy than like hard rock.) They tend to deform at microscopic levels along atomic slip planes and a macroscopic levels along bedding planes. Slow deformation that produces folds takes millions of years as the rocks are gradually displaced. But at various stages the deformation may exceed the strength of the rock so that it does not bend but fractures instead and undergoes displacements along faults.

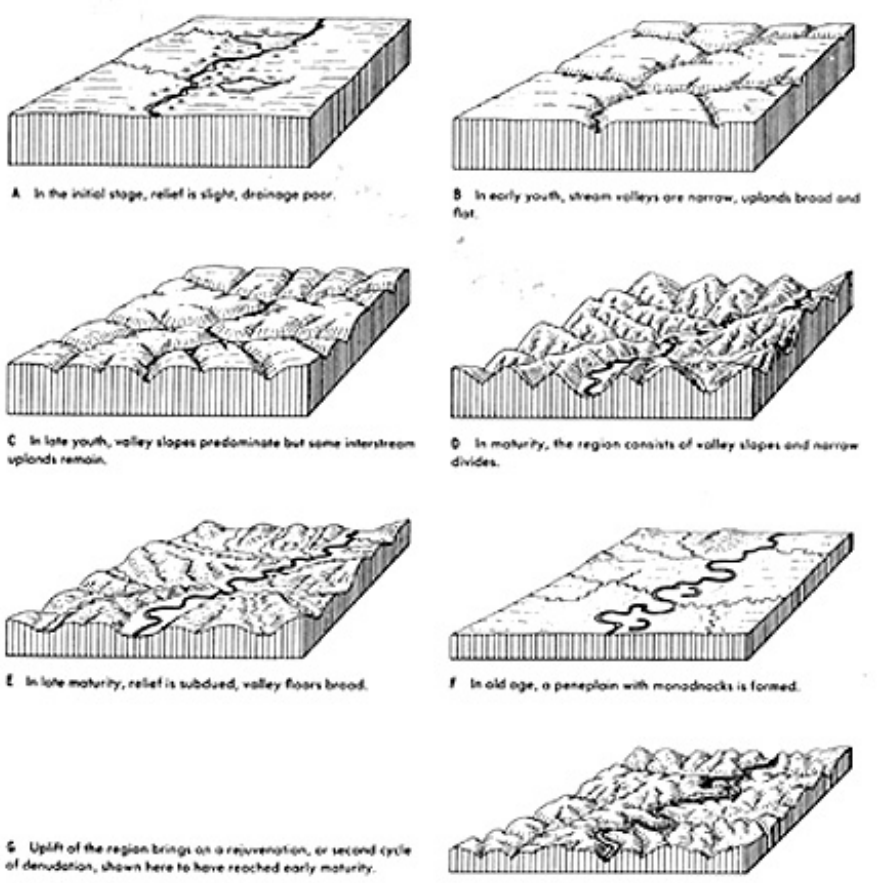
4. The distortions (lateral and/or with up-down movements) of crustal materials combine with physical and chemical reactions between atmospheric constituents (mainly oxygen and water) that weather (breakdown and/or dissolve) rocks which are then eroded, transported (by running water, ice, wind, gravity) and deposited in low surficial locations on land or in water bodies (oceans and lakes). These actions contiually modify the shape of the land and ocean surfaces producing a wide range of continental and oceanic landforms (mountains, valleys, plateaus, plains, volcanic edifices, etc.), developing a wide variety of landscapes.



This illustration shows the four fundamental continental landforms:

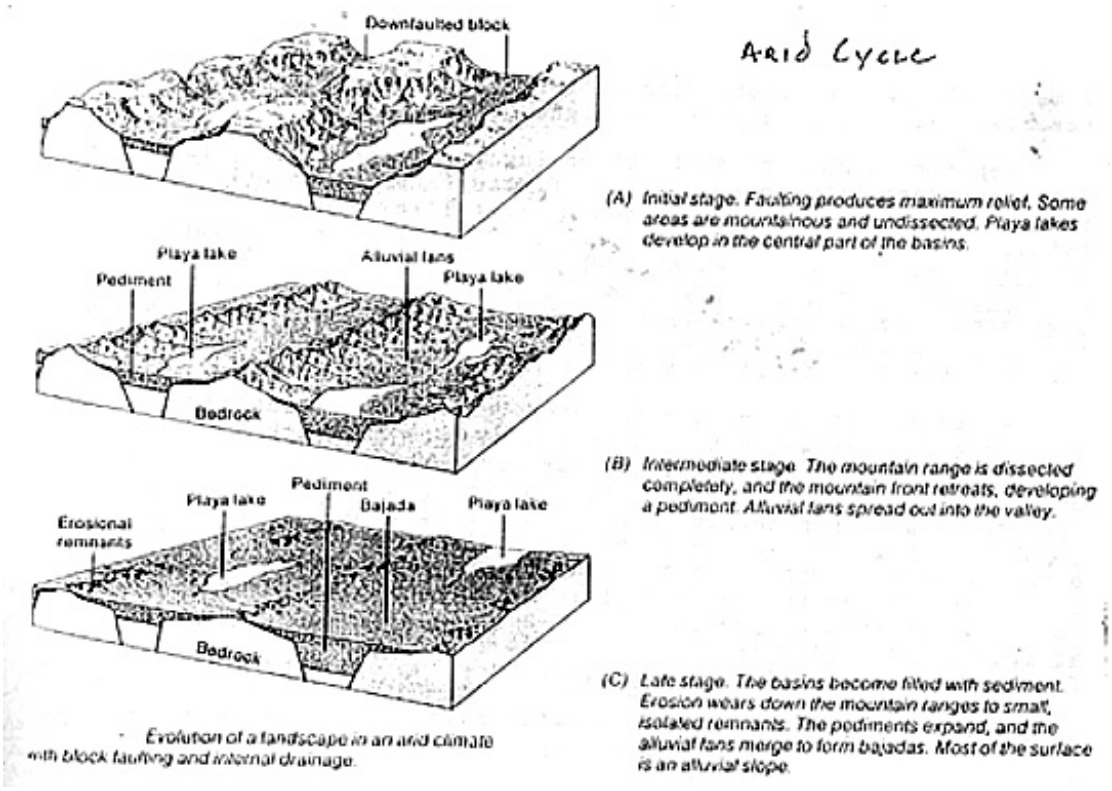
This is a panorama of many of the common landform types:

Landforms development is often a complex process requiring long time periods during which specific landform types take shape, evolve, and disappear. Factors involved, besides time, are the actions of shaping forces such as running water, etc., the type(s) of climate a region experiences (can change from humid to arid or reverse), the nature and resistance to erosion of the various rock type present and their structural configuration, the history of deformation over time, and rises and falls of the regional elevations (through isostasy - a tendency for the crust to assume altitudes that maintain balance [equilibrium] within the Earth's gravitational field). Modern theories of landform development are diverse but most trace their ideas back to 19th Century specialists (Geomorphologists) such as William Davis. While the details have changed, his notions of landform cycles remain largely valid. This illustration generalizes the changing landscape in a humid environment:

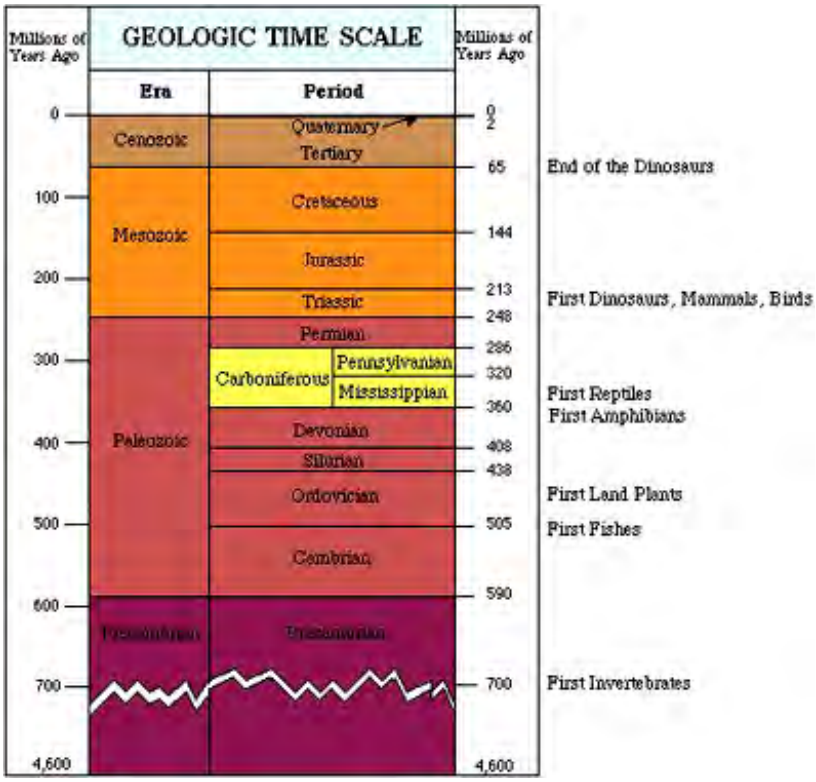


The starting point is the emergence of flat-lying sedimentary rocks from the sea as a coastal plain made up of flat-lying sediments. Streams that develop during Youth follow the gradient (slope) from the highest land to the ocean shoreline start to cut down narrow, steeper-walled valley slopes. The progression then is towards valley widening that leaves uplands as hills and mountains. By a stage called Maturity, the uplands have been carved by enlarging valleys so as to leave only the original uplands at narrow ridges. Thereafter, as Old Age is approached, gravity-driven erosion by sheet flow (thin spread of water over a surface) and by mass-wasting (loose rock movements) over the mountain surfaces slowly reduce these uplands to local hills with the landscape. The mountain terrain, having been generally lowered over time, finally becomes one of low relief (small differences in elevation). Davis called this end product a "peneplain", a term not now used except as an idealization of what a final stage would be like; uplift (block diagram G) is likely to occur before then, which causes a repeat of the overall process cycle (rejuvenation). If the rocks had been inclined (folded) rather than flat, the cycle would have been modified, with hard, less easily eroded rocks maintaining the upland mountains. Rejuvenation has now acted on the present-day Appalachian Mountains such that the ridges (see page 6-3) represent hard rocks and the valleys occur in softer, more easily removed rocks.

A different cycle can be specified for erosion under arid climatic conditions. The end result depends on the structural nature of an eroding region. One case, shown below, relates to mountains uplifted along high angle faults (producing "block-fault mountains") such as in the Basin and Range of the western U.S. (see page 6-8). Here the sequence of change seems simple: from the starting point of high mountains and low valleys, the mountains wear down and their eroded debris fill the valleys, so that the final outcome is a subdued topography with low remnant uplands (pediplains) and deeply filled, raised valleys.

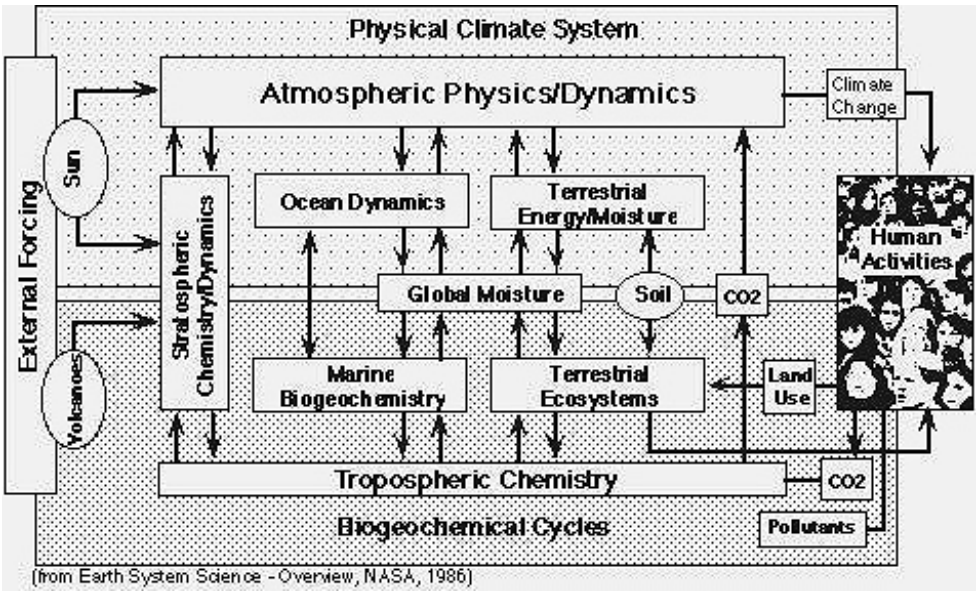


5. Since its beginning, the Earth has been an active, dynamic planet that experiences continual changes in its interior and especially its ouer lithosphere and surface. Its continents have grown relative to oceanic crust and have shifted in position (as referenced to a standard global surface) the movements are called continental drift. Most of Earth's history (expressed sequentially as the Geologic Time Scale) is best deciphered from its rocks, particularly sedimentary ones, that record sequences of modifying events (deduced in part through patterns of lifeforms [usually as fossils] changes (by evolutionary processes) and from rock age measurements (based on fixed rate radioactive decay). Geologic time, especially as expressed in the concept of "stratigraphy" is discussed in detail on the next page.



6. The Earth's surficial environments operate as a complex, interrelated **system** of units and features best categorized in terms of the physical/chemical components of the **Geosphere**. **Atmosphere**, **Hydrosphere**, and **Biosphere** powered by solar and internal heat that interact

at, just below, and above the global surface to produce a series of conditions that aid, inhibit, and otherwise affect *Humans* and all living creatures. The study of how these "Spheres" interact, exchange energy, and produce positive and/or negative feedback is called *Earth Systems Science*. This version of the definitive Bretherton diagram suggests some of these inputs and effects.



On the next page of this primer on the basics of Geology, we will explore in more detail three primary topics: The Rock Cycle; Geologic Time; Plate Tectonics. The subject of Landscapes is more fully treated in Section 17, entirely devoted to Landforms from both ground and space prespectives.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

The second Section concentrates on applications of remote sensing to geological studies. A list of principal uses begins this page. Special attention is given to ways in which remote sensing (especially through image classification) can aid in making geologic maps. The notion of "formation" is discussed and reasons given as to why this standard geologic map unit cannot be recognized as such in imagery alone. One of the pitfalls of making these maps solely from imagery - namely, the presence of soil and/or vegetation cover - is mentioned. Some typical spectral signatures of different rock types are displayed.

NOTE TO EDUCATORS: THIS SECTION, IF WORKED THROUGH IN ITS ENTIRETY, CAN ALMOST BE TREATED AS A *MINICOURSE* IN THE FUNDAMENTALS OF GEOLOGY. Many readers of the Tutorial and particularly Sections 2, 5, 17, and 19, will find the geological concepts and underlying principles unfamiliar or even unknown to them. To assist these individuals in building a quick background, a primer or review of the essential ideas of Geology is offered as an added page, which you can access by clicking here. A similar opportunity is provided in Section 14 dealing with Principles of Meteorology.

Sections 19 and 20 are likewise equivalents of short courses in Planetology and Astronomy/Cosmology.

GEOLOGIC APPLICATIONS:

STRATIGRAPHY AND STRUCTURE

General Background

The synoptic character of Landsat and other large-area coverage satellite remote sensors is proving especially favorable in geologic mapping and other geologic applications. Land cover/use, agriculture, urban monitoring, and similar non-geologic activities related to human endeavors for the most part relate to areal coverage at smaller scales. Geologic phenomena typically are spread over wider scenes, so that the ability to see the "regional picture" is a powerful attribute of space imagery. This perception is supported by the discussion of landforms analysis that makes up most of Section 17.

This view is somewhat tempered by the "interference" of vegetative cover in trying to single out geologic features. Aerial photography and space imagery work especially well on geologic subjects when the scenes they portray are minimally covered with vegetation. Early spring pictures are usually optimal (before leafing). Also, best results obtain when rocks are exposed (outcrop), are covered with little soil or are bare, and are not buried by alluvial deposits. The best images seen in this Section are in the western United States and arid regions elsewhere in the world.

Geologists have used aerial photographs for decades to serve as databases from which they can do the following:

1. Pick out rock units (stratigraphy)

2. Study the expression and modes of the origin of landforms (geomorphology)

3. Determine the structural arrangements of disturbed strata (folds and faults)

4. Evaluate dynamic changes from natural events (e.g., floods; volcanic eruptions)

5. Seek surface clues (such as alteration and other signs of mineralization) to subsurface deposits of ore minerals, oil and gas, and groundwater.

6. Function as a visual base on which a geologic map is drawn either directly or on a transparent overlay.

With the advent of space imagery, geoscientists now can extend that use in three important ways:

1) The advantage of large area or synoptic coverage allows them to examine in single scenes (or in mosaics) the geological portrayal of Earth on a regional basis

2) The ability to analyze multispectral bands (especially hyperspectral data sets) quantitatively in terms of numbers (DNs) permits them to apply special computer processing routines to discern and enhance certain compositional properties of Earth materials

3) The capability of merging different types of remote sensing products (e.g., reflectance images with radar or with thermal imagery) or combining these with topographic elevation data and with other kinds of information bases (e.g., thematic maps; geophysical measurements and chemical sampling surveys) enables new solutions to determining interrelations among various natural properties of earth phenomena.

While these new space-driven approaches have not yet revolutionized the ways in which geoscientists conduct their field studies, they have proven to be indispensable techniques for improving the geologic mapping process and carrying out practical exploration for mineral and energy resources on a grand scale.

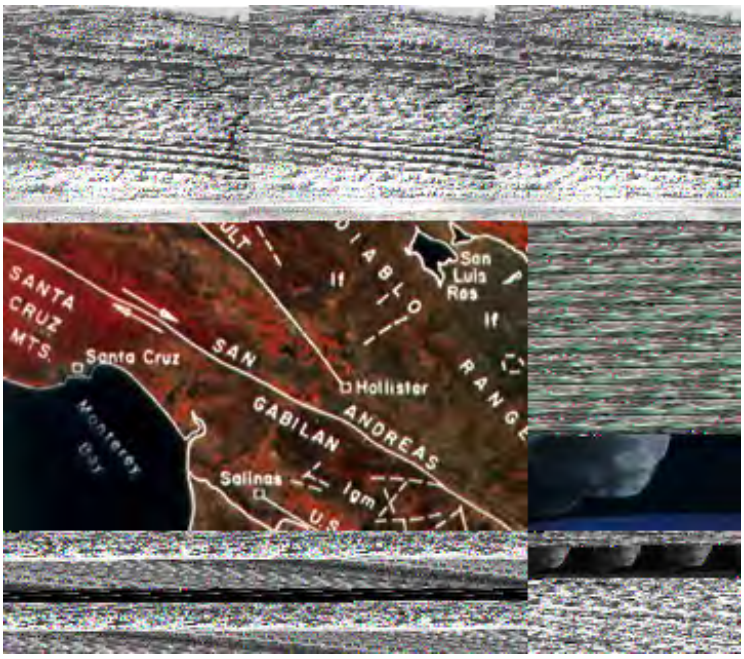
The very first practical use of ERTS-1 (Landsat-1) imagery in any discipline was the drawing by Dr. Paul D. Lowman, Jr, of a geologic structures map superimposed on the first color composite image, based in part on already known field information and in part on his interpretation of this scene. He is a geologist at Goddard Space Flight Center, and an expert on space photography (he prepared Section 12 on Astronaut Imagery in this Tutorial). The image was of the central California coast around Monterey Bay, acquired 3 days after launch.

This map confirmed predictions from his studies of astronauts' photos that Landsat would be an efficient tool for recognizing faults and other known structural trends in small-scale imagery. In spite of lower resolutions, these images excel in portraying regional geologic settings and are easily enhanced by digital processing.

We now consider several examples of geologic applications using these new approaches. We concentrate initially on how Landsat Thematic Mapper (TM) data for a local region in Utah are manipulated to identify different rock types, map them over a large area using supervised classification, and correlate their spatial patterns with independent information on their structural arrangement. Next, our focus changes to examination of geologic structures, particularly lineaments, as displayed in regional settings in the U.S., Canada, and Africa. Then, in Section 5, we will look at how space-acquired data fit into current methods of exploring for mineral and hydrocarbon deposits by considering a case study of a mineralized zone in Utah and at a large-area Landsat scene in Oklahoma. In Section 17, we will return to a geologic theme by examining landforms at regional scales, (so-called Mega-geomorphology), as a prime example in considering how remote sensing is used in basic science studies.

Most geologic maps are also stratigraphic maps, that is, they record the location and identities of sequences of rock types according to their relative ages. The fundamental rock unit is the **formation** (abbreviated as Fm or fm), defined simply as a distinct mappable set of related rocks (usually sedimentary) that has a specific geographic distribution. A formation typically is characterized by one or two dominant types of rock materials.

The term "formation" is most commonly associated with strata, namely layers of sediments that have hardened into sedimentary rocks. Under most conditions, sediments are laid down in horizontal or nearly so layers on sea floors, lake bottoms, and transiently in river beds. Here is a typical set of sedimentary layers exposed in a road cut (note that the layers have been cut and slightly offset by a break which is termed a "fault"):

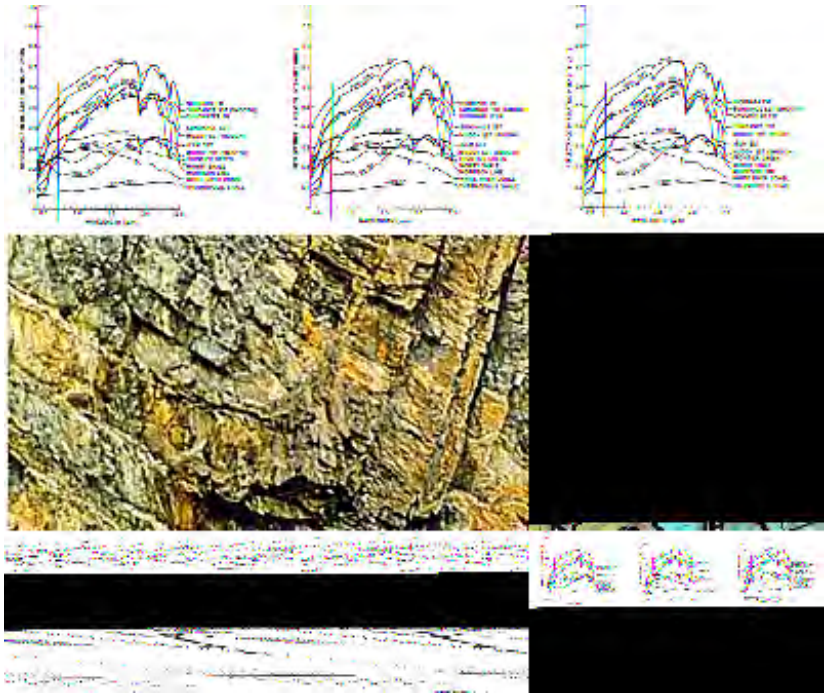


If we see sedimentary rocks inclined at more than a few degrees from the horizontal, we should suspect that these are involved in displacements from their original horizontal state by forces (tectonic) that cause the rocks to bend and curve (folds) or break (faults). Here is a roadcut along a Maryland highway that is passing through the fold belt of the Appalachians.

As we shall see later in this Section, inclined layers can produce curved structures called anticlines (uparched) and synclines (downarched), Here is an example of the latter:

Any given formation is developed and emplaced over some finite span of geologic time. We can approximate its age by the fossils (evidence of past life) that were incorporated within the soft layers (which become strata or beds) during the time in which these life forms existed. Age dating by determining the amounts of radioactive elements and their decay-daughter products can usually produce even more accurate age estimates. Another, less precise, approach to fixing the age (span) of a rock unit is to note its position in the sequence of other rock units, some of whose ages are independently known. We can correlate the units with equivalent ones mapped elsewhere that have had their ages worked out. This method tends to bracket the time in which the sedimentary formation was deposited but erosional influences may lead to uncertainties. The association of sedimentary layers with specific time intervals constitutes the field of *stratigraphy*. Igneous and metamorphic rocks also have time significance and are treated as rock units on geologic maps (which show all stratigraphic and crystalline units in a legend).

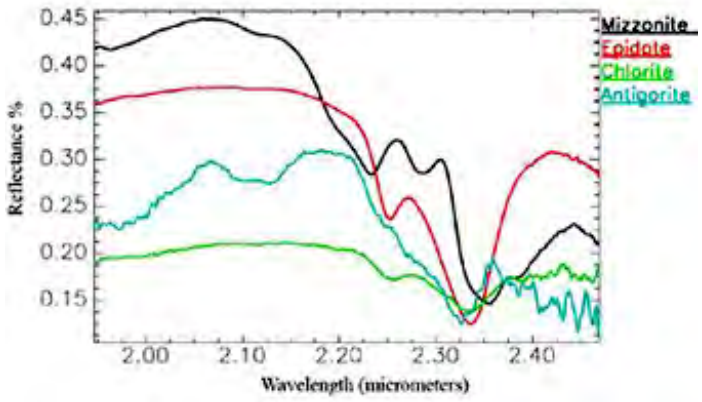
Remote-sensing displays, whether they are aerial photos, space-acquired images, or classification maps, show the surface distribution of the multiple formations usually present and, under appropriate conditions, the type(s) of rocks in the formations. The formations show patterns that depend on their proximity to the surface, their extent over the surveyed area, their relative thicknesses, their structural attitude (horizontal or inclined layers), and their degree of erosion. Experienced geologists can recognize some rock types just by their appearance in the photo/image. They identify others types from their spectral signatures. Over the spectral range covered by the Landsat TM bands, the types and ages of rocks show distinct variations at specific wavelengths. This is evident in the following spectral plots showing laboratory-determined curves obtained by a reflectance spectrometer for a group of diverse sedimentary rocks (collected from their field formations [named from a geographic locality near where the fromationas unit was first described) from Wyoming:



2-1 From these spectra, predict the general color of these four rock units: Niobrara Fm; Chugwater Fm; Frontier Fm; Thermopolis Fm. ANSWER

2-2: **What spectrally distinguishes the Mowry Fm from the Thermopolis Fm (both dark in the field); the Jelm Fm from the White River Conglomerate? ANSWER**

Several of these spectral signatures are composite, in that more than one mineral substance is present. Spectra of individual minerals can be quite different, as shown in the next illustration:

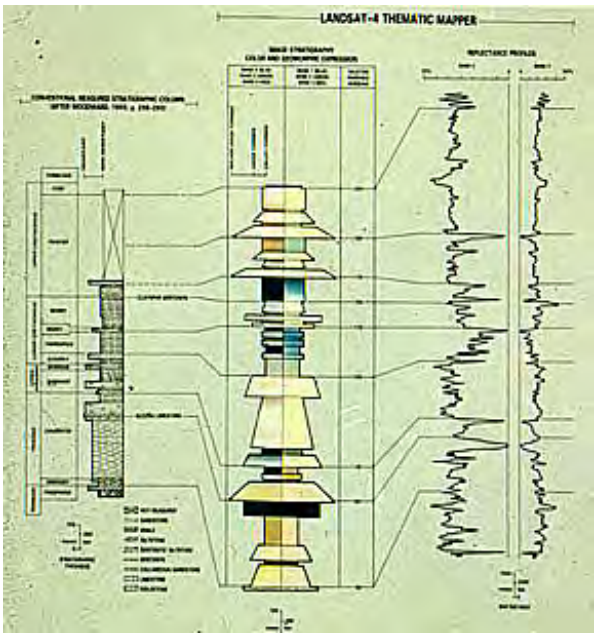


The Near-IR wavelength interval from 2.0 to 2.4 µm is especially sensitive to absorption of radiation; note that all four minerals share a negative peak (trough) around 2.32 µm. But a word of caution: there are several thousand mineral species known from Earth. The spectra of many of these can be similar, making field and even laboratory distinction of just what mineral species is being looked at difficult to assess (X-ray diffraction patterns are usually better identifiers). When remote sensing is involved in onsite surveys, knowledge beforehand of the particular minerals likely in the sensed scene is usually a benefit. Thus, when an area is being surveyed for specific mineral content (as in exploration for ore minerals; Section 5), this information aids in identifying the mineral species being mapped. This is also true for rock types; whenever possible, field spectra such as shown above for Wyoming formations are gathered as Ground Truth (Section 13) so that their diagnostic characteristics can be used to calibrate the image analysis (as an example, referring to the rock spectra above, the averaged reflectances of the 6 non-thermal TM bands can be calculated for, say, the Muddy Sandstone; this knowledge could then be fed into a computer classification as a discriminator function which looks for all pixels with similar multiband values).

A common way of mapping formation distribution is to rely on training sites at locations within the photo/image. Geologists identify the rocks by consulting area maps or by visiting specific sites in the field. They then extrapolate the rocks' appearance photographically or by their spectral properties across the photo or image to locate the units in the areas beyond the training sites (in effect, the supervised classification approach).

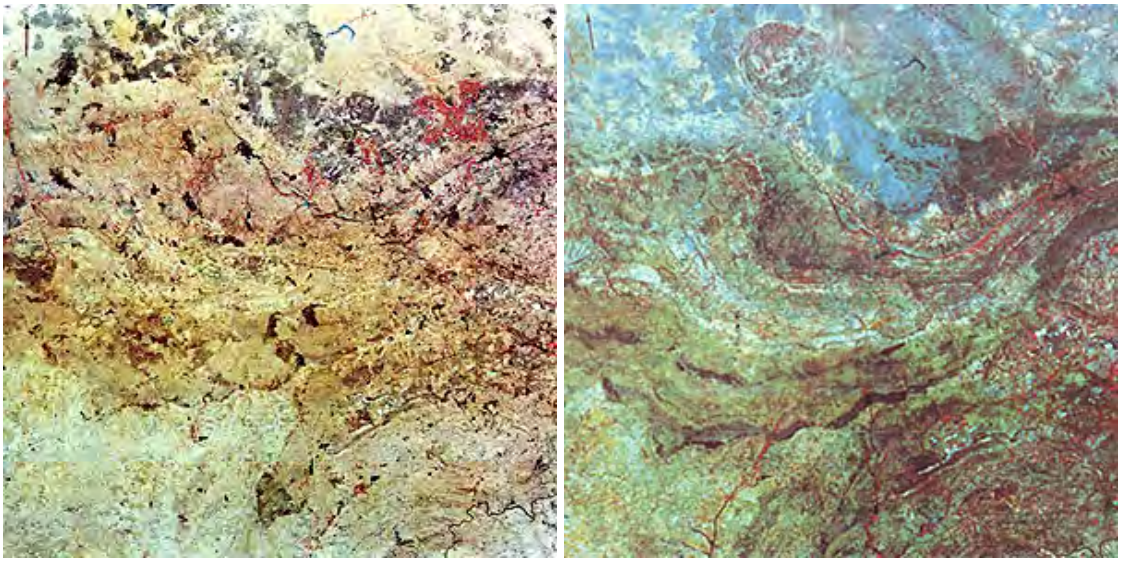
In doing geologic mapping from imagery, we know that formations are not necessarily exposed everywhere. Instead they may be covered with soil or vegetation. In drawing a map, a geologist learns to extrapolate surface exposures to underneath covered areas, making logical deductions as to which hidden units are likely to occur below the surface. In working with imagery alone, these deductions may prove difficult and are a source of potential error. Also, rock ages are not directly determined from spectral data - only material types are determined, so that identifying a particular formation requires some independent information (knowledge of a region's rock types and their sequence).

In exceptional instances, such as those to be shown on the next three pages, when geologic strata are turned on their side (from folding; discussed on page 2-5) so that the successive geologic units are visible as a sequence, the changes within and between each discrete unit can be measured in terms of some spectral property, as for example, variations in the reflectance of a given band, or a ratio of bands. When plotted as shown below the result are tracings that resemble (analogously) those made from well logging of such properties as electric resistivity, permeability, magnetic intensity and other geophysical parameters. Here are two figures, the top showing the succession of sedimentary strata exposed along the Casper Arch in central Wyoming; the bottom being reflectance "logs" derived from spectral traverses along one of the lines in the upper image:



In the lower diagram, the bottom unit is the Permian Phosphoria Formation, extending upward from the Triassic Chugwater Formation to the Frontier sandstone (Cretaceous) at the top. On the right the left tracing is of TM Band 3 (red), with 0% reflectance on the right base extending to 70% on the left, and the right tracing, for Band 1, goes from 0% on the left to 50% on the right.

Before looking at some specific examples of the use of space imagery for geologic structure analysis, this is a good point to introduce one particular advantage of having space observing systems that can repetitively cover the same large regions over the four seasons. Two Landsat images appear below: one taken during the southern Winter in South Africa; the other during the height of Spring. The area includes Johannesburg, some of the gold mines in the Witwatersrand district, and the Pilanesburg pluton (near the top). In the wintertime, some of the underlying rock units fail to show distinctly because the entire scene has its vegetation (mostly grasslands) dormant. But with plant reawakening in Spring, different units have different vegetation types and these variably modify the colors displayed, thus revealing the more complex structures in the region.



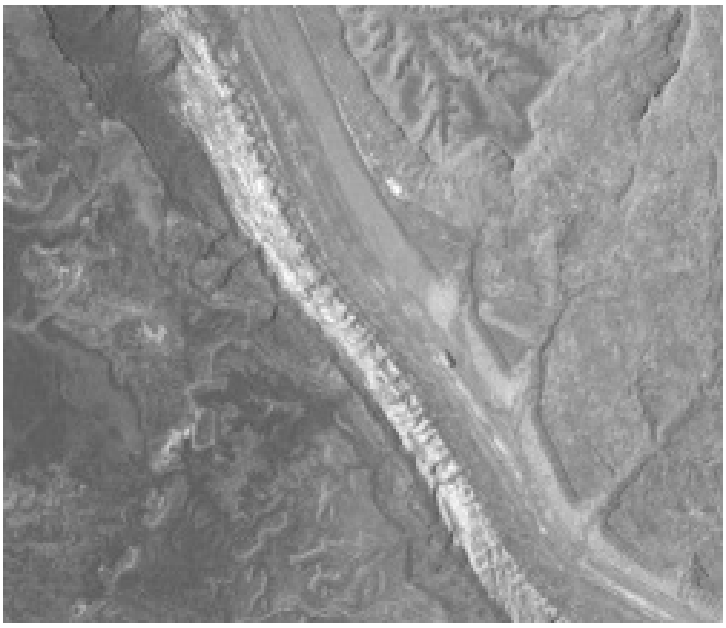
This geobotanical phenomenon (the differential distribution of various plant types as a function of soils developed on different rock types) is sometimes used as a mapping or prospecting device (example; the element Selenium is associated with Uranium; certain plants thrive on Selenium enrichment and are thus indicators of subsurface Uranium enrichment).

Primary Author: Nicholas M. Short, Sr.

The making of a geologic map with the aid of remote sensing imagery is illustrated for an area in Utah known as the Waterpocket Fold (a monocline as defined on this page). Its location with the Colorado Plateau is specified and that major geomorphic/structural unit is described and illustrated. A Landsat TM 1 image of the Fold area is introduced and several aerial oblique photos and ground shots help one to visualize the geologic units and terrain found there.

Using Landsat for Geological Studies

With these provisos in mind, we view the first scene to be analyzed. It's a demonstration of the mapping capabilities of a subset from a Landsat TM image dominated by the Waterpocket Fold that lies within the Capitol Reef National Park in south-central Utah about 35 miles south of Caineville on Highway 24.



2-3: **The Waterpocket Fold is shown in this image just above. Examine the Landsat TM Band 2 scene closely. How many distinct and different units do you think you can distinguish in this band, based solely on gray level variations?** **ANSWER**

The Capitol Reef Park is well within the Colorado Plateau, a major physiographic province that extends across parts of Colorado, New Mexico, Arizona, and Utah. The Plateau's best known natural feature, of course, is the Grand Canyon. There, the Colorado River has cut nearly a mile deep into horizontal rock strata of Paleozoic age. These horizontally layered sedimentary rocks are the hallmark of the Plateau. Over most of the U.S., great forces from lateral interactions between the North American and Pacific tectonic plates and other plates have caused strata in the upper continental crust to be folded, faulted, and crushed into numerous mountain belts, of which the Rocky Mountains, the Wasatch, the Pacific Coast Ranges and many others are prime examples This is one of the central ideas in the great paradigm known as Plate Tectonics (plates are subcontinental scale blocks or segments of the Earth's lithosphere; the term "tectonics" refers to the large-scale deformation and movements of/in the Earth's crust, leading to such structural features as folds, faults, thrust blocks, and fractures). However, in one region in the West - the Colorado Plateau - the crust responded differently by being lifted vertically

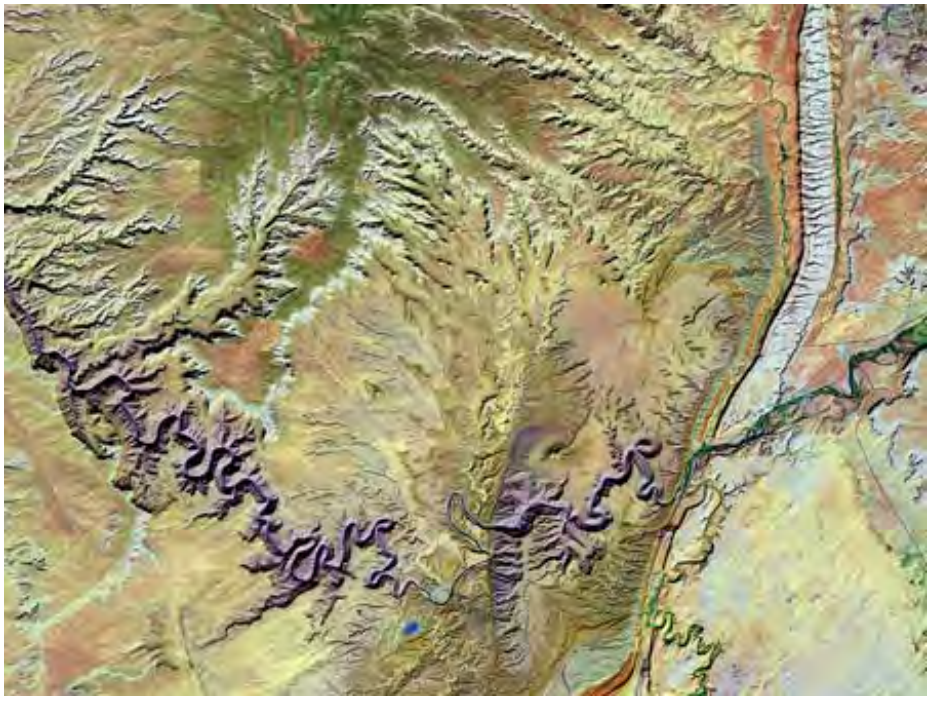
while maintaining the initial horizontal orientation of the cover strata, rather than being compressed laterally into deformed units.

In the Plateau, there is one structural exception to this. In places, the rocks are gently folded in broad warps called anticlines and synclines, in which rock layers become inclined (we say they dip) in opposing directions. Even more common are a type of folds called monoclines.

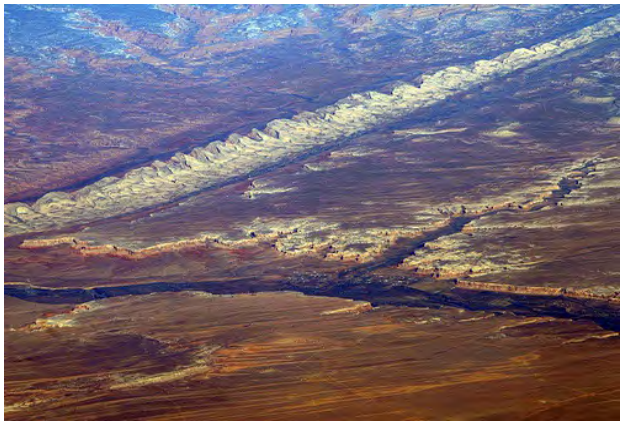
The scene below is a beautiful Landsat MSS full scene taken on June 13, 1977. That shows the northwestern section of the Colorado Plateau, including parts of the so-called Canyonlands. The subscene you will study is near its upper left corner; the black line marks the flight line of a NASA aircraft overflight during which a multispectral scanner acquired data (see next page).

Most of the upper scene lies in Utah but the lowermost areas fall within northern Arizona. Passing through the scene is the Colorado River that, backed up from the Glen Canyon Dam (lower left) near the town of Page, Utah fills the steep-walled canyons to form Lake Powell. The San Juan River joins the Colorado from the east. The prevalent yellow color in this image (printed on Cibachrome, which improves both color quality and definition of features) corresponds mainly to sandstones and some shales that are reds and oranges in nature. The brown tones associate with rocks that are dark reds and browns in the field. Some fields are tinged with reds in the image because vegetation in this IR version. Blues associate with shales that in the field are often grayish. Areas with strong red (IR) tones include the Abajo Peaks (upper right), the Henry Mountains (upper left), and the smaller, isolated Navajo Mountain east of Lake Powell. The northernmost end of the forested Black Mesa appears along bottom center. The cluster of isolated mountains to its north are the collection of mesas and buttes, known as the Hopi Buttes district. Instead of being made from sedimentary rocks, as occurs in Monument Valley to the north, these are actually eroded volcanic necks and diatreme fills. They will be discussed on page 17-3.

In the upper left is the monoclinal fold (crossed by a black line) known as the Waterpocket Fold, which is the topic of the next few pages of Section 2. Along the right side of the Landsat image is a long band of folded rocks that comprise Comb Ridge. These are shown again in these two Landsat images:

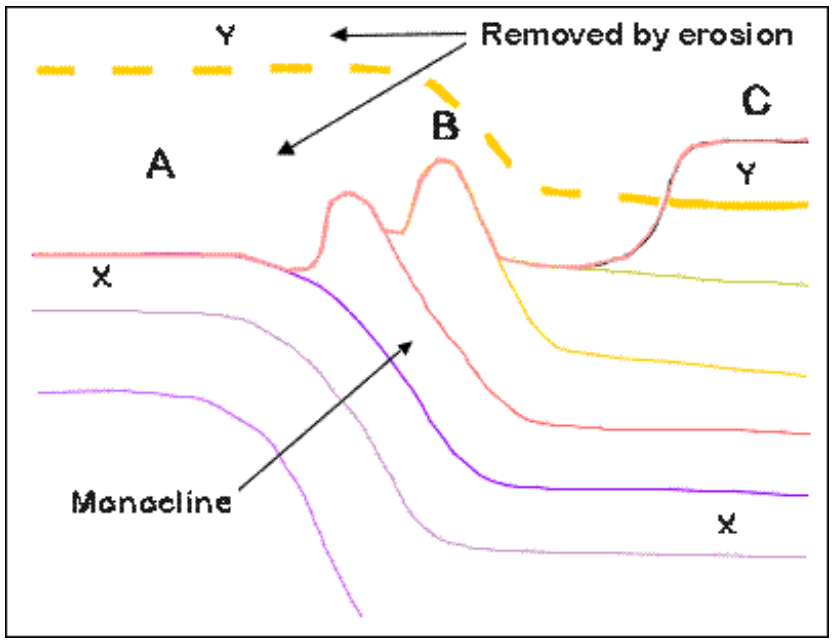


Comb Ridge, near Bluff, UT, is sometimes called the "Great Wall of Utah", as is suggested by this aerial view. The ridge is actually a geologic structure called a monocline.



In a monocline, forces cause rocks to change their inclination in only one direction and then resume the initial direction, as depicted in this sketch which presents a side view (as though it

was along a vertical cut into the Earth's surface layers) of a monocline. At the top is a line which represents a profile of the ground surface. At B, above the inclined layers, rocks that are more resistant to erosion are shown as ridges (often called "hogbacks" or "flatirons"). At A, a rock unit labeled X lies just beneath the surface, but at B it has dipped downward such that it is now well below the ground. A unit labeled Y at C is well on top of X. These rock units and the rocks in between follow the stratigraphic Law of Superposition, which states the obvious: younger rocks are deposited on, and thus rest above, older rocks. If we follow Y to the left along a dashed line, it bends upward at B and also lies above the ground at A. In other words, it was once there but has been removed by erosion.

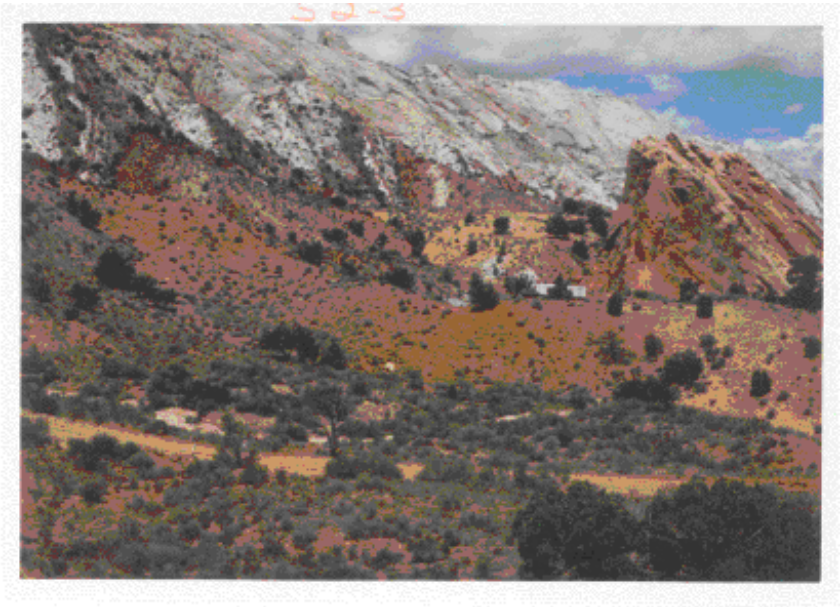


2-4: **The two most common types of folds are called anticlines and synclines. Do you know what these structural terms mean? If not, check the answer.** **ANSWER**

This is the situation at the Waterpocket Fold, a monocline which runs across this subscene. A black and white view followed by a color view, both taken looking to the northwest from low-flying airplanes, show a hogback ridge dominated by a whitish unit, the famed Navajo Sandstone (the principal rock formation at Zion National Park) and several other (reddish) units near its base. The thick red unit beyond (west) the Navajo Sandstone is the Wingate Formation. These several units dip at angles as steep as 45 degrees. Younger gently dipping rocks in the foreground (bottom of photo) are part of the lower bench of the Tarantula Mesa whose top (east of the photo; equivalent to C in the above diagram) is actually higher in elevation than the fold. We can barely see older rocks (equivalent to A) below the distant western horizon. These make up the Circle Cliffs.

A third aerial view taken from the west side of the fold looking to the northeast, presents these older rocks to the left and younger rocks in the distance, with the Fold highlighted by the Navajo Sandstone and standing out as a narrow band of outcrops along the structurally-controlled topographic ridge.

Now, we take two close-up looks from the ground, one looking up against the Fold with the Navajo Sandstone near the top and the Entrada Sandstone (reddish) below it but stratigraphically young.



The second photo looks eastward at Tarantula Mesa (a stepped, flat-topped, plateau) with the Mesaverde Sandstone capping it; the famous Henry Mountains (where Gilbert conceived the idea of a laccolithic intrusion) against the skyline; and the Dakota Sandstone as a blocky outcrop in the foreground.

2-5: **In the first aerial oblique color photo (which the writer made literally hanging out the window of a light aircraft), how many rock units within the steeply inclined limb of the monocline can you separate visually. In the black and white aerial oblique photo, the slightly inclined rock units along the right side margin seem blackish; why?**
ANSWER

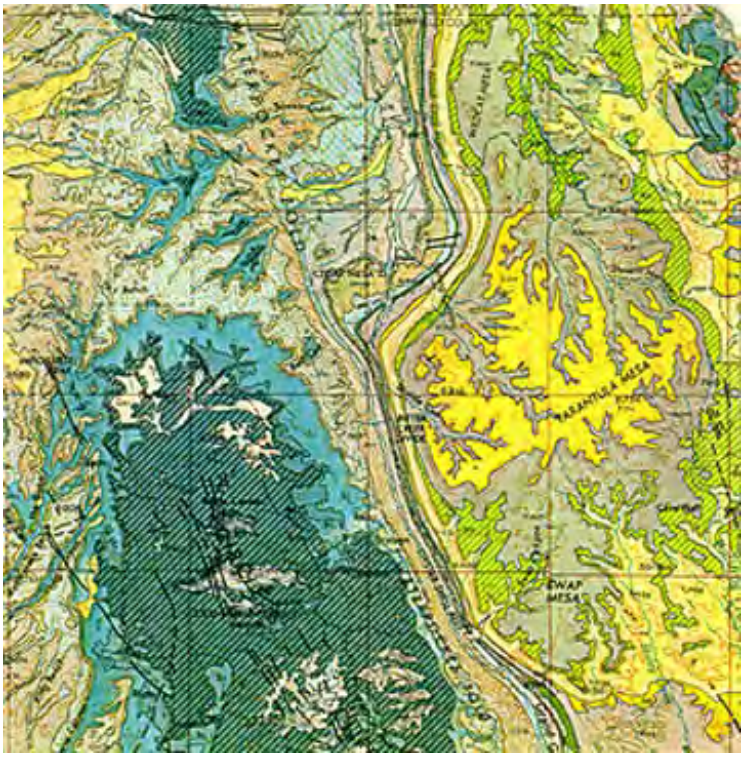
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

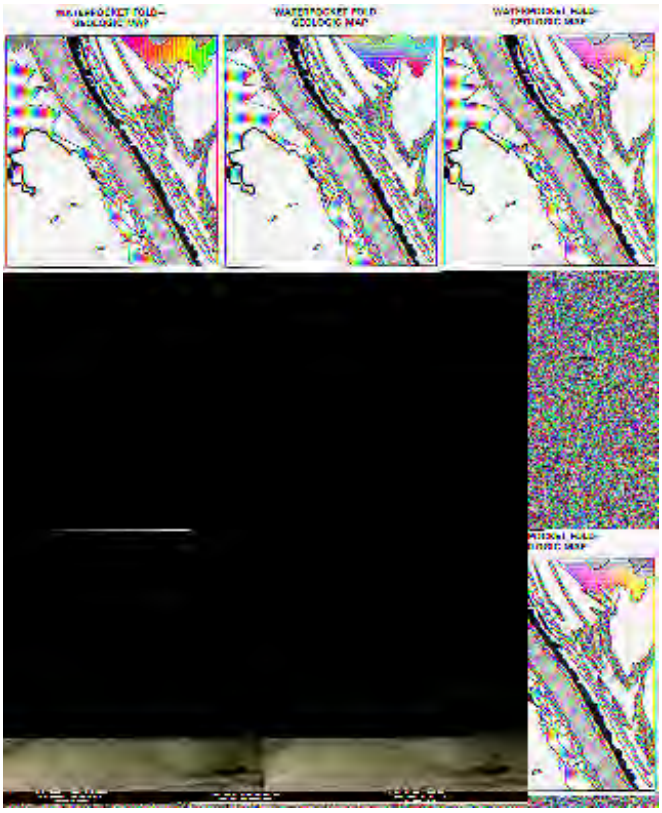
The details of the geology of the Waterpocket Fold, displayed in an aerial photo, and by a airborne multispectral scanner and Landsat imagery, are covered on this page. A published Geologic Map and one devised from the imagery depict the main stratigraphic units present, some of which are located on the Landsat image used as a frame of reference. Various color renditions of the scanner and Landsat images bring out different aspects of the ground classes.

Geologic Maps and Images of the Waterpocket Fold

We can understand these rock-unit relationships by examining two geologic maps. The first was extracted from the Utah state geologic map. The blue units are Upper Paleozoic limestones and shales and Triassic sandy shales. Jurassic to Cretaceous sedimentary rocks make up the thin band outcrop patterns (just right of center) within the Waterpocket Fold. To the east are Cretaceous shales (green stripe pattern) and a capping sandstone, the Mesaverde (in yellow).



The writer made the lower map, generalizing the rock units shown on a large scale map published by the U.S. Geological Survey, to depict the approximate area we cover in the Landsat subscene. The individual units, or formations, are indicated, with their symbols, in the legend, in which the youngest strata are placed in the upper left (first column) and the oldest in lower right (second column). Refer to this second map when we display the scene images. You may want to print it if your browser supports this.



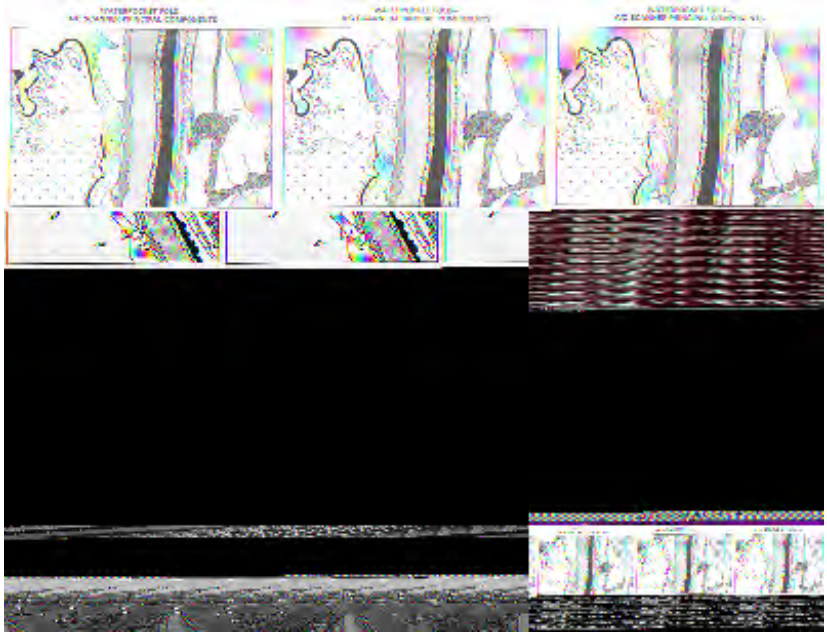
The Circle Cliffs are located in the lower left of the map, the Waterpocket Fold at the center, and the Tarantula Mesa at the upper right of this map.

The Waterpocket Fold was one of 11 Utah sites imaged by the Bendix 24-channel (band) multispectral scanner flown for the writer in June 1976 as part of NASA Aircraft Mission 340. The purpose then was to simulate the TM sensors to determine if the multiband data were capable rock-type identification and geologic mapping. The flight included an onboard aerial camera with color film which obtained high resolution photographs at a scale of 1:24000 (one inch on the picture = 24,000 inches on the ground), one of which we reproduce here to detail the area that we will examine in a Landsat scene.

You can also compare this photograph with the color PCA image made from some of the 24 channels (bands) on the Bendix airborne scanner; the bright area towards the right in the photo corresponds to the red patch in the scanner image.

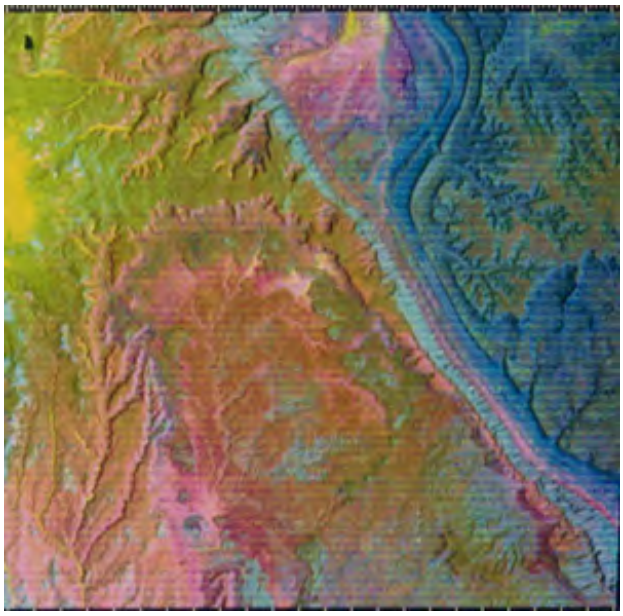
The resolution in this image is 5 m (17 ft). Data from seven channels in the visible and near-IR were combined to generate a series of images based on six Principal Components. This color composite was made from the first three components. Using this composite as an indicator of the rock unit classes, an interpreted geologic map (not shown here) was drawn, in which the symbols for each formation are the same as the field-based geologic map, which we already examined. A strikingly strong correspondence between the two maps (the published geologic and the scanner data maps), is evident which substantiates the premise that high resolution, narrow band multispectral imagery can be a powerful tool for mapping.

A supervised Maximum Likelihood Classification of the Bendix scanner image was produced, using 8 of its bands to produce 8 Principal Components and choosing the training sites from the writer's onsite knowledge of the Waterpocket Fold at the area shown in the aerial photo. Here is the resulting map (use legend above):

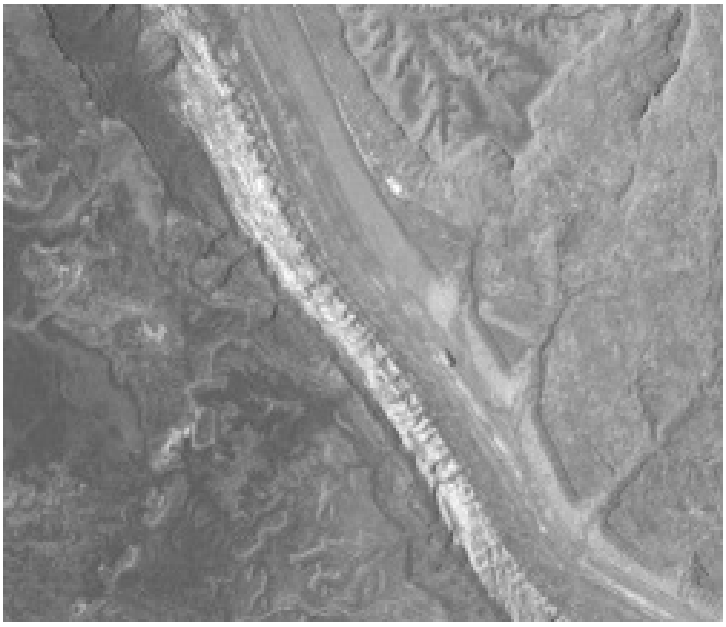


How well can space imagery capture this scene? The next image is a false color composite (band 4 = blue; band 5 = green; and band 7 = red), made from data acquired by the MSS onboard Landsat 1 on August 6, 1972 (scene 1014-17370). This image was produced using JPL's VICAR imaging software program. Except for the patch of red vegetation near the top, the colors of the rocks are roughly similar to their natural colors. The Triassic units (Moenkopi; Chinle) and the Jurassic Wingate Formation are orange-brown here. The Navajo is white and the Entrada and younger units to the east are yellow-brown. The Masuk and Blue Gate shales with interbedded sandstones are bluish, while the Mesaverde is brownish.

The same August 1972 data set was processed by JPL's VICAR software using the Karhuen-Loeve transformation program to calculate four principal components. In the PCA composite image below, PC1 was assigned to blue, PC2 to green, and PC3 to red. The units east of the Fold show mainly in blue colors but so does the Navajo Formation. The Triassic rocks appear largely in shades of red. The Kayenta is expressed in green. The yellow patch at the left is of uncertain identity but probably coincides with thick stands of cedars and some pines.

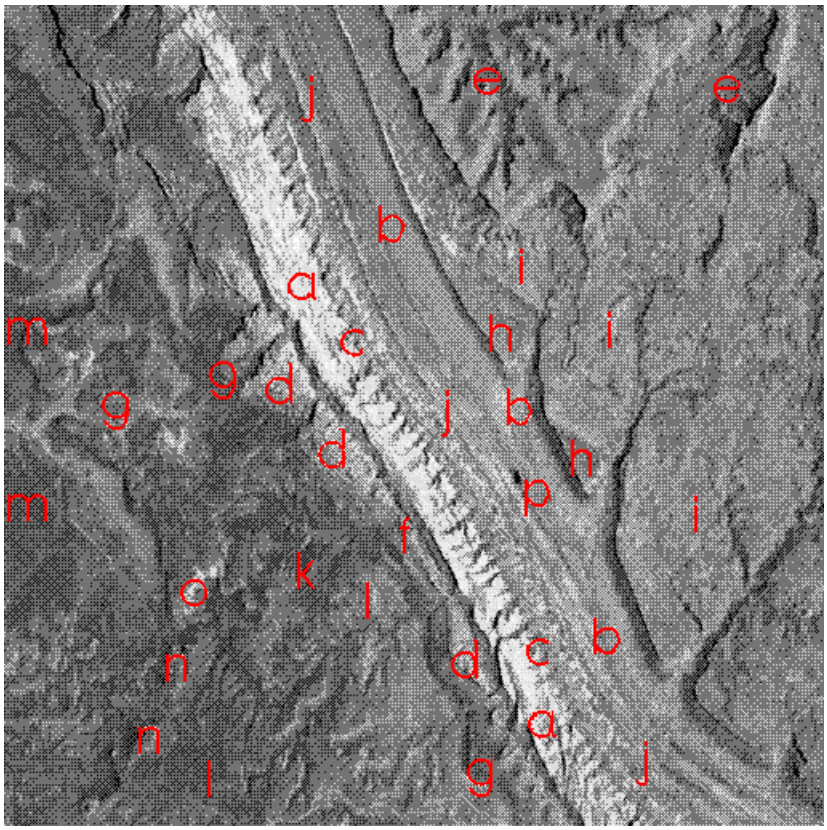


Next, we show a TM Band 2 image acquired by Landsat 4 in the summer of 1984.



Consider the following observations and refer to the geologic map as necessary: The terrain east (right) of the Fold, occupied by units from the Blue Gate Shale to the Mesaverde Sandstone, generally has brighter tones in keeping with the facts that the two sandstone Formations are lighter and more reflective and the shales have weathered, bright surfaces that mask their darker, fresh colors.

Look at the labeled (and degraded) version of this band to examine the tonal responses for various other units. You may want to print this map for easier reference.



The Navajo Sandstone (**a**) is the brightest single unit in the scene, followed by the valley floor to the east, much of which is covered by alluvium (**b**). The Entrada (**c**) and Wingate (**d**) Sandstones are notably darker, as is the top surfaces of the Mesaverde (**e**). Note that the Wingate "pinches" out along the Fold towards the bottom of the image, which suggests that its dip steepens making its outcrop pattern thinner. The Kayenta Formation (**f**) is barely discernible in the image as a thin, darker tone, but the Chinle (**g**), also darker, has a wider outcrop pattern, because it is thicker and less steeply dipping. The Emery Sandstone (**h**) has a darker tone than the overlying Masuk Shale (**i**) because, in the field, it is dark brown (similar to the Dakota Sandstone seen in the foreground of this picture from the ground).

Units in the sequence from the Salt Wash to the Ferron Sandstone (around **j**) are not visually separable, a consequence of their similar colors and relative thinness. The oldest formation, the Moenkopi , appears as a single flat-lying unit on the geologic map, but in the scene we can visually subdivide its outcrop areas (mainly in lower terrain) into two subunits. In fact, in the field the Lower Moenkopi (**k**) is a darker reddish than the more brownish, lighter Upper Moenkopi (**l**) but these two stratigraphic members are not thus distinguished on the map. Other darker areas around (**m**) lie within flat-floored valleys and are presumably alluvial fill. Very light thin bands as at (**n**) correspond to the Shinarump Formation on the map. These in the field make up steep faces

along the Circle Cliffs that reflect much more light. But, patches at (**o**) are extraction pits into the Shinarump, which has been mined for its uranium content. The very dark blotch at (**p**) is a cattle watering hole made by damming the intermittent stream. The shales and alluvium in the valley support clumps of sage and other vegetation.

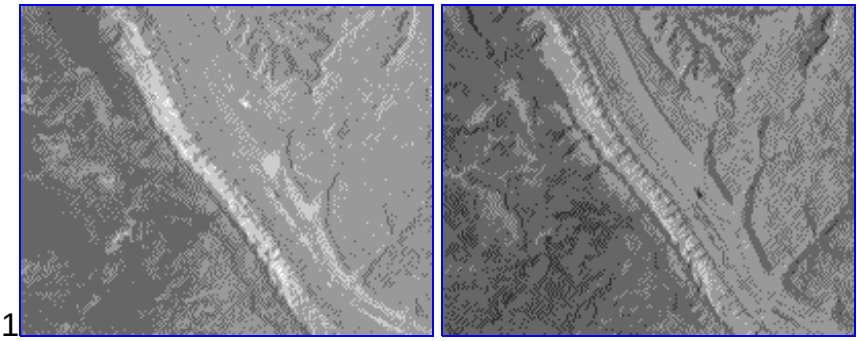
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The Waterpocket Fold is displayed in several varied representations, including TM Bands 1, 6, and 7 true and false color composites, and several PCA composites. Information evident in each is discussed.

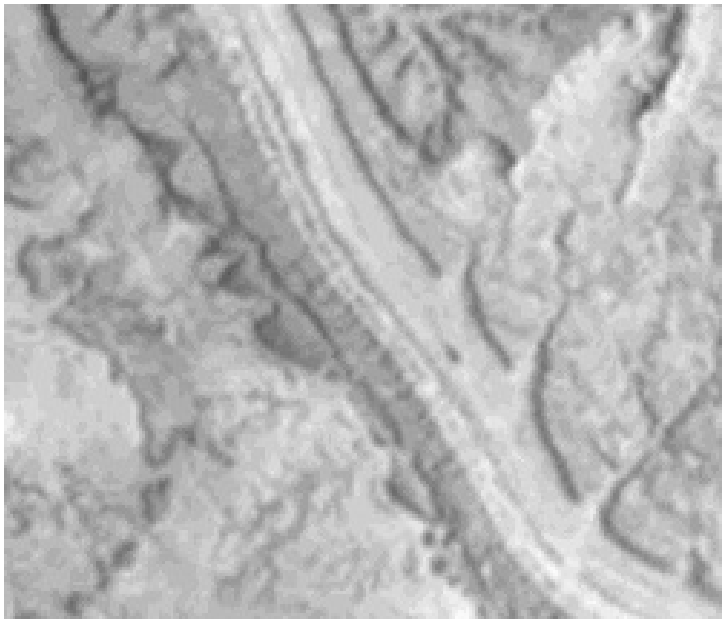
Some Specialized Images of the Waterpocket Fold

Compare the Band 1 image with that of Band 7 noting changes in gray levels at corresponding points. The other reflective bands (not included here) show similar variations.

Band 1  Band 7

2-6: **What is/are the main differences between Bands 1 and 7?** **ANSWER**

But, Band 6 above is quite different from all others:

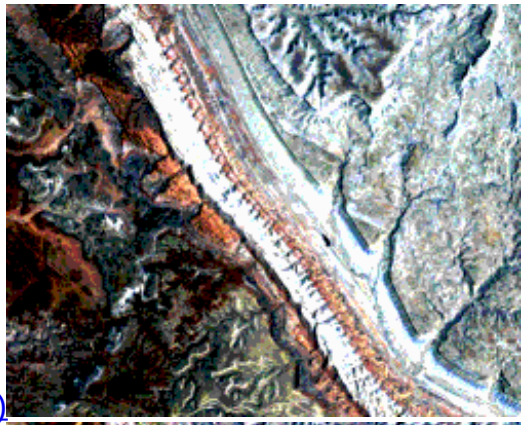


Two mildly contrasting tones (medium grays) are partly the result of orientation of surfaces with respect to the sun's rays but are mainly controlled by the absorptive character of the rock units. The Navajo formation is moderately dark, as are the Emery and Mesaverde sandstones. These units reflect more of the sunlight and hence don't absorb enough to be warm relative to other, darker (in the field) units that absorb solar radiation, become warmer, then reradiate the heating as more thermal radiation. Darkest in the image is the Wingate formation although its eastern dip slopes are less so; a darker shading just to its right may be a combination of the Kayenta

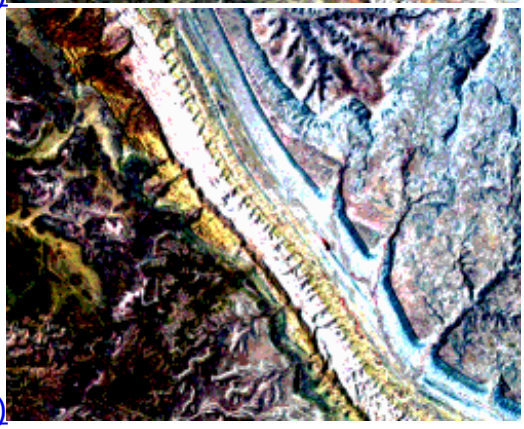
formation and shadows from Navajo hogbacks. A dark pattern near the Blue Gate Formation is a cliff shadow effect. Most shaly units are warmer (brighter) after having absorbed this radiation. Dark areas at the left of the image (beyond the map coverage) consist of wide outcrops of the Shinarump Formation. Similar thermal properties give the Upper and Lower Moenkopi the same medium tones.

The natural color composite (Bands 1, 2, and 3), shown below, is a roughly realistic rendition of this scene as it might appear from the air. The Wingate and Entrada formations are actually orangish in the field and the Lower Moenkopi and Carmel are maroon when seen on site. The units below Tarantula Mesa have a bluish undertone in the image, roughly a counterpart to bluish-gray and tan tones they possess in nature. In the image, it is hard to distinguish the bright area that corresponds to an extension of the Blue Gate formation into the valley from the alluvial fill as mapped; this implies that fill may have spilled onto the Blue Gate, covering it there. Note the dark orange-brown alluvium in the western valleys.

(Bands 1,2, and 3)

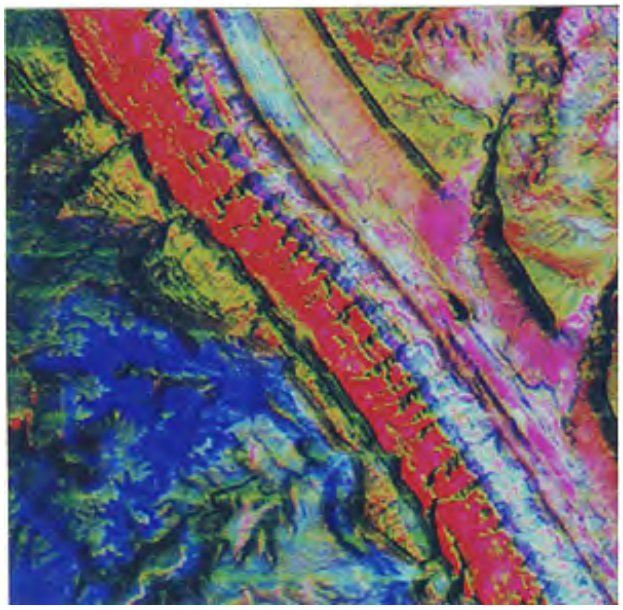


(Bands 2,3, and 4)

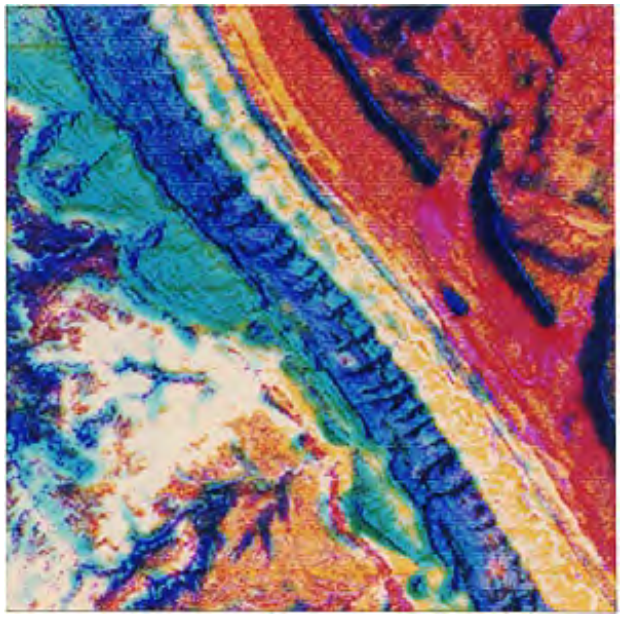


In the above false color composite (Bands 2, 3, and 4) most of the formation units have similar colors (although somewhat darker) to those in the natural composite. Because vegetation is sparse, it contributes little reflectance brightness in Band 4, so that the rock and soil colors result from proportionately related DN levels in each composite. However, bright reds at the watering hole and along the canyon drainage (look in lower right quadrant) are a certain indication that these intermittent creeks support active vegetation growth.

A PCA data set for the Waterpocket Fold was calculated using all seven TM bands from the summer 1984 Landsat overpass. In the VICAR-generated composite shown below, PC1 = red, PC2 = green, and PC4 = blue. The Navajo is uniquely discriminated by red (but note the isolated red patch [top center] of unknown nature). The Moenkopi divides into upper and lower areas, which are clearly visible. Wingate stands out as green, and Entrada as blue.



The same data were used to produce a color hybrid ratio. In the image below, the ratios of Band 3 to Band 2 are displayed in green and Band 4 to Band 5 in blue. Thermal Band 6 is assigned to red in the color composite. Once again there are distinct separations. The Navajo is singled out as blue; the Wingate is blue-green; and the Entrada is a blotchy mustard yellow. Most of the units to the east are dominated by red colors, resulting from their ability as dark shales to absorb solar energy and re-radiate it. The light tan tones coinciding with the lower Moenkopi suggest bright contributions from the thermal band and the two ratios, blending these three primary colors into this whitish composite.



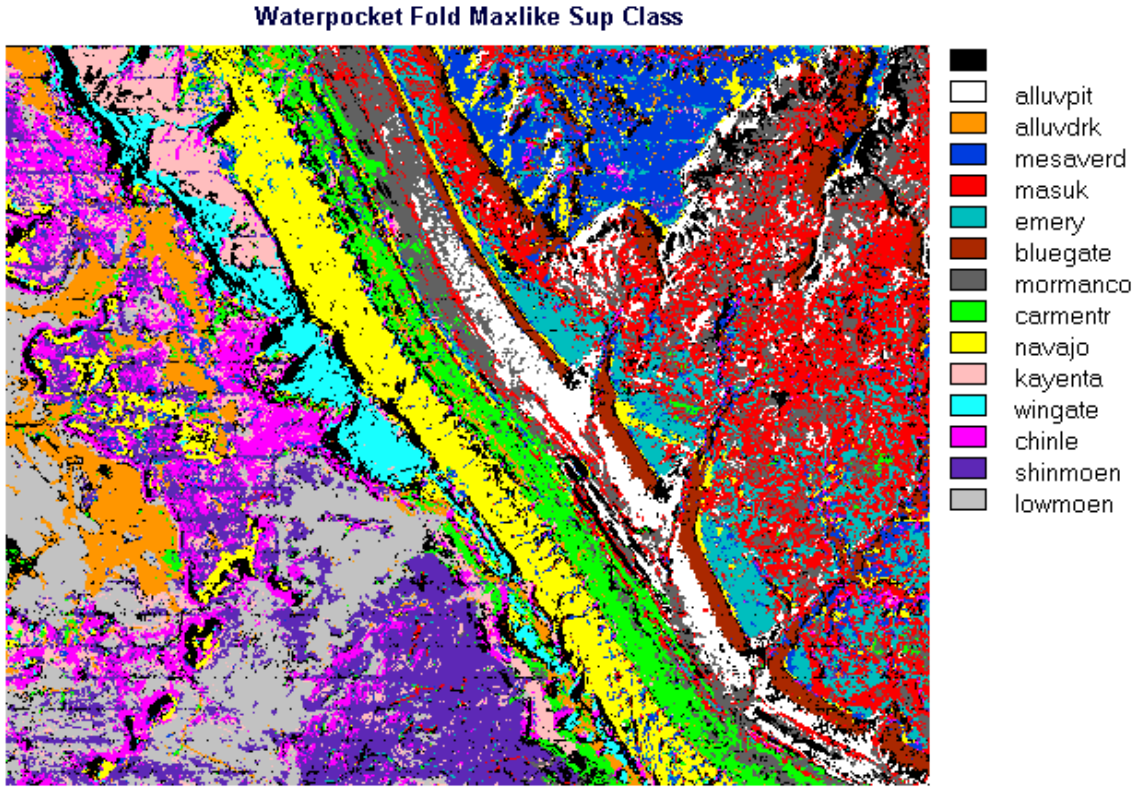
2-7: **Of the PCA and Color Hybrid images, which seems to better separate units?**
ANSWER

Primary Author: Nicholas M. Short, Sr.

The Waterpocket Fold subscene affords an excellent test of the ability to classify a Landsat geology scene because most of the stratigraphic units are distinctive, have almost no vegetation cover, and have thin soil cover much like their parent lithologies. A Maximum Likelihood Classification Map made on IDRISI shows generally a favorable level of accuracy for most units. Accuracy is less for shale units in flatter terrain and fairly high on the Fold itself. Earlier classifications of the TM data made using different training sites and the classification algorithms on the IDIMS processor indicate comparable quality of identification for the summer 1984 subscene but poorer accuracy in a winter 1985 subscene that has reduced and lower angle illumination and more shadowing. A trio of images of another scene, the Saline Valley in eastern California, made by the ASTER sensor on Terra, shows that various broad classes of rocks can be distinguished and separated by using visible, SWIR, and Thermal IR bands to make discriminating color composites.

Maximum Likelihood Classification of the Waterpocket Fold; the Saline Valley; Saudi Arabia

This Waterpocket Fold scene was part of a 1984 study of how accurately the writer could identify rock units using Landsat TM data from arid (vegetation-poor) terrains. First presented is a Maximum Likelihood Supervised classification map of the scene from the summer of 1981, without Band 6, made on the IDRISI processing system.

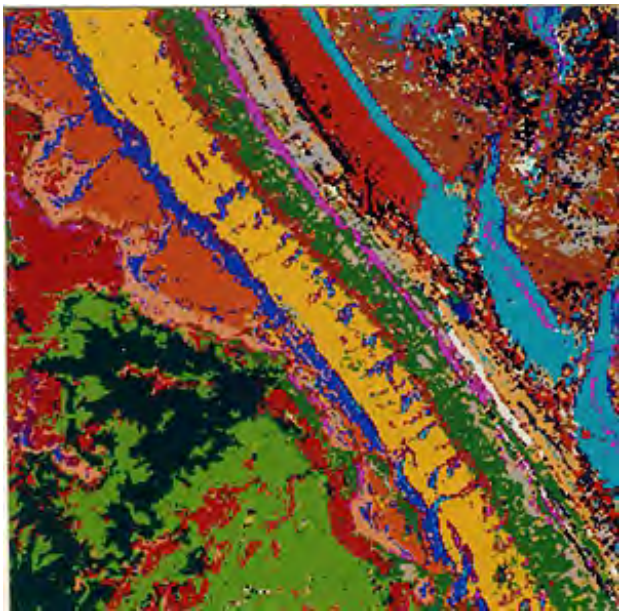


Its classification accuracy is an estimated 78%; that is, the rock units or formations are correctly identified over at least that percentage of surface exposures. This is high in view of shadow and slope effects, soil cover, and other sources of error. Among misidentifications and conflicts: The

Navajo legend color also occurs in the prospect pits and along canyon walls on Tarantula Mesa; black areas such as associated with the localized Carmel/Entrada (combined because of a 15 class limitation) are scattered throughout the scene and generally represent misclassification ambiguities; Kayenta almost disappears in the south except where it shows up incorrectly west of the Chinle (which itself is not well mapped); colors assigned individually to the Morrison/Mancos and the Masuk and the Emery Formations are dispersed within each other. Note too that the Upper Moenkopi and Shinarump Formation were combined into a single training site because the latter is too thin in outcrop area to be mapped discretely at this resolution.

2-8: **What other explanation(s) for the black pattern distribution can you think of; account for the apparent misclassifications associated with the Navajo, since its color appears elsewhere besides its outcrop ridge on the fold; the Morrison-Mancos unit grades into alluvium in the valley; why? ANSWER**

There can be considerable subjectivity in setting up and interpreting any classification. Two more classifications of the Waterpocket Fold are shown below to demonstrate how changing factors lead to modified results. Both were made on the IDIMS system at the Goddard Space Flight Center.

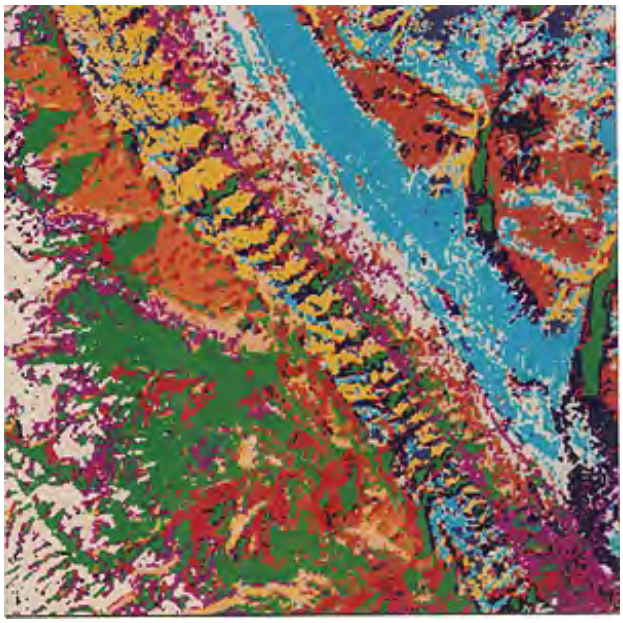


The above classification shows the same summer of 1984 scene but applies a different set of training sites and, unlike the IDRISI one above, includes thermal Band 6. The formations are exceptionally distinct. Red was assigned to alluvium, which, for the most part, was placed correctly (the large section of red occupying the Blue Gate shale member [in blue] of the Mancos formation corresponds to strongly weathered slope wash along steep slopes against Tarantula Mesa). The Dakota formation is a wine purple color (but the white strip near it is of unknown character). The Kayenta is singled out in reddish-brown, and the Chinle is pinkish-peach.

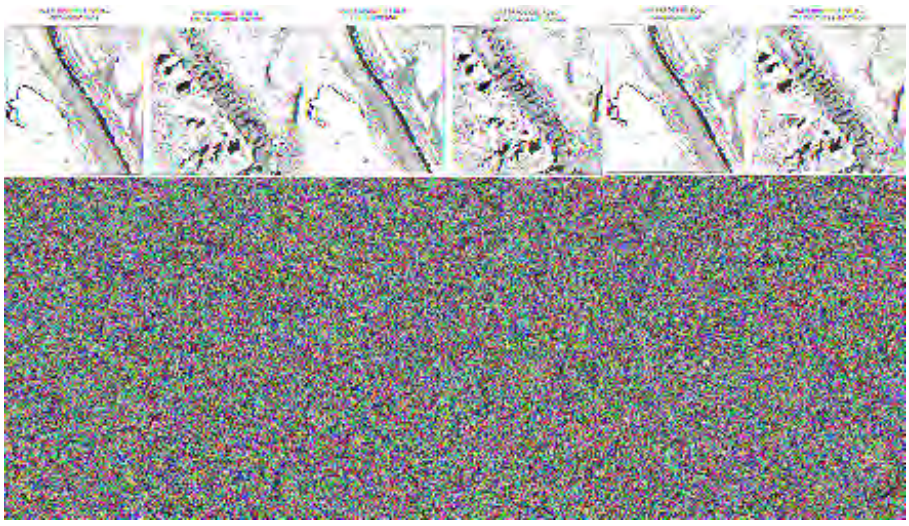
An accuracy assessment of the TM mapping was carried on on classifications using 1) all 7 TM bands, and 2) a 7 PCA image set. The b & W geologic map (page 2-3) was digitized as a comparison base and the classification maps were registered to it. On this basis, the overall accuracy for the 7 band classification was 69.4% and the PCA classification 85.8%. Maximum accuracies were 97, 94, 88, 86, and 85 percents for the Carmel, Moenkopi, Kayenta, Lower Chinle, and Navajo formations respectively, in the TM data set. The Masuk and Tununk Formations had low accuracies in the 20% range.

2-9: **Compare the summer IDRISI to the summer IDIMS classifications; why does the second (lower) IDIMS classification seem less accurate than the one above it; account for the black patterns in this second version. ANSWER**

The illustration below is a classification of a Landsat scene acquired during the winter of 1985. Results are less positive, although many of the units are where they should be. The various formations within the monoclinal limb are discernible but with more irregular outcrop patterns and they include more patches of other units. Note the black patches which are shadows in erosional alcoves. The Moenkopi is not well subdivided into upper and lower areas in this winter version. The light grayish-tan patches along the left side of the image mark snow cover.



Here is a map of the winter image using the legend units symbols (page 2-3) next to the geologic map for this scene. Overall accuracies of 48.5 and 46.1 percent were obtained for the 7 TM and 7 PCA winter maps respectively. One might suppose that the summer imagery is superior in classifying a geologic scene over one in winter. But there can be a tradeoff: in much of the world, summertime means active vegetative cover not present in winter.



On the whole, the Waterpocket Fold has proved an exceptional test site. Its field characteristics are especially suited to enhancements like PCA and ratioing. The abundant exposures of rocks rather than soil and vegetation cover make discriminating them easier and hence improve the classification's accuracy over most geological units, particularly those in areas where the units are hidden by weathering products and organic growth. As we have always known for aerial photographs, space imagery has limited value for direct mapping in vegetated terrains, so that field observations requiring on-site inspection and even digging remain the traditional way to map rock units.

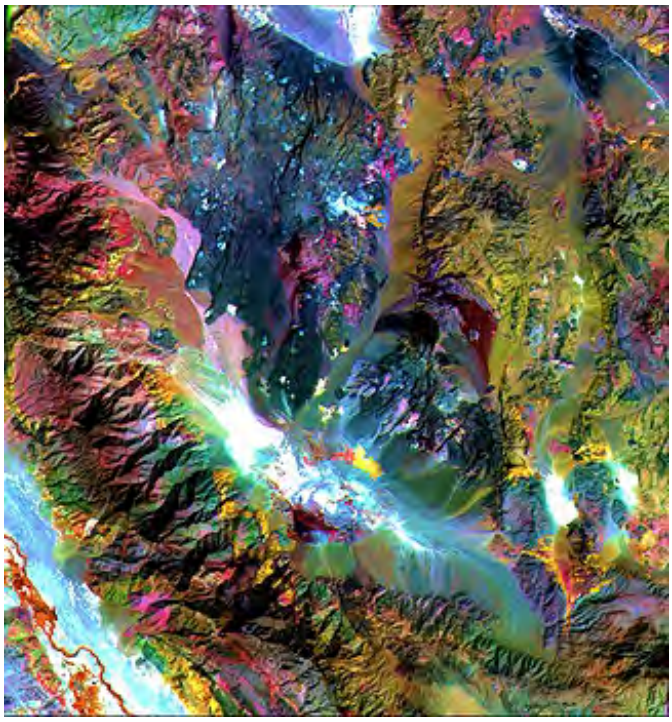
2-10: **Why aren't rock units (formations) as easy to map in the eastern U.S. as in parts of the West; what might cause mapping problems in desert areas such as in Utah?**

ANSWER

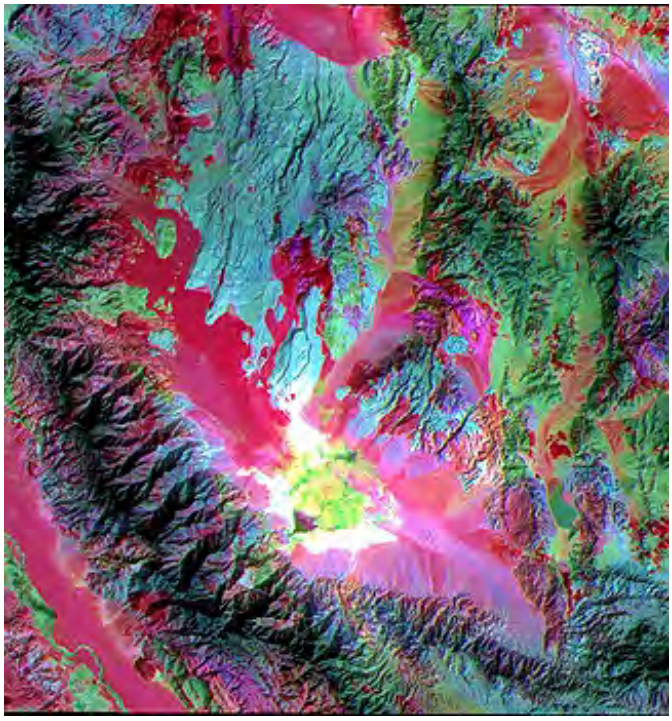
After this Section was prepared, a new set of very impressive images that deal with rock type identification were released from the on-going Terra satellite program which is the kingpin of the Earth Systems Enterprise (see page 16-6). Although the main objectives of the ESE international effort is to get co-ordinated information about the atmosphere-oceans-biosphere interactions, with geologic studies remaining subordinate, several of Terra's sensors are proving very sensitive to certain characteristics of the land surface that translate into still another approach to stratigraphy and mineral exploration.

This is supported by three images from the Terra's ASTER sensor, shown below as color composites. The area imaged is the Saline Valley which lies to the west of Death Valley and is hemmed in by the Inyo Mountains to its west and the Eureka Range to its east. At the top is a false color composite (15 m resolution) made from ASTER Bands 3 (0.81 µm), 2 (0.76 µm), and 1 (0.61 µm) registered as red, green, and blue (RGB) respectively. In the image, red is vegetation, snow and sand are white, and rocks are gray, brown, yellow, and blue.

The next ASTER image is made with the SWIR Bands 4 (1.65 µm), 6 (2.205 µm), and 8 (2.23 µm) as RGB. Limestone appears in yellow-greens and rocks altered to kaolinite as one of the clays are purple. Image resolution is 30 m.



ASTER also has several thermal channels. The image below is made from Bands 13 (10.6 µm), 12 (9.1 µm) and 10 (8.3 µm) as RGB. Resolution is 90 m. In this rendition, silica-rich rocks are red, carbonates green, and silica-low (mafic) rocks are purple.



With ASTER and other Terra sensors now on active duty, and hyperspectral sensors (Section 13) becoming spaceborne in the near future, the ability to distinguish (and *identify*) various rock types has become an effective reality.

However, proper processing of the Landsat MSS and TM imagery can often separate similar appearing rock units as seen in natural and false color. In the top image (false color) below is a part of the Arabian Shield (see page 17-3) in which there are several dark rock types, including basalts, that are hard to separate. But when a color ratio image is made, units at C are separated from those at D, and two plutons at E are now emphasized..

Hyperspectral remote sensing, first treated in the Introduction, promises to provide far better capability for identification (classification) of rock lithologies. This will be covered in some detail on pages 13-5 through 13-9.

We now move on from our evaluation of using remote sensing data to distinguish different rock types and where possible assign them to specific stratigraphic units (formations) to a consideration of the structural expressions of geologic units.

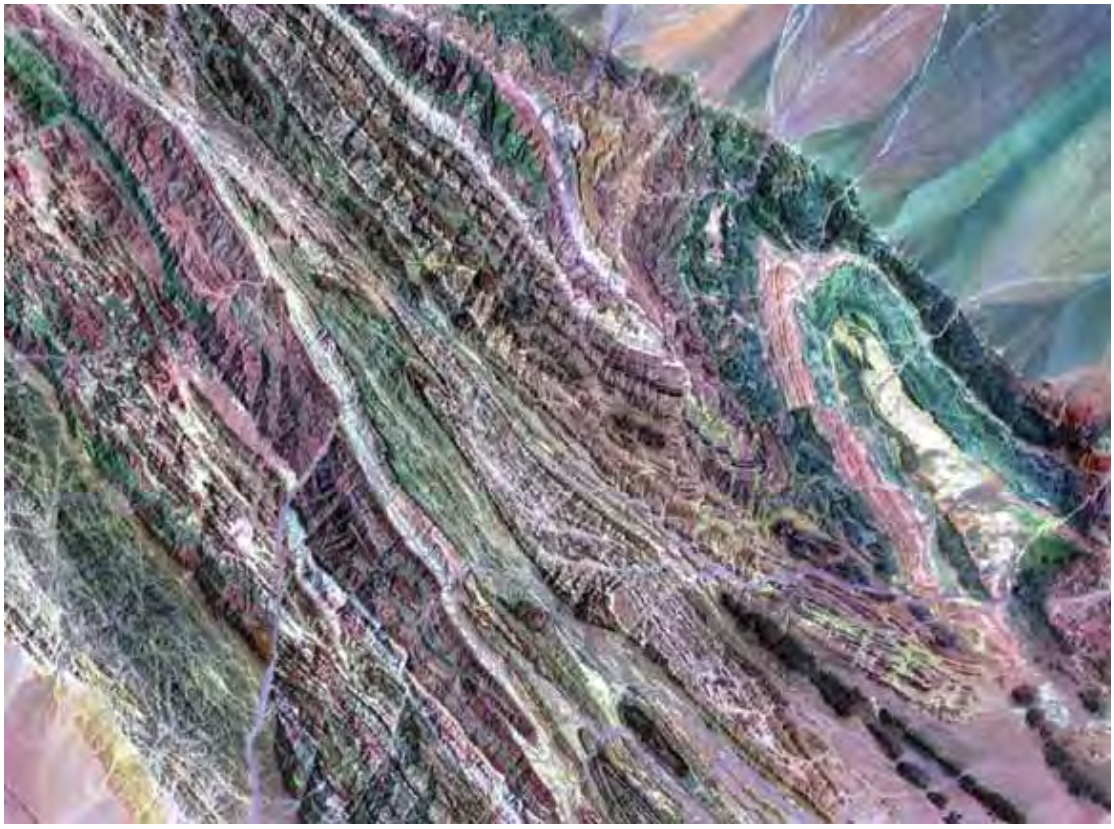
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

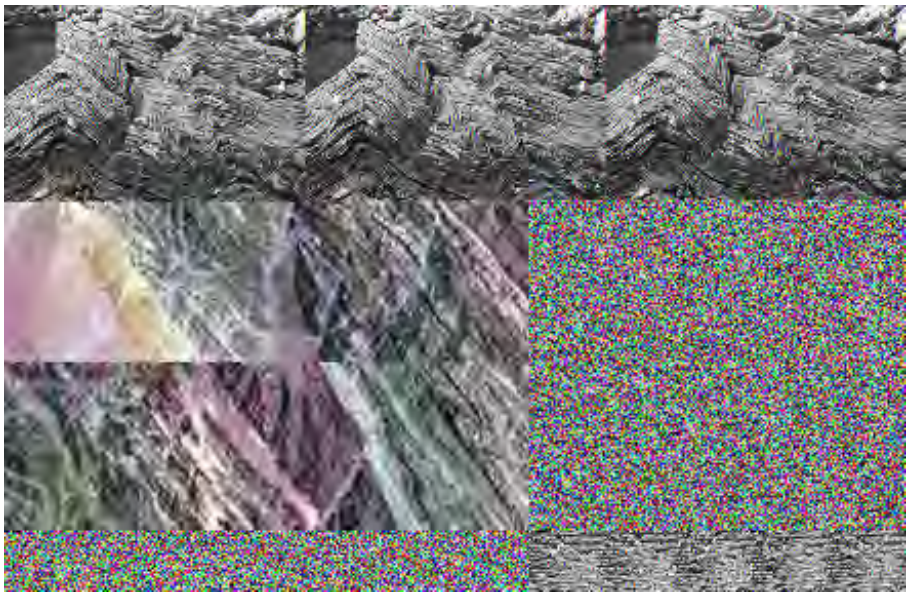
One of the geologic phenomena that shows up strikingly in space imagery is that of the structural deformation of crustal rocks which produces folds and faults and usually accompanies intrusions of magma. Warping or arching of units tend to stand out because of differential erosion which leaves the more resistant layers as ridges rising above the more easily eroded units. Several examples of folds as seen from the ground are compared with typical examples of space-observed folding from above.

Geologic Folds and Intrusions as seen from Space

Space imagery is well-suited to recognizing and interpreting types of distortions of layered strata that produce such geologic structures as folds, faults, and fracture sets (joints). The images can also display the structures and deformation associated with intrusions of magma that dome or raise the near surface rocks. To be recognizable, the strata need to be well-exposed and largely free of vegetation. As an example of readily evident layering in small-scale imagery, look at the strata that are highly tilted from horizontal in a mountain region of Afghanistan:



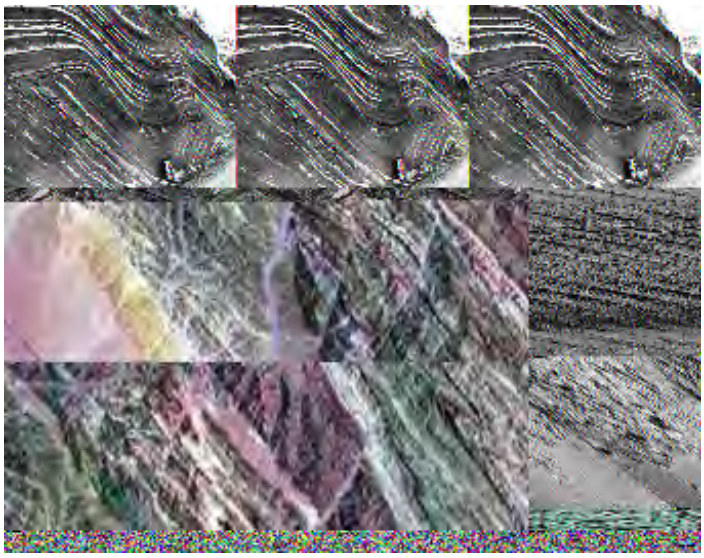
Some tectonic structures are so small that they must be identified on the ground, i.e., they do not show up well from space. As an example of what lies below the resolutions achieved in Landsat/SPOT-type imagery, consider this photo of a small outcrop exposing crumpled shale layers, that form miniature anticlines and synclines (this same pattern occurs on grander scales such as the major folds of the Appalachians).



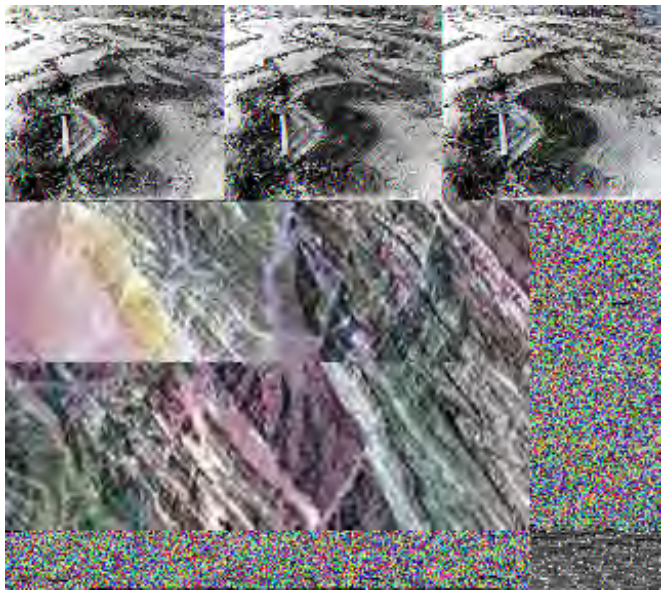
Anticlines are upfolds (arched upwards) whereas synclines are downfolds (shallow- to steep-sided U shape). In a series of folds anticlines are always next to synclines that in turn are next to anticlines (unless disrupted by faulting).

A group of folds, shown here in color, is at about the same scale as the above outcrop.

On a somewhat larger ground scale, look next at the folding shown in this outcrop photo of volcanic ash layers in Japan. There is an erosional discontinuity (unconformity) that separates earlier folding in the lower half from folding (above) after later ash flows were deposited.



The folds shown above occur in supracrustal rocks, those above what is commonly called the "basement" rocks - those that make up the underlying (generally older) core of a continent (in the craton) and are composed of metamorphic rocks mixed with igneous rocks. At one time, before reaching the present surface, these deeply buried rocks were hot and "soft" (plastic-like). Under those conditions, layers in these crystalline rocks are deformed by squeezing and crumpling of heated units under high pressure as metamorphism proceeds. Here is an example of this type of folding (technically, called ptygmatic); note the hammer for scale. This folding is normally not detectable by space sensors (except those with very high spatial resolution).

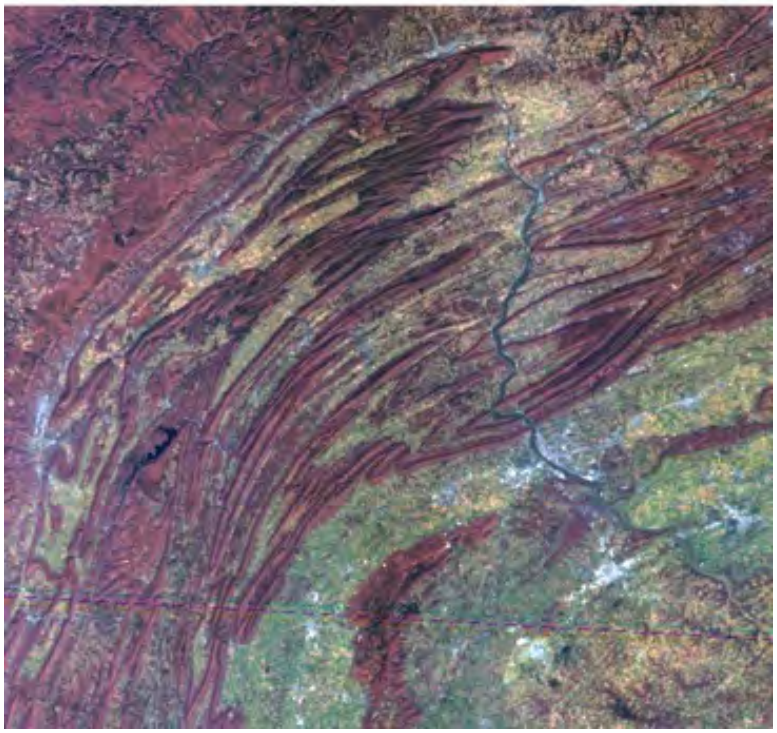


However, larger fold features at regional scales often show obvious patterns of geometric curvature or displacement that stand out in relation (context) to neighboring features, as best displayed in space images covering extensive areas. We used the Waterpocket Fold in the preceding pages of this Section to introduce this idea. Folds that extend over large areas (e.g., a single anticline may be one or more kilometers/miles in width and nearly as high) are quite evident in space imagery. But when seen on the ground, typically only a small part of the arching or downfolding is visible at any local exposure of the folded strata, so that it is usually necessary to measure variations in inclinations (called "dip" by geologists) at separated locations in order to perceive the full nature (as a fold) of such large structures.

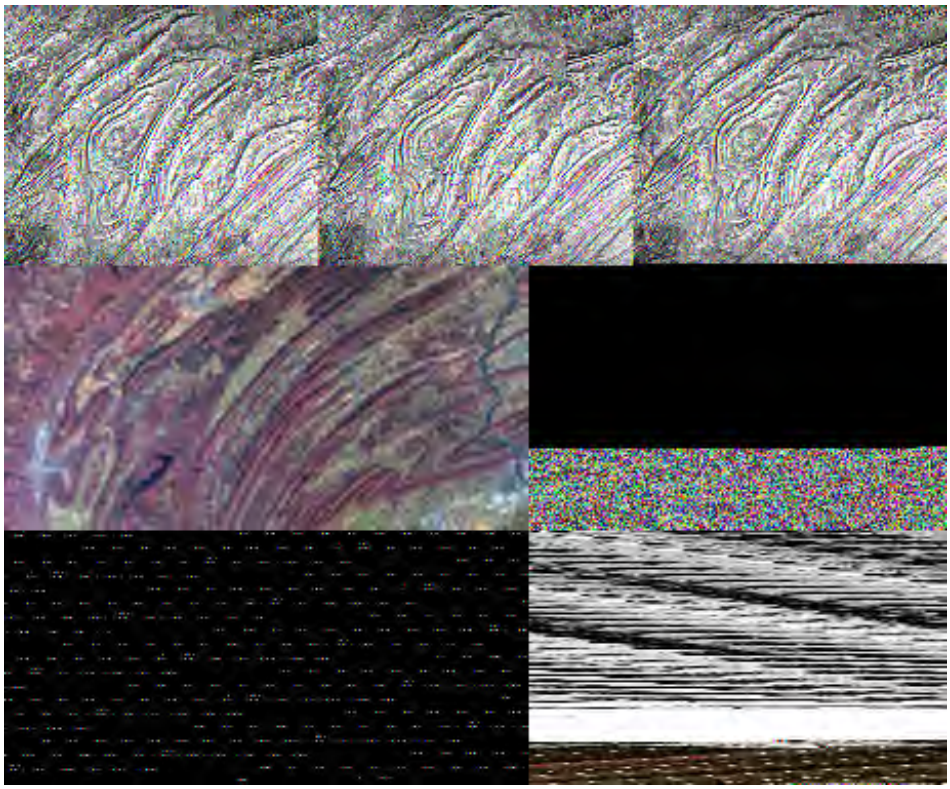
Thus the Maryland Highway synclinal fold shown first on page 2-1 and reproduced here is such a case since it has one inclined limb (its western [left] segment) tilted down to the right (east), then an inflection point (rocks horizontal), and a companion limb inclined to the west. Such folds have limb pairs inclined in opposing directions.

Folds have distinctive appearances when seen from the air or space. Most tend to have maximum arching or downfolding in the center of their structural depressions and then progressively less folding at either end, so that they die out. This diminution is referred to as "pitching". This aerial photo of the Sheep Mountain anticline in Wyoming is illustrative of this effect:

We saw examples of large folds - both anticlines and synclines - in the Section 1 Exam you may have completed. There, the Folded Appalachians were highlighted. These folds tend to be elongated but do end, and are thus referred to as closed (or pitching) when viewed from above. This image shows the extended ridge effect resulting from folds that can remain continuous (without notable pitching) for tens of kilometers or more - but they ultimately die out or fold around:



An obvious example of folds that either pitch or extend for some distance are these next group found in the Ouachita Mountains of Oklahoma-Arkansas (the full scene is shown near the bottom of page 6-3). The Ouachitas are an extension of the Southern Appalachians (the fold belt in between is buried under the Mississippi Embayment). The folds range here from those nearly circular to elongate folds plunging in two directions (causing a curvature at each end; thus closing) to areas where the folded units remain parallel for a considerable distance.

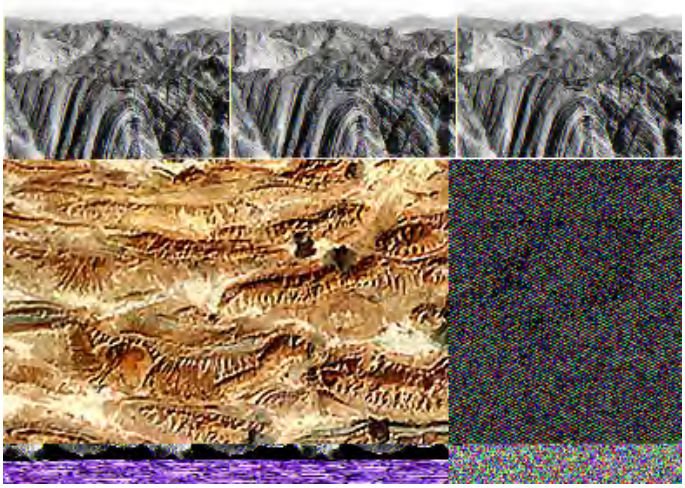


A similar style of folding is displayed in the Sierra Madre Orientale (fold belt) near Monterrey, Mexico, as shown in these two Landsat images. Some structural geologists consider this belt to be an extension of Appalachian folding, but inclined rocks of younger age (early Mesozoic) are present here.

Next, we take a quick look at classic folds in Iran and northwest Africa. Elsewhere in this Tutorial (e.g., in Section 8 on Radar; Section 17, page 17-3), we describe other examples of folding depicted at regional scales.

The full Landsat scene below covers part of the Zagros Mountains along the southwest coast of Iran by the Persian Gulf. These mountains consist mainly of elongate folds which arch upwards as anticlines and downwards as synclines. The anticlines here make up distinct landforms as high hills with central ridges that taper at either end (a condition referred to as a closed fold [one that pitches]).

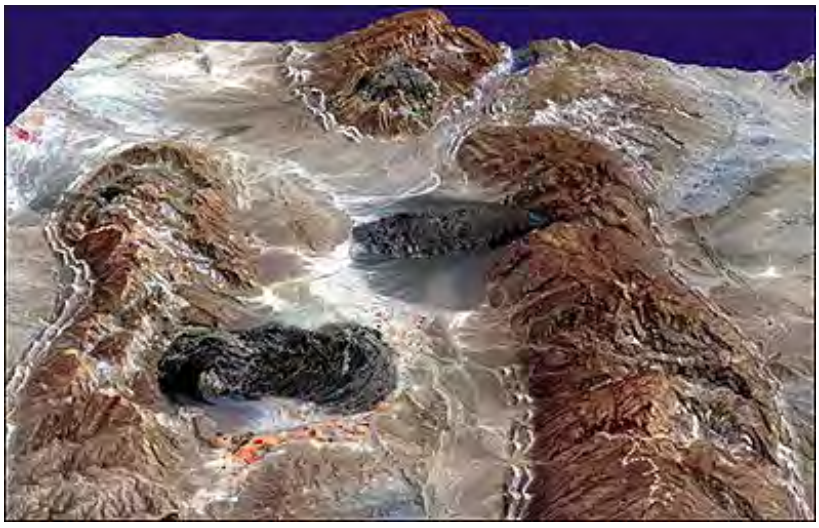
A simple analogy is to imagine cutting a watermelon in half through its longest dimension and laying the flat side on the floor. From above it resembles some of these Zagros anticlines. If we cut through it again across the long dimension at the mid-pointer, the exposed cross-section through green outer skin, white rind, and reddish center would appear similar to the folded strata within the anticline seen here eroded to create a cross-sectional view.



These elliptical anticlinal folds in western Iran comprise a belt that is unsurpassed anywhere else in the world for their symmetry, extent and quality of exposure. Here is a more detailed look using a Landsat-7 ETM+ image:

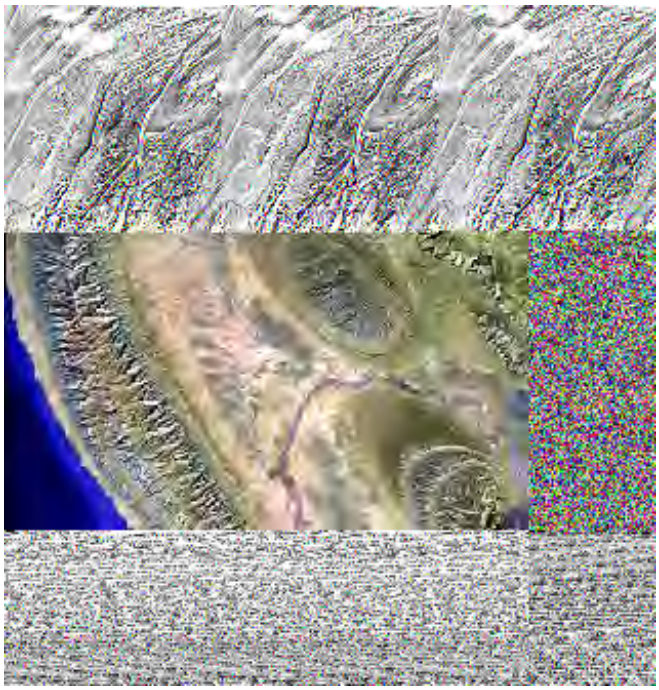
2-11: **Why don't you see the synclines in the Landsat images? ANSWER**

An unusual phenomenon occurs in parts of the Zagros Mountains. Go up to the first image of this region; note patches of very dark gray material. These show up better in this perspective image made from ASTER and DEM data:

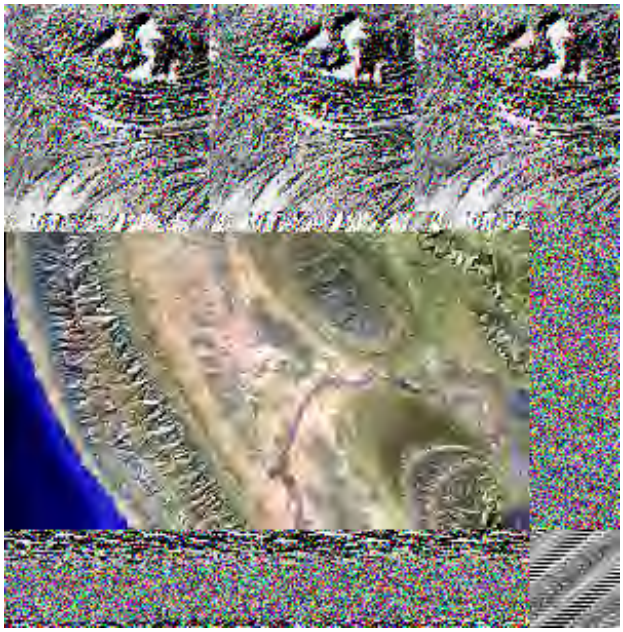


These dark features are outspillings of salt that have been called "salt glaciers". Rock beds composed dominantly of salt (NaCl) can be produced, often in thick layers, in a marine environment in which salinity exceeds a certain value and direct precipitation removes the mineral halite and sometimes other mineral species that make up the class *evaporites*. As salt beds become more deeply buried, the overburden pressure or pressures associated with folding cause the salt material to flow like a very thick liquid under conditions that produce "plasticity". The salt may be pushed upwards, piercing overlying rocks, making salt domes or diapirs (often excellent traps for gas and oil). The salt may reach the surface and "pour out", moving slowly to make the "salt glaciers" observed here.

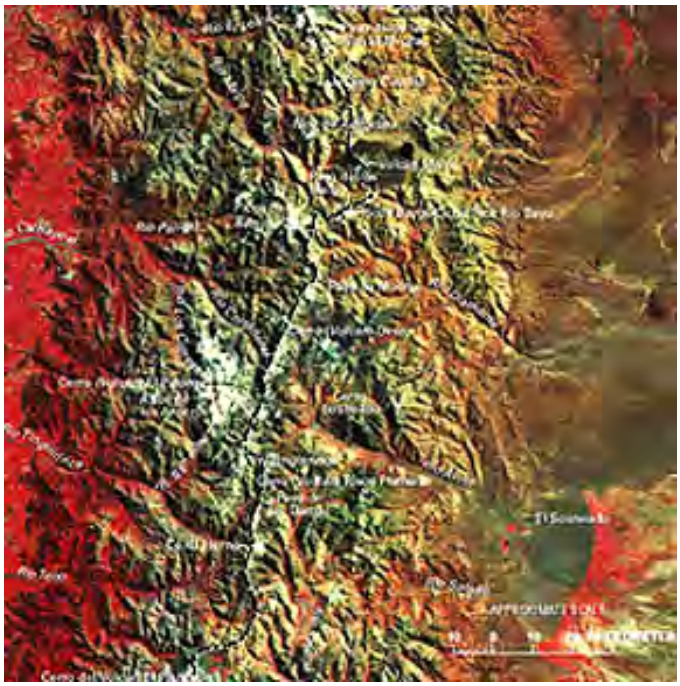
Sometimes anticlinal folds form ridges or linear mountains that have widely separated interfold segments (synclines may not be well-formed). An example is this next style of folding that makes up a *decollement* - a French term that describes the detachment of a sheet of upper crustal rock in which the folds are likened to creases in a slip rug that has been wrinkled. The area shown in this Landsat-1 image is in the Sichuan Basin of central China:



Many folds do not stand out as individuals but are a larger part of continuous folding (and faulting) called orogenic belts or, more commonly, fold belts. Even where carefully mapped, the size of such belts is so great that their overall characteristics are often hard to appreciate. This next image (Landsat-2, Band 6) illustrates again the value of large-scale or large region coverage: The Tapa Shan mountain belt in west-central China seems to have split, with the two western branches swerving south and north; such tectonic behaviour is unusual:

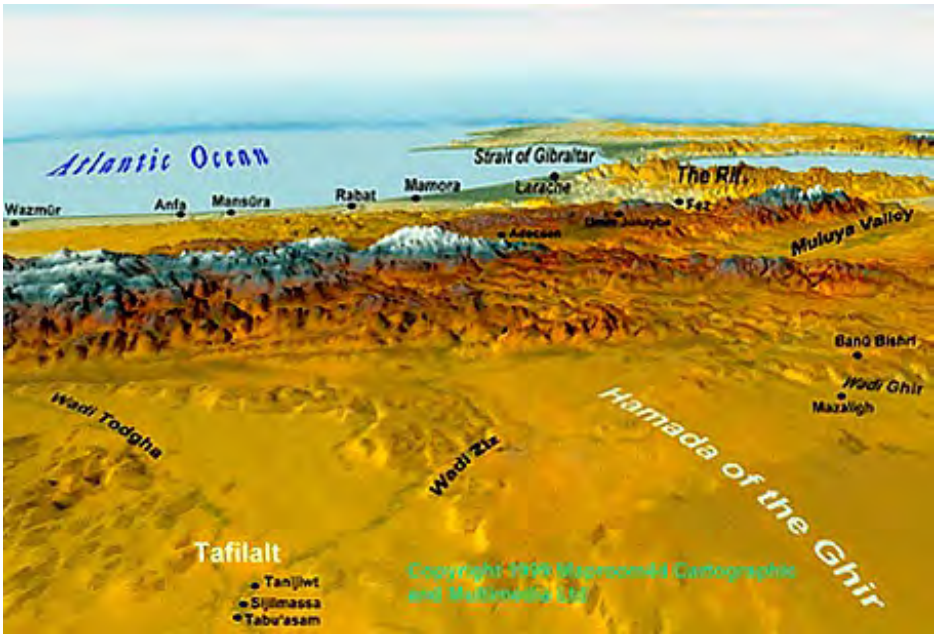


There is a more or less continuous fold belt, of varying widths, running from Alaska to the tip of South America. The term "Cordillera" can be applied to this general mountain trend. The belt system is formed along the zone of convergence in which the Pacific tectonic plate is subducted near the western edges of the North and South American plates, causing sedimentary rocks along the edge of these plates to crumple (fold) and be lifted up into mountain chains. In South America, the belt makes up the Andes Mountains. In this Landsat image, near Santiago, Chile, the Andes belt has narrowed to less than 120 km, with high plains on either side.



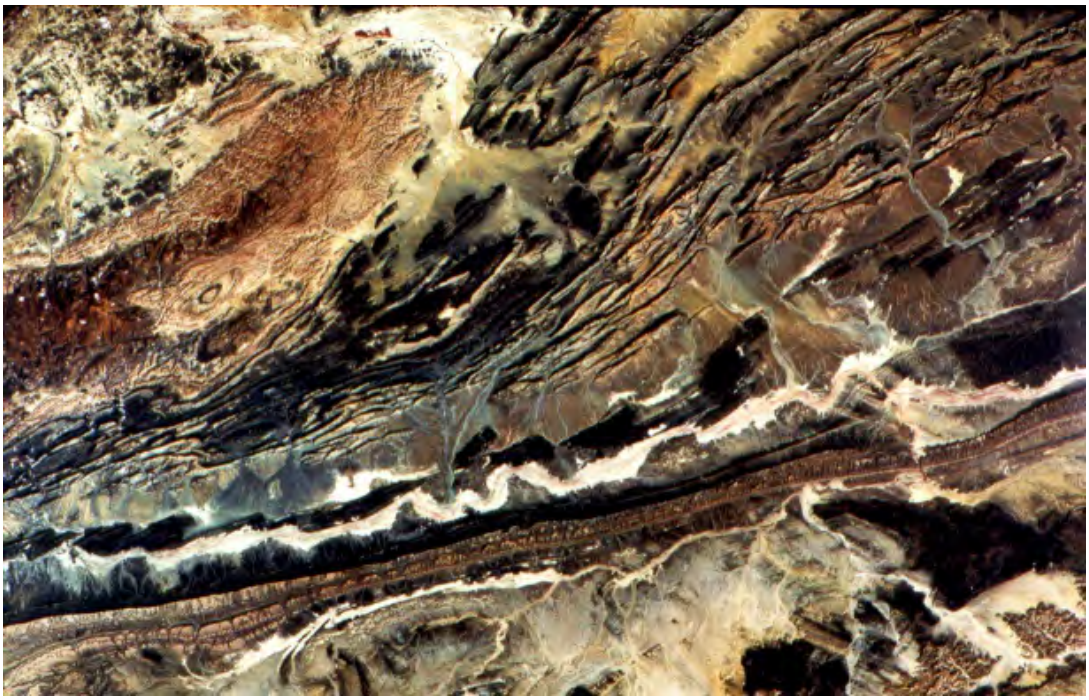
One of the best exposures of a complexly folded mountain belt anywhere occurs in the Atlas Mountain system of northwest Africa. This group is part of the great orogenic belt that includes the Alps, Appenines, the Betic Cordillera (southern Spain), and other chains that we can trace eastward through Turkey into the Zagros Mountains. These belts began to form about 70 million years ago, when the Tethys Ocean (precursor to today's Mediterranean) started to close as the African Plate moved northward against the Eurasian set of plates. The orogeny climaxed in the late Cenozoic period and is still active.

This perspective sketch map of northwest Africa shows the Atlas Mountains in context with the coast of Morocco and the extension into Algeria. Beyond the left end of the map lies the Anti-Atlas segment of this belt. The map includes the Middle Atlas and High Atlas segments, and the Rif mountains east of Tangiers. The mountain chain cuts out moisture coming in from the Atlantic such that this orogenic effect produces the Sahara Desert to the east of these high ranges (much like the coastal mountains, Sierra Nevada, and Cascades do in the United States; rain on the west, rain shadow [dry] on the east).



The Landsat scene below covers part of the Anti-Atlas mountains of southern Morocco. In the upper left is a deformed and metamorphosed core of Precambrian rocks against which the tight disharmonic folds of lower to mid-Paleozoic rocks (center) have been shoved northward along

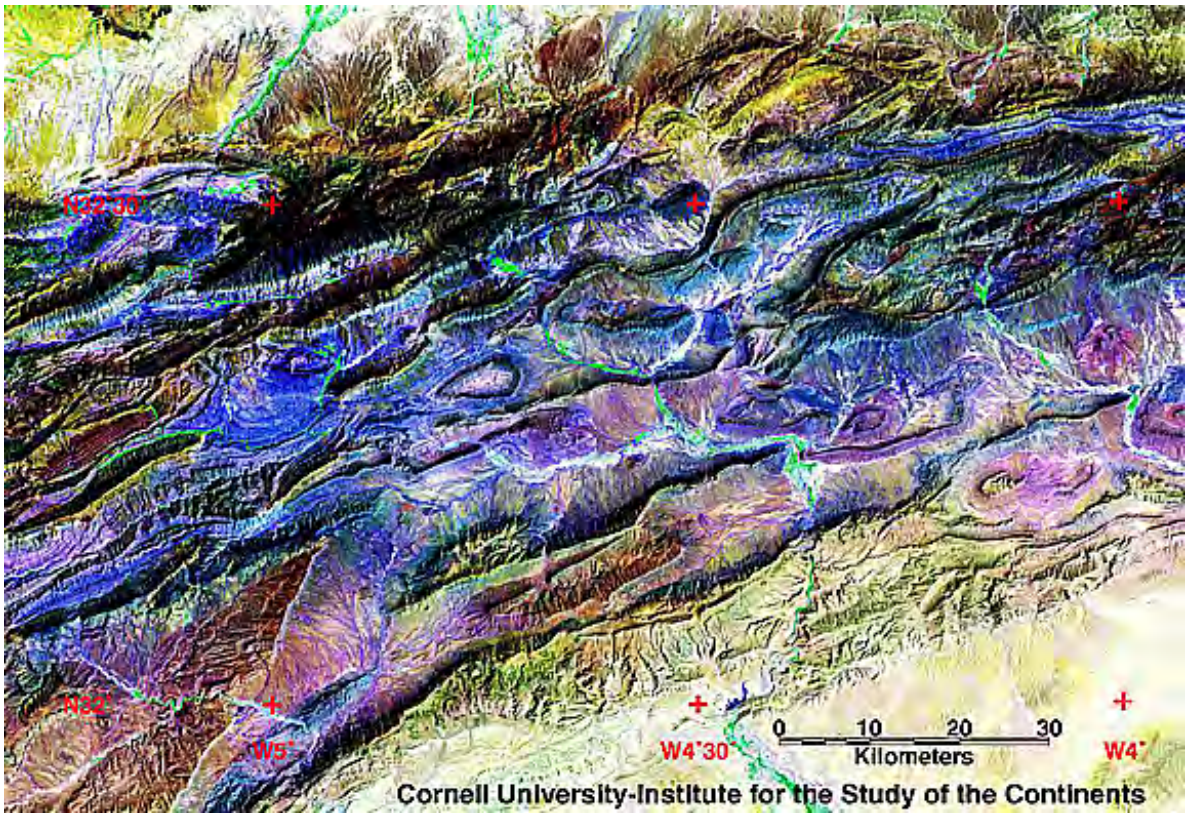
thrusts. The white sinuous band against a fold ridge is a dry stream or wadi.



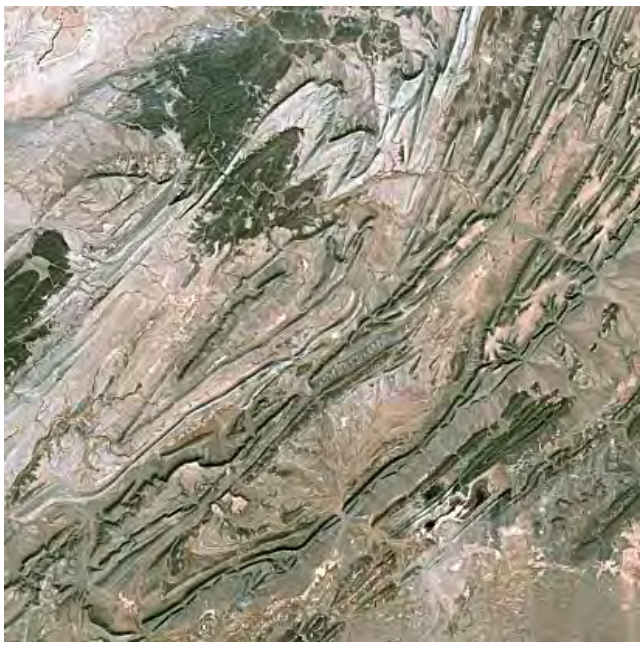
2-12: **What does part of this Moroccan scene remind you of structurally in some other part of the world? ANSWER**

The Atlas mountain belt was also involved in the major closing phase of the Tethys Sea about 80 million years ago, when the African Plate shoved northward, producing the Alps and other European mountains. A portion of the Anti-Atlas is imaged below using three SWIR bands on Terra's ASTER (see Section 9). Again, the folded structures stand out. The color composite shows various colors associated with rock units: yellow, orange, green, and dark blue denote limestone, sandstone, gypsum, and granite respectively. This is another strong confirmation that IR and thermal remote sensing can distinguish and identify major rock types (where vegetation cover is sparse) with considerable validity.

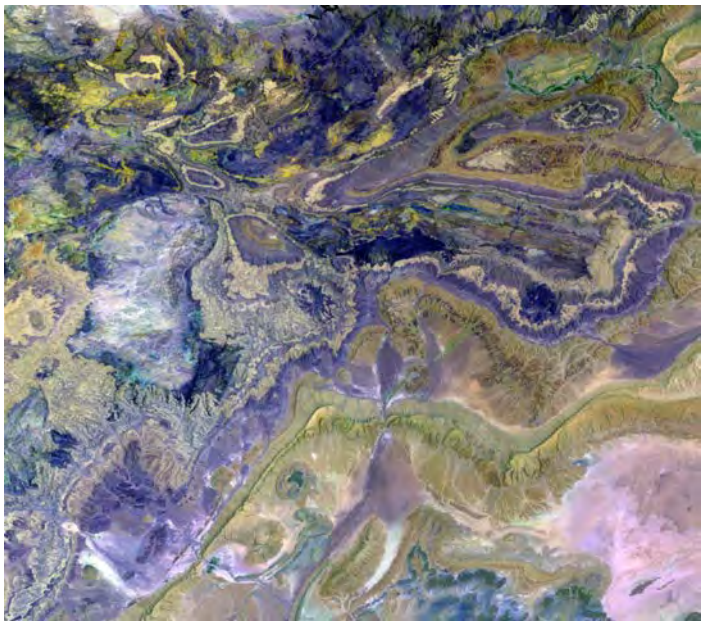
The High Atlas has peaks reaching above 4500 m (13000 ft). Here it is seen from space in a Landsat mosaic:



These are photogenic mountains as seen from space. Witness this natural color view made by the SPOT satellite:



Even more striking is this false color composite Landsat image:



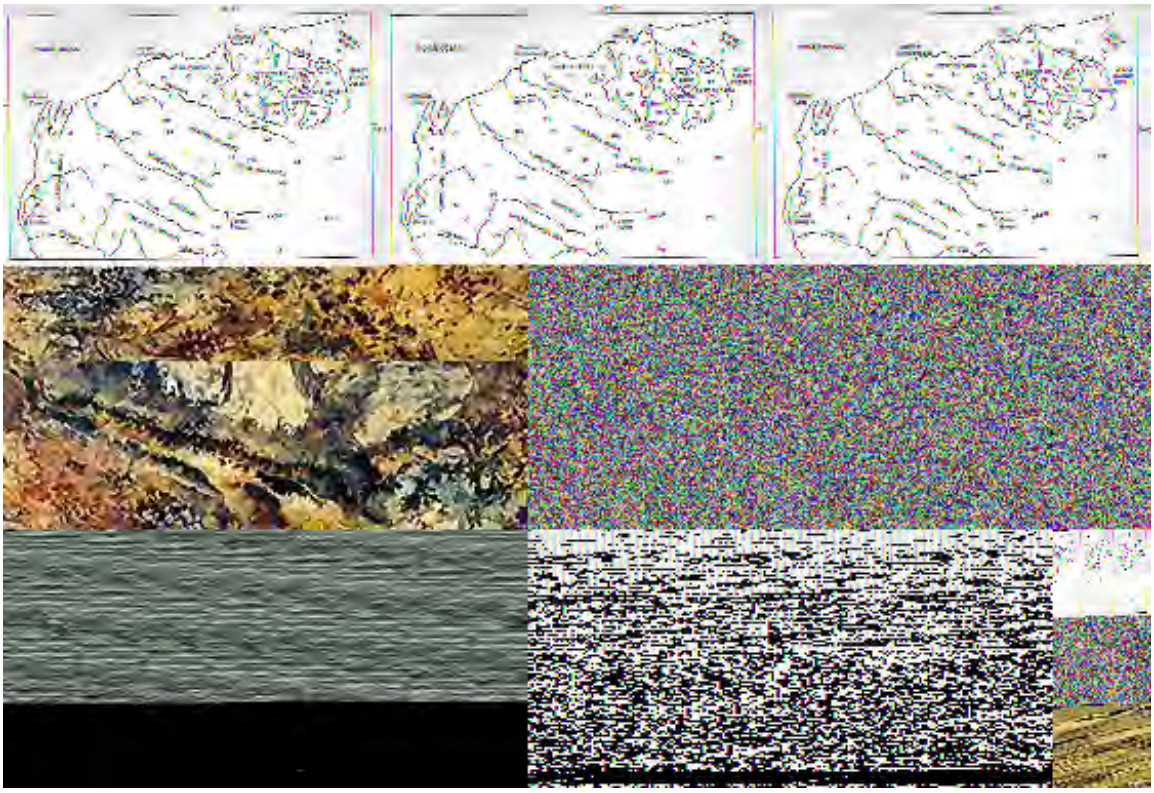
And, seen again, this time in another SPOT image:

These are the highest tectonic mountains in Africa and resemble parts of the Alps except that the vegetation is distinctly different. Here is a view taken from Marracech in the south interior of Morocco that typifies the terrain of the Anti-Atlas; below it is a view within the High Atlas segment:

It may come as a surprise to read this claim that the Anti-Atlas and Atlas Mountains of northwestern Africa are geologically tied to the Appalachian Mountains of North America. When Africa split from North and South America following the closing of the Iapetus Ocean, part of the rumpled collision area resulting from closure was detached onto the African plate. Here is a map of the Appalachians (shown as elongate black patches) before the Pangaea breakup:

We close this page with a look at intrusions that accompany mountain building and related tectonics. One of the classic regions on any continent is the Pilbara craton in northwestern Australia. We will introduce this region to you now but will return to look at geologic scenes from selected areas on page 6-15. The remarkable mosaic below covers much of northwestern Australia, a region of limited vegetation so that the rocks and valley-fill stand out and reveal much of their underlying structure. The mosaic has been divided into left (west) and right (east) halves stacked vertically to fit on the page without much reduction. This is the Western Australian shield, containing mostly Precambrian metasedimentary and metavolcanic rocks, interlaced in places by igneous rocks. At the top of the (right half, lower image) mosaic is the Pilbara block, a leading candidate for the classic expression of an ancient greenstone-granite complex anywhere on Earth.

The two halves do not quite match because of copying problems. This annotated sketch map applies to the mosaic.

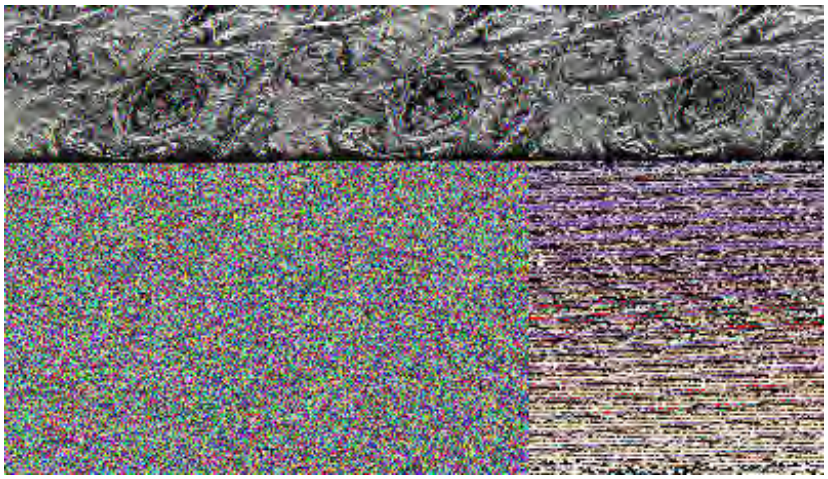


This next image enlarges the area known as the Pilbara block that contains these remarkably exposed (little vegetative cover) batholiths. In this image is perhaps the best single exposure on a denuded surface with little vegetation of the geologic phenomenon know as "batholiths". These are intrusive igneous rock masses - often 100s of kilometers in dimension - that came as melt into the crust but did not reach the surface (until eventually revealed by downward erosion). Between these giant bodies are metamorphosed rocks containing basalts that are now called "greenstones". This is probably the classic example of an Archean Greenstone Complex, and especially of batholiths, anywhere on Earth.

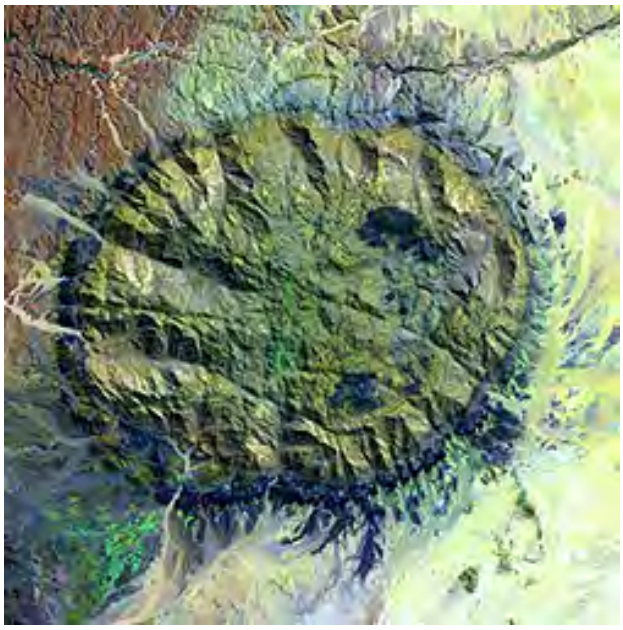
For further guidance, look at the above Pilbara features in relation to the geologic map of this part of Australia that covers the right half of the sketch map.

The granite appears as batholiths, up to a 100 km (62 mi) long. These light rocks are diapiric intrusions into the dark greenstones (metamorphosed basalt). To the south is the Hamersley Range (blue area on the map) and the smaller Opthalmia Range (red), bordered on the south by the Ashburton Trough (left) and the Bangemall basin (right). Low-relief hills mark much of the region. The highest area (1,235 m, 4,051 ft) is in the Hamersley Range.

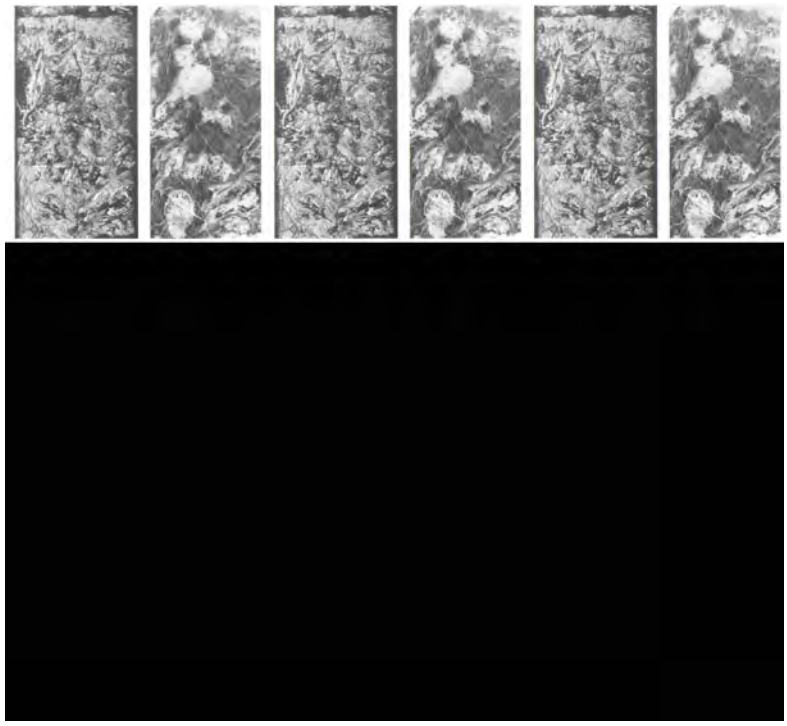
We witness the radar expression of folding with this look at a SIR-A image of closed structures composed of metasedimentary rocks in the Hamersley Range of northwest Australia.



In Namibia in southwest Africa is the Branberg Massif, a 21 by 31 km dome that reaches 2.5 km (1.6 miles) in height. A granite batholith intruded into the sedimentary rocks along its side.



An extension of the Arabian Shield is found in eastern Egypt. The left image below is a radar image of its surface. The same area is shown in a Landsat image on the right. The radar image brings out the fractures and the continuous nature of the crystalline surface. The Landsat image shows this terrain to contain intrusive plutons midst metamorphic rocks.



On the next page, we look at illustrations of how we can detect faulting from ground offsets and, more subtly, from discontinuities of landforms. Following that, we illustrate the advantages of space imagery in picking out lineaments (usually fractures in the outer crust that may be faults), together with an appraisal of how these can often be misidentified and misinterpreted. We then close this Section with a practical example of how fracture analysis leads to a successful search for a valuable natural commodity water.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Another regional scale feature usually easy to recognize in suitable space imagery is that of faulting. A fault is a fracture in the Earth's crust along which there has been some relative movement of rock on one side against the other side. This appears as some form of displacement or offset of once contiguous units.

Recognition of Faults

When earth materials within bodies (such as blocks or layers within the crust) are subjected to forces (or stresses, which are forces applied to finite areas), they tend to fail - either by fracture or by plastic flow. Fracture is a general term but it has a second meaning - breaking along irregular, often curved, surfaces. Breaks that tend to create planar failure surfaces are called faults or joints. Most earthquakes result when the blocks on each side of the fault suddenly slip against each other and move (are displaced).

Any fracture that lies beneath the surface and also intersects that surface will produce a line or linear trace (in geometry, the intersection of two planes is always a line).

Because of their role as the prime cause of earthquakes, faults - and their detection (as from space) - are of special interest and concern. Faults are fractures along which there is *relative sliding movement* (offset) of the blocks in opposite directions on either side. We recognize them by various criteria:

1) Layers of different types and ages of rock units sit side-by-side (offset)

2) Abrupt topographic discontinuities of landforms

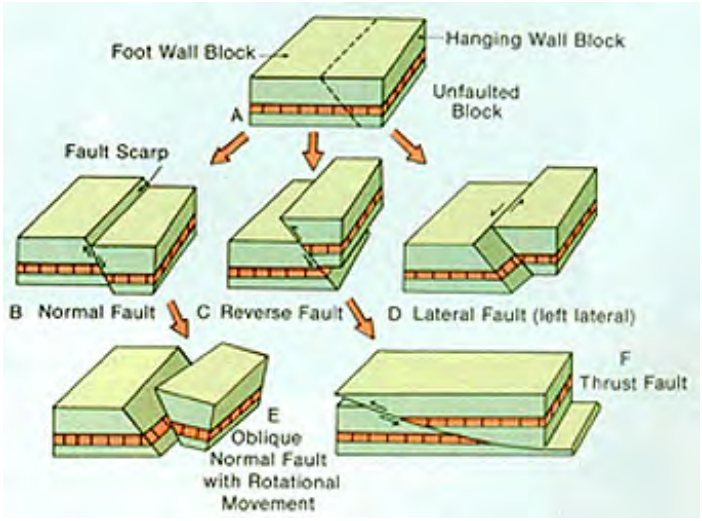
3) Depressions along the fault trace (broken rock is more easily eroded)

4) Scarps or cliffs

5) Sudden shifts of drainage courses.

6) Abrupt changes in vegetation patterns

The relative movement direction along the plane of fracture (fault plane) and the angle of that plane determine the type of fault that is involved. Nomenclature of fault types is shown here:



From Contemporary Physical Geology, H.L. Levin, Saunders Publ.

For normal faults (caused by tensional forces), the hanging wall block moves down the fault plane, so that marker beds within the block drop downward relative to their once contiguous counterparts in the footwall. Reverse and thrust faults (compressional forces in the crust) involve upward hanging wall movements; thrust faults also carry their blocks outward, overriding the rocks on the footwall side. Strike-slip faults (shear forces) result in largely horizontal movements along a steep to vertical fault plane.

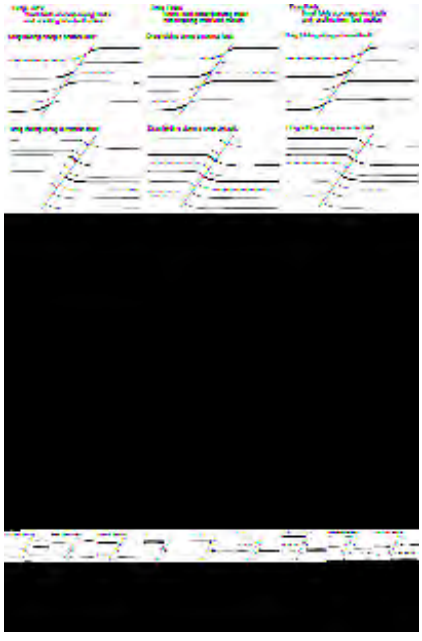
A ground example of each of the four most common types follows. First, a normal fault:

This is a small reverse fault in a roadcut in Japan:

And then, a thrust fault:

And last, a strike-slip fault.

If the displacement has only been a few meters or so, the sense of movement for normal and reverse faults is possible, where seen in outcrops, by matching the patterns of bedding. If the displacement is large (tens to hundreds of meters), chances are the outcrop is too limited in size to show the now separated beds. Often, the sense of movement can be gained another way: as the blocks slide past each other, the layers on either side are bent as "drag folds". This diagram depicts the expected bending geometry:



As a general rule, strike-slip and thrust faults - which usually involve considerable displacements - are often easy to pick out in space imagery (assuming little vegetation or soil cover). Normal, and especially reverse, faults are much less frequently visible.

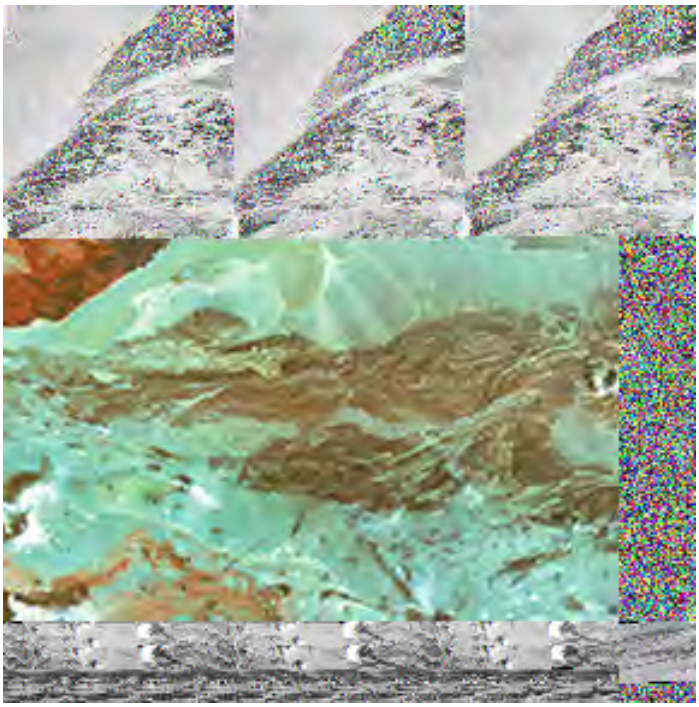
Strike-slip (wrench) faults are usually the best displayed when viewed from above (space or aerial) because they can cause lateral shifts of initially continuous rock units to move cumulatively over millions of years for many kilometers. This brings into juxtaposition units that generally are quite different as now positioned. Thus, abrupt change in lithology across the fault plane at any location along the fault are the hallmark of this kind of displacement.

The discontinuity owing to offset of once continuous layers or other rock units along the strike-slip plane is often remarkably displayed in a space image. The scene below of the Kuruk Tagh fault in part of the Tian Shan mountains of westernmost China shows a sharp fault trace running east-west through the Kuruk Tagh mountains. It's composed of folded sedimentary strata, metamorphic rocks, and igneous intrusions. The block of crust on the north side (top of image) has shifted sub-horizontally to the west at least 60 km relative to the block containing the corresponding segment of mountains to the south. This type is a left-lateral wrench fault (also called a strike-slip fault). (Left lateral means that the block on the far side of the fault plane moves left relative to an observer standing on the near side; right lateral describes a movement to the right.)

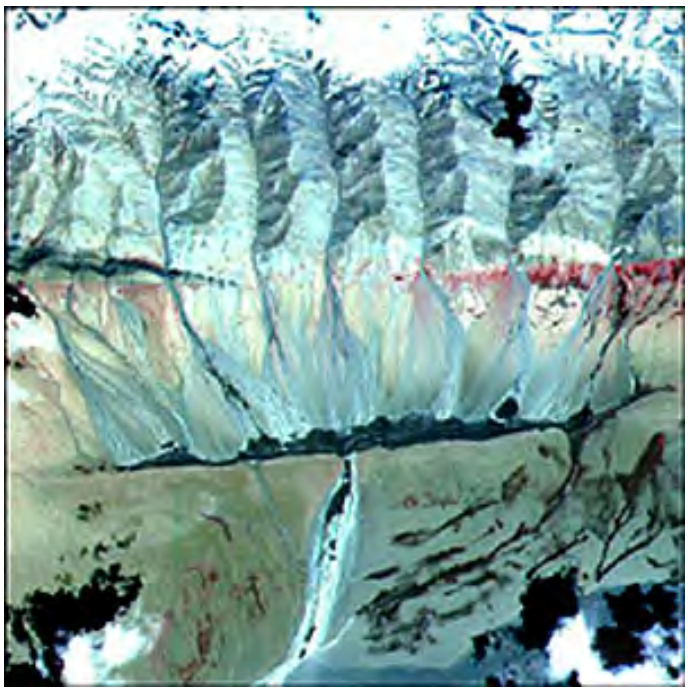
The Kuruk Tagh fault is similar to the San Andreas fault, which is a right-lateral fault, with the Pacific plate moving northward against the North American plate (see below). The Kuruk Tagh fault is easy to identify because of topographic offset (as well as equivalent parts of the strata and metamorphosed rock units), fault scarps, and displaced drainage. This and similar major wrench faults in south-central Asia represent crustal adjustments to the stresses induced by the collision of India against Asia (see mosaic in Section 7).

2-13: **Can you find a second major fault in this scene? ANSWER**

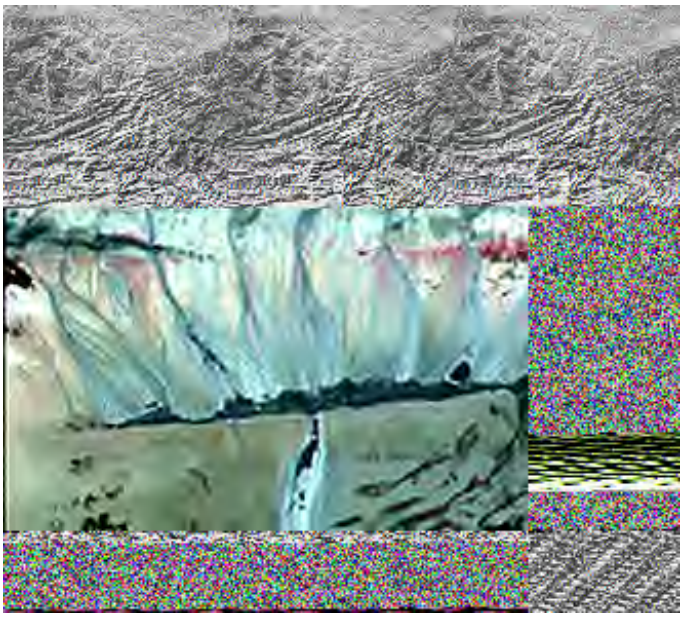
Lateral displacement (one mode of offset) is also evident in this Landsat image of the Altyn Tagh fault, also in western China near Tibet and the Kunlun Shan (mountains):



Another major fault zone in western China near Tibet is the Kunlun strike-slip (left lateral) system. In the ASTER image below, the fault has split into two parallel segments. The lower one has produced a topographic barrier against which water (black) has been impounded to form a long, narrow lake. Above a series of alluvial fans is the upper segment from within which water has emerged to flow downslope and to enable vegetation (red) to grow.



Discontinuities are revealed at space observation scales when a sequence of layers abuts abruptly against another sequence trending in a different direction. This is evident in this Landsat subscene in western China:

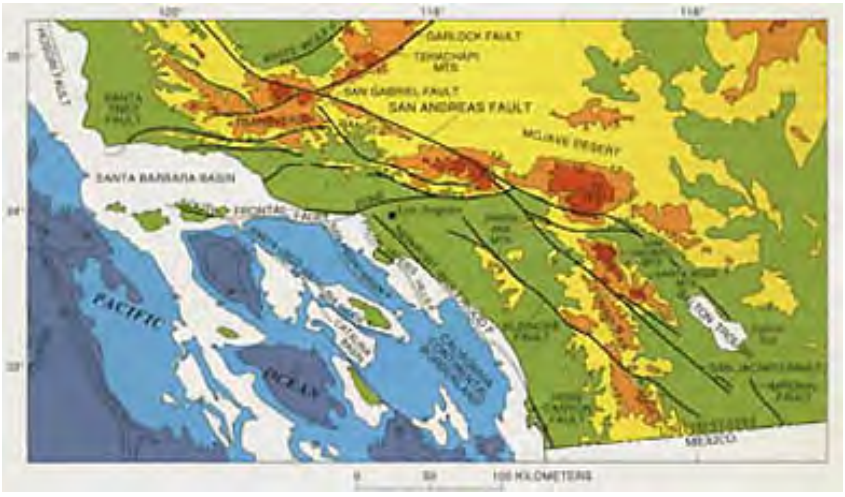


Comparable to the Chinese faults is the Dead Sea Fault that runs from just below the mountains of east Lebanon southward through the Sea of Galilee and the Dead Sea (both actually lakes), thence into the Gulf of Aqaba. This right lateral fault marks one of the three arms of the Afar Triple Junction (see page 17-3). Here it is shown in a mosaic made from two Landsat images.

The fault is recognized in part by topographic discontinuities (mountains in Jordan not fitting with those on the western side) and because the Jordan River and the two lakes follow zones of weakness that are erodes so that they are lower than their surroundings (land adjacent to the Dead Sea has the lowest elevation on land anywhere on Earth; in places below - 300 feet).

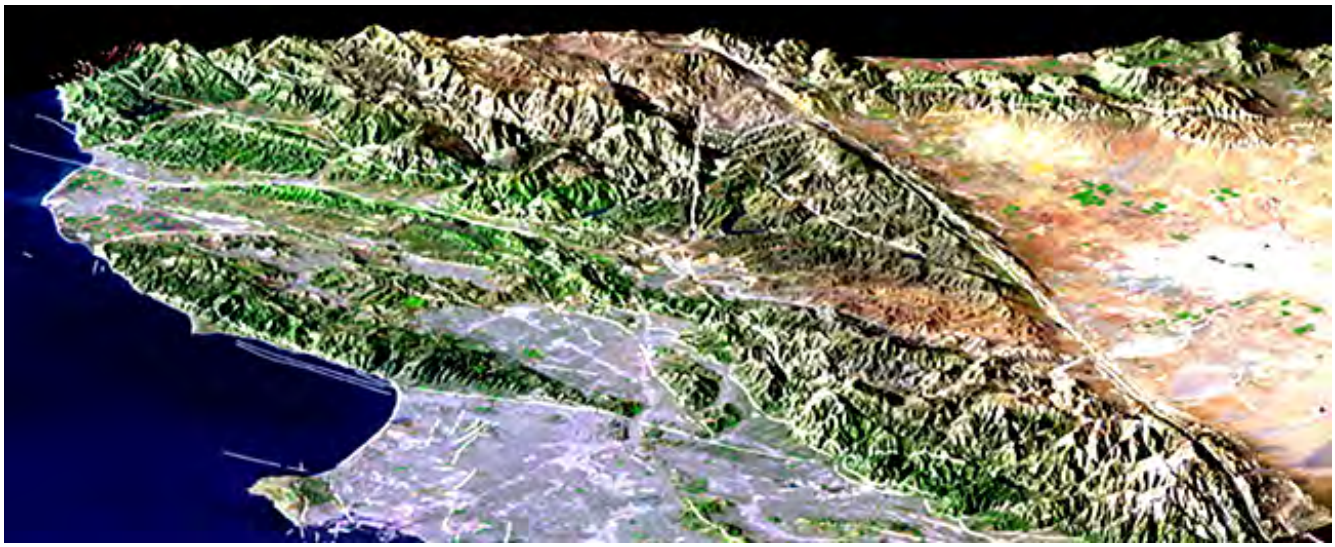
Now, take a look at this Landsat image and 1) try to identify its location (w/o peeking at the caption), and 2) see what evidence you find for possible faulting.

This region in the U.S. famed for its propensity to experience frequent earthquakes is much of California from the Mexican border to about 100 km (62 miles) north of San Francisco. North of the Los Angeles Basin is a series of mountain ranges trending towards an east-west orientation. These collectively are known as the Transverse Ranges, and include the San Gabriel and Santa Monica Mountains north of Los Angeles. They are bounded by faults that have a strong effect on their topography. These faults appear in this physiographic map of southern California:

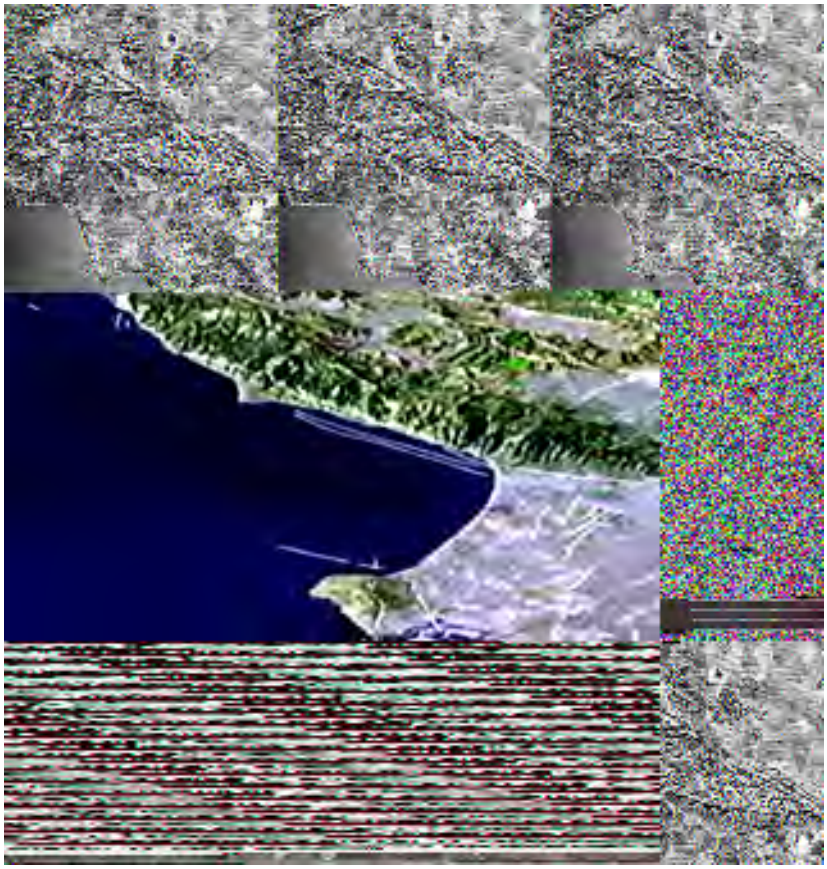


Best known is the famed San Andreas Fault, the earthquake-maker running from the Gulf of California through the California Coastal Ranges north of San Francisco (see the Los Angeles mosaic further down this page and the L.A. image in Section 4 and the San Francisco images in Section 6).

Although most of these western California faults are wrench or strike-slip types (dominantly horizontal slip motion), they are capable of influencing the fronts of ranges in a manner similar to the normal-type of fault. Here is a perspective view of the Transverse Ranges made by combining Landsat imagery with topographic data acquired by JPL's SRTM mission (radar altimetry). Fault lines (in white) have been drawn on the resulting image, which also depicts the Los Angeles Basin (bottom center) and the Mojave Desert (right).

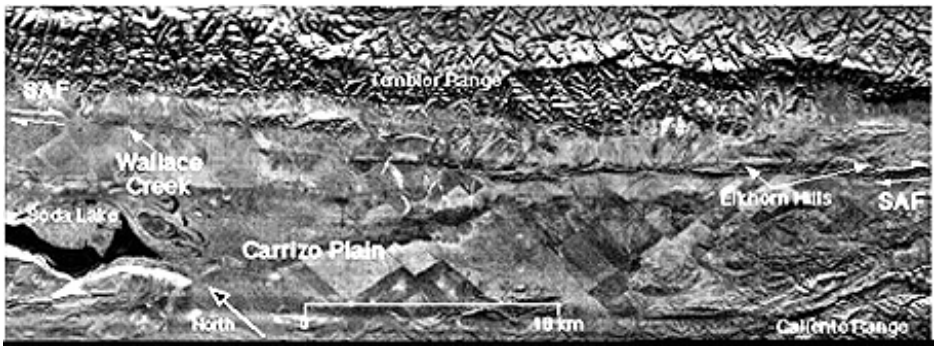


No doubt the most famous fault in North America is the San Andreas, a major strike-slip type, that runs from the Gulf of California northward more or less parallel to the California coastline until it finally passes out to sea as a transverse fault in Bodega Bay north of San Francisco. In the ranges north of Los Angeles the San Andreas marks a prominent straight boundary with the southern Mojave Desert. This is evident in this image which is actually a beautifully produced aerial photomosaic that includes the Los Angeles Basin, with its many cities, to the south. In the mosaic, note the arcuate fault trace which is the topographic expression (makes a narrow valley) of the San Gabriel fault.

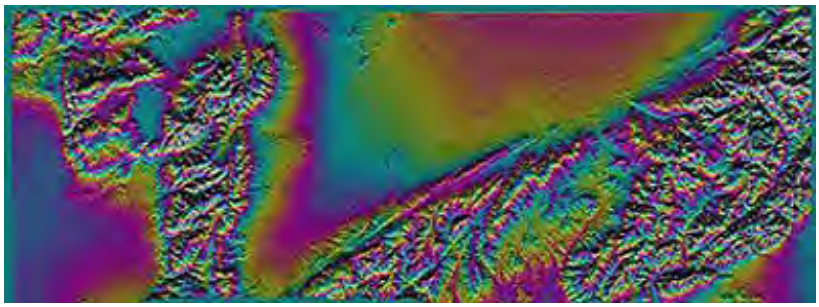


An aerial oblique photo shows the fault in an area of the Coast Ranges in the Carizzo Plains northeast of Morro Bay.

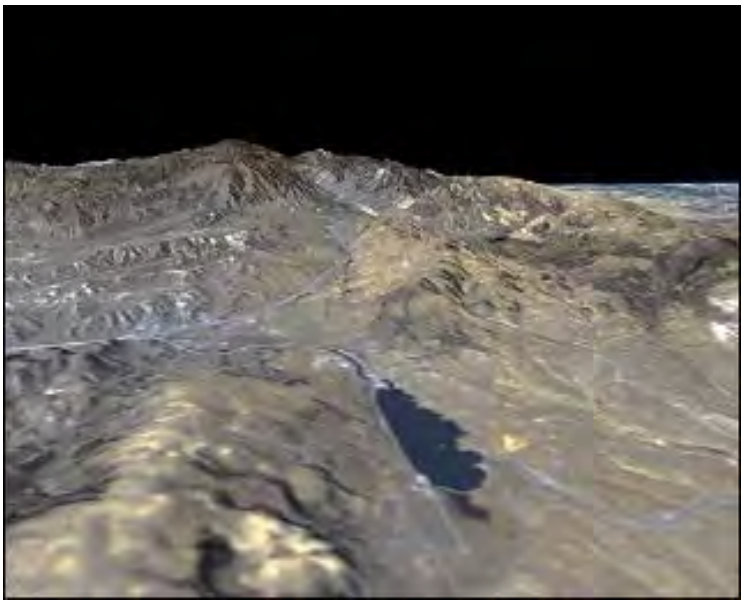
Segments of the San Andreas have been imaged by Landsat, SPOT, and radar systems many times. This is an airborne radar mosaic of the fault in the Carizzo Plains area:



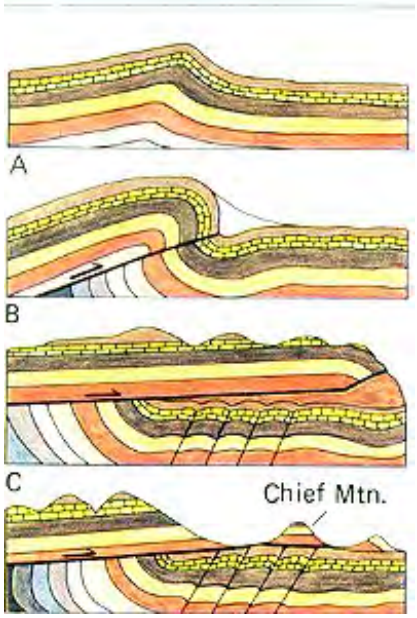
Now we show an unusual portrayal of its appearance in a C-band image specially processed to give landform information, with elevation variation shown as a series of color bands. The imaging instrument is the Shuttle Radar Topography Mission (SRTM) that is discussed more fully in Section 11. Compare this image with the aerial photomosaic shown above: the straight boundary along the northern Tranverse Ranges in both images stands out. The image orientation is shifted somewhat from the mosaic, with the Mojave salient apex now pointing down at the left.



As is described in Section 11 and elsewhere, imagery coupled with elevation data (in STRM's case, from its own stereo-like capability) can be recast in the perspective mode as though it is being viewed much like an aerial oblique photo. Here is an STRM construction of the topography west of Palmdale, Calif. (an earthquake-active area) in which a somewhat straight valley (holding the lake) roughly coincides with the San Andreas fault.



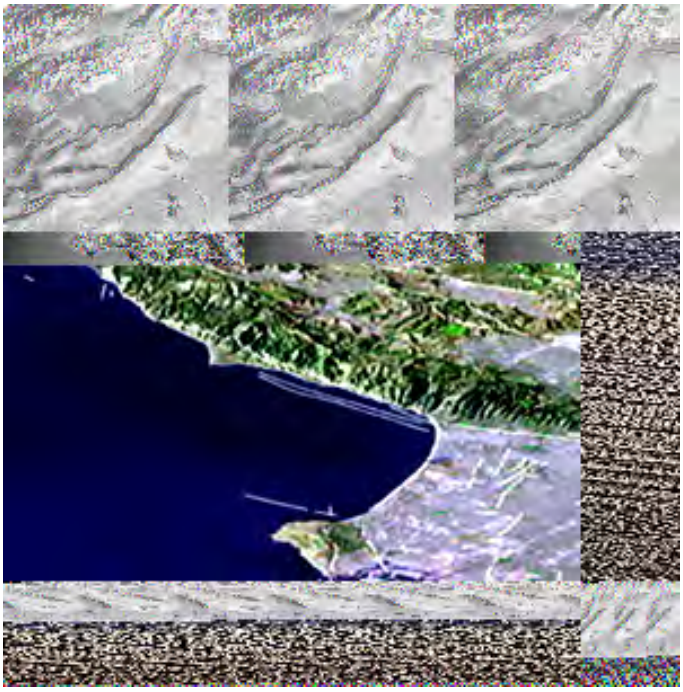
Moving on: A second fault type often well-displayed in space imagery is the thrust fault, in which the fault plane is at low angles relative to Earth's surface and the usual direction of movement carries the upper (near-surface) block over the lower block, causing rocks of different ages to be juxtaposed. Thus, a shallow, tabular slice of crust slides (thrusts) over the fault plane and on top of the surface ahead of it. Thrusting can occur on a large scale when a terrane attempts to "dock" on a continental margin and slides up and over. Another way is for a fold to overturn onto one limb, then break off, and slide over the terrain (landscape) ahead, as illustrated in this diagram"



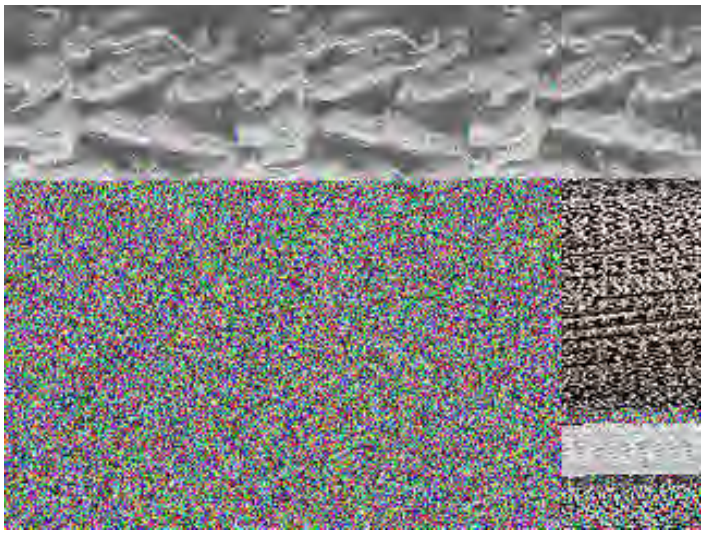
From Contemporary Physical Geology, H.L. Levin, Saunders Publ.

If each block consists of rock types that are different in composition and erosive response, these will appear at the surface as intervals of rock with contrasting topography. When seen in the field, the disparity between rock types can be dramatic. Here is a scene witnessed many times by the writer (NMS) as he flew into Las Vegas from San Francisco enroute to the Nevada Test Site. On top of Mesozoic sandstone (buff-yellow) are much older Paleozoic limestones (left; dark gray); this superposition of old on top of younger defies the Law of Superposition (young over old) unless the overlying older unit was emplaced by thrusting:

We return to western China to see an excellent example of topography related to multiple thrust sheets; fault planes dip to the south, the thrusting is towards the north:

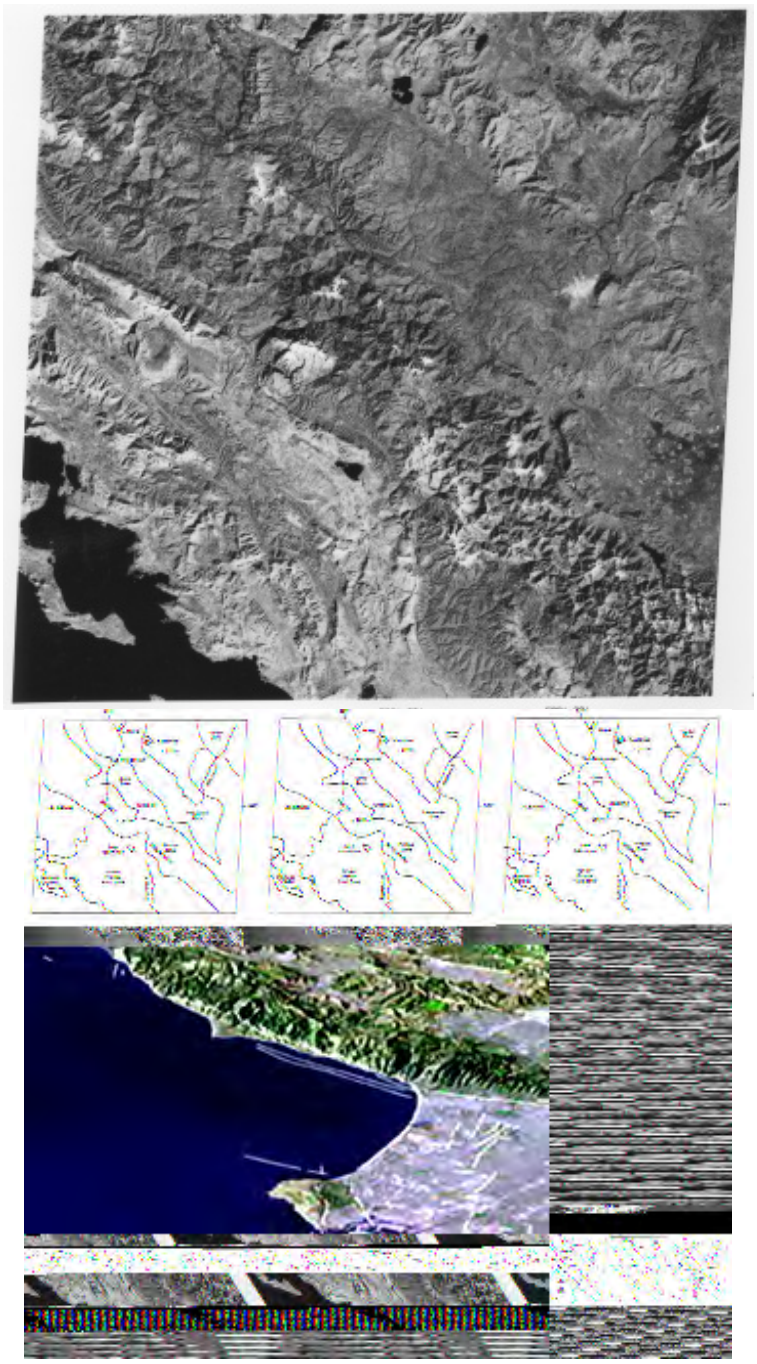


This radar image shows the Kalpin Tagh thrust in China; its boundaries are a pair of wrench (tear) faults on either side.



Several thrust slices (sheets) may stack one on top of the other in a staggered pattern horizontally, leading to a sequence of side-by-side thrust block bands that outcrop in

mountainous terrains. This topography is superbly displayed in the Landsat scene below of the Pindus Mountains of western Greece and Albania.

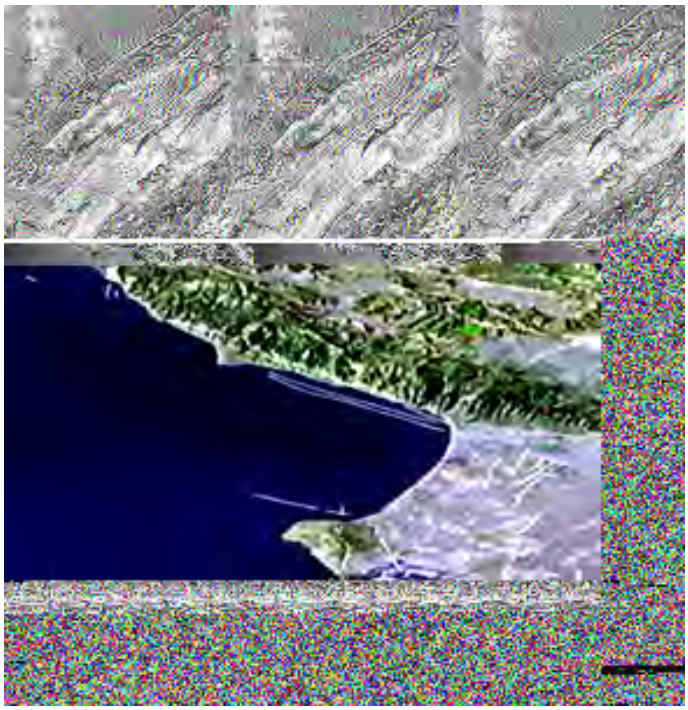


One can differentiate five tectonic zones, named in this generalized map, by sight because of distinct topographic variations and tonal differences related to contrasting rock types. The direction of tectonic transport is from east to west (right to left) causing the sheets to partially overlap below the surface, but each front edge occupies a different geographic position relative to the one it overrode and the one overriding it. This zone of thrust belts is part of the Balkan Alpine system, an offshoot of the European Alps that runs sub-parallel to the Apennine Mountains of Italy (see Alps mosaic in Section 7). As part of the tectonic adjustments caused by the African Plate shoving northward against the European Plate, a small tectonic plate underlying the Tyrrhenian Sea (off Sardinia) is squeezed against the Adriatic plate. The Adriatic plate then pushes it eastward against the Aegian plate, underthrusting it and causing the thrust slicing shown here.

2-14: **Which two tectonic zones are hardest to recognize and separate in the above image? ANSWER**

As you would expect, thrusting is common in the Appalachians, as a result of compression

tearing of sedimentary units during terrane docking and again as the African-North American plates collided. This thrusting is strikingly displayed in this Landsat image of the Pine Mountain thrust in eastern Tennessee; the thrust front is located at the upperleftmost ridge (see caption):



Both normal and reverse faults are far less likely to appear in a recognizable format on space imagery - even high resolution aerial photos can fail to detect them. There are several reasons for this: 1) reverse faults are uncommon (in tectonic settings requiring vertical uplift(; 2) many of these faults are relatively small (hundreds of meters or less; 3) the vertical displacement along a normal fault may be only a few to a few hundred meters, so that rock units differing in lithology are not always juxtaposed; and 4) the downdropped side of a normal fault, if large and resulting in significant displacement over time, produces a tectonic basin (the graben and horst style of faulting); this basin is the destination of sediments carried in from higher surrounding fault blocks, fills up (covering the downdropped block rock units), and laps against the trace of the fault plane.

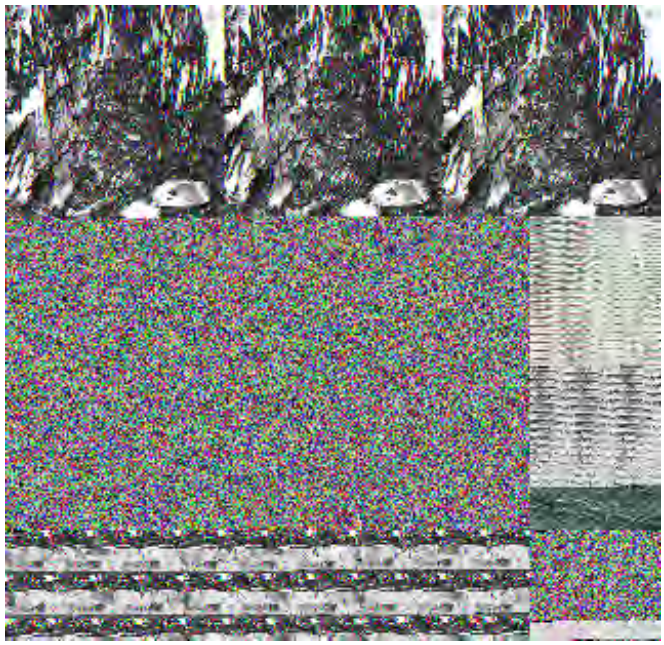
Here is an aerial photo that shows how disparate lithologies on either side of a normal fault surface expression can appear:

Active normal faults (those still developing; earthquake sites) can drop down the hanging wall in one or more movement events a few to ten or more meters. The result is an exposed fault plane - called a fault scarp. This is an aerial view of the scarp along the active Wasatch Front fault in Utah; this fault has been moving for millions of years (intermittent; a few meters at a time), so that the mountains in this photo are the upthrown side - now much higher - and the foreground is the downdropped hanging wall block:

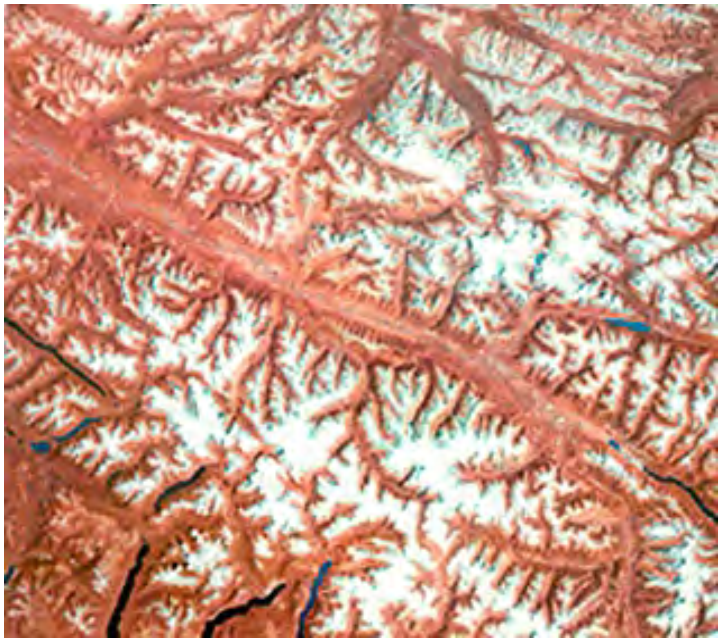
The scarp can be visible in higher resolution space and aerial imagery. But in three decades of viewing Landsat images, the writer has never seen evidence for normal faults as juxtaposed rock units. IKONOS-type images probably can show this.

There is one setting where normal faulting can be inferred. This is the tectonic situation in which large sections of the crust are uparched or downdropped where the near surface rocks are being subjected to tension. Under tension, normal faulting is expected, leading to the classic graben structure (downdropped block) and complimentary horst (higher block left behind or actually pushed up) (see page 2-9 and page 3-2 for pertinent diagrams [look for Kenya callout). In East Africa, the zone of the crust that is rupturing from tension as an overriden spreading ridge is pulling off part of the eastern continent is responding by having melted basaltic magma reach the surface as lava outflows. As these cool into basalt units, those are pulled apart to form normal faults with some slices of rock between faults dropping down as grabens while others remain behind as horsts. This SPOT image shows a small segment of the Rift Valley in which dark lines mark the fault traces and the blue depicts basaltic blocks on either side:

This image, extracted from a photo taken by astronauts using the Large Format Camera (LFC), shows a number of step faults (of the 'normal' type) which cuts into basaltic flows making up rift valleys:/p>



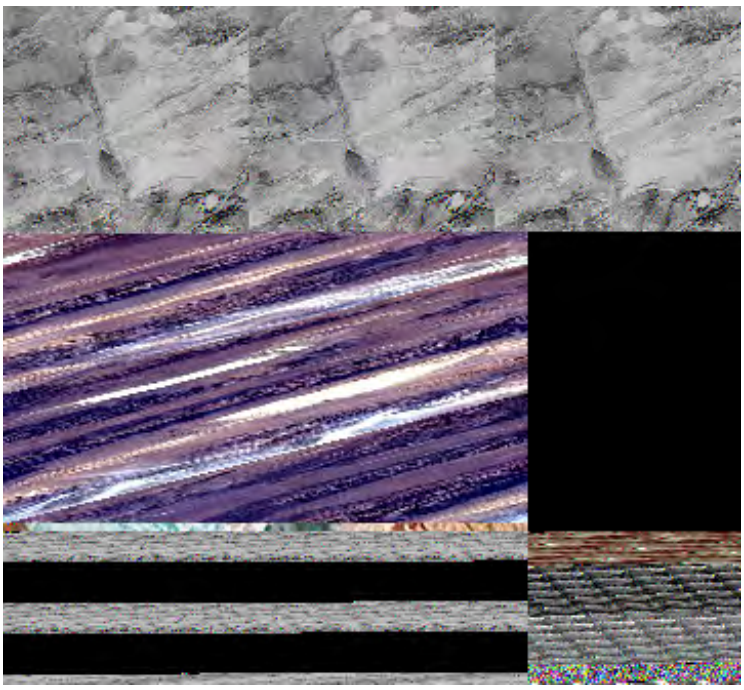
A spectacular example of a single downdropped block bounded by normal faults on either side is the Rocky Mountain Trench in Alberta, Canada. Here is a Landsat-1 view of this structural graben that is a flat valley midst high mountains:



The faults have been verified in the field. But if all one has to go on is the Landsat image, this faulting would be deduced from its effect on the topography. While there are rivers within the Trench, they do not cause the valley, since rivers seldom are straight over long distances and seldom carve out straight bluffs (here, these are fault scarps).

In the United States, the classic region where normal faulting abound is the Basin and Range of Nevada and neighboring states (see page 6-8. The geologic factors that have caused this faulting are discussed on that page (suffice to state that the entire region has been uplifted owing to thermal expansion so that the arched crust has failed in tension). Here are typical horsts (the ranges) and grabens (basins). The infilling of the basins masks the bounding fault planes. Reproduced here is a Landsat subscene that includes several ranges and the alluvial-filled intervening basins:

Sometimes it is difficult to identify a structural feature in a Landsat-type image. Examine this scene:



The area shown is in the eastern Sinkiang desert of western China. The Pei Shan is the mountain system at the bottom. Cutting across the left third of the image at a slant is a thin zone that separates different land cover types (desert surfaces; hills). Although the writer (NMS) has no maps of this region, he photointerprets the feature to be a fault. However, close inspection of the original image indicates some low topographic expression along the fault line. This could be explained as a dike (rock intruded along the fault plane) but there is obvious uncertainty here.

Thus, from the above evidence we can conclude two things: 1) faults can often be discerned in space imagery by their expressions as discontinuities, differences in topography and lithology on either side, and signs of displacement; and 2) the type of faulting cannot always be inferred until the rock ages are known; reference to maps or even field work may be necessary. One last item: Use of space imagery has led to the discovery of new, previously unknown faults; this can be important because it can reveal potential locations of future earthquakes.

Primary Author: Nicholas M. Short, Sr.

Joints are fractures that do not involve differential movement; they are usually too small in size to be directly recognizable from space but their effect on topography may reveal their presence. Two examples are shown. Lineaments are any linear features that can be picked out as lines (appearing as such or evident because of contrasts in terrain or ground cover on either side) in aerial or space imagery. If geological these are usually faults, joints, or boundaries between stratigraphic formations. Other causes of lineaments include roads and railroads, contrast-emphasized contacts between natural or man-made geographic features (e.g., fence lines), or vague "false alarms" caused by unknown (unspecified) factors. The human eye tends to single out both genuine and spurious linear features, so that some thought to be geological may, in fact, be of other origins. Examples are shown.

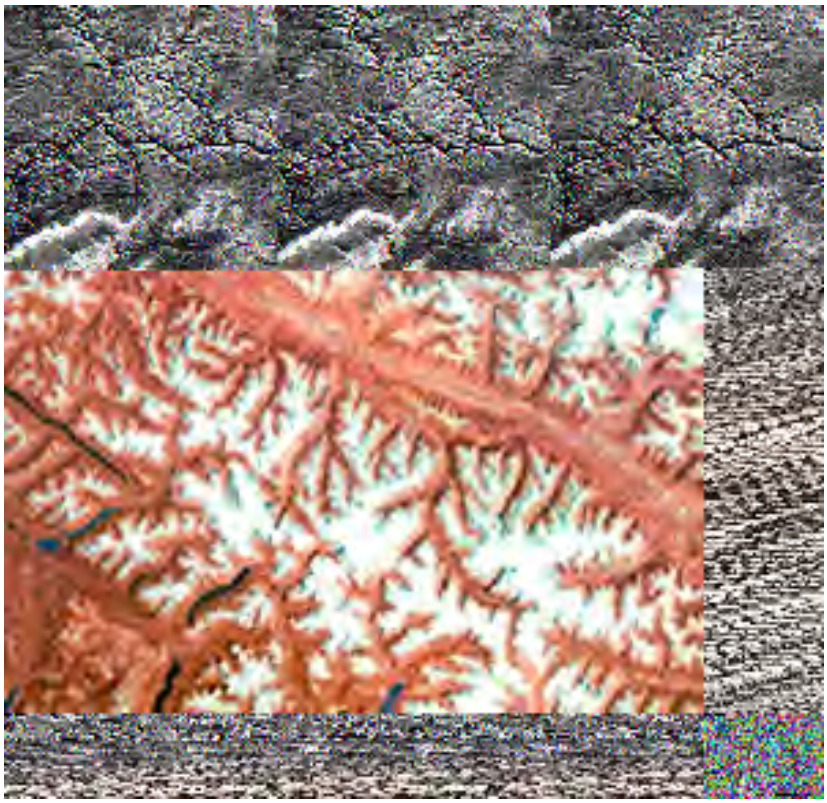
Joints and Lineaments

Joints are fractures that represent rock sprung apart (usually fractions of a centimeter unless/until widened by weathering) but next to which the rock on either side has not moved laterally. Most rock units, when seen in the field, show extensive breaking without any displacement. Every outcrop seems to be replete with cracks. These two photos show jointing as seen under exceptional viewing conditions:

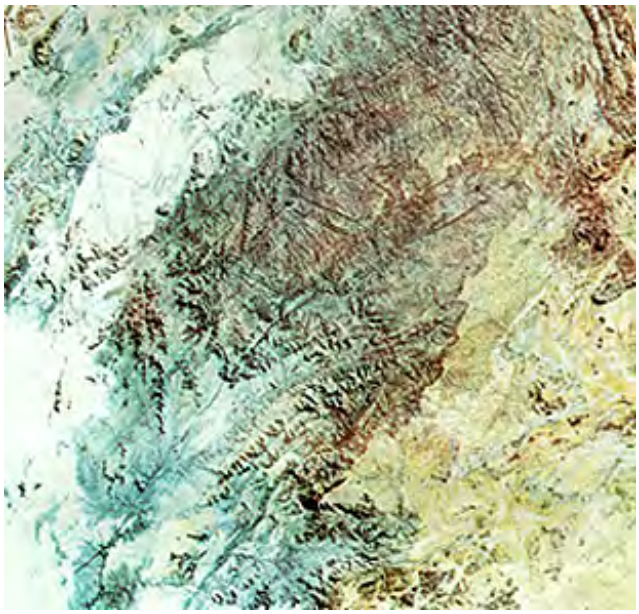
">

In space imagery, major joints can be conspicuous under certains conditions, where they are long, continuous, wide-spaced and enlarged by erosion. They can be conspicuous in Landsat imagery. This is well shown by the joint system cutting a basalt flow in Zambia over which flows

the Zambesi River (Victoria Falls).



Several sets of long, intersecting joints are evident in this Landsat subscene that center on a plateau underlain by Precambrian metasedimentary rocks southwest of Windhoek, Namibia:



Megajoints also control the topography in this Landsat image of the Nigerian Shield (southern Nigeria), a region of metasedimentary and basaltic rocks. At least 6 fracture sets are discernible. Surface and ground waters have opened up the fractures so that they now are narrow valleys in some places. Shadows cast by valley walls help to enhance fracture expression. Part of the upland erosion surface is vegetated (red).

2-15: **Do you know the difference between a fault and a fracture? How might the two types be separated visually on an aerial photo or space image?** ANSWER

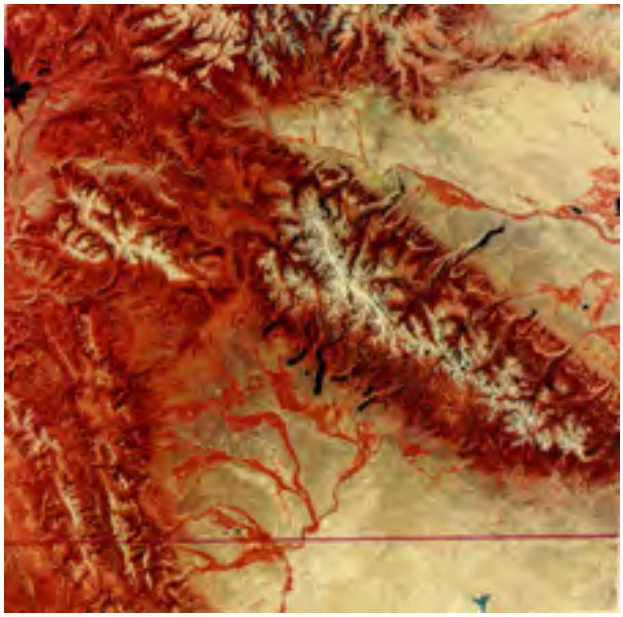
We turn now to lineaments. This subject has been a leading point of contention over much of the history of use of space imagery for geologic studies. In the early days of Landsat, perhaps the most commonly cited use of space imagery in Geology was to detect linear features (the terms "linears" or "photolinears" are also used instead of lineaments, but 'linear' is almost a slang word) that appeared as tonal discontinuities. Almost anything that showed as a roughly straight line in an image was suspected to be geological. Most of these lineaments were attributed either to faults or to fracture systems that were controlled by joints (fractures without relative offsets). Lineaments are well-known phenomena in the Earth's crust. Rocks exposed as surfaces or in road cuts or stream outcrops typically show innumerable fractures in different orientations, commonly spaced fractions of a meter to a few meters apart. These lineaments tend to disappear locally as individual structures, but fracture trends persist. The orientations are often systematic meaning, that in a region, joint planes may lie in spatial positions having several limited directions relative to north and to horizontal. For example, 60% of the joint planes might fall into a cluster of fractures with azimuths between N40W and N60W and inclinations between 30 and 40 degrees).

Where continuous subsurface fracture planes that extend over large distances and intersect the land surface produce linear traces (lineaments). A linear feature in general can show up in an aerial photo or a space images as discontinuity that is either darker (lighter in the image) in the middle and lighter (darker in the images) on both sides; or, is lighter on one side and darker on the other side. Obviously, some of these features are not geological. Instead, these could be fence lines between crop fields, roads, or variations in land use. Others may be geo-topographical, such as ridge crests, set off by shadowing. But those that are structural (joints and faults) are visible in several ways. They commonly are opened up and enlarged by erosion. Some may even become small valleys. Being zones of weak structure, they may be scoured out by glacial action and then filled by water to become elongated lakes (the Great Lakes are the prime example). Ground water may invade and gouge the fragmented rock or seep into the joints, causing periodic dampness that we can detect optically, thermally, or by radar. Vegetation can then develop in this moisture-rich soil, so that at certain times of year linear features are enhanced. We can detect all of these conditions in aerial or space imagery.

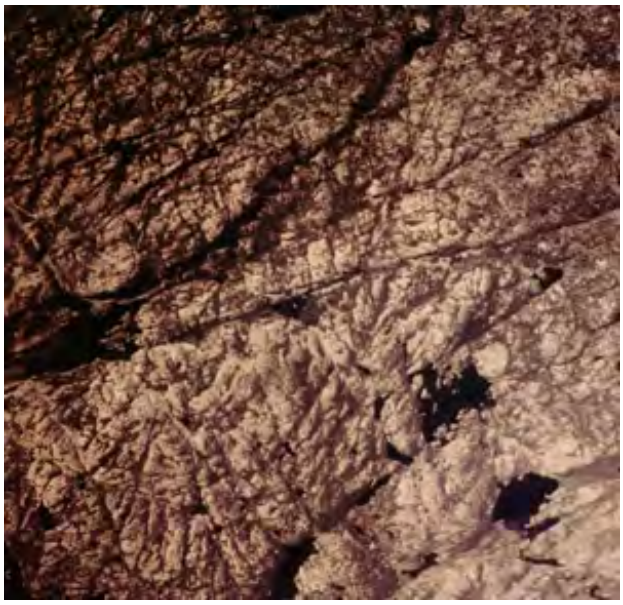
When we illuminate a space image, particularly a transparency on a light table, we can often spot numerous linear or boundary discontinuities that we can record on tracing paper. The resulting lineaments map may have dozens, even hundreds of straight or slightly curved lines. But, scientific skepticism can raise key questions: Are they real? Do they all correspond to the same feature or phenomenon? If not, what is the proper identity of each one? And, finally, how

do we verify their existence and determine their identities? These are important issues since we know lineaments (particularly faults) play key roles in location minerals and some oil or /gas deposits. They may also give clues to structural activities that bear on earthquakes. We save responses to these queries until the end of this section, after we cover how lineaments appear in images. Near the end of Section 5, we will again consider how human operator variations affect the correctness and consistency in picking valid lineaments amidst phony lines.

Within a month of the Landsat launch, images of the central Rocky Mountains were delivered to a team of investigators at the University of Wyoming. On this team, Professor Ronald Parker, a structural geologist, had been field mapping lineaments and other deformation features in the Wind River Mountains in the west-central part of the state. This great block of metamorphic and igneous basement rocks, flanked and patchily-covered by Paleozoic sedimentary rocks, is part of a segment of the Rockies where sections of the crust had been uplifted six km or more above other segments. Other segments stayed put or had downdropped to form deep (up to six km below the present surface) basins now back-filled by erosional debris from the surrounding mountain groups. As shown here by this Landsat image, the Wind River Mountains (center right) lie between the Wind River Basin (upper) and Green River Basin (lower), with the Gros Ventre and Hoback Ranges to the west. The Wind River Mountains rise to more than 3940 m (13,000 feet) above sea-level and have been strongly dissected by rivers and glaciers that exposed and carved into many major faults and lineaments.



It had not proved easy to map these lineaments in the field or from the limited aerial photos. As part of the project, NASA had previously flown a U-2 on a single pass, taking pictures of terrains that included this next image. The photo shows an eroded granitic surface with numerous lineaments, exposed in a rugged glaciated high mountain surface, with almost no tree cover on its higher elevations.

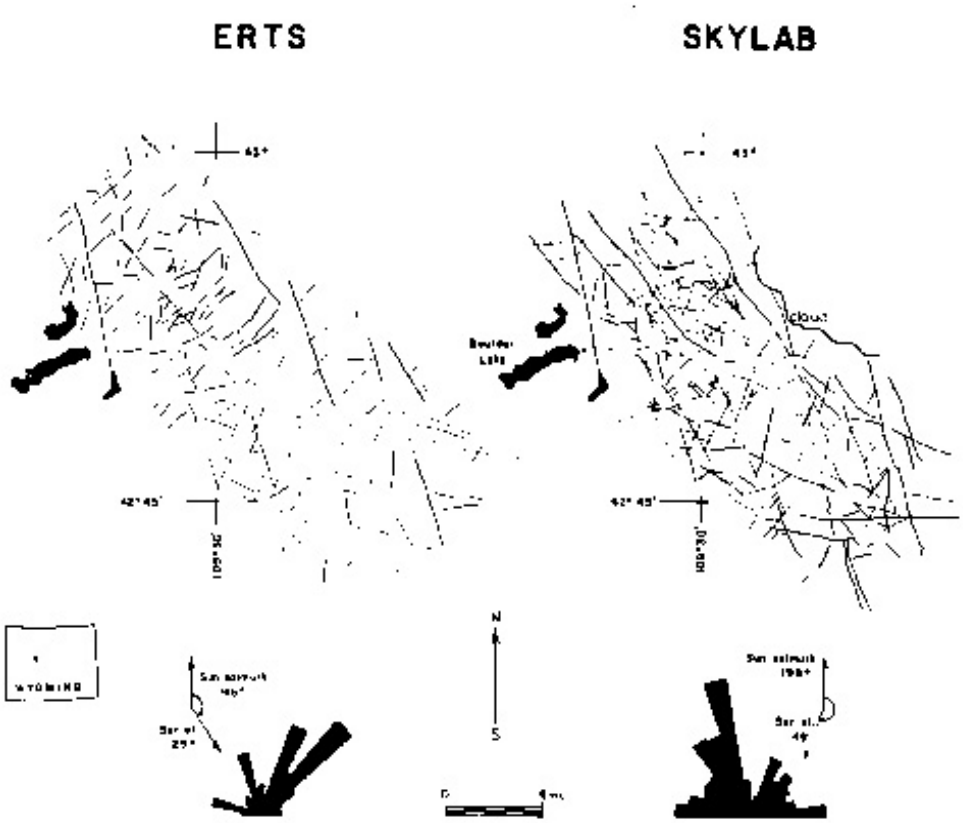


2-16: **The above aerial photo displays lots of fractures - a real hodgepodge. Can you think up a good way to show their relative orientations as well as the relative number in any given direction? ANSWER**

Dr. Parker, during five field seasons (camping in the high country, and supported by pack mules) had completed ground mapping faults, shear zones, and filled dikes in about 20% of the Wind River uplifted block. He supplemented his work by photointerpretating the U-2 strip. His map prior to receipt of the Landsat imagery is shown below on the left.

After recciving a Band 5 Landsat (ERTS) image as a transparency, using a light table he produced the lineaments map on the right in just three hours (including a break). Some of the newly plotted lineaments had been discovered earlier by other geologists, but most were heretofore unknown, and some of those have since been verified in the field. This great breakthrough effort now meant that we could map the synoptic overview of a large section of fractured continental crust with acceptable reliability in less than a day rather than months of difficult field work in poorly accessed terrain. This new method also meant great savings in time and money.

But, this mapping is of necessity incomplete and somewhat misleading. The maps below compare the ERTS and Skylab (Mission S-190B) Lineaments maps from the central Wind River Mountains. We plotted the orientation (azimuth or compass bearing) of the dominant direction relative to north, of each linear feature in a rose diagram (at the bottom of the map).

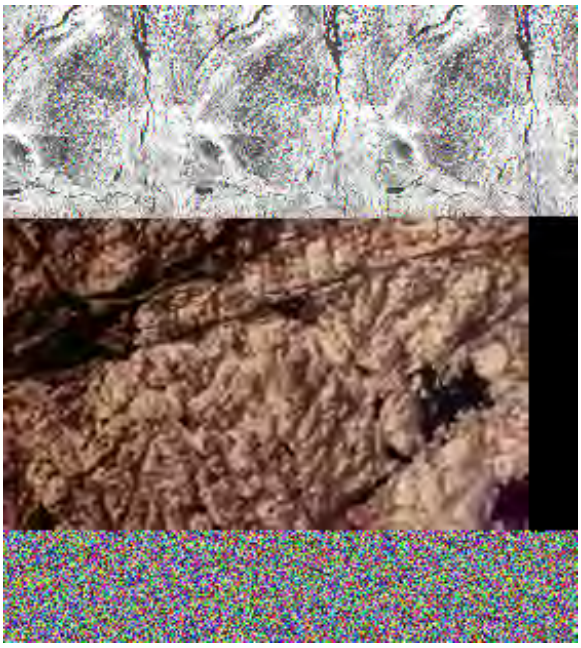


This diagram shows directions for all linear sets that we grouped in intervals (here 10 degrees), such that the length of the tapering bar in each interval is proportional to its frequency distribution (relative proportion) among all lineaments in the intervals distributed over the 180 degrees; that encompass west to east trends. Note that the dominant directions for ERTS lineaments are NE, whereas those for Skylab are NNW.

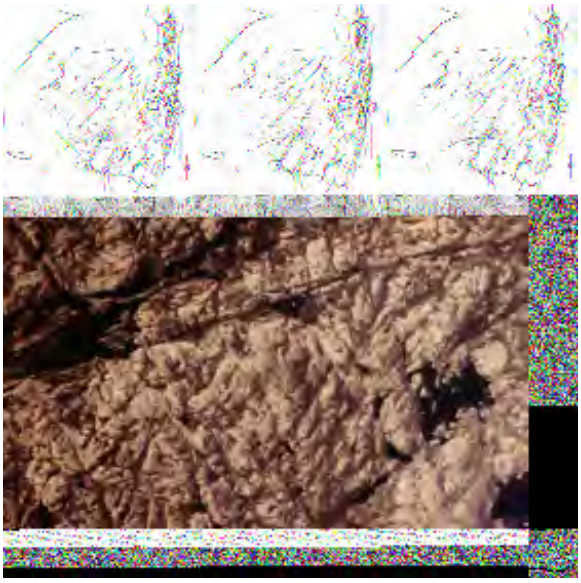
The cause of this difference between the two observations is simply time-of-day. The ERTS image was acquired around 10:30 A.M., local time, when the Sun's rays came from a SE position at a moderate elevation angle. Fractures occupying depressions that trend NE are shadowed on their NW side, and hence, stand out in the image as shadow relief. Whereas, those trending NW are equally illuminated on both sides and hence largely invisible and easily missed. On the other hand, the Skylab photo was taken in mid-afternoon, when the Sun was shining towards the NE at a higher elevation, so that shadows would maximize along NW trending lineaments. This illumination bias is a well-known effect (see page 8-4) and forces us to decide cautiously about structural trends when only data sets taken at one time of day are available. The obvious solution is to use multiple data sets, obtained at different times. Unfortunately this is not an option with Landsat. However, a geostationary satellite that can image at any time would circumvent this problem, but no high resolution systems [superior to meteorological ones] are operational yet.

2-17: **How could you get around (overcome) this time of day or sun angle bias, using a satellite?** **ANSWER**

Even as Dr. Parker and his cohorts were testing ERTS-1 as a means of speeding up their mapping of the Wind River Mountains, another geologist, Dr. Yngvar Isaachsen of the New York State Geological Survey, was examining his test site (he was one of the original Principal Investigators in the ERTS evaluation program in Geology). His mission: to test his hypothesis that ERTS imagery would significantly improve his long-standing studies of fracturing of the Precambrian massif known as the Adirondacks Mountains. (This large uplands in upstate New York is a surface expression of the Canadian Shield exposed after uplift and erosion south of the edge of the Shield near Montreal.) Here is what he saw in the first imagery he received:

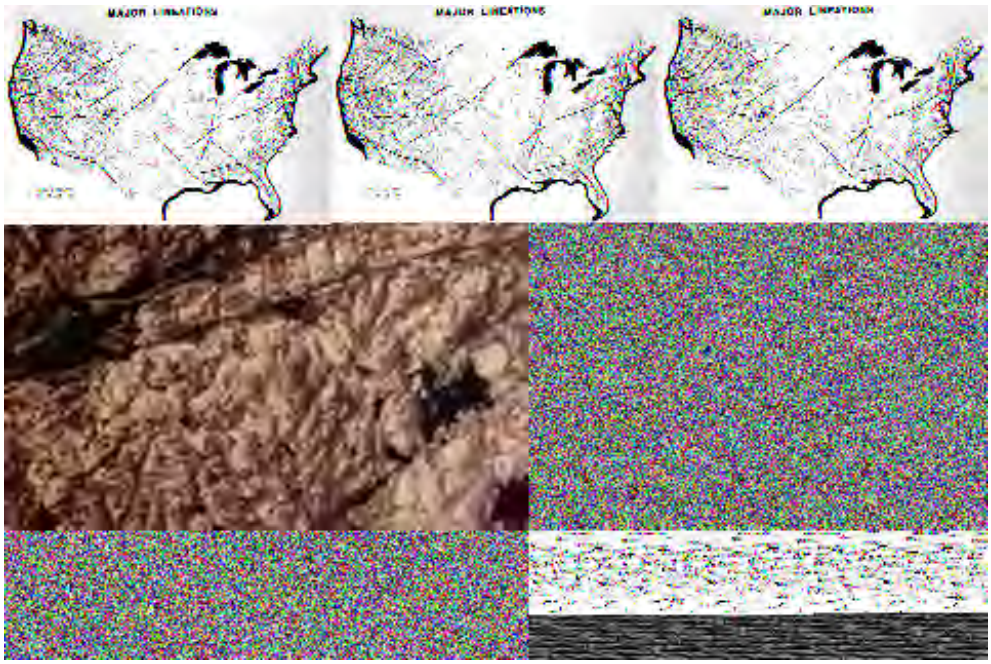


Within days, Isaachsen had drawn this provisional map of lineaments in the Adirondacks:



Isaachsen is a true geologist - a field man. He proceeded to spend a half year field-verifying ("ground truthing") the existence and nature of the lineaments he had drawn on an improved map. Some of these proved to be shear zones, suggesting faults. Most could only be construed as fracture zones with no evident displacements (the majority of lineaments were hidden, many covered by glacial deposits). Some he discovered were artifacts - fence lines that kept trees on one side, farmers fields, etc. But after this verification study, the tectonic map of New York was judged to have been improved in the Adirondacks by at least 50%.

With some trepidation, the writer has added this map which plots the major lineations across the 48 U.S. states. The map was made by Mr. Douglas Carter of the U.S. Geological Survey, a leader in their remote sensing program and a respected structural geologist. He used both the Soil Conservation Service U.S. MSS Band 5 mosaic and individual transparencies of nearly all individual Landsat frames (he observed using a light table). The reason for caution: In the early days of ERTS/Landsat-1 geologists were finding lineaments everywhere. Nearly all were never field-checked, so their veracity remains unknown. In the writer's (NMS) opinion, some of Carter's lineaments are valid; others ???



Other studies on the next page speak to both the efficacy and the shortcomings of using space imagery to map lineaments.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

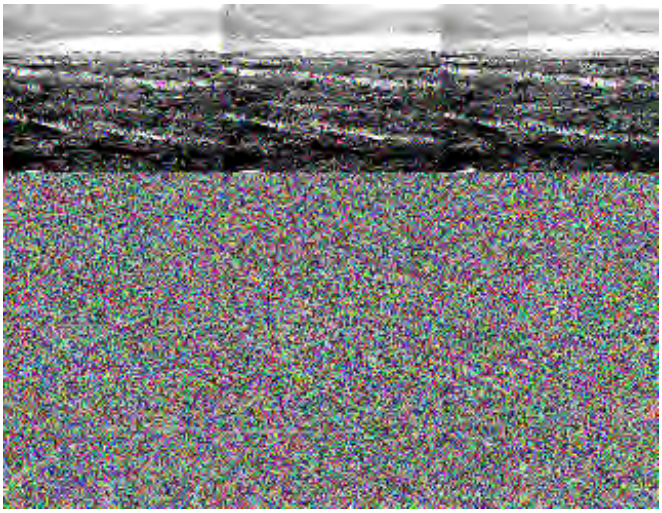
The value and applications of lineament studies with space imagery are indicated by three case studies. The first compares lineaments detected in Landsat and SPOT images with those from four case studies: 1) lineaments caused by normal faulting in the African Rift Valley; 2) lineaments in the Canadian Shield; 3) the surface expression of the San Andreas Fault in southern California; and 4) a study of joint expression in the Piedmont near Atlanta, GA.

Several Case Studies of Fracture Analysis

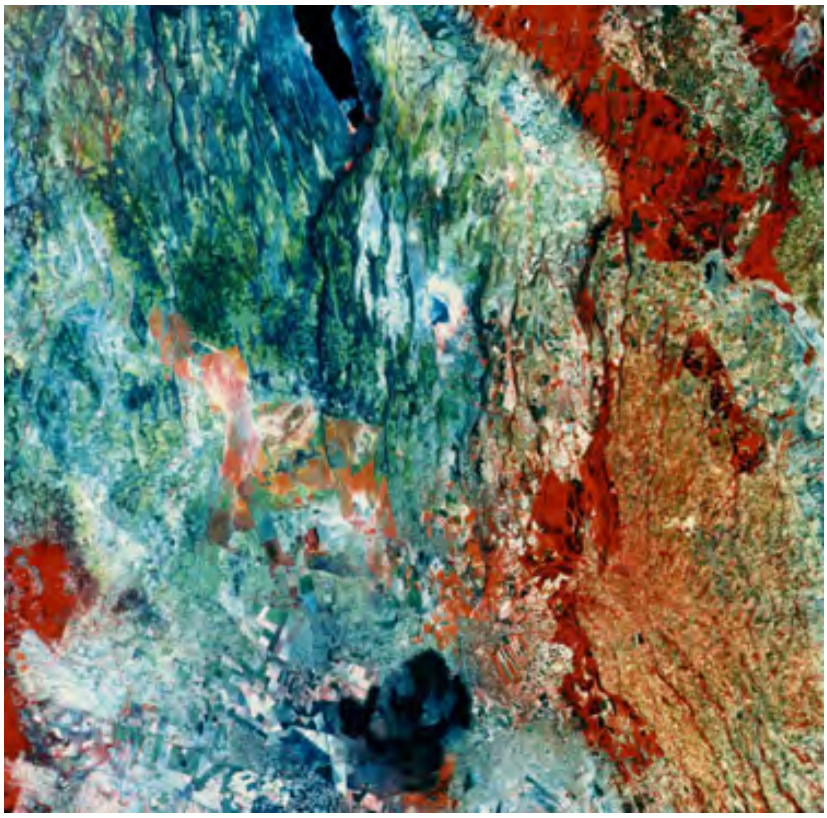
The East African Rift

To test the merits of space imagery for finding fractures on Earth's continents, we examined a region, in which these are clearly visible and reliably mapped, as a ground-truth site using Landsat and SPOT scenes. The East African Rift is a zone of crustal extension, in which part of the eastern African continent is pulling away from its parent along one arm, separating the divergent blocks that stem from the triple junction, centered in Ethiopia (the second arm includes the Red Sea and the third, the Gulf of Aden off the Arabian Peninsula). Much of the rift consists of downdropped crustal sections, bounded by deep-rooted normal faults (forming grabens) that cut into the basaltic lavas, extruded in the resulting depressions. (In the fault type called "normal", imagine a rectangular block broken by a fault that inclines downward at some angle; the upper part of the block then slides down the fault plane; if there are two normal faults facing each other but the second inclined in the opposite direction (so the two converge), then the part of block between the two faults drops as a unit [the graben] becoming lower and leaving the blocks on either side of the fault pair higher; see illustration near top of page 3-2)

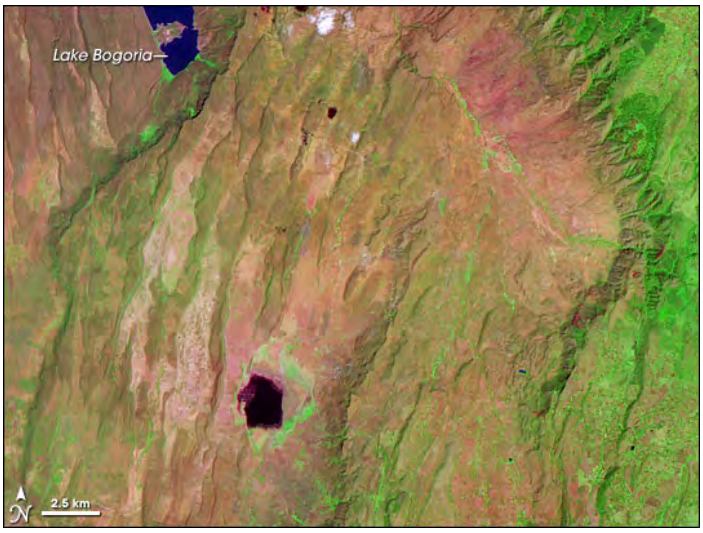
These faulted lava fills, because they are recent and not severely eroded, have strongly-expressed surface scarps. They don't show tonal differences on the two sides of a fault, because the faulting affects the same rock type as seen at the surface. The next ground photo shows a part of this East African Rift in Ethiopia, with prominent individual faults having scarps all facing in the same direction.



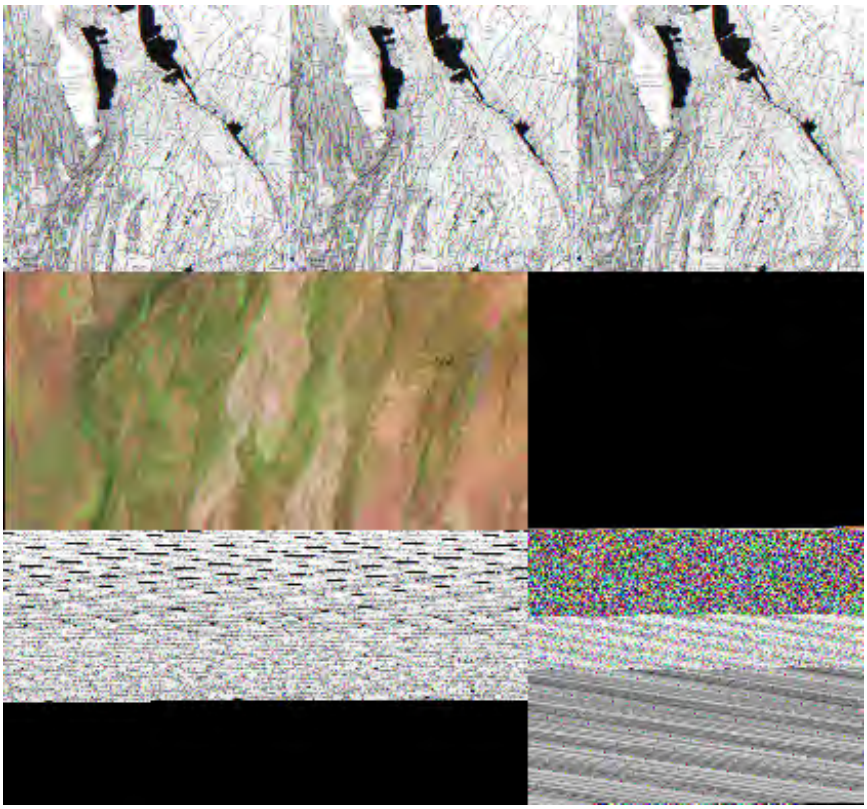
The scene below shows part of the rift in Kenya (Lake Hannington is at the top) as imaged by SPOT (described in the next Section).



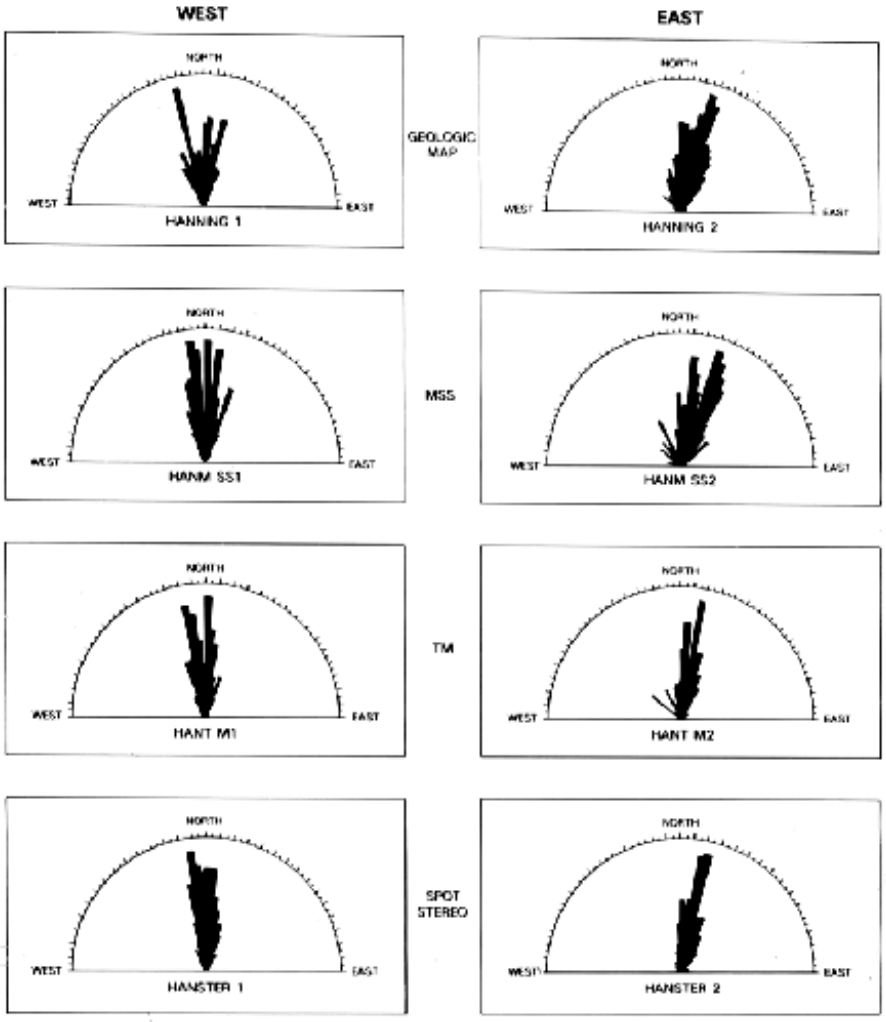
This ASTER natural color image shows the topographic expression of these rift faults:



A map of rock units and fractures (most are faults) of the upper half of this image was made by Kenyan geologists by field mapping and aerial photointerpretating. We reproduce it here (at a quality that precludes reading details) to illustrate the reference database.



The next figure shows rose diagrams prepared by the author from his interpretation of Landsat Multi-spectral Scanner and Thematic Mapper images and a stereo pair from SPOT (see page 11-9).



Areas west and east of Lake Hannington were treated separately. Hence, there are two sets of

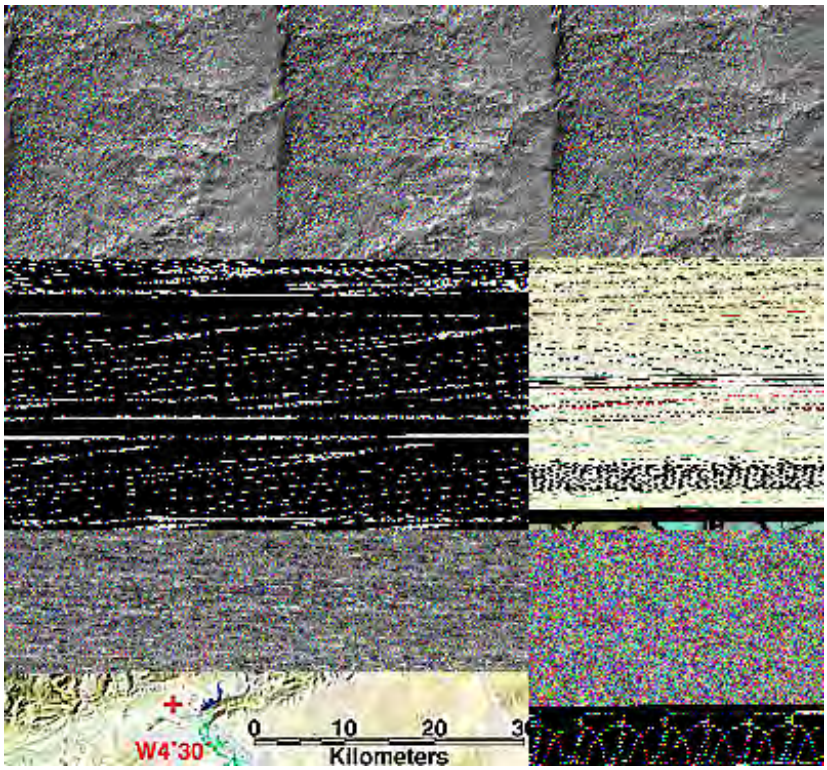
diagrams. Judging from the similarity of the orientations in the ground-map diagrams, Landsat and SPOT do an excellent job defining larger fractures. These topographic expressions are so pronounced that shadow bias is minimized, thus allowing accurate measurements of their trends.

2-18: **Although the different satellite images show pretty much the same orientations, there is one difference when these are compared to the geologic map. What is it and why? ANSWER**

The Canadian Shield

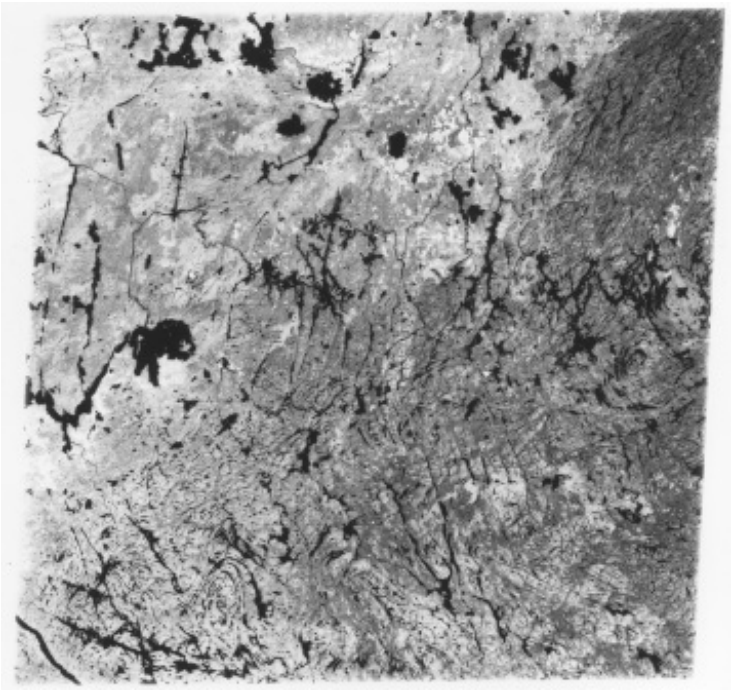
Lineament analysis using space images has been particularly valuable in determining regional and subcontinental fracture patterns that reveal some of the stress history imposed on large crustal units. Surface exposures of deep crustal rocks are known as shields which are mostly igneous and metamorphic, and form the basement core of the continental craton. Shields typically contain major sets of fractures developed mainly at the time of emplacement of the individual terrains that collided and assembled into these primary crustal units. In northern Canada, where soil and glacial cover are minimal, the Shield there contains a very high density of fractures, shown in the image below.

A segment of this image has been enlarged, contrast-stretched, and enhanced by filtering to emphasize the linear expression of the fractures.

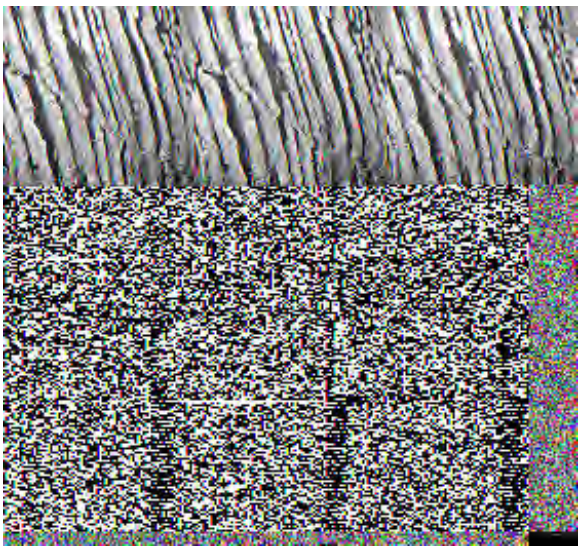


These are highlighted by three factors: 1) in this winter scene, the Sun's elevation angle is only nine degrees;, so that the low-angle rays emphasize small differences in topography associated with linear depressions, 2) the snow cover acts as a dusting which emphasizes these features similar to sprinkling a powder on a fingerprint, and 3) many of the fractures have been gouged out by glacial scouring.

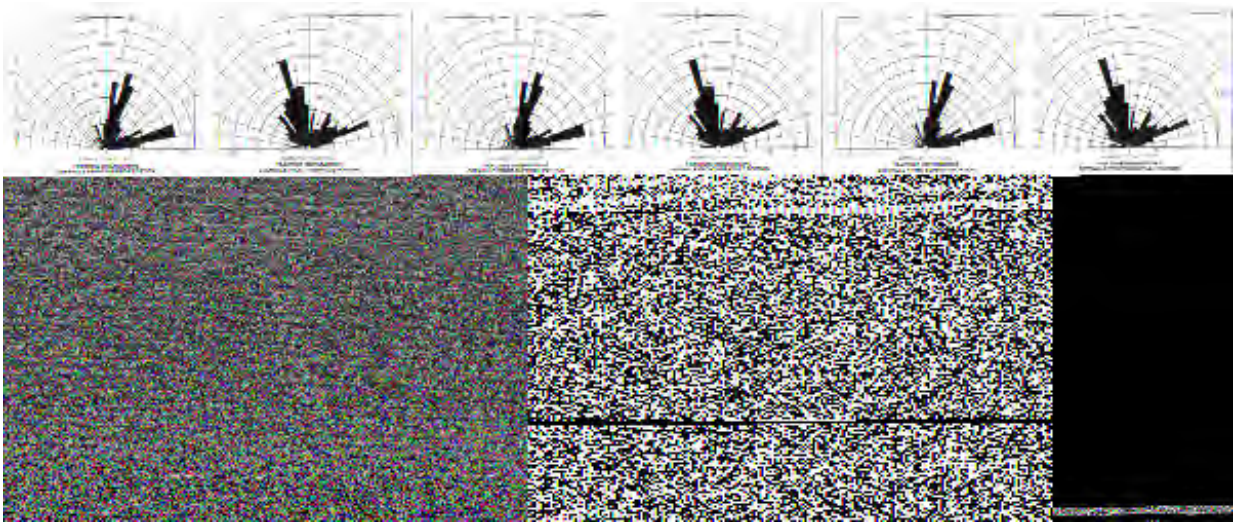
The writer (NMS) was part of a team at Goddard Space Flight Center that produced lineaments maps for selected areas within the Canadian Shield, which comprises most of the eastern 2/3rds of Canada. Their goal was to better understand its structural history. Because glacial scouring has removed covering soils and has sculpted lineaments into lakes, fractures over large parts of the shield are clearly evident, as demonstrated in this example.



It is the frequent filling of scours by lakes that brings out the presence of the lineaments; the lake waters are very dark in IR images. This aerial photo illustrates the control by more easily eroded lineaments on lake distribution after scouring.



Simply by looking at the Landsat image, we can tell that the terrain in the middle of the image (below the upper-left corner) is different from the remaining terrain to the south. The terrain in the upper left corner marks the Superior Province of the Shield, consisting of meta-igneous rocks, whose ages go back beyond 2 billion years. The terrain to the right is the Grenville Province, whose rocks are generally younger and were emplaced northwestward against the Superior Province some 900 million years ago. The line, is the Grenville Front, made of thrust faults that are seen from above as a NE-trending boundary. Many of the fractures in each province stand out as linear lakes. These were glacially-scoured gouges that are now filled with water. When the orientations of the lineaments are plotted in rose diagrams for each province, two characteristics emerge:

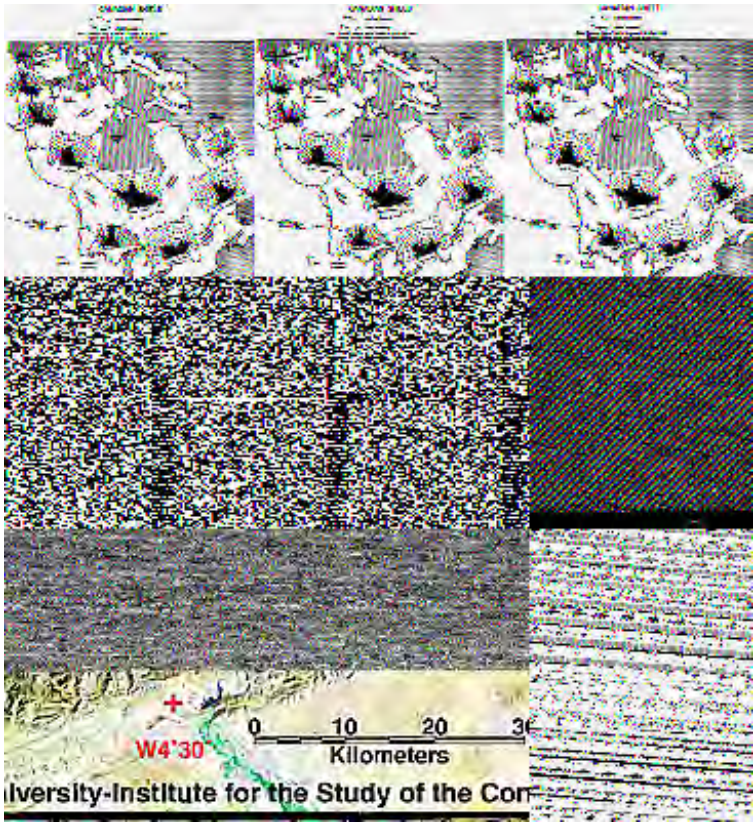


1) both provinces have a dominant ENE fracture trend and 2) the second maximum is NNE for the Superior (left diagram) and NNW for the Grenville (right diagram) Province. This example underscores the power of space imagery in providing observations that are well-suited to scientific analysis of regional structures.

2-19: **There is another notable difference in the fracture distribution in the two provinces. What is it?** **ANSWER**

Over part of a one year span lineaments orientation and frequency measurements were carried out by a group at NASA Goddard for the entire Canadian Shield (this probably was the largest area so studied by regional fracture analysis attempted up to that time [1986]). One objective was to see if fracture patterns differed from one tectonic province to the others. From this general rose diagram for the crystalline shield, the answer appears to be a firm yes. This suggests that as each province was emplaced as an accreted terrain (see Section 17), the

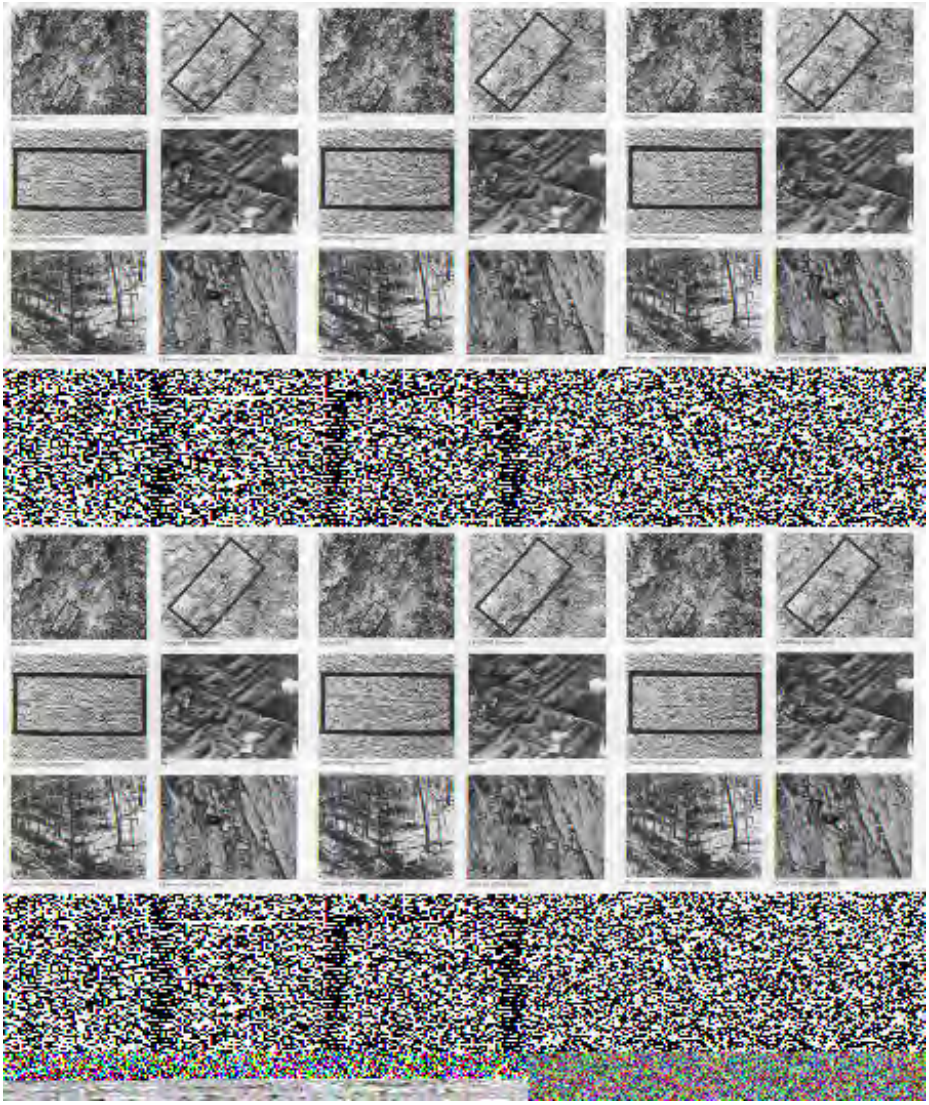
stress field at the time was responsible for the observed lineaments pattern for that province; stress fields for different provinces as they docked and added to the continental craton had different orientations relative to today's geometry.



Lineament analysis has a practical side, as well. For example, fractures and faults can serve as channelways for circulating solutions that eventually deposit ore minerals. They also are involved in traps for oil and gas and can be instrumental in storing and moving ground water. Original or improved lineament maps, therefore can payoff economically.

Atlanta Groundwater Survey

Our next example emphasizes this economic point. In the mid-1970s, the State Geologist of Georgia was commissioned to show where in a small area northwest of Atlanta conditions would be favorable to locate a new industrial park, which would require large amounts of groundwater for plant use. He included space imagery as a means of locating fracture patterns that could control water accumulation. The results of this effort are summarized in the sequence of image panels shown here.



The upper left panel is a Skylab photo of Atlanta with the survey area outlined in a box. The image to its right is part of a Landsat MSS image that shows layered rock units in the Piedmont Province of the Appalachians. In the center-left image, data from Landsat were computer-processed by a technique known as edge enhancement that highlights linear features by bounding them with a light-dark contrast. The strata are out clearly visible from this technique, but also exposed are numerous parallel lines, running across the strata from upper left to lower right (almost east-west when properly reoriented). These parallel lines looked suspiciously similar to a single, prominent, joint set. When we see them in a high altitude aerial photo, taken from an RB-57 aircraft, they appear as small linear valleys (right center). When visited on the ground, they appear as joint valleys, such as in the lower left panel. When we walked out several of these joint valleys until they intersected the Interstate 285 Beltway, we found that they flowed into road cuts, where exposures such as that shown in the lower right panel, identified them as fracture zones. The porosities associated with these zones made them favorable as potential aquifers for groundwater storage.

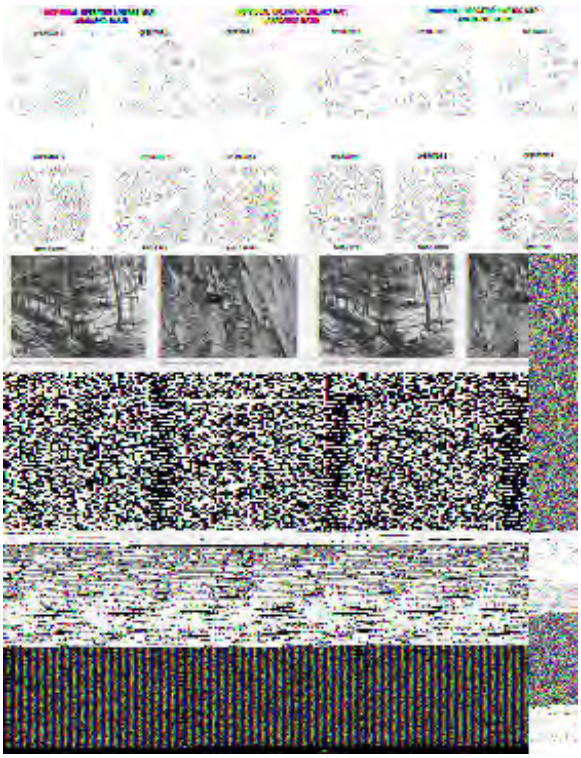
The area was drilled and large supplies of groundwater were proven. This led to an approval for developing the industrial park. The results of this study were brought to the Governor's attention, and he was so impressed that he insisted on going into the field with the State Geologist to see the evidence. A year later that governor became President Jimmy Carter, who was always a great friend to the Landsat program. This study also was cited in Congress as *prima facie* evidence of the value of space technology to the nation.

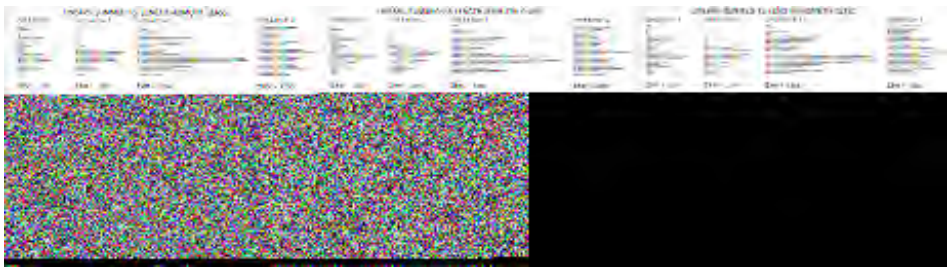
2-20: **What was the key step in the investigation process applied to the Atlanta survey that led to its eventual success?** **ANSWER**

Oklahoma Lineaments Analysis

As mentioned on the previous page, the writer (NMS) was for several years at Goddard Space Flight Center the Landsat Discipline Leader in Geology. As such, I was responsible for evaluating 57 initial Principal Investigator studies. One, known as the Eason Oil Study, is summarized on page 5-5. Its main concern was to look for surface indicators of subsurface oil deposits in the Anadarko Basin of southern Oklahoma. One clue that the investigator team reported was the widespread occurrence of lineaments, some of which they argued were possible factors in the development of these deposits (their principal claim, stated on page 5-5, was that over many of the developed oil fields leaks along the fractures led to chemical changes in surface rocks and soils which were expressed as tonal anomalies in ERTS images).

The writer (NMS) has tended to be a skeptic about much that was reported in the early days of geologic studies of satellite imagery. That attitude was paramount in my evaluation of the Eason Oil results (as will be presented on page 5-5). It was doubly re-enforced by my doubts about their lineaments mapping in Oklahoma. So, an experiment was designed. I and three other geology colleagues at Goddard each made our own lineaments map from the prime Landsat image (black and white Band 7 transparency) that included the Anadarko Basin study area in Oklahoma. Prior to this effort, all of us met for an hour to try to establish some common ground for our decision-making in selecting the linears we would then map. The results are summarized in these two illustrations:

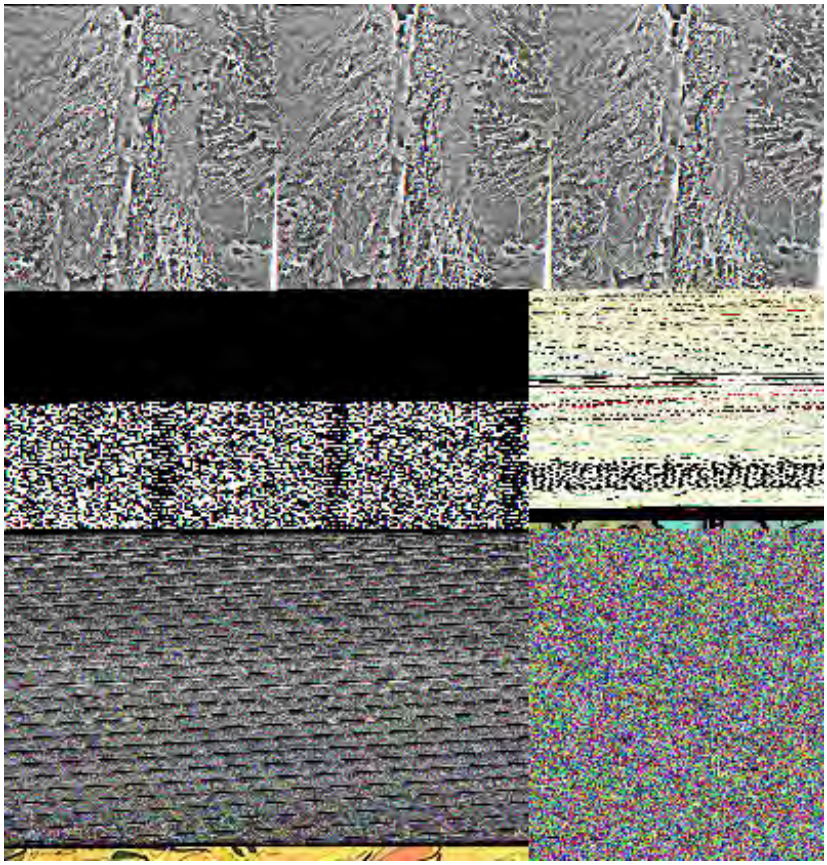




The writer, Operator 3, was the most experienced of the four in lineaments mapping. Thus, I found the most. Operator 4 was confused about some of the mapping criteria and probably found too many spurious linears. Operators 1 and 2 - both experienced structural geologists - were more conservative in what they considered bonafide linear features. One positive note, all the first three identified the principal aximuth trend of linear directions in the region. Comparing our results with those reported by Eason Oil resulted in a 59% agreement in picking out the larger lineaments. Conclusion: once more the subjectivity and variability of individuals in mapping

lineaments was confirmed.

To close, we give you a mini-minitest to see if you have learned how to search a space image for specific information. Look over this image.



2-21: **What structural feature, the continuation of an area we saw before in this Section, can you pick out in this enhanced Landsat image? What gave it away?**
ANSWER

A closing remark: During the first ten years of Landsat, its fullest use, as measured by sales of data to the user community, was in geology. Landsat, and later SPOT and other space systems, including radar, became respected new sources of information of particular interest to those who explore for oil and gas, mineral deposits, and groundwater. We will describe remote sensing applications to geological resources exploration, such as metal deposits and petroleum fields, in Section 5. But first, in the next two Sections we introduce two other major application regimes: in monitoring vegetation and assessing land use and urban planning .

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Vegetation can be distinguished using remote sensing data from most other (mainly inorganic) materials by virtue of its notable absorption in the red and blue segments of the visible spectrum, its higher green reflectance and, especially, its very strong reflectance in the near-IR. Different types of vegetation show often distinctive variability from one another owing to such parameters as leaf shape and size, overall plant shape, water content, and associated background, e.g., soil types and spacing of the plants (density of vegetative cover within the scene). Even marine/lake vegetation can be detected. Use of remote sensing to monitor crops, in terms of their identity, stage of growth, predicted yields (productivity) and health is a major endeavor. This is an excellent example of the value of multitemporal observations, involvig several looks during the growing season, allows better crop type determination and estimates of output. Vegetation distribution and characteristics in forests and grasslands also are readily determinable.

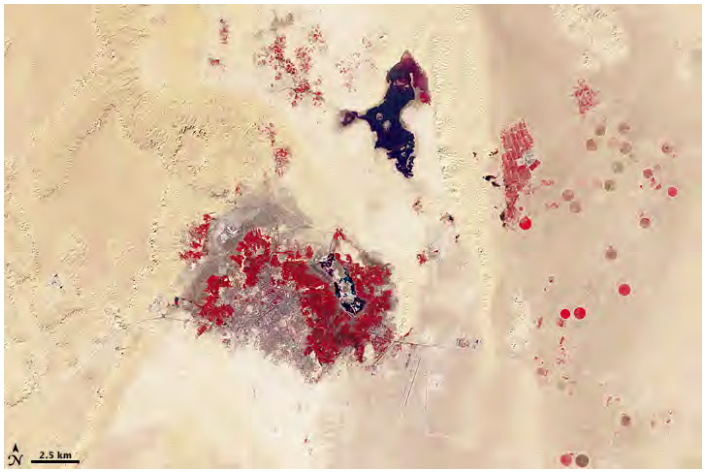
VEGETATION APPLICATIONS:

AGRICULTURE, FORESTRY, AND ECOLOGY

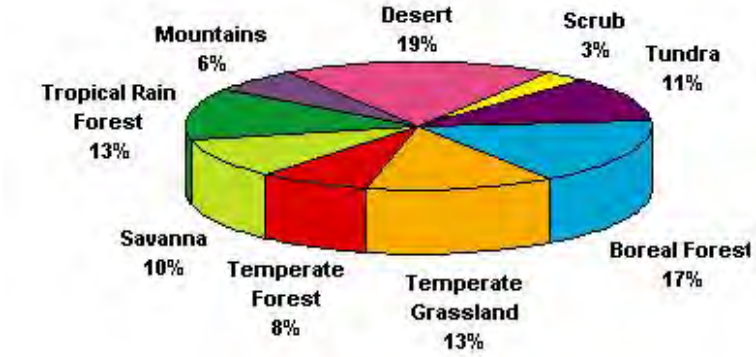
General Principles For Recognizing Vegetation

Planet Earth is distinguished from other Solar System planets by two major categories: Oceans and Land Vegetation. The oceans cover ~70% of the Earth's surface; land comprises 30%. On the land itself, the first order categories break down as follows: Trees = 30%; Grasses = 30%; Snow and Ice = 15%; Bare Rock = 18%; Sand and Desert Rock = 7%.

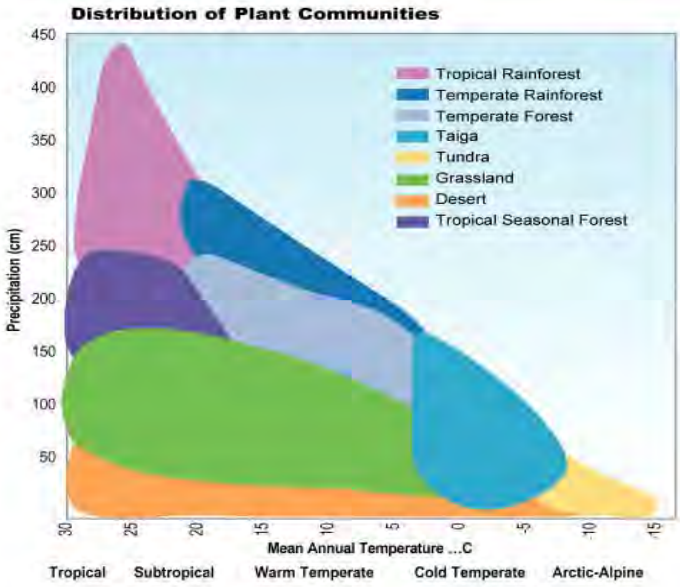
We have already seen in previous Sections and in the Overview that in false color imagery the remote sensing signature of vegetation is a bright red. The landscape shown in this first image could almost be on Mars except for the presence of this bright red sign of vegetation. This is the Ouargla Oasis in the Sahara Desert of southern Algeria, a concentration of trees and plants where groundwater reaches the surface:



On Earth, the amount of vegetation within the seas is huge and important in the food chain. But for people the land provides most of the vegetation within the human diet. The primary categories of land vegetation (biomes) and their proportions is shown in this pie chart:



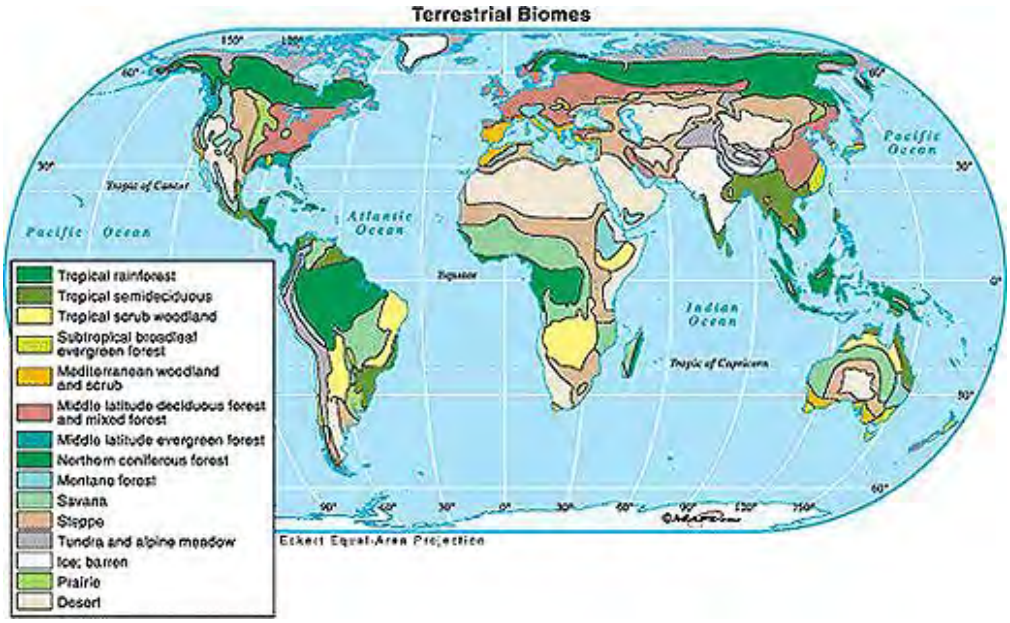
These biomes are defined in part by the temperature and precipitation controls that differentiate them:



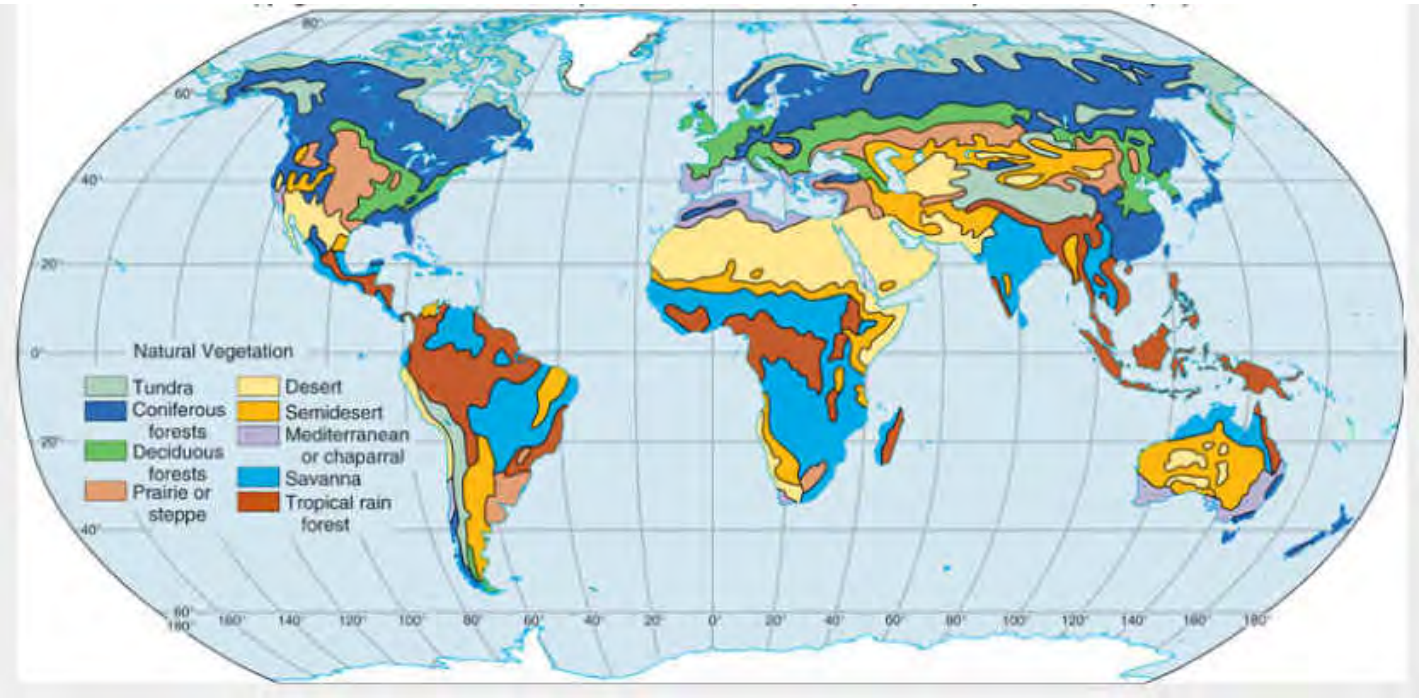
Both temperature and precipitation limit the distribution of plant communities. The climate (temperature and precipitation) zones of some of the major plant communities (such as temperate forests, grasslands, and deserts) in the US are shown in this figure. Note that grasslands' zone encloses a wide range of environments. This zone can include a mixture of woody plants with the grasses. The shrublands and woodlands of the West are examples of grass/woody vegetation mixes that occur in the zone designated as grasslands.

With climate change, the areas occupied by these zones will shift relative to their current distribution. Plant species are expected to shift with their climate zones. The new plant communities that result from these shifts are likely to be different from current plant communities because individual species will very likely migrate at different rates and have different degrees of success in establishing themselves in new places.

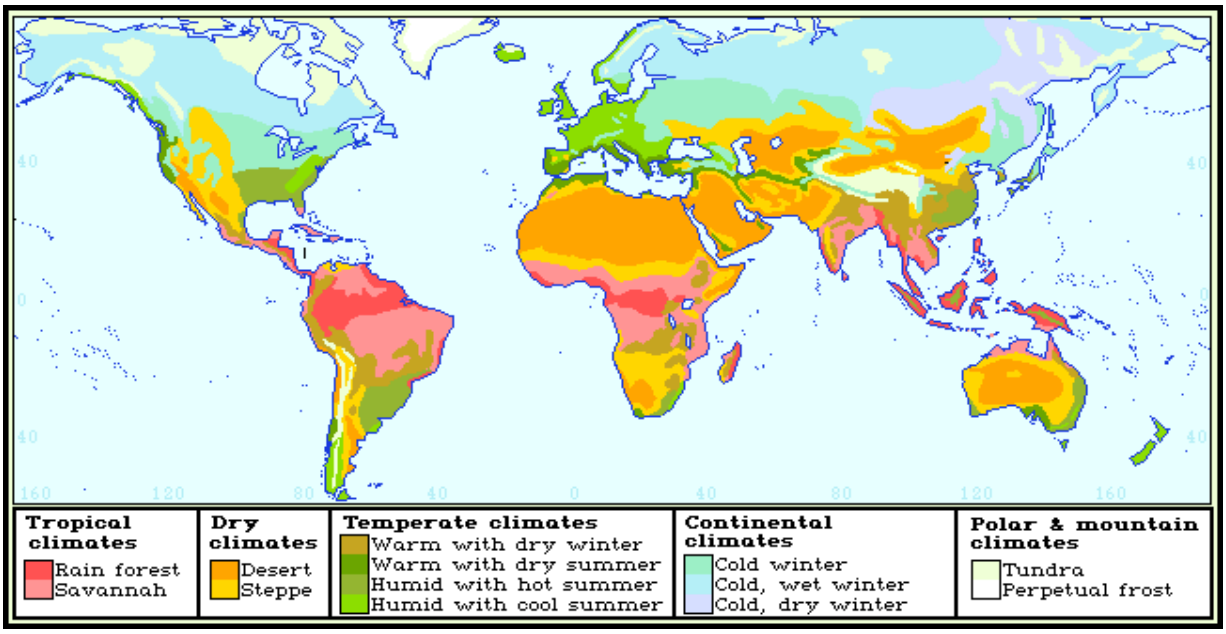
Global maps of vegetation biomes on the continents show this general distribution:



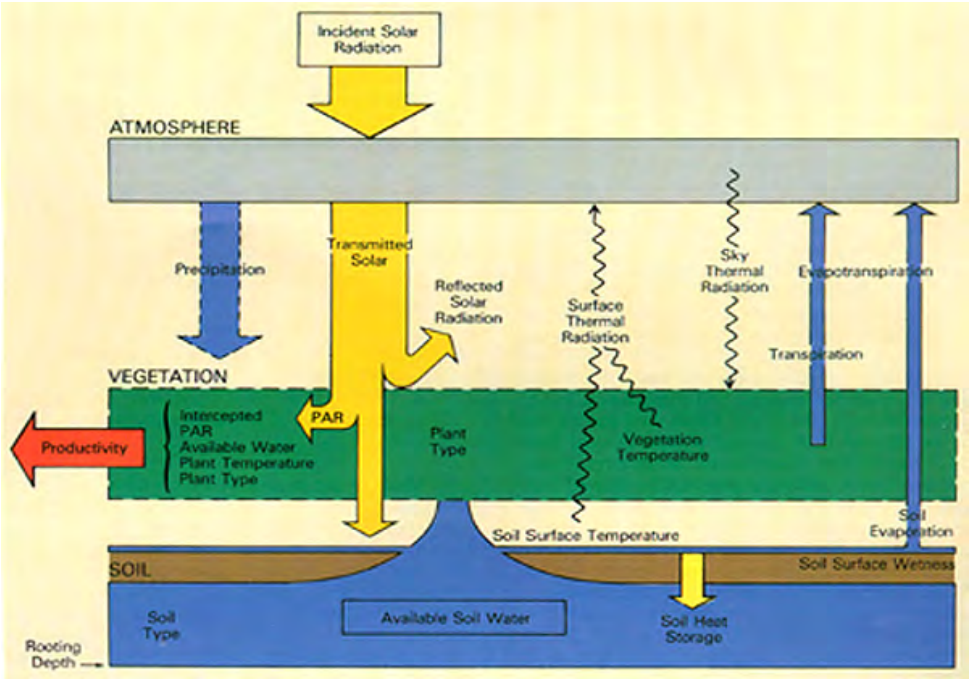
A fair number of global vegetation maps have been published. These usually show slight to moderate differences, depending in part with the types and numbers of classes established in the classification. Compare the map below with the one above:



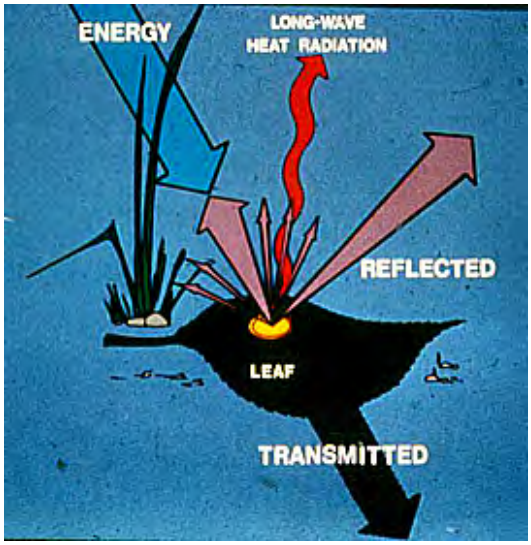
There is a notable correlation between vegetation classes and climate. Once again, compare the climate map below with the two vegetation maps.



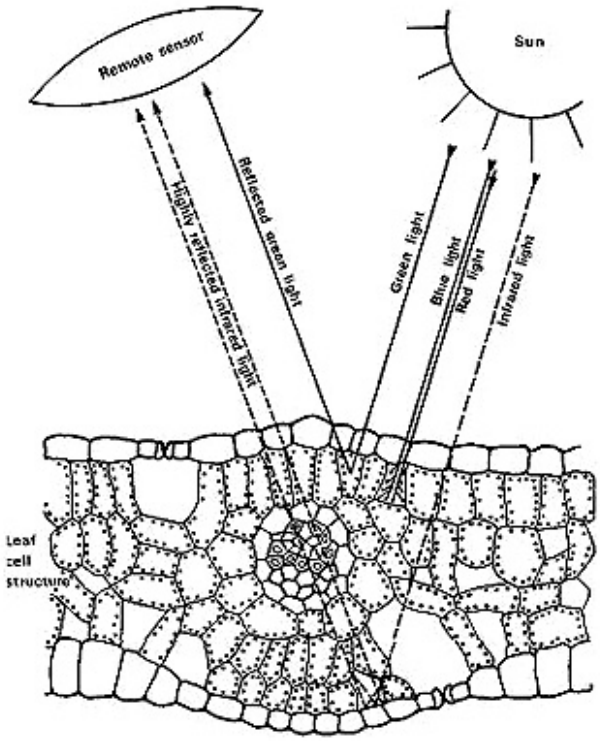
Remote sensing has proven a powerful "tool" for assessing the identity, characteristics, and growth potential of most kinds of vegetative matter at several levels (from biomes to individual plants). Vegetation behavior depends on the nature of the vegetation itself, its interactions with solar radiation and other climate factors, and the availability of chemical nutrients and water within the host medium (usually soil, or water in marine environments). A common measure of the status of a given plant, such as a crop used for human consumption, is its potential productivity (one such parameter has units of bushels/acre or tons/hectare, or similar units). Productivity is sensitive to amounts of incoming solar radiation and precipitation (both influence the regional climate), soil chemistry, water retention factors, and plant type. Examine the diagram below to see how these interact, keeping in mind that various remote sensing systems (e.g., meteorological or earth-observing satellites) can provide inputs to productivity estimation:



Because many remote sensing devices operate in the green, red, and near infrared regions of the electromagnetic spectrum, they can discriminate radiation absorption and reflectance properties of vegetation. One special characteristic of vegetation is that leaves, a common manifestation, are partly transparent allowing some of the radiation to pass through (often reaching the ground, which reflects its own signature). The general behavior of incoming and outgoing radiation that acts on a leaf is shown here:



Now, consider this diagram which traces the influence of green leafy material on incoming and reflected radiation.

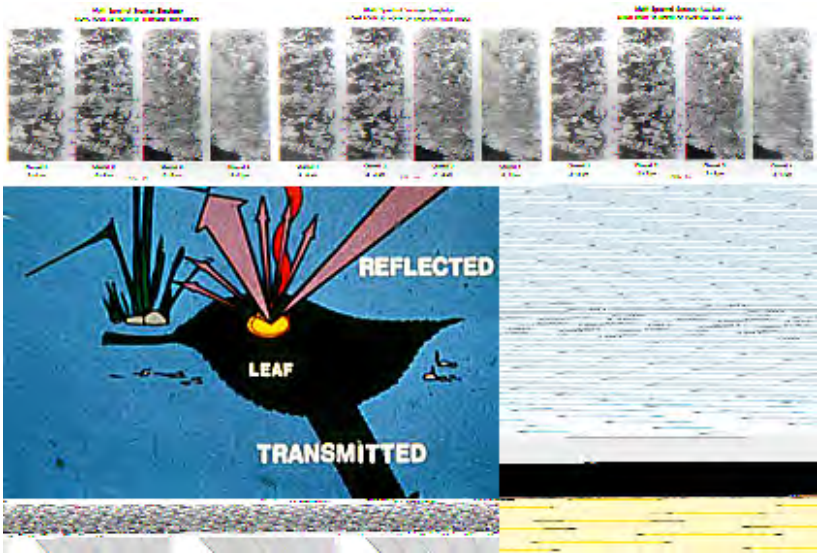


Absorption centered at about 0.65 µm (visible red) is controlled by chlorophyll pigment in green-leaf chloroplasts that reside in the outer or Palisade leaf. Absorption occurs to a similar extent in the blue. With these colors thus removed from white light, the predominant but diminished reflectance of visible wavelengths is concentrated in the green. Thus, most vegetation has a green-leafy color. There is also strong reflectance between 0.7 and 1.0 µm (near IR) in the spongy mesophyll cells located in the interior or back of a leaf, within which light reflects mainly at cell wall/air space interfaces, much of which emerges as strong reflection rays. The intensity of this reflectance is commonly greater (higher percentage) than from most inorganic materials, so vegetation appears bright in the near-IR wavelengths (which, fortunately, is beyond the response of mammalian eyes). These properties of vegetation account for their tonal signatures on multispectral images: darker tones in the blue and, especially red, bands, somewhat lighter in the green band, and notably light in the near-IR bands (maximum in Landsat's Multispectral Scanner Bands 6 and 7 and Thematic Mapper Band 4 and SPOT's Band 3).

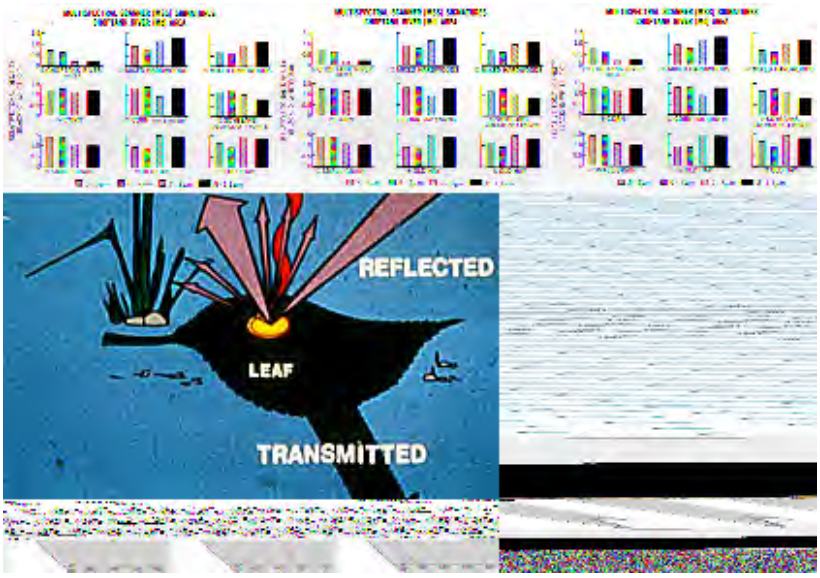
Identifying vegetation in remote-sensing images depends on several plant characteristics. For instance, in general, deciduous leaves tend to be more reflective than evergreen needles. Thus, in infrared color composites, the red colors associated with those bands in the 0.7 - 1.1 µm interval are normally richer in hue and brighter from tree leaves than from pine needles.

These spectral variations facilitate fairly precise detecting, identifying and monitoring of vegetation on land surfaces and, in some instances, within the oceans and other water bodies. Thus, we can continually assess changes in forests, grasslands and range, shrublands, crops and orchards, and marine plankton, often at quantitative levels. Because vegetation is the dominant component in most ecosystems, we can use remote sensing from air and space to routinely gather valuable information helpful in characterizing and managing of these organic systems.

The ability to distinguish different types of vegetation was brought home to the writer (NMS) through a simple study using a densitometer to examine multispectral images of a strip of agricultural land near the Choptank River in the eastern shore of Maryland. These images were part of an experiment by my first "boss", Dr. Warren Hovis, at Goddard. He had built a multispectral sensor to fly on an aircraft that would simulate images made by the same four bands on the ERTS-1 (Landsat-1) Multispectral Scanner (MSS). Here are the images:

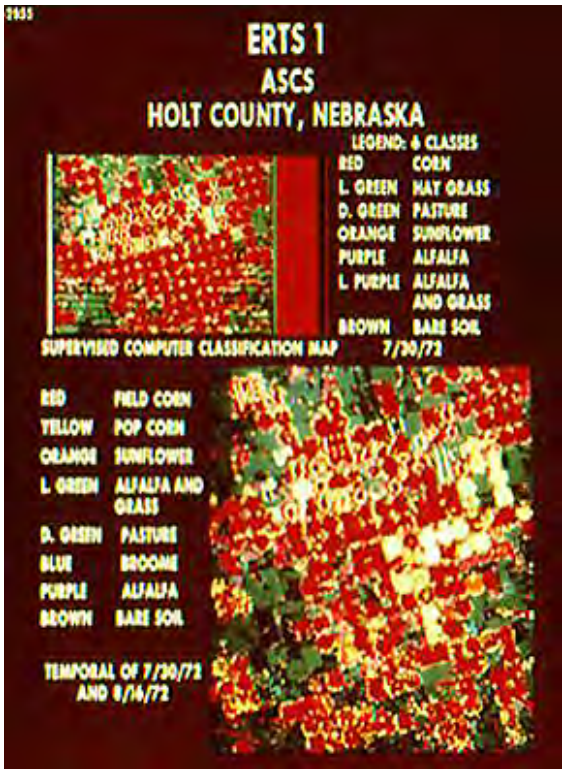


The relative gray levels are plotted as a four band histogram for each of the numbered features in the above image. It should be evident that there are real differences in these band signatures among the vegetation and other features present; thus Mixed Hardwoods have different relative "brightness" patterns from Soybeans, from Old Hay, etc..



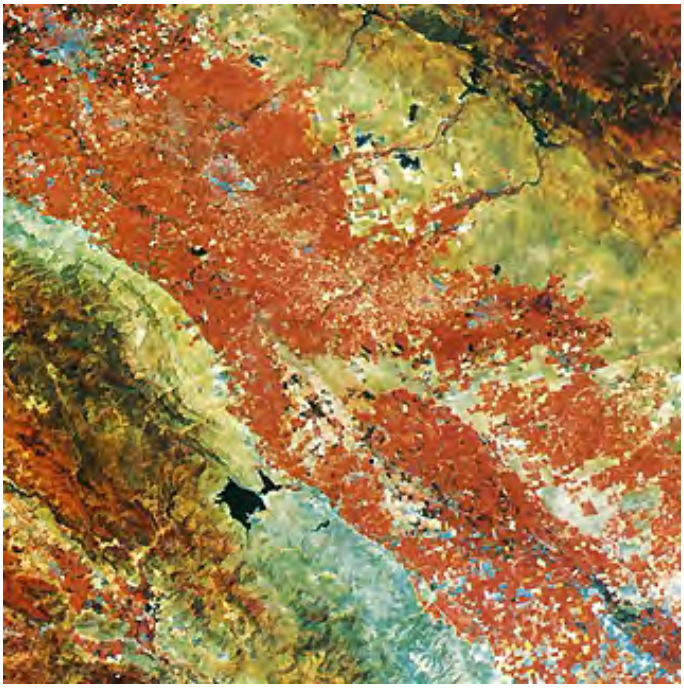
This discrimination capability implies that one of the most successful applications of multispectral space imagery is monitoring the state of the world's agricultural production. This application includes identifying and differentiating most of the major crop types: wheat, barley, millet, oats, corn, soybeans, rice, and others.

This capability was convincingly demonstrated by an early ERTS-1 classification of several crop types being grown in Holt County, Nebraska. This pair of image subsets, obtained just weeks after launch, indicates what crops were successfully differentiated; the lower image shows the improvement in distinguishing these types by using data from two different dates of image acquisition:



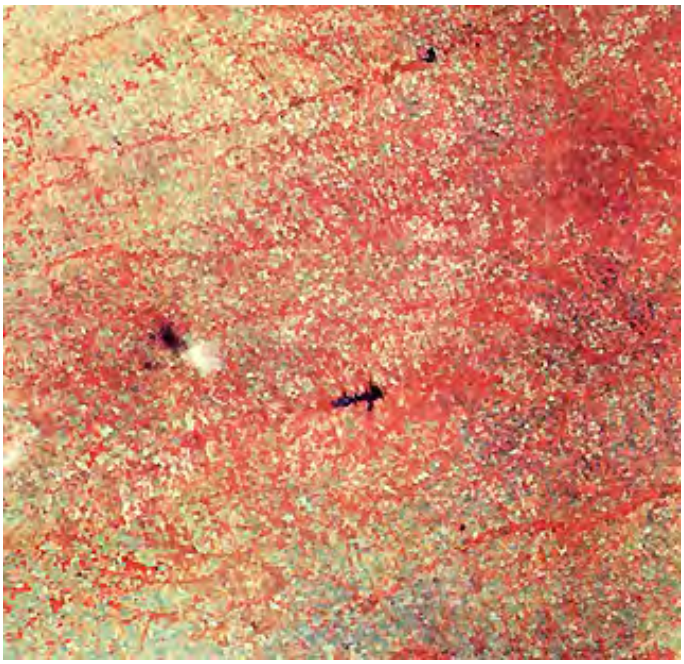
This is a good point in the discussion to introduce the appearance of large area croplands as they are seen in Landsat images. We illustrate with imagery that covers the two major crop growing areas of the United States.

The first scene is part of the Great or Central Valley of California, specifically the San Joaquin Valley. Agricultural here is primarily associated with such cash crops as barley, alfalfa, sugar beets, beans, tomatoes, cotton, grapes, and peach and walnut trees. In July of 1972 most of these fields are nearing full growth. Irrigation from the Sierra Nevada, whose foothills are in the upper right, compensates for the sparsity or rain in summer months (temperatures can be near 100° F). The eastern Coast Ranges appear at the lower left. The yellow-brown and blue areas flanking the Valley crops are grasslands and chapparal best suited for cattle grazing. The blue areas within the croplands (near the top) are the cities of Stockton and Modesto.

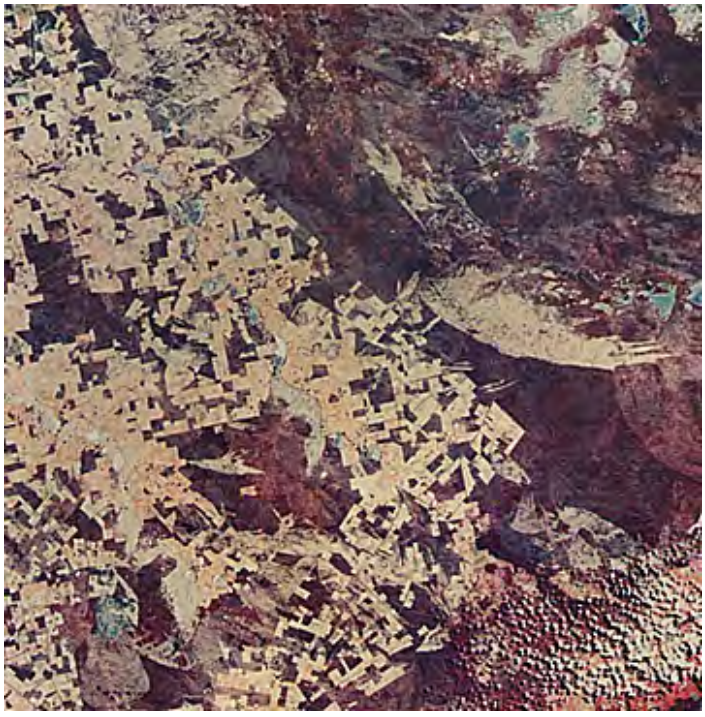


The second Landsat image is in the Wheat Belt of the Great Plains. The image below is of western Kansas in late August. Most of the scene consists of small farms, many of section size (1 square mile). The principal crop is winter wheat which is normally harvested by June. Spring wheat is then planted, along with sorghum, barley, and alfalfa. This scene is transitional, with

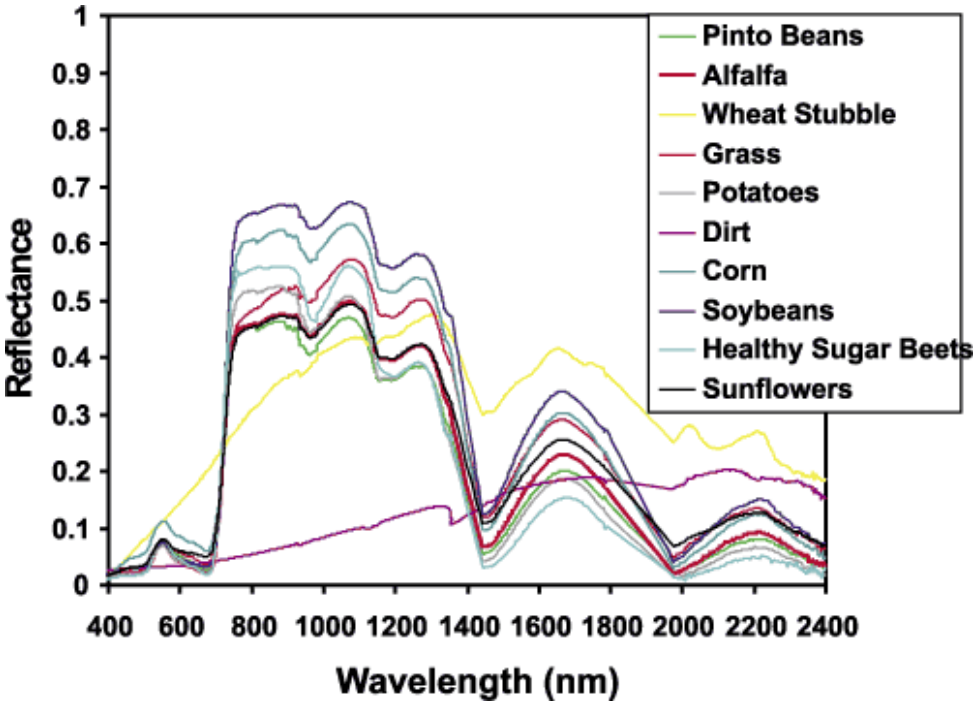
nearly all of the right side being heavily planted, but the left side (the High Plains, at higher elevations) contains some unplanted farms and cropfree land, some used for grazing.



Still another example of winter wheat in early growth is this scene in southwestern Australia, east of Perth. Some of the wheat fields are quite large - 5 km (3 miles) or more on a side. The prevailing color is tan but with a faint red cast, implying initial growth. There is a sharp line dividing many fields from the mallee scrub (dark brown) growing on soils derived from Precambrian rocks. This line marks an electrified rabbit fence, keeping these "pests" from nibbling on the wheat and other crops being grown.



Many factors combine to cause small to large differences in spectral signatures for the varieties of crops cultivated by man. Generally, we must determine the signature for each crop in a region from representative samples at specific times. However, some crop types have quite similar spectral responses at equivalent growth stages. The differences between crop (plant) types can be fairly small in the Near-Infrared, as shown in these spectral signatures (in which other variables such as soil type, ground moisture, etc. are in effect held constant).

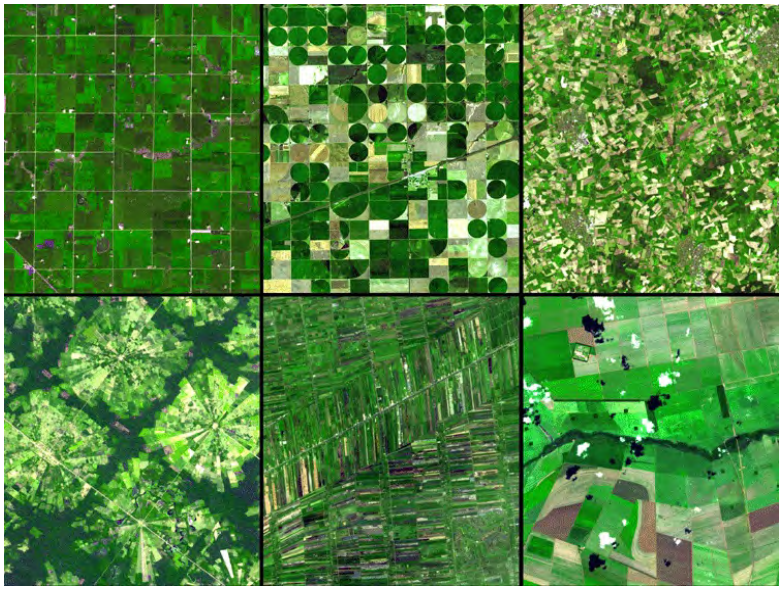


The shape of these curves is almost identical when each crop type is compared with the others. The big difference is in the percent reflectance. The similarity in shape is explained by the fact, discussed earlier on this page, that most vegetation matter has the same basic cell structure and similar content of chlorophyll. Yet remote sensing is reasonably effective at distinguishing and identifying different crop types. Why is that possible? These next two illustrations afford several clues:

3-1: **Drawing on your experience and common sense, make (or think) a list of the factors that will affect the spectral signatures of field crops, and thus help to separate crop types.** **ANSWER**

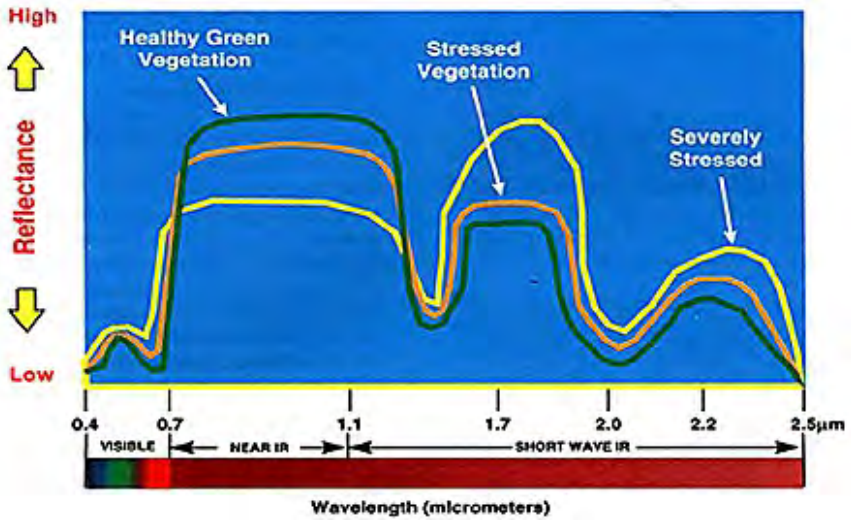
Read the answer to this question - it is important. The list is incomplete, but the main factors are discussed. But with so many variables involved, it is difficult to claim that each crop has a specific spectral signature. This means that, in order to identify the several crops usually present in agricultural terrain in any particular area, the most efficient course is to establish training sites, as was discussed on the pages dealing with supervised classification in Section 1.

As we learned in the Introduction Section, spectral characteristics are one means of identifying and classifying features in a scene. We will see how reliable this is by itself as this Section unfolds. We also learned in the Introduction that shape and pattern recognition are valuable inputs in determining what a feature is. The geometric shape of a field of crops sometimes is helpful in determining the actual crop itself. But field shapes tend to vary both within regions of large countries like the U.S. and in different parts of the world. This variation is evident in the illustration below (read the caption to find out which country goes with a particular panel):

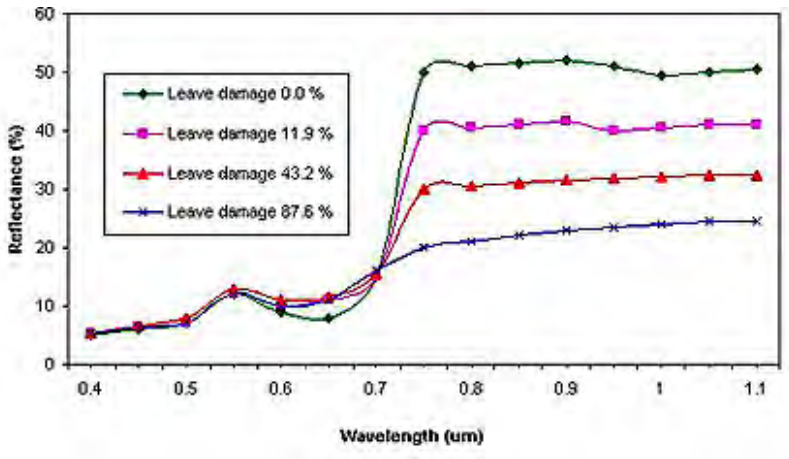


Through remote sensing it is possible to quantify on a global scale the total acreage dedicated to these and other crops at any time. Of particular import is the utility of space observations to accurately estimate (goal: best case 90%) the expected yields (production in bushels or other units) of each crop, locally, regionally or globally. We do this by first computing the areas dedicated to each crop, and then incorporating reliable yield assessments per unit area, which agronomists can measure at representative ground-truth sites (in the U.S., county farm agents obtain routinely from the farmers themselves). Reliability is enhanced by using the repeat coverage of the croplands afforded by the cyclical satellite orbits assuming, of course, cloud cover is sparse enough to foster several good looks during the growing season. Usually, the yield estimates obtained from satellite data are more comprehensive and earlier (often by weeks) than determined conventionally as harvesting approaches. Information about soil moisture content, often critical to good production, can be qualitatively (and under favorable conditions, quantitatively) appraised with certain satellite observations; that information can be used to warn farmers of any impending drought conditions.

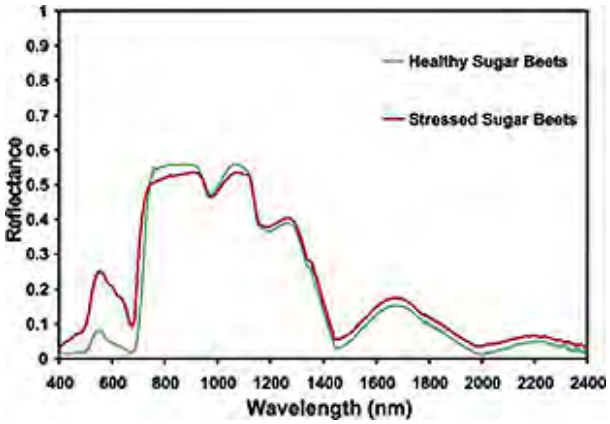
Under suitable circumstances, it is feasible to detect crop stress generally from moisture deficiency or disease and pests, and sometimes suggest treatment before the farmers become aware of problems. Stress is indicated by a progressive decrease in Near-IR reflectance accompanied by a reversal in Short-Wave IR reflectance, as shown in this general diagram:



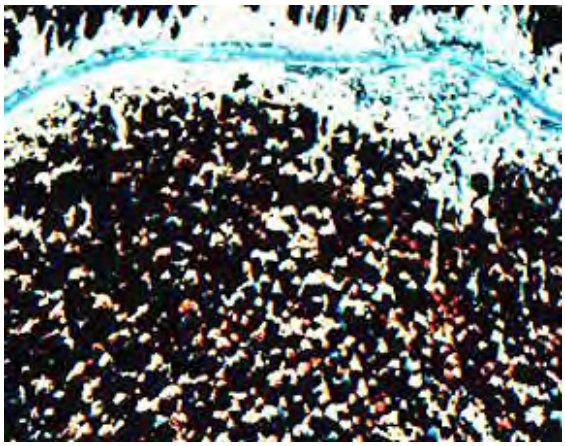
This effect is evidenced quantitatively in this set of field spectral measurements of leaves taken from soybean plants as these underwent increasing stress that causes loss of water and breakdown of cell walls.



For the soybeans, the major change with progressive stress is the decrease in infrared reflectances. In the visible, the change may be limited to color modification (loss of greeness), as indicated in this sugar beets example, in which the leaves have browned:



Differences in vegetation vigor, resulting from variable stress, are especially evident when Near Infrared imagery or data are used. In this aerial photo made with Color IR film shows a woodlands with healthy trees in red, and "sick" (stressed) vegetation in yellow-white (the red no longer dominates):

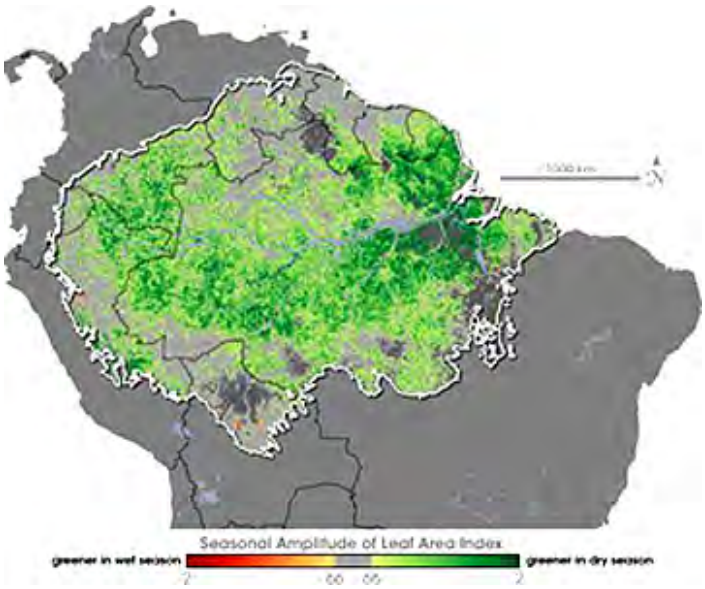


For identifying crops, two important parameters are the size and shape of the crop type. For example, soybeans have spread out leaf clumps and corn has tall stalks with long, narrow leaves and thin, tassle-topped stems. Wheat (in the cereal grass family) has long thin central stems with a few small, bent leaves on short branches, all topped by a head containing the kernels from which flour is made. Other considerations are the surface area of individual leaves, the plant height and amount of shadow it casts, and the spacing or other planting geometries of row crops (the normal arrangement of legumes, feed crops, and fruit orchards). The stage of growth (degree of crop maturity) is also a factor. For example during its development wheat passes through several distinct steps such as developing its kernel-bearing head and changing from shades of green to golden-brown (see below).

Another related parameter is Leaf Area Index (LAI), defined as the ratio of one-half the total area of leaves (the other half is the underside) in vegetation to the total surface area containing that vegetation. If all the leaves were removed from a tree canopy and laid on the ground, their combined areas relative to the ground area projected beneath the canopy would be some number greater than 1 but usually less than 10. As a tree, for example, fully leaves, it will produce some LAI value that is dependent on leaf size and shape, the number of limbs, and other factors. The LAI is related to the the total biomass (amount of vegetative matter [live and dead] per unit area, usually measured in units of tons or kilograms per hectare [2.47 acres]) in the plant and to various measures of Vegetation Index (see below). Estimates of biomass can be carried out with variable reliability using remote sensing inputs, provided there is good supporting field data and the quantitative (mathematical) models are efficient. Both LAI and NDVI (page 3-4) are used in the calculations.

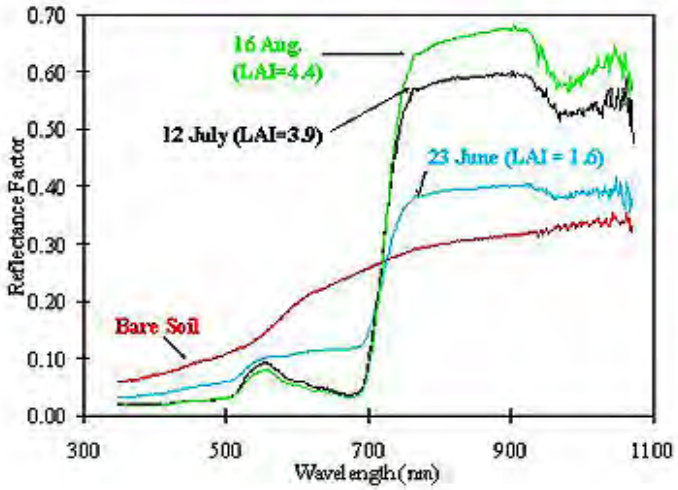
Satellite remote sensing is an excellent means of determining LAI on a regional or subcontinental scale. Here is an LAI map of northern South America and southern Central America obtained using Landsat data. The Amazon rainforest (see page 3-5) and the tropical forests on the eastern slopes of Costa Rica and Panama are well-defined by the LAI distribution (they diminish rapidly, in part because of rainfall patterns):

Studies of LAI for the Amazon vegetation led to an unexpected discovery. Examine this map:

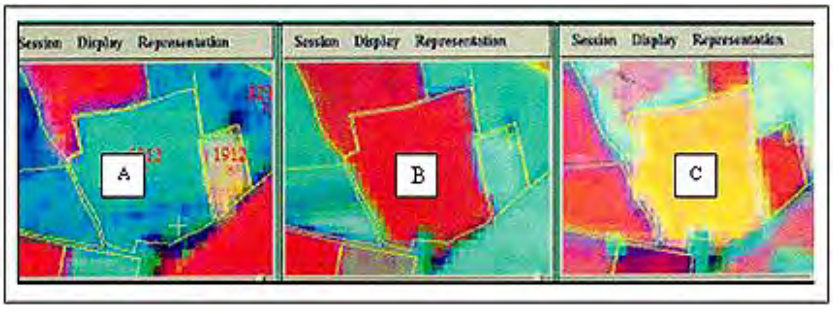


The largest change in the "greeness" of the tree leaves in the Amazon occurred, not as anticipated in the wet season, but in the dry season. The explanation: water is stored in the near surface vadose zone (part of upper soil in which groundwater percolates downward) and below the top of the water table. That water is tapped by the plants during the dry period so that leaves are able to continue growth.

In principal, actual LAI must be determined on site directly by stripping off all leaves, but in practice it can be estimated by statistical sampling or by measuring some property such as reflectance. Thus, remote sensing can determine an LAI estimate if the reflectances are matched with appropriate field truth. For remotely sensed crops, LAI is influenced by the amount of reflecting soil between plant (thus looking straight down will see both corn and soil but at maturity a cornfield seems closely spaced when viewed from the side). For the spectral signatures shown below, the Near IR reflectances will increase with LAI.

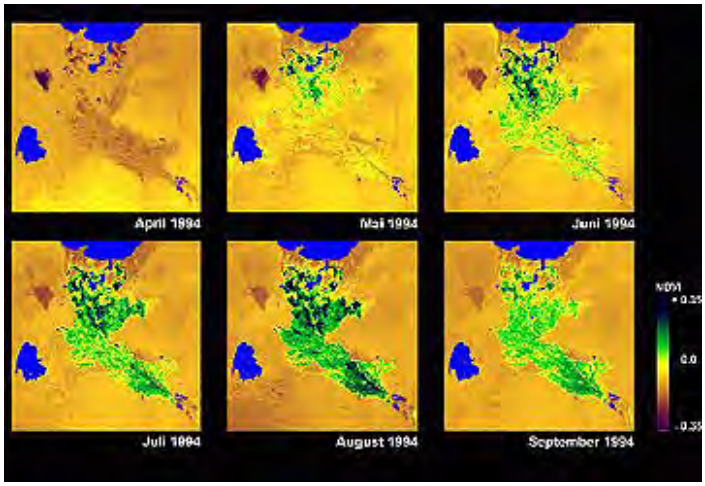


This change in appearance and extent of surface area coverage over time is the hallmark of vegetation as compared with most other categories of ground features (especially those not weather-related). Crops in particular show strong changes in the course of a growing season, as illustrated here for these three stages - bare soil in field (A); full growth (B); fall senescence (C), seen in a false color rendition:

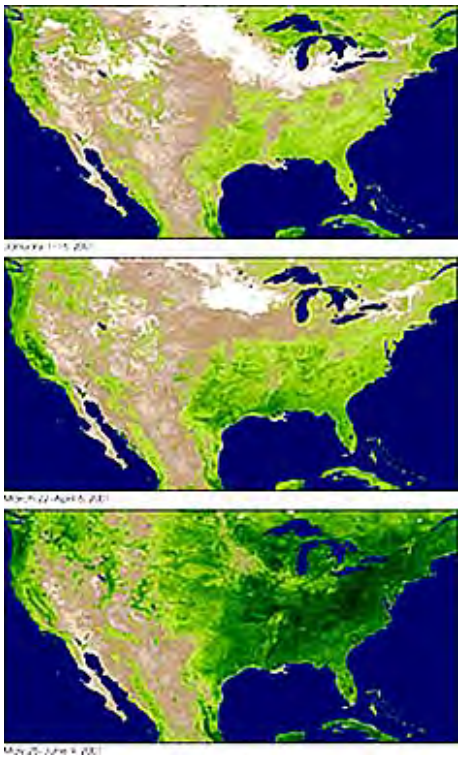


3-2: **How would non-growing or dead vegetation (such as crops in senescence) be detected by Landsat? ANSWER**

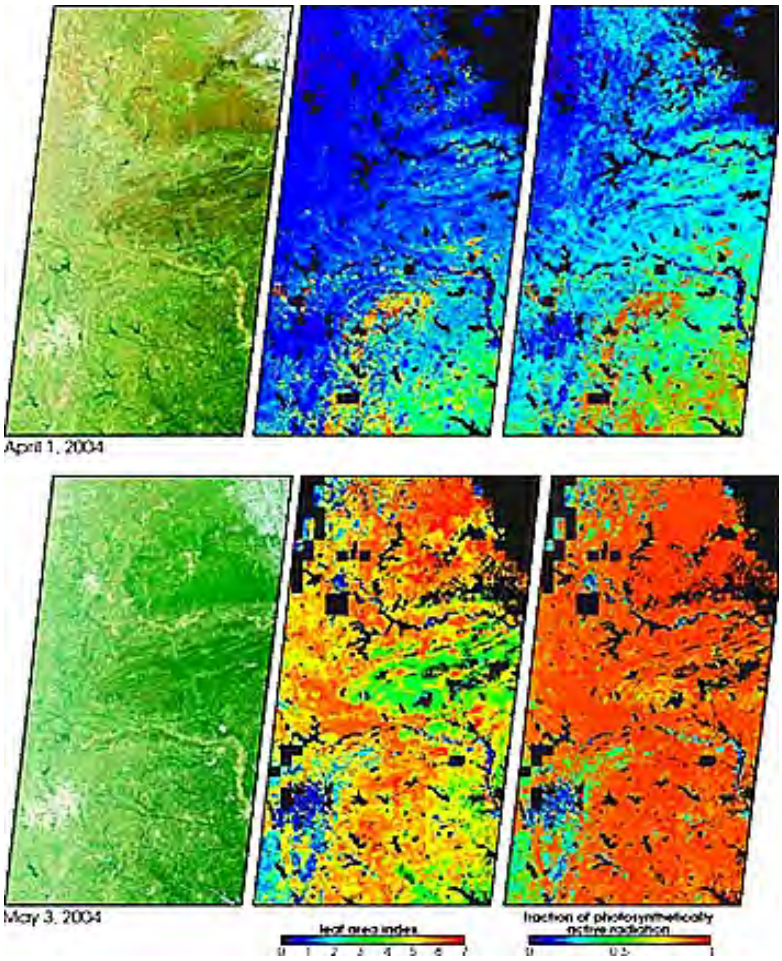
The study of vegetation dynamics in terms of climatically-driven changes that take place over a growing season is called *phenology*. A good example of how repetitive satellite observations can provide updated information on the phenological history of natural vegetation and crops during a single cycle of Spring-Summer growth is this sequence of AVHRR images of the Amu-Dar'ja Delta just south of the Aral Sea in Ujbekistan (south-central Asia).The amount of vegetation present in the delta (a major farming district for this region) is expressed as the NDVI, an index defined on page 3-4. The Aral Sea - a large inland lake - is now rapidly drying up (see page page 14-15).



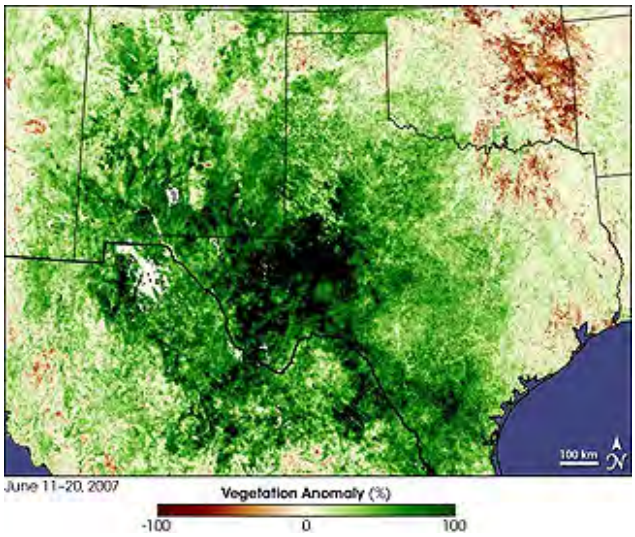
More generally, seasonal change appears each year with the "greening" that comes with the advent of Spring into Summer as both trees and grasses commence their annual growth. The leafing of trees in particular results in whole regions becoming dominated by active vegetation that is evident when rendered in a multispectral image in green tones. The MODIS sensor on Terra has several vegetation-sensitive bands used to calculate a variation of the NDVI called the Enhanced Vegetation Index (EVI). This trio of images (dates in the caption) shows the spread of growing natural vegetation across the U.S.



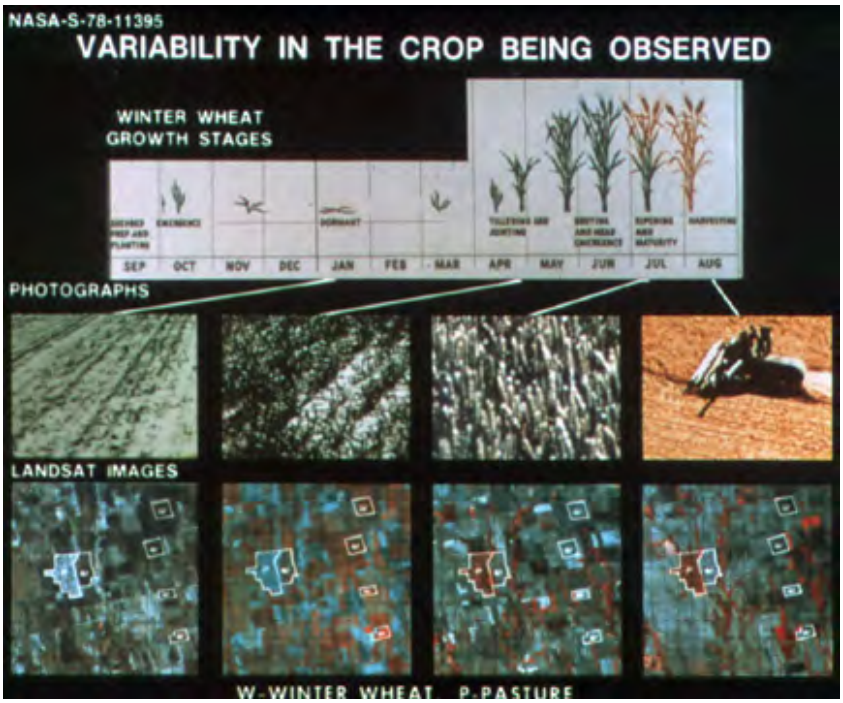
During the first stages of growth in the Spring season in the eastern half of the U.S., drastic changes in vegetation signatures will ensue. In the 6-panel figure below, the left two panels (in the top and bottom triads) show a near natural color MISR image of a strip in the central U.S. that includes the western Ouachita Mountains of Oklahoma-Texas on April 1, 2004 and May 3, 2004. In the center pair (top and bottom), LAI values are calculated for both dates - by May the barren trees and grasslands had almost fully leafed. On the right is shown a map that displays the fraction of photosynthetically active radiation for the two dates - photosynthesis, as expected, has neared full term by May



A variant of these ideas is the abnormal greening of an area as a result of an extended period of excessive rainfall. This happened in Texas in June of 2007 when as much as 20 inches of rain fell during the month. This mosaic made from SPOT imagery (see next page) shows that most of the western two-thirds of the state has a far greener vegetation response than normal at this time of year (the blackish area, around the Midland Basin, is the extreme):



Now, to emphasize the variability of the spectral response of crops over time, we show these phenological stages for wheat in this sequential illustration:

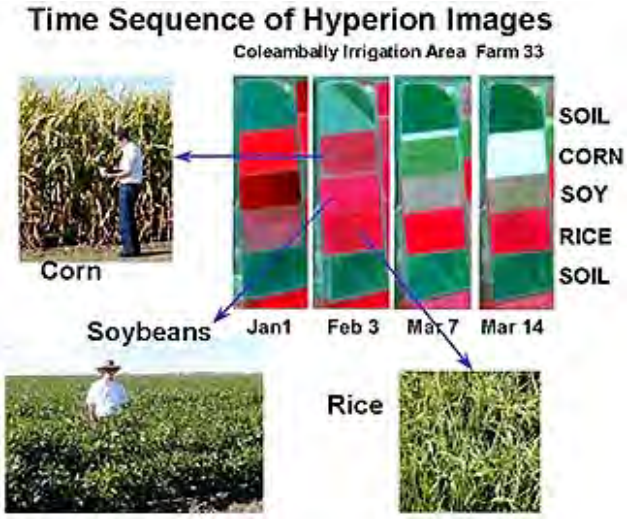


Note that, in the Landsat imagery, the wheat fields (particularly the light-blue polygon in the far-left image) show their brightest response in the IR (hence red) during the emergent stage but become less responsive by the ripening stage. The grasses and alfalfa that make up pasture crops mature (redden) much later.

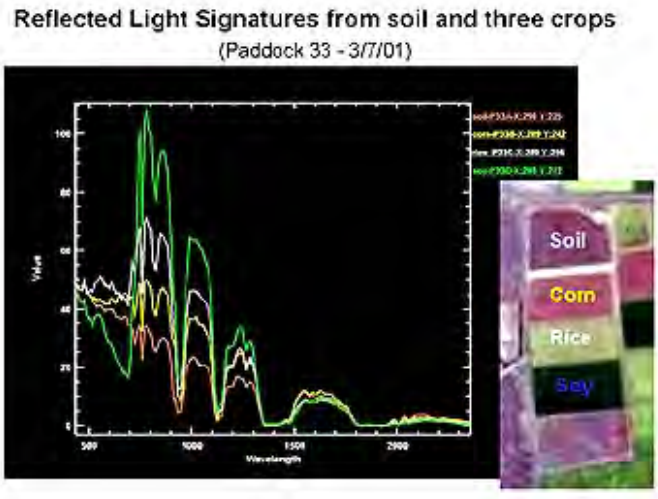
3-3: **The display above was made as part of the LACIE program, designed to demonstrate that crop history can be monitored by satellite and that productivity (yield in bushels) and prediction of total yield in a season from regions of major productions can be quantitatively assessed. How were the specific crop types used as training sites identified (determined)? ANSWER**

With this survey of the role of several variables in determining crop types, let us look now at one of the most successful classifications reported to date. These are being achieved by hyperspectral sensors such as AVIRIS (page 13-9) and Hyperiorn. The Hyperion hyperspectral

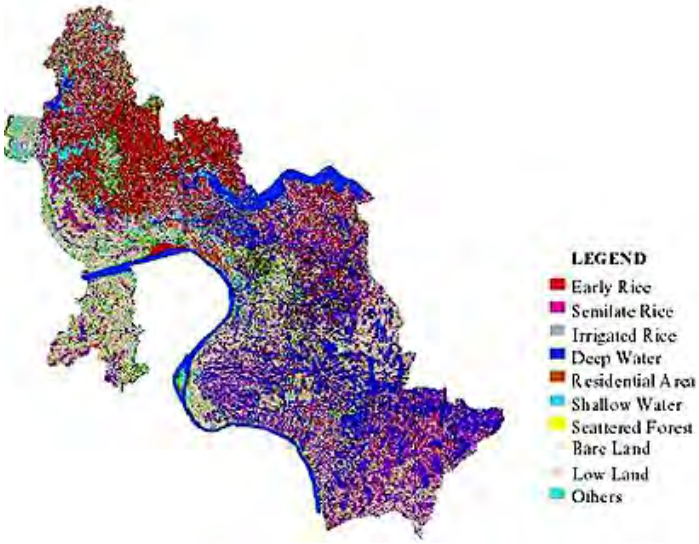
sensor on NASA's EO-1 ((Page Intro-24) has procured multichannel data for the Coleambally test area in Australia. This image, made from 3 narrow channels in the visible-Near IR, shows how the fields of corn, rice, and soybeans changed their reflectances during the (southern hemisphere) growing season: Notice the pronounced differences in crop shapes which is a big factor in producing the reflectance differences (as said above, healthy leaf vegetation generally has a spectral response that does not vary much in percent reflectance from one plant type to others, so that differences in crop shape become the distinguishing factor).



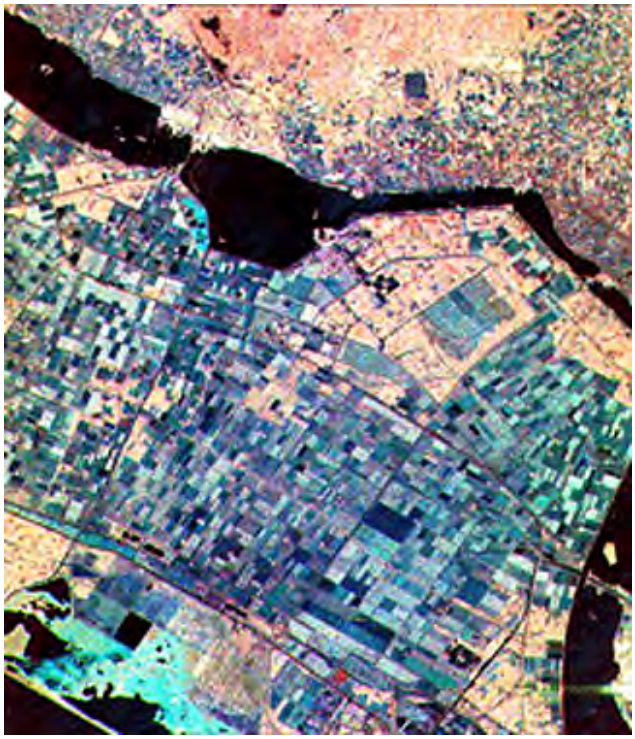
The multichannel data from Hyperion were used to plot the observed spectral signatures for the soil and three crops, as shown here (the curves identified in the upper right [the writing is too small to be decipherable on most screens] are, from top to bottom, soil, corn, rice, and soybeans):



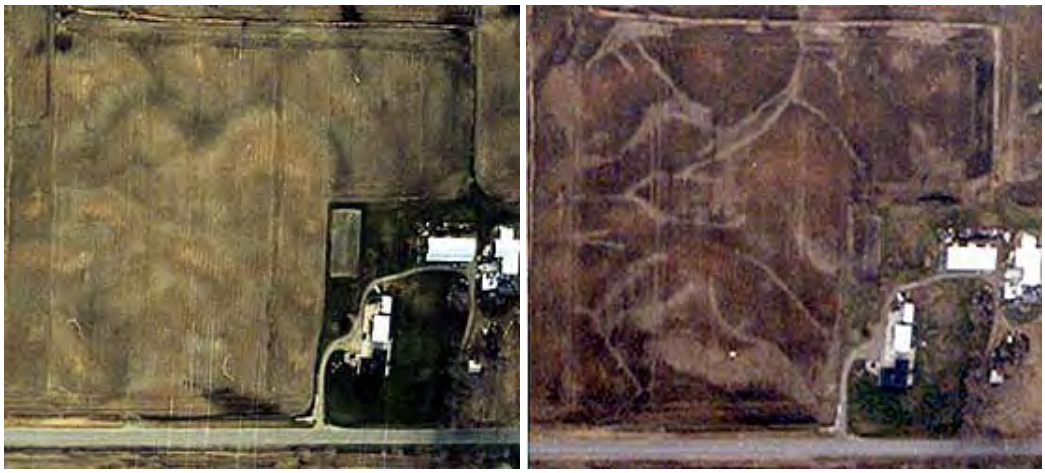
Using a large number of selected individual Hyperion channels, this supervised classification of the four classes in the subscene was generated; this end result is more accurate than is normally achievable with broad band data such as obtained by Landsat:



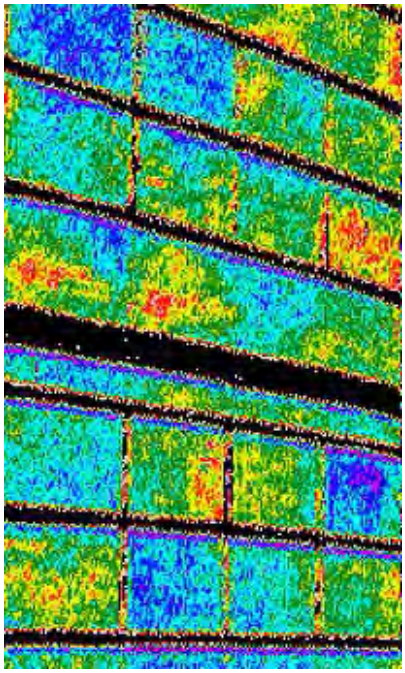
Active microwave sensors, or radar, can use several variables to recognize crop vegetation and even develop a classification of crop types. Here is a SIR-C (Space Shuttle) image of farmland in the Netherlands, taken on April 4, 1994. The false color composite was made with L-band in the HH polarization mode = red; L-band HV = green; and C-band HH = blue (see page 8-5).



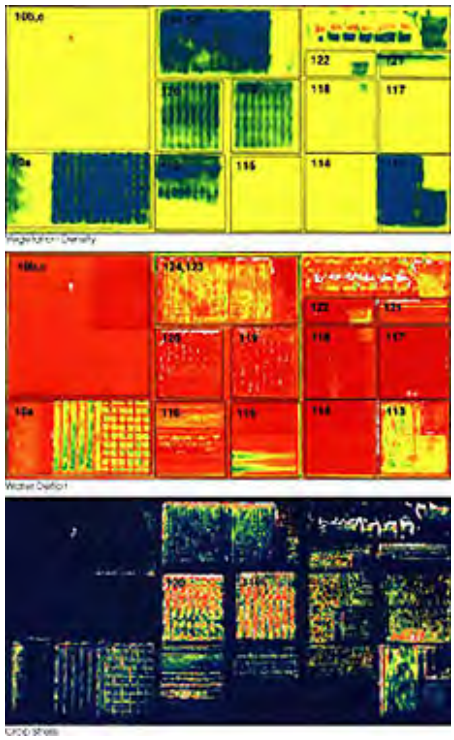
An additional image variable is the crop's background, namely the nurturing soil, whose color and other properties can change with the particular soil type, and whose reflectance depends on the amount of moisture it holds. Moisture tends to darken a given soil color; this condition is readily picked up in aircraft imagery as seen in this pair of images:



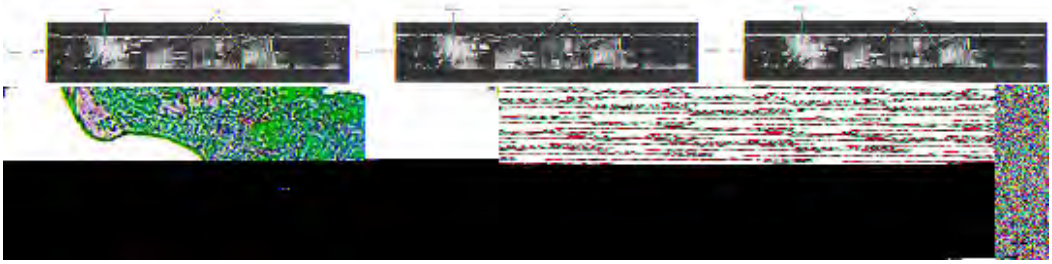
Often, the distribution of moisture, as soil dries differentially, is variable in an imaged barren field giving rise to a mottled or blotchy appearance. Thermal imagery (Section 9) brings out the differential soil moisture content by virtue of temperature variations. The amount of water in the crop itself also affects the sensed temperature (stressed [water deficient] or diseased crop material is generally warmer). Soil water variations are evident in this image made by an airborne thermal sensor of several fields, where high moisture correlates with blue and drier parts of the fields with reds and yellows:



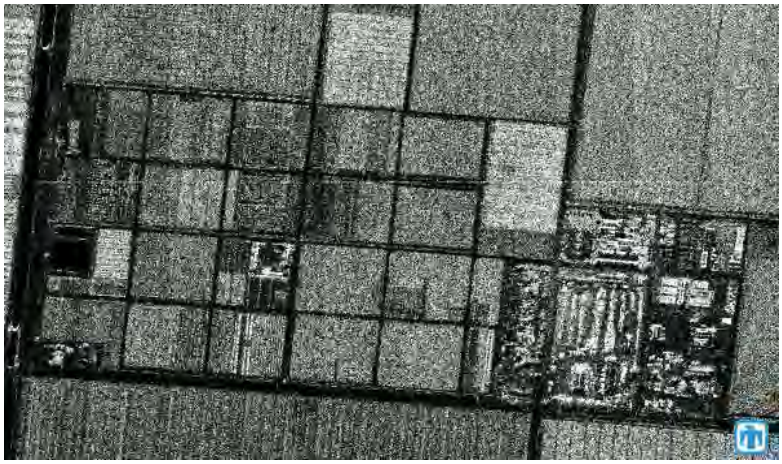
A combination of visible, NIR, and thermal bands can pick up both water deficiency and the resulting stress on the crops in the fields. This set of three images was made by a Daedalus instrument flown on an aircraft. In the top image, yellow marks unplanted fields and those in blue and green are growing crops. The center image picks up patterns of water distribution in the crop fields. The bottom image shows levels of stress related in part to insufficient moisture.



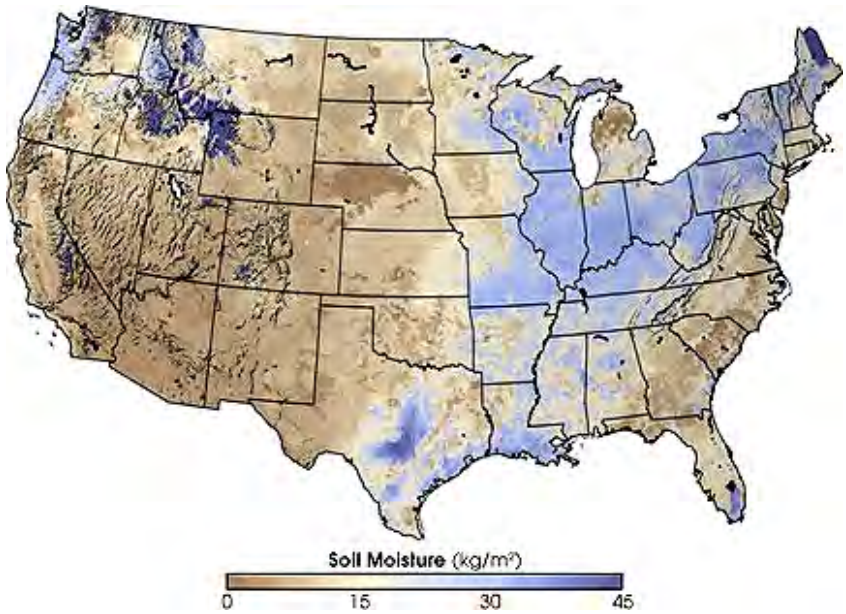
A passive microwave sensor (page 8-8) also picks up soil moisture. Cooler areas appear dark in images of fields overflown by a microwave sensor - although other factors, such as absence or presence of growing crops (and their types) besides moisture can account for some darker tones:



Radar likewise can detect variations in soil moisture in agricultural fields. Below is a C-band airborne SAR image of an experimental station at Maricopa, AZ near Phoenix. The darker fields are those with both higher moisture and growing crops which, in this case, result in less signal returns to the SAR receiver.



A major goal in the EOS program (Section 16) is to produce soil moisture maps on a short-term basis (say, for two weeks running, available almost immediately thereafter). Various sensors on Terra and Aqua can provide data needed to calculate regional soil moisture distribution. Here is a map for the entire United States covering the period July 1-15, 2003:



This map indicates an abundance of moisture in the northeast and upper Central Lowlands (various areas therein have been affected by one of the wettest seasons in recent decades) and a continuing (for several years) drought in most of the western half of the U.S. (a condition responsible for an abnormally high number of forest fires (see Overview).

Before moving on to some specific examples of vegetation analysis, we append this interesting image of a variant of the topic of Crop Signs - which refers to strange, often unexplained circles and other figures cut into maturing crops (commonly corn fields). In this case, the perpetrator was not some mysterious "alien" source but a patriotic farmer named Fritzler whose farm is outside Greeley, Colorado:

And since this page is ending with the unusual, here is another "oddity" that defies categorization (which means it is hard to find a proper page to display it). Farming calls to mind crops that are 'vegetation'. In a broader sense, animals are said to be farm-raised. But fish!! Yet aquaculture is a growing industry. For example, catfish are now mostly harvested from ponds on 'farms' that produce this as a sole (no pun intended) crop. On a large scale, Egyptian farmers grow fish in collection ponds in the Nile Delta, as displayed in this astronaut photo:

This part of the Delta contains about a half of Egypt's aquacultural industry. That industry sprang up after the Aswan Dam was built. The dam's ecological impact has diminished the nutrients that helped to feed fish in the Nile River. The managed Delta fisheries are the replacement of the depleted river supply.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The SPOT satellite, operated by the French, first mentioned in the Introduction, is here described in detail. Its earlier version produced three bands of data in the green, red, and near-IR segments of the spectrum. It also has a panchromatic band and can be pointed off nadir to give stereo imagery (dependent on passes near the scene on different days). SPOT is a competitor to Landsat but differs by having fewer bands but (compared to Landsats 1-5) higher resolution. Both SPOT and Landsat images covering an area in the East African Rift Zone of Kenya are examined, with special attention given to the appearance and classification of vegetation and crop fields. A second classification of a notable number of different crops growing along coastal Holland is included on this page as an indication of the versatility of Landsat TM data to identify crop types.

The SPOT Satellite: A Case Study in Africa; Further Applications to Vegetation Mapping of Crops and Rangelands

We interject the following synopsis of another operational satellite which, like Landsat, can provide very sharp and timely images well suited to agricultural analysis. We will then use SPOT to conduct a study of vegetation in Africa.

SPOT is an operational series of Earth-observing satellites, sponsored initially by the French government and managed by the French Space Agency, which, since the first launch in 1986, offers commercial multispectral data to users worldwide. From its 831 km (516 mi), sun-synchronous (10:30 A.M. equatorial crossing), near-polar orbit, SPOT revisits the same 117 km-wide (73 mi) ground tracks every 26 days, when its sensors look vertically downward (at nadir). However, operations can move its mirror sideways in both directions up to 27 degrees off-nadir, enabling it to view a 950 km-wide (590) corridor along its ground track. With careful planning, and optimal cloud conditions, it can revisit an area within this corridor seven times during a 26 day orbit cycle.

Images of the same area taken from different lateral angles on separate orbital passes can be paired, so that their parallax offsets permit stereoscopic viewing, thus facilitating a 3-D perspective of the area (see Section 11). These images are of great value to topographic analysts, geomorphologists, and others.

The payload sensor on the first SPOTs is the HRV (High Resolution Visible) imager. This name is slightly misleading, because in its multispectral mode, Band 3 operates in the near-IR wavelength over the interval 0.79 - 0.89 µm (Band 1 = 0.50 - 0.59 µm [green] and Band 2 = 0.6 - 0.68 µm [red]). A separate unit in the HRV can also image in the panchromatic mode as a single band encompassing 0.51 - 0.70 µm. Instead of an oscillating scan mirror, the tiltable mirror (set either on or off-nadir) sends light from a scene to a linear array of tiny detectors, known as charged coupled devices (CCDs). (See page I-5a for details on how these devices function.) The line of CCDs receives photons simultaneously and very rapidly discharges in sequence into an electron current stream. This array is then reactivated with photons almost instantly, as the array line moves to the next ground position along the forward track. In this way, the scene is developed line by line. In the panchromatic mode, there are 6,000 detectors in the linear array, each receiving light from a 10 x 10 m ground area, thus providing a resolution of 10 m for an image of 60 km (37 mi) width. The multispectral HRV contains 3000 detectors in the array, providing a ground resolution of 20 m. On each spacecraft there are two such HRV arrays, mounted side-by-side, each viewing its own 60 km swath width but the pair of images sidelapping to give an effective 117 km (73 miles) wide coverage.

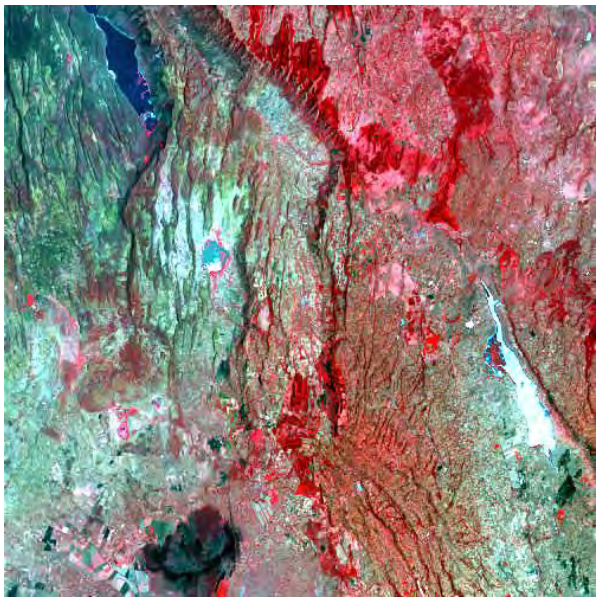
Later SPOT satellites have improved resolution. They cover smaller scene areas. Many SPOT images are near ideal for certain applications such as viewing a city in context with its surroundings. As an example, look at this SPOT image of Marseille, France:

3-4: **What are the major advantages and disadvantages of Landsat compared with SPOT? ANSWER**

SPOT has a large number of image on the Web, under the categories of Nature and Cities, that can be accessed on SPOT gallery.

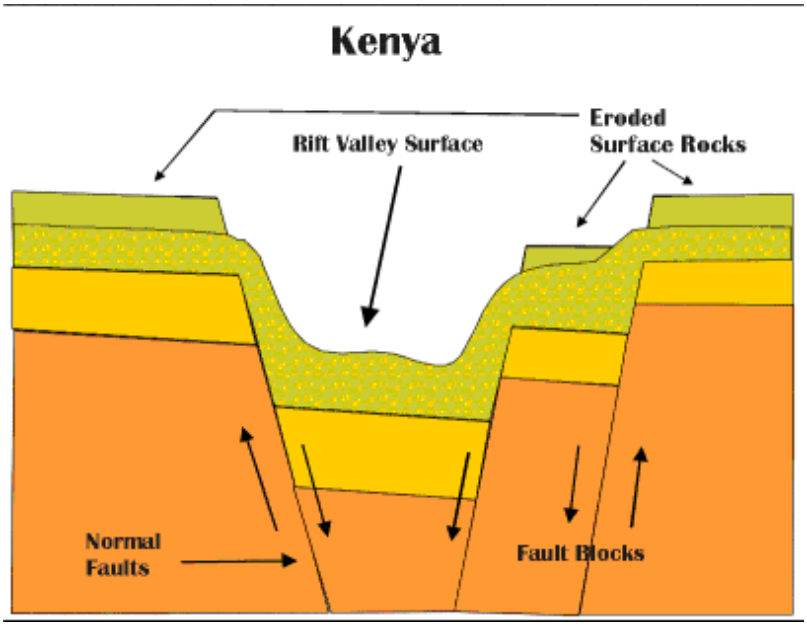
Kenya and the African Rift Valley

Now that we've described SPOT, lets look at one of its images and see what you can spot (couldn't pass up that one!). In Section 2 you have earlier examined a SPOT scene of the Great African Rift valley which we will repeat here for context.



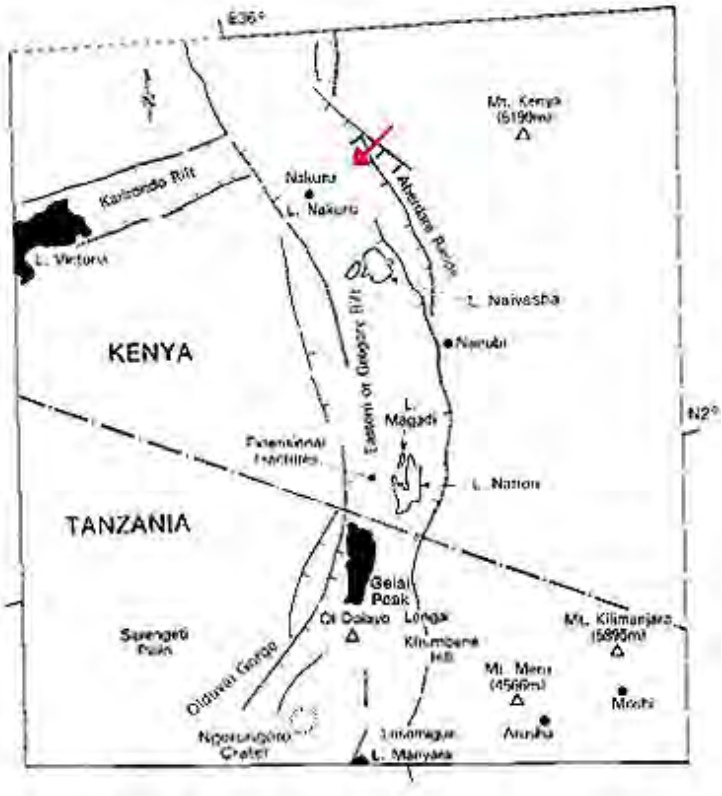
To put the SPOT image in context, we need first to say something about the Rift Valley. This rift zone, one of the largest in the world, is astride a regional uplift, where part of the eastern African continent is pulling away (an incipient divergent boundary in the plate tectonics model). Within the rift, parts of the crust are dropping down along normal faults in a series of one or more steps as shown in the figure below. If normal faults are paired on opposite sides of a valley, this is

evidence of a strutural graben, shown diagrammatically here:.



Some faults are young enough to remain relatively uneroded, so that they present a steep, often abrupt, face (called a scarp), separating the upthrown (upland) block from the downthrown (valley) block, as evident in the ground photos below.

Along much of the rift, the land is an arid to semi-desert savannah (top photo, above). In other segments, the scarp is older, and weathered, and may be covered in part with forest or dense shrub, such as in the lower (or right) photo above. Most of the rift zone is underlain by a sequence of basaltic (volcanic) flow units. The next scene is a regional mosaic made from all or parts of nine Landsat MSS images, of much of Kenya that extends over an area about 500 km (311 mile) wide, east-west.

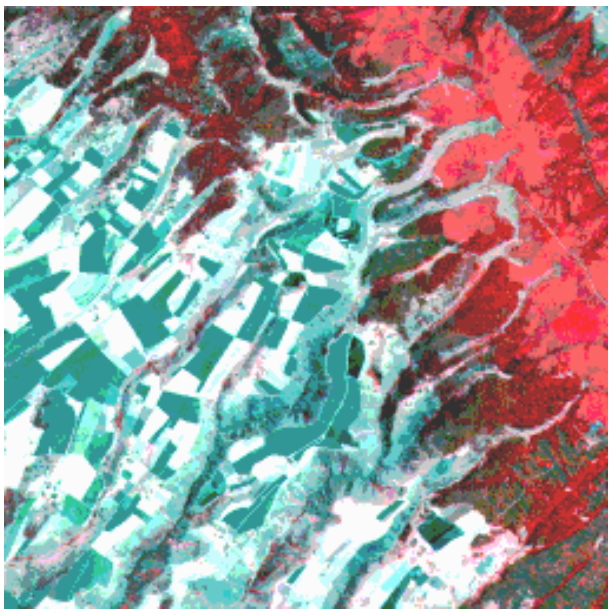


Several of the volcanic mountains, in particular Mt. Kilimanjaro shown at the bottom of the map, are displayed as a nearly horizontal perspective view using combined Landsat-SRTM data:

Match major landmarks in the above mosaic, with those in the index map including Mt. Kenya (in Kenya) and Mt. Kilimanjaro in Tanzania, the latter being over 5,230 m (17,154 ft) high and

therefore, snow-capped. Kilimanjaro supports active glaciers, as seen from this International Space Station as photographed by an astronaut:

Now to the specific subscene that we will examine for vegetation and classify for crops. This example is an agricultural and forestry scene, just east of the town of Nakuru in Kenya, about 100 kms (60 mi) northwest of Nairobi, the capital. That town (shown on page 6-13a) lies within the Great African Rift Valley of East Africa (described earlier on page 2-9). The 10 km-wide subscene below was extracted from a full 60 km SPOT image, which was produced as a false color composite. We will analyze this subscene in the next paragraph.

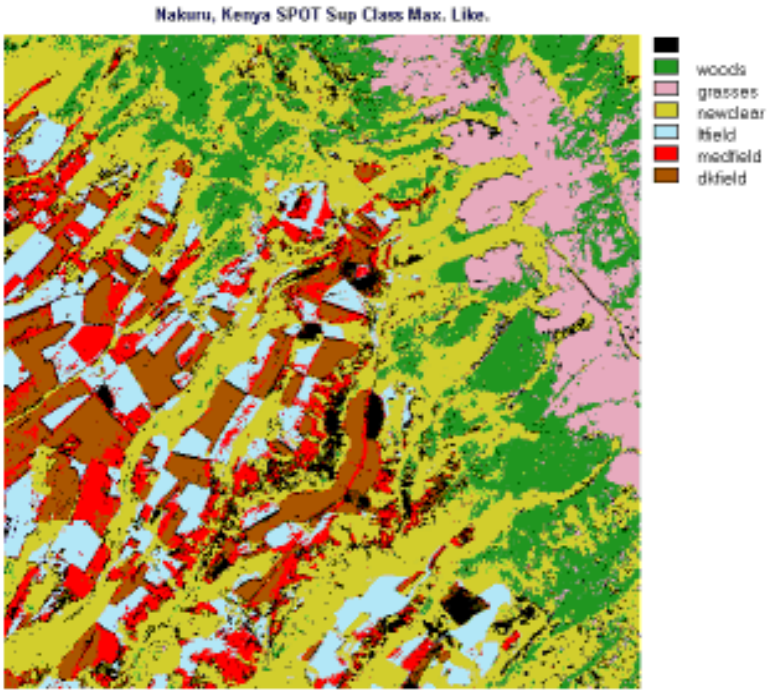


Whereas details are lacking in the mosaic, the SPOT image shows a plentiful amount of surface information. The edge of the Rift Valley passes from upper center to lower right, as marked by dark red areas that are associated with forest lands. To its right is an area of rather uniform red, without any treelike textural pattern, which are interpreted as grass and shrub lands. The area to the left (west) is occupied mainly by irregular-shaped fallow fields, many of which are about 0.8 km (0.5 mi) long. Some fields appear in this color version as almost white, whereas others are in three dominant shades of blue (darker, medium, light blue). Inspecting the three individual SPOT bands (not reproduced here), the white fields are nearly white in each band and the blue fields are medium gray in band 1, darker in Band 2, and very dark gray in Band 3. The lack of any red areas in these two types of fields confirms that crop growth (mainly of coffee and wheat in

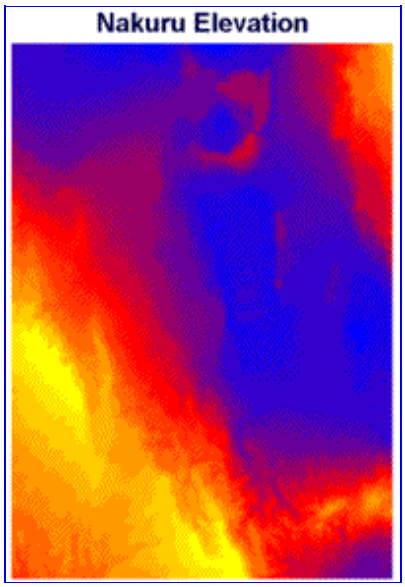
this part of Kenya) is nil or too early to appear. Not having been there, we do not have any adequate explanation for this pronounced difference in tone between these fields. The dark tones could represent the prevailing color (also dark) for soils derived from the very dark basalts. But this fails to explain the light blue tones for the other fields. They might be the signature of a grain crop, such as unharvested mature wheat (hence appearing bright straw yellow as seen onsite), thus being very reflective in Band 1, much less so in Band 2 (ground reds), and least reflective in the IR band. Any further speculation without supporting ground truth would be frustrating. Note the pink tones in several areas of the scene including along stream courses and in the narrow stream canyons that are cut into the rift wall slopes. These pink tones may be due to two conditions: the foliage of some trees has only begun to develop new growth in this March 14, 1986 scene and the vegetation there is experiencing some stress due to deforestation as local farmers progressively clear the land.

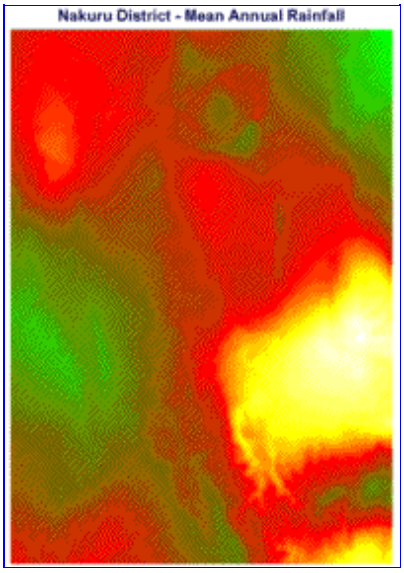
3-5: **Which part of the scene - the reddish grasslands and forests or the areas of fields (shades of blue) - is lower in elevation? There is a clue to this. What is it? ANSWER**

Some of the above ideas are implicit in the supervised maximum likelihood classification of this scene, in which the classes chosen are: woods, grasses, newclear (new growth and cleared land), ltfield, medfield, dkfield (light, medium and dark fields). The results are quite believable, in that the color patterns tend to be homogeneous and their distributions are logical. The extent of newclear is somewhat greater than we can discern in black & white images. There are some black areas in the classified image. Those black areas that appear scattered are simply unclassified pixels. But several areas have definite shapes and are probably small lakes or ponds (there are flamingo breeding grounds in this region).



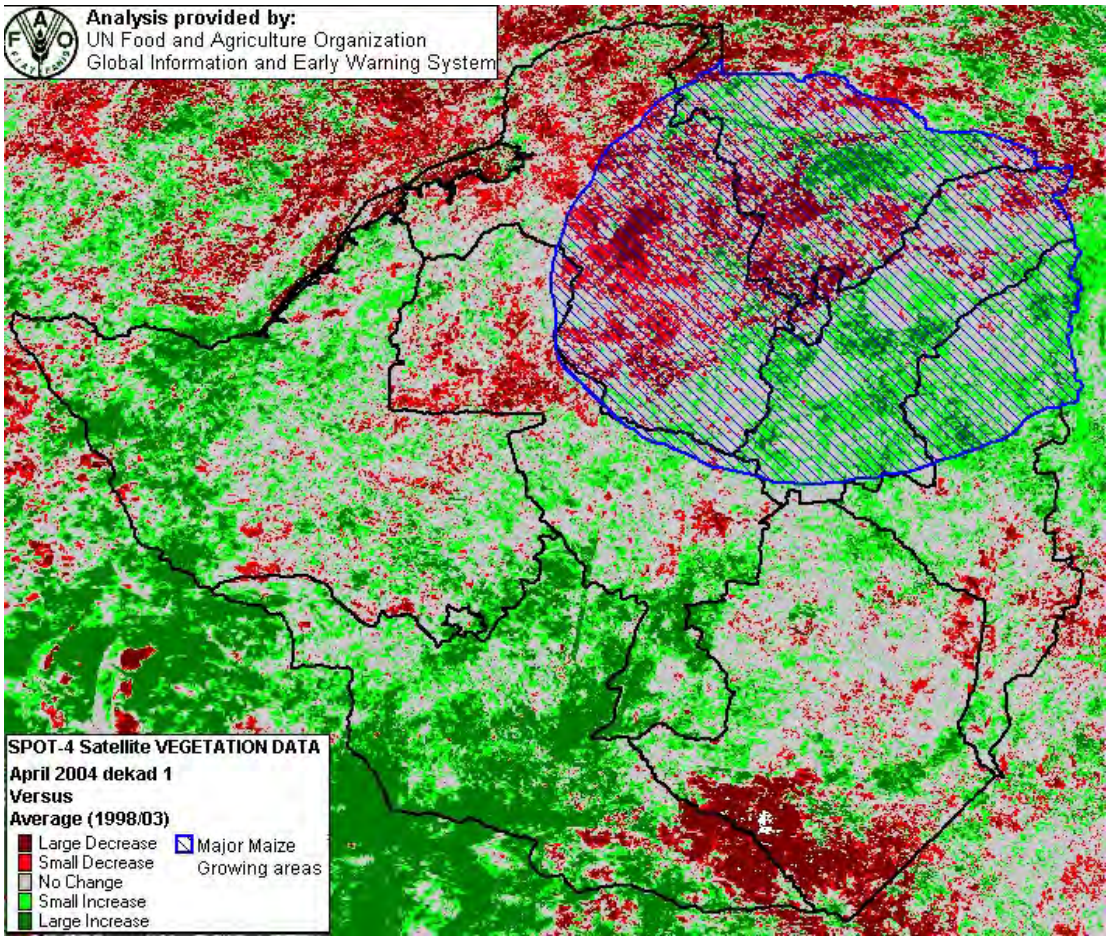
Quite by serendipity, during the preparation of this caption, we discovered that the IDRISI program uses two special theme maps of the Nakuru region as part of its training exercises. With permission, we repeat these here. On the left is an image map of elevations, in feet, showing the Rift Valley to be as low as 5,690 ft (1,734 m), at a point within the blue pattern, and the western uplands to reach to greater than 9,850 ft (3,002 m) in yellow. The SPOT scene fits somewhere in the upper right. The map on the right is another map, which plots the annual rainfall in millimeters (green is wettest, then red, yellow, and white progressively driest). (Note that in this equatorial region of East Africa there are two rainy seasons: winter and summer.) Try drawing your own relationships to the SPOT scene from these maps.





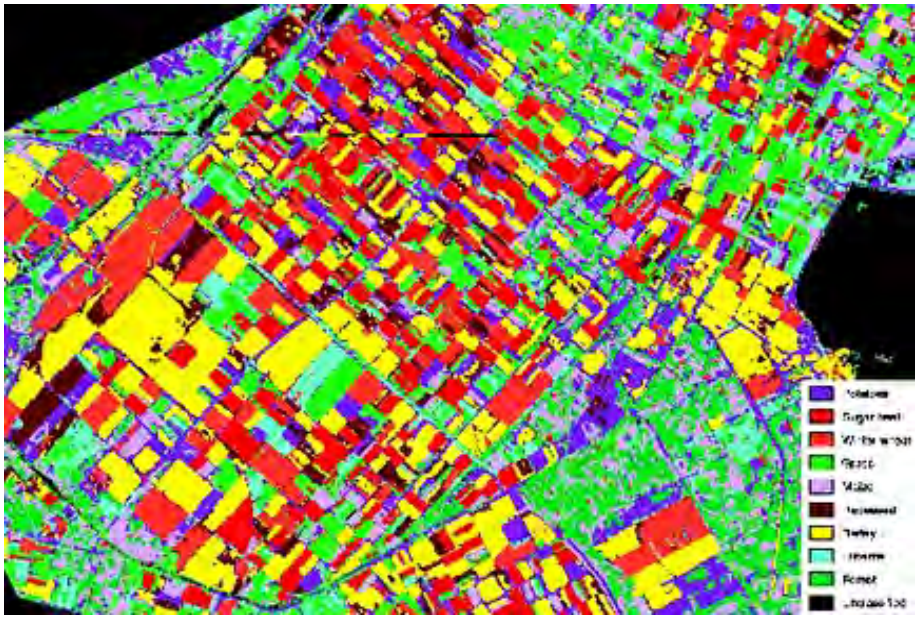
3-6: **Making a plausible guess from logic, in the Nakuru elevation map, where do you think the SPOT scene is located? ANSWER**

SPOT is effective at one mode of change detection using multitemporal imagery, applying to the same scene from two dates a point by point merging of pixels and then a classification algorithm, to determine in the case study shown below the variations in productivity of maize in Zimbabwe:



While in Africa, make note of these unusual scenes, in the desert of Sudan drained here by the Upper Nile. Water irrigated from this river has turned a once barren area into a vast array of farms (mostly cotton and millet). These are elongated in the style of the French Long Lot patterns (see also the wider scope image of this farming development on page 6-13).

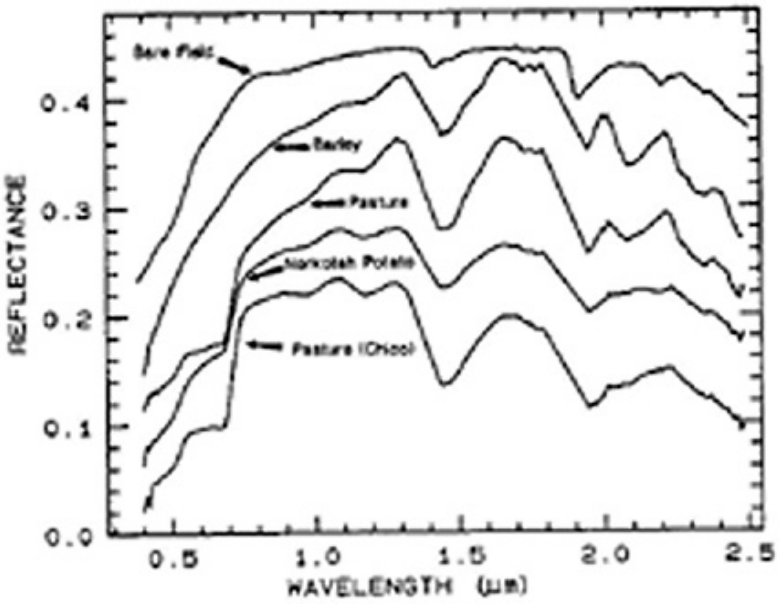
Here is a stellar example of classification of crops, this time using a Landsat TM data set for July 7, 1993 that covers numerous farms along coastal Holland.



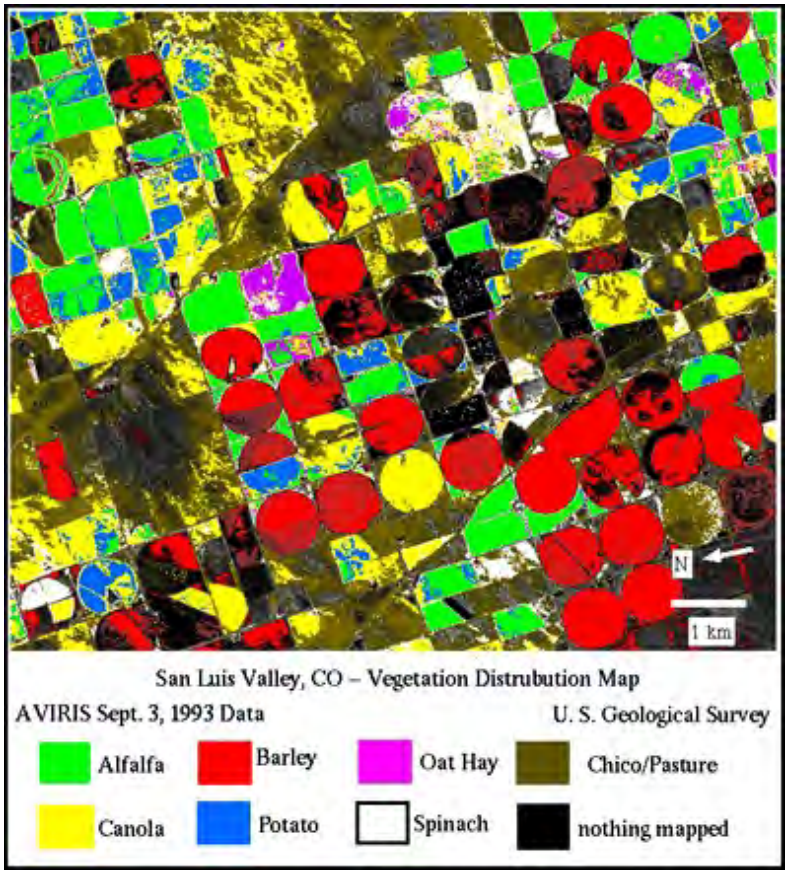
Unfortunately, the legend writing is obscured, but from top to bottom the crops are: Potato; Sugar Beets; Wheat; Grass; Maize; Rapeseed; Barley; and Lucerne (a type of alfalfa), and forest. Although colorful and rather convincing in its patterns, the accuracy of identification was only 56% - not uncommon when certain conditions needed to achieve higher accuracies are not met.

We will encounter this scene again in page 13-4c during the discussion of multitemporal monitoring of crops.

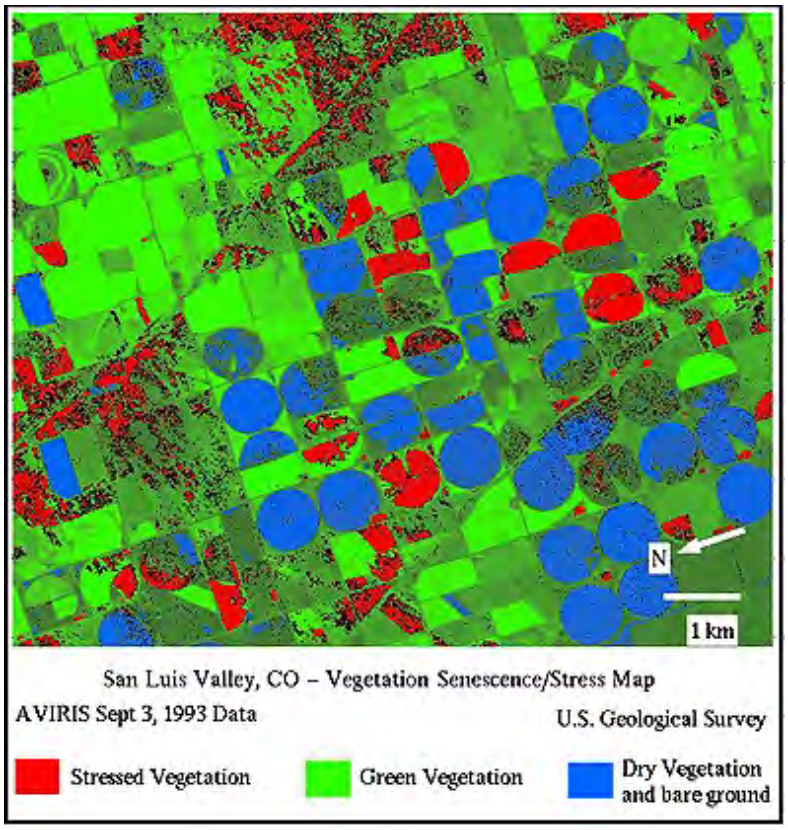
As asserted on the previous page, the spectral response of vegetative matter does not show much difference between various species or types. The spatial aspect of crops, for example, in plant shapes, leaf arrangement, stage of growth, etc. are often the best distinguishing factors. However, as was also stated, detailed spectral curves show subtle variations that can be diagnostic - hence, hyperspectral remote sensing is commonly the best tool for crop identification, as was demonstrated using the Hyperion sensor. We underwrite that postulate with this look at the use of AVIRIS to identify and differentiate crops in the San Juan Valley near Summerville, Colorado. First, examine this set of spectral curves.



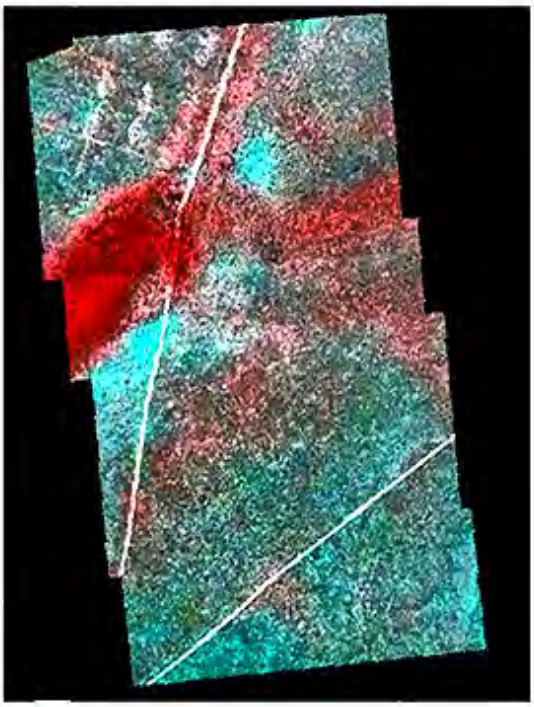
Using those curves, the fields (many circular for adaptability to pivot irrigation; see next page) were identified as to crop type (see caption) using supervised classification, giving this result (>80% accuracy):

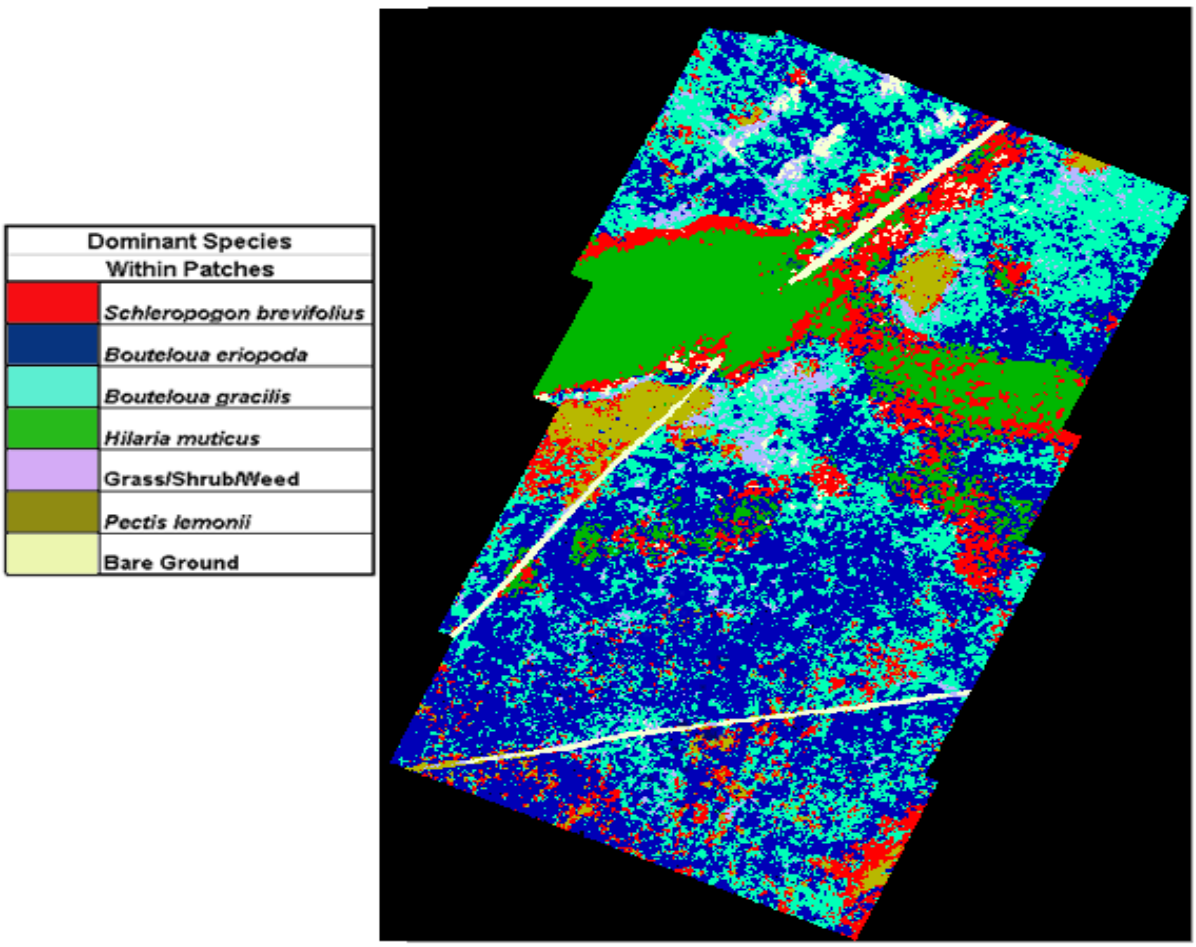


These plots are actual reflectances, rather than offset (as on page 3-1). Certain absorption bands and the height of a spectral curve for a given crop type will provide evidence of a deficiency of plant moisture that leads to a state of crop stress. Some of the individual field crops in the Summerville scene had insufficient moisture (thus, are stressed) at the time AVIRIS sensed them during an aerial flyby. This map shows this condition:



Many legum crops such as wheat and barley resemble grasses. Both natural and planted grasses are commonly the prime crop in rangelands, where they serve as feed for grazing animals. Landsat TM data can identify and map grasslands, as is demonstrated in this next example, reported by Dr. Paul Tueller of the University of Nevada-Reno, a world expert in this application. The TM subscene occur near the Organ Mountains in the piedmont covered by the open rangelands that occur within Fort Bliss, Texas (where NMS served his military career). First, the subscene, then the classification:





This classification shows six categories of vegetation, plus a modicum of bare land, a feat of identification that is impressive indeed.

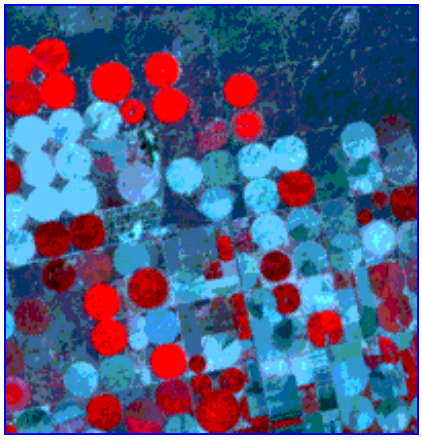
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Two more SPOT scenes, and two Landsat images, are examined mainly for their agricultural information content. The first, in Kansas, introduces an unusual, but now fairly common practice, that of pivot irrigation, recognizable by the circular field patterns. The second, in the western plains of Morocco, shows a variety of field crops in various stages of growth during a pair of observations two months apart. Differences in the degree of "redness" in the false color composites indicates the extent to which certain crops have matured. Examples of change detection using ratioing and temporal differencing are also presented, using data from the U.S. Southwest and England.

Additional Agriculture Examples in the Great Plains, New Mexico, Morocco, Guinea-Bissau; Vegetation in Wetlands

We turn momentarily to a second SPOT subscene (10 km [6.2 mi] on a side), taken on May 11, 1986, over an area near Garden City in southwestern Kansas, a major wheat producing area in the Great Plains agricultural belt. The most striking feature is a series of circular patterns ranging from less than 0.4 to about 0.8 km (1/4 - 1/2 mile) in diameter (about the same sizes as the Kenyan farm plots). Some people are astounded when they see these features from an airliner flying across the Midwest.



These fields have an effective innovation for watering known as center pivot irrigation systems.

Center Pivot Irrigation System, Northern California, Mount Shasta in the Background

The system uses a long water pipe that is mounted on motorized wheels and has one end connected to the water line at the center of the field. When operating, the irrigation system swings in a circle,sprinkling water as it rotates. (Several fields actually show a dark linear radius which is this interesting system.) The red circular fields are mostly plantings of winter wheat that are nearing harvest. Those circles in various shades of blue (note the darker patterns of moisture) are fallow at this time, but most of them have probably been seeded for spring wheat,

or will be. Note the road patterns that outline squares. These roads follow survey lines that coincide with the one-mile squares, representing sections in the American Township and Range surveying system.

In some parts of the U.S. Great Plains, circular fields become the dominant type of agricultural pattern. This is strikingly illustrated in this Landsat subscene of a part of southern Nebraska:

3-7: **What is an obvious disadvantage to the circular irrigation system?** **ANSWER**

This striking pattern of field growth is shown in this next scene - a natural color image made by Terra's ASTER instrument. The darker greens generally bespeak of standing stalks of corn. Lighter greens are various other crops. The yellows are associated with now mature spring wheat.

Circular irrigation is a worldwide practice, and can show up in unlikely places. A wadi (very intermittent stream) in the Great Desert of Saudi Arabia is the source of surface drainage and subsurface stored water from infiltration that are utilized in growing crops for distant villages:

Identification of crops using space imagery works especially well. Most fields are large (although those in parts of Asia tend to be small) - hence these will be defined by a fairly large number of uniform (homogeneous) pixels, thus favoring greater accuracy of classification. It proves very efficient to rely on Landsat-scale scenes to monitor the regional distribution of crops. This favors efficient crop yield estimates as well. Although a lower resolution sensor system, the AVHRR on some of the meteorological satellites has been effective in identifying and monitoring crops. Consider this AVHRR-based classification of crops and other classes in part of Oklahoma:

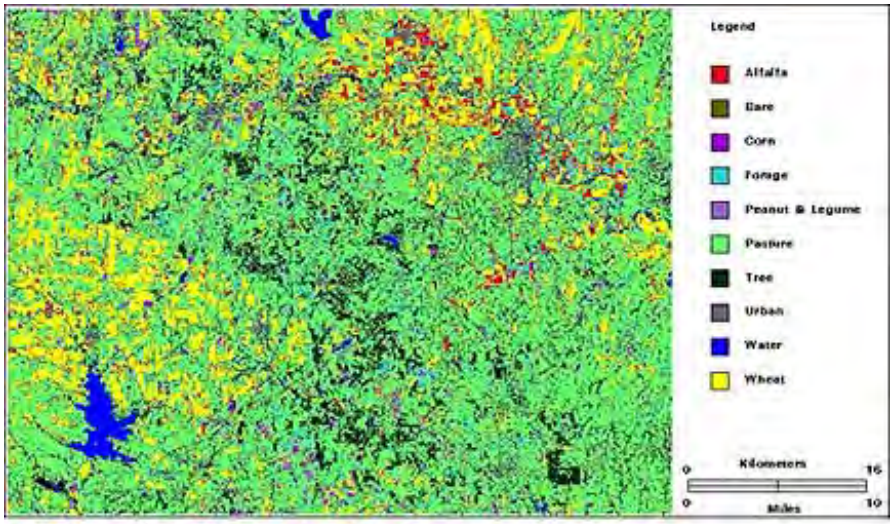
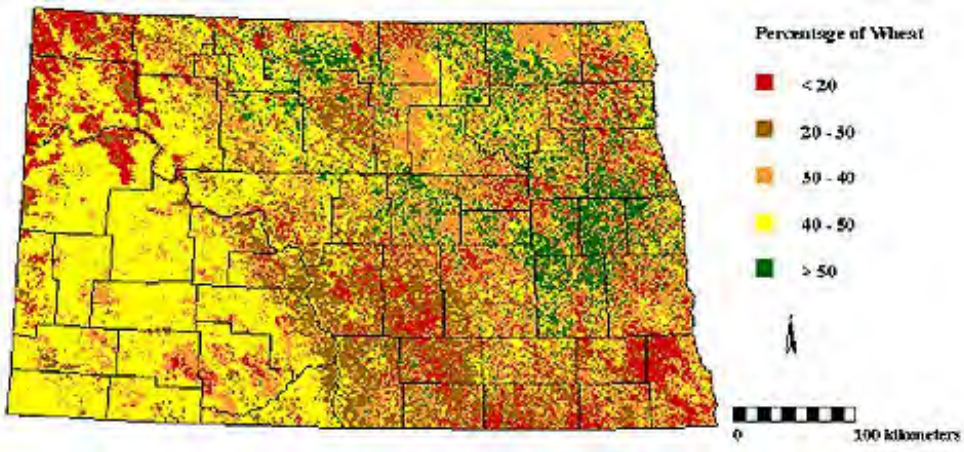
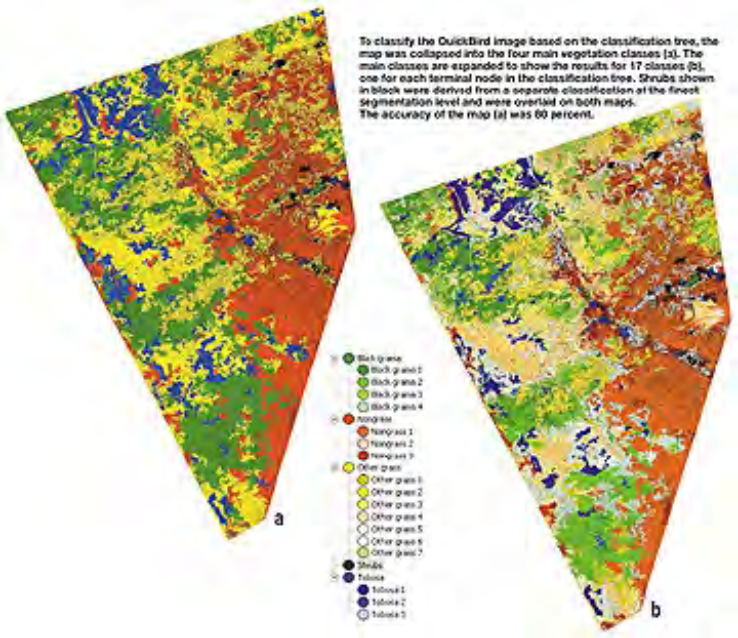


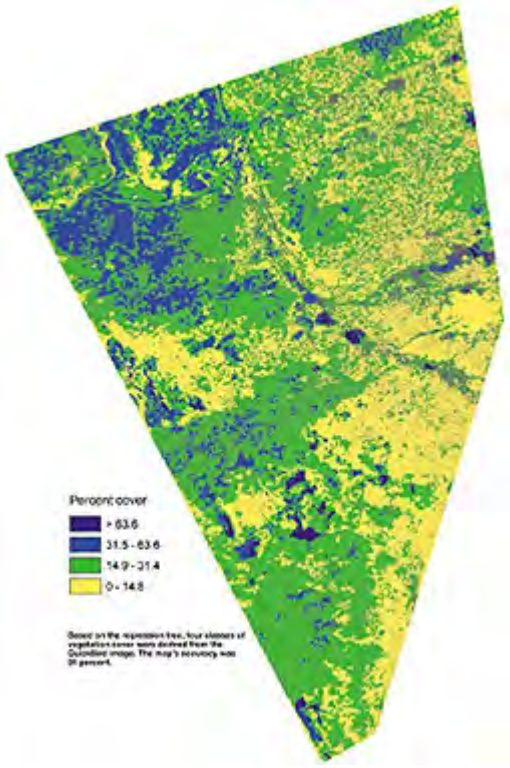
Figure 1. Land Use/Land Cover Map of the Study Area

When an AVHRR classification is coupled with independent data on productivity, a reliable estimate of percent of a crop - here wheat - in various parts of a scene - here all of North Dakota - is achievable in a very short time frame:



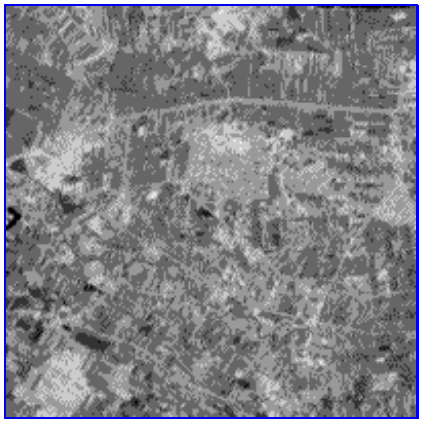
The Great Plains and the High Plains, especially their western parts, are rich in natural grasslands. These have been widely used as rangelands on which to graze cattle. The areas involved are vast; hence, knowledge of their extent and the grasses vigor are important over the short run. Early in the ERTS program, a study indicated the effectiveness of monitoring rangelands as an aid to managing their productivity. Recent studies confirm this. A group at the USDA field station near Las Cruces, working with New Mexico State University students, tested high resolution Quickbird multispectral imagery as to its capability for distinguishing grasses types and extent of cover. This illustration shows two results: the map on the left was a "first cut" at separating grasses into 4 categories; the right map divided these into subclasses. Below it is a map that establishes four levels of grass cover:

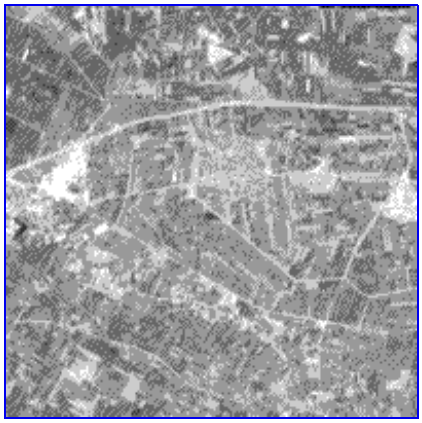




Our final agricultural scene is actually two SPOT images of the same area, obtained two months apart during late winter and spring in northwest Africa. The purpose here is to illustrate changes in vegetation as the crop calendar progresses. The area is the Ghard plains, not far from the Atlantic coast in western Morocco, near Casablanca. This granery belt is a main producer for that country. The crops include beets, sugar cane, wheat, and rice.

Monitoring growing crops is an excellent example of using multidate imagery for change detection. Let's start by looking at the SPOT Band 1 (green) image (left image) acquired on March 14, 1986. A first reaction may be that the landscape is mostly farmland, with almost all plots being smaller than in the Kenyan and Kansan scenes. The next obvious characteristics are the small river (Oued Oum er Rbia) and the major roadway to El Jadida on the west and Settat on the east, met by several feeder roads. Note also some very dark areas in the image.



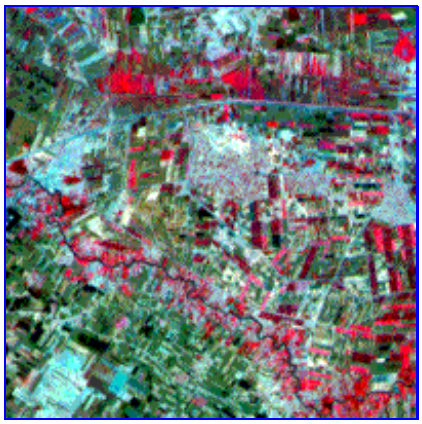


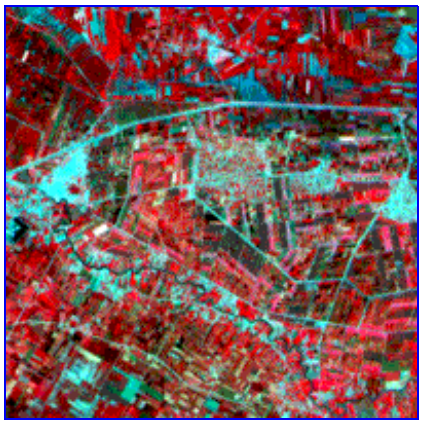
These dark areas are nearly black in Band 3 (IR) (the right image above). They are probably bodies of water, with some being irregular in shape that represent standing water from spring floods and others in field-like shapes are probably irrigated rice fields. Only a fraction of the fields (particularly along the river) are rendered in bright tones, indicating that most crops are not growing much yet.

But, when the same area is imaged in SPOT Band 3 on May 10, (the right image above) a larger percentage of fields are in medium gray tones, suggesting many crops are growing. The black

areas of flooding are reduced, but rice fields remain about the same tone. The image shows towns in medium-gray tones. Tiny bright patches dot the one in the center, indicating numerous trees.

The pronounced differences in the extent of active crop and tree vegetation, are much more evident when comparing false color images first for the March scene (left) and then the May scene (right).





In the March image, an estimated 30% of the fields have actively growing crops, while in the May image, this increases to about 80%. Some of the areas thought to be water in the March scene show up as dark red (typical of rice fields) in the May view. In the March scene, a large part of the land not in crop growth appears in a dark brownish-green color which in false color is the expression of soils that are naturally reddish-brown.

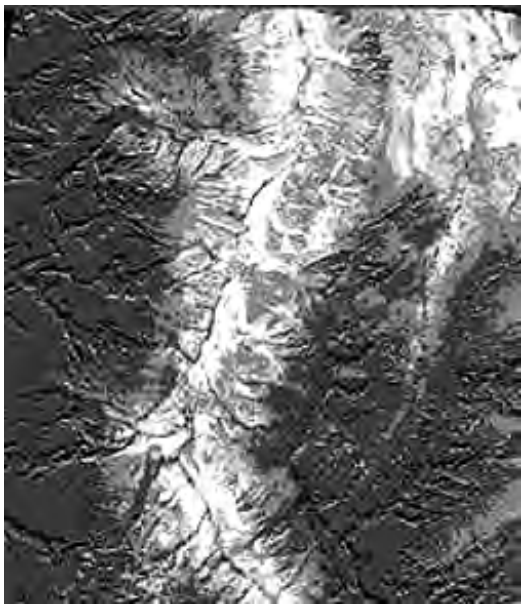
3-8: **There is at least one town in this scene, maybe more. Where? ANSWER**

The signature for rice will depend on the season in which the imagery is acquired. Rice fields during dormancy can take on a brownish color in this Landsat-7 ETM+ natural color image of an area in Guinea-Bissau, a West African country (the green areas are evergreen mangroves).

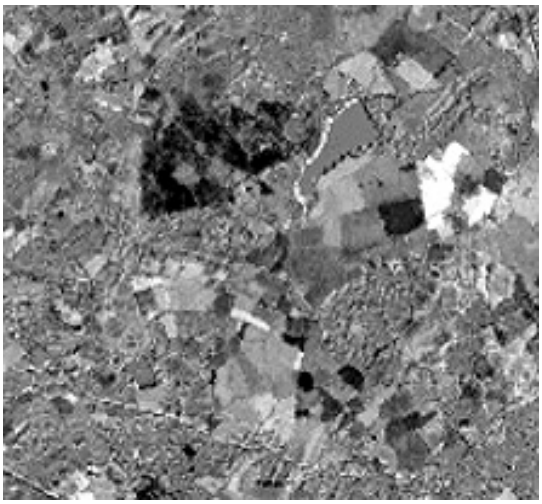
Image-processing techniques for change detection can help to emphasize major distributions of active versus dormant, or absent, vegetation. One simple approach is to make a ratio image (divide the DNs of one image by the other's DNs) using the same bands, usually in the IR, for two dates. This results in very bright tones wherever vegetation is prevalent in the dividend's date (high DN values) but not in the divisor's date (low DNs). (See page 1-15 for explanation of ratioing.) The opposite is true if the vegetation distribution is reversed because the dates are reversed. If vegetation is present to similar extents on both dates, the ratio image tones may be

moderately gray. We attempted to do a ratio image on IDRISI for the March and May scenes but failed because the two scenes did not coincide (the images were taken on dates when the orbits were not exactly aligned). Co-registering images requires a computer algorithm that can execute a "rubber-stretch fit"). IDRISI, while it can generate ratio images, does not have a simple program for co-registering (aligning) two closely related images where one is slightly offset from the other (but it can do certain kinds of registration).

Ratioing of the vegetation-sensitive IR band to the corresponding red band will cause the vegetation to express in bright tones. To illustrate how well vegetation can be made to stand out by ratioing, we operate here on a same date TM Band 4/Band 3 image of an (unidentified) area in which a valley is heavily vegetated (appearing whitish in the ratio image) and the adjacent plateau is mostly rock and soil (dark).



More than any other class, it is vegetation that has strong seasonal variations (such as deciduous forests; croplands) which shows the largest differences in brightness in the infrared. Another change detection approach applicable to locating vegetation is to computer-register two scenes (IR is optimal; red band can be responsive) acquired on different dates (such as height of growing season versus winter) and then subtract the DNs of one from the other. Large changes give rise to big differences (can be assigned white in an image that spreads the differences out in the gray scale); small changes would lead to small differences rendered black; moderate changes in gray. Below is an *image differencing* rendition of mostly fields near Leicestershire, England, made by subtracting a February IR scene from a May IR scene. The prevalence of medium grays in this difference image means that the fields have "greened up" fairly uniformly. If the dates were summer - winter, the differences would be larger, and the tones whiter; the very white patch is probably a crop that matures early whereas the black patch is a barren field.

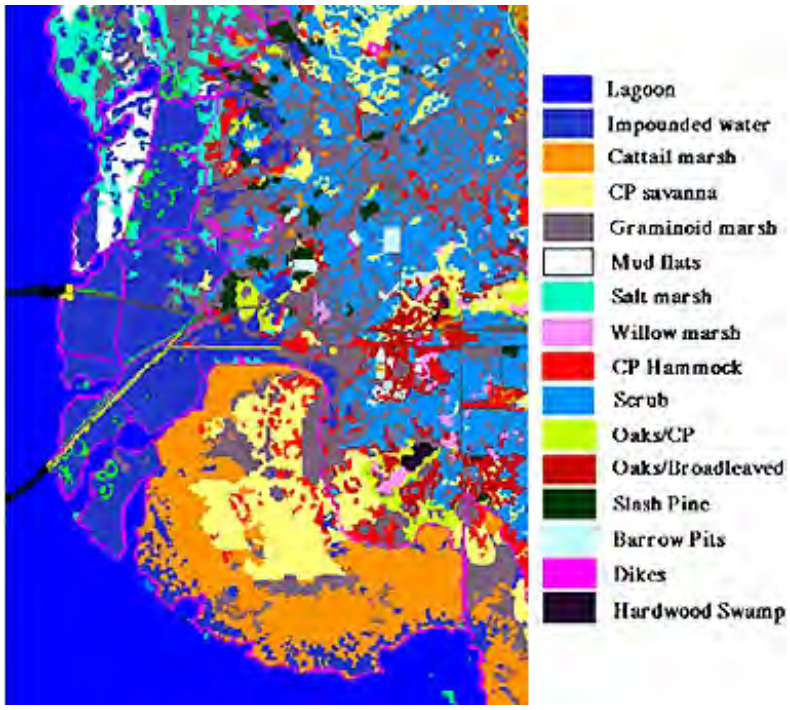


Another vegetation-rich environment which benefits considerably from satellite and aircraft monitoring by remote sensors is the general category of *Wetlands*. These are water-rich impoundments such as lakes, etc. or natural bodies of water, many at and in from the edges of ocean shores, lakes, or winters. Depending on whether the water is saline (salty seawater) or fresh, there will be a biotic assemblage that is characteristic of the particular conditions of growth. The monitoring is interested mainly in noting the shrinkage or expansion of the wetlands water and in the health and abundance of the supported vegetation. Often wetlands will follow a natural course that tends towards drying up as sediment fills the depressions holding the host water or the supply of water diminishes, or vegetation becomes so abundant as to supplant the water. Another, ever more common, cause is the deliberate draining of the wetlands by humans bent on reclaiming the area for farming, housing, recreations, etc. We shall show the ability of remote sensing to determine the characteristics of wetlands from satellites.

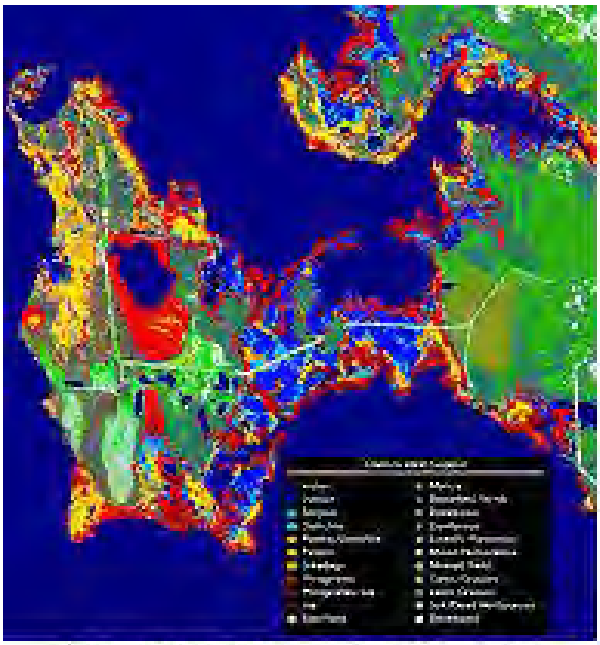
First up is this subscene from a Landsat TM image showing part of Cape Canaveral on the east-central coastline of Florida. Note the land area (with brownish tones) in the upper left quadrant. This is a mix of brackish (less salty than ocean water) water and land plants that are typical of this area.

The two dominant land plants are loblolly (slash) pines and palmetto plants, as seen is this ground photo.

This is the supervised classification of the small land bulge seen in the Landsat subscene

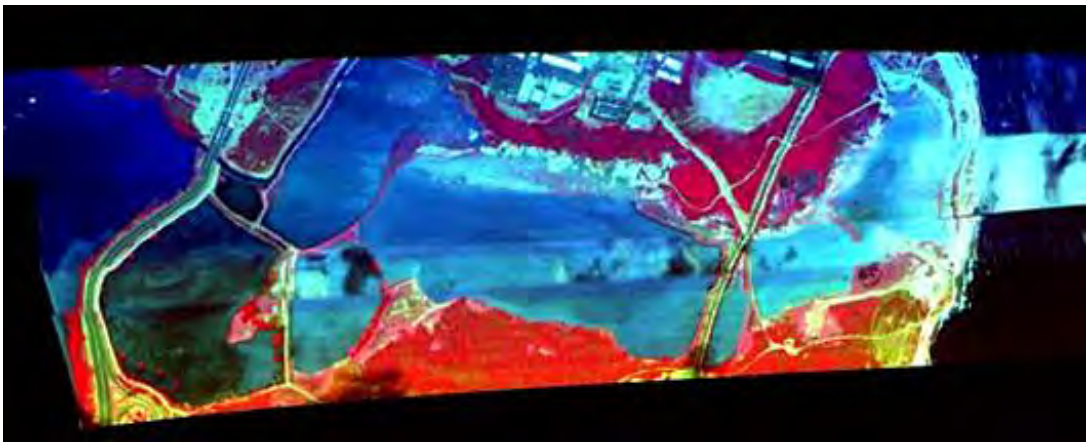


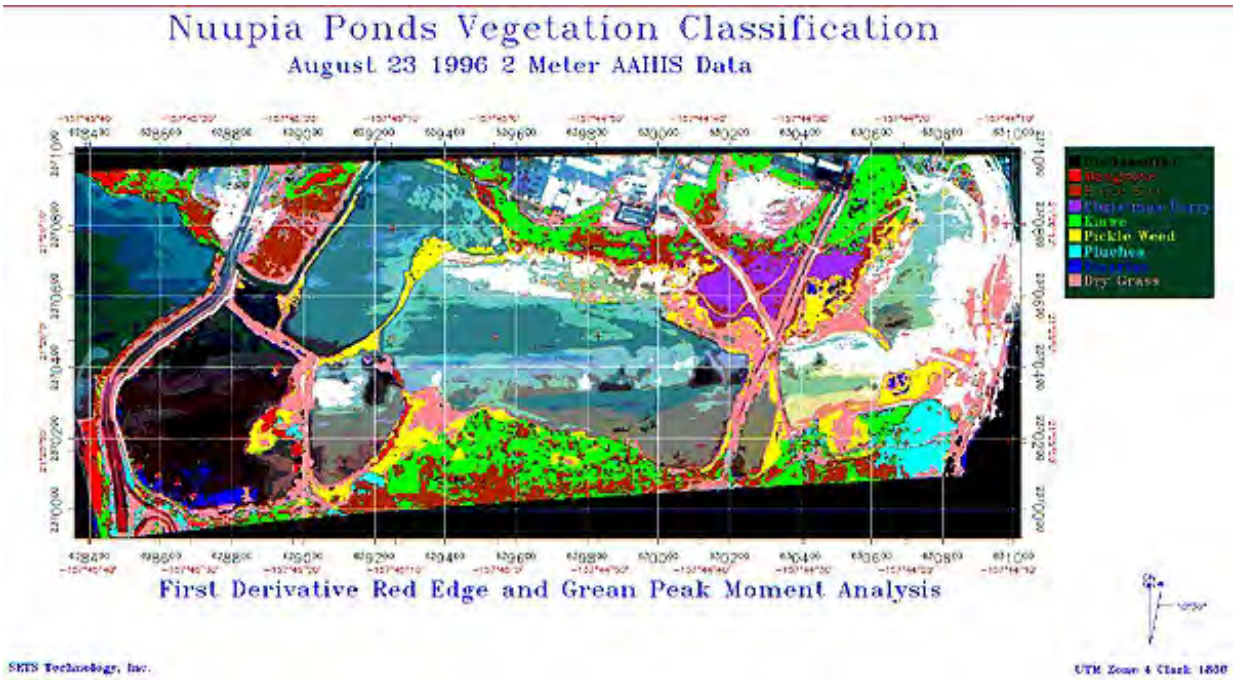
This next illustration has been degraded in the reproduction process, with the legend unreadable, but it indicates the amount of information leading to feature/class identification possible from an aircraft-mounted hyperspectral sensor. It is a classification of a neck of land tied to an "island" jutting into the Chesapeake Bay near the Bay Bridge on the eastern shore of Maryland.



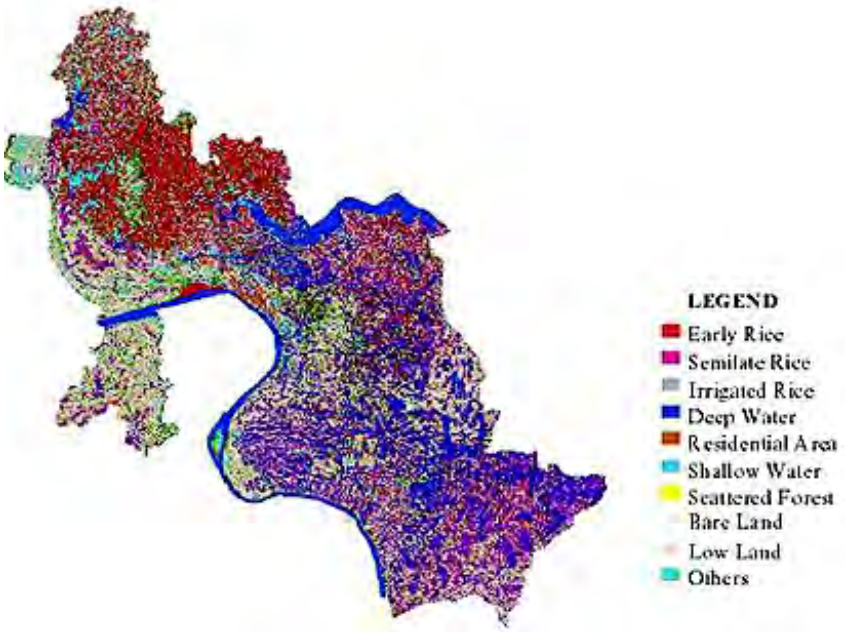
Vegetation Community Mapping
Horsehead Wetlands Preserve
Stevensville, MD - June 1999

A hyperspectral sensor system, AAHIS, built and operated by International Science and Technology, Inc., has 228 very narrow bands in the spectral range from 0.437 to 0.840 µm. Flown from an airplane, it also achieves high spatial resolution. Here is a false color image of the Nuupia Ponds on Kaneola Bay, Oahu, Hawaii. Beneath it is a specialized classification that distinguishes most classes present in the scene

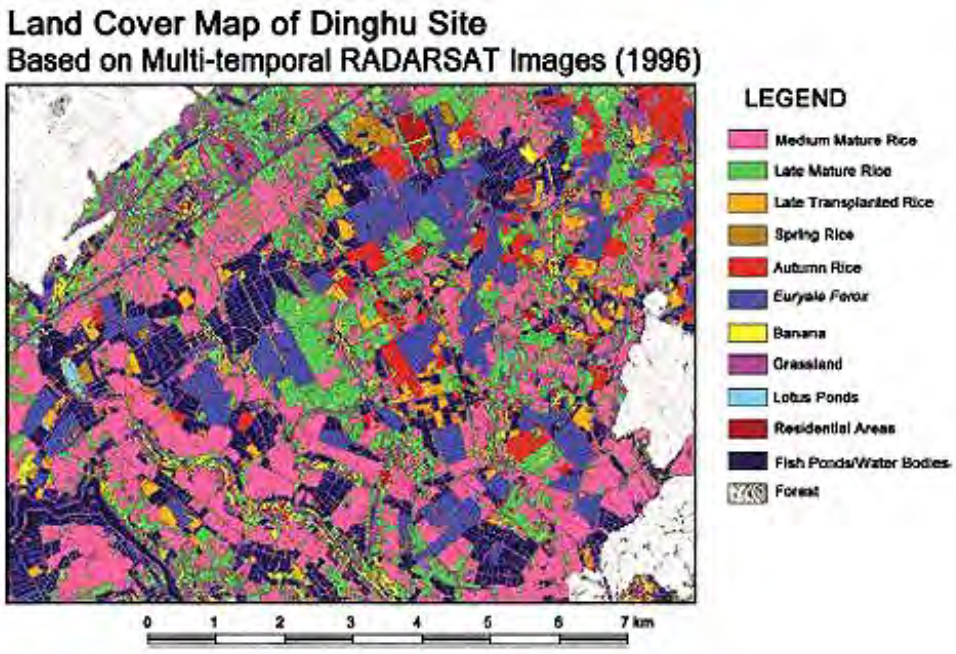




Wetland conditions are sometimes characteristic of and suitable for growing crops that require considerable available water. Rice is such a crop and is a major staple for huge populations on Earth - mainly in Asia. Below are two classifications that include rice crops. The first is a Landsat supervised classification in China:



Multitemporal and multiband radar data can also yield quality classifications, as for this site near Dinghu, China:



PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

This page considers more examples of practical uses of remote sensing in studying several aspects of vegetation. The first introduces the concept of "Vegetation Index" in which bands sensitive to chlorophyll absorption and cell wall reflectance are treated by simple mathematics (usually ratios of individual bands or of band sums or differences) to accentuate recognition of and variation within types and densities of growing forests, fields, and crops. A now famous example of vegetation changes over growing seasons in the entire African continent, as imaged by the AVHRR sensor on a NOAA satellite, is presented. Other AVHRR examples display NDVI results for Canada, Mexico, and the whole Earth's landmasses. Another example looks at wheat fields along the Volga River in southern Russia, as seen in Landsat images taken at dates about 3 weeks apart in 1974 and 1975, in which crops in the latter year are greatly reduced in vigor and extent owing to a major drought. A similar situation affected East Africa in 1984 and was monitored by the AVHRR. The final scene includes a Landsat view of the California-Mexican border just above the Gulf of California, in which a striking contrast in number of fields planted and degree of growth results from different crop practices and water uses. Two ASTER scenes covering the same region are also depicted. All these observations, especially at AVHRR scales, are needed as input to integrated data systems that monitor crops at regional and global levels.

The Vegetation Index; Other Vegetation Scenes

The Landsat, SPOT and other systems participated in crop control and inventories of various kinds. As stated earlier, Landsat Thematic Mapper (TM) 4 and Multispectral Scanner (MSS) 6 and 7 bands (and SPOT Band 3) are the most sensitive for detecting IR reflectances from plant cells (modified by water content). TM Band 3 and MSS Band 5 (and SPOT Band 2), which measure reflectances in the visible red, provide data on the influence of light-absorbing chlorophyll. Ratio images using these bands help to quantify the amount of vegetation, as biomass, involved in signature responses. For the NOAA series, Bands 2 and 1 of their Advanced Very High Resolution Radiometer (AVHRR) sensor are roughly equivalent to TM Bands 4 and 3 (see Section 14 for a review of metsat systems). The ratio of TM Band 4 to Band 3, MSS Bands 6 or 7 to Band 5, or AVHRR Bands 2 to 1 is a simple approximation of the Vegetation Index (VI). Other VI variants depend on other combinations of these variables. Three of the most commonly used are shown here; read their captions for more information:

$$VI = \frac{NIR}{R}$$

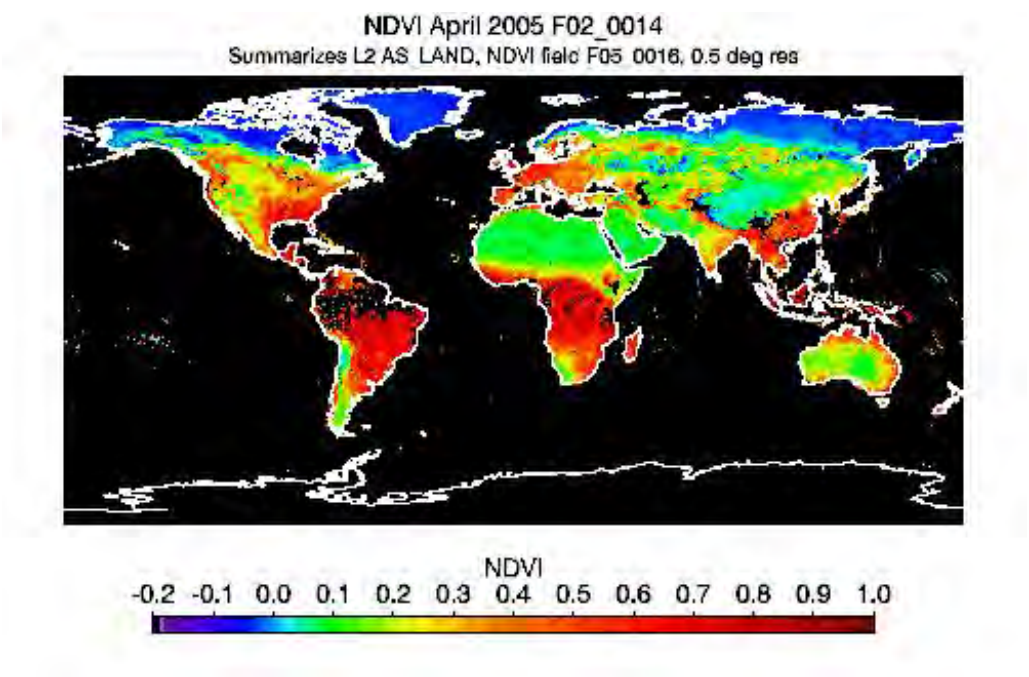
$$NDVI = \frac{(NIR-R)}{(NIR+R)}$$

$$SAVI = \frac{(NIR-R)\ \ (1+L)}{(NIR+R+L)}$$

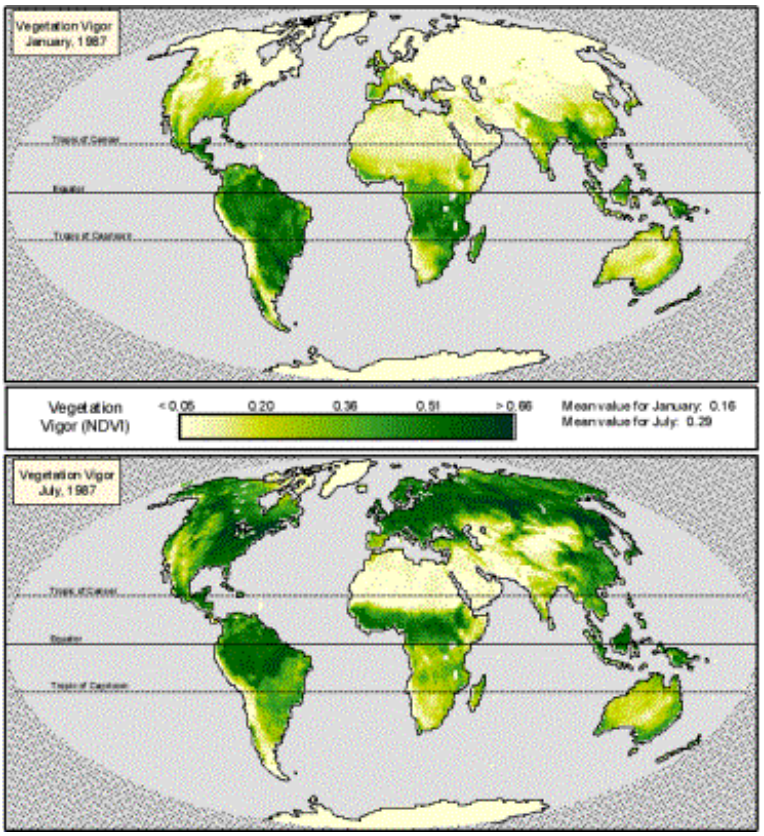
Most commonly used is the Normalized Difference Vegetation Index (NDVI), which is defined as (Near IR band - Red band) divided by (Near IR band + Red band). For TM this is (4 - 3)/(4 + 3); for AVHRR this is (2 - 1)/2 + 1).

NDVIs can be produced for the entire world. These can be for short periods (a month or so), entire years, or years apart. The top image below is a general global NDVI map for a summer. The bottom map shows a similar period but a different (time of) year and uses more contrasting

colors to highlight the variations.



When two NDVI maps of the world, one from northern winter and the second from northern summer, are compared, notable changes in the temperate and polar zones would be expected because of the seasonal contrasts; changes in the equatorial zone should be much less. This pair of images exemplifies this idea:



Changes in vegetation over shorter periods, say 2-4 weeks apart, can be displayed as NDVI maps. This series of maps applies to the 48 mid-continental United States, and progresses from mid-summer to late-fall:

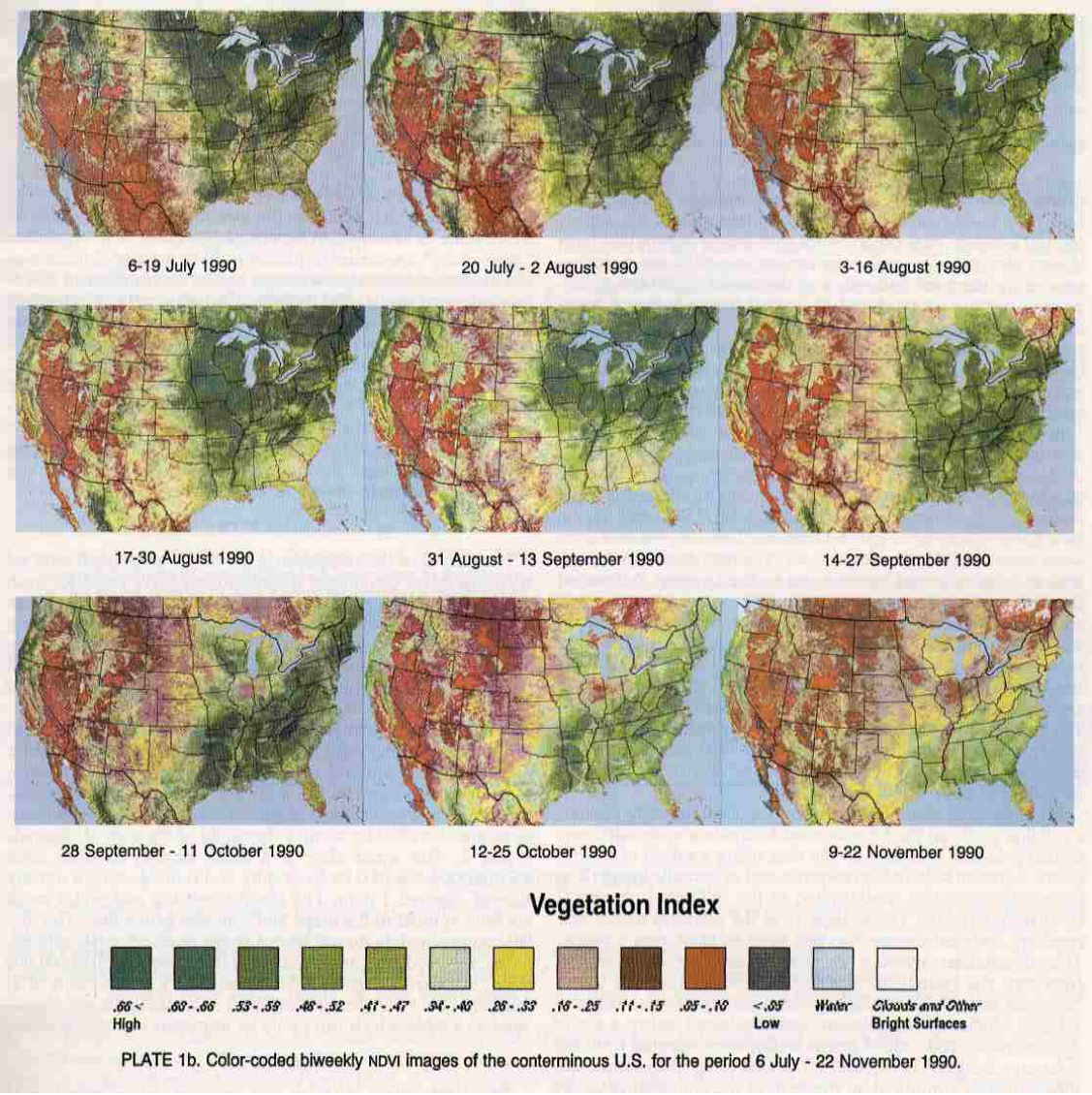
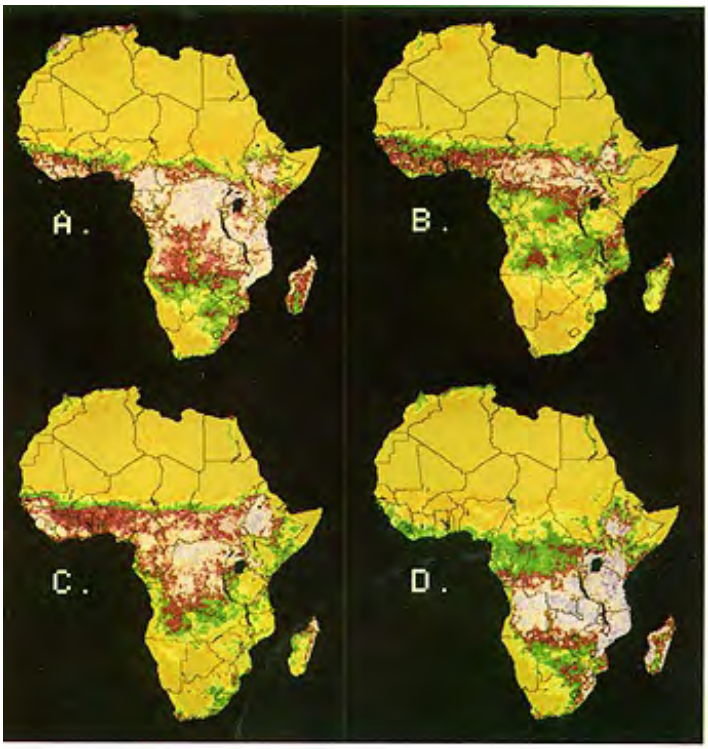


PLATE 1b. Color-coded biweekly NDVI images of the conterminous U.S. for the period 6 July - 22 November 1990.

NDVI distributions for entire continents can be monitored in one view from geostationary satellites such as the meteorological satellite (this is the origin of the common descriptive term, "metsat"). Using AVHRR data (and supporting ground truth) grouped into 21-day periods for eight observing intervals (by NOAA-7) April 1982, through mid-February 1983, J.C. Tucker and his associates at NASA's Goddard Space Flight Center have produced, using Principal Components Analysis, a general classification of land-cover types for all of Africa. In the class map below medium blue = thickets and bushlands; dark blue = interspersed tropical forests and grasslands; purple = Sudan-type grasslands and woodlands; medium green = semi-arid wooded grasslands and bushlands; dark green = woodlands; yellow = deciduous bushlands and wooded grasslands; orange = semi-deserts to deserts; and red = tropical rainforests and montane forests.

This group and others have continued to apply metsat AVHRR to observe seasonal changes in biomass ("green wave") over all of Africa, as illustrated below for the following dates: A. April 12-May 2, 1982; B. July 5-25, 1982; C. Sept. 27-Oct. 17, 1982; D. Dec. 20, 1982-Jan. 9, 1983. This has proved invaluable in determining crop shortfalls and drought conditions in Ethiopia and in areas of the Sahel (northern desert regions) during periods in the last decade where starvation was a mass threat. This observational technique has now been applied worldwide.



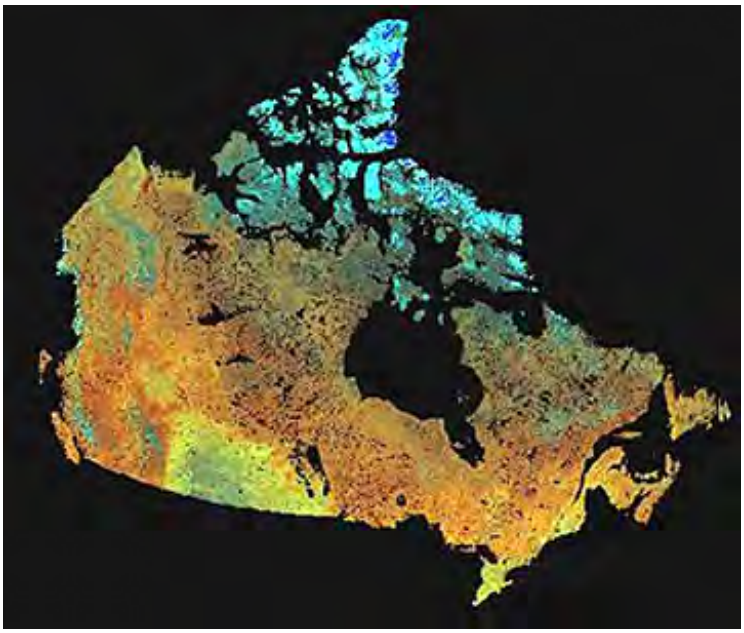
3-9: **Knowing that the semi-desert south of the Sahara in northern Africa is called "The Sahel", in which of the four above panels has the "drought" moved farthest south; which panel marks the most significant northern encroachment of vegetation, as disclosed by a higher VI? ANSWER**

Analysis of a mosaic made from SPOT images defines the broad categories of natural land cover. In the image below, greens indicate heavy vegetation (biomass), reds show mostly pampas grass cover, rose refers to very arid regions, and blue-white demarcates areas of barren, high, often ice-covered land:

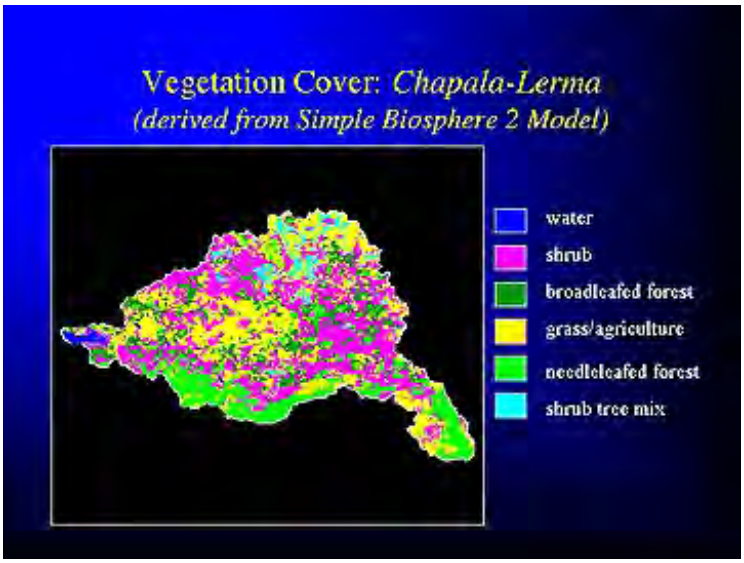
Satellite images covering large areas can give quick and direct impressions of the overall distribution of vegetation, without the need for special processing. Look at this Orbimage rendition in near natural color of Australia. It shows at a glance that most of the vegetated (forested) areas, in green, are in SE Australia, and along the coasts elsewhere. Most of the continent's interior is sparsely vegetated, as shown by the prevailing browns:

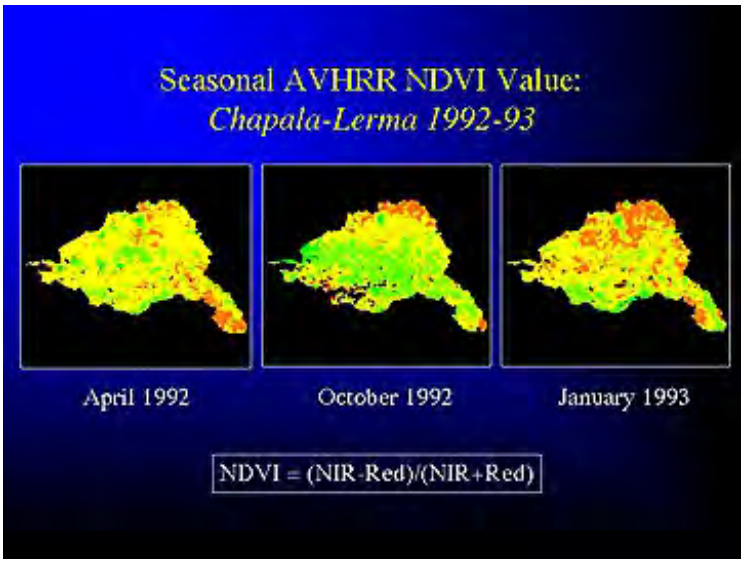
AVHRR data allow monitoring of vegetation over scales ranging from continental, as above, down to regional or up to global. Here is a color mosaic made from Band 1 = blue; Band 2 =

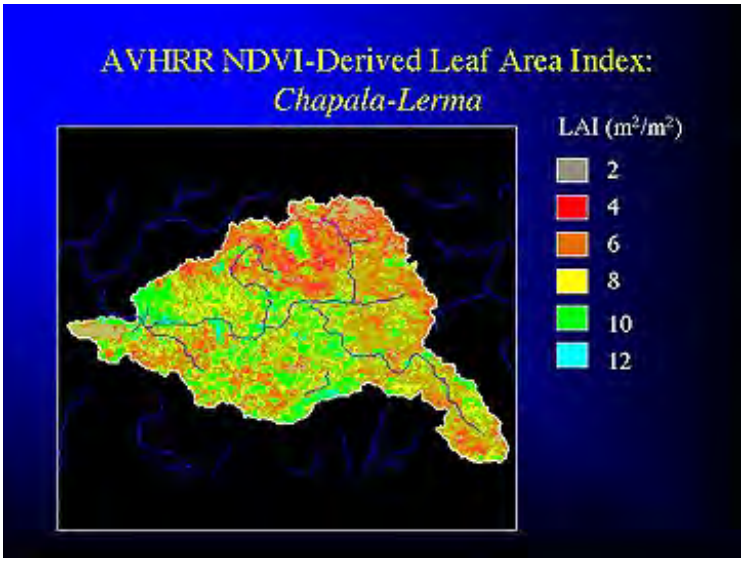
green; and NDVI = red for all of Canada. The browns that result indicate the wide extent to which that country is forested; in this case, greens denote areas of grasslands, such as the Canadian Great Plains.



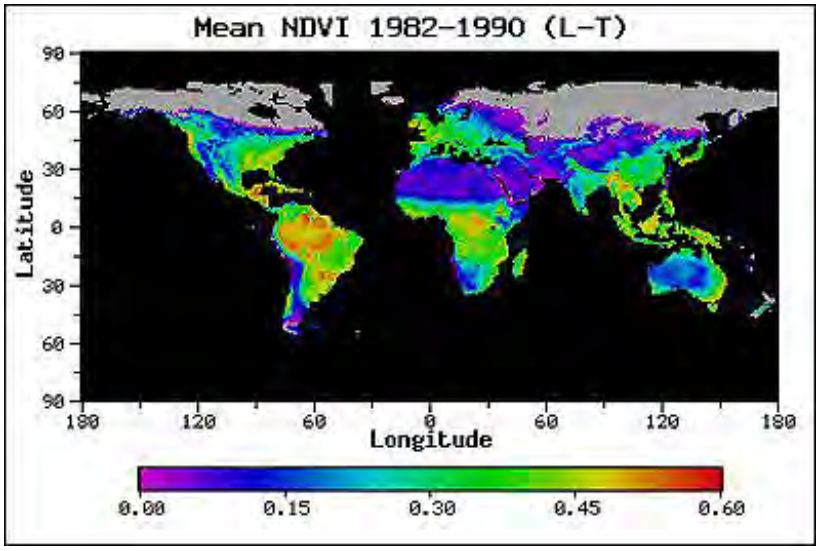
Much smaller areas can be studied with AVHRR, Landsat, SPOT, and other systems. The next three images, whose information content is given in their titles, cover a large drainage basin in the central Mexican Highlands. The Lerma River flows through the scene into Chapala Lake. This region is south of Guadulajara. This study was made by Joseph White of Baylor University.





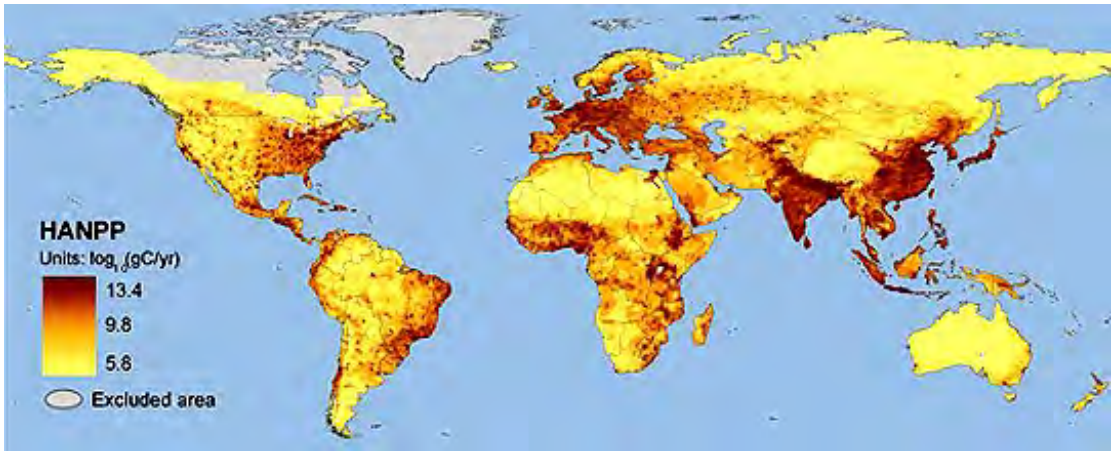


At the other extreme, cumulative AVHRR data allow a plot of mean NDVI distribution on a global basis. This next illustration shows values averaged between 1982 and 1990, using the Los-Tucker model. Note that the highest values are in the Amazon region of South America and the largest low values are in the deserts of North Africa eastward through the Middle East into central Asia.

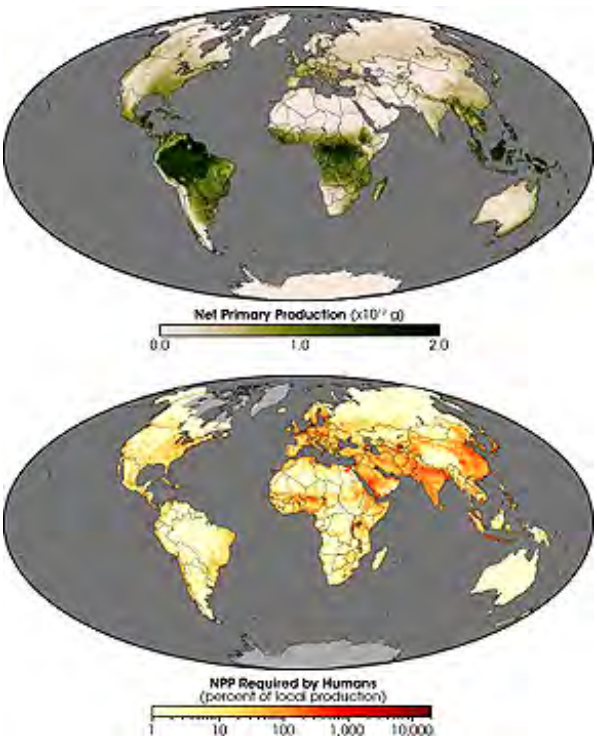


Such data can be reworked to indicate a global parameter called HANPP (Human-Appropriated Net Primary Production). The map below, released in 2004, and developed by NASA scientists in cooperation with the Dept. of Agriculture, shows the extent to which plant food, wood, and fiber

are harvested in different parts of the world. As expected, maximum amounts are gathered in the eastern U.S., southern Brazil, most of Europe, parts of Africa. and eastern China, Japan, India and Indonesia. Note the sparcity in Australia - this in part accounts for the low population in that island continent, confined largely to the coast.

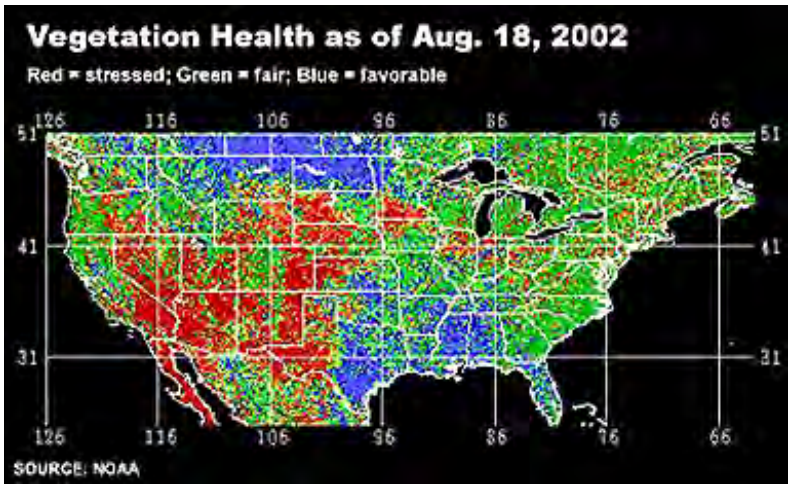


This type of assessment has been refined in 2007 by Dr. Marc Imhoff and colleagues at NASA's Goddard Space Flight Center. Their analysis is summarized in this pair of data plots:

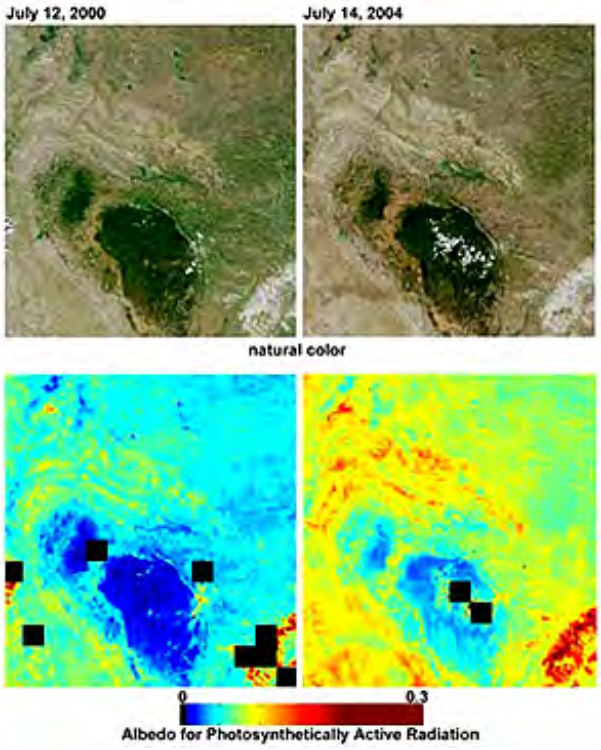


The top map shows the distribution of all vegetation, including agricultural crops, worldwide. The bottom map is very similar to the one above, showing in percentage, the amount of vegetation - mainly crops - need by the indigenous population. Note that many regions, such as India, require much more foodstuffs than can be grown locally, thus requiring imports.

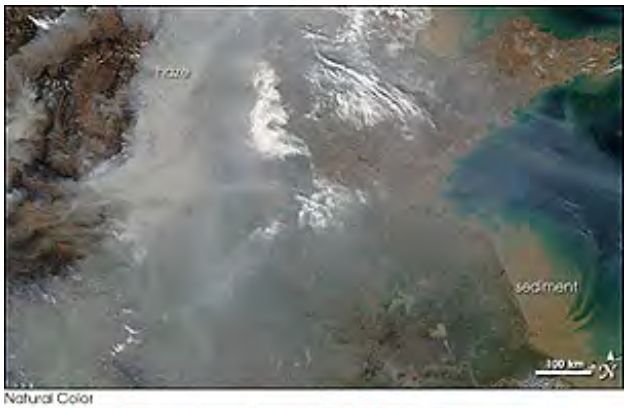
Stressed vegetation can be revealed by NDVI calculations or more directly by other indices. In the summer of 2002, the United States was beset by one of the worse and widespread droughts in decades. Sparsity of rainfall in early August is discussed at the top of page 14-15. This image is a plot of AVHRR data processed to indicate vegetation stress:



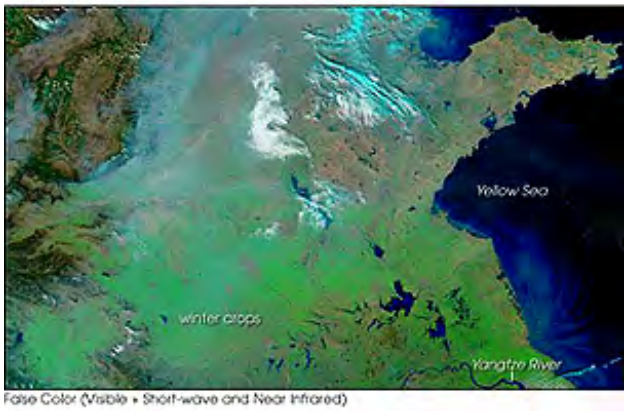
Vegetative stress will commonly turn green vegetation into brownish tints as trees and plants wither and perhaps die. This will cause reflectance (albedo) values for the vegetation to increase. That change is evident in this pair of Terra images of the Black Hills (heavily forested) and surrounding grass-covered high plains for the years 2000 and 2004. The latter year was one of worsening drought conditions over much of the western U.S. The higher reflectances extracted from image analysis for 2004 confirm this approach to detecting adverse conditions as rainfall diminishes.



Stress of another kind can have a significant impact on vegetation health and productivity. China, which needs to feed 1.3 billion people, is especially sensitive to droughts and other atmospheric factors that affect (stunt) growth. With its rapidly growing number of automobiles and industry, China has been experiencing thick blankets of smog and haze over large areas. This reduces beneficial solar radiation that promotes growth. The pair of images below, made with Terra's MODIS instrument, shows the widespread dirty haze spreading over much of eastern China as seen in natural color. The image beneath taken at the same time uses UV and Near IR bands to penetrate the haze, thus showing the extent of green (wheat mostly) in this late winter image. If this haze repeats often enough lowered productivity will result.

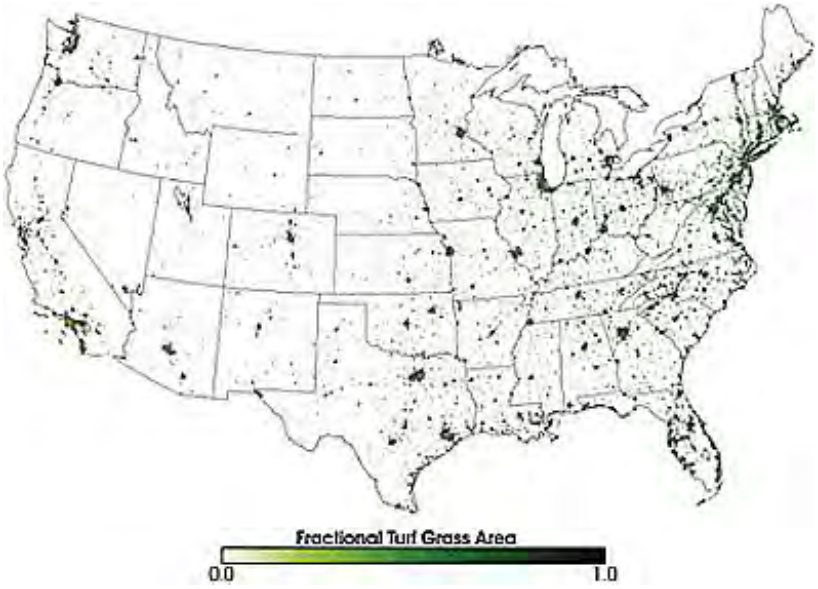


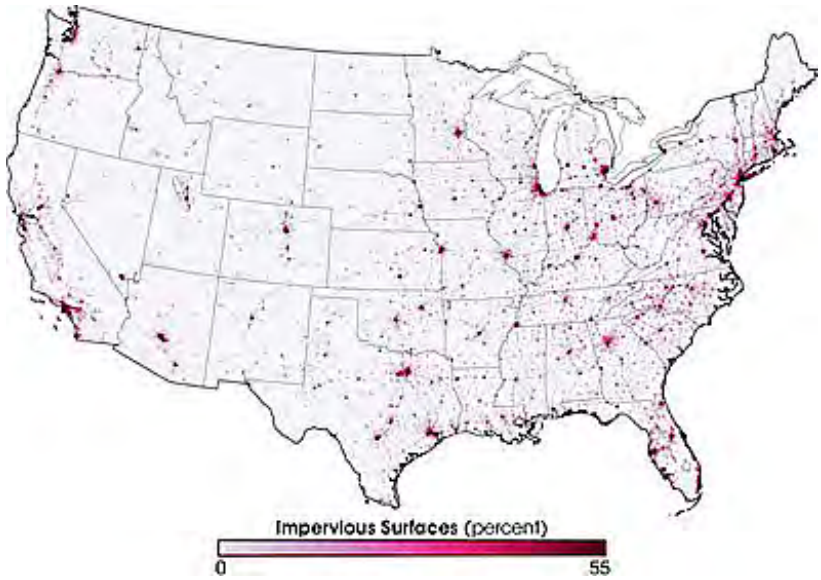
Natural Color



False Color (Visible + Short-wave and Near Infrared)

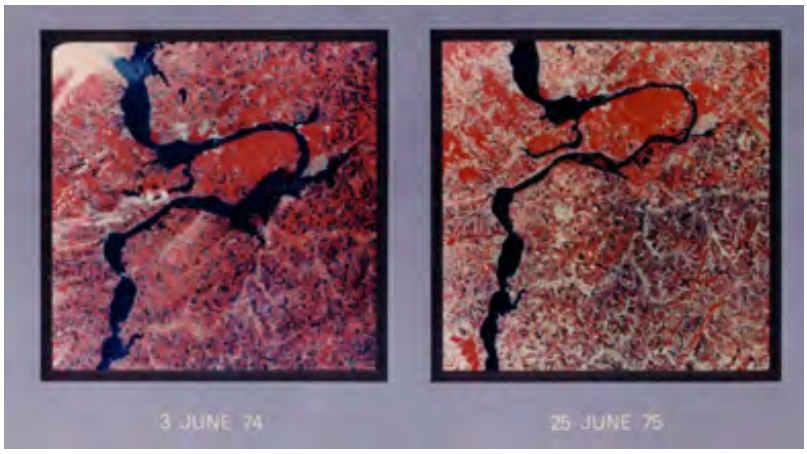
One thing that perturbs home owners in America is the effect of adverse weather on their lawns. Cristina Milesi, a fellow at NASA's Ames Research Center, chose as her Ph.D. thesis to study the distribution of lawns in the United States using both space and aerial imagery. She has published the map below which indicates in greens where the larger proportions of lawns are found in the U.S. She also produced a map showing areas that are impervious to rain - these are principally in cities, towns, and metropolitan districts. The patterns are closely coincident, demonstrating that most lawns in the U.S. - as one would expect - are found in urban and to a greater extent suburban regions of the country. Locate the green patch nearest you.





The Volga Wheat Drought; Africa's Drought; the Salton Sea, California/Mexico

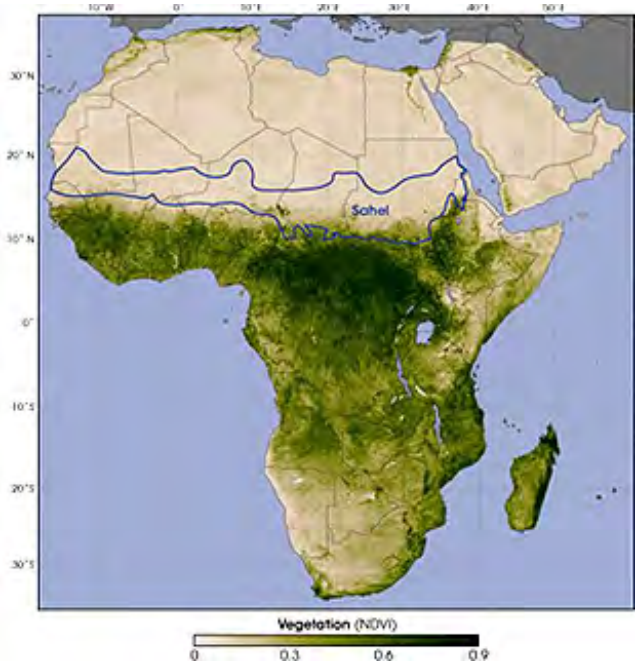
In the mid-70s, another example of crop failure on a grand scale, but not in Africa, made media headlines. In the Volga and other key wheat growing areas of the former Soviet Union, a severe drought in parts of Russia led to a threatening production shortfall that forced the leaders to seek help from outside wheat markets. They approached the U.S. and Australia governments, in particular, to furnish enough wheat and other grains to forestall possible starvation in several regions. Some critics claimed that the leaders were faking the shortage to take advantage of good prices elsewhere. But, the camera doesn't lie. Landsat images proved the veracity of the Russian plea for help, as is clearly depicted in this before and after image:



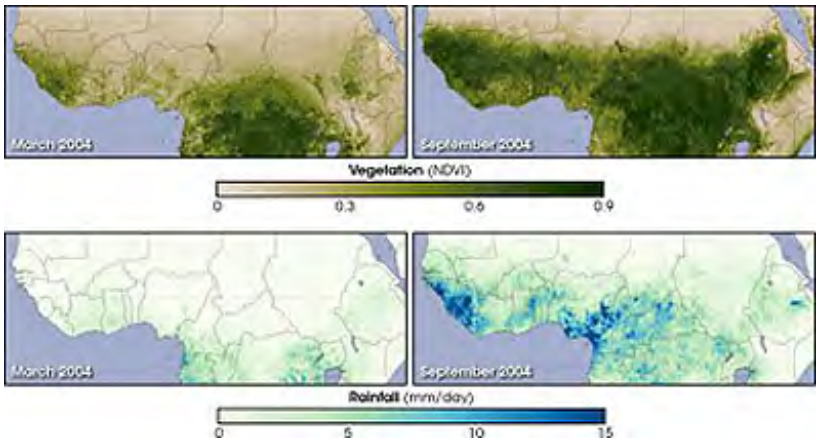
In the 1974 subscene that embraces a large bend in the Volga River, the fields are already in normal crop stages. A year later, and three weeks beyond the 1974 time, when mature crops should have increased the scene redness, instead much of the farmland is fallow (darker grays and tans), confirming the drought claims.

3-10: **Does the drought appear localized or regional; what is the nature of the red colored area within the great curved bend of the Volga? ANSWER**

Drought is one of the conditions affecting vegetation that can be sensitively measured by NDVI calculations. Repetitive space imagery, such as AVHRR, Landsat, and MODIS, has been the main data source for measuring NDVI on a regional and even continental scale. Such data can be coupled with TRMM and other meteorological satellites to correlate rainfall with vegetation distribution and health. This can also be a regional "first alert" when growth is abnormal and drought ensues. Here is an NDVI map of vegetation in all of Africa; the Sahel region, which is the zone of grasslands and farming between the Sahara and the Congo jungle, is indicated:



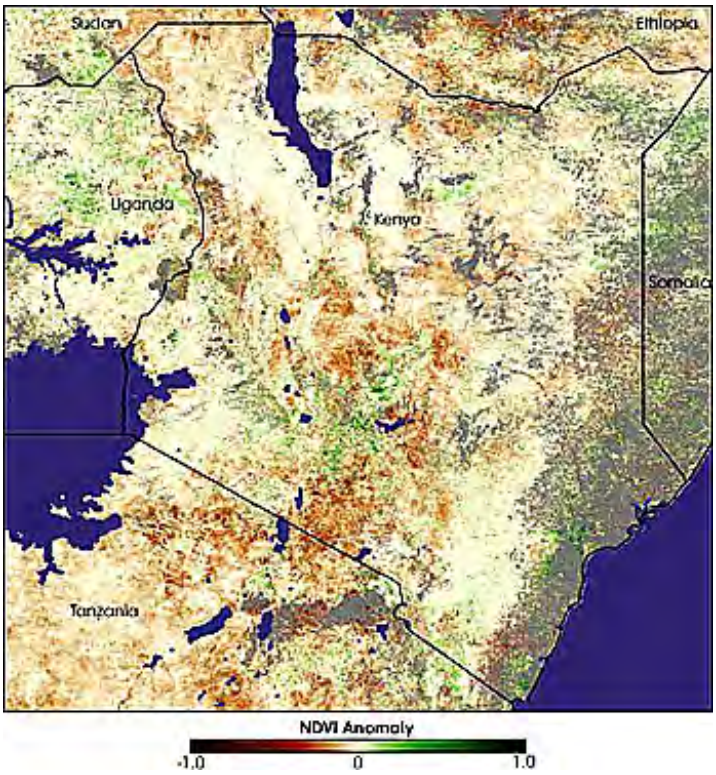
If all goes well each year, the wet season is marked by a greening of the Sahel. Here are dry and wet NDVI maps of the Sahel showing a normal change:



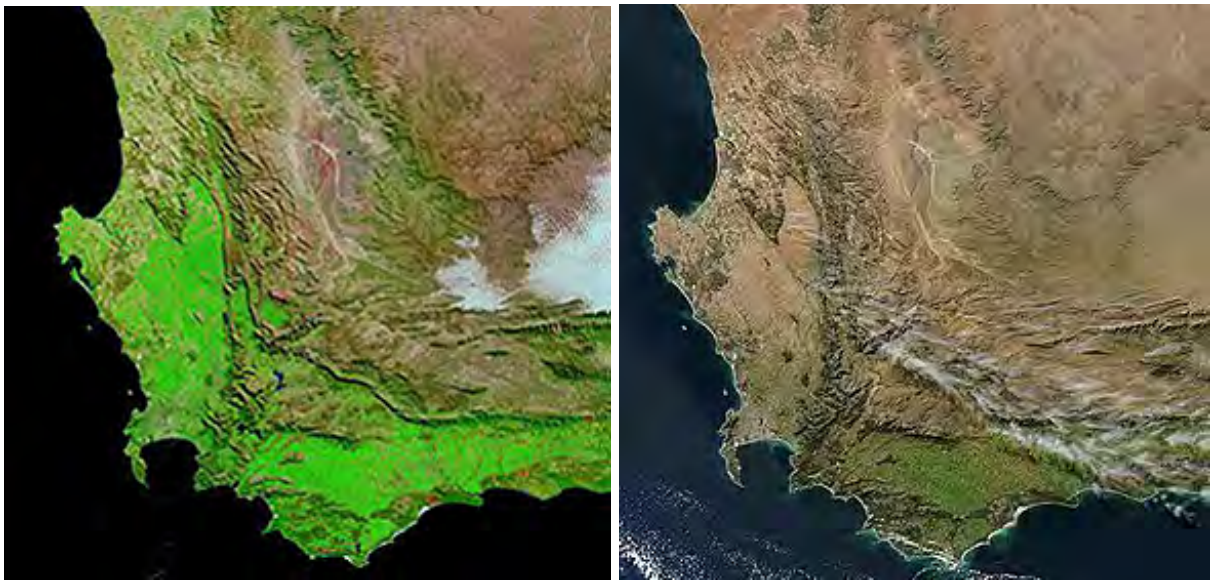
But droughts can replace the pattern of normal growth. This next scene covers part of East Africa around the Horn extending from Ethiopia and Somali. The browns indicate abnormally severe stress conditions in the vegetation; yellow is a bit worse than normal; and green correlates with local areas not suffering from the effects of low rainfall.

Drought intensified in 2004 and has threatened starvation for more than three million people in Kenya and neighboring countries. The rainy season normally starts in March and ends in June. In

2004 the rains fell mostly between mid-April and early May, amounting to less than 60% of normal, and exacerbating the previous years shortfalls. This MODIS-based NDVI map indicate the extent of the drought in June of 2004:

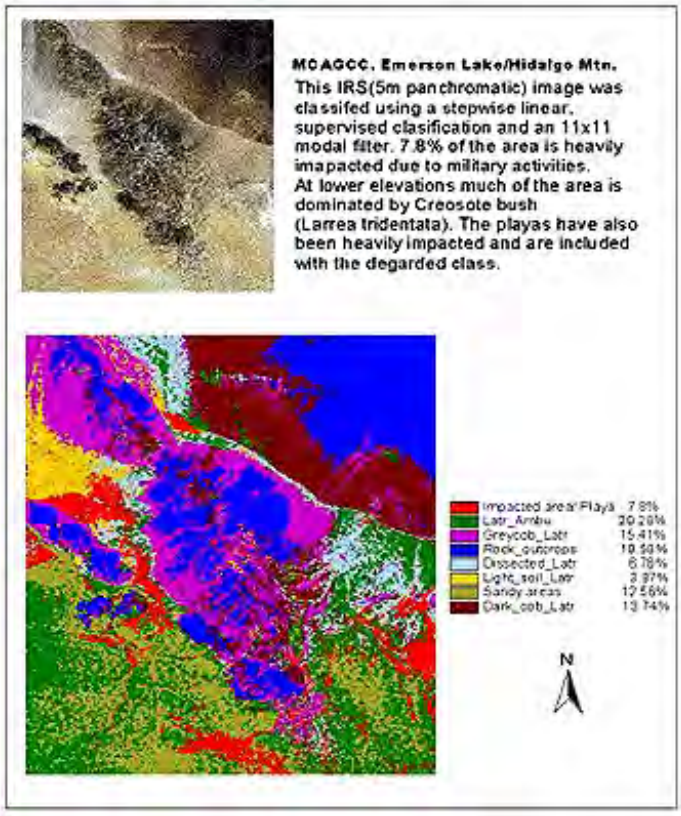


Winter wheat is a staple around Cape Town, South Africa. The next pair of MODIS images were taken exactly 1 year apart on July 21, 2002 and July 21, 2003 during the peak of southern hemisphere growing for this type of wheat crop. The greatly reduced green tones in the 2003 natural color image is a strong indication of poor crops owing to a pronounced drought in southern Africa; also affected are the natural grasslands that abound in this climate.



As an aside from the theme of this page, look at this SRTM perspective image of Cape Town, which show how it and its harbor are nestled within a mountain range; the scene is rotated 90° counterclockwise from the above pair:

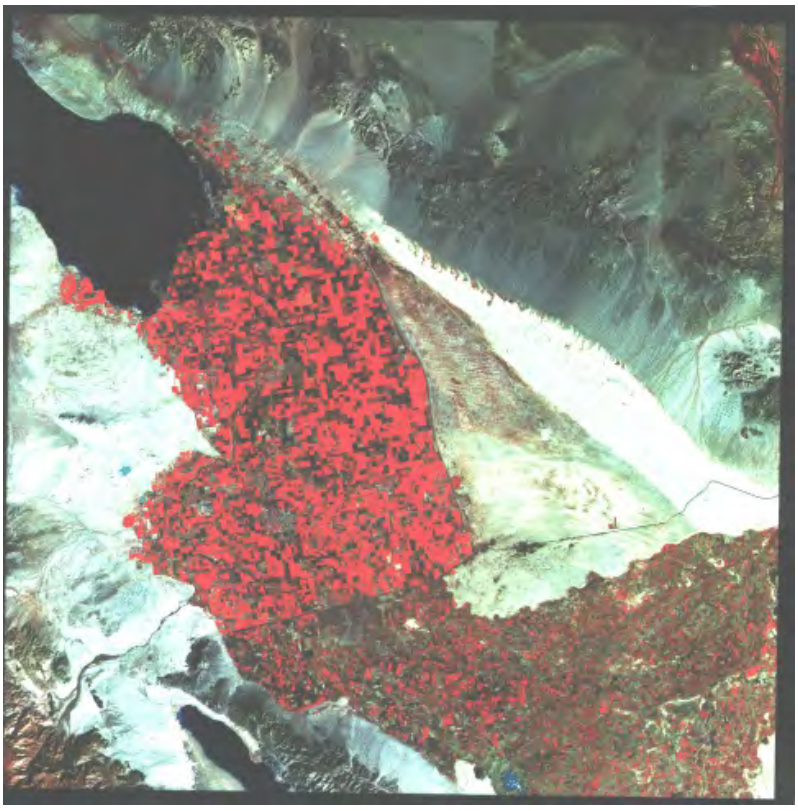
Dry or desert-like regions do actually have considerable vegetation. Arid country is characterized by grasses and brushland. Here is a IRS scene in the Mojave Desert in which the main land cover types have been classified:



The principal vegetation is the Creosote bush (Larrea Tridentata) which occurs in several settings.

Before leaving this agricultural theme, we want to look into one more example to show one of the most fertile and prolific growing areas in North America. which lies just a hundred miles or so south of the Mojave Desert. This astronaut photo shows much of southern California including the Mojave Desert, the coastal ranges east of San Diego, the Salton Sea and Imperial Valley further east, and a part of Mexico at the northern end of Baja California.

In this enlarged part of a Landsat scene, the Salton Sea and the Imperial Valley are hightlighted. The town of El Centro (bluish-black patch) lies about 16 km (10 miles) north of the Mexican border (very evident because of the sharp contrast with the agricultural activity). South of El Centro is Mexacali.

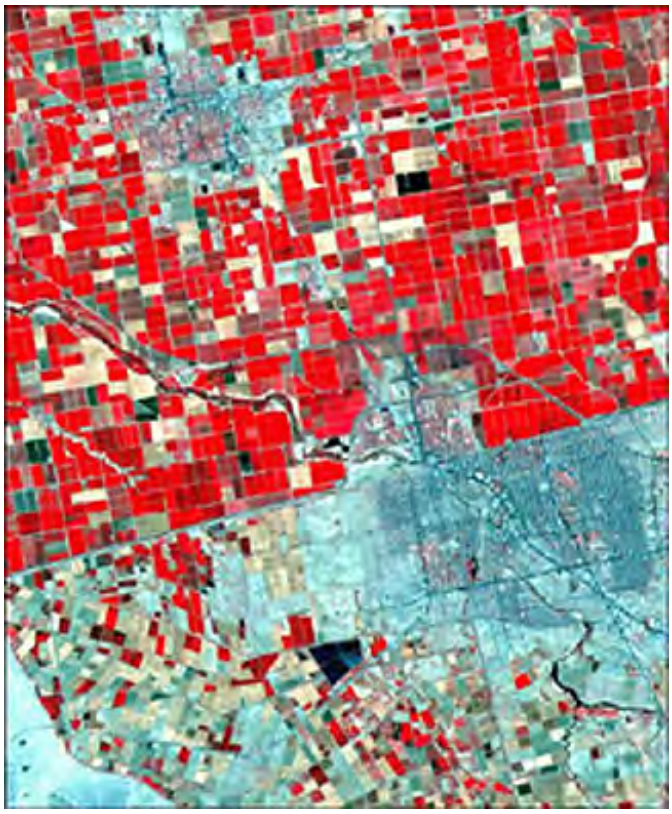


This agricultural region, one of the main producers of winter vegetables in the U.S., extends north and south of the Salton Sea, a saline body of water more than 49 km (30 mi) long, that fills a basin about 82 m. (269 ft) below sea level. This "Sea" was created by an overflow of water from the distant Colorado River (a small segment is visible in the upper right corner) shortly after the beginning of the 20th Century. Floodwaters poured through low dry washes, traveling westward more than 64 km (40 miles) to empty into the lowest part of the Coachella Valley. In this desert climate, that water is slowly evaporating and turning brackish (moderately salty) and is thus not suited for direct irrigation.

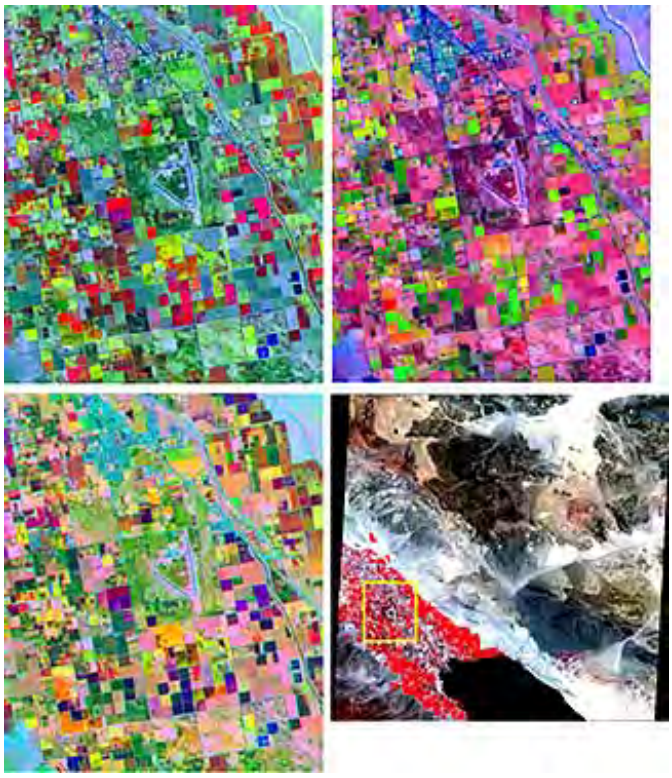
The lake-like water at the image bottom is Laguna Salada, which undergoes seasonal drops in level that at times reach a dry state, exposing playa lake beds. East of the Valley are the Chocolate Mountains, part of the Basin and Range system, against whose flanks are conspicuous alluvial fans. The bright strip in the right half of the subscene is the Algodones

Dunes field, derived from beach sands left at the surface after an ancient predecessor to the Salton Sea had occupied the Salton Trough, a structural basin between the Coast Ranges (lower left) and the eastern mountains.

Today, canals from the Colorado River transport water to the sea. The biggest canal in this scene is the prominent All-American canal. Mild winters promote year-round farming (up to three harvests) in the Valley, with cotton, sugar beets, lettuce, and citrus being the main crops. Most fields are in full growth in this April scene, as indicated by the bright, uniform reds. Differences in land use practice and availability of water (no major canals) account for the pronounced decrease in agriculture on the Mexican side. This is visually quite striking when this ASTER close-up of the border land use is examined (El Centro is the town in the U.S.; Mexicali is in Mexico).



The individual fields, which show small differences in the two false color composites above, are hard to pick out in these traditional renditions. In the image below, there are three renditions of the Coachilla Valley farmlands north of the Salton Sea made by different band combinations from the ASTER spectral bands (see page 16-10). These are listed in the caption.



3-11: **How might a sequence of Landsat or SPOT scenes taken over the months or even years be beneficial to the economic and environmental management of this region?**
ANSWER

Clearly, the differences in expression using different combinations of just these bands suggests the ability to identify, differentiate, and classify the types of crops being grown. The additional bands on ASTER improve this capability, allowing greater accuracy in specifying crop types.

Before leaving the topic of conversation of arid land by irrigation into highly productive agricultural farms, which is so evidently successful in the Imperial Valley, here is a dramatic "before and after" pair of Landsat images that show the extent of change possible when water becomes abundant. The area shown in the images below is in southern Turkey just north of the Syrian border. The 1993 image shows that some croplands have been developed using subsurface water. Since then, the completion of the Ataturk Dam and Reservoir has permitted water to be spread over the area from irrigation ditches. The end result seen in the 2002 Landsat image - back to back cotton farms - is dramatic and obvious. Because such capability is still large for much of the world's arid lands to be developed using dam water, the prospect of famines over the global seems small for the foreseeable future.

August 23, 1993

August 24, 2002

Crop monitoring to control health and estimate harvest output has become a sophisticated system using various multisource data sets. Here is one scheme used by federal agencies and private commodity firms to make predictions or issue warnings about crop damage and other problems.

August 23, 1993

The LACIE, Agristars, and FIFE field experiments were early examples of proof-of-concept. Crop identification can be as high as 90% accurate. The programs require considerable onsite inputs but depend on satellites to provide high resolution multispectral data to extrapolate from the few training sites needed for classification of relatively small areas to maps of the vast regions that are needed to make reliable estimates at country or global scales. We shall consider this theme again in the first part of Section 13.

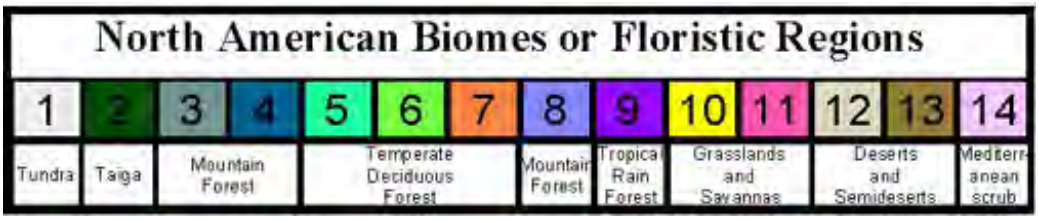
PREVIOUS NEXT

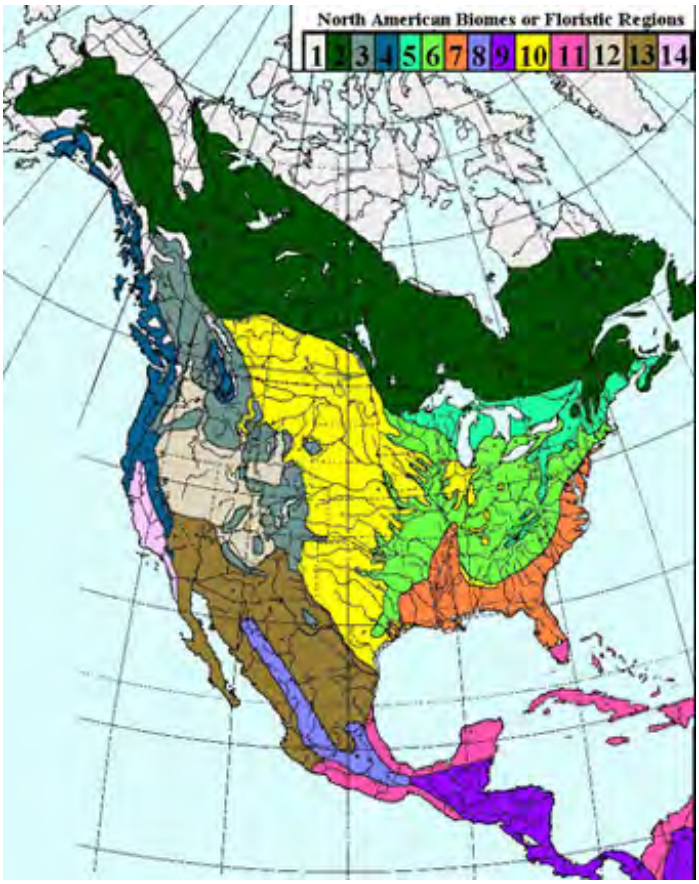
Primary Author: Nicholas M. Short, Sr.

Global warming will obviously affect natural vegetation. One indication that this is a growing problem is the increased activity of various beetles that invade and infect various conifers, especially pines. A dramatic example is the spread of the Mountain Pine Beetle into much of western North America, most emphatically in British Columbia. Millions of acres of lodgepole pine trees are being infested resulting in severe economic loss. The main defense is early spraying. Earth-observing satellites such as Landsat and IKONOS provide timely monitoring that guides foresters in decisions to spray. We preview the next page - ecological damage - with this case study of a forest hazard. **The study demonstrates two special capabilities of satellite remote sensing: interseasonal repeat coverage and contributions from a variety of sensors providing various spectral bands, different resolutions, and scales of resulting imagery.**

A Case Study: Monitoring Diseased Pines

Nearly all the vegetation biomes identified in the global map at the top of page 3-1 are found in North and Central America, as evident in the next illustrative map:





Credit: University of Tennessee

Some of these biomes are predominately composed of evergreens. There are two types of evergreens: Evergreen Needle Leaf (pine, spruce, fir, etc.) and Evergreen Broadleaf (laurel, magnolia, etc.). The pines are usually coniferous (cone-bearing) and in North America are widespread. Pine habitats fall into three main groups: Boreal Forests (mainly in Canada); Mountain Forests (further to the south); and Southern Forests (southeastern U.S.). Although examples from all three are included on this page, the concentration will be on Boreal Forests. In Asia, mainly Siberia, these are also called Taiga. These are the second most widespread plant biome type in the world, as indicated in the next map.

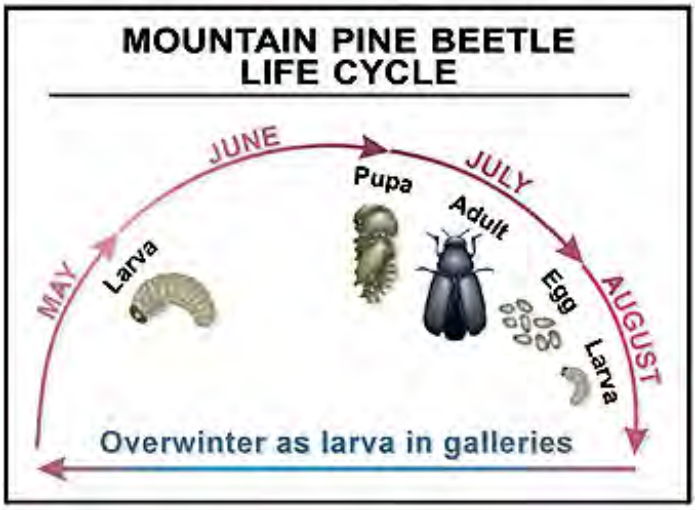
In western Canada's Boreal Forest the most common type of tree is the Lodgepole Pine. Its distribution is vast and many forests are almost continuous stands of these trees (next figure). They are a main source of lumber (as are the other pines mentioned above) and a very valuable commodity. Threats to their normal existence, such as fires and disease, can have a huge economic impact. On this page, we shall dwell upon just one specific and serious problem - the infestation of the Mountain Pine Beetle (MPB) - as a case study showing how remote sensing is helping.

Various species and genera of beetles are known to attack pines and other evergreens. Among common ones known by their non-taxonomic names are the Southern Pine Beetle, the Douglas Fir Beetle, the Spruce Beetle, the Turpentine Beetle, annd the Pinion Ips. Here are five different species in two genera.

The beetle ravishing the western Canadian boreal forest is the Mountain Pine Beetle, Dendroctonus ponderosae, a small (less than 0.5 cm) shown here:

A good, reasonably brief webpage that provides an overview of the MPB problem is this Beetle Infestation website prepared by the Alberta (Canada) Forest Service. The main stage in the killing of a pine tree by the MPB is known as a Red Attack, a name evident from the color of this Lodgepole pine after the disease runs its course.

The MPB usually begins as an early stage known as the Green Attack. The Red Attack is concentrated during the second growing season (sometimes over two years in high elevations). The life cycle during the main season is shown here:



Here is a tree trunk surface showing larva evolving into pupae:

During much of the attack, the MPB dwells in the inner bark. It then proceeds into the inner tree by tunneling to deposit its eggs. This is a sawed surface exposing the interior with borings and eggs:

The tree fights the MPB by secreting resin on its surface producing pitch tubes seen in this outer surface view.

Later in the process further damage is done by the invasion of Bluestain fungi, which respond to the activity of the MPB, as evident in this exposed cut into a dead pine:

The final stage, in which the needles become inert and fall is known as the gray stage.

Now to field views of MPB infestation. The next group of field aerial photos shows stands of Lodgepole pine in British Columbia that have been infested by the MPB. Read the captions for further information:

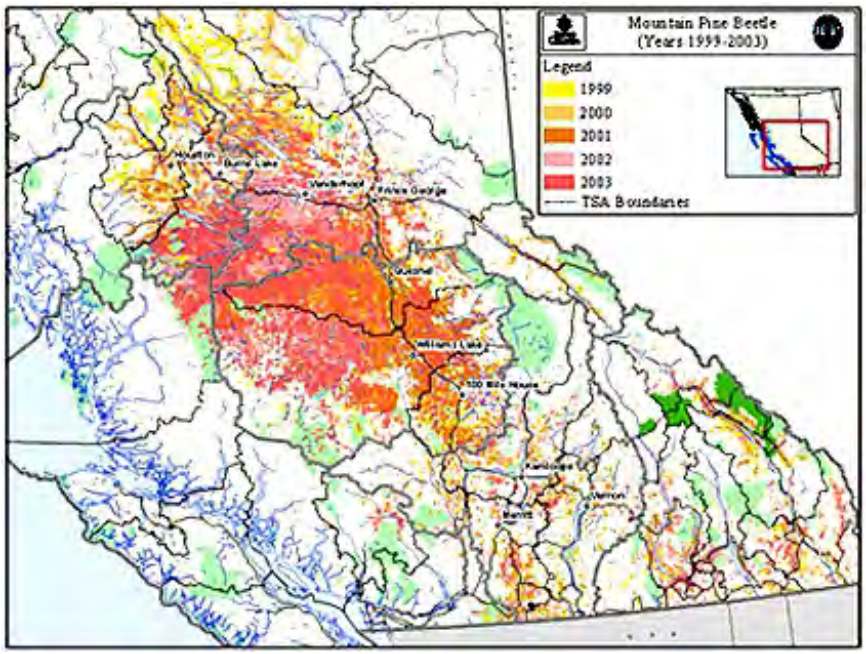
Once the MPB damage becomes widespread, the main recourse is to cut down the diseased trees, thinning the forest or even making it barren. Some of the trees can be partially salvaged for lumber or other uses.

Others are too far gone and must be burned:

The "trick" is to catch the onslaught of the MPB disease before it progresses too far. One obvious approach is to spray individual trees. The two most effective sprays are Carbonyl (Sevin) and Permethrin (Astro). An alternative that often helps is to inject the tree with chemicals that provide some resistance.

But the humongous numbers of pines involved make individual tree spraying labor-intensive and usually impractical. The standard solution is aerial spraying, preferrably in the early stages of the MPB infestations.

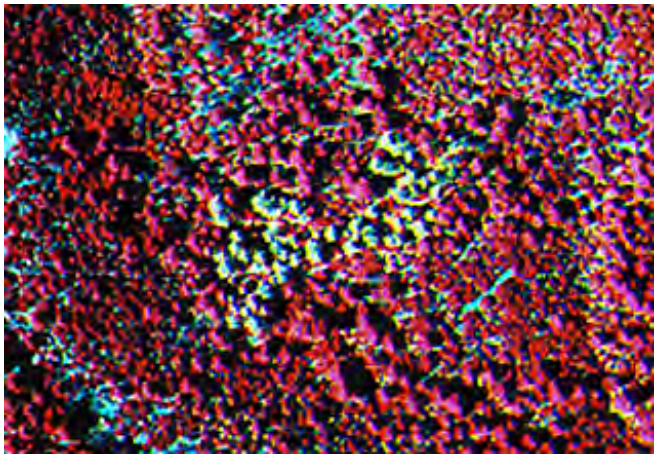
British Columbia at present is the most threatened part of the Boreal Forest in Canada. Here is a map showing the extent of the region affected.



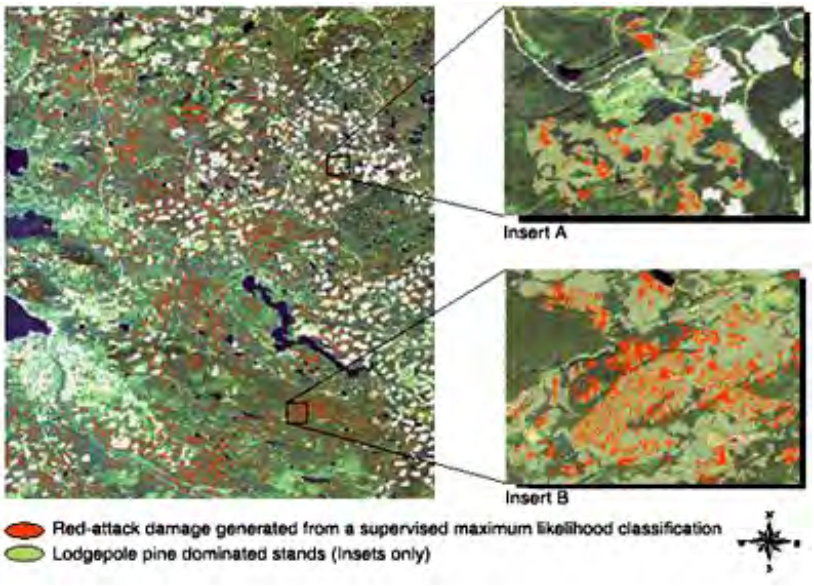
As stated above, infestations by various beetle species is ubiquitous in North America and worldwide. Here are two examples of damaged pine forests in several regions of the United States:

Prior to the plethora of earth-observing satellites, detecting and monitoring the onset and spread of MPB tree-destroying activity was carried out through aerial surveys and reconnaissance flights. Here is what a forest agent would see from a plane passing over early stage infestation.

Experience with color IR film showed a greater sensitivity to the progress of the disease inflicted by the MPBs; diseased trees in blue.

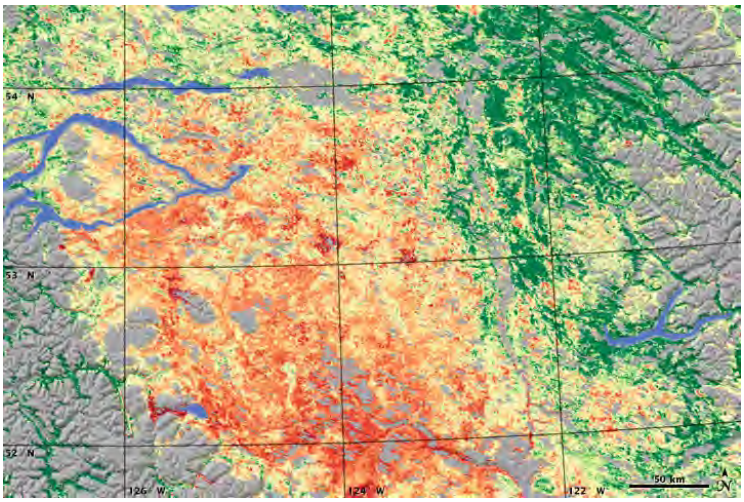


Landsat has been a major means of watching the progression of MPB blight over large areas. Here is a Landsat subscene in which Maximum Likelihood classification has pinpointed areas suffering MPB damage at the time.



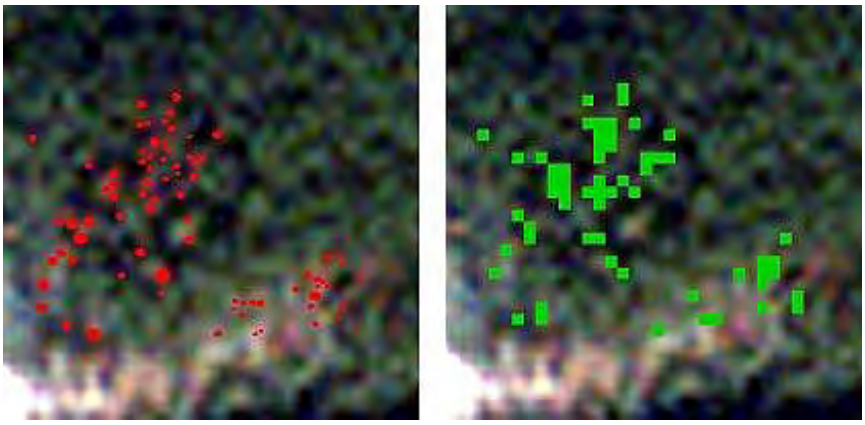
Terra's ASTER sensor provides broad area coverage of MPB infestation, shown in shades of olive brown, in part of British Columbia.

Terra's MODIS sensor has been used to produce and periodically update maps enclosing areas experiencing beetle damage:

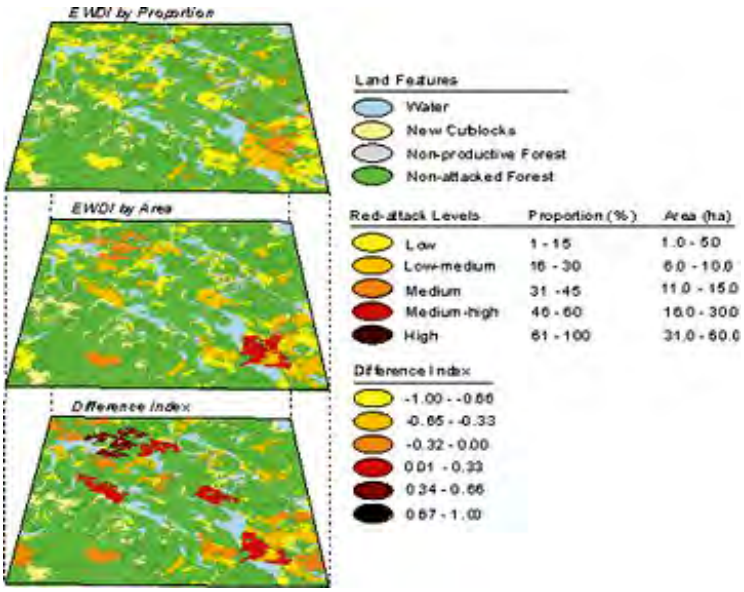


IKONOS high resolution imagery covers smaller areas but provides high detail, so that these images are comparable to aerial photos.

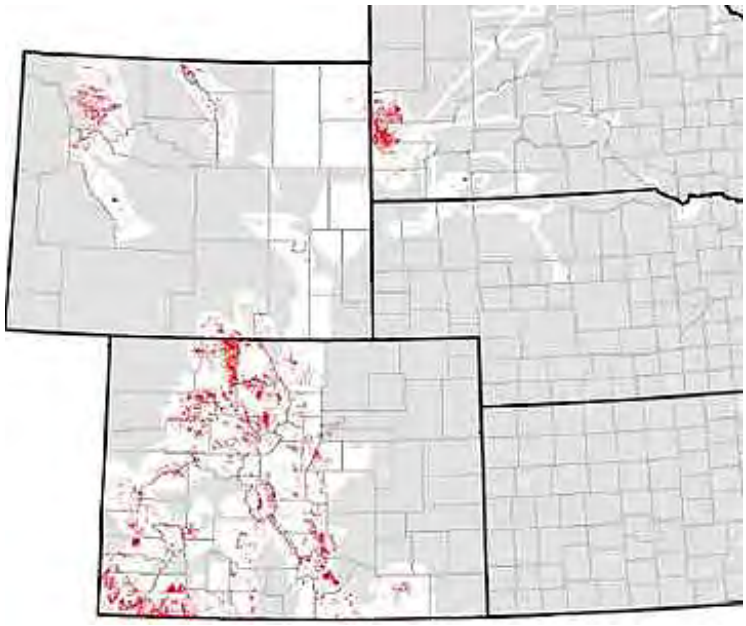
This pair of images compares areas identified as in the Red Attack phase as extracted from an aerial photo (left) and an IKONOS subimage right.



Satellite imagery is used to calculate NDVI (see page 3-4) whose values deviating from normal are indicative of tree stress. For the British Columbia infestation, Canadian foresters have devised the Enhanced Wetness Difference Index (EWDI), which uses temporal change detection (multidate) imagery to determine MPB damage. Here is a typical result:

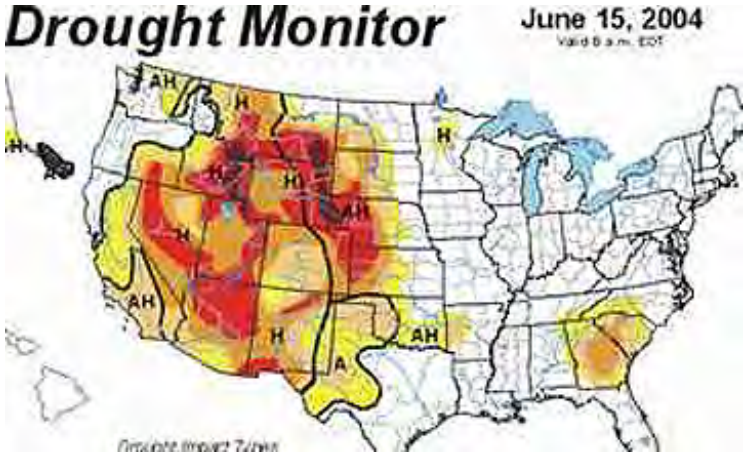


The current infestation in Canadian pines has been intensifying in the past decade. This has been attributed to warmer average winters (intense cold can kill the beetles) and less summer moisture (which decreases wetness). This is now a general condition in western North America. Foresters in the U.S. began to note MPB problems in some of the western states:



By 2005 this condition had affected the entire western U.S.

Coincident with this increase is a conclusion reached by experts monitoring drought conditions that the U.S West is experiencing the worst sustained drought in 500 years. Consider this map:



Many experts attribute this drought situation to global warming - the MPB infestation being one of many indicators. We shall examine other evidence for global warming on page 16-2.

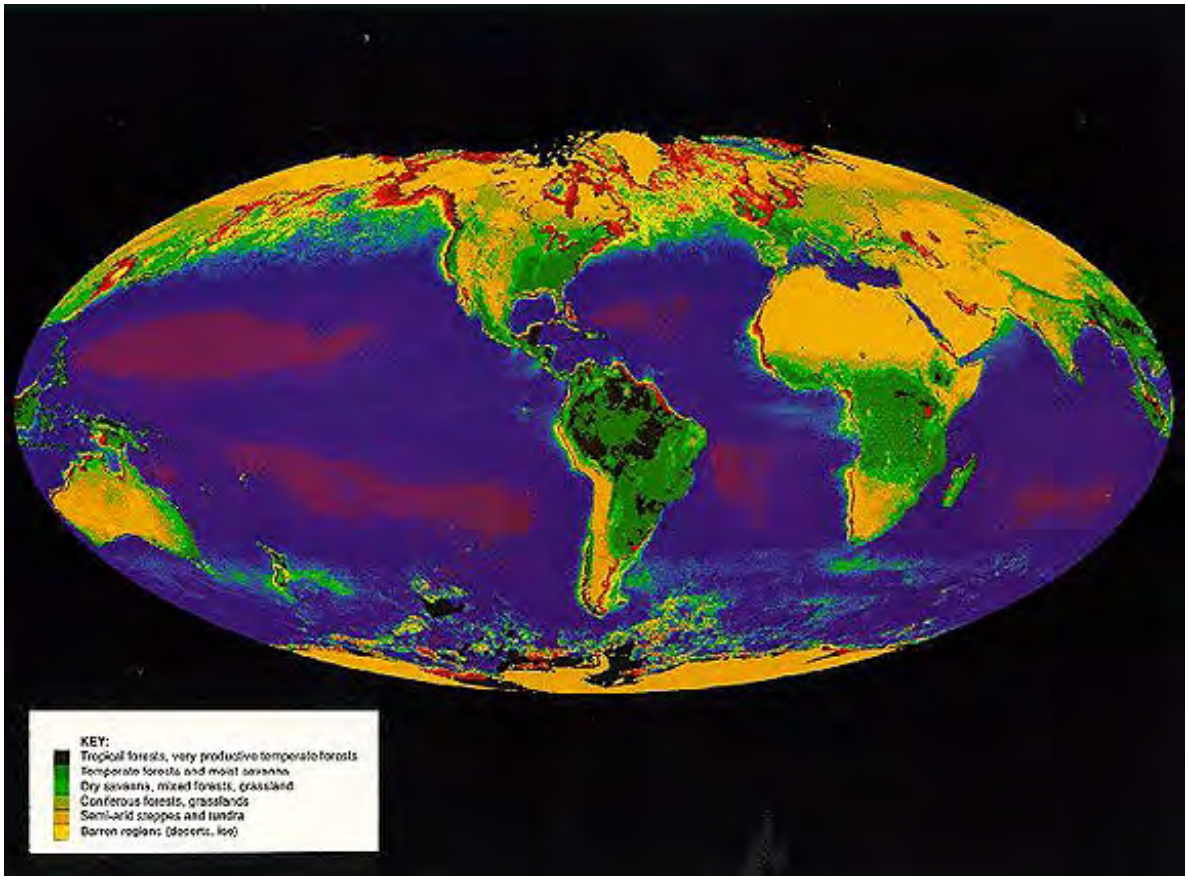
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

In terms of areal extent, the vast majority of vegetation that shows up in images acquired by satellites is not agricultural but forests and grasslands. Some regions, such as the Tropics, display green, active vegetation all year round; others, such as temperate (humid) and boreal, do not appear as reds in the false color mode during the dormant season. Deciduous trees can usually be differentiated from evergreen-dominated forests. Globally, the amount and variety of vegetation cover changes with the seasons. On this page, examples of tropical and boreal forests in the Amazon and Canada are presented. Removal of forest cover by clearcutting is also illustrated for both regions. Forest classification and biomass estimates are touched upon.

Forest Applications

Let's switch attention now to the world's forests, natural and planted and maintained for future harvesting. The distribution and condition of these forests, along with other principal types of vegetation in the global biosphere, can be determined and refined periodically, by analyzing Vegetation Index data acquired by satellites. As an example, here is a plot of the distribution of general forest classes, as they were identified in AVHRR data from NOAA 7 and 9 metsats. Colors that are not identified in the legend include red = shallow, coastal or inland waters; blue = ocean waters; purple = deep ocean; black in the arctic and antarctic regions = not classified.

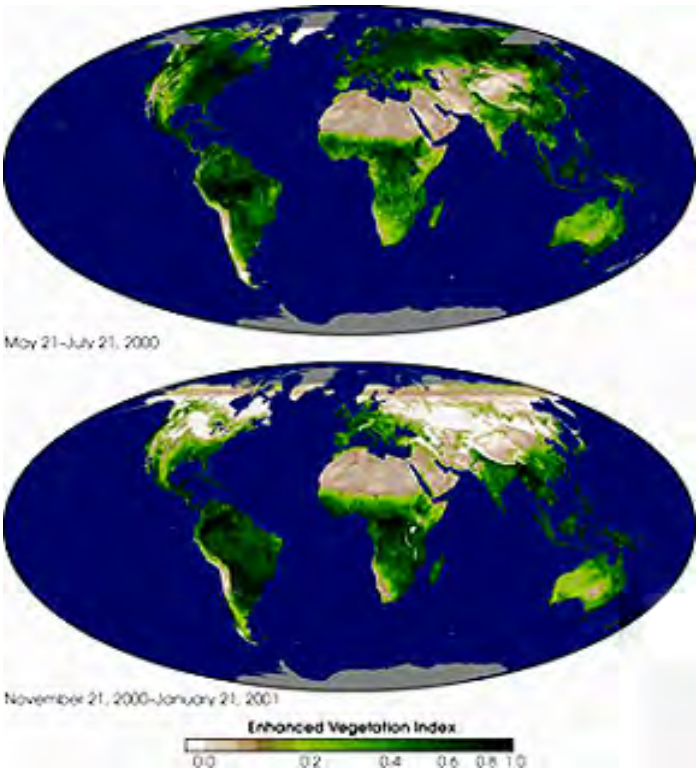


3-12: **Which continent appears to be most widely vegetated; least widely?** **ANSWER**

One obvious application lies in determining the geographic and areal distribution of major types of sylvan vegetation and their ecological settings. Generally, Landsat and other remote sensing

systems can readily distinguish between deciduous and coniferous forests, and can recognize shrublands, savannahs, and various desert ecosystems. But, identifying most tree types at the species level is much more difficult, unless high resolution images are available and the trees have distinctive crowns and leaf shapes that produce fairly explicit spectral signatures. Easier tasks, that have widespread applications, involve discerning defoliation - either in the extreme with tree removal by clearcutting or through denuding trees of their leaves by insects, and assessing forest fire or storm damage.

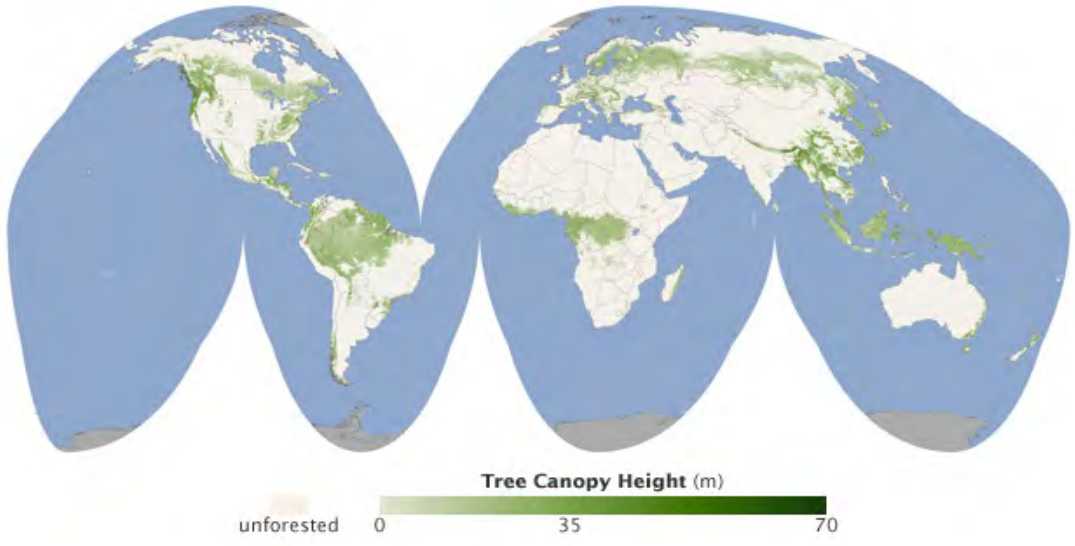
We all know from experience that vegetation cover varies seasonally. This reaches maximum differences in active growth in the higher latitudes (and altitudes) and is less changed in the equatorial and tropical regions of Earth. Here are two global vegetation cover maps made from MODIS data, with the growth data collected over time frames of weeks:

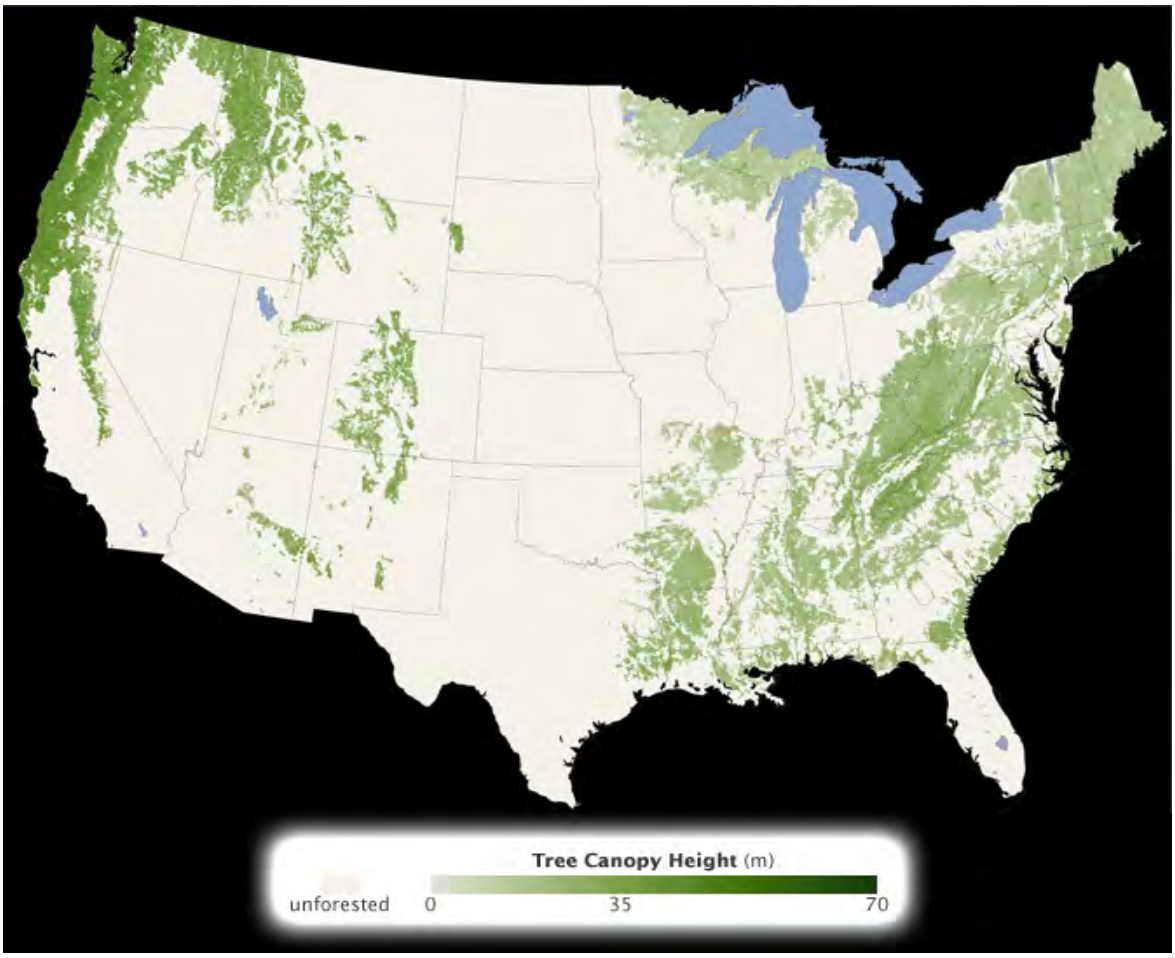


This map indicates that parts of the world have extensive, in places continuous, forest cover. One such region is the Congo in central Africa (where the notion of "jungle" was fostered in the 1930s by the Tarzan movies). This next image shows a nearly complete forest canopy in Africa so dense that, in the false color rendition, the tone is a uniform red. That type of cover marks true jungle, matched elsewhere mainly in the Amazon basin described later on this page.

Look carefully at this image - the effects of the underlying topography is expressed in the canopy by variations in the shading..

Two maps published in 2010 show tree canopy heights worldwide and in more detail for the U.S. MODIS and a Lidarlike instrument were used. The maps do not display areas of small forest patches or groves of trees, but only the *continuously forested* areas. In the U.S. this coverage is associated mainly with mountainous regions where slopes are too steep for building.

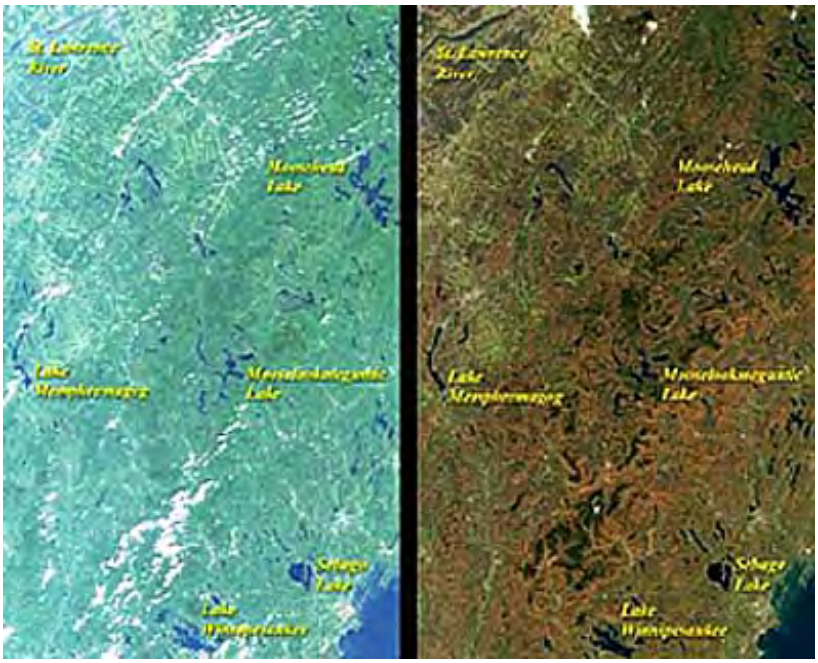




In the U.S. the most extensive and continuous forests are found in the Pacific Northwest and the Atlantic Northeast. Here is a Landsat image of forestland in Maine:

There is an esthetic side to watching tree foliage from space at different times of the year. In Fall, New England is famed for its marvelous leaf colors, bringing millions of tourists to witness this panorama of visual beauty. Such imagery helps TV Weather commentators in advising residents in their listening area as to the optimum times and places to view the countryside. This Terra MISR image pair of northern N.E. (western Maine into New Hampshire and eastern

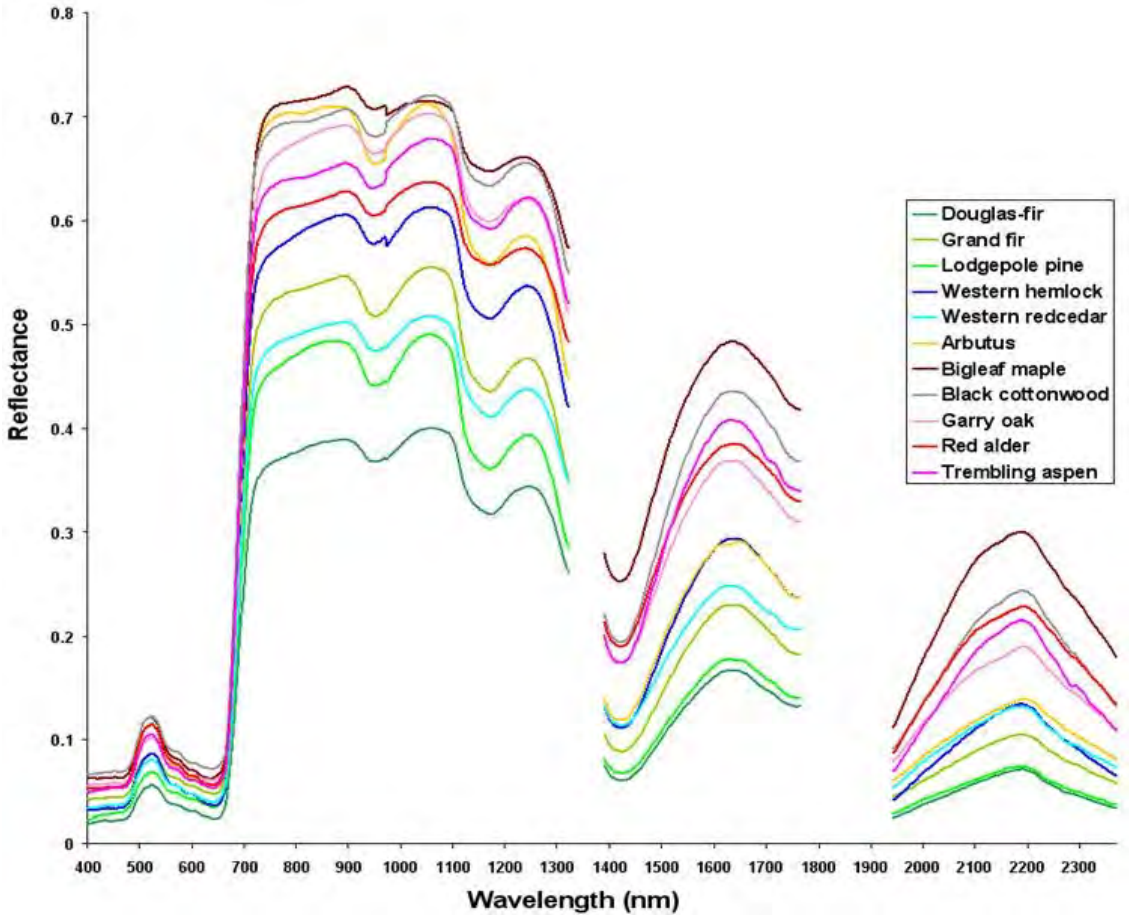
Vermont) portray almost true color views of foliage in late August 2000 when the forests were near their height of greeness (left) and past their peak (right) - most leaves are now dropped - in mid-October 2000:



Let's also look at a companion MISR image showing southern New England. Here, being further south, the image shows more of a red tone, in keeping with the fact that the leaves (red and yellow in color) have not yet fallen.

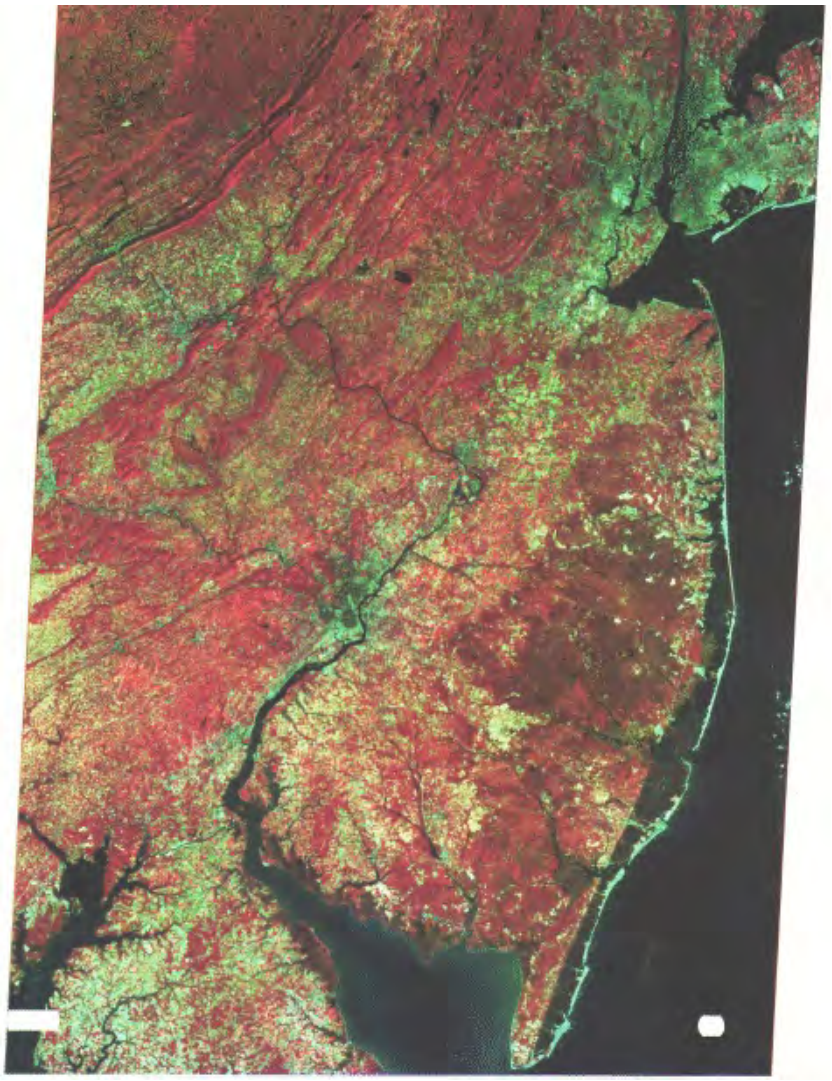
Tree and Forest Classifications

By now, the question should have crossed your mind as to how well individual tree species (and the individual trees themselves) can be identified by various remote sensors and data processing techniques. Generally, if complete spectra over Vis, NIR, and SWIR regions of the EM spectrum are available, the discriminatiion of species can be accomplished with acceptable accuracy. This set of spectral signatures support that claim:

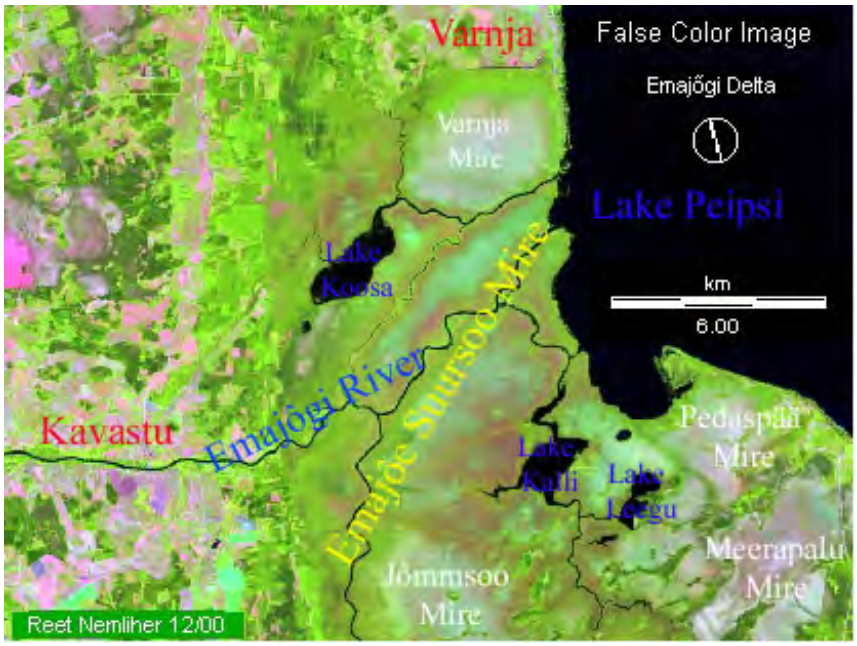


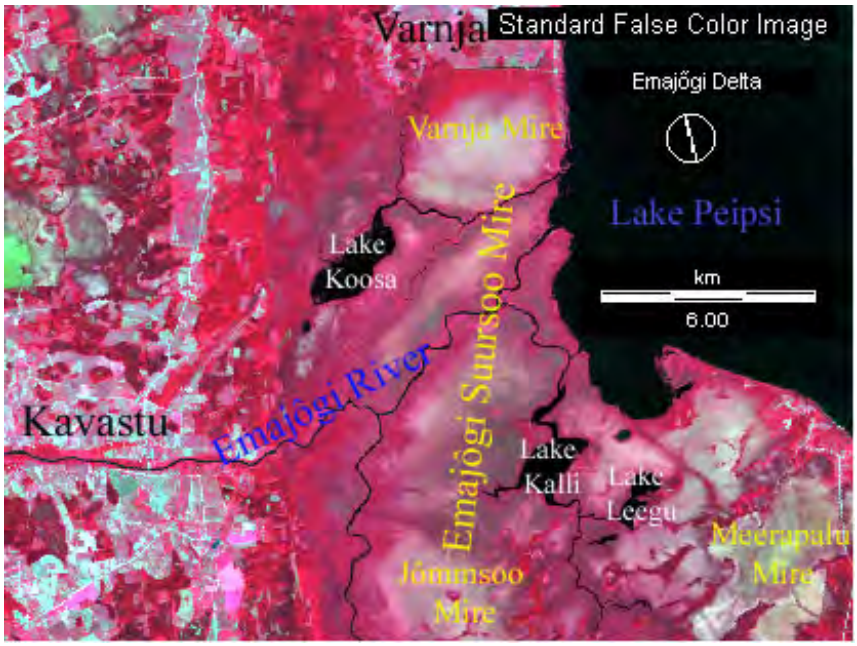
Note that the reflectances for the several coniferous tree species (fir, pine, cedar) are lower than those for deciduous trees (maple, oak, aspen)

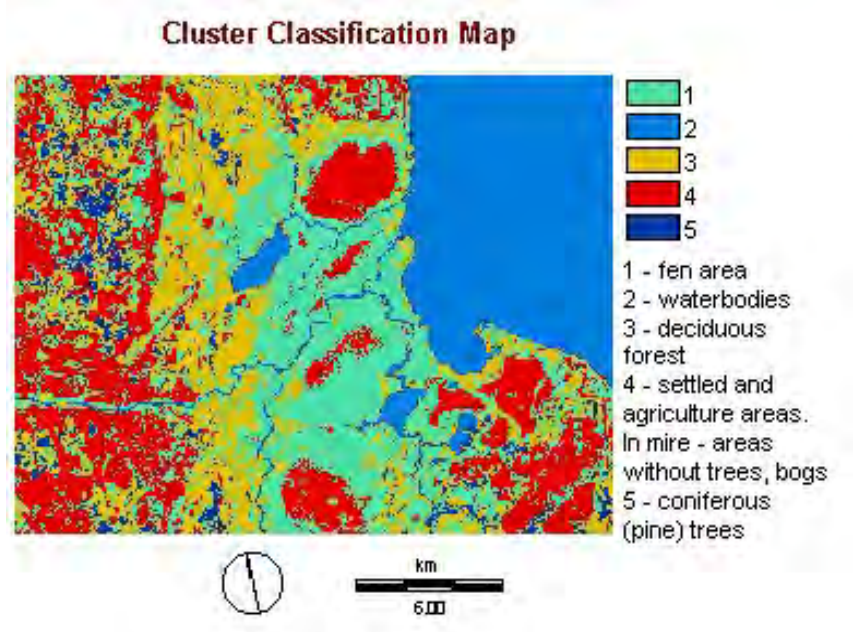
For Landsat, the answer to species identification is generally "not as well as desired" because of both spatial and spectral resolution. In general, at lower resolutions dark evergreens can be separated from deciduous trees in temperate and boreal ecosystems. The pine forests of eastern New Jersey, which we saw earlier in the Introductory Section, are distinguishable from the deciduous forests in that state by their darker reddish-brown tones. We reproduce this image here; note other areas in the scene, such as in the Appalachian segment, that are darker, and thus indicative of the occurrence of coniferous trees:



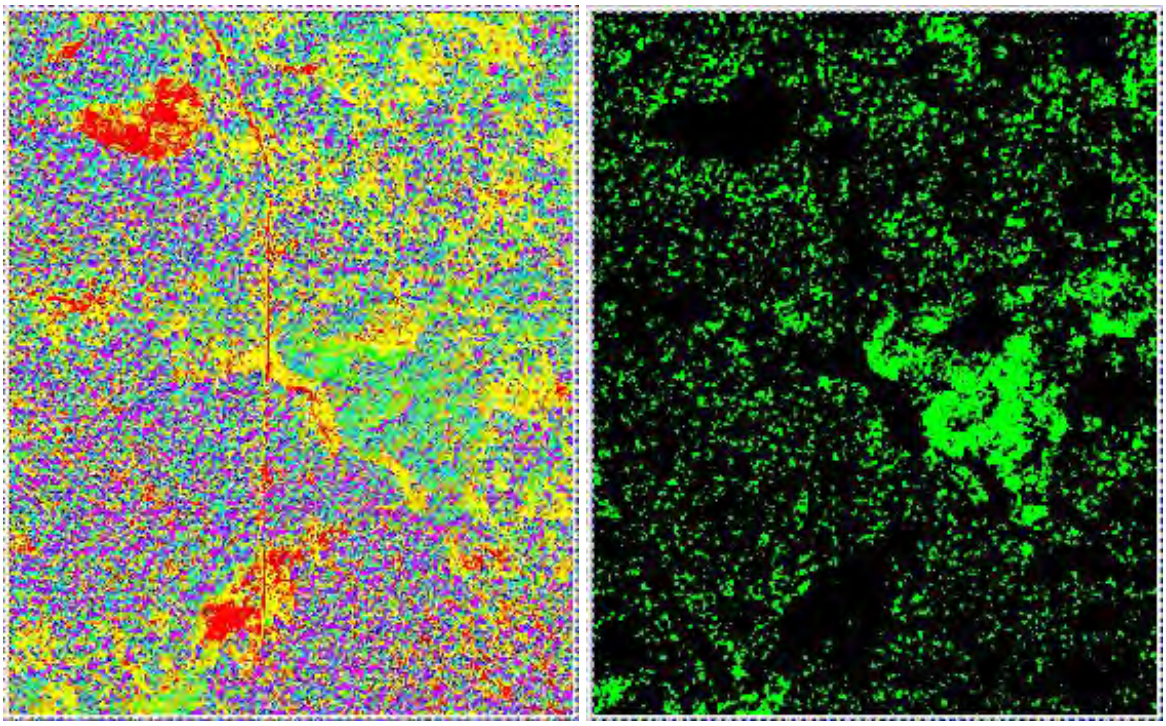
This next Landsat7+ ETM study of evergreens and deciduous trees in an area in Estonia shows the former appear darker in both natural and false color images relative to the tone brightness of a stand of deciduous trees; the terms "mire" and "fen" refer to wetlands and bogs. The band data were used to produce a classification - at the type rather than species level - of this scene:



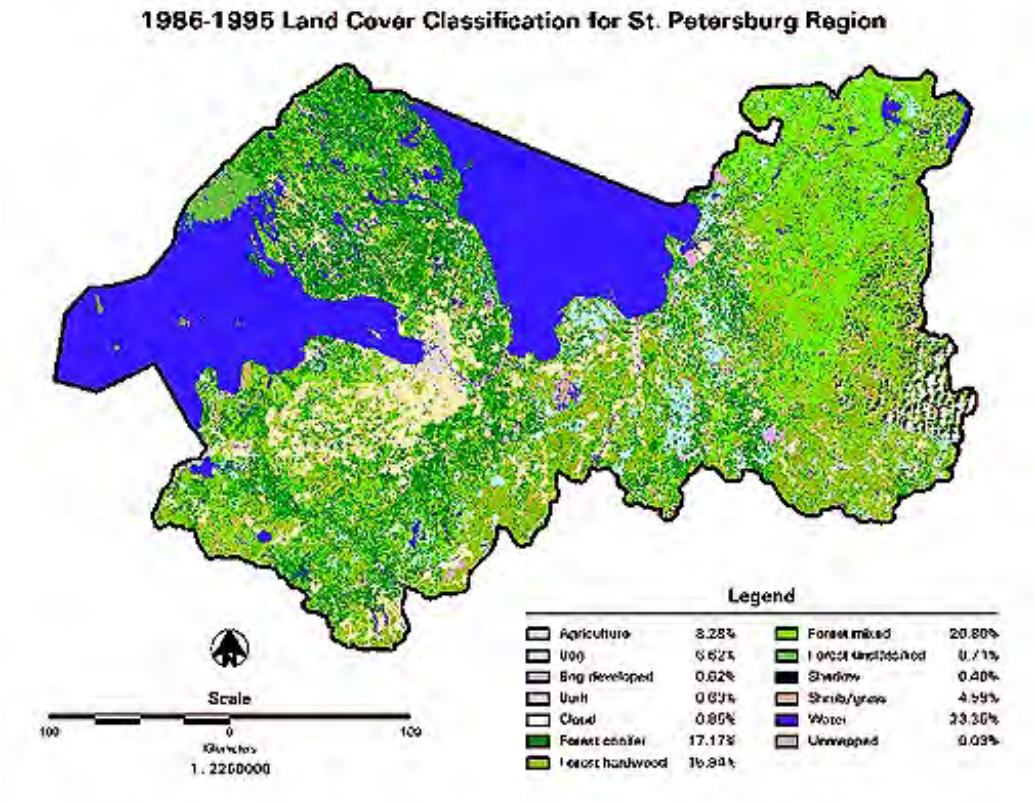




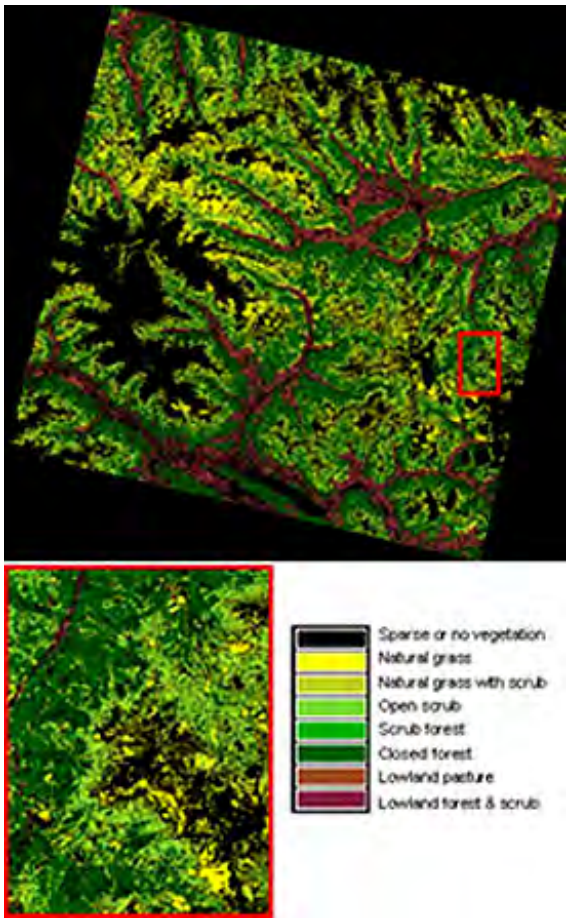
A similar result is also the case for trees in the Pike National Forest in central Colorado near Hayman. An unsupervised classification of a forest area in this Colorado preserve imaged by the Landsat TM yielded the result shown in the top image below. When ground truth was added to the process, evergreens are shown in blue; aspens in green, clear areas in red. A new map (bottom) in which just the aspens are located, as shown in green, proves that some individual tree speciation can be reliably determined.



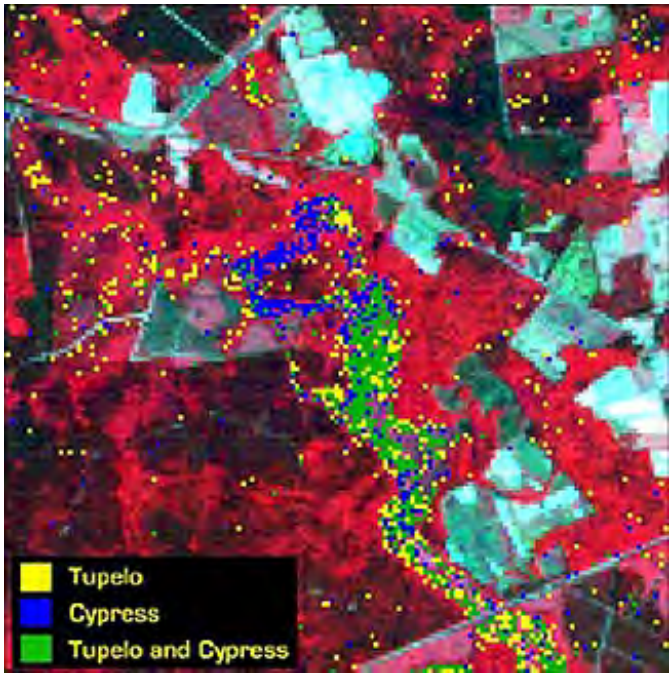
Broad categories of vegetation and other land cover types can be classified as indicated in this illustration which shows a fair number of categories - most vegetation-related - in the region around St. Petersburg in Russia.



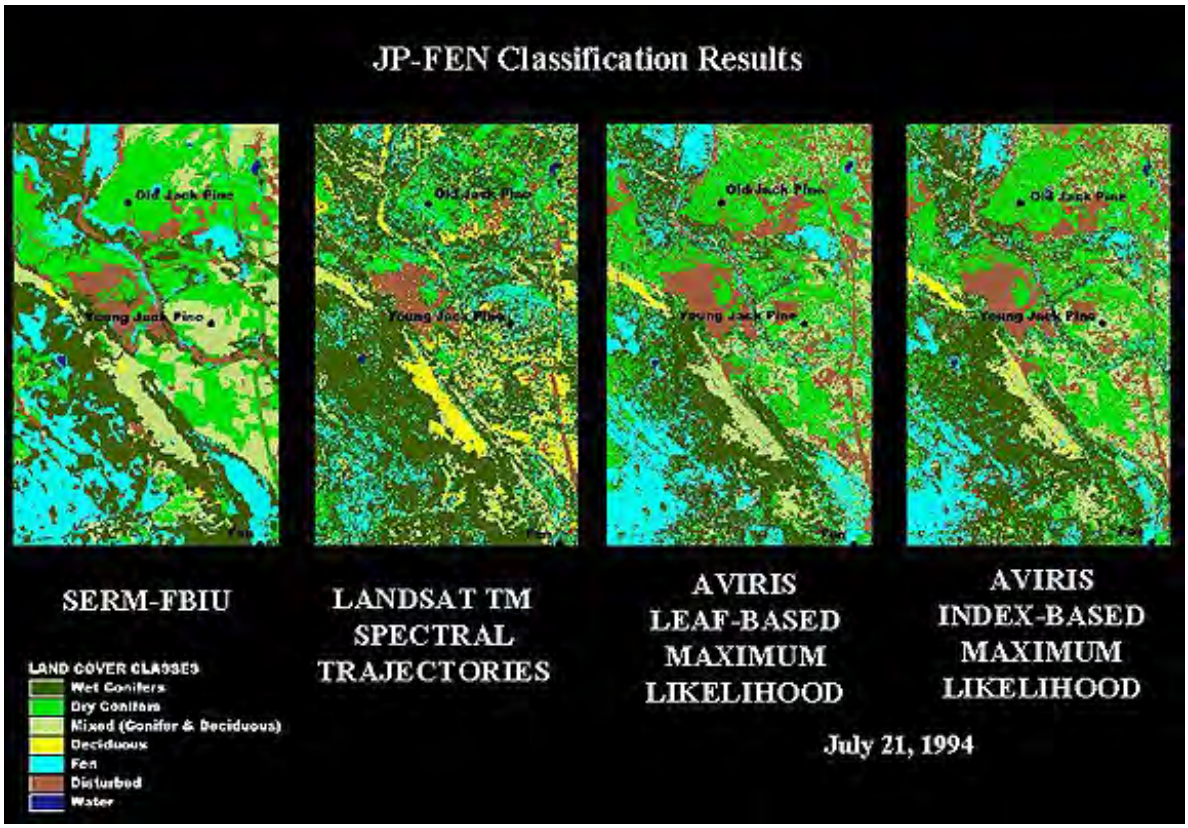
Staying in Europe for the moment, examine next this supervised class map of vegetation in hilly terrain in Austria:



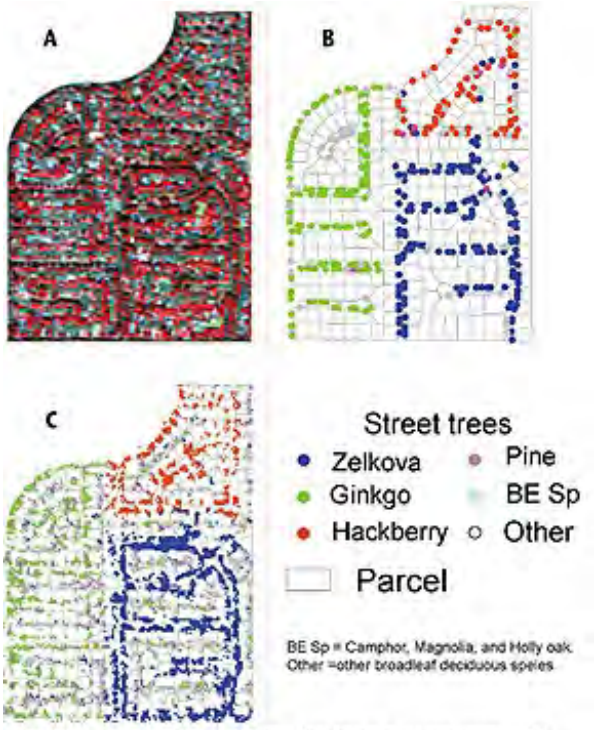
Again, as was true with crops, high resolution imagery greatly improves the ability to identify and locate individual tree species. In this IKONOS 4 meter subscene of a forest near Charleston, South Carolina, individual Tupelo and Cypress trees, and a class of the two mixed, have been pinpointed by classification.



Perhaps the best example on this page of the efficacy of remote sensing to classify vegetation and other ground cover types is this series of images showing a forest in Canada that contains conifers (mostly Jackpines), deciduous trees, and fens (bogs). On the left is a map made from ground and aerial photos. Next to it is a surprisingly good Landsat TM classification. The right two images are made from AVIRIS data; the left one is based on using the Leaf Area Index values and the right uses the NDVI approach to calculate the classes.

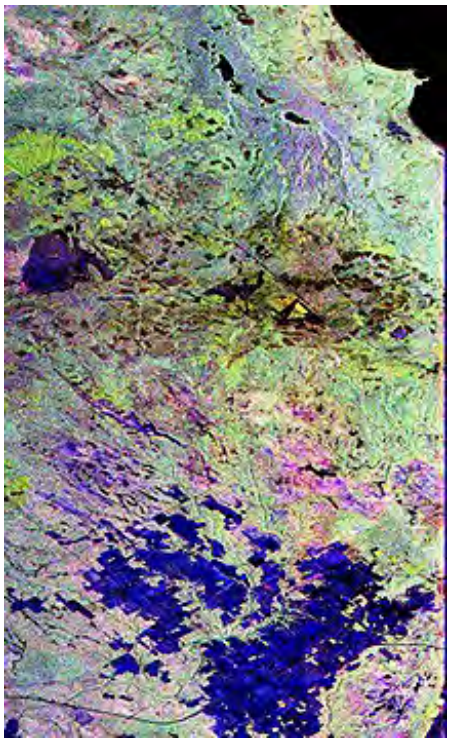


The above AVIRIS images imply that high resolution hyperspectral data, coupled with adequate ground truth, can be quite effective in classifying trees down to the species level. This study of urban tree classification along the streets of Modesto, California supports that conclusion:



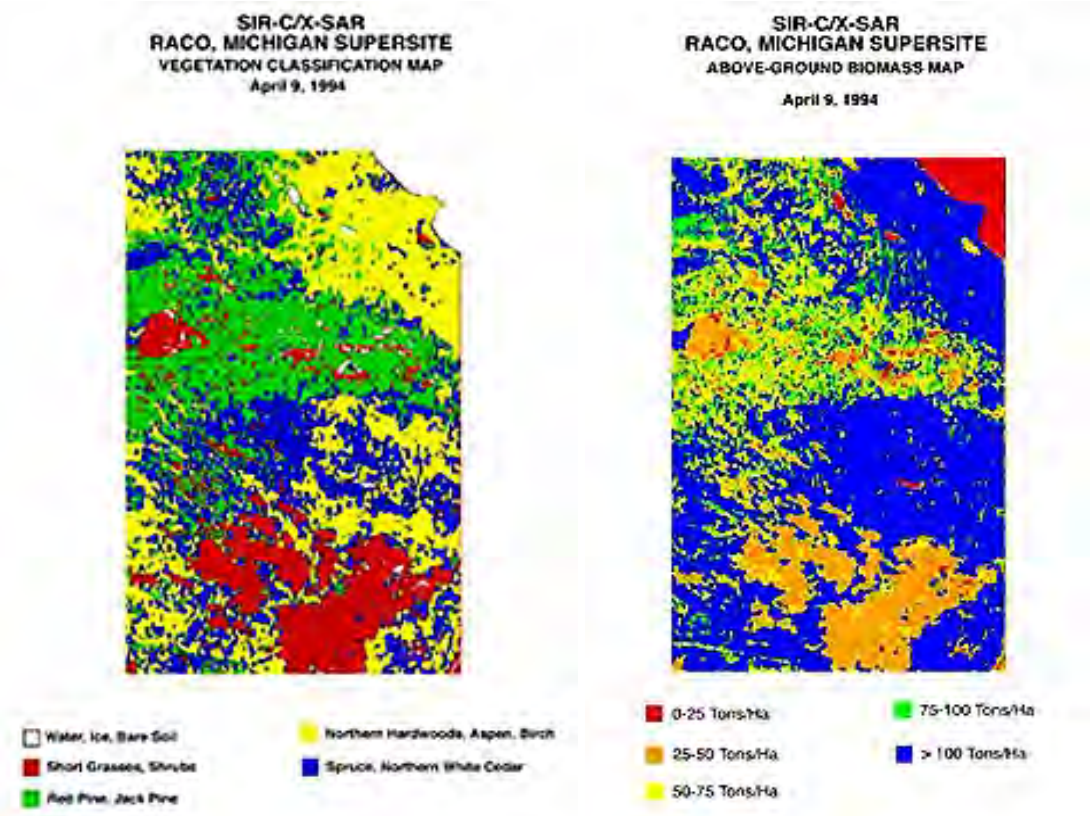
A tree species distribution of a selected Modesto, Calif., area includes a color-infrared AVIRIS image (A), a tree species identification from the GIS database (B), and a tree species identification from AVIRIS data (C).

Radar has been used to determine tree types. as shown in this SIR-C investigation of a NASA Supersite at Raco, Michigan on the Upper Peninsula next to Lake Superior. The host Shuttle was STS-59 which flew in April, 1994. The target area was heavily forested, as seen in the multiband color composite:



The blue area is mostly cleared fields. Light greenish-yellow marks red pines; brownish-yellow denotes jack pines, and purple locates deciduous trees.

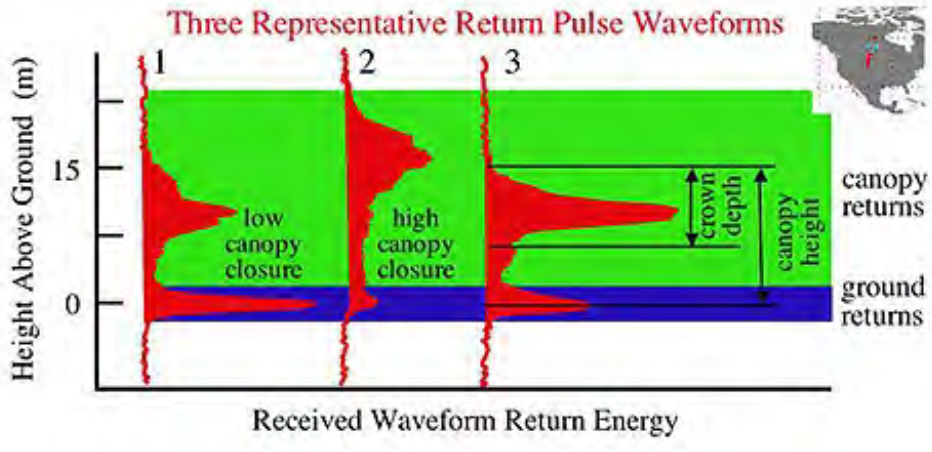
Using ground truth and other information sources, the above image was classified as shown in the left illustration below. Estimates of biomass, using field data and a mathematical model, are shown in the right illustration.



Both Visible-NIR and radar images have met with moderate success so far in estimating biomass in forests, grasslands, and crops. The degree of success is strongly influenced by the quality of ground truth and the specific model used. But as higher resolution imagery and better models are developed, these results should improve significantly.

An important parameter in determining biomass is the "crown height", i.e., the height to the top

of a forest canopy from the ground. ICESat's GLAS laser (see page 14-14) can make such a determination along a profile line as it orbits over non-polar regions. Here is an example of measured heights in an area in the Rocky Mountains:



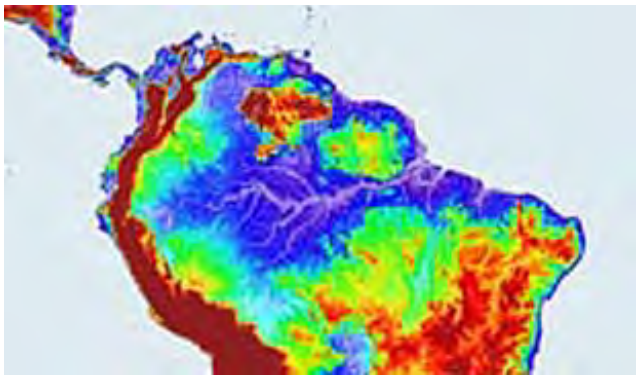
The Amazon Rain Forest; Deforestation

Drained and nourished by the mighty Amazon river and its tributaries, the Amazon rain forest is the largest such tropical jungle in the world. It hosts myriads of plant and animal species unique to the region. Some plants have proved invaluable as sources of chemicals that can be converted into medicines. Until recent decades it was inhabited almost exclusively by native aboriginal peoples. This photo shows a typical stretch of unspoiled jungle canopy:

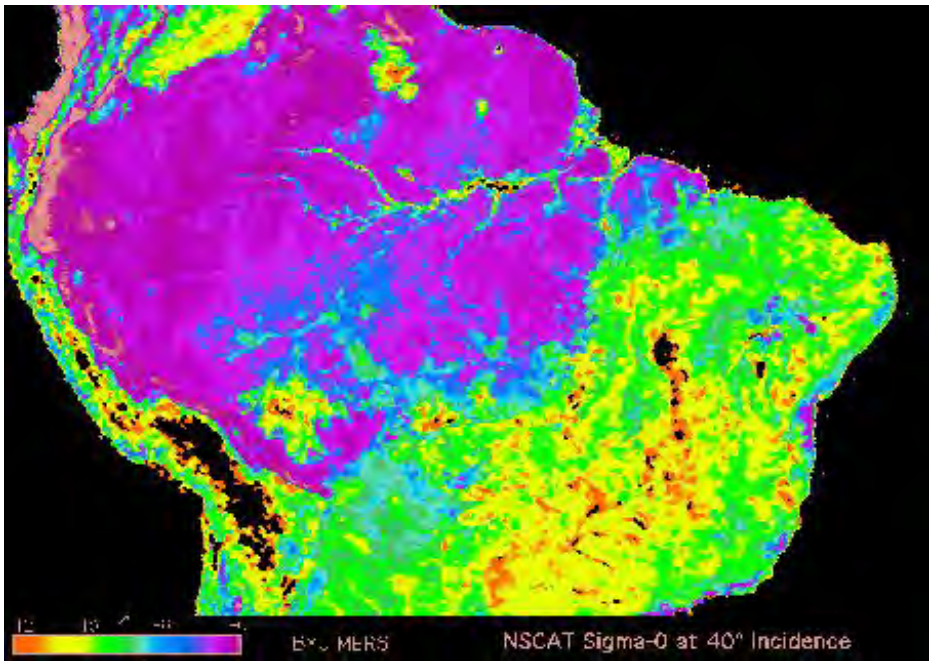
Clearcutting in parts of the world has become a major problem and is the subject of much environmental concern, because of habitat destruction and biochemical threats to the atmosphere. Clearcutting disturbs the balance between the oxygen and carbon dioxide produced and used in photosynthesis. The systematic attack on rain forests in the tropics, to provide open land for farming and other development, is currently the greatest source of alarm. The top concern is for the vast jungles of the Brazilian Amazon, which transition into savannahs at their edges. Over the years, farmers, ranchers, and other entrepreneurs have cleared about 200,000 square miles (about 10%) of the 1.93 million square miles in the Amazon Basin. Fortunately because of an outcry from the rest of the world, the rate of removal is slowing somewhat.

This image map made by the ASAR (radar) on Envisat shows the Amazon drainage basin to be generally low in elevation and flat. The absence of blocking mountains to the east, so that Atlantic marine moisture can move westward unimpeded, is a key factor in the reason for the Amazon basin forests covering most of the landscape. That, plus the tropical conditions near the Equator, which means warm, steady tempertures (no significant cooling since seasonal changes are minimum) and abundant rainfall, have led over the eons to the build-up of thick,

luxurious, species-varied, and persistent vegetation of jungle-like character.

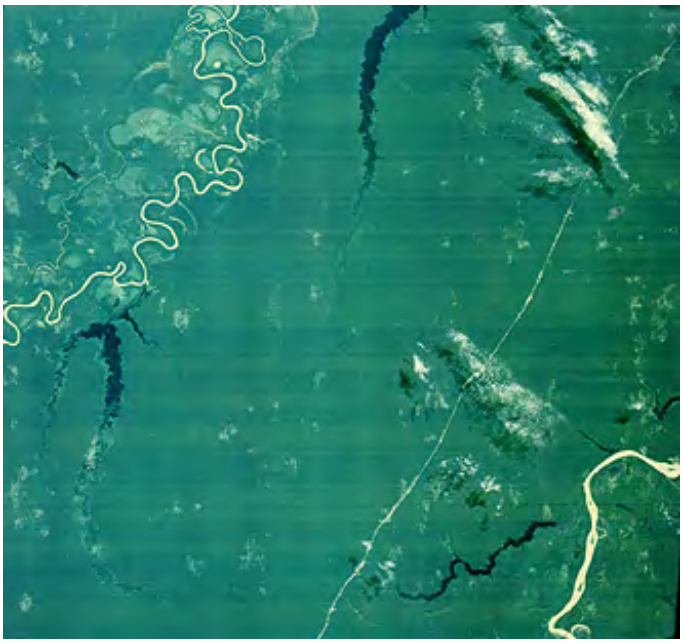


Radar data from the NASA Scatterometer (NSCAT) (see page 14-12) operating over South America shows (below) the distribution of several general, land-cover classes in the Amazon. In this image-based map, rain forest appears in blue/purple, woodlands and savannah in green or yellow, and farmlands or undeveloped mountain lands in black.

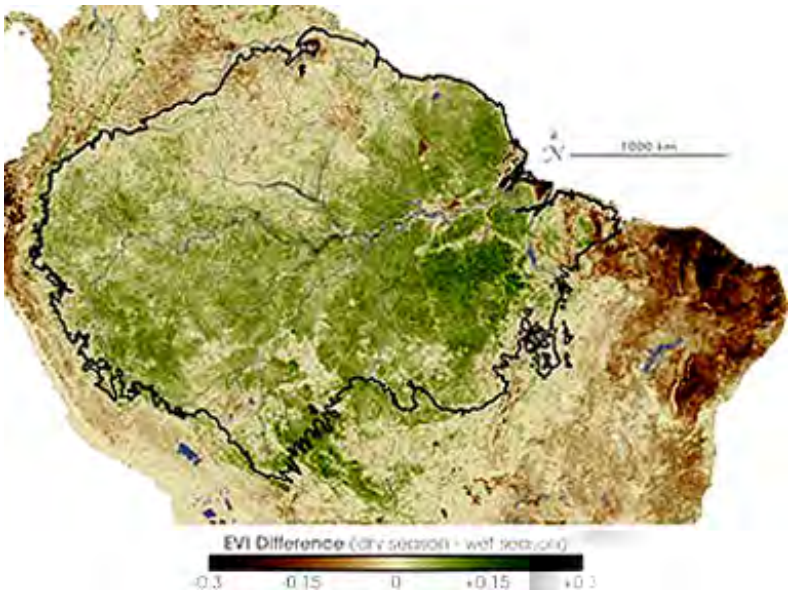


Where the Amazon forest has not been clearcut, it is extremely dense - a true jungle. This Landsat-1 false color demonstrates the uniformity of the forest canopy.

This next Landsat TM image is a natural color composite that further illustrates the uniformly dense tropical forests. These forests are the largest continuous tract of thick woodlands anywhere on Earth - and provide a haven for diverse plant and animal species, now threatened with destruction.



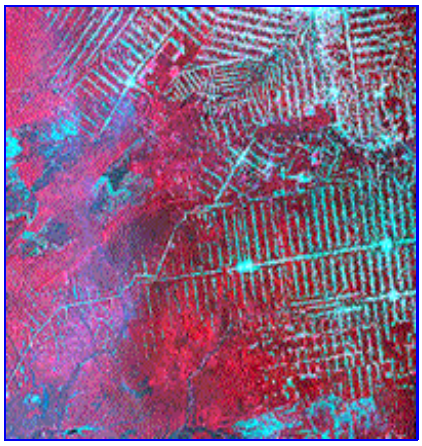
One of the surprises resulting from analysis of multitemporal space-acquired images of the Amazon Basin is that, according to determination of EVI (Enhanced Vegetation Index) (Dry - Wet NearIR images), the Amazon forests are "greener" in the Dry Season than in the rainfall-deluged Wet Season:

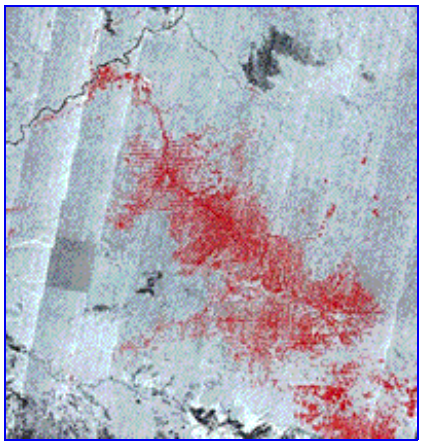


The most plausible explanation for this: During the Dry Season, when increased sunlight means greater photosynthesis, the trees receive adequate moisture from roots within the saturated soils, thus causing more leafy growth.

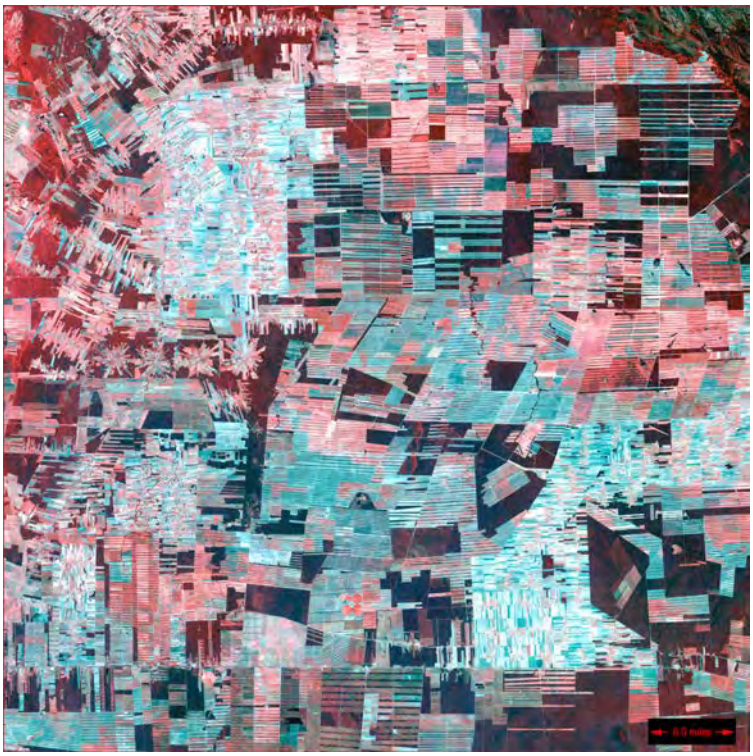
The extensive, dense woodlands of the Amazon have made them targets for systematic cutting down, either to supply wood for building or to clear land to convert it into farms or other use. This image below shows the early stage of deforestation in the Amazon:

One of the most active programs for converting rain forests to cultivated land persists within the State of Rondonia in the southwestern Amazon. In fact, this area is justly cited as the prime example of land stripping. The next three images support this conclusion. The first (left) image is a Landsat TM subscene taken in 1986. The neat rows or strips in blue are cleared land amidst the surviving forest. In the right image is a Synthetic Aperture Radar (SAR) image taken by the Japanese JERS-1, in which areas of clearcutting show in red.

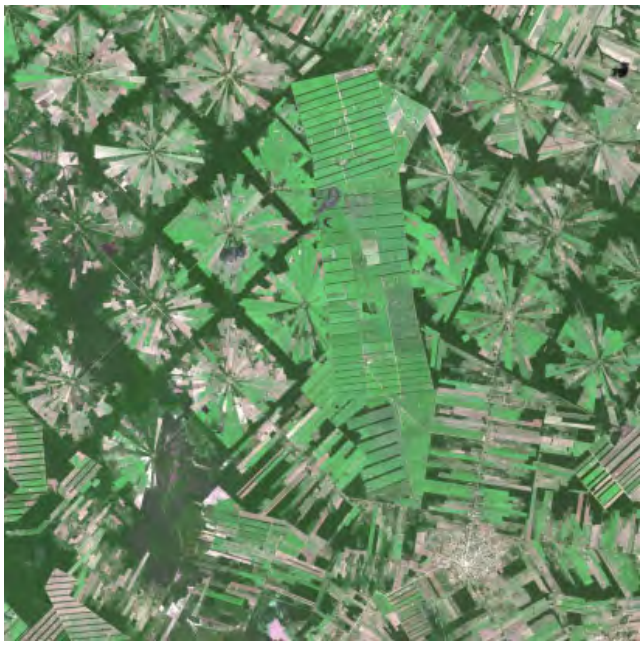




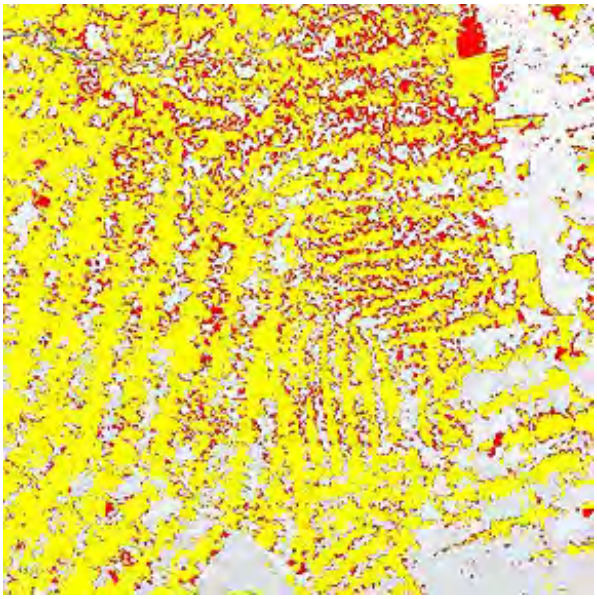
This pattern is among several shapes present in this image of Bolivian deforestation:



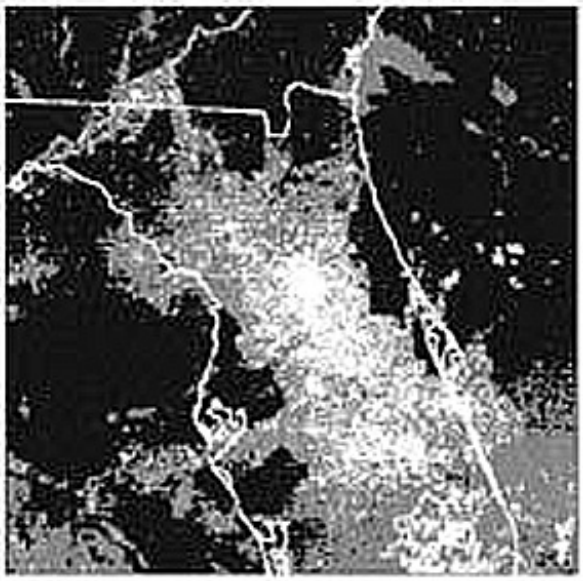
The rectangular forest stripping pattern is common. Another pattern, which could be called stellate, involves deforestation outward from a common center, as shown in the astronaut photo of an area in eastern Bolivia, and then in a SPOT image:



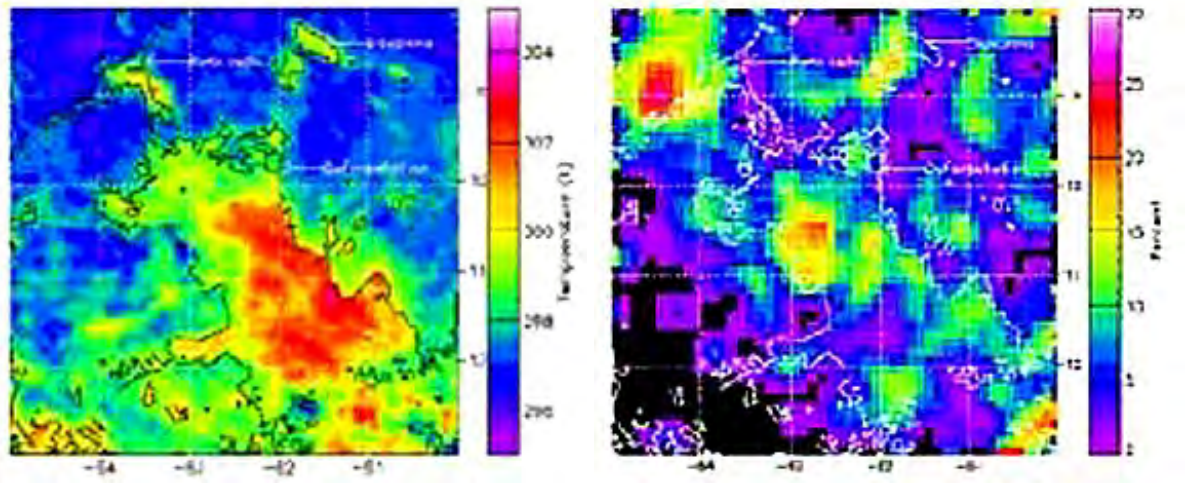
Radar imagery, which can penetrate cloud cover (see page 8-5), is especially effective for short-term monitoring of changes in deforestation. This next image uses two JERS-1 SAR images taken 15 months apart to pinpoint the enlarged edges of cleared strips. Preserved rain forest is shown in white. Deforestation as of July 1992 appears in yellow, and additional stripping up to October 1993 is marked in red. While the changes are not gross, the encroachment on remaining natural forests is obvious



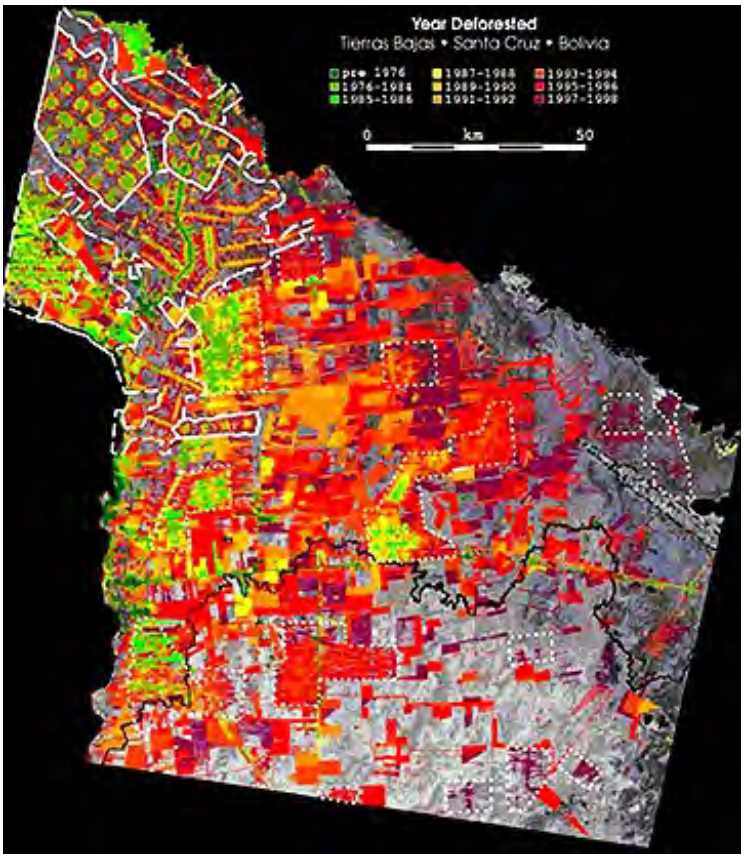
Rondonia has become a sort of "type locality" for typifying and monitoring large-scale deforestation. This GOES image shows just how large the clear-cut area is, as the light-toned deforestation is set out against the outline of the State of Florida to elucidate the scale:



Removal of this much of the rain forest produces an area capable of locally modifying the weather, and overall climate. On the left is a TRMM image indicating a rise in temperature; on the right the TRMM image shows that rainfall has actually increased because the extra heat is causing more evaporation above the scarred landscape:



Over the last 30 years, imaging of the rain forests of South America from various satellites has accumulated a lot of history about deforestation as it progresses. This next image uses color coding to indicate removal and some reforestation in the Tierras Bajas jungle in the Santa Cruz region of eastern Bolivia:

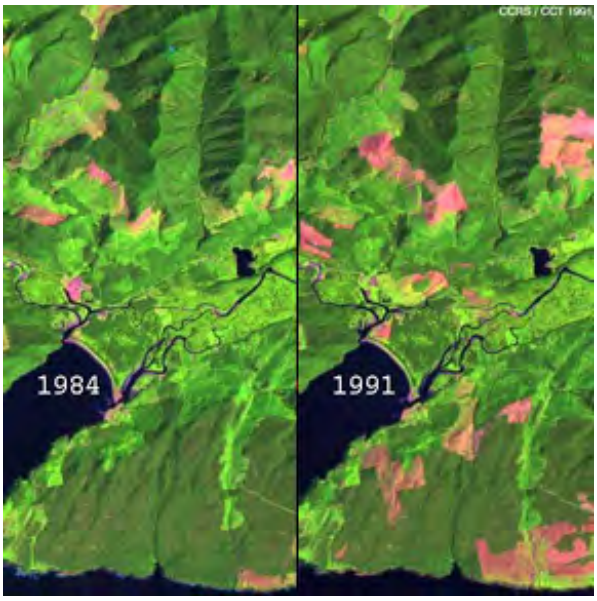


3-13: **Make a guess as to the percentage of South American rainforest that has been cleared or defoliated. ANSWER**

Many environmentalists, especially the "greens", have deplored what appears to be wanton destruction of the South American rain forest. Although the percentage of cleared land remains small (<10%), strong efforts continue to attempt to curb these losses. Yet there are certain mitigating circumstances that should be considered in optiing for blanket restrictions. At the heart of the counterarguments are these facts: 1) population growth will inevitably lead to incursions into the rain forests as food stuffs and other materials are needed; 2) the nature of the soil (e.g., widespread lateritic soils that are poor in nutrients) over much of the region is poor for long-term use; 3) farms established on such soils may be usefully productive for only a few years; 4) when production becomes limited and unprofitable, the custom is to burn off the last crop remnants, abandon the land, and strip or burn off forests to obtain new land for similar short-term use. The bottom line: reworking parts of the rain forests is inevitable; what is needed is a master plan and regulations to restore natural forest growth in the abandoned lands. Conservation laws should be written to restrict deforested areas to at most no more than 20% of the rain forest, by maintaining land use equilibrium through procedures that put back natural conditions through mandatory reforestation (and also including attempts to improve soil quality by appropriate fertilization).

Other Examples of Clearcutting

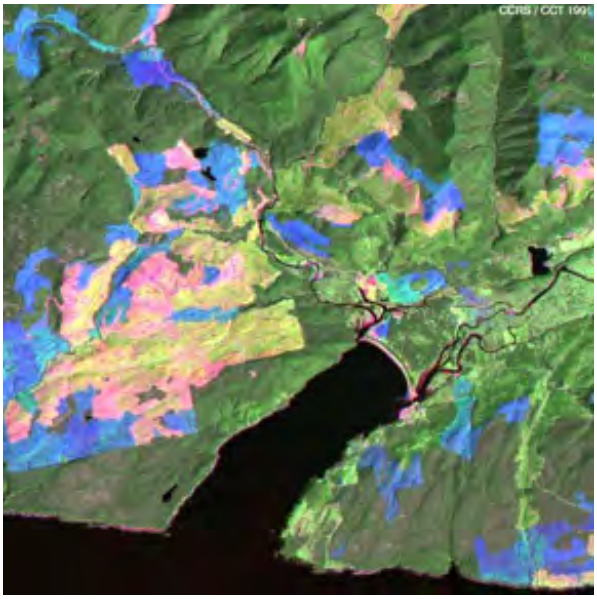
At a smaller scale, but nevertheless subject to continuing debate, is the stripping of whole tracts of woodlands in temperate and boreal forests to provide lumber and paper pulp. Loss of the redwoods in the western U.S. coastal forests is a typical situation. While current laws require reforestation of most of these lands, the rate of regrowth in some regions is not keeping pace with removal. Some of these forestry applications are nicely portrayed with several Landsat subscenes (below) of heavily timbered areas of Canada, as prepared by the Canadian Centre for Remote Sensing. The first two cover an area near Port Renfro on Vancouver Island in British Columbia, in part of the Pacific Coastal Ranges with fir, spruce, and some deciduous trees. In this next subscene are a pair of TM images, each 7 x 15 km (4.3 x 9.3 miles), taken on July 17, 1984 (left) and June 19, 1991.



Each is processed with TM Band 3 = blue, Band 4 = green, and Band 5 = red. This color scheme makes active vegetation green. Clearcutting of the mixed forest is in pinkish-red. Mature vegetation is darker green and regenerated forest appears in light green.

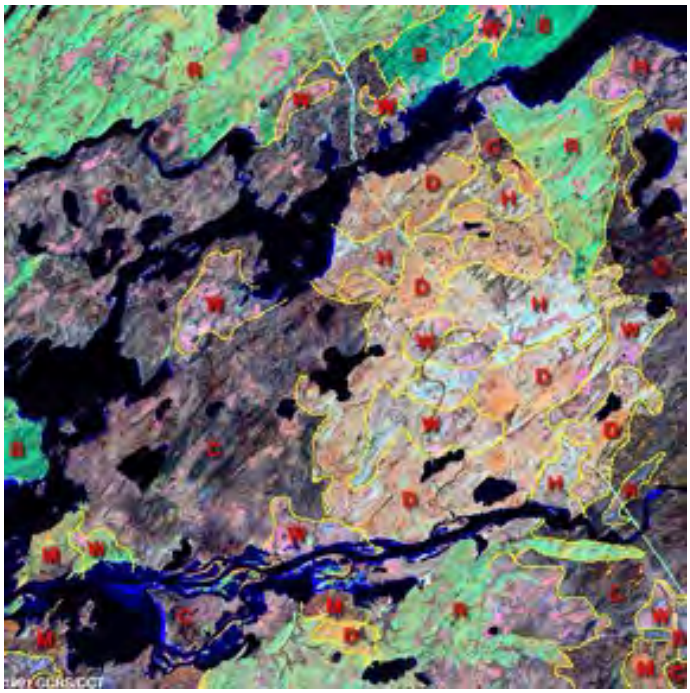
3-14: **Estimate the increase in clearcutting from 1984 to 1991. ANSWERS**

Another way to carry out this change detection is to use two bands from one year and one band from the other year in making the color composite. Note this combination in the following image:



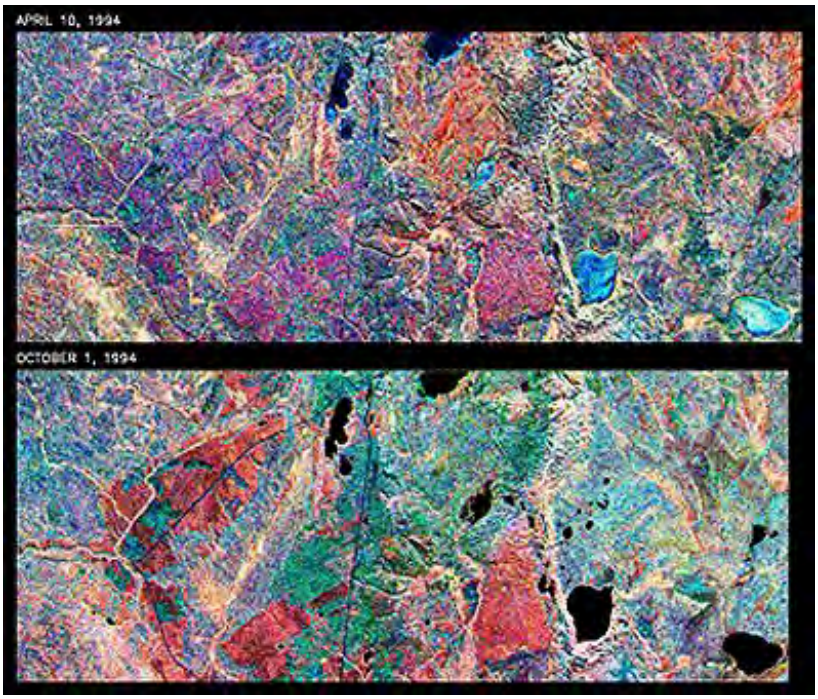
In this 15 x 15 km (9.3 x 9.3 miles) version, TM Bands 4 and 5 from 1984 are assigned to green and red, and Band 5 from 1991 is shown in blue. Once again, mature forest is rendered in dark green and reforested land in light green but here pink denotes areas that were recently deforested in 1984 and blue refers to areas that were clearcut between 1984 and 1991 but have yet to make a strong recovery.

We shift now to part of the boreal forestlands that are widespread on the Canadian Shield. The area (15 x 15 km, 9.3 x 9.3 miles) is in an isolated timber wilderness near Lac Nemiscau in northern Quebec:



In this August 20, 1991 subscene, TM Bands 5, 4, and 3 have been assigned to red, green, and blue, respectively. The Riviere de Rupert and associated water offshoots appear in black. Shades of bright green are areas of recent, partially recovered fire burns (B in red lettering) or regenerated forest (at R). Purple-gray denotes coniferous forests (at C) and orange is deciduous forest (at D). Small areas of wetlands and bogs (at W) show as pink; M refers to mixed forest, and H indicates heath (evergreen shrubs, similar to Scottish heather) with trees. Try to pick out the power line running to the north. This false color rendition is almost as graphic and definitive as a supervised classification, because of the incorporation of the longer wavelength, TM Band 5.

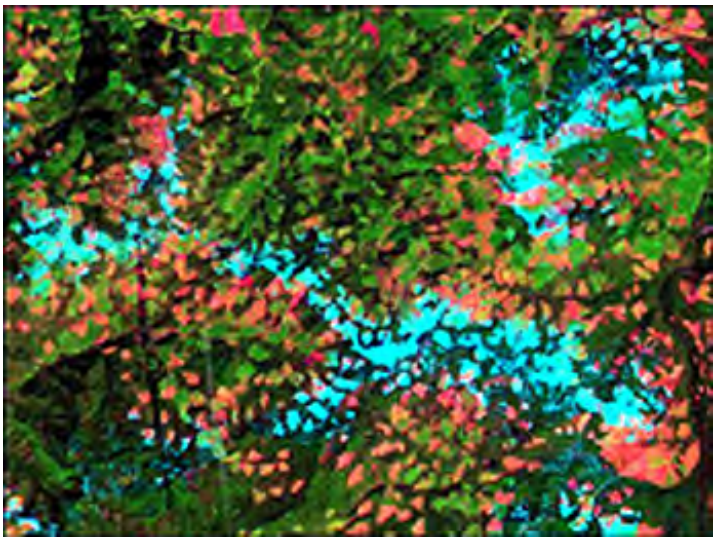
Staying in Canada which has a vast area of boreal forest, let's look at another multitemporal example of change as monitored by the multiple bands on SIR-C radar:



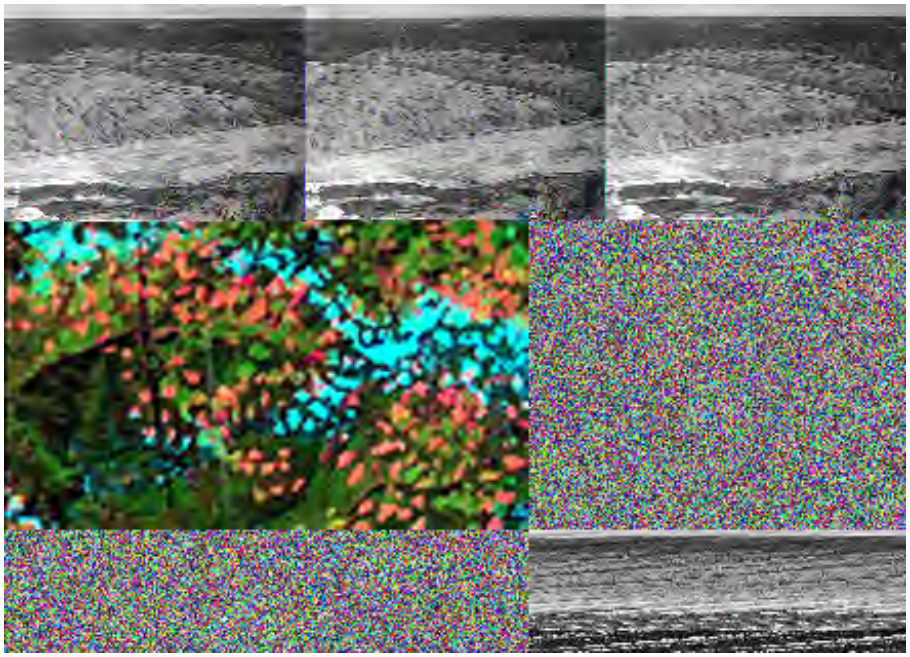
There are notable changes between April and October of the same year. In April Old Jack Pine is purple but red in October; Spruce and Aspen are red and green in April and blue in October. A small clearcut area made during the summer is also blue in the October image.

The Terra satellite is contributing in many ways to better our knowledge of forests and other

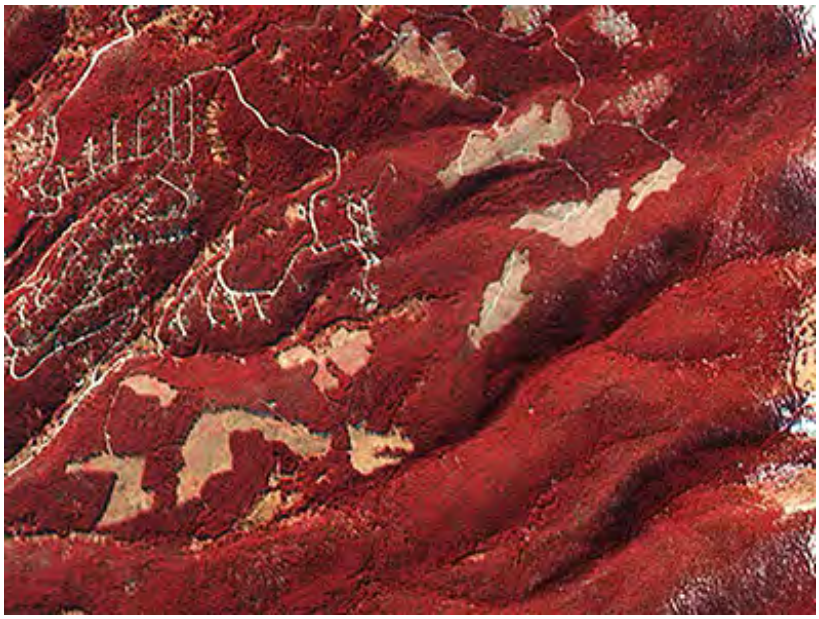
vegetation biomes. It, too, provides striking images of deforestation. Perhaps no other part of the U.S. shows the nature and magnitude of clearcutting than the Cascades (and nearby Coast Ranges) in the Pacific Coast states. Here is an ASTER image that shows recently clearcut forests in red (not the actual color of the denuded land, but chosen to emphasize their widespread location). Replanted earlier clearcuts appear in light green while old forest is in darker green; the blue is residual snow surviving into this May, 2000 subscene.



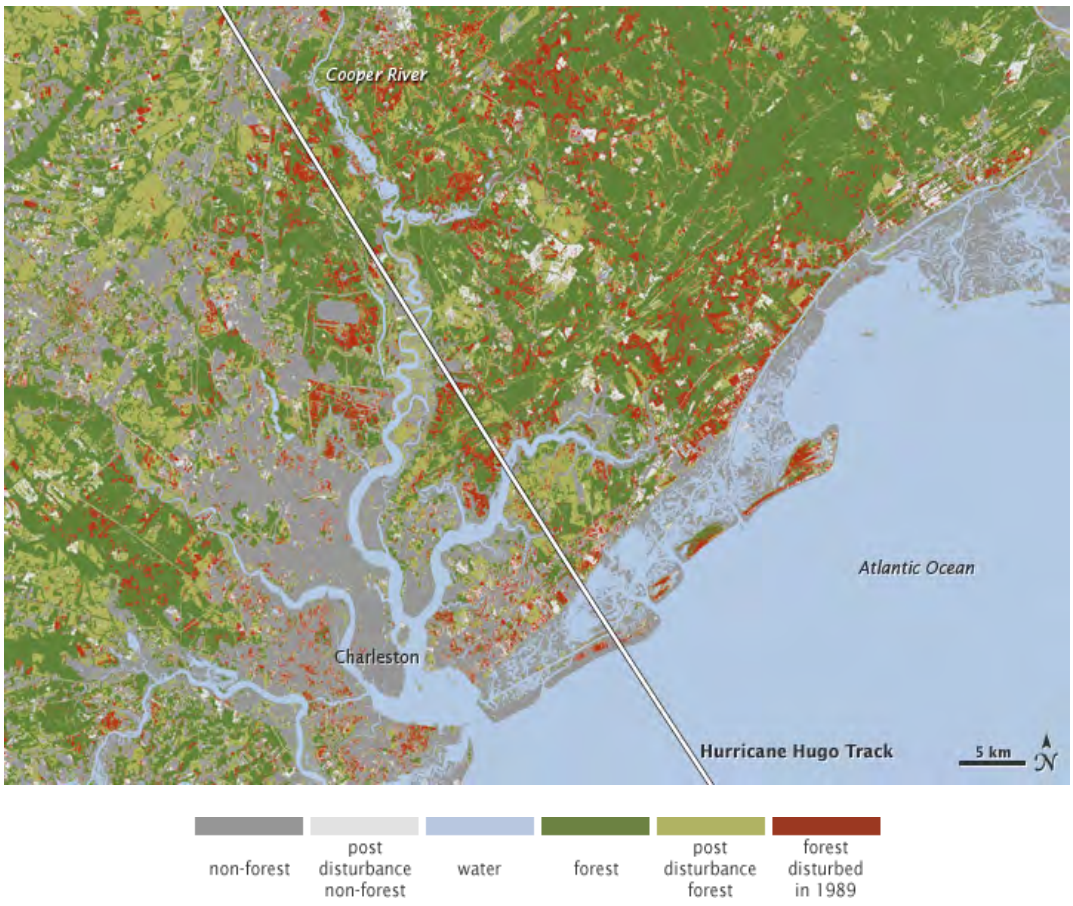
This photo shows a ground scene in the Oregon Coast Range foothills in which a once-heavily forested surface has been largely denuded by clearcutting.



With the advent of high resolution space imagery, details of the clearcuting can now be deciphered for small areas. This IKONOS-2 image is of the Copper Mountain area of the Colorado Rockies. With its 4 m resolution, a texture defined by individual evergreens can be made out. Logging roads and cleared sections are readily defined for easy mapping.



Nature has its own way to accomplish "clearcutting". For example, hurricanes can down trees or strip off foliage on a grand scale. Hurricane Hugo slammed into South Carolina just north of Charleston in 1989. Millions of trees were severely damaged. By comparing Landsat TM imagery from the year before with an image taken soon after the hurricane hit, areas of major damage could be mapped, as shown in red in this illustration:

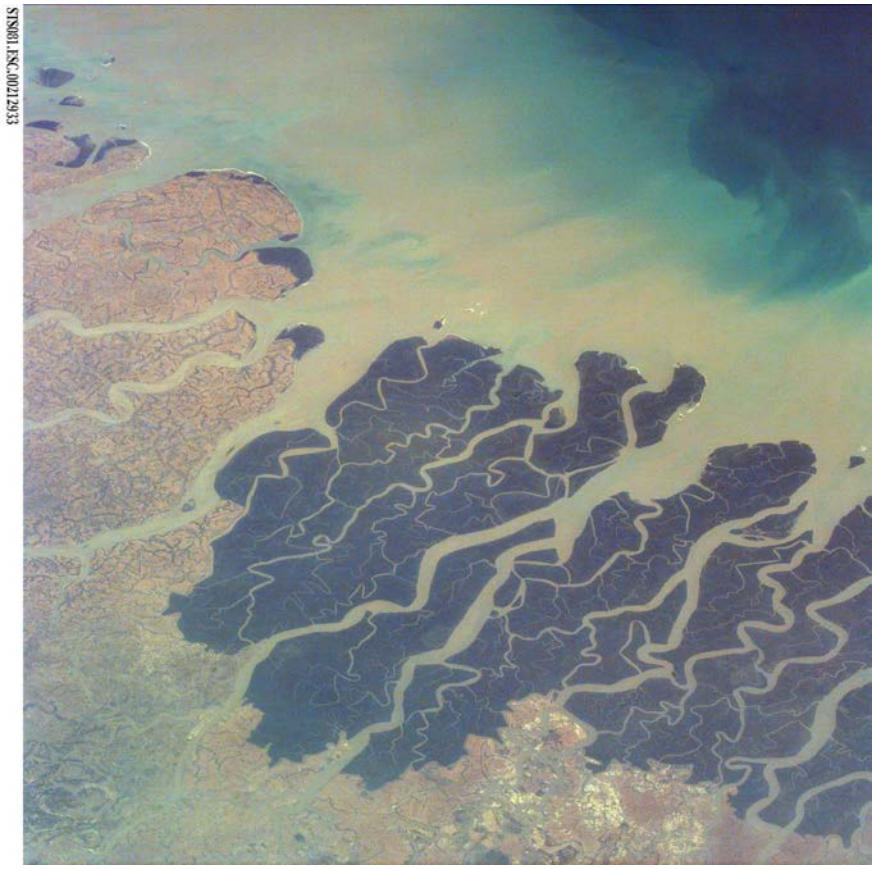


Clearcutting and other modes of ecological damage are placing the world's mangrove forests at ever-worsening risk. Mangroves are situated along marine coastlines. Mangrove vegetation has adaoted to saltwater and other conditions unique to coastal environs. This map shows the distribution of the Earth's mangroves; note that they are more common in tropical regions; beneath is a typical coastal mangrove:

Mangroves appear as dense green tree canopies viewed from the air. Here are mangroves in the Florida Everglades:

Mangroves usually show up as medium deep green in natural color space images:

One of the most expansive mangrove forests is the Sundarbans in the Ganges delta, Bay of Bengal. The green areas are still untouched but the brownish-pink areas have been deforested.



With this overview of forest applications, we can proceed through a case study of the use of space imagery to monitor beetle infestation in evergreens (specifically, pines).

Primary Author: Nicholas M. Short, Sr.

One of the most dreaded of ecological problems - often disasters - is uncontrolled leakage of oil from and around a well drilling site. But for marine offshore drilling it is often much more difficult to cap a well and end the leaking. Catastrophic offshore well failures are rare but have occurred. In the western hemisphere the biggest such failure happened off Mexico in 1979. But in 2010 another failure approaching this in severity took place in deep water in the Gulf of Mexico off the Louisiana shoreline. It had a devastating effect on coastal wetlands and on offshore fisheries. Remote sensing played a major role in monitoring the effects of this oil spill in the Gulf.

Ecological Damage from Natural and Manmade Events

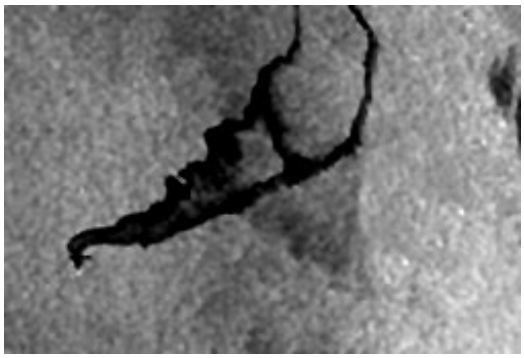
Oil Spills

Accidents involving petroleum and natural gas extraction, transportation, or processing are frequent enough to be of major concern to environmentalists. Oil spills on land, where most oil wells are located, usually can be stopped quickly. Most spillage is confined to the site immediately around the wellhead. The area may appear "messy" for awhile but the damaged environment can eventually be cleaned up.

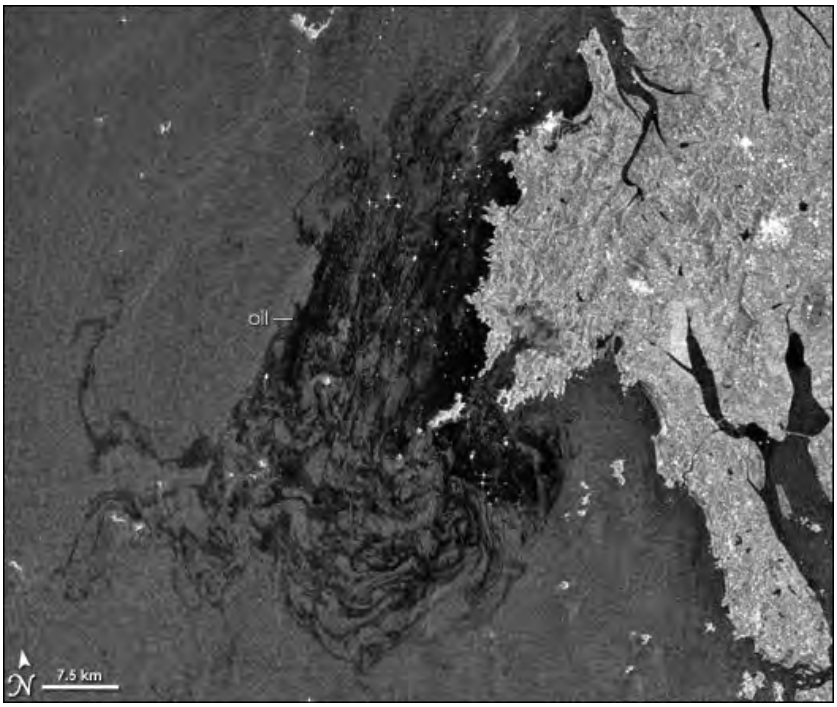
A much more dreaded oil spill is one that contaminates the ocean from a failure at an offshore drilling site, from tanker accidents, or from submarine pipelines. Such events are not uncommon and some can be catastrophic. Fortunately, such spills can be easily monitored from ships, the air, and from orbiting spacecraft. For ocean spills, in particular, remote sensing data can provide information on the rate and direction of oil movement through multi-temporal imaging, and input to drift prediction modelling and may facilitate in targeting clean-up and control efforts. Remote sensing devices used include the use of infrared video and photography from airborne platforms, thermal infrared imaging, airborne laser fluourosensors, airborne and space-borne optical sensors, as well as airborne and spaceborne SAR. Oil is noted for "calming the waters", i.e., reducing the degree of wave disturbance, making it visible in certain spectral bands. Multiple oil leaks in the Arabian Sea, west of Bombay, India, are obvious in the SIR-C radar image below. Their darkness is not solely due to black oil color involving light absorption but also to the decreased backscatter of the radar beam (see Section 8 for the principles).

Radar imagery, especially in the black and white mode, is very effective at locating and monitoring oil spills over their full extents. This next image, made by the radar unit on Envisat, shows how radar graphically highlights the large oil spill resulting from the breakup of the oil tanker Prestige in late November, 2002. Most of its cargo of 20 million barrels went down with

the ship as it sank into deep waters. But 1.5 million barrels of oil eventually reached the northwest corner of Spain, contaminating almost 300 km (200 miles) of beach. Here is a radar image taken soon after the spill leaked out.

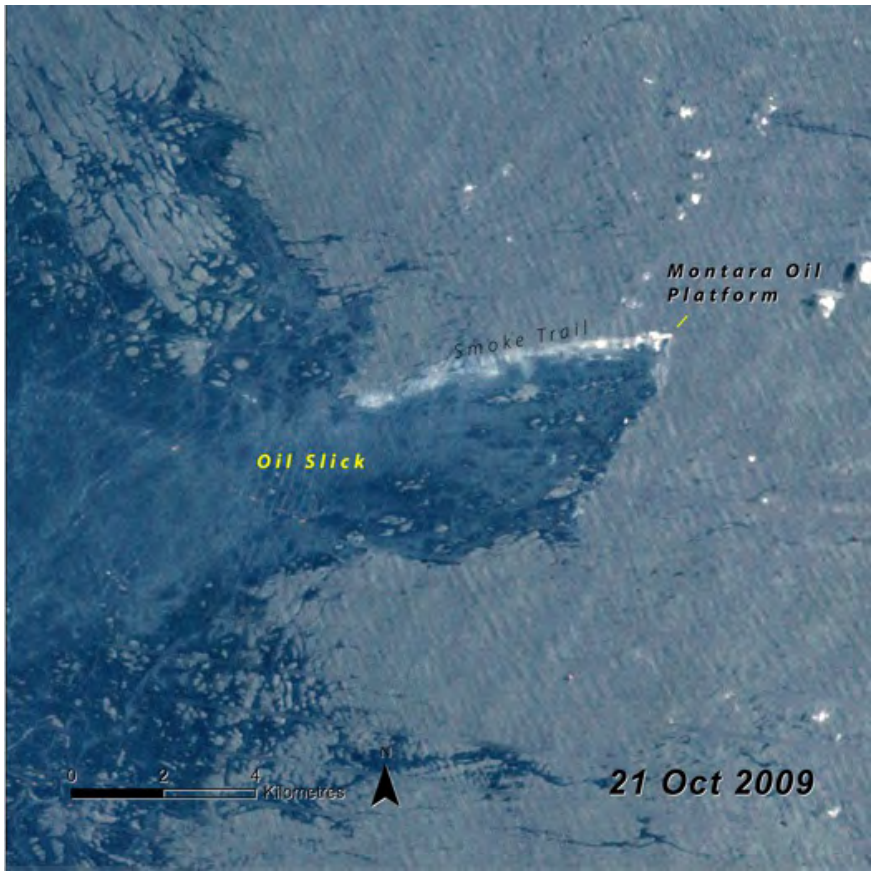


This next image, of an oil spill from a tanker off the west coast of the Korean peninsula, was made by the ASAR instrument on ESA's Envisat:

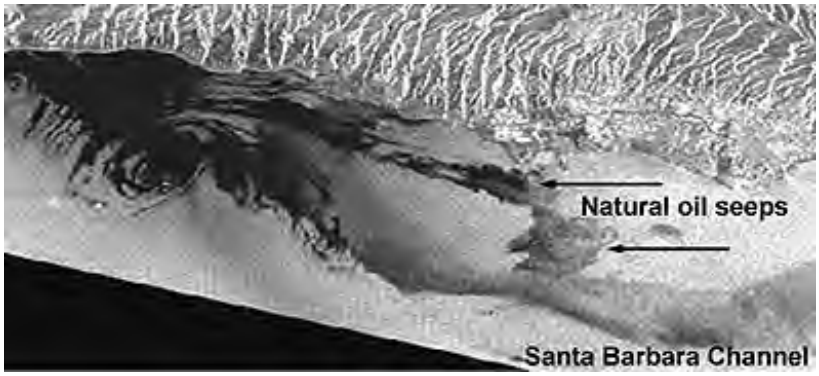


3-16: **Can oil spills be detected in Landsat imagery?** **ANSWER**

Surprisingly, Landsat is not the prime satellite used in monitoring oil spills. Images made by Landsat don't usually show up the oil as "contrasty" against a water background. Here is one example that requires some effort to recognize the oil-affected surface; it is a spill from a drilling platform in the Sea of Timor:



Oil can also reach marine surfaces from so-called "oil seeps", which are natural effusions of oil from fractures underlying the ocean floor. Off Santa Barbara (California) is a persistent natural "spill" shown here as an ERS-1 radar image:



Oil is also commonly added to the sea surface as an "acceptable" steady leak from ships that use oil instead of coal as propulsive fuel. This radar image shows oil trails from two ships off the west coast of Italy:

Ultraviolet (UV) radiation is an effective tool for monitoring oil on water because petroleum responds by fluorescing. Various companies now offer oil spill monitoring services. One is Optimare of Denmark. Its website lists these applications of its Imaging Airborne Laser Fluorescence Scanner (IALFS) instrument:

* Detection of laser-induced fluorescence of crude oils, petroleum products and water constituents.

* Classification and mapping of crude oils, petroleum products and chemicals spilled at sea.

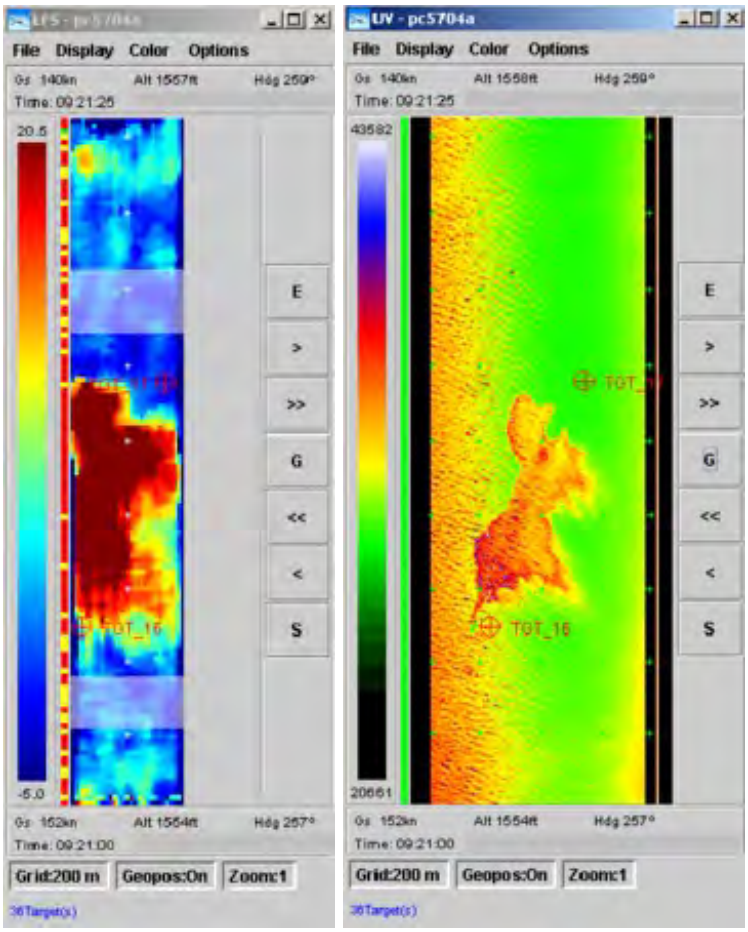
* Detection of crude oils, petroleum products and chemicals floating underneath the water surface.

* Measurements of oil film thickness over very thin (optically thin) oil layers.

* Distinction of naturally occurring biogenic slicks from oil spills.

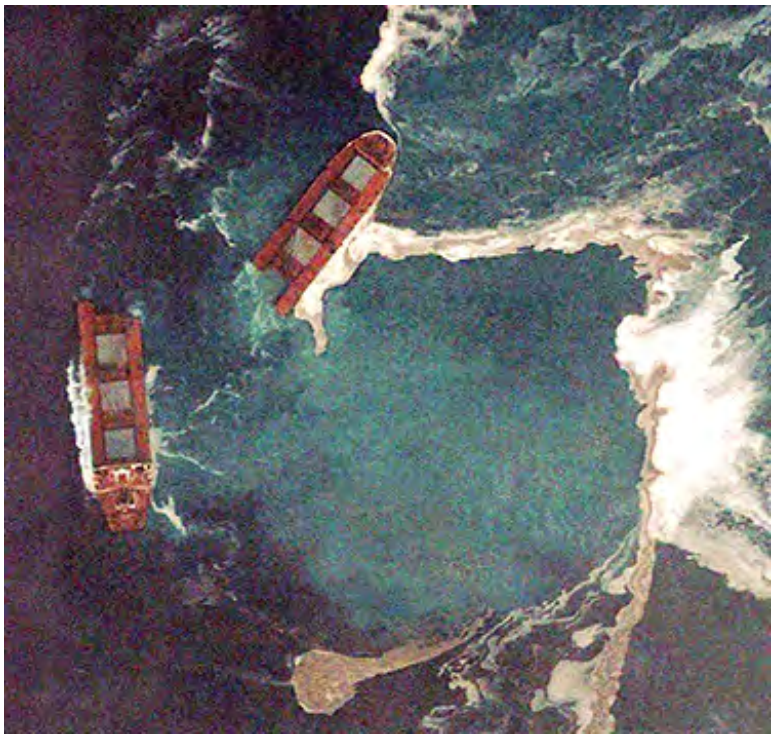
* Hydrographic measurements (CDOM, turbidity, chlorophyll-a).

Here are two images of an oil spill taken first by Optimare's IALFS and then by their IR/UV line scanner instruments:

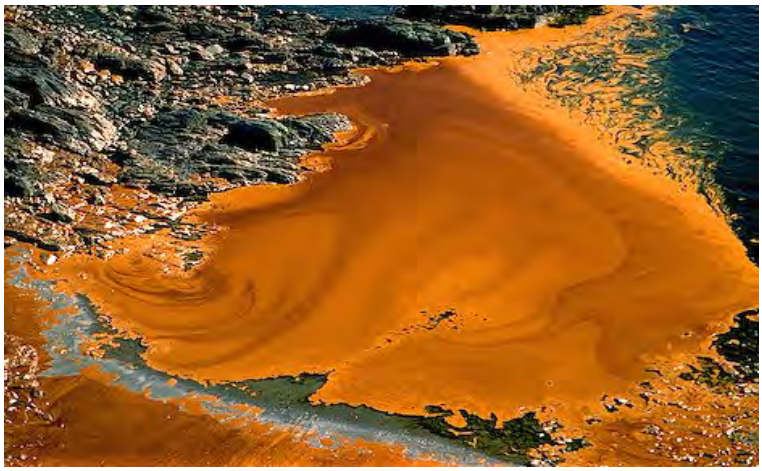


Infrared detection of oil slicks depends on the fact that oil appears cooler than surrounding water most of the time. Here is a mid-gray patch of oil in the river passing by a refinery in Baytown, TX near Houston:

With the advent of high resolution satellites, much more information is now available over short turnaround times above major oil spills. One occured in the Aleutians off Unalaska in mid-December, 2004. A tanker split in half, leaking 10s of thousands of gallons of crude into offshore waters. The boat wreakage and oil-coated waters were imaged by Quickbird:

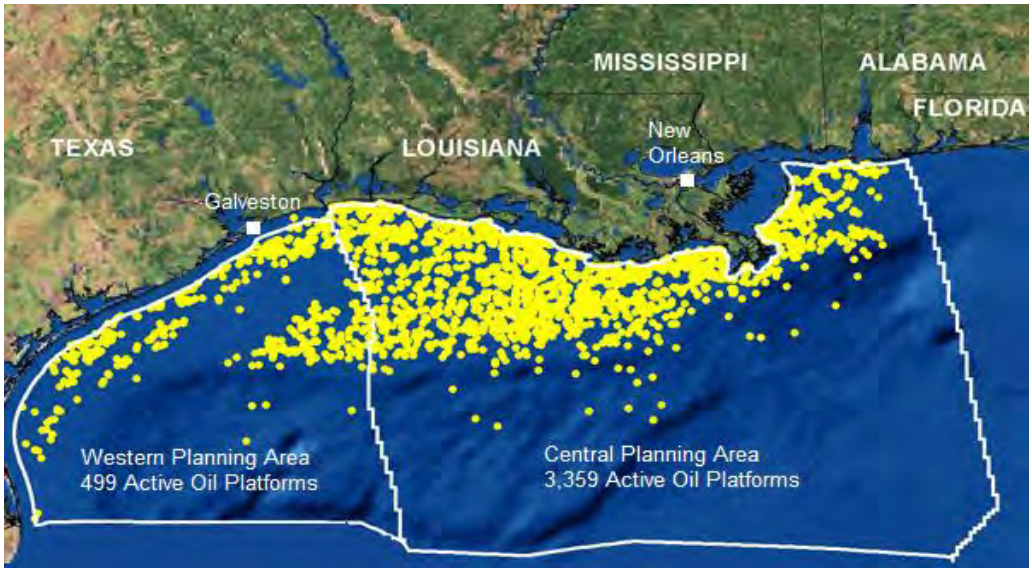


Until the BP oil spill of 2010 (see below), the most infamous U.S. oil spill was in 1989 off Prince William Sound on the Alaskan coast. The tanker Exxon Valdez ran aground against a reef and split open. Almost 11 of its 55 million gallons of oil escaped to open waters. Its story is told pictorially in these two photos:

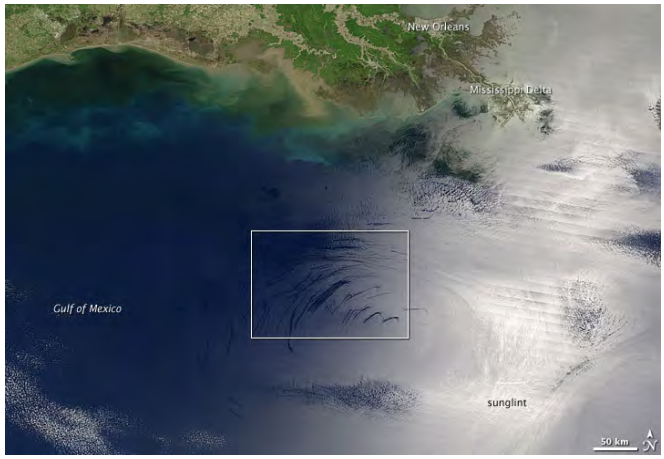


Oddly, this Alaskan oil spill was not well documented by satellite imagery. Most tracking of the migrating oil was done from aircraft and ships.

As seen from the above discussion, oil and other petroleum products can be accidentally introduced into the ocean in several ways. One source is at or below drilling platforms. For the U.S., the most numerous of these is in the western half of the Gulf of Mexico, as shown by this map; as would be expected, relatively few of these are in deep water far out into the Gulf:



At any given time, there is also oil reaching the Gulf surface simply from natural seeps, as shown in these two images:



However, the ever present fear is some kind of catastrophe failure at the drilling platform or on the sea bottom at the well head. What has become the worst oil spill in American history occurred about 50 miles offshore from Louisiana's Mississippi River delta. A drilling platform - known as the Deepwater Horizon. operated by British Petroleum (BP) exploded from a gas leak on April 20, 2010 and sank, killing 11 workers. The causes of the catastrophe are technical, but involve ignition of escaping methane gas; bad safety practices are also a major factor.

Here is the platform as it was towed to its final site.

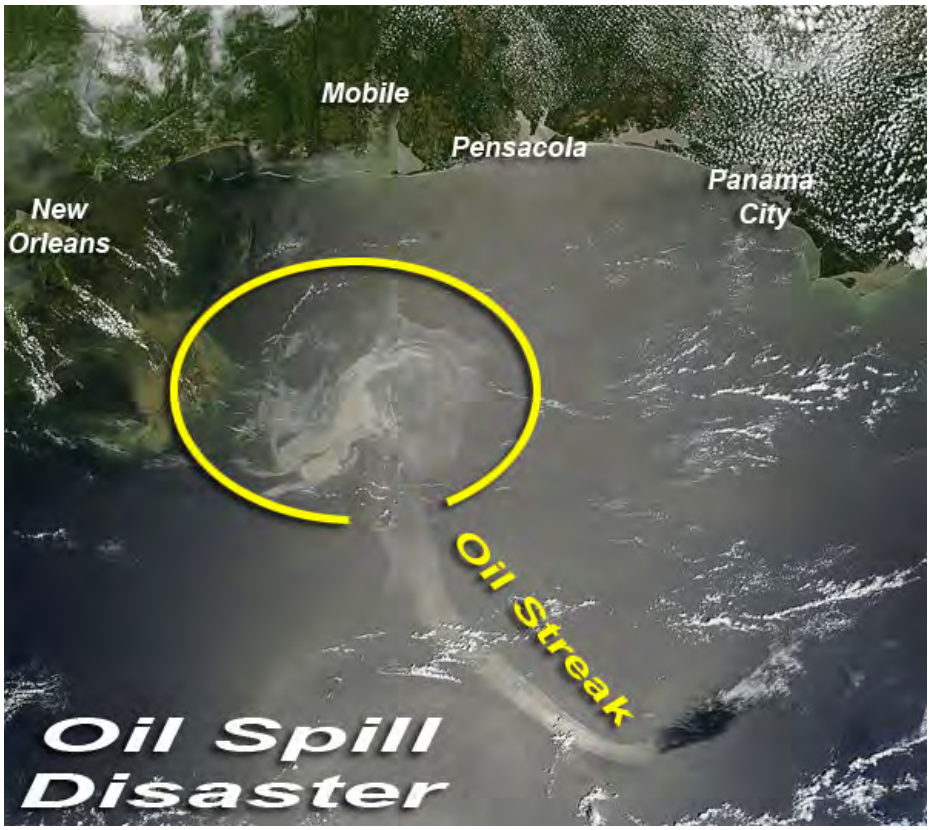
Once at the site, the platform floats freely and its position is maintained using motors called thrusters (fixed propellers). Deepwater Horizon was built in Korea and is operated by Transocean. It can drill in up to 2400 meters (about 8000 feet) of water. At the time of the accident a crew of 115 persons were aboard. Here is a photo of the emplaced platform; below that is a view of the intense fire consuming the platform; it collapsed 36 hours after the initial explosion::

At the time of this event, the drilling, which starts at the ocean floor some 1500 meters (about 5000 ft), had reached a rich reservoir of oil at a depth exceeding 4000 meters (13000 ft). The oil was mixed with natural gas and under high pressure. It flowed readily to the wellhead, thence through pipes to the surface platform. But conditions existed that would lead to failure, full-scale leakage, and an environmental disaster the magnitude of which the U.S. had never experienced before.

Apparently, the Deepwater Horizon well was inadequately sealed, and natural gas built up inside it. This drove both drilling mud and oil itself out onto the ocean floor, initiating the leak. When workers on the rig tried to activate the well's blowout preventer (BOP), it failed. An attempt to activate the blowout preventer after the fact, using undersea robots, also proved unsuccessful. The oil, driven by the embodied gas, leaked from the well head some 1600 meters below the surface. Hundreds of thousands of barrels of oil came the surface to form a slick that headed towards the Gulf coast. Here are three images of that slick taken a few days into the disaster:

Satellites tracked this oil spill for months. In fact, this is a disaster in which satellite observations (and remote sensing in general) played a primary, pivotal role in monitoring the spill over time and its consequences for the environment. The ability of many satellites with wide fields of view to detect and encompass the entire spill in single images showed clearly their advantage over aerial photography.

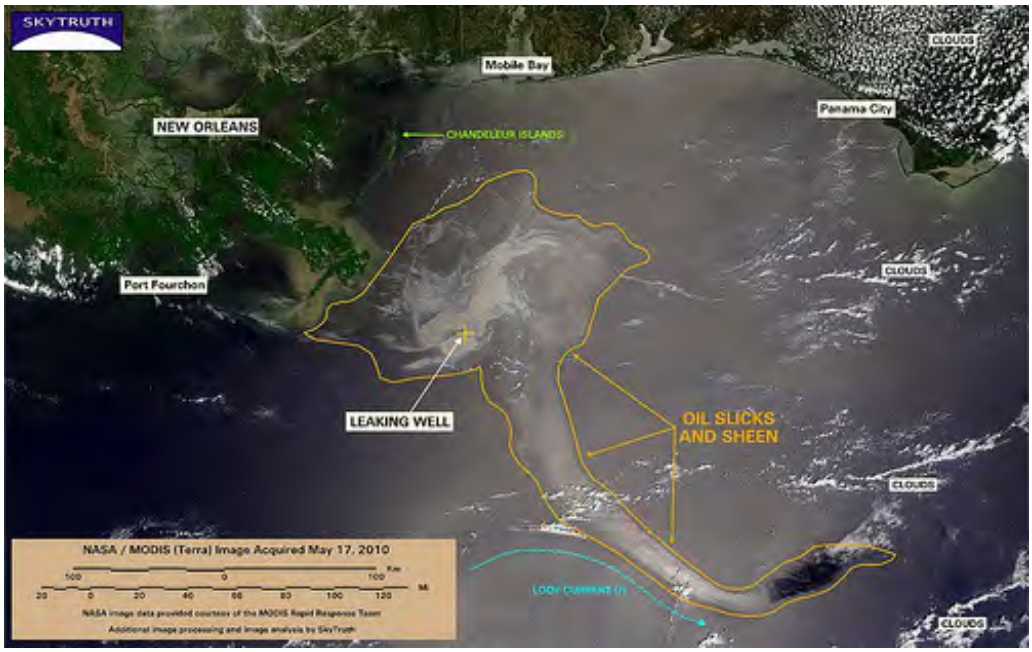
For a while, winds kept most of the oil from reaching the Gulf states shores until mid-May. But, a month after the leak began, it still had not been shut off. Tens of thousands of barrels have escaped into various levels of the Gulf waters. One fear is that the Loop current, which begins in the Gulf, swings around the tip of Florida, and then joins the Gulf Stream up the Atlantic coast, will carry the oil beyond the Gulf - greatly extending the damage both to the fishing industry and to tourism. That this may happen is evident from this MODIS image made on May 17th, in which a streamer of oil is moving southward towards the Loop current:



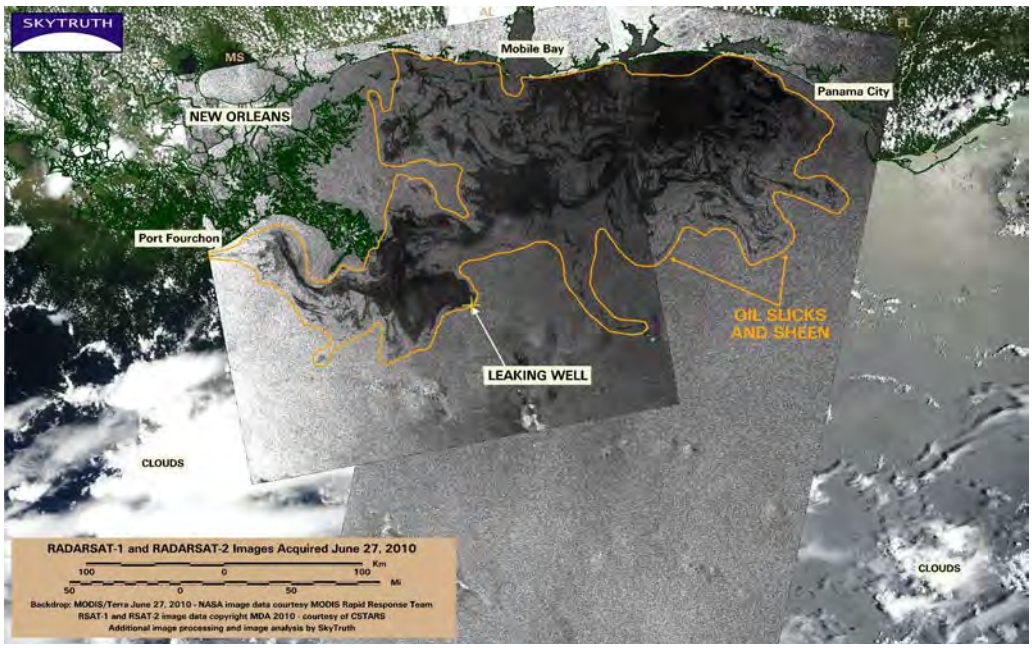
Satellites have followed the growing spill (as of June 1st, about the same areal extent as the state of Connecticut; it grew even larger, exceeding the size of Florida). A series of maps that track the spill's location, size, and shape can be accessed at this CNN website, provided it is still active. While the size keeps enlarging, the shape and coherence changes constantly, even as its location shifts owing to guiding currents and wind effects. Here is an image obtained nearly five weeks after the initial failure; part of the spill appears bright because of sunglint:

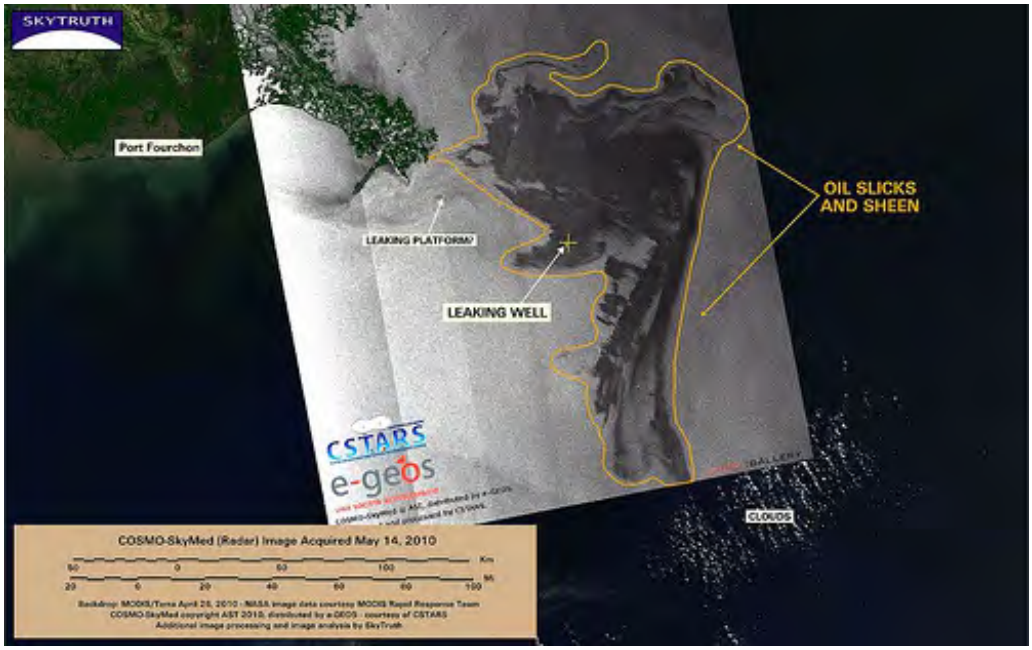
The next series of images indicate the diversity of views obtained from different satellites. The first pair shows that, as seems logical, similar sensors produce nearly identical images; a MODIS view is on top of a view made by Envisat's MERIS:

The next pair compares the scene on May 18 as captured by the ASAR radar on Envisat and the MODIS optical instrument on Terra:

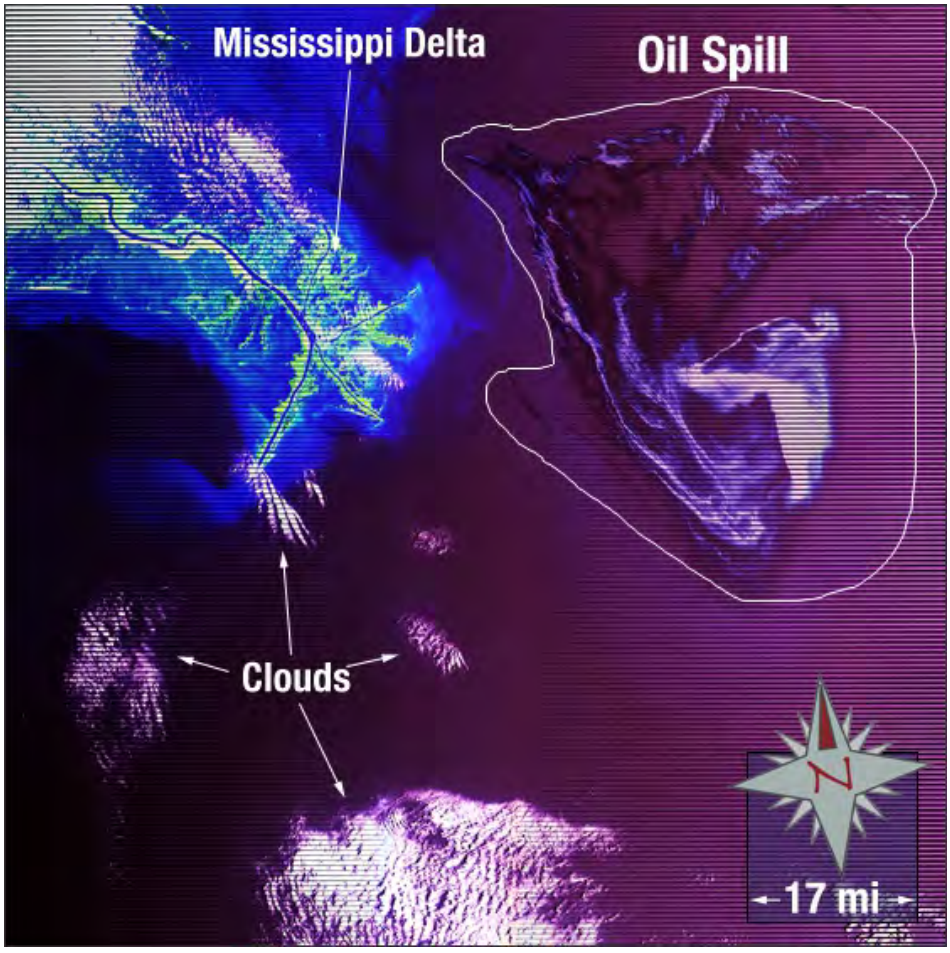


Radar is especially effective at imaging oil on water. Here are a Radarsat and a COSMO Sky Med (Italian satellites) image of the Gulf spill:

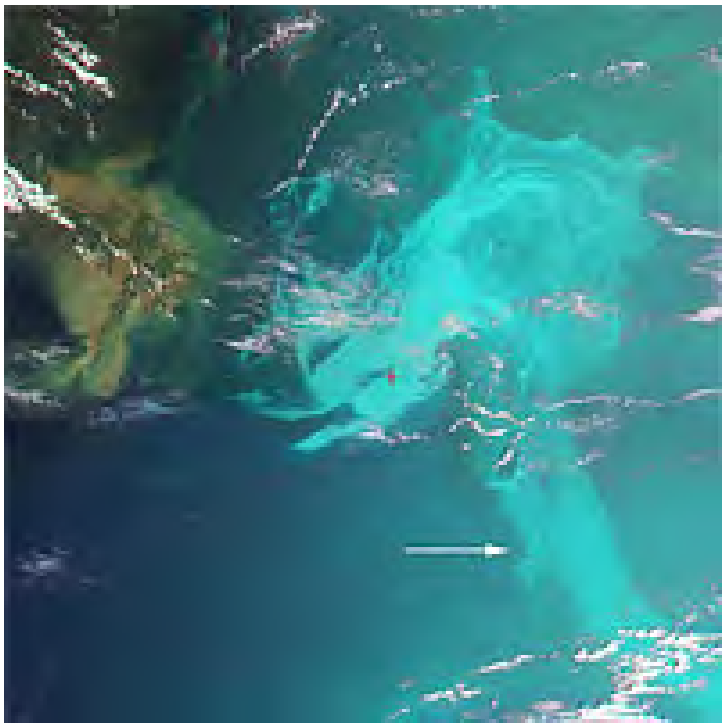




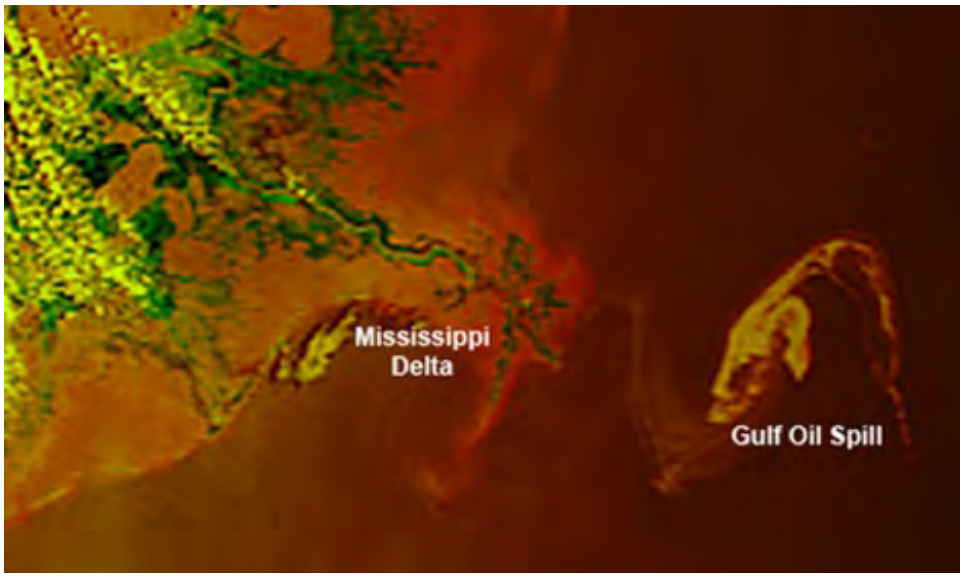
The "old reliable" Landsat-7 took this image of the spill:



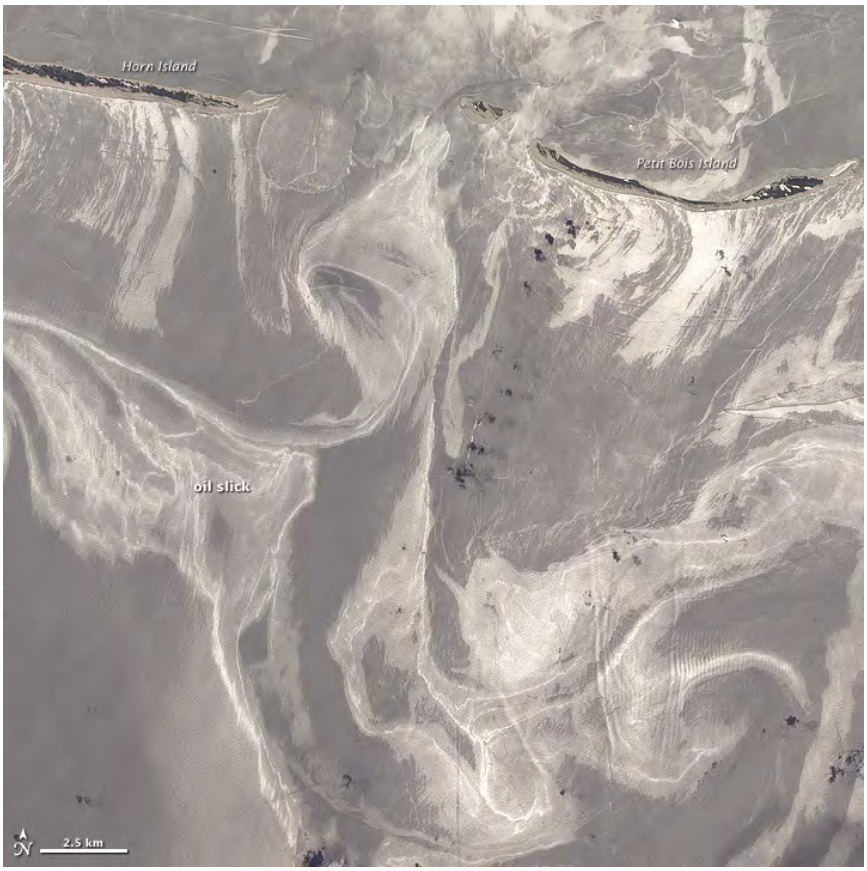
Specialized sensors can enhance the appearance of the oil spill, making it easier to establish boundaries. The MISR (Multi-angle Imaging Spectro-Radiometer) on the Terra spacecraft has produced images like this:



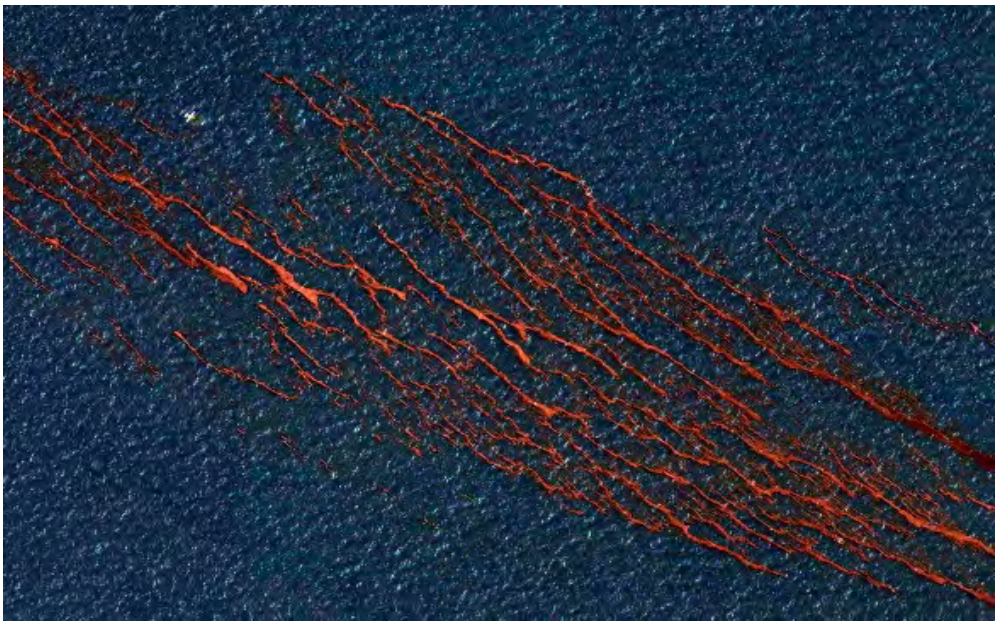
Because of its size, meteorological satellites can easily detect the spill as evidenced in this Eumetsat image. Their wide field of view (FOV) provides a unified look at the entire extent of the spill.



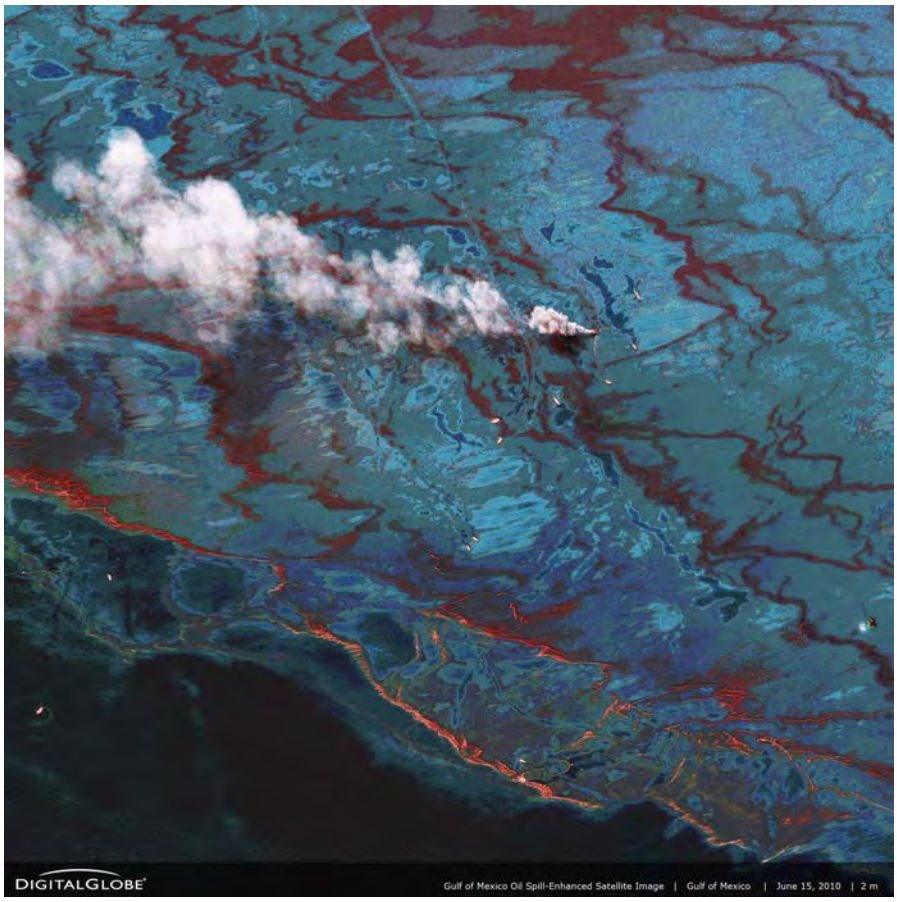
But, smaller FOVs are useful in monitoring parts of the spill that are at or nearing shorelines. This EO-1 ALI image of the coast off the state of Mississippi is a good example:



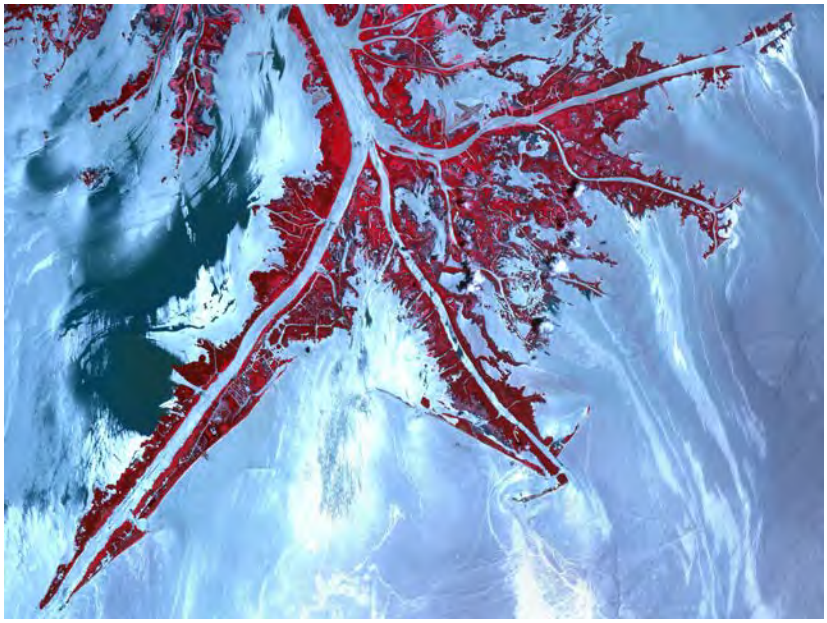
The high resolution satellites such as IKONOS and Quickbird can see small areas of the spill in detail; this is the case in this Quickbird image:



Digital Globe's WorldView-2 satellite obtained this view:



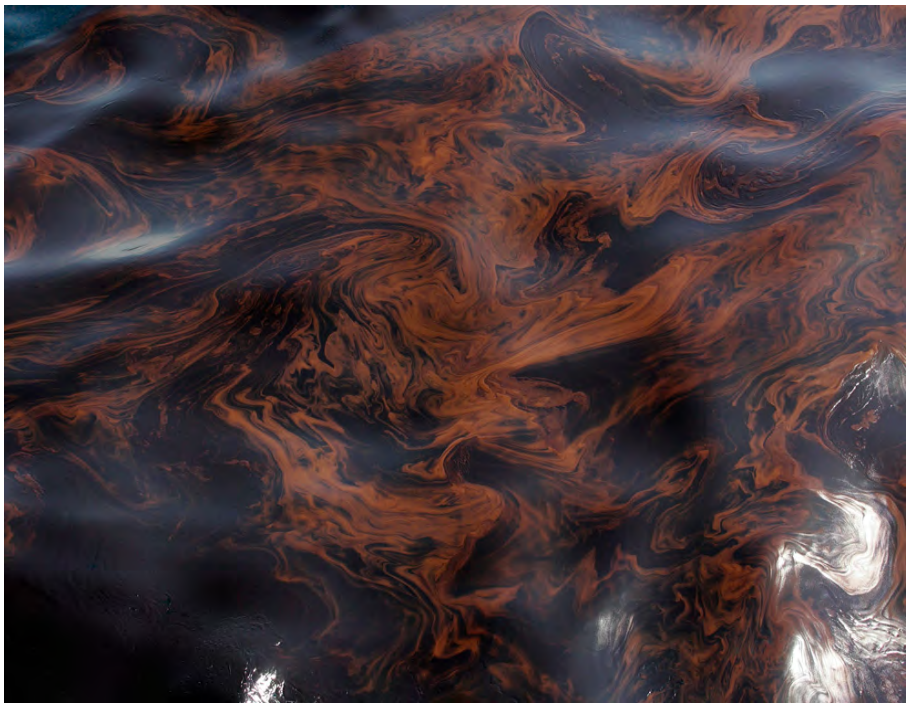
In between these scales, Terra's ASTER produces images such as this, which shows the oil (rendered here as silver streaks) moving onto the birdsfoot distributaries of the Mississippi River delta:

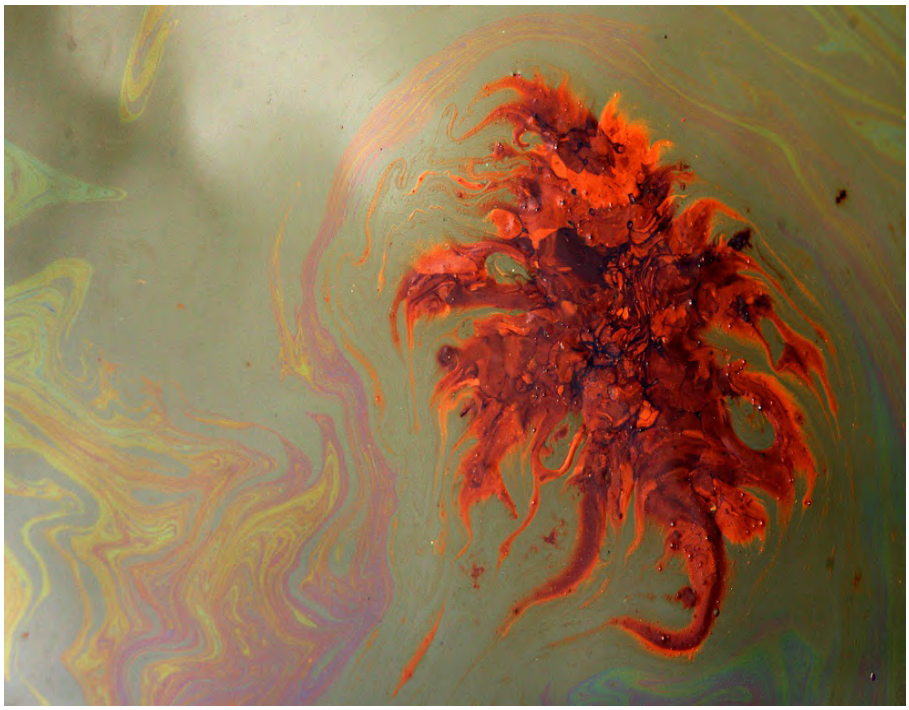


The land in the distributaries is just above sealevel - low and flat - and thus is easily breached by high waves during a storm. This aerial photo shows that condition:

Humans in space have joined in the observations. An astronaut photo taken from the International Space Station gives a different impression of the spill since it is oblique (visualized on a slant):

Photos taken from airplanes and helicopters allow specific problem areas to be checked out. A small part of the spill is seen in these aerial photos. Note its distinct brownish-red color.

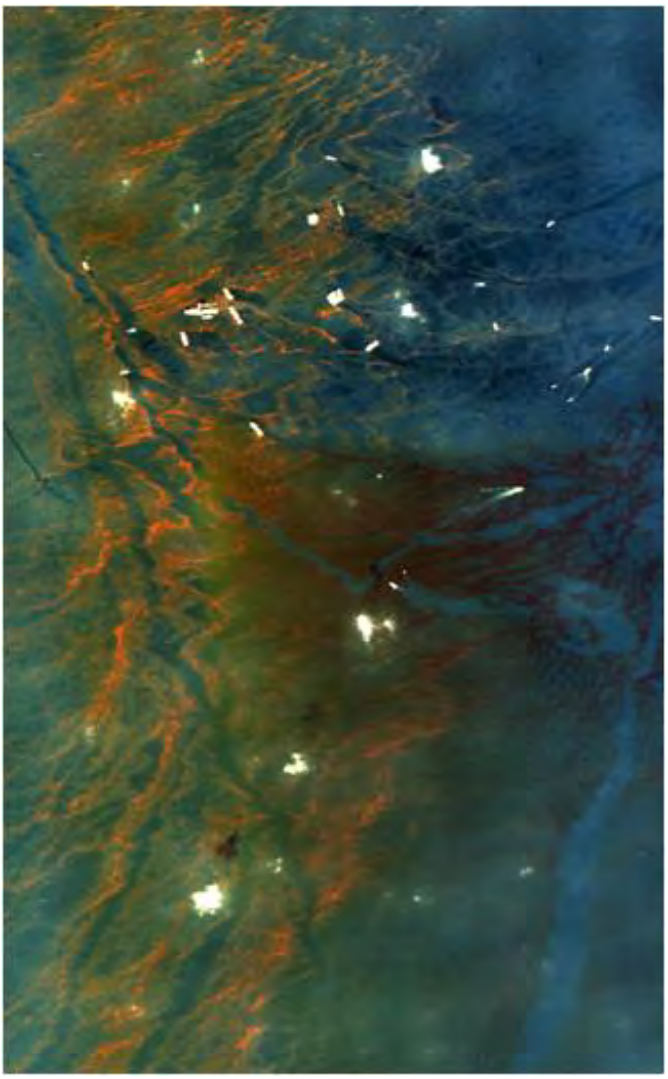




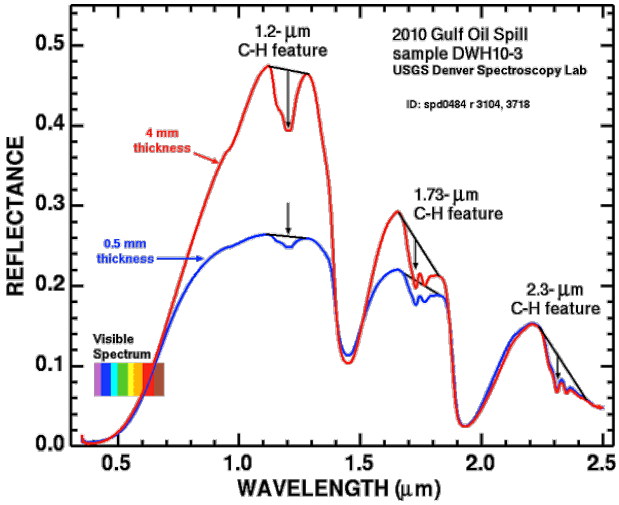
NASA has conducted its own aerial flights over the oil spill, designed to obtain detailed spectral measurements of the properties of the oil as it rests on the Gulf waters. The plane involved is ER-2, a modified U-2 (the famed spy plane used in the Cold War era):

The instrument that gathered the data is AVIRIS (Airborne Visible/Infrared Imaging Spectrometer), described on page 13-9, which is a hyperspectral imaging system as well as one

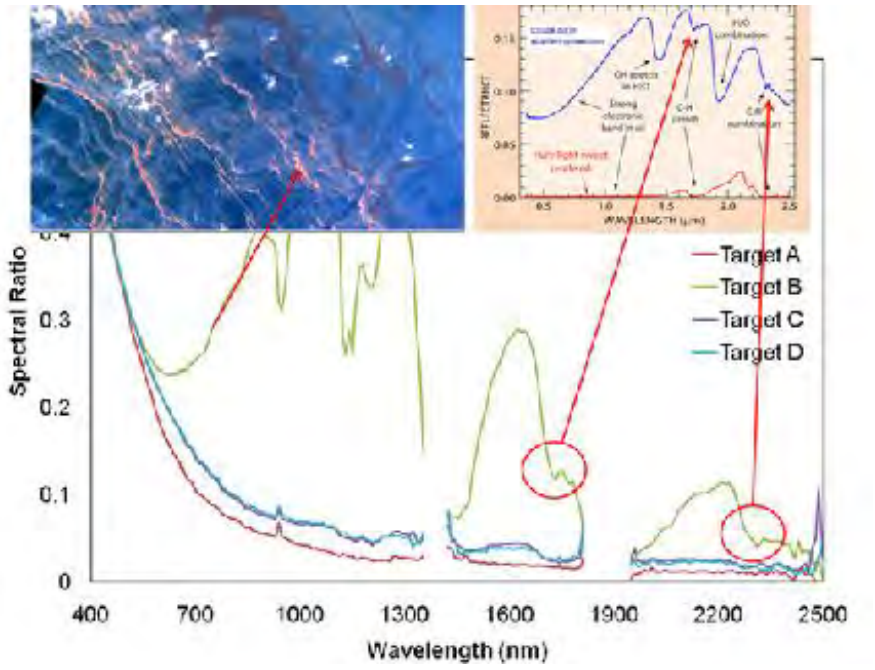
capable of producing a continuous spectral curve. AVIRIS can produce both natural and false color images such as this true color one of the oil streaks within the spill.



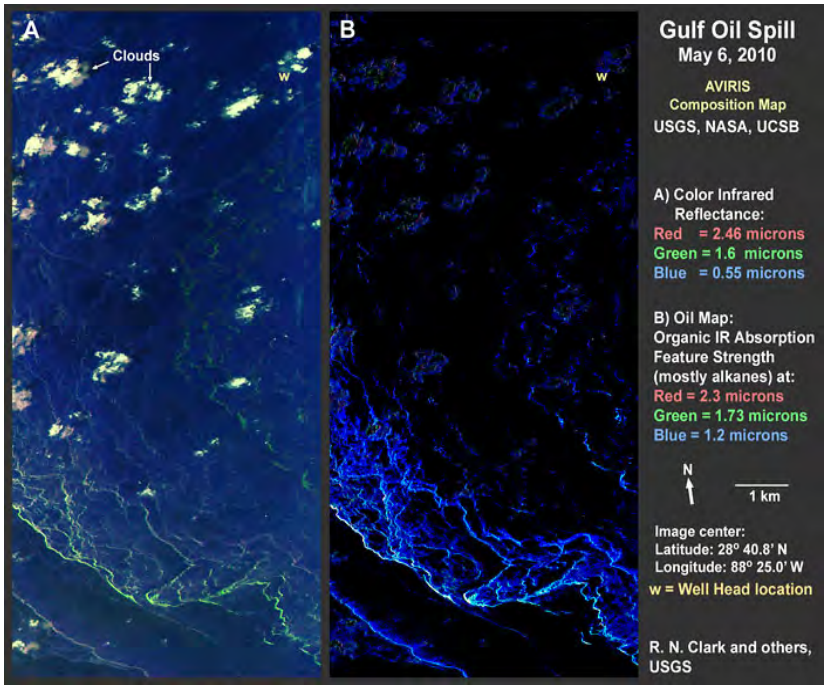
Spectral curves for the oil itself have been measured in the Spectroscopy Laboratory of the U.S. Geological Survey in Denver, CO.



Compare that plot with the one produced by the AVIRIS overflight at 28000 feet (~8.5 km).



The AVIRIS data have been used to produce images in false color using individual hyperspectral bands:



The following is a description of these images given in a NASA press release: AVIRIS measures a spectrum of the surface at each pixel from 0.35 to 2.5 microns (the visible spectrum is: blue: 0.4 microns, green 0.53 microns, deep red 0.7 microns) in 224 wavelengths. This fine spectral sampling allows discrimination of absorptions due to specific compounds in the scene. A, a color-infrared composite image in which reflectance at 2.46 microns is assigned the color red, reflectance at 1.6 microns is assigned green, and reflectance at 0.55 microns is assigned blue. B, an image produced by assigning the measured strength of the absorption at 2.3 microns as red, assigning the absorption strength at 1.73 microns to green, and assigning the absorption strength at 1.2 microns to blue. The absorption at 2.3 microns is intrinsically the strongest organic absorption in the oil in this spectral range and is sensitive to thin/small amounts of oil. The absorption at 1.73 microns is sensitive to greater amounts of oil. The absorption at 1.2 microns is weaker and needs the greatest amount of oil to register. For such thick layers of oil, the stronger absorptions are saturated and do not change significantly (as shown in figure on right) leaving the 1.2-micron feature to probe the thickest oil. The image on the left (A) shows the position of the oil well head (marked with a "w"). Pixel spacing is 8.5 meters so each pixel

covers about 72 square meters. The images shown cover about 67 square kilometers, and are a subset of a long flight line. Data analysis indicates oil was detected in a total of more than 4 million square meters (more than 57000 pixels) in this scene. Small dots, some of which appear white in A and red or green in B, are boats whose paint reflects light with organic spectral signatures.

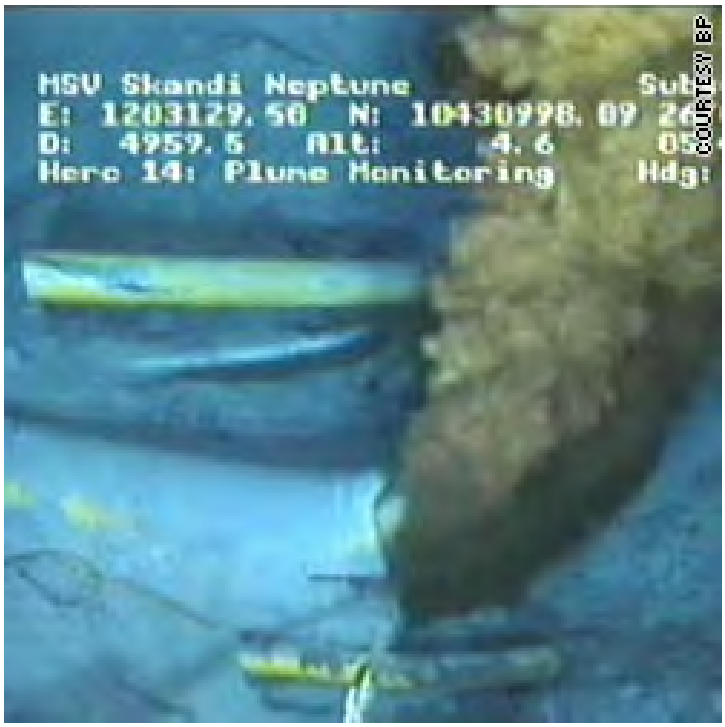
Satellites and aircraft are the primary platforms for observing the effects of the spill over the open ocean and along the beaches and coastal wetlands. But the scope of the spill catastrophe has justified some other novel ways of monitoring these effects. A Navy blimp, the MZ-3A, and helicopters, allow observers to remain over small areas at a time, noting details that are immediately relayed back to ground workers engaged in setting up protection barriers or in cleaning up contaminating oil. Here is the blimp and a helicopter, being used in tandem:

A NASA web site summarizes how satellites play(ed) a role in the BP spill monitoring.

To see how the spill's main segments have moved from day to day, click on this CNN website. Note that there are significant variations from one 24-hour period to the next. These changes are due in part to the influence of winds and currents but also due to shortcomings in the satellite detection of boundaries, etc.

Of course, satellites can only see oil at the surface. There have been reports that much oil remains beneath, at various depths, dispersed and also in plumes. The evidence for this is ambiguous but such subsurface distribution is likely, despite oil's tendency to rise through the water.

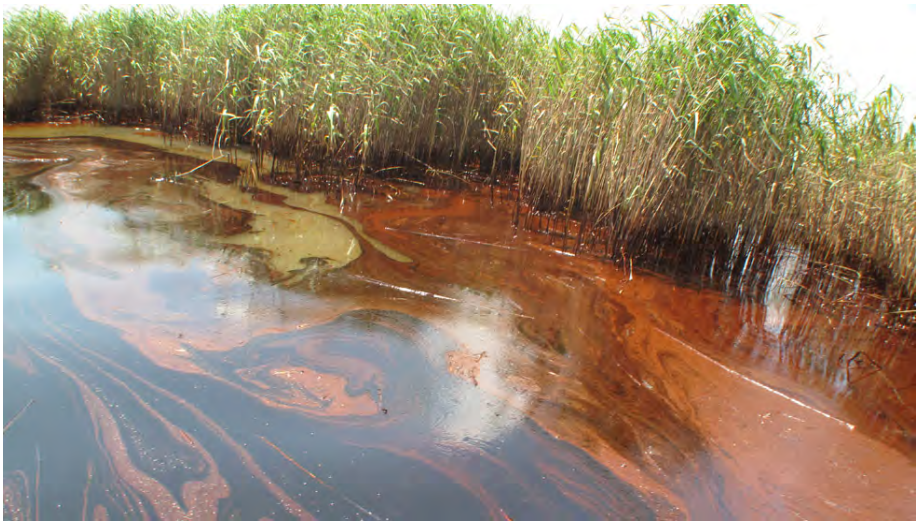
Now, with this overview of how the BP oil spill has been observed by various remote sensing endeavors, mainly satellite observations, let us examine in some detail the ecological results of the spill and then the efforts that have been made to stop or at least contain the spill at the wellhead, in part by capturing some of the escaping oil and bringing it to ships at the surface. This was the first failure of an oil well in deep water. That depth, nearly a mile, has made it extremely difficult to cap or otherwise contain the escaping oil. Remote cameras from nearby submersibles capture the gushing oil at the wellhead:



Many Internet sites posted live pictures of the oil gushing from the wellhead. Here is one such site that should work:YouTube.

The oil appears dark, almost blackish as seen underwater. In sunlight it has a reddish overtone, as visible on a worker's hand in this photo:

By mid-May, oil had now reached en masse the fragile wetlands of the Louisiana and Mississippi coasts, as exemplified here:



The effects of the spill on both the marine and the coastal ecosystems has been devastating (but this is generally true for most such spills). This is brought home by this photo of an oil-covered Brown Pelican - the state bird of Louisiana:

Hundreds of pelicans have been retrieved, cleaned up, and released elsewhere in Florida and Texas. But many are not being found and are presumed to have died. Oil-drenched pelicans cannot fly off the water:

Among marine animals seriously affected by the oil are several species of endangered sea turles:

The deaths of turtles, dolphins, birds, and many marsh dwellers is evident in this image (but keep in mind that fish and countless smaller marine organisms have died as well):

One of the most dire consequences of the spill has been its impact on the fishing industry. Especially vulnerable are the tuna, the shrimp, and the oyster beds. Louisiana is the largest supplier of oysters in the U.S. The oyster industry, some firms more than 130 years in business, has been financially devastated. Here is an example of a compromised oyster.

Crabs, another mainstay of the Gulf seafood industry, is also much affected.

Dead fish are washing up on shore. Some are covered with oil, as seen below. But did the oil itself kill the fish. Post mortems suggest yes, in some cases. But not always: other factors may be involved and the oil is just a coincidental coating.

The BP spill is now judged to be the worst petroleum drilling disaster in American history. It has closed fishing and shrimping in a large part of the northern Gulf of Mexico and is inhibiting tourism. Its price tag will eventually exceed tens of billions.

The amount of oil that has leaked so far has proved hard to estimate. BP originally said that 5000 barrels (one barrel contains 42 gallons) come out each day. Others believe the amount is (much) higher (that has been confirmed; see second paragraph below). For the lower estimate, this yields about 7.9 million gallons in the first 35 days. However, as time progressed and better observations of the gushing at the wellhead led to more exacting calculations, the dailing spill size estimate kept rising, first to between 11000 and 19000 barrels, then 25000 barrels, then 35000 barrels, and thereafter to between 40000 and 60000 barrels (a few claim even more).

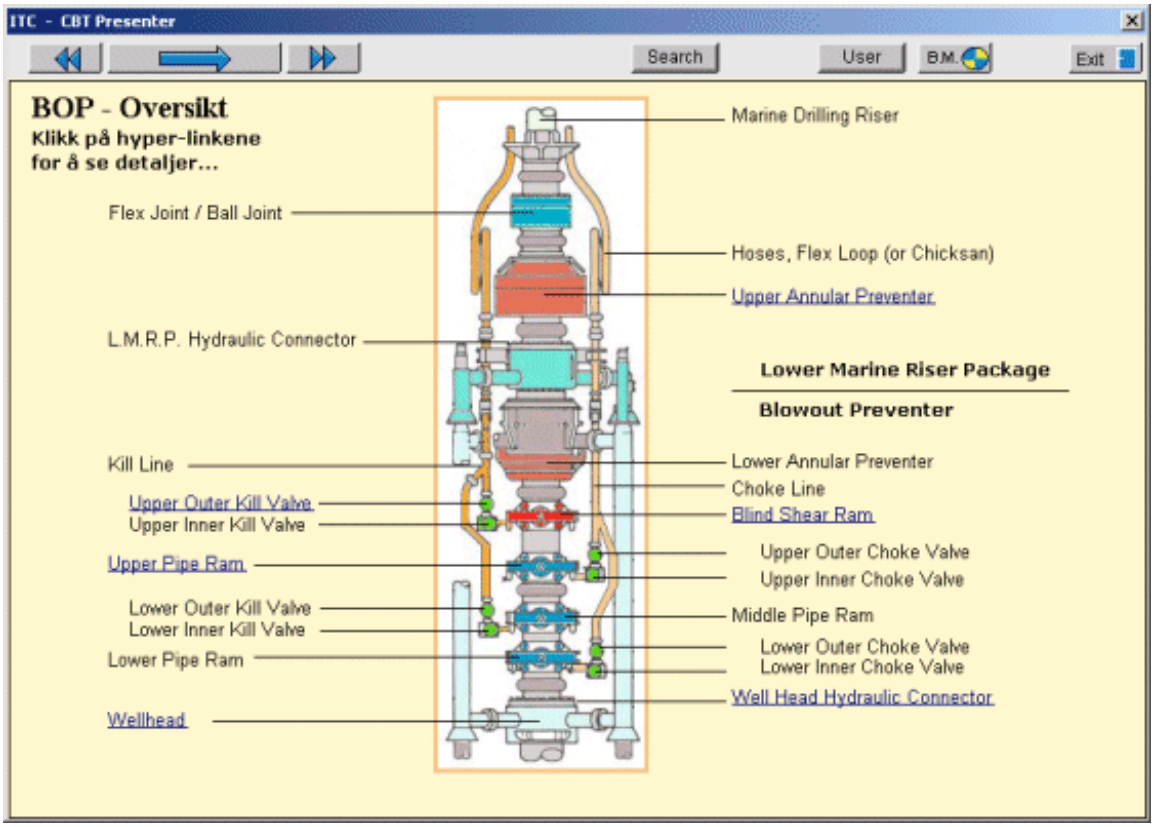
Until 2010, the largest spill in the western hemisphere was in the Gulf of Mexico off the Mexican coast. The Ixtoc drill platform, operated by Pemex, in the Bay of Campeche off the Yucatan had a blowout failure that released 138 million gallons of oil over a 10 month period (time needed to finally shut it off) between June of 1979 and March of 1980 (see photo below). The oil drifted northward and eventually produced moderate contamination of the Texas coast. But in time its initial ecological impact has lessened to the now almost nil traces of damage. Most of the oil stayed at sea and was gradually destroyed by a combination of dispersants and petroleum "eating" marine micro-organisms. Several methods were tried to stop or contain this spill but the only thing that eventually worked was the "relief" well that intersected the Ixtoc well and sealed it off with cement. Because this spill has similarities to the Louisiana spill, you can gain some perspective on these spills - and spills in general by clicking on this Wikipedia web site.

Based on three separate methodologies, the independent conclusion made by the federal-

directed Flow Rate Technical Group, which has analyzed both the areal satellite coverage and the wellhead output as monitored by cameras, has determined that the overall best initial estimate for the lower and upper boundaries of flow rates of oil is not the 5000 barrels reported by BP, but is in the range of 12,000 and 19,000 barrels per day. This is, however, still not a firm number. But 19000 barrels translate into 798000 gallons per day, and for 38 days would amount to about 30000000 gallons, greatly exceeding the Exxon Valdez spill.

So, what has been done to stop the flow and seal off the well. To help in answering this, here is a quick primer on a common type of failure in an oil or gas drilling exercise.

Petroleum occurs in openings - pores and fractures - within rocks, usually sandstones. Liquid oil is generally accompanied by varying amounts of natural gas. after the reservoir is penetrated by a well, the local petroleum products move under pressure to the well pipe and are driven up to some extent by the gas and/or more customarily by a partial vacuum in the hole created by surface pumping. The emerging material can often be under such pressure that it gushes out. This needs to be controlled to avoid what is called a blowout. To accomplish this, a large apparatus, called a Blowout Preventer or BOP is commonly mounted onto the wellhead. The BOP has various "fail safe" devices designed to close the pipe stream if too much gas, or oil under high pressure, is sensed and detected. Here is a cutaway diagram with parts of a typical BOP (these can have a variety of sizes and shapes and cutoff components) and a photo of a BOP at the surface of the drilling rig:



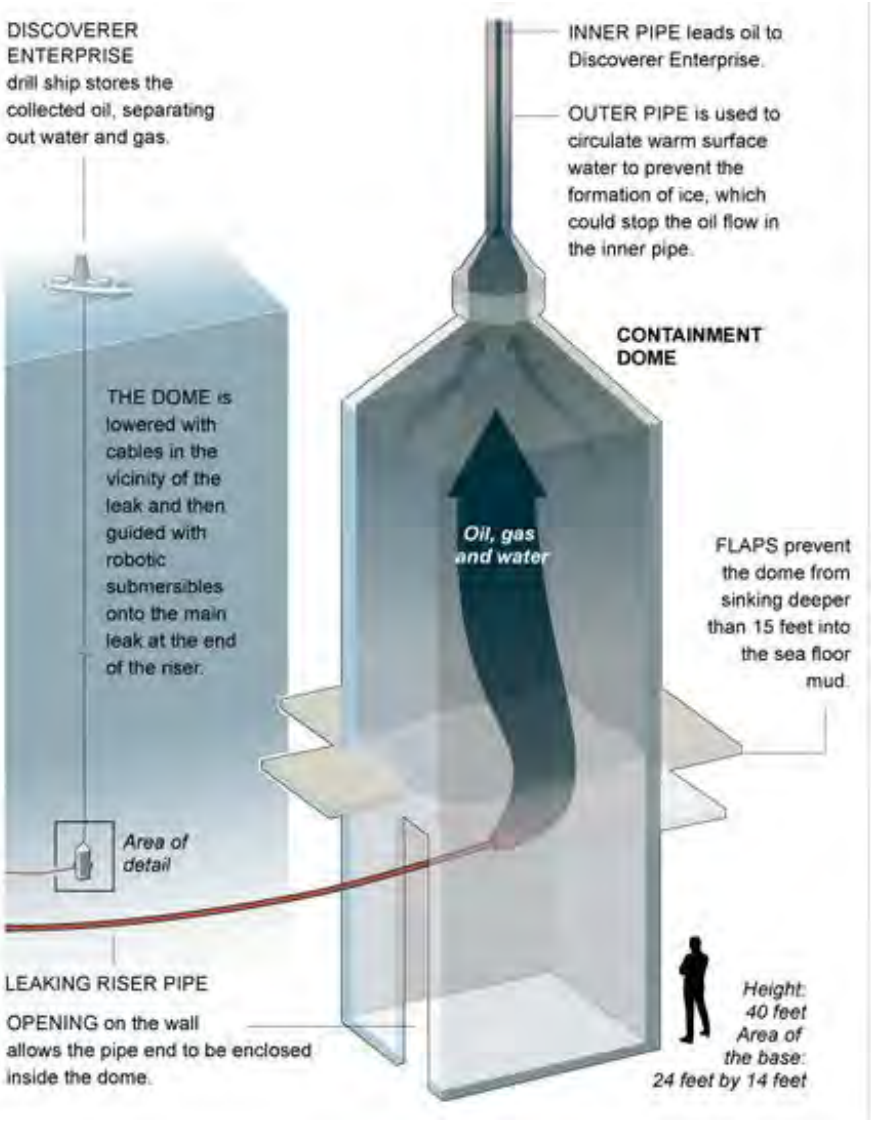
The Deepwater Horizon accident (location referred to platform number mc252) occurred because pressurized oil, accompanied by entrained gas, reached a danger point that should have triggered the BOP action. The problem was compounded by a bad decision a few hours earlier to replace heavy drilling mud with seawater. At a critical pressure, the BOP failed to function properly because of several factors including a leak, a dead auxiliary battery, and ineffective maintenance. The gas/oil mix reached the surface, was ignited by something - perhaps a friction spark - and set off the devastating explosion that eventually doomed the platform. Oil escaped from the wellhead complex. For weeks, millions watched on TV news the camera images such as we saw above, and seen again as a reminder:

Once this scenario was set in motion, BP initiated plans to gain control over the gushing oil even as the spill grew in size and volume. This graphic depicts various options available to BP and its contractors (including Transocean and Halliburton).

New ocean-going platforms were sent in as staging units from which to conduct several efforts to stop the spill. ROVs, or Remotely Operating Vehicles, were lowered to the well both to observe the effluence and to assist in conducting steps involved in the fix. ROVs, developed both for oceanographic science studies and for drilling monitors, have TV cameras and apparatus for perform tasks. They are, in fact, classic examples of remote sensors comparable to the types considered in this Tutorial. Here is a typical ROV:

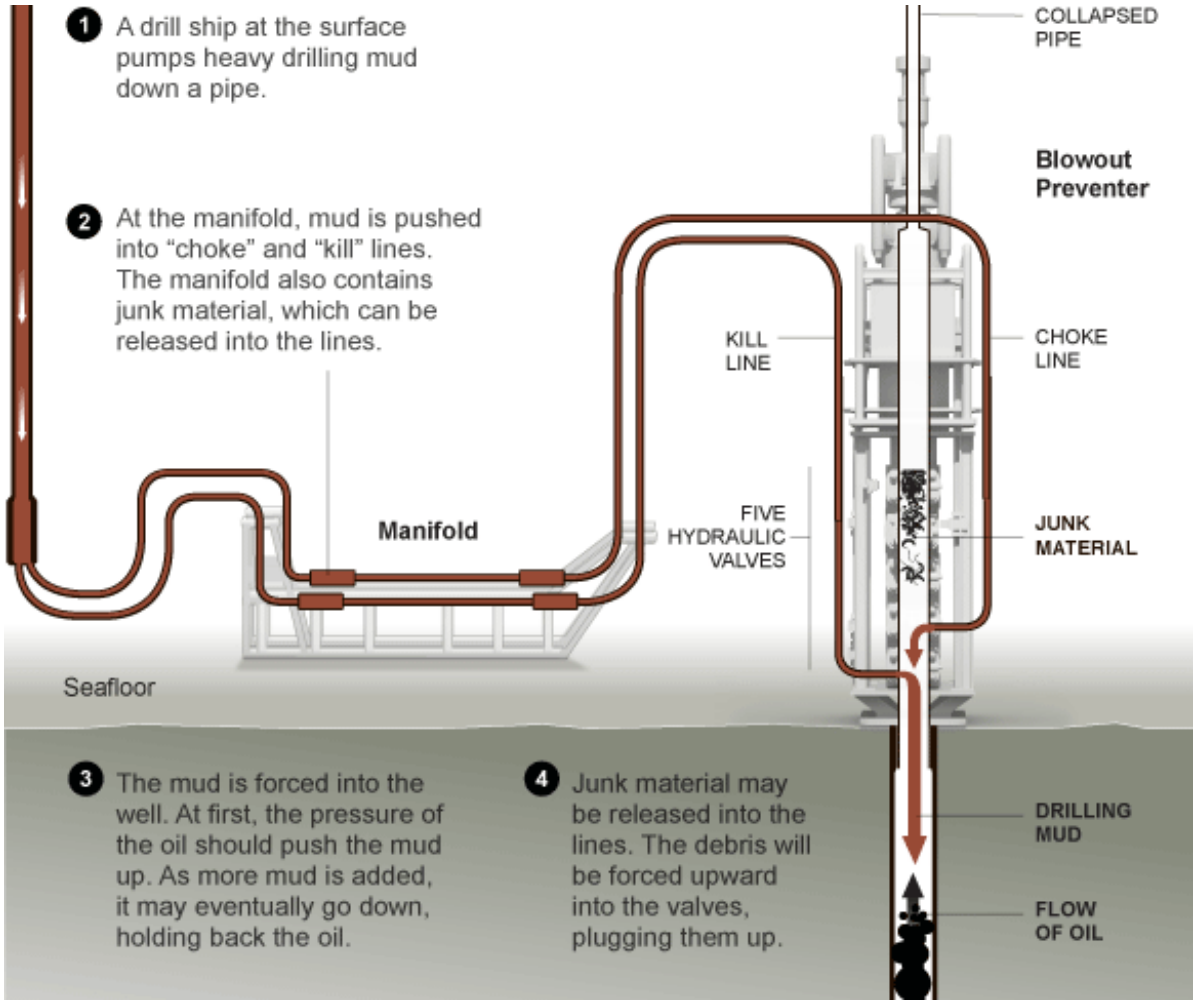
The site around the original Deepwater Horizon rig has been populated with several relief well rigs and other installations, supporting nearly 1800 workers (in 3 shifts):

The first attempt to stem the oil flow was to place a large containment shell over the BOP, collect the oil inside, and pump it to the surface. This was the arrangement:



This try failed within a day. At the high ocean bottom pressures, the physicochemistry caused the shell to be lined and clogged with methane hydrates.

The next effort, in later May, was the so-called Top Kill scheme, which has about a 60% chance of succeeding. This diagram explains the procedure.

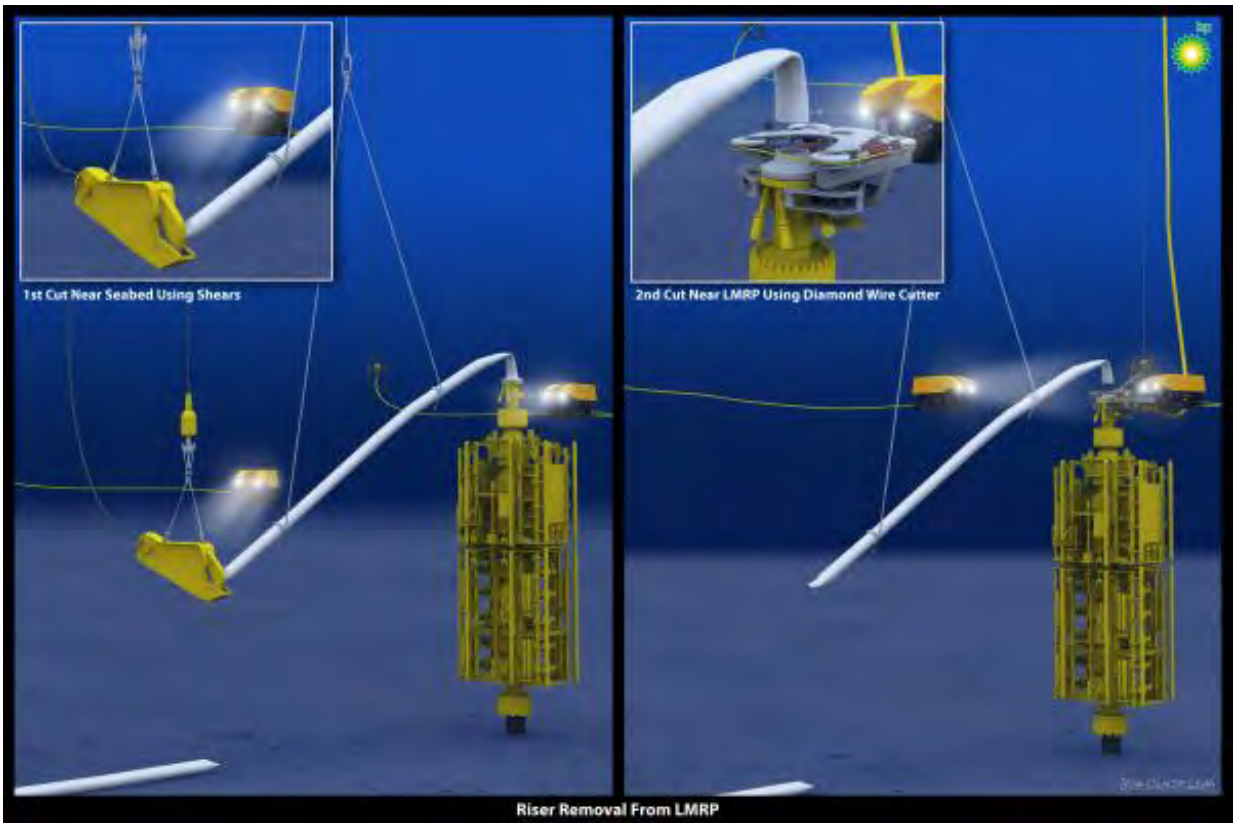


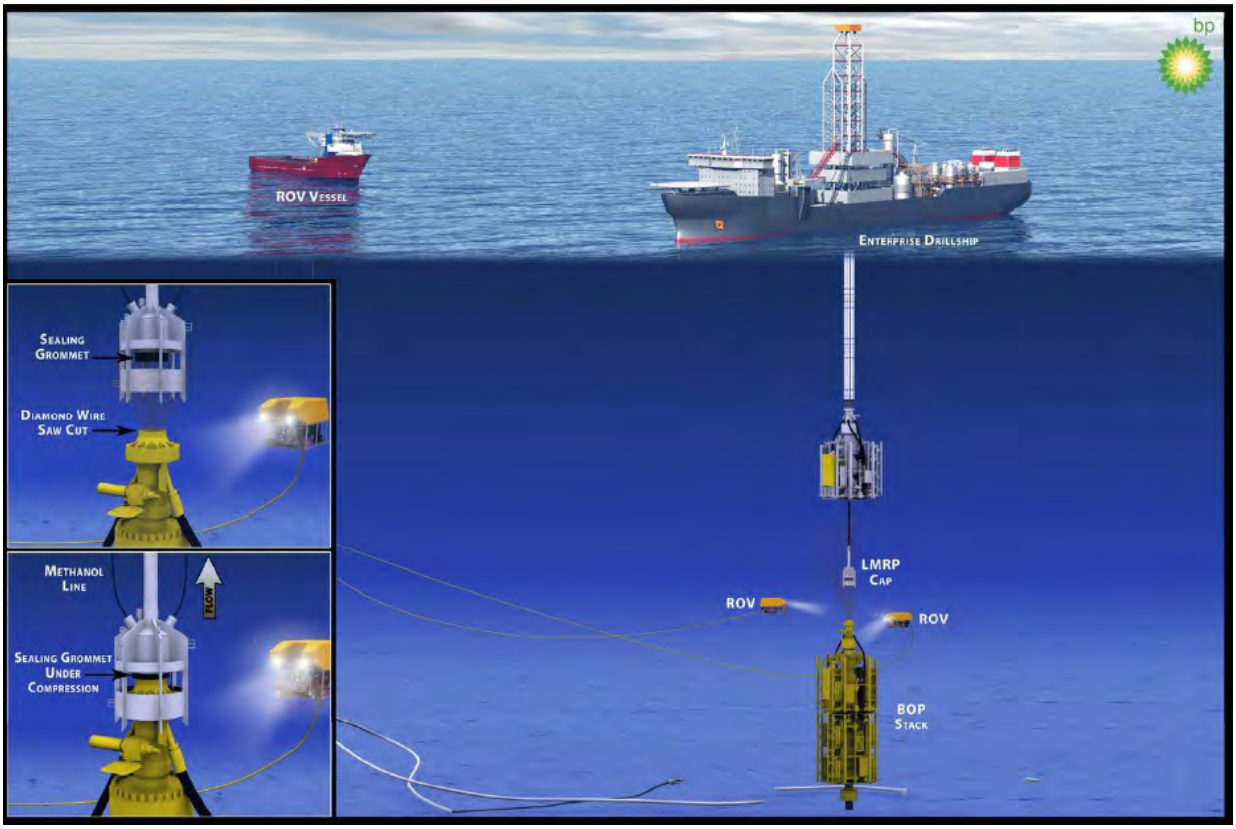
The basic concept was to inject the well with heavy mud mixed with rubber fragments, golf balls, etc. that would counter the pressure from the gas. Millions of TV viewers watched in awe the live feed pictures supplied by BP. The writer predicted failure when this mode of imagery became commonplace:

The color of the effluent was that of the drilling mud. Thus, the pressure of the gas/oil mix was just too great for the mud injection to overcome. Top Kill was abandoned in late May.

As June opened, BP decided to attempt a maneuver that would only siphon off part of the oil. Known as DMRP (Deep Marine Riser Package), its purpose is to remove broken pipe, cut a clean condition around the main exit pipe (riser), and cap the remaining pipe with an assembly that would direct the oil into pipe reaching a surface ship. This is the setup, with BP's description

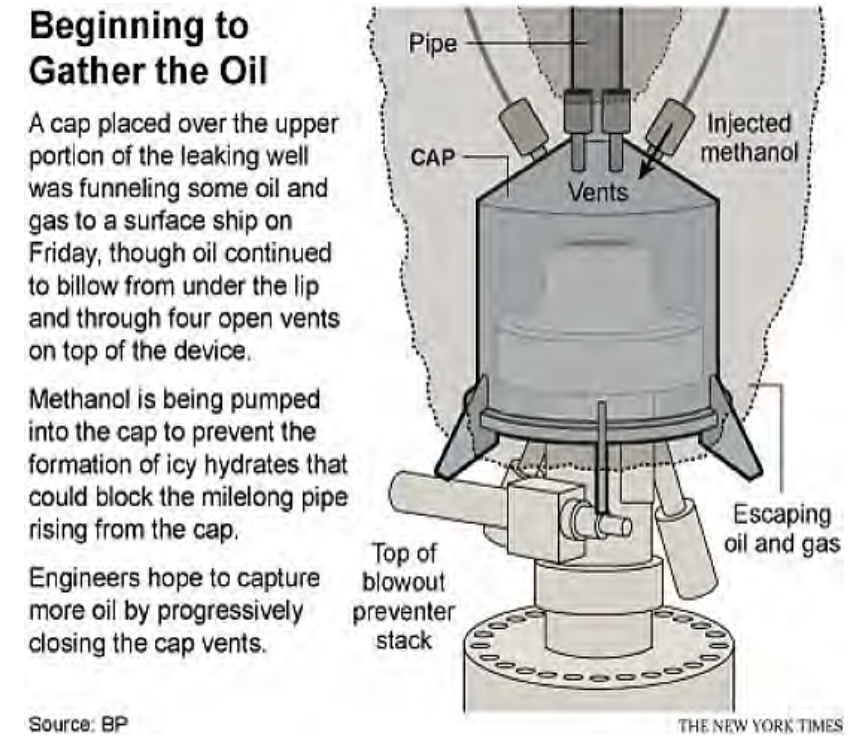
beneath it:

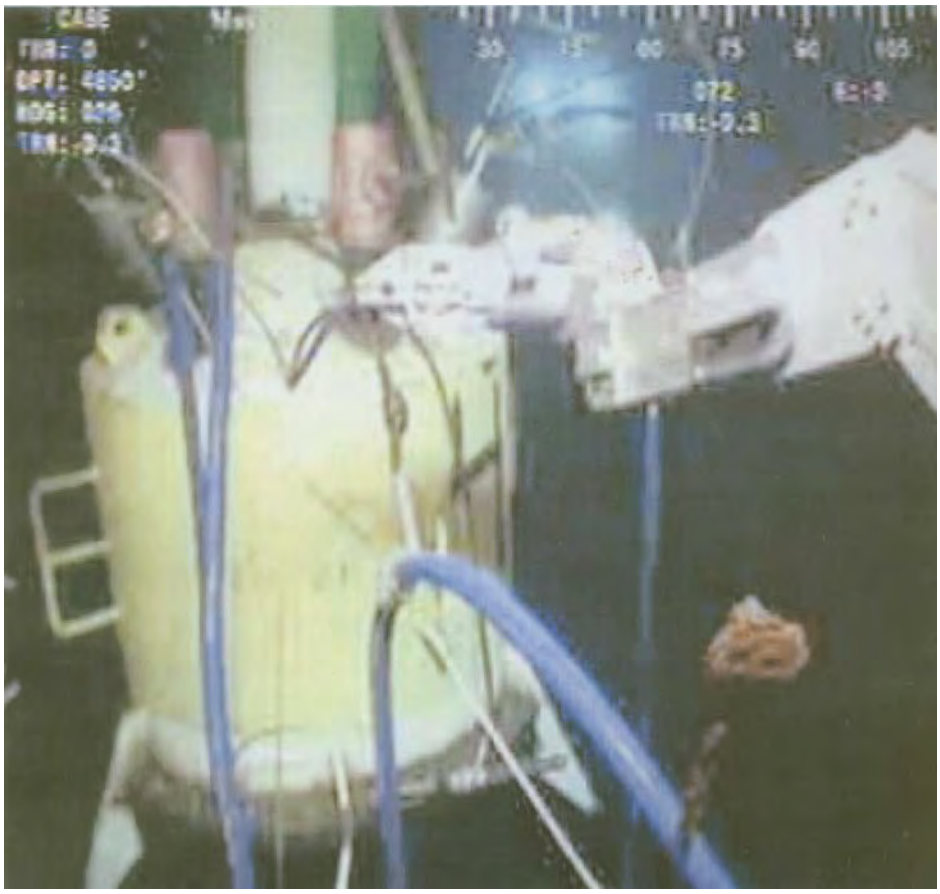




Installing a Lower Marine Riser Package (LMRP) Cap is a containment option for collecting the flow of oil from the MC252 well. The LMRP is the top half of the blow out preventer (BOP) stack. The installation procedure first involves removing the damaged riser from the top of the BOP. A remote operated hydraulic shear made two initial cuts and then that section was removed by crane. A diamond wire saw was used to cut the pipe close to the LMRP and the final damaged piece of riser has been removed. The LMRP Cap is designed to seal on top of the riser stub. The seal will decrease the potential of inflow of seawater as well as improve the efficiency of oil recovery. Lines carrying methanol also are connected to the device to help stop hydrate

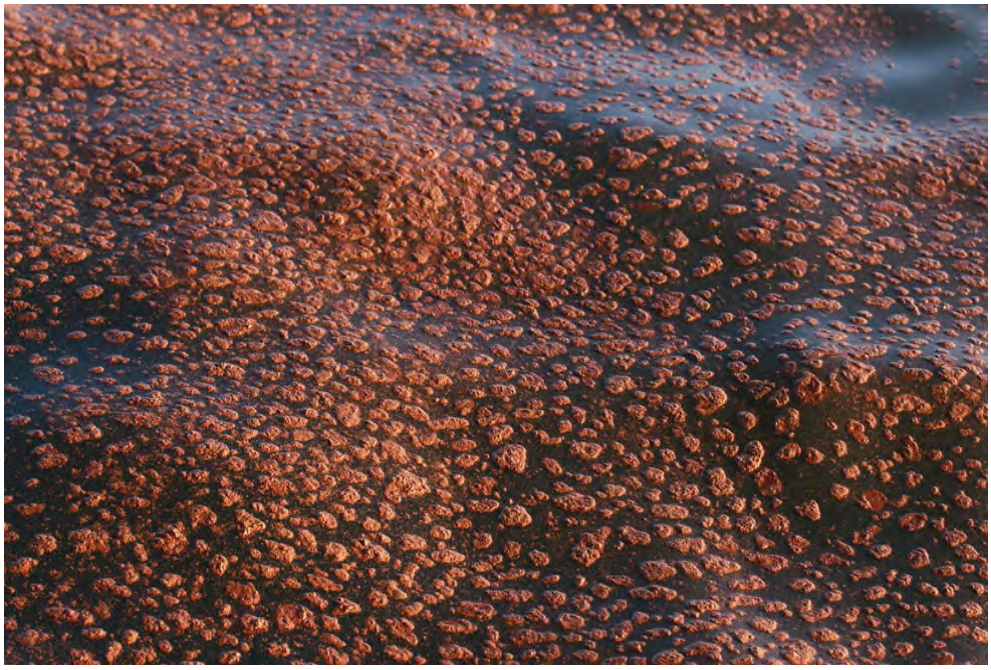
formation. The device will be connected to a riser extending from the Discoverer Enterprise drillship. The cap, sometimes called the Top Cap, is shown schematically and then in place. It has succeeded to siphoning off up to 25000 barrels a day, and BP hopes to increase the amount:



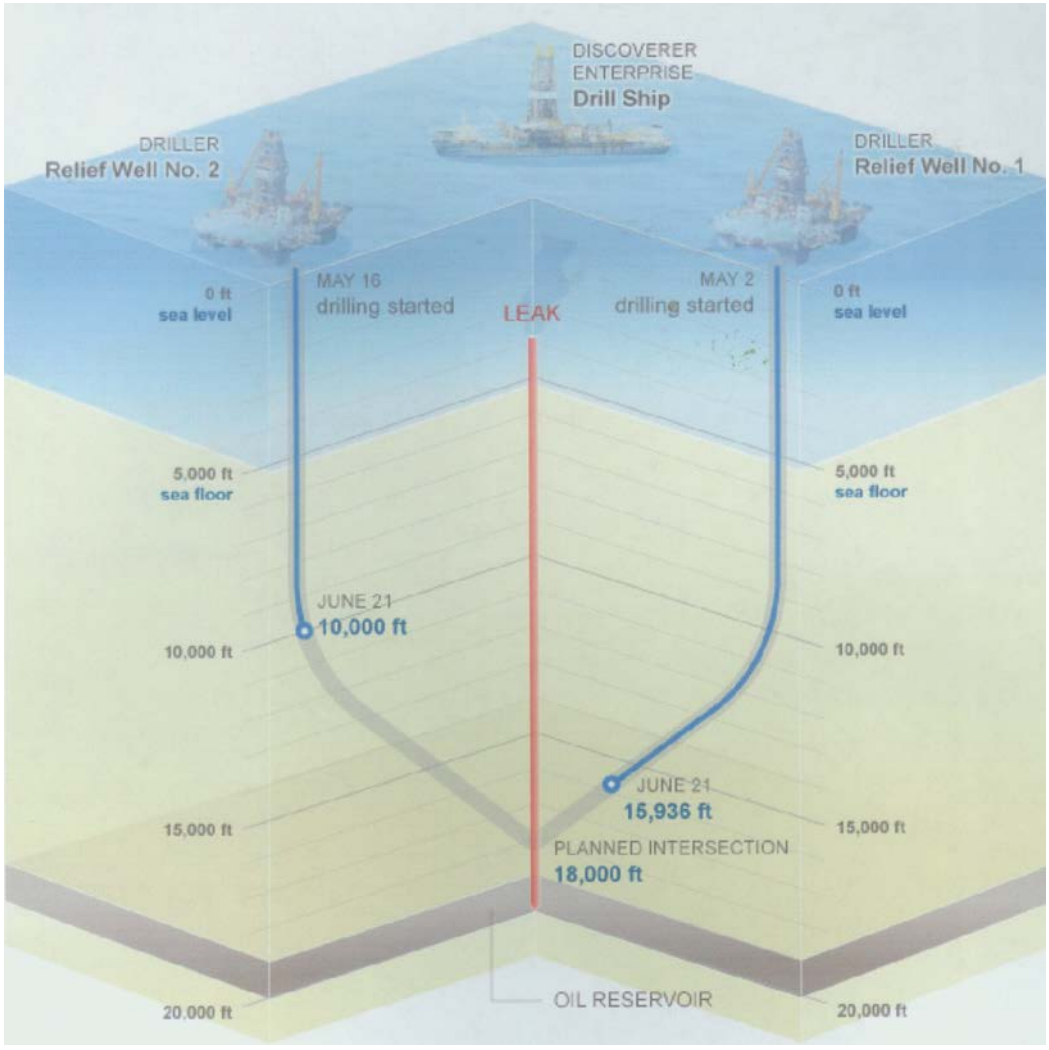


While these several attempts were going on, BP, the Coast Guard, the National Guard, and private citizens (including the fishermen whose livelihoods are now totally compromised) were mobilized to clean up oil reaching the shores. Attempts to forestall this direct assault on the wetlands were along five lines: first, to break up the oil at the wellhead by injecting an emulsifier into the escaping oil; second, to spray the emulsifier (Corexit) from aircraft onto parts of the spill; third, to burn off water surface oil by controlled burns; fourth, to suck up the oil out in the Gulf, using small to large boats; fifth, to skim the oil off the surface using skimmer boats, sixth, to

prevent oil from reaching the shoreline using a continuous line of booms, and seventh, once the oil has reached shore (such as the famed white sand beaches of the Gulf), remove it in a variety of ways including "brute force" shoveling by hand.



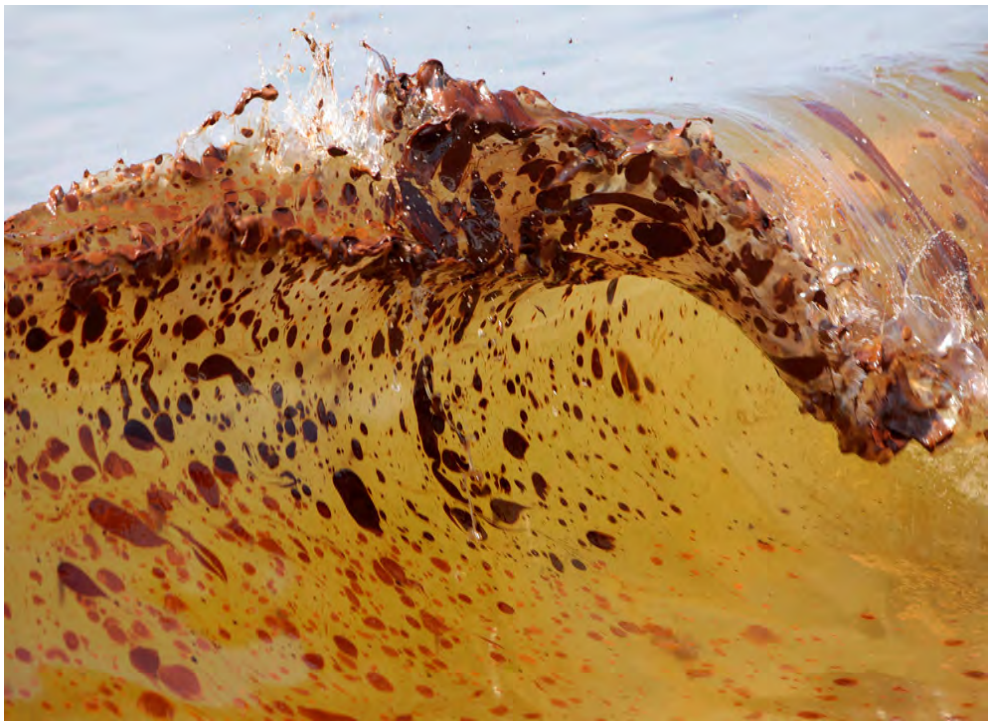
As June opened, BP and the federal government, along with many experts, shifted their attitude to one that the spilling may likely go on indefinitely. The best hope would be intersecting the original well with two relief wells (the second as backup) that should occur sometime in August. This diagram shows progress as of mid-June:



The destruction to Louisiana's Bayou "way of life" has just had to be accepted as unrelievable, with the hope that over the second decade of the 20th century the ecological damage will ameliorate, with eventual recovery. The fears for the summer of 2010 were that the coastline

effects will spread to other states, driven on by the hurricanes forecast for this period, the relentless growth of the spill (by June 1 it covered 6000 square miles), and, heaven forbid, the entrainment of oil into the Loop current. Sure enough, the first hurricane of the season, Alex, in late June effected the efforts to clean up the oil by forcing the skimmer boats to cease operations for several days. Alex slammed ashore just south of the Texas-Mexico border but its winds stirred up most of the Gulf:

Everyday waves crashing on shore are often dominated by an oil coating, as seen in this photo:



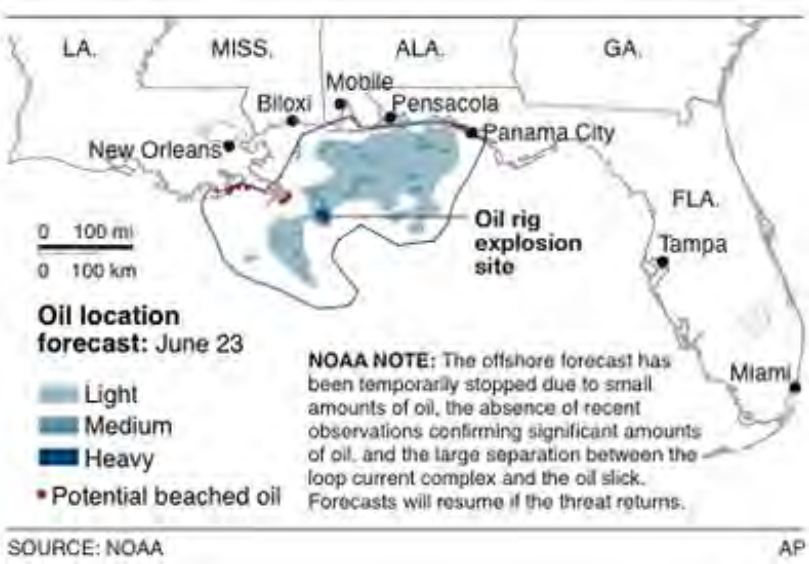
By June oil had reached some wetlands and beaches in Mississippi, Alabama, and western Florida (particularly around Pensacola):

Tar balls are common along most of the affected Gulf Coast beaches:

The oil is so pervasive it stains individual quartz sand grains:

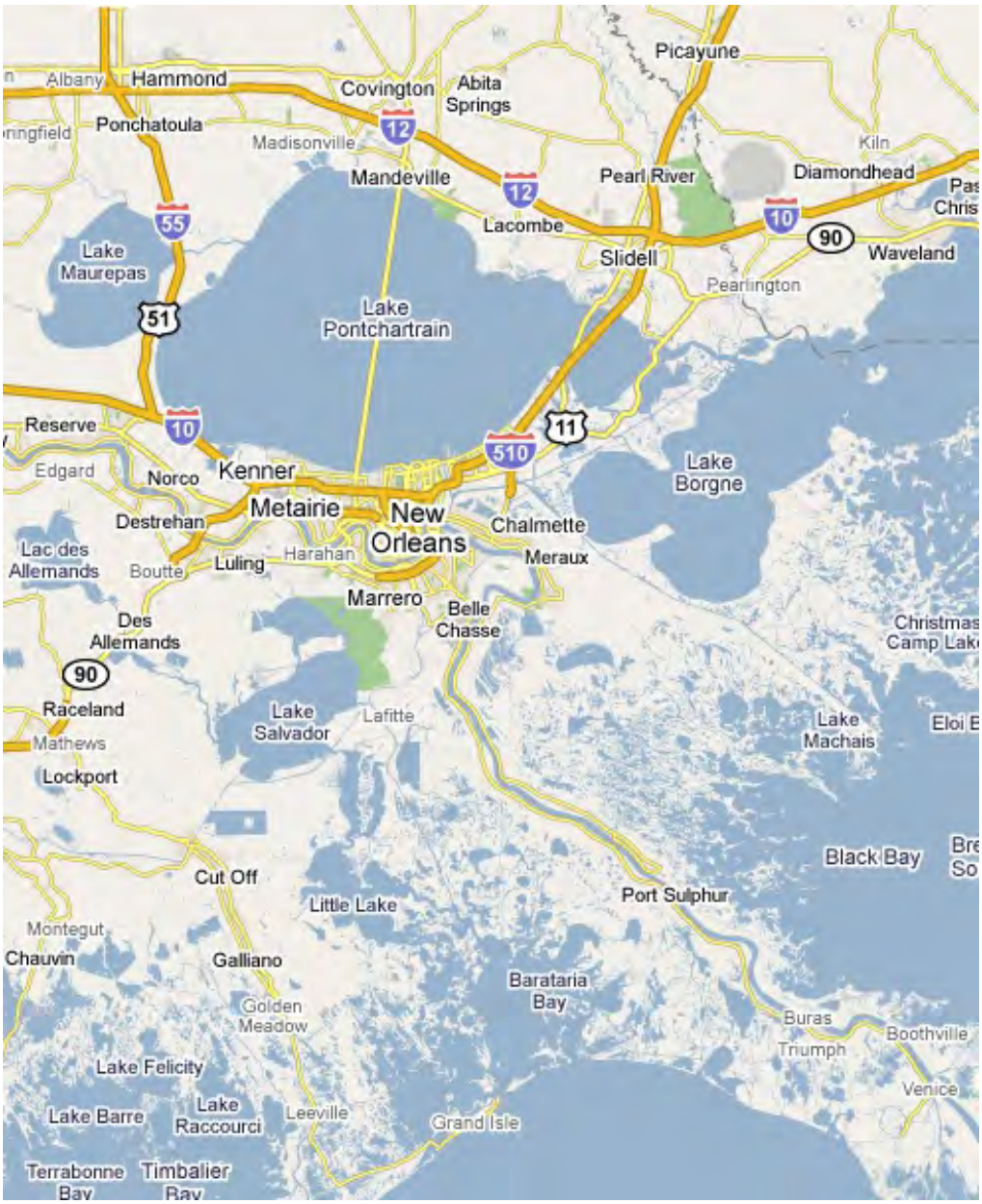
Several of the barrier islands have been severely affected. As an example, here is the tip of one of the small Chandeleur Islands of Louisiana:

As of early July, the BP well disaster continued almost unabated, although some oil was now being captured. One of the great fears is the prediction of several severe hurricanes which could spread oil well inland. Using satellite data, daily forecasts of where the oil is and is expected soon are released.

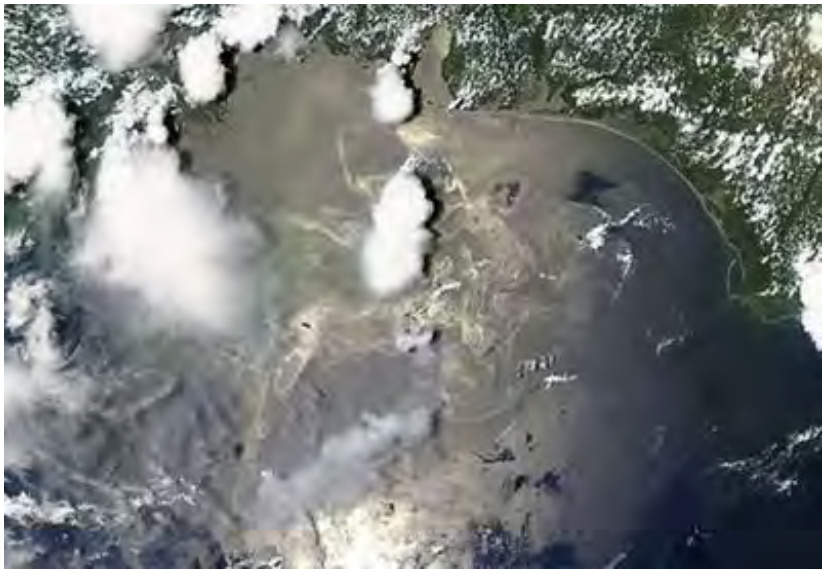


In early July oily balls reached Galveston, TX; they may have been carried there by clinging to ships. Much more alarming is the appearance, just after the 4th of July, of tar balls in the eastern waters and shore of Lake Pontchartrain, along which New Orleans is located:

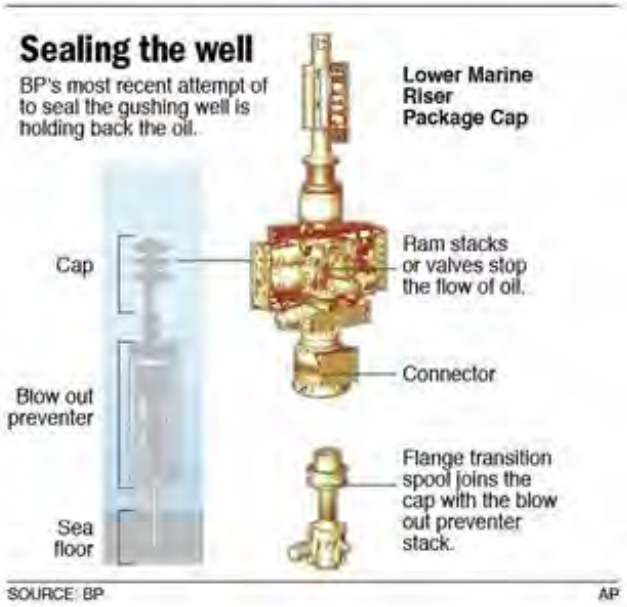
This is an unexpected development since the lake is inland. There is a narrow connection via Lake Borgne, but the pathway to that lake is itself devious relative to the Gulf's open waters.



The flow continued until mid-July:



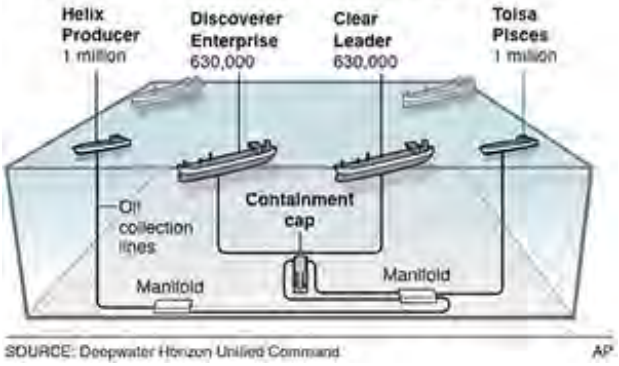
With so much oil still escaping, BP decided to put a new, more sophisticated containment (stack) cap on the well. It was successfully mounted on July 11 and testing to determine its stability (can it withstand the pressure from the flowing oil?) is underway. The cap weighs 75 tons, as shown below. It has four outlets to which pipe will be attached; the four pipe stems will connect with 4 surface ships. It is hoped that most, if not all, the oil will be captured in the interim before the well is sealed by one (or both) relief wells.



BP says it's close to containment

BP hopes to fully contain the oil leak gushing since the Deepwater Horizon exploded April 20. The plan calls for a larger cap over the blowout preventer, and a total of four vessels at the surface that can process oil and gas.

Capabilities of ships to process oil, in gallons

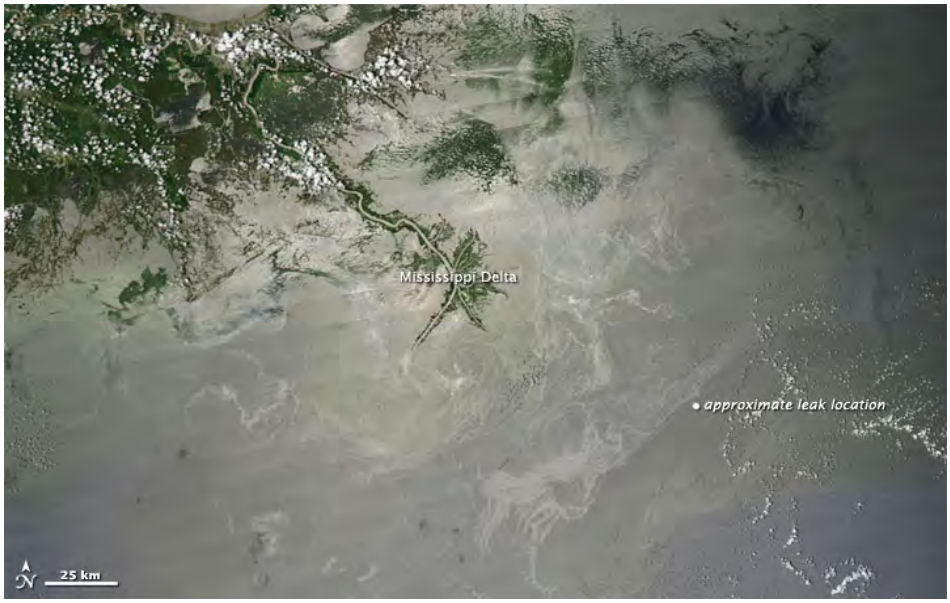


SOURCE: Deepwater Horizon Unified Command AP

By July 15, the 4 main valves on the cap were closed. Soon thereafter, oil ceased to gush, as shown in this ROV image of the top of the riser cap:

On July 19th, BP reported that there was some (perhaps minor) leakage in the vicinity of the wellhead. If confirmed, this would be the result of escape from rock fractures that intersect the well.

Wonder of wonders! The cap has held into August. No leakage has occurred. And, a combination of a storm and the action of dispersants has greatly reduced the amount of oil visible on surface waters. The last day in which oil could be detected by satellite was July 24th, as seen in this MODIS image:



This rapid disappearance may also be attributed to a factor that some scientists had predicted wouod occur: Marine waters contain bacteria and other organisms that feed on oil. They are abundant and "welcome" the oil as a "bonanza". This, if indeed it is happening, can account for much of the vanishing act that has removed the spill as the lead item on the network news.

To sum up efforts to contain the Deepwater spill, between April 20 and July 20 of 2010, this itemized summary was copied from CNN:

1. Robots shut off blowout preventer valves. Outcome: Failed

2. Drill a relief well. Outcome: In progress

3. Install four-story containment dome. Outcome: Failed because of formation of hydrates, similar to ice crystals, that made the dome buoyant 4. "Top hat" containment device. Outcome: Abandoned before start

5. Four-inch insertion tube: Outcome: Succeeded, but had limited effect

6. Top kill (clogging well riser with drilling mud). Outcome: Failed

7. Junk shot (plugging riser with debris). Outcome: Failed

8. "Cut and cap" (cut off riser and place new containment cap). Outcome: Failed when diamond wire saw got stuck 9. Revised "cut and cap" (device placed atop jagged riser). Outcome: Some oil funneled to surface, but more still escaped

10. Second containment system (connection directly to blowout preventer funnels oil directly to ship on surface). Outcome: Oil reaches ship; operation continues

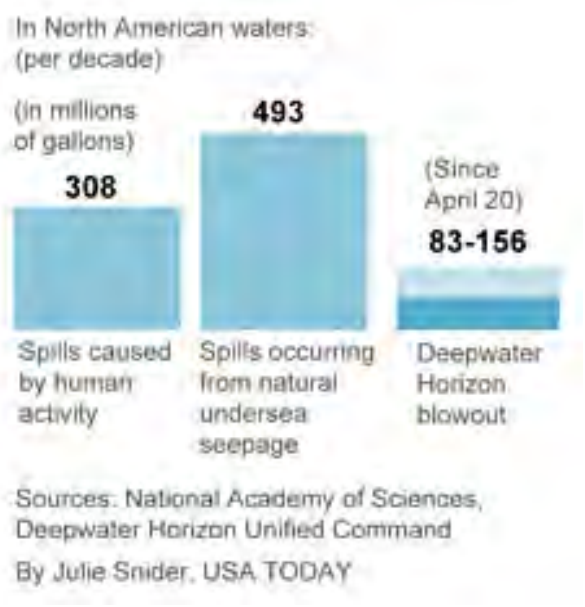
11. Better-fitting containment cap. Outcome: Oil flow shut off Thursday afternoon; test to continue for 48 hours

As of August, 2010 the estimates of the total amount of oil spilled from the Deepwater Horizon well range from 206 million gallons (4.9 million barrels) to 252 gallons (6 million barrels). If the upper limit is valid, the Gulf of Mexico spill is even more massive than the Persian Gulf spill in 1991, then the worst on record (5.7 million barrels). At its maximum outpouring, 62000 barrels a day were introduced to Gulf waters but at times this number was lower (thus the uncertainty in the range).

The good news, reached on July 15, is that the well has at last been effectively sealed. The process used is called "static kill". Heavy drilling mud has been injected through the cap and blowout preventer. It has forced the oil down the hole because it is capable of overbalancing the pressure of about 6900 psi at the well head. Afterwards, cement will be poured in to fully seal the top of the well pipe. This provides containment but as a guarantee of permanent sealing the relief wells will later seal off the bottom of the well. As September started, BP has removed the "temporary" cap used for the static kill and has put on a more effective cap.

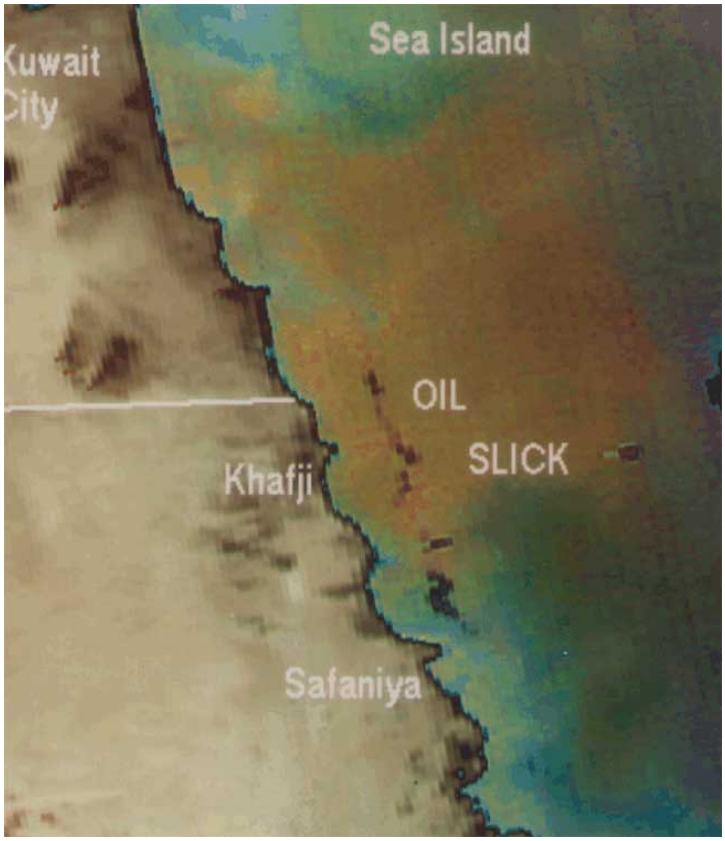
The relief wells reached and intersected the Deepwater well in mid-September, at a depth of about 5.7 km (3.5 miles). Cement was injected into both inner and outer casings. On September 19th, 2010 the well was officially declared "dead" (permanently sealed). Gulf Coast residents breathe a huge collective "sigh of RELIEF".

But the region will need years to recover. This next chart helps to put the BP spill in perspective; it shows that over long periods cumulative leaks from natural seepage exceeds that of accidental manmade mistakes but since the seeps are widely dispersed and often far offshore, their effects usually don't threaten coastlines.



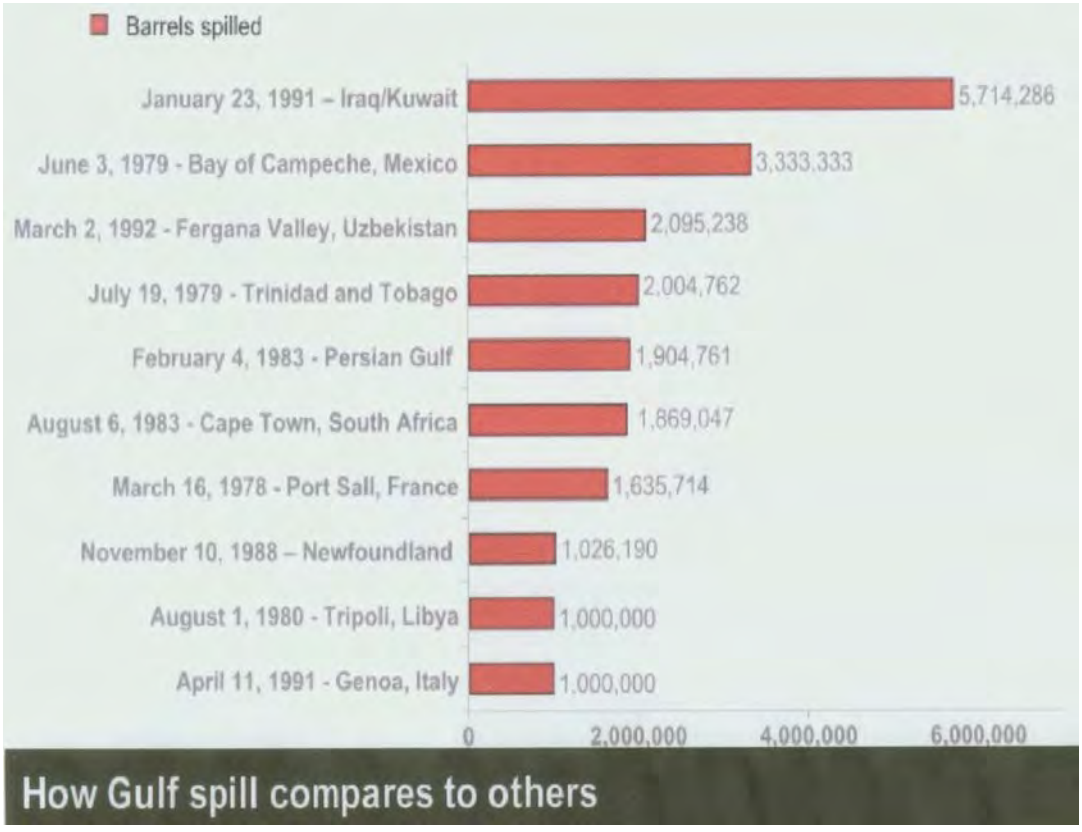
And, oil spills around the world, although infrequent, have happened repeatedly and will do so in the future as long as gasoline and motor oil continues to power automobiles. The Niger delta in Nigeria in West Africa has experienced thousands of such spills, many inland and some off coast. More than 30 million live in the Niger Delta and the Niger river floodplain. They are among many millions worldwide who have been victimized by inevitable oil spills. Check this website: Niger spills for a review of Niger's history of spills and an overview of spills in general.

So, bad as the BP spill has been - it has dominated the news and TV for parts of 5 months - Americans are not alone in their suffering. And there have been bigger spills in the past. Well over 200 million gallons were deliberately spilled by the Iraqis into the Persian Gulf during the 1991 war over Kuwait. On Jan. 23–27, 1991 in southern Kuwait, during the Persian Gulf War, Iraq deliberately released an estimated 240–460 million gallons of crude oil into the Persian Gulf from tankers 10 mi off Kuwait. Iraqi forces opened valves and emptied tankers, mostly at Kuwait's Sea Island terminal, with the intention, in part, of thwarting a landing by U.S. Marines arriving.



Other spills also resulted in higher magnitude releases, as summarized in the map below and in a bar chart; note that most spills in the map are not included in the chart. (Also, see list on this

Wikipedia web site.)



But, these statistics give little comfort to the residents of the American Gulf coast.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

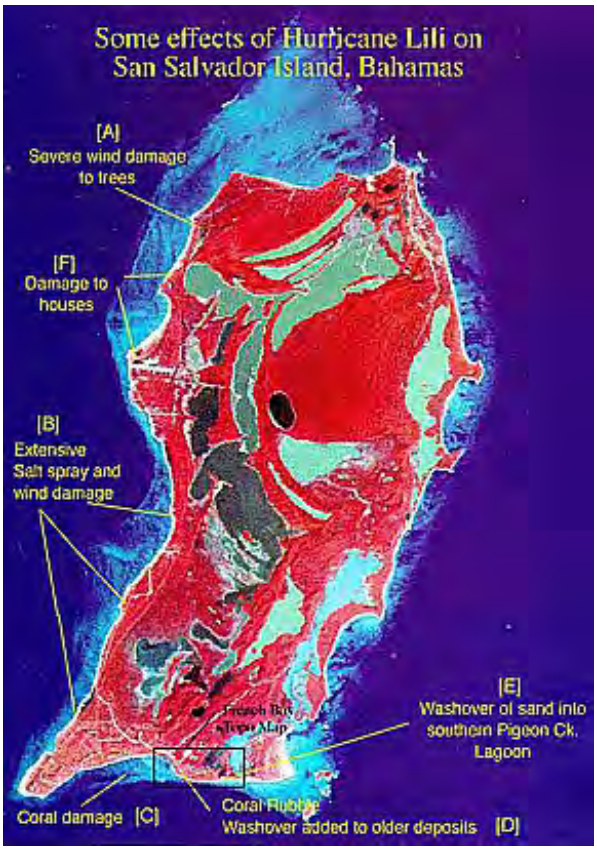
Satellites afford an excellent means to monitor on-going and potential ecological threats and damage, as well as long term after effects, to the Earth's natural surface and to areas relevant to human activities. Sometimes, ongoing ecological problems can be watched in near-real time using both high resolution and geostationary satellites. This is true for assessing the damage done to wetlands, shorelines, and forests hit by strong hurricanes. Much obvious damage is imposed on vegetation by forest fires and grassland burns. Also, normally very easy to see are the destruction and deposits associated with sand storms. Dust storms brought about by nature but sometimes made more severe because of human land practices can affect large regions. Oil spills constitute another ecological catastrophe that often is detectable in Landsat-type imagery and under favorable conditions in radar scenes. Damage deliberately attributed to human decisions includes strip mining. Illustrative examples are given.

Ecological Damage from Natural and Manmade Events

Non-Geological Events

Perhaps the most devastating natural event that affects both human infrastructure and natural landscape and vegetation is the massive storms called "hurricanes" or "typhoons". We show examples of these on page 14-10. Here we will show just a few illustrative examples of the kinds of damage that result. Hurricane Lili occurred in October of1996, starting in the warm Atlantic, passing over the Bahamas and other islands, causing considerable destruction of natural features as well as property.

Here is a satellite view of labeled damage on the small Bahaman island of San Salvador:

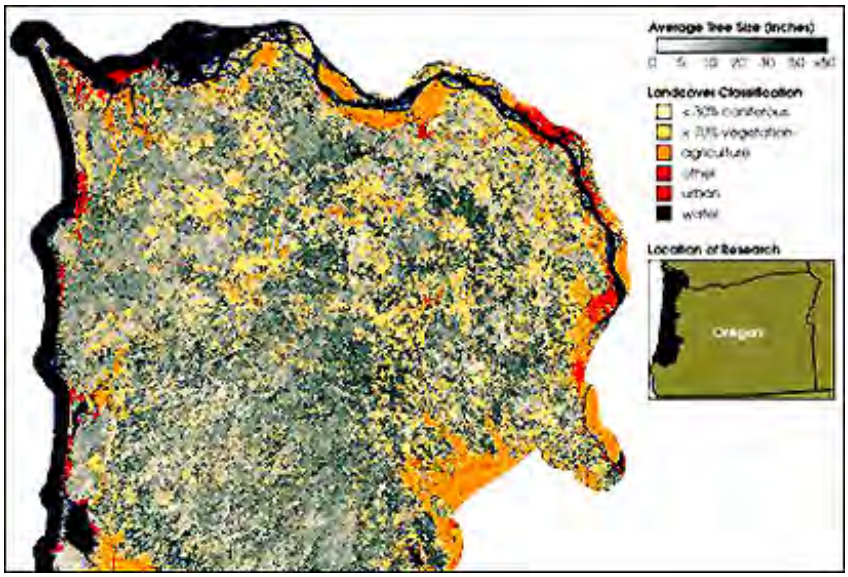


Ivan, in September 2004, did similar damage to Florida coastline dunes and vegetation

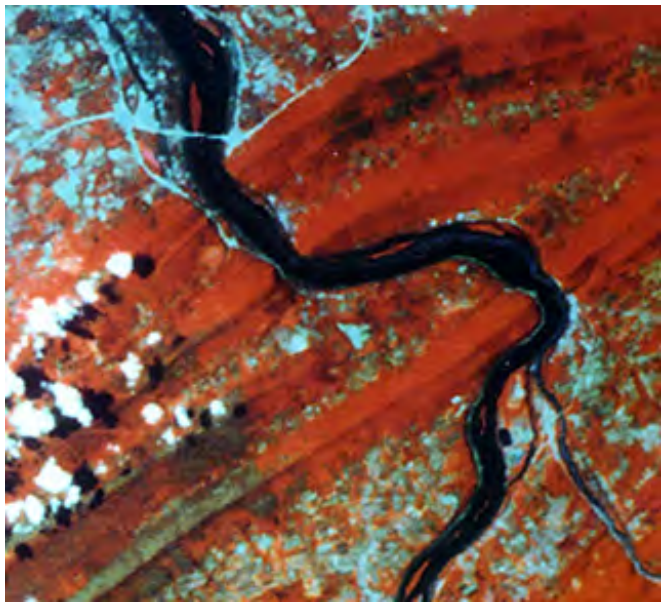
Trees are very vulnerable to being toppled by hurricane-force winds. Here are trees in New England blown over during a 1938 hurricane (they weren't given names then):

Landsat and other satellites are very effective in characterizing ecological habitats and in monitoring any changes that threaten plant or animal life, especially species considered endangered or subject to undesirable influences. One of the most famous habitat focal points in recent years is the battle between environmentalists and the lumber industry over what timbering, such as the clear-cutting considered in Section 2, is doing to the very specialized conditions in which the Spotted Owl (shown below) prefers:

This species favors dense, particularly first growth, fir forests. Although not yet rare on the West Coast, the owl is losing some of its prime habitat. At this time, there is a partial moratorium on harvesting certain types of forest ecology that support the Spotted Owl. The U.S. Forest Service is using Landsat in two ways: 1) to produce a basemap (shown below with the dark green being Spotted Owl territory), and 2) to ensure that these areas are not subjected to illegal clearcutting. Here is a classification of a large tract in northwest Oregon that singles out areas favored by the owl:

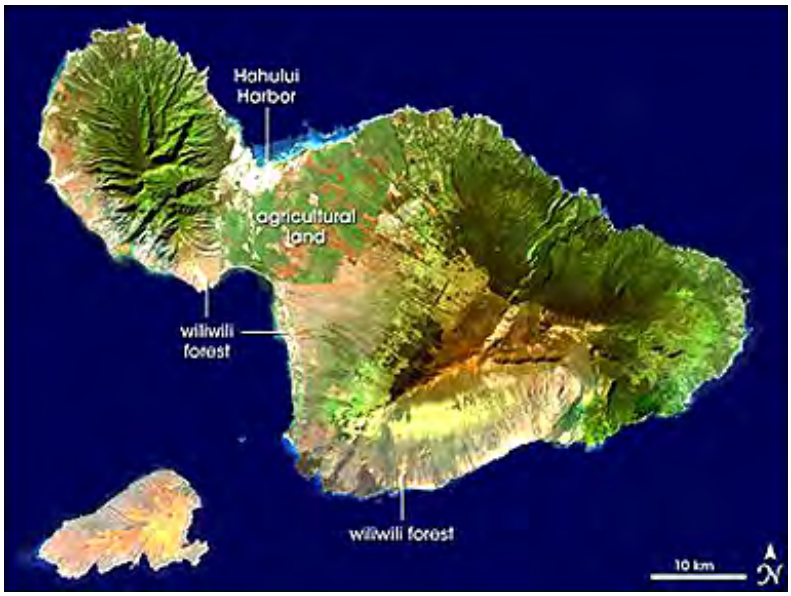


Landsat and other space observation systems efficiently monitor transient ecological maladies such as insect defoliation (as you saw in the Pennsylvania "Exam" at the end of Section 1). Drawing upon your recollections from that exam, in which you learned to pick out gypsy moth defoliation, you should have little trouble in spotting the defoliation telltales in this subscene enlarged from the Landsat-1 MSS image that includes Harrisburg (just to the south):

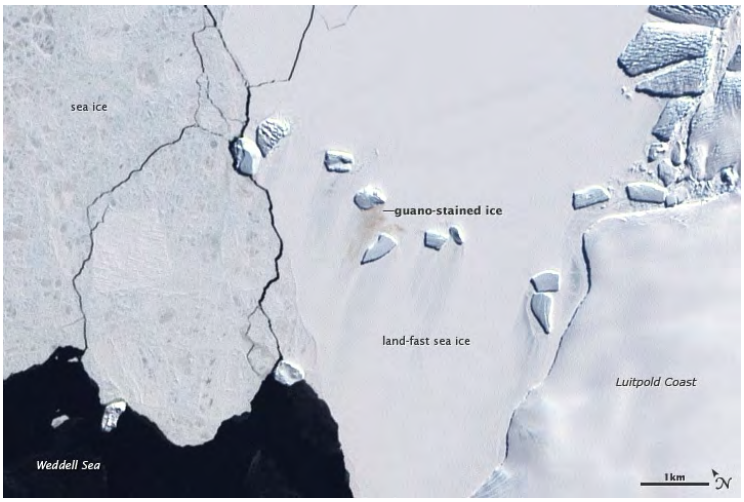


On page 3-5a, we examined in detail a widespread defoliation threat to pines in the western U.S. One more example: The Gall Wasp, from Asia, has reached the Hawaiian Islands by "hitching a ride" on ships. It has a 'taste' for the coral-colored leaves of the Wiliwili tree (sp.: Erythrina) which is found on several of the big islands. In this Landsat image of Maui, yellowish areas are defoliated Wiliwili trees. Both short- and long-term imagery, from various satellites, afford a means of observing the spread of wasp damage and aid in selecting areas for spraying.

Courtesy: David Cornwell



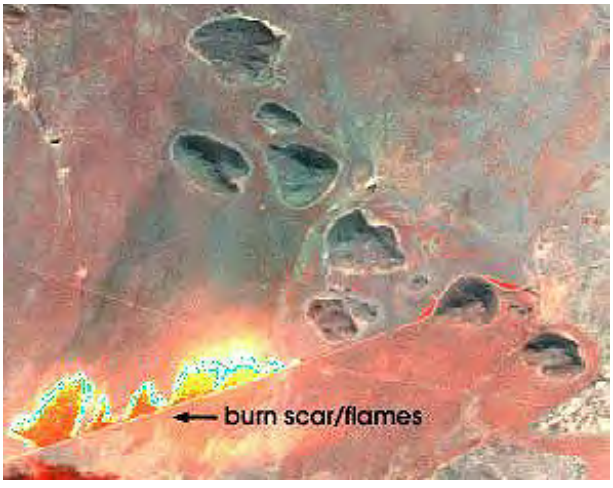
This next pair of images calls attention to a real oddity. An item on the Internet stated that space imagery was capable of finding "penguin poop" (the proper name for this is "guano"). It seems that penguins habitually pass their feces in areas that can cover acres or more, and they seem to choose a confined area rather than dump anywhere. The waste appears reddish brown on the ground, which shows up in contrast to the light blue-white ice in high resolution satellite images. Here is a "bathroom" for Emperor penguins, seen first from ground level, then from space:



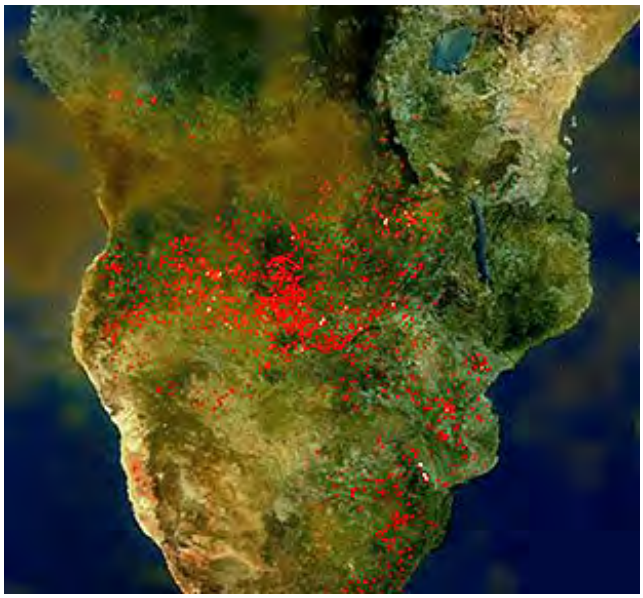
These observations are not inserted here merely to be facetious. Locating the penguin "poop parlors" helps environmentalists to monitor the location and migration of the penguins.

Many ecological problems (some are actual disasters) occur within very short time frames. A prime example is vegetation fires in grasslands and forests. Monitoring from space is especially well suited to watching these fires as they occur and after they are extinguished to gauge their effects. Both those set purposely for beneficial purposes and those which are unwanted are proper subjects of this surveillance.

In the first category, controlled burning (including the cut and slash approach to management of crops and grasslands) on a grand scale is especially commonplace in the African savannah. We show first a pair of images made by the MODIS Airborne Simulator (MAS) (see Section 16); the one on the left simulates natural color and on the right depicts the ability of using longer wavelength infrared to penetrate the smoke to see the actual flames.

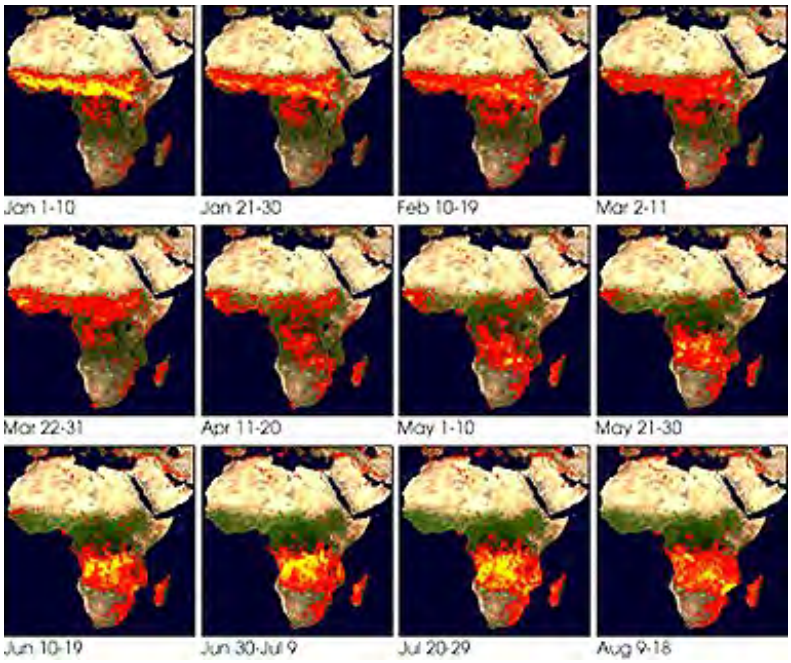


In September, 2000 much of south-central Africa was covered by major controlled fires. The image below was made using a combination of registered imagery from the NOAA-14 AVHRR and the TOMS sensor (see page 14-9).

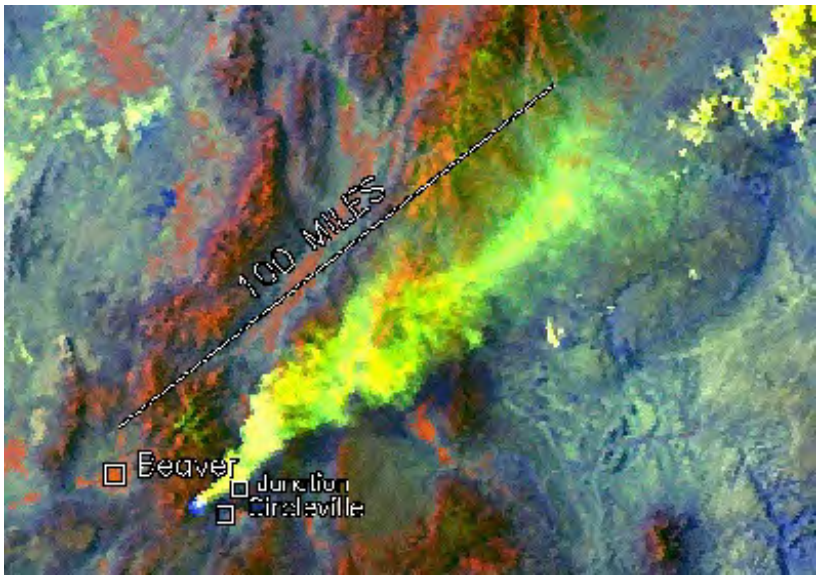


In this next higher resolution image made by Aqua's MODIS on May 6, 2004, more than 4000 individual control-burn brush fires in the Democratic Republic of Congo and Angola are underway (not all visible in the image below but countable in a transparency made from this source):

This mode of agricultural management is widespread during the dry season in Africa, but peaks at different times in various parts of the continent. MODIS has taken a series of observations at 2-week intervals during 2005, with these results (the solid reds and yellows indicate regions of widespread burning, being subcontinental in scope):

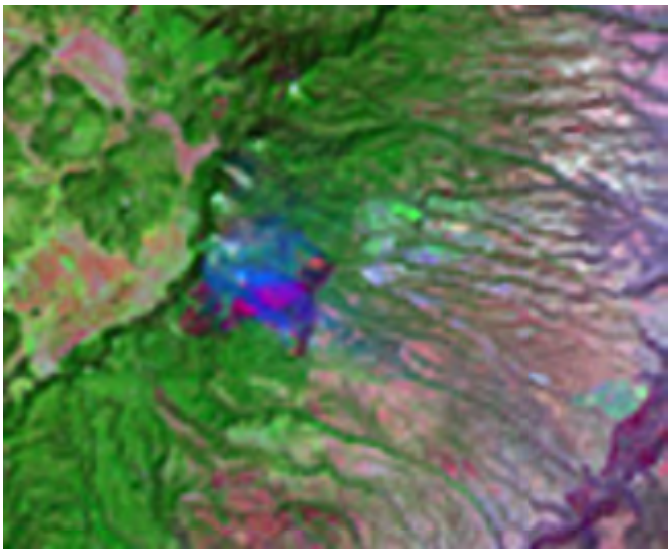


Forest fires usually burn for a few days to several weeks (especially in isolated areas) until brought under control. A fire in the Fishlake National Forest near Beaver, Utah (about 60 km [40 miles] ESE of the White-Mountain scene you will examine in Section 5), was imaged by the NOAA-9 AVHRR (1 km ground resolution), on June 17, 1996.

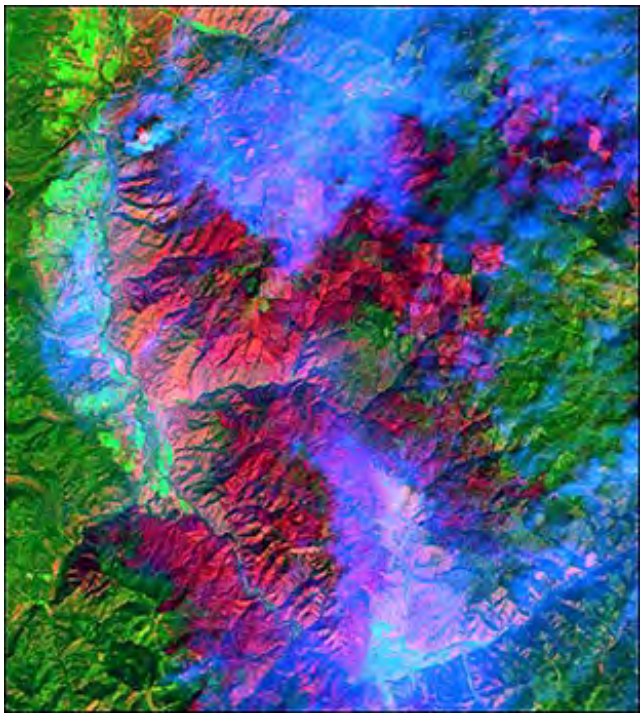


This false color composite was made by assigning the channel (a synonym for band) 1 image to red, channel 2 to green, and channel 3 to blue. In this version, the grayish smoke shows as yellow and the fire, at the base of the smoke column, appears as bright blue. Major fires are commonly imaged by meteorological and land-observing satellites, and the extent of the burn damage is easily assessed afterwards by the dark patterns in the visible bands. These dark patterns are usually evident as bright patches in thermal imagery because of the blackbody effect (see page 9-1).

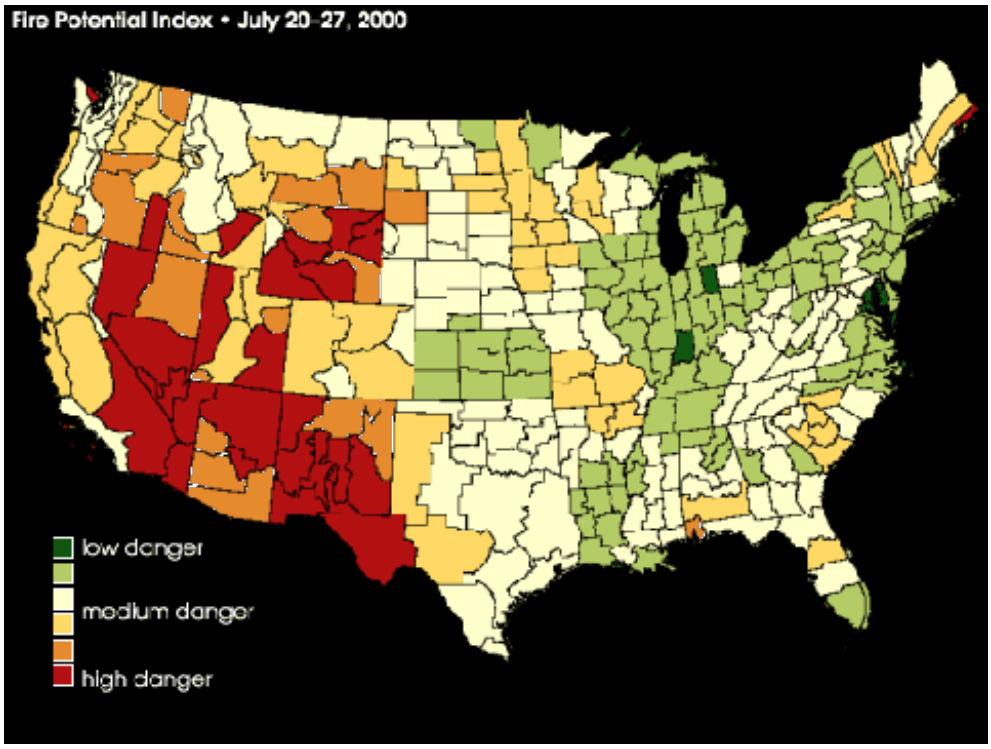
The great wildfires of 2002 in Arizona, Colorado, and elsewhere in the western U.S. were examined in the Overview. In 2000, a major fire in New Mexico consumed more than 46000 acres and destroyed 260 homes in Los Alamos, the town supporting activities at the Los Alamos National Laboratory where technology dealing with uses of nuclear energy (including the nuclear bomb which was developed there in 1945) is the principal industry. Landsat 7 captured spectacular views of this fire. The image below was made from TM Band 2 = blue; Band 4 = green; and thermal Band 6 = red. The dark red spots in the resulting scene are actual flames or very hot burn areas.



Fire outbreaks continued through the summer of 2000 when at one time more than a million acres in 7 western states made this the worst forest fire season in more than 50 years. In mid- to late-August, the most widespread burning was in Montana. This Landsat TM scene shows mountainous terrain whose forests are on fire at several places; the already burned areas are in red.

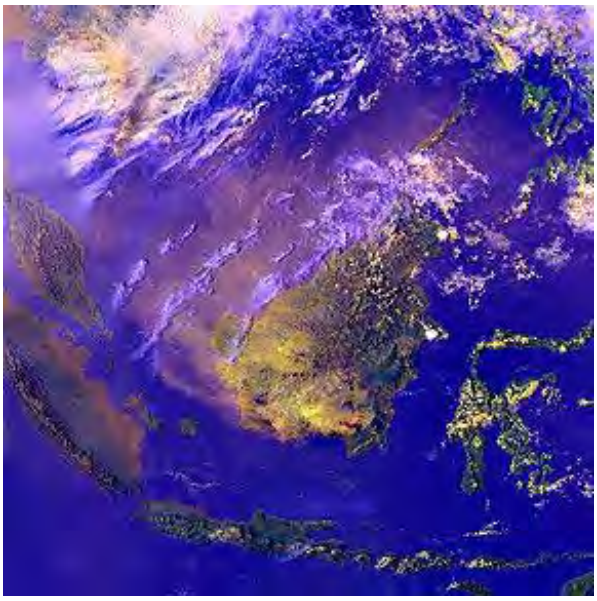


Much of the U.S. has been experiencing mild to severe droughts since the late 1990s. Satellites can not only sense fires that break out but can monitor conditions that indicate the likelihood of future fires. The Fire Potential Index (FPI)uses data from meteorological and other satellites to measure Soil Moisture, which is a predictive indicator of stages of dryness in western forested or grasslands areas. Most fires there start from lightning strikes; some are manmade (carelessness or deliberately set. The calculated FPI for a period between July 20 and 27, 2000 is shown below:

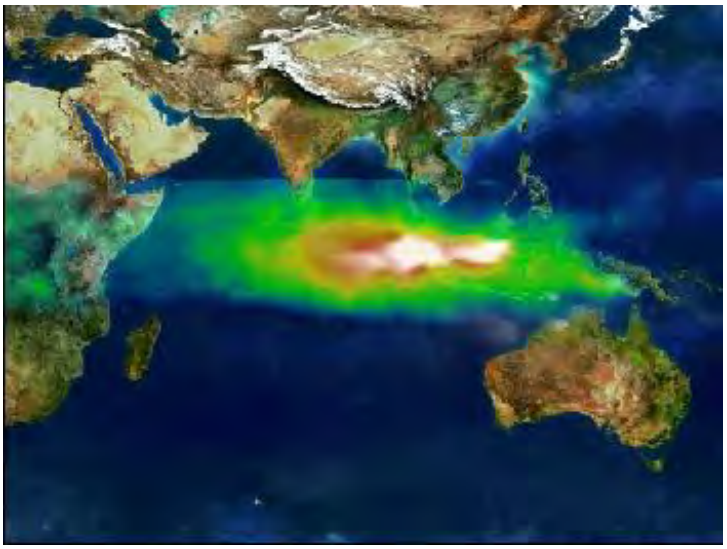


Forest fires are common targets of opportunity for astronauts and cosmonauts during their photo sessions because these are usually so obvious when seen from space. Here are fires in the tropical north of Australia:

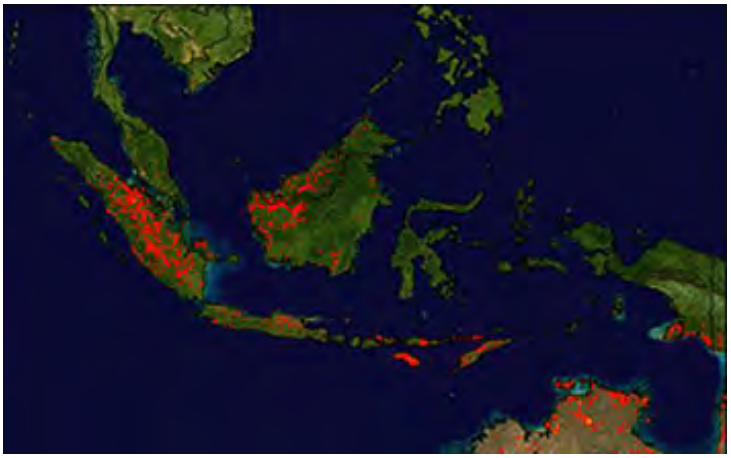
On a wider scale of view, forest fires in three separate areas, one in Borneo, north of Java, and two in Sumatra, in 1997 were imaged by NOAA-14 (at a resolution of 4 km [2.5 miles]), as seen here:



Air pollution on a grand scale is often easily visible from space. An imposing example is the mixture of smoke and smog created by fires spread over many islands in Indonesia. Using imagery acquired by TOMS, and set against a backdrop developed from NOAA data, a huge smoke plume (white), mixed with smog (colors represent variations in ozone amounts) is seen to be heading westward from the islands across the Indian Ocean.

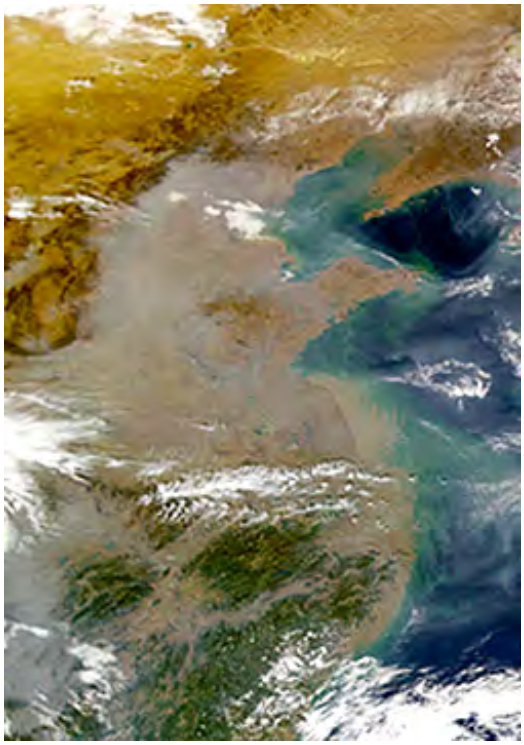


Indonesia is a country that practices slash and burn. This is evident in this MODIS image, in which some fires are natural, others manmade:

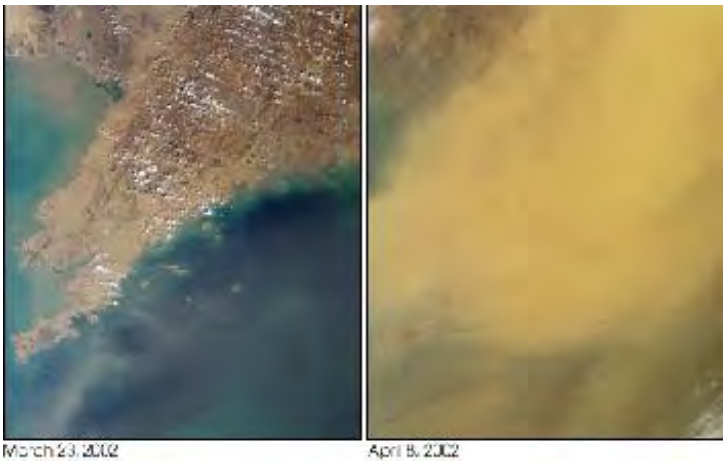


China has had many problems with widespread pollution owing to its rapid transition to heavy industry and to increased automobile usage over much of the country. This next image, taken by

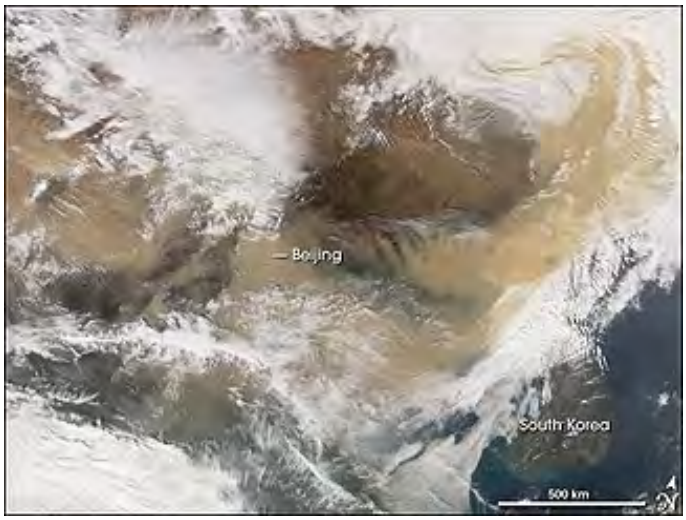
SeaWIFS (page 14-13), shows a brown haze covering much of northeastern China to the extent that land surface features have been blotted out. Many of China's huge cities, such as Shenzhen and Beijing, are becoming notably unhealthy for their inhabitants. Pollution from China spreads over much of SE Asia and Japan, and has been traced as far away as western North America.

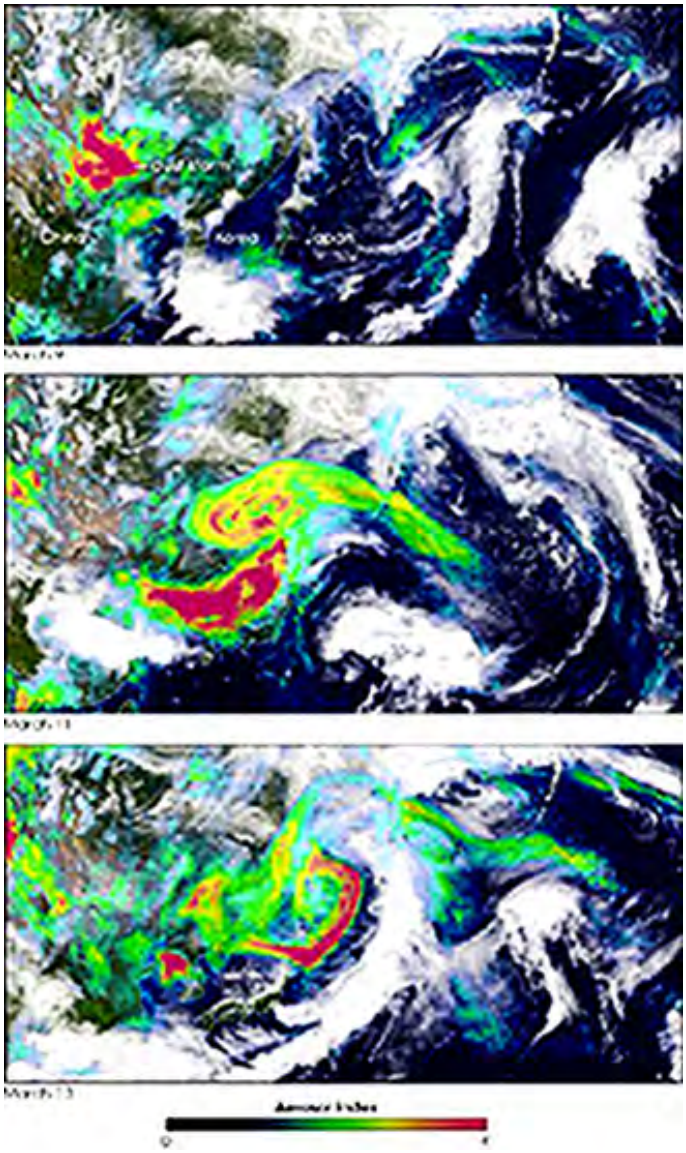


Natural air pollution from intense dust storms also besets China from time to time, especially when winds remove fine materials from the Sinkiang and Mongolian deserts. This MISR image pair shows a part of eastern China near the Korean border; on the left, a relatively clear day and on the right the almost total obscuration from a thick cloud of dust:

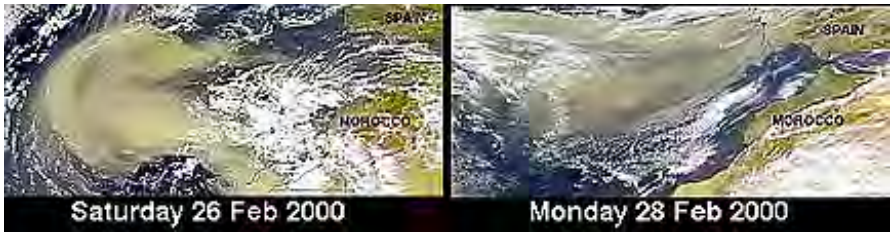


A major dust storm, coming from the Gobi Desert, passed over much of northeast China in mid-March of 2006. Here it is in an Aura MODIS image taken on March 9. Some snow that fell during this passage actually had a yellowish tone owing to intermixing with the dust. Beneath the image are three views of its progress (follow the red patch), in which Aura' OMI (Ozone Monitoring Instrument) allows variations in cloud densities to be assigned colors.

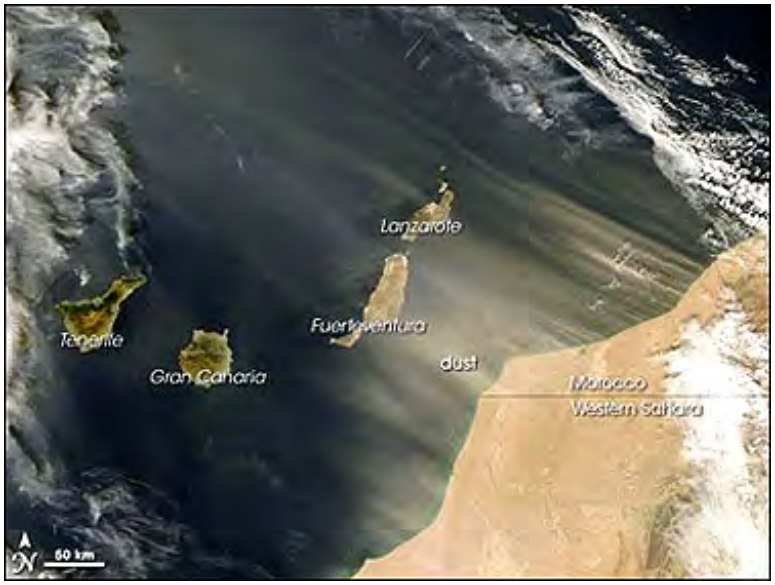




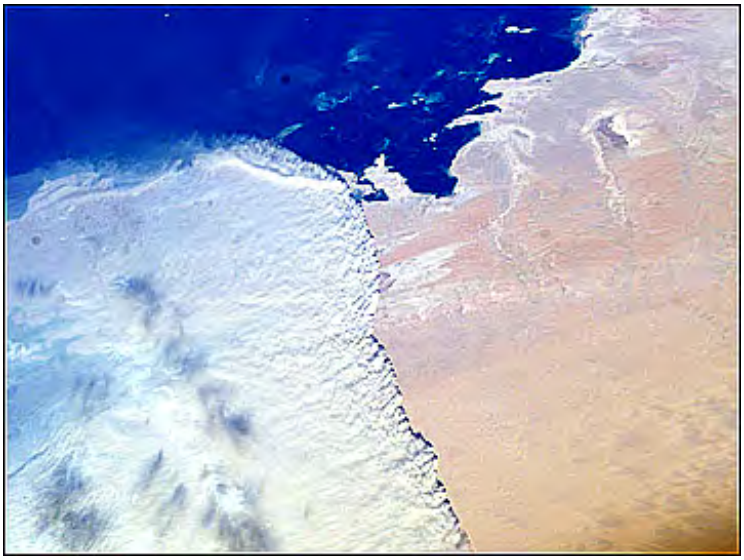
Dust storms are also common off the Sahara Desert of Africa, as is displayed by this pair of images taken two days apart:



This Terra image shows a strong dust storm on November 11, 2006 coming off the western Sahara and extending out to the Canary Islands.



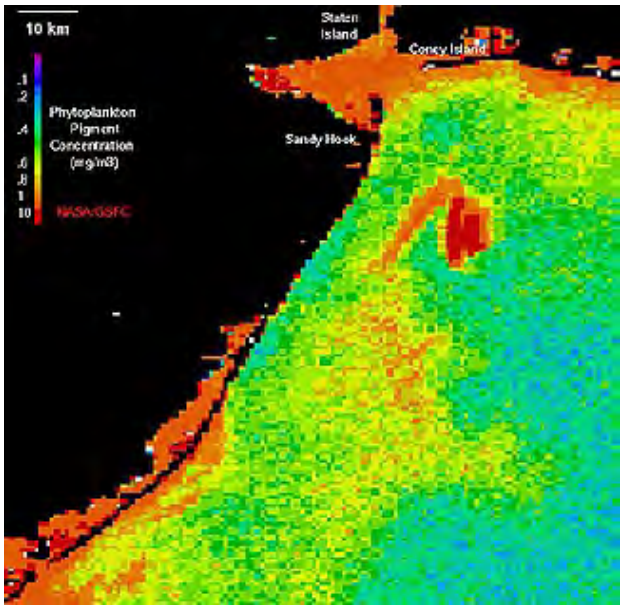
Sometimes strong sand storms can be imaged in the act of their formation and movement, especially in the deserts. Astronauts aboard the International Space Station caught this storm as it advanced on February 15, 2004 across the desert of Qatar along the Gulf of Arabia.



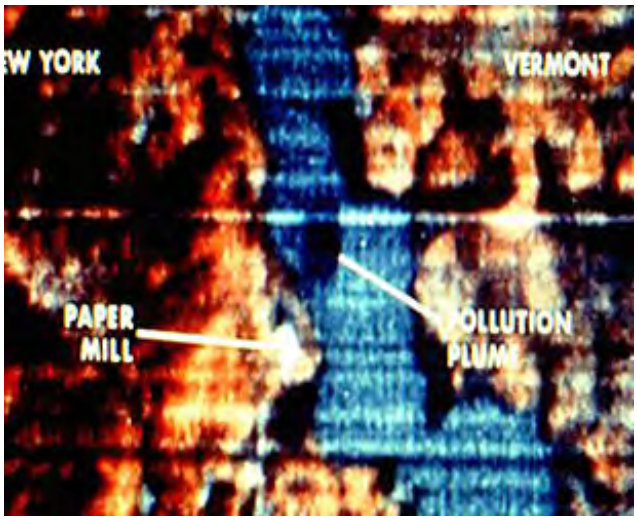
Another example of air pollution was detected just two months after ERTS-1 was launched. The area affected included Goddard Space Flight Center the operational home of the ERTS-1 satellite. On a particular day in September the writer remembers a severe smog (pollution) alert for the Washington and Baltimore region. Later, in checking through some of that early imagery, I found the image shown next that dramatically revealed the extent of pollution on that day. The area shown is east of these cities in the Upper DelMarVa peninsula (mainly Maryland's Eastern Shore); the loss of red is due to the thick blanket of smoke (in white, with light yellowish tint) above the dominantly agricultural and forested terrain.

A more localized form of air pollution derives from SO_2 and other gases released from smokestacks onboard ocean vessels such as freighters. This condenses along with water vapor to form cloud trails similar to those made by jet airliners. Here is such a phenomenon seen over the Pacific Ocean:

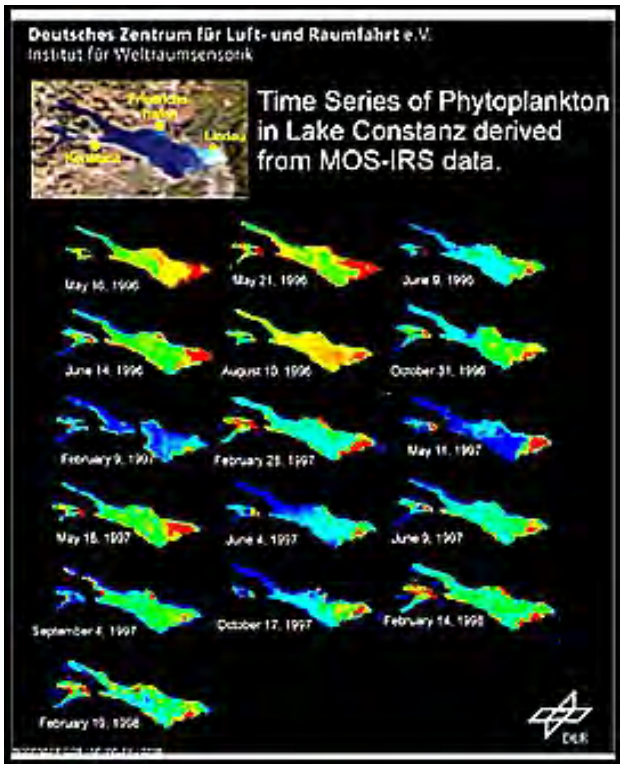
Turning now to water quality issues: Two telling examples show water pollution in remote sensing images. The image below, made by the Coastal Zone Color Scanner (CZCS) on Nimbus 7 (see page 14-13), includes the shallow shelf region of the Atlantic Ocean in the Long Island Bight, south of New York City. For decades, a train of barges has carried waste materials (euphemism for garbage) several times a week to an offshore dumping site. The barges move in a zig-zag pattern that leaves a distinctive curlicue pattern of pollution, which is readily apparent (the orange pattern in the water) in this color image.



Other types of water pollution can be detected from space. One of the first demonstrations of this is the example of waste pulp and hot water from a paper mill being dumped (illegally) into Lake Champlain from a New York site. Its effect was to reduce water wave action. The ERTS-1 subscene shown below was used in a legal court case in which a judgment against the mill owners was won by the State of Vermont. This space imagery application was given much attention in the newspapers.



Lakes are also subject to natural variations which affect their "health" and quality. Phytoplankton are algal-like plant that can thrive in freshwater lakes, changing in quantity with the seasons. The Modular Optoelectronic Scanner (MOS) on India's IRS-P3 satellite was used to monitor amounts of phytoplankton in Lake Constanz (also known as the Boden See) along the Swiss-Bavarian border (in northeast Switzerland) for several years, as seen in this succession of images (time-series) from May, 1996 through February, 1998:

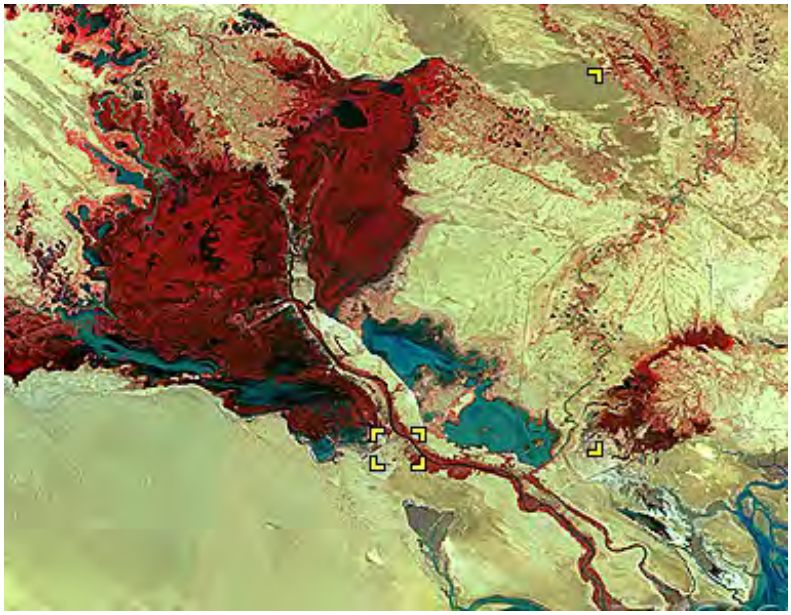


One of the major concerns about humans' destruction of the environment is the alarming reduction in wetlands (see page 3-5) - marshes within landmasses, particularly those associated with rivers or near the coasts, that are ecological habitats for various plant species and many specialized animals (particularly birds) adapted to living in these watery conditions. Extensive wetlands tracts occur at ocean/land boundaries. One of the best known in the U.S. is wetlands and bayous along the Delta country where the Mississippi River meets the Gulf of Mexico, as seen here:

Another major wetlands is found behind the Outer Banks of North Carolina (see page 17-4 for an image that includes nearly all of this landform) which is a classic example of a regional offshore bar. This astronaut photo shows the water bodies behind the oceanic sand bar (Bodie Island). Note that a second, older bar (which once was the main barrier island) lies to the west; agriculture has taken hold on this land as wind has removed some of the sand leaving silty soil. The widest expanse of wetlands is the area around the words "barge wake":

">

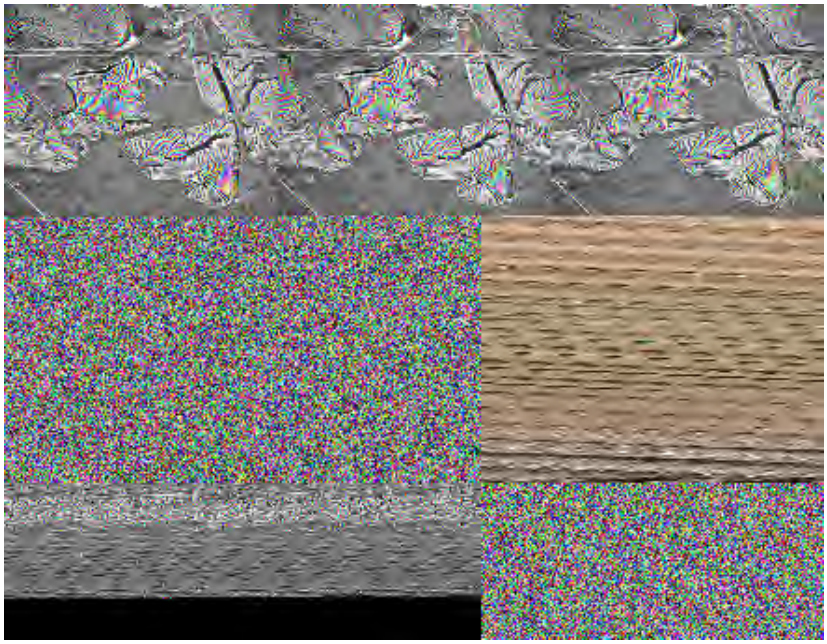
Wetlands are often deliberately destroyed to recover land for agricultural development or for home building. One of the most famous wetlands - the marshes of Mesopotamia, in Iran and Iraq which were a key factor in the rise of cultured peoples in the "Cradle of Civilizations" - has undergone drastic draining and dessication in just the last 30 years. Below are montages made from Landsat imagery in 1972 (top) and a 1997 RESURs image (bottom) of an area where the Tigris and Euphrates rivers meet around Shatt-al-Arab. Damming of the rivers and deliberate diversion of their waters have reduced the marshlands that once covered more than 20000 sq. miles to now less than 4000 sq. miles.



This land reclamation was done largely to increase available land use for Iraqi development around Basra and the port of Umm Qasr where oil from the Rumaila fields is shipped to tankers plying the Persian Gulf. When the second Gulf War took place, it was feared that the Iraqi forces would blow up dams upstream to reflood the marshes, possibly affecting populated areas. This did not happen. Here is a satellite view of central Basra, Iraq's second city (about 1,000,000 people):

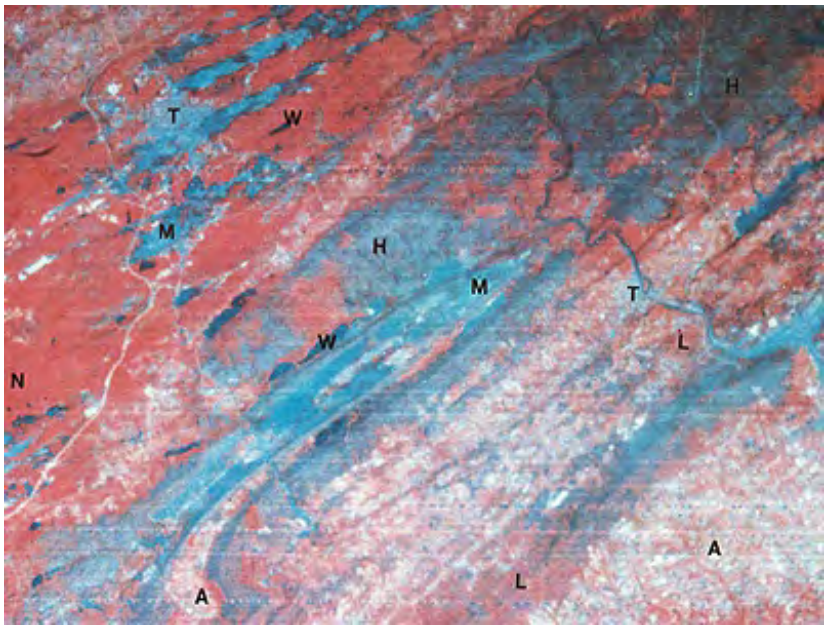
After the war, and with a new Iraqi government in place, the decision was made to restore the Mesopotamian wetlands near Basra, as this should also improve ecological conditions upstream. These two MODIS images tell the story: In 2002, the wetlands were nearly gone but by 2007 they had been revitalized to conditions prevalent at the start of Saddam Hussein's policy of draining the marshes:

Similar to our clearcutting examples already examined above, satellite systems can follow long term changes in resource use, such as strip mining and the progress of land reclamation (as you noted in the Pennsylvania case study in the first Exam). Strip mines often produce a very distinctive pattern, as in this aerial photo that shows the bench method of strip mining in a near-surface metallic ore extraction operation.



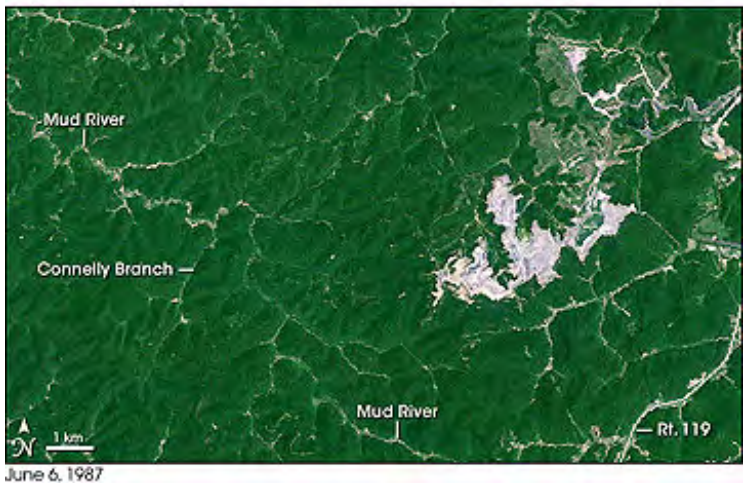
In standard false color images strip mines in coal are often indicated by a distinct signature (bluish to bluish-black) that represents the barren exposed surfaces with a residual covering of coal lumps and dusts causing the ground to appear dark. This is the case in this Landsat subscene in the anthracite belt of northeast Pennsylvania (part of the image shows lighter reds

indicating some reclamation):



This is evident again in the image below, which shows strip scars (light to medium blue), replanted areas (red stripes), and farmland in the coal measures of western Maryland (Allegheny Plateau). The image was made by the Linear Pushbroom Radiometer (LAPR) developed at Goddard Space Flight Center as a prototype to test the CCD array concept and flown on a NASA aircraft.

Satellite imagery is particularly good at monitoring changes over years in the progress of strip mining - both in expanding development and in reclamation. Note the differences between 1987 and 2002 in the Hobet mine on a plateau mountaintop in West Virginia:



June 6, 1987

May 22, 2002

As a transition image to the next Section, the TM scene below shows an urban area near Leipzig in Germany, surrounded by numerous large fields, both fallow and with diverse crops, and several large white areas (those being clouds have associated shadows) which are the unfilled and poorly reclaimed scars of major surface pits from which soft coal was dug out for heating and power production in the past.

Elsewhere in Germany, in North Rhine Westphalia, this ASTER natural color image shows three

very large strip mines:

We will illustrate other agricultural, forestry and ecological applications in this Tutorial (several appear in Sections 6, 13, 16, and 17). The text will introduce more of the possibilities that space imagery affords in monitoring and inventorying crops, grasslands, forests, wetlands, and other ecological niches. But now, let's examine ecological and other forms of damage brought on by geological events.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Volcanoes can be very destructive to inhabited areas and to natural land cover when their eruptions are widespread and violent. This will be illustrated with an in-depth treatment of the famous (to Americans) eruption of Mount St. Helens in 1980. Several other volcanic eruption consequences will be briefly examined. Also considered on this page are observable effects from strong earthquakes as monitored by ground and space sensors.

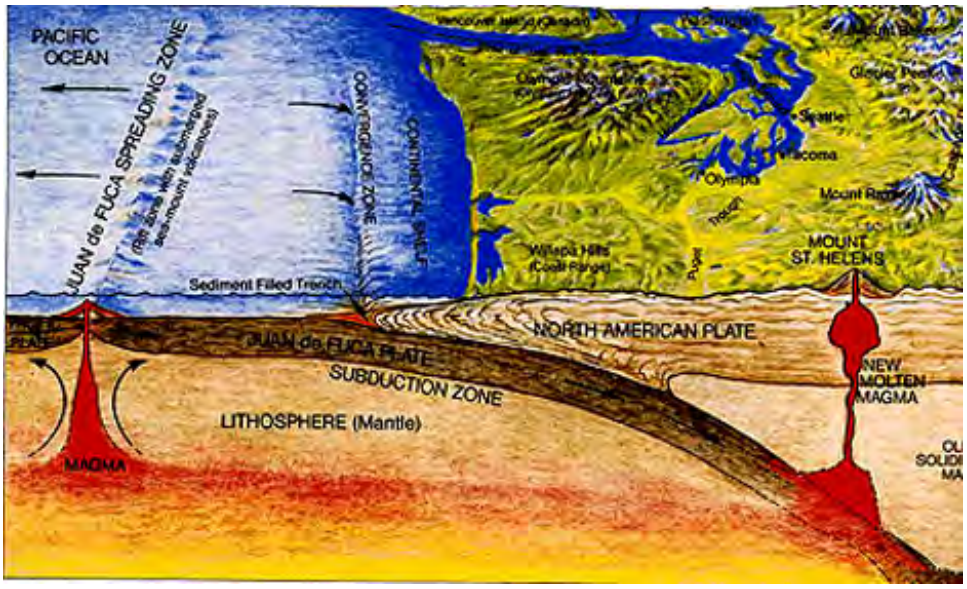
Ecological Damage from Natural Events

Geological Events

Natural catastrophes of a direct geologic nature, such as volcanic eruptions and earthquakes, are a much feared and often unavoidable calamity to people living in proximity. Although hurricanes have caused damage into the many $billions, potentially a major earthquake in a metropolitan area such as Los Angeles can be even more costly. We will review how remote sensing contributes to monitoring pre-event conditions and assessing post-event damage on this page.

Most of the page will be devoted to one event that has captured the American imagination and still leads to fears and concerns about repetitions. This is the major eruption at Mt. St. Helens in Washington State in May of 1980. We will go into enough detail to qualify this page as a Case Study.

Mt. St. Helens (to which we will usually refer hereafter as MSH) is part of the Pacific Ring of Fire, where stratovolcanos of (most commonly andesite-dacite composition) form on the upper tectonic plate at a subduction zone, as indicated in this diagram, which suggests that frictional heat during the underthrusting causes melting that rises upward as magma and, if the melt reaches the surface, as volcanic structures whose repeated eruptions over millions of years leads to the distinctive cones that are among the most impressive of all mountains.



MSH is one of the major volcanic cones making up the Cascades that start in Northern California and continue just into British Columbia. The volcanoes are a direct and normal consequence of classic subduction, in this case of the small Juan de Fuca plate caught between the Pacific and

North American plate. These volocanoes are shown in this map:

Until 1980 there have been no major eruptions of the Cascadian volcanoes since colonial times. MSH last erupted about 4000 years ago. Mt. Hood, Mt. Rainier, and several others have erupted in the last 10000 years. Mt. Lassen had a small eruption in 1915.

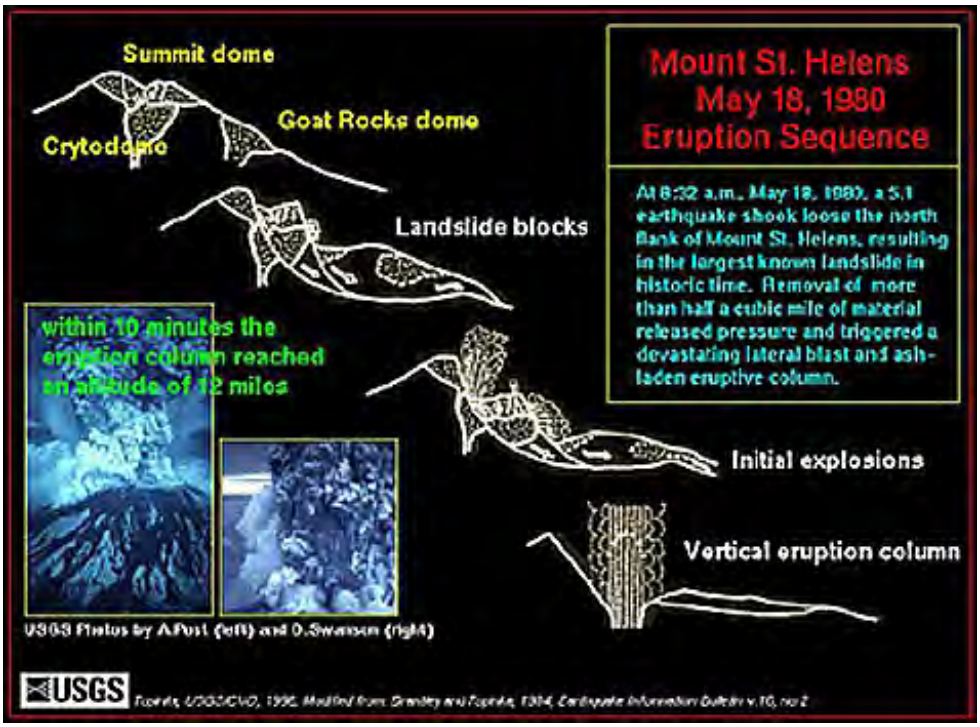
MSH lies to the west of the main line of stratocones, as shown in this 1973 Landsat image.

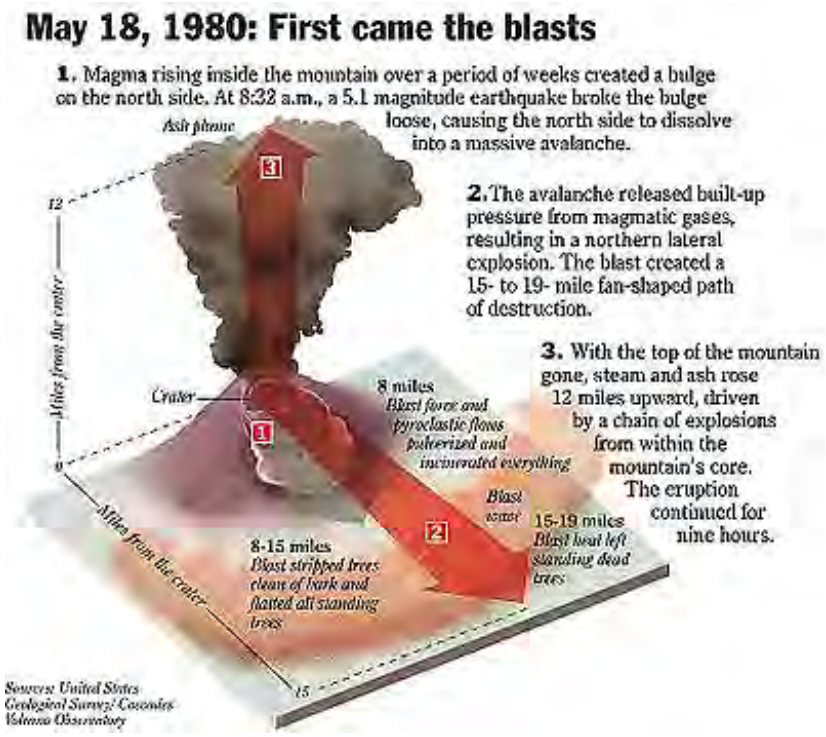
MSH had always been praised as one of the most "handsome" stratocones in the Cascadian chain - some considered it comparable to Mt. Fujiyama in symmetrical beauty. Below is a ground photo and an aerial photo of this mountain, which in pre-1980 times reached an altitude of 2970 m (about 9800 ft) above sealevel.

Although MSH had undergone several minor steam and ash explosions in the 1800s, it was considered by most volcanologists to be inactive or at least quasi-dormant. But in March of 1980, seismic records showed a marked increase in small magnitude (3 to 4) earthquakes, and some steam venting. Thermal measurements suggested a near surface warming. An old, small dome near its top began to rise an average of 1-2 m per day. Volcanologists interpreted these signs as indications of some moderate level activity that might occur that year.

The majority opinion of those at the Cascades Volcanological Observatory (CVO) favored only insignificant activities with a low likelihood of any dramatic eruption. However, in mid-May a warning was issued for voluntary evacuation of the region around the volcano as a precaution. Many campers, and the soon-to-be famous Harry Truman (yes, that was his name) of Spirit Lake, ignored this since it was not mandatory. There was a brave hero, trapped in the advancing lateral blast: Dr. David Johnston of the USGS was a forward observer who radioed what he was seeing until buried by the debris. The U.S. Geological Survey had several other observers stationed near the volcano on Sunday morning, May 18, 1980, and a light aircraft was circling the structure at that time.

Before documenting individual stages, we will describe narratively the sequence of the first minutes of eruption, using these two diagrams and suggesting that the reader examine the table at this CVO website.

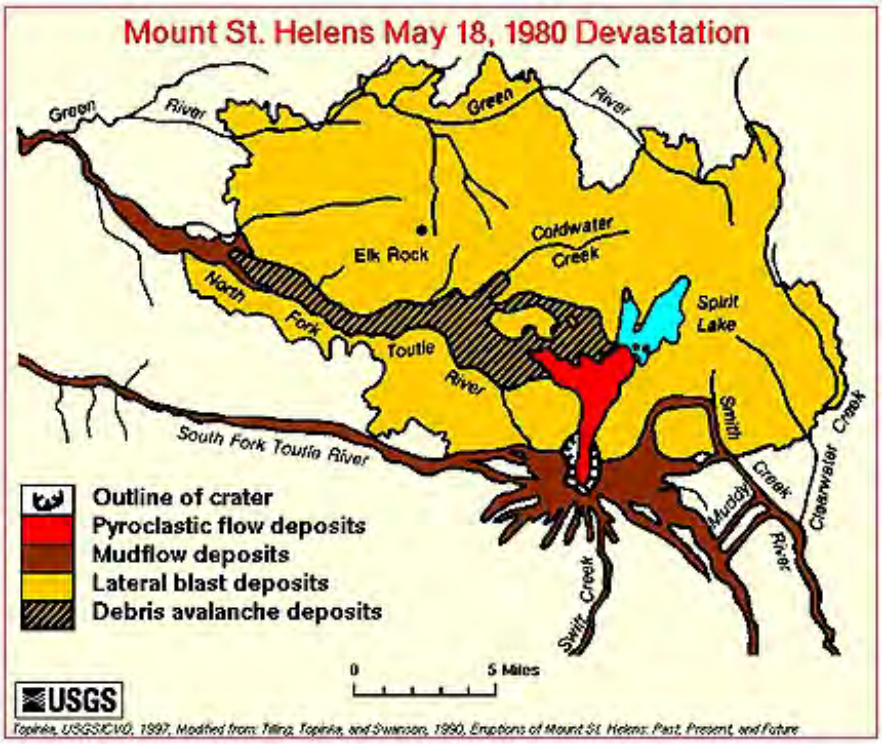




The ground around Great Rock Dome failed first, followed quickly by detachment of rock below it as a debris avalanche of blocks up to house size. This released interior pressure which expelled material as a lateral blast and a surge of ash, blocks, gases, and steam that moved northward at speeds up to 1000 km/hr. The surge moved indiscriminantly over rough terrain out to 12 km, carrying debris for the next 30 seconds into the North Toutle River and covering part of Spirit Lake (which retains both the body and spirit of Harry Truman). Air overpressures flattened trees up to 18 km from MSH's peak. Over the next ten minutes the mountain, after being lowered 400 m to form a huge ampitheater (much like a breached crater) 2 by 3 km wide and 680 m deep, continued to expel tephra (ash and pumice) mostly vertically up to an altitude of 18 km where winds carried the fine-sized particles for great distances to eventually reach the ground as ashfall. On the first day much of the ash mixed with melted snow to produce lahars (volcanic-particle mudflows), some being channeled into rivers, choking them for many miles. Pyroclastic flows continued intermittently for days, along with vented steam that mixed with this ash to generate more mudflows carried into the North and South Forks of the Toutle River. We shall document these actions in illustrations below.

The blast in particular was fatal to 57 people nearby. All said, 540,000,000 tons of ash came to rest over 55000 km^2. Ash beyond MSH reached locally up to a meter thick and to a few centimeters in Yakima, Moses Lake, and other towns in central Washington. Small eruptions continued through July 1980 (the writer and family reached the area in June but missed seeing any of these events) and sporadically through 1986. MSH, having done more than $1,000,000,000 in damage, quickly became a major tourist attraction in the Pacific Northwest.

Now, let us document visually much of what was described above. We start, for reference, with a map produced by the U.S. Geological Survey, of the MSH area, showing the major deposit types:



The great eruption began at 8:32 AM {PDT), probably triggered by a 5.1 magnitude earthquake, causing an abrupt failure around the bulge, as it seemed to tear loose into the debris avalanche, even as the first ash was expelled from the peak, a moment captured on film.

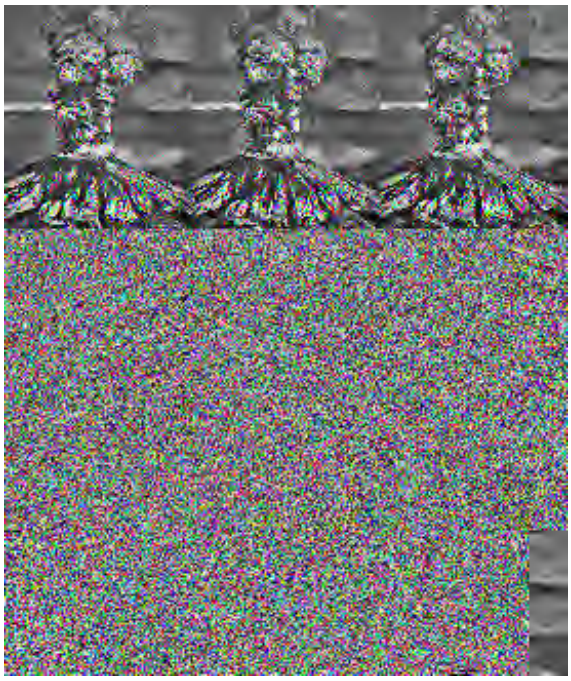
Only a few seconds later, we see the lateral blast as it blew out the side of the mountain, initiating the collapse of the peak.

The effects of that blast were widespread, causing trees to be stripped of branches and closer in uprooted to lie prone on the ground.

The next pair of images focus on the modification of the ash and debris into what is called in its French terminology a "nuees ardente" or "glowing avalanche". A synonym is "pyroclastic flow" - entrained tephra (the clastic particulates such as ash and pumice) propelled and supported by steam and other gases - that moves downslope under the influence of gravity. This is an airborne avalanche of hot, thick volcanic debris composed, in the case of MSH, of material from the escaping lava mixed with now fragmented parts of the pre-blast volcano wall.

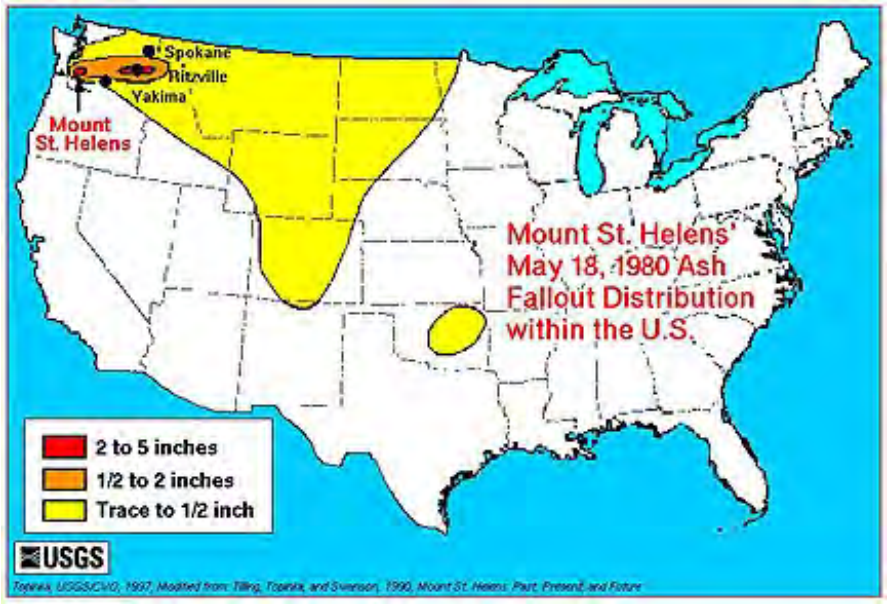
This post-eruptive photo shows some of the components present in a pyroclastic flow at MSH:

Over the next 10 minutes, MSH sent ash vertically by means of thermal uplift of heated gases into a plume that rose to a height of 18 km before starting to disperse downwind. One of the most famed photos taken of this is shown here. (The picture is hanging in my study just to the right of the computer desk; I have seen it attributed to Austin Post or to Robert Kimmel). Note the dark flows which are not lava but mudflows (see below):

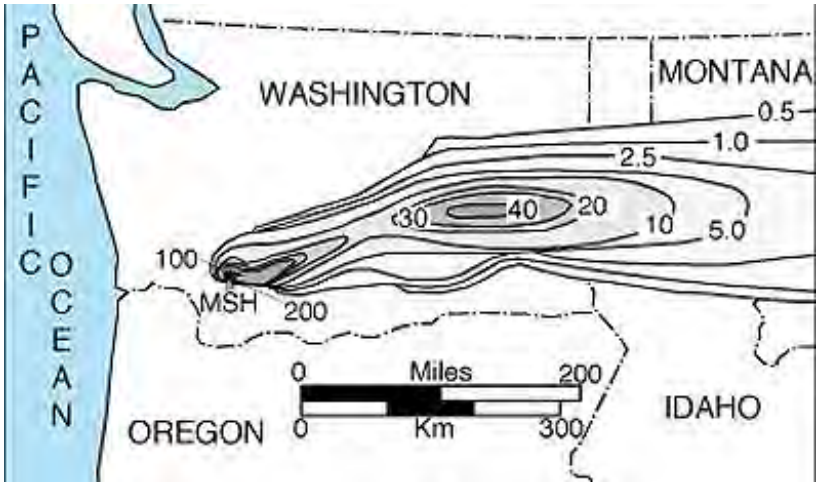


The plume continued to rise throughout the day. This next photo, in color, shows the curtain of falling ash as highlighted by the sunset glow.

The ash was widely dispersed, mainly to the east, as shown in this ashfall map.



Another map shows the thickness (isopach map) of the main lobe of ash.



Streets and lawns in towns east of MSG, such as Yakima, received up to 6 cm of ash, that was removed from roadways, as shown here, but eventually was washed away by rain or carried into the soil.

Looking from north to south towards MSH, the nature of the debris avalanche deposits is evident from this ground photo

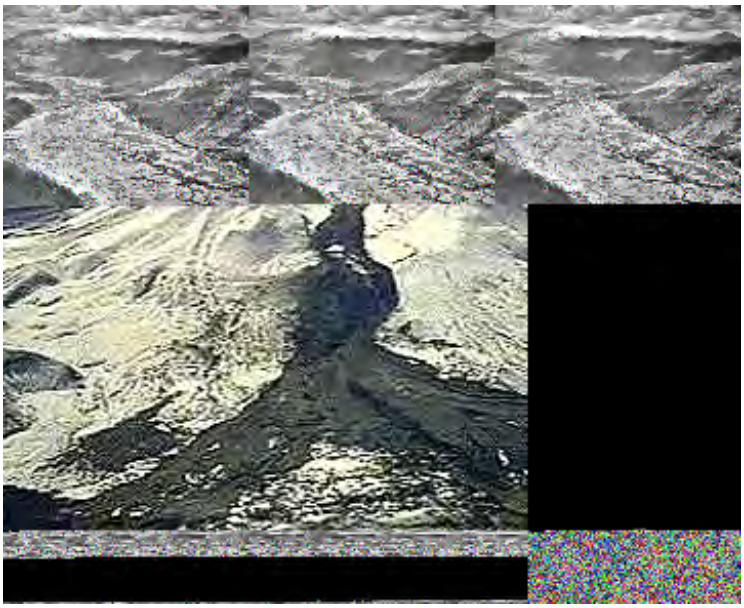
Deposits in the debris avalanche zone were hummocky and often in layers containing coarse boulders and smaller blocks:

Layered tephra - mostly of large blocky fragments - had this appearance from the ground near the volcano:

Mudflows (lahars) began to develop on MSH almost from the beginning and continued (occasionally) for years. The mudflow deposits can be both dark and light-toned.This photo shows a dark mudflow that happened in 1982

A light-colored ash flow that has entered a stream valley appears from above:

The North Toutle River was choked with ash debris and lahar deposits:



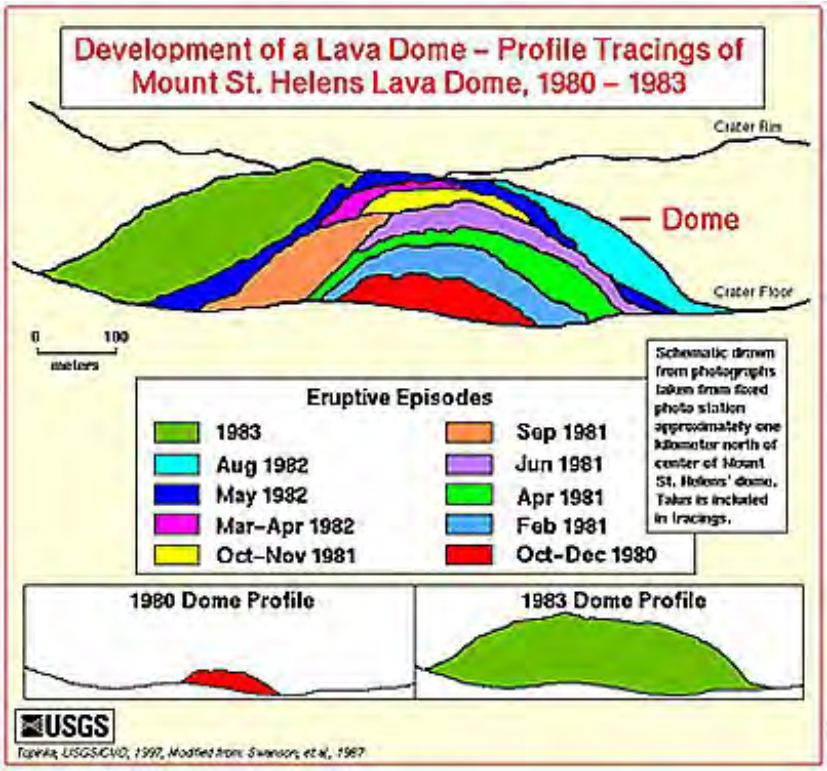
Upstream the lahars carried logs (precut at a timber mill) that jammed against a highway bridge. Beset by further mudflows and lahars this bridge was eventually crushed and buried.

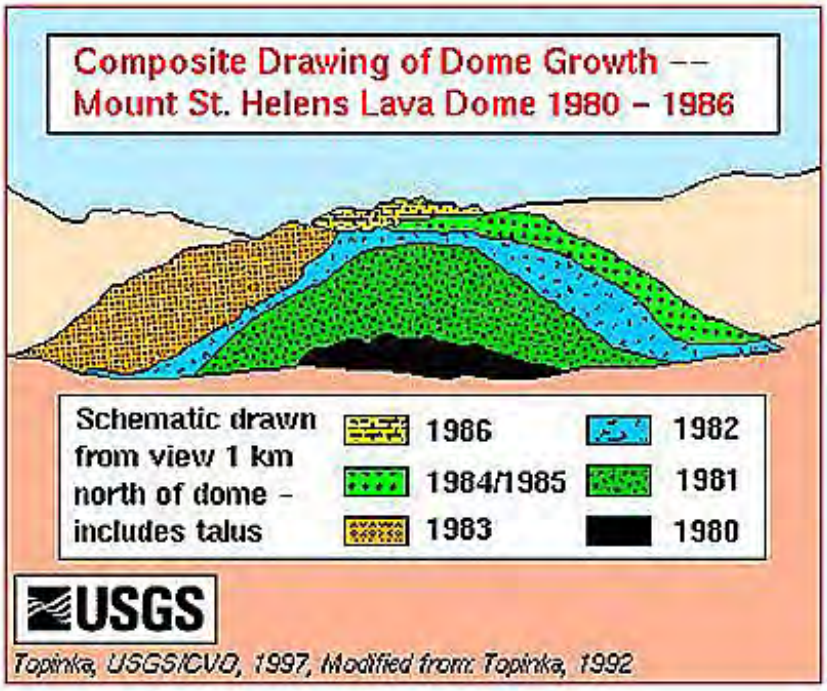
Further downstream another bridge survived:

Subsequent rain runoff has brought water into the Toutle River, causing stream dissection that produces channels.

Let us return to the center of the ampitheater where most of the volcanic activity was concentrated between 1980 and today. After the blowaway of the upper mountain, thick lava re-entered the crater floor as deposits that once more established a central dome composed of dacite (quartz-bearing andesite). A July 1980 view shows this dome; below it is a much more recent photo which indicates how much bigger the dome has grown:

Two diagrams depict dome growth: the upper covers the period from 1980 through 1983 while the lower diagram extends this to 1986.





The dome now is about 300 meters higher than its starting height and covers more than two kilometers at its base.

Before we turn to the role of remote sensing in monitoring MSH, let us take two more views of the post-1980 mountain. The first image was taken from the rim looking down into the ampitheater, the dome, the debris and pyroclastic deposits, partially exposed Spirit Lake, and the massive Mt. Rainier to the north. The second looks at the mountain from the south, where no obvious signs of an eruption are evident except for the loss of the peak.

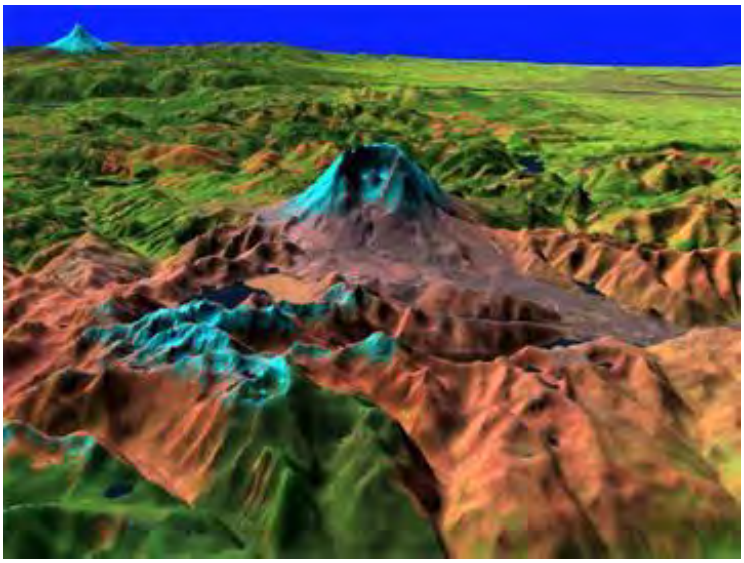
Since the full force behind the 1980 eruption was directed northward, one would expect a near absence of deposits and other signs on this south side. However, field work has found some

much older deposits from pervious eruptions of Mt. St. Helens that resemble debris avalanche deposits formed in 1980.

As you would expect, the eruption and its after effects can be effectively monitored with remote sensing tools by various satellites and aircraft. Just a month and a half after the May 18, 1980 eruption, a cloud-free Landsat-3 image was obtained that showed the depositional pattern.

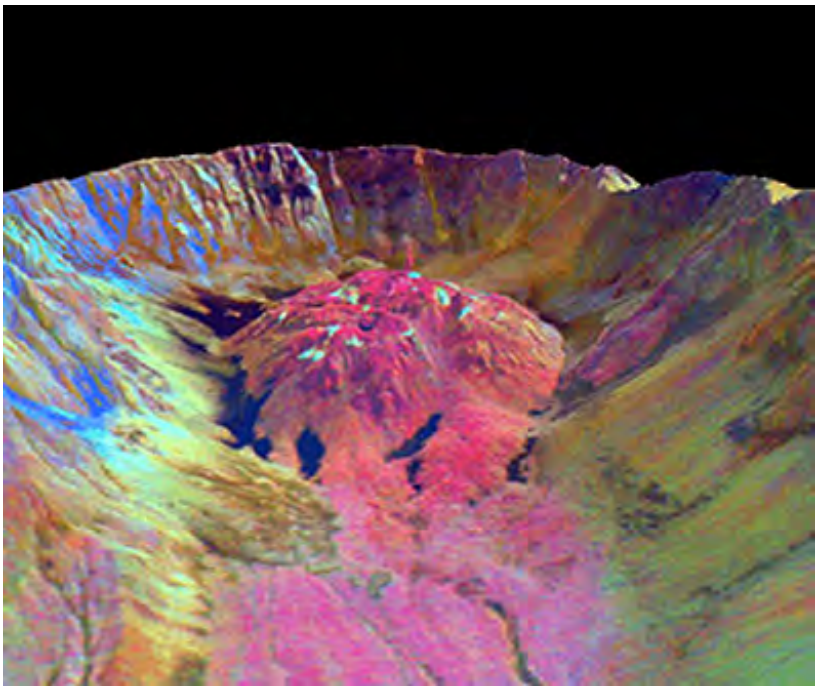
As soon as a computer-compatible tape (CCT) was made available to the writer (NMS), I and a colleague (Charles Bohn) worked well into the night on the IDIMS image processor at Goddard to obtain this next classification map. Compare the main features shown in different color, which we did not categorize at the time because of lack of familiarity with the category names now used, with those shown in the deposits map shown above. This exercise verifies that with computerized data of this nature, one can produce meaningful interpretations in a very short time.

Landsat imagery combined with DEM data produces this perspective (oblique or side-looking) view of Mt. St. Helens (middle center) and its surroundings; the red color helps to define areas of heavy ash fall now undergoing regrowth.



Compare the above perspective view made in the 1980s with this late 1980s version made using SRTM elevation data derived from two radar bands. The view here is to the southeast.

The central area of the excavated (blown-out) part of Mt. St. Helens, especially the extruded dome, remains hotter than its surroundings, as shown in this late 1980s thermal image made by the airborne TIMS instrument (see Section 9):



Repetitive Landsat coverage of MSH over the years has indicated that grasses and trees have gradually gained a foothold so that reforestation and other vegetative recovery is underway. These three Landsat images (1973; 1983; 1988) illustrate that:

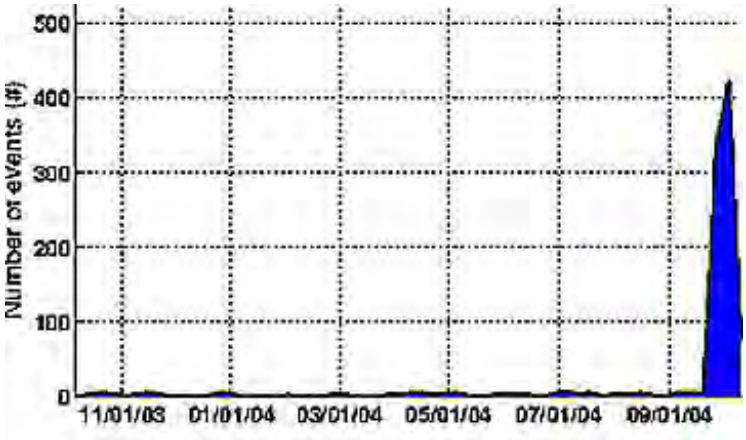
Ground photos taken in 1983 and 1998 at nearly the same scene show how trees have reestablished in the outer parts of the volcanic deposits, with ash apparently having decayed to early stages of soils.

The gradual plant regrowth in the ash deposits around the site has been followed over the years in Landsat imagery. Below we show a Landsat TM natural color subscene taken in August 1999. Although the gray ash is still widespread and dominant, careful inspection indicates various shades of green within the deposits where trees, bushes, and even grasses have made a substantial comeback in the two decades since eruption.

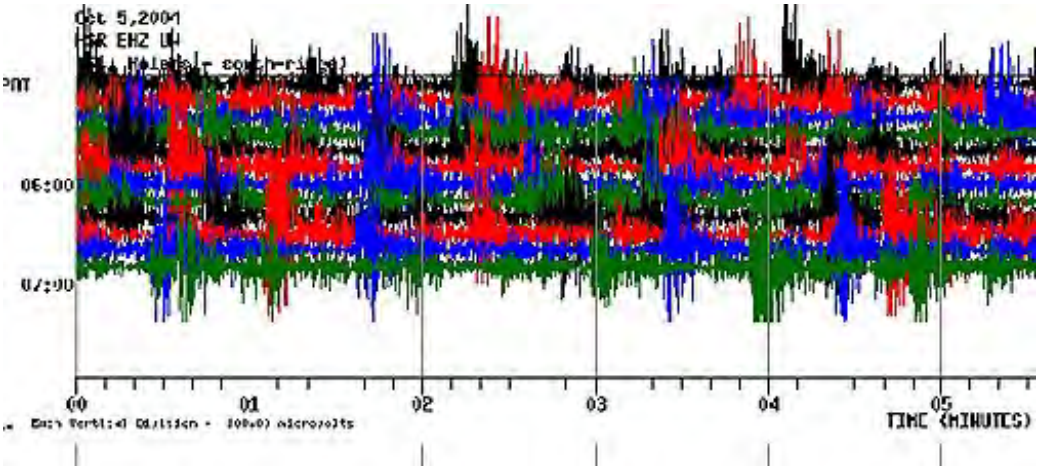
One of the last photos taken on October 1, 2004, by the astronauts on the International Space Station does not capture the still limited extent to which vegetation is returning to MSH. This photo is in near natural color:

By no means has MSH gone dormant. Minor but notable eruptions were occasional in the '80s. Until 2004 that activity subsided to some degree. But watchful eyes of experts, mainly those at the Johnston Cascades Volcano Observatory, are using many tools to monitor this and other volcanoes in the chain for signs of impending eruptions. The main monitoring approaches can be accessed at the USGS's CVO site. One example: GPS is being used on the volcano to measure any vertical displacements on the volcano's flanks or within the ampitheater.

In September of 2004 there was a dramatic increase in the number of small earthquakes from beneath MSH's surface.

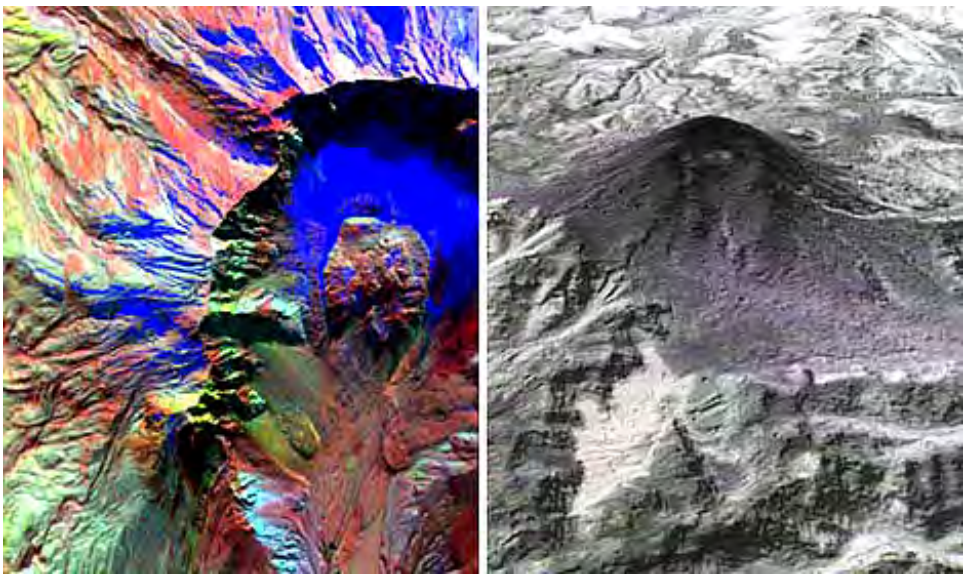


Here are seismograms typical of this flurry of activity:



During this period, the dacite dome has risen about 30 meters, as shown in this aerial view.

An airborne flight of the TIMS (thermal) instrument over MSH show in the left image hot areas in red and cold snow as blue. Other hot areas as appear as light tones in the right black and white image.



As September, 2004 progressed, ash and steam began to exude in significant amounts, leading the CVO to issue a Level 3 alert (strong likelihood of an eruptive event of moderate magnitude).

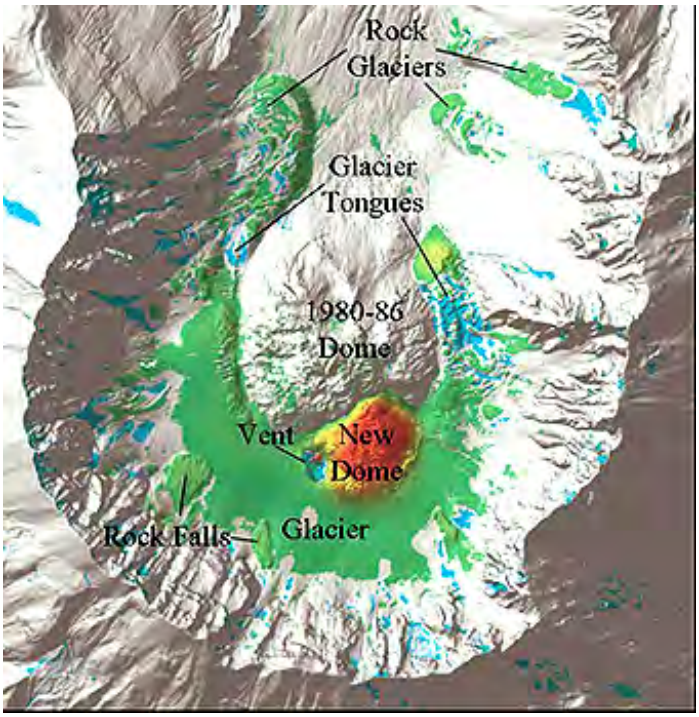
An especially esthetic eruption occurred at sunrise on October 10, 2004:

By mid-October, magma had reached the surface in the ampitheater (thus, to use correct terminology, it is now lava). A new smaller but rapidly growing dome, just south of the main dome, has been forming. Here are twp AP photos of the new dome in the day (left) and at night (right) in which the red glow is emerging lava.

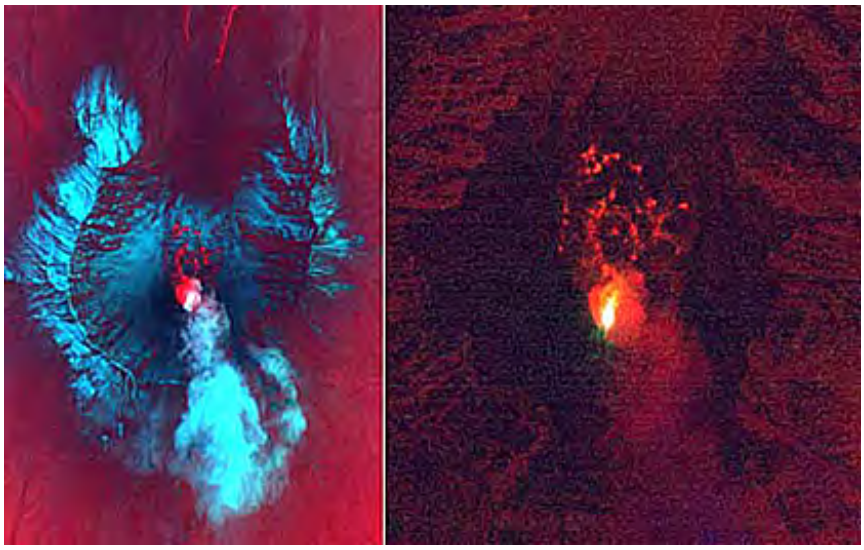
A closer view taken from an airplane shows the new dome in its early growth stage on October 1. The label "glacier" is misleading; it is actually a large patch of snow. Below the aerial view is an IKONOS image of the now paired domes as seen from space.

A lidar altimeter instrument was flown over the amphitheater to measure upward elevation changes due to the upwelling magma. These are plotted on an IKONOS image as the frame of reference. Color code for elevation increases: Blue = 0.1 to 3 meters; Green = 1.5 to 40 m; Yellow

= 40 to 80 m; Orange = 80-120 m.

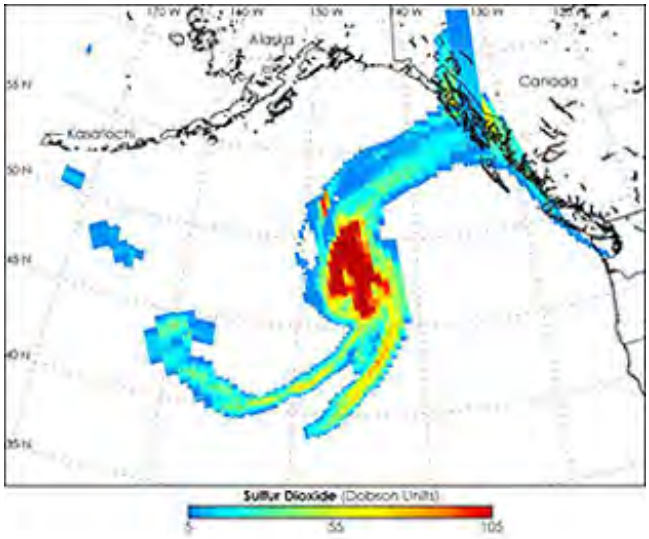


An aircraft thermal remote sensor, called MASTER (using components found on the MODIS and ASTER sensors on the Terra spacecraft), was flown over MSH to show the lava hot spots in detail:



Since its eruption MSH seems to have calmed down but some small earthquakes and a few steam clouds indicate possible activity in the foreseeable future. A good source to check out for updates on any MSH situation is the CVO Home Page.

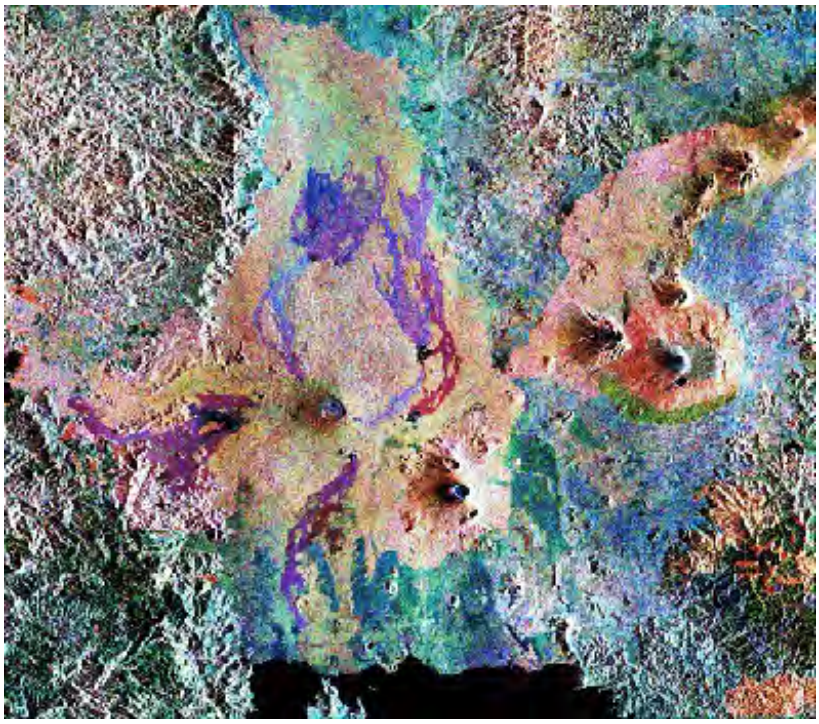
Alaska's Aleutian Islands are a continuous chain of active to dormant volcanoes lying in the upper plate side of a major subduction zone. On August 8, 2008 one of the westernmost Aleutians, the 2.5 km wide Kasatochi Island, erupted, sending a major ash plume south and east, causing airplanes to avoid the area for more than 24 hours (ash can clog both prop and jet engines):



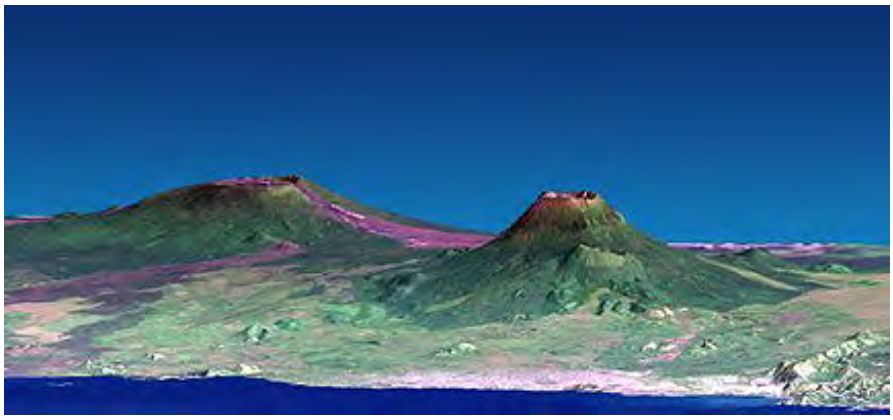
Just a few more examples of destructive volcanic events that were monitored by remote sensing. Overseas, at least one major eruption affecting populations seems to occur every few years. Currently considered one of the most dangerous in the world is Nyiragongo in the easternmost edge of the Democratic Republic of the Congo in the African Rift valley (see page 3-2).

Its "sister" volcano, Nyamuragira, centers about 20 km to the north. To the south is Lake Kiva, above which lies the city of Goma, in normal times with about 500000 people. Nyiragongo is a stratocone, with lavas containing high amounts of alkaline elements (sodium; potassium). The lava has remarkably low viscosity, traveling as fast as 100 km/hr (62 mph), which means it can descend on Goma and nearby villages in minutes. The top of the volcano collapsed in the past, producing a large caldera. This is evident in the ground photo that makes Nyiragongo look as though it has a flat top (this summit is 3400 m [11370 ft] above sealevel):

This SIR-C radar image of the Nyiragongo region shows this volcano and its "sister"



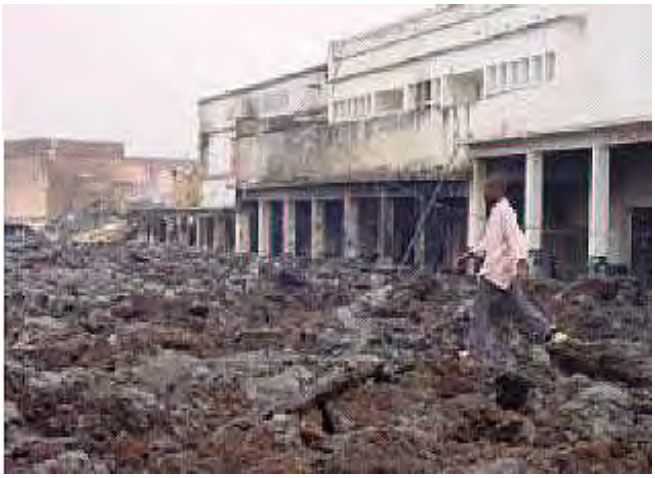
The two volcanoes appear in this radar/DEM SIR-C perspective image:



Nyiragongo is one of the most active volcanoes in the East African Rift. It has erupted, mainly by expelling lavas from the caldera and from fissures beyond the rim, at least twice in almost every century since records have been kept. The last eruptions occurred in 1948, 1977, and 2002. The eruption on January 17, 2002 caused lava to descend the volcano's south flank so that within 3 days it had flowed through Goma, destroying much property and killing more than 600 people.

That wide stream of lava was imaged by ASTER on Terra. The view below uses a Landsat TM image as the general picture, superimposed on which is this lava flow.

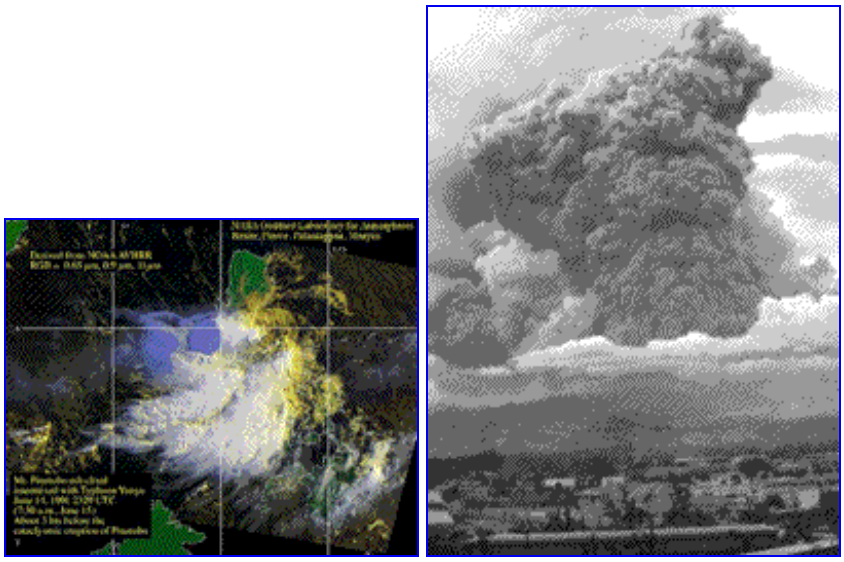
Parts of the town were covered with meters-thick lava, as seen here:



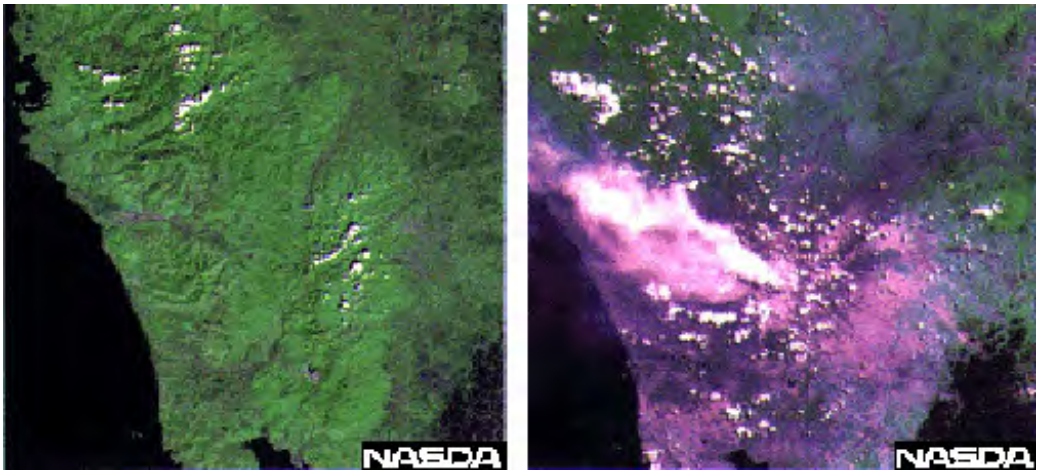
The volcano continues to show threatening activity, as seen in this photo into the caldera center, with a red hot lava lake.

Volcanologists now consider Nyiragongo to be a strong threat for a major eruption in the next several decades. The dilemma facing the Congolese government and the Goma officials is whether to evacuate the city in the near future or await more conclusive signs of an impending eruption.

The June 15, 1991, event at Mt. Pinatubo on the island of Luzon in the Philippines is a classic example of an awakening volcano, portending eruption. After almost 600 years of dormancy, earthquakes around it began in April of 1991. Their increase prompted evacuation of more than 50,000 people. Just three hours before the main eruption, the AVHRR on NOAA 7 imaged the billowing clouds of water vapor mixed with ash, that foretold the impending climax. This three-band color composite shows the eruption material interspersed with clouds associated with Typhoon Yunya (left image).



The peak eruption (right image), of the violent Plinian type, sent pumice, ash, and chemical aerosols above 12,200 m (40,000+ ft), leaving behind a 2 km (1.2 mile) wide caldera (cavity) that displaced most of the summit. Sulphuric acid was a major constituent of the aerosols that spread worldwide, affecting the weather and climate for the next five years. In the immediate vicinity, ash deposits surrounded the volcano for 20 km, destroying most of the heavy forests. That is evident in this before and after pair of images, obtained by the Multispectral Electronic Self-Scanning Radiometer (MESSR) on the Japanese MOS (Marine Observation Satellite) operated by NASDA (National Space Development Agency, of Japan).

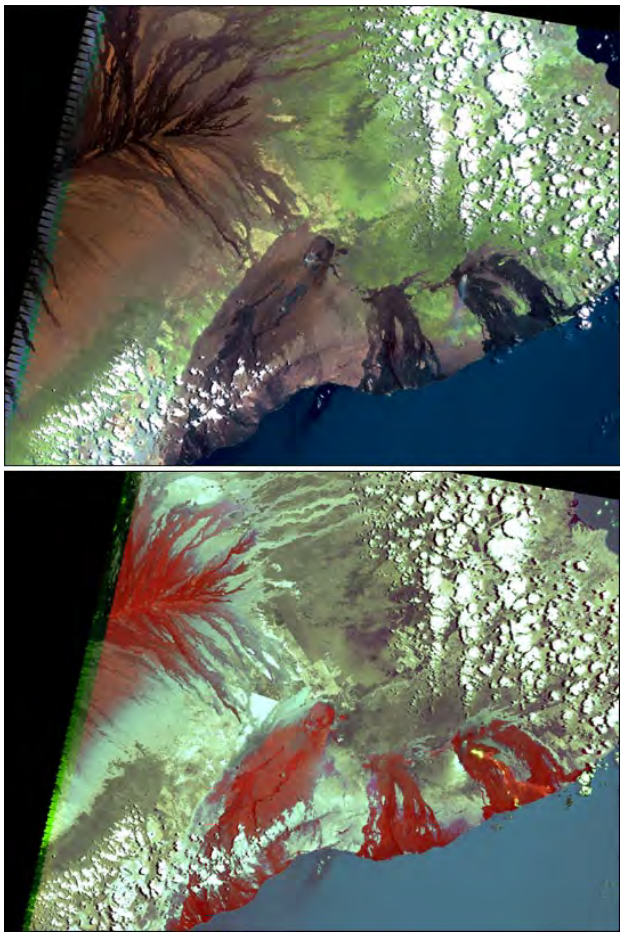


3-15: **What is the broad area of pink in the "after" image?** **ANSWER**

Europe's most active volcano, Mount Etna in the northeast corner of Sicily, began a series of strong eruptions in mid-July of 2001. Lava flows (dark red) threatened villages perched astride this stratovolcano's slope. One of its more violent eruptions was captured by a sensor on the Terra spacecraft, as seen in this image:

This is all we wish to say about Mt Etna now, since it will be the subject of page 13-4d which will summarize the "multi" concept.

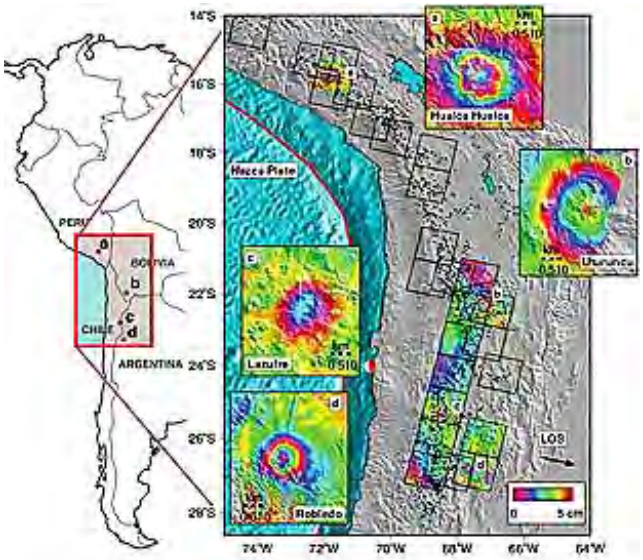
The most active volcanic area in the world is the Kilauea rift on the east flank of Mauna Loa on the big island of Hawaii (see page 17-3 and page 9-7). A more or less continuous period of eruptions of basaltic lava began in 1983. Lava has moved off island into the ocean at several points of entry, in effect building on to the land. Landsat-7's ETM+ sensor captured these two images (top, natural color in the visible; the bottom, a thermal image processed to show the warm residual heat in the lava as a red tone). While the paths and progress of lava (typically, 300,000 to as much as 600,000 cubic meters/day when active) are closely followed by park rangers and volcanologists on the ground (thus permitting early warnings to villagers that a lava stream may come their way), the space imagery give a long term running account of the spread of lava over the years.



Mud flows often accompany volcanic eruptions. These, and landslides, also occur after heavy rains. A landslide/flow is depicted in detail by the IKONOS satellite for an area in Venezuela subjected to a deluge and subsequent flooding.

Landslides are common by-products of ground shaking resulting from earthquakes. A 7.2 magnitude quake occurred in Pakistan on October 8, 2005, causing perhaps as many as 70000 fatalities in that country's north near Kashmir and parts of neighboring India and Afghanistan. Space Imaging's IKONOS has acquired images of the damage. In the town of Muzzaffarabad, a landslide removed part of a mountain slope carrying the debris into a river. The area moved is dark in the lower image but the upper image now shows a light-toned scar.

Turning now to a direct accounting of the effects from earthquakes. Considerable scientific manpower and money have been dedicated to discovering ways to predicting damaging and life-threatening natural events - mostly of a geologic nature. Earthquakes and volcanic eruptions lead the list. Radar interferometry (see page 11-10) offers real promise since one of its accomplishments is to be able to measure vertical ground displacements (such as occur when a volcano begins to swell upwards and outwards as its magma chamber fills) and, under appropriate circumstances, lateral (subhorizontal) displacements (as occurs along strike-slip or wrench faults). In the illustration below is evidence of ground swelling at four volcanic sites in South America as determined by analysis of 8 years of data from the radar on ERS-1 and ERS-2; while this may not guarantee an impending eruption, it does fit the pattern known to foretell many eruptions.



As one might expect, high resolution space imagery can be effective in assessing earthquake damage. On December 26, 2003 a magnitude 6.5 quake occurred near the city of Bam, in

southeast Iran. This ancient city's buildings are constructed mostly from sun-dried clay blocks. Many of these, poorly suited to withstand earthquake waves, failed catastrophically during the main temblor and subsequent aftershocks. Casualties in the town and surrounding areas have exceeded 36000 and may reach 50000 when all debris has been searched through (another quake in northwest Iran killed 50000 on June 21, 1990). This IKONOS image gives an idea of the destruction in Bam, in which >80% of the buildings in this metropolitan area (about 80000 lived in Bam and another 100,000 in nearby areas) were destroyed:

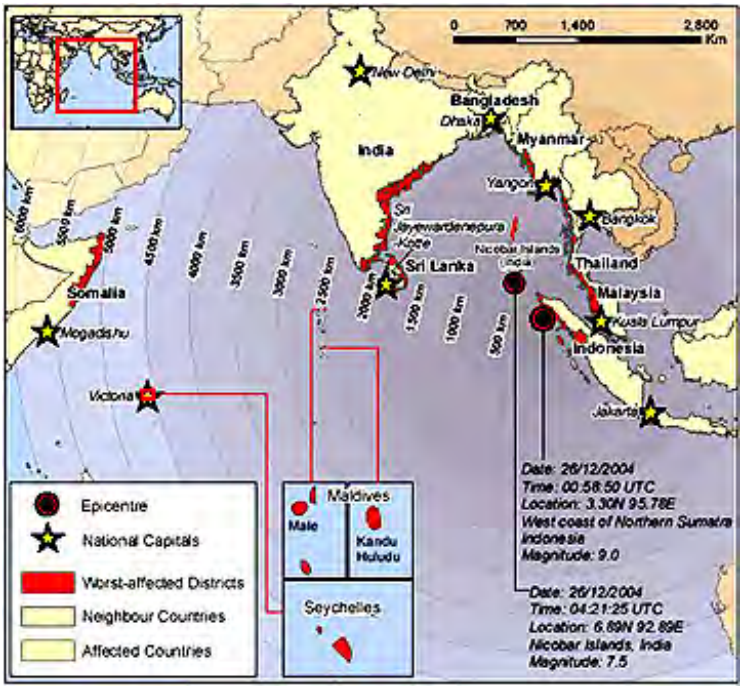
This aerial view zeroes in on the nature of the damage:

Surviving the destruction was the Citadel in Bam, the largest adobe structure in the world, seen here in a pre-earthquake image.

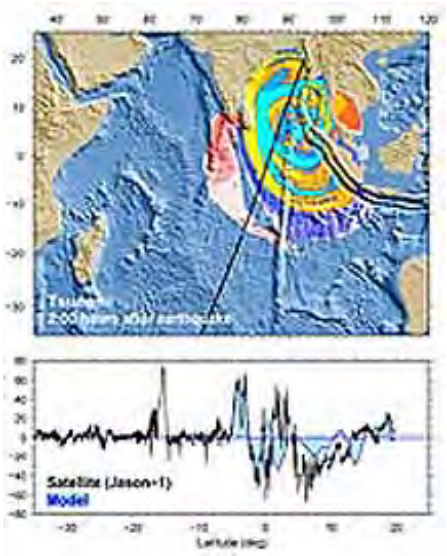
This next pair of MISR (Terra) images reveals another subtle manifestation of the aftermath of an earthquake of magnitude 7.7. This, the Gujarit quake which killed more than 20000 people, took place on January 26, 2001 along an active fault line in western India. This arcuate trace just above the large white sand area on the right is visible in both the January 15 and January 31, 2001 images. Look closely at that line in the Jan. 31 image, along its left half. Small black blotches are evident - these are not present in the Jan. 15 image. They are the marks of saturated soils resulting from liquefaction releasing groundwater that seeped to the surface after the earthquake.

The largest number of countable human deaths from an earthquake, as occurred on land near the epicenter, in the last hundred years was approximately 242000 near and in Tangshan in Eastern China. However, the submarine earthquake (magnitude 9.3, the second largest on record) off the northern Sumatra coast on December 26, 2004 (see second page of Overview) may have killed more. The deaths were caused almost exclusively by tsunamis (huge "tidal" waves produced by upheaval of ocean water above the epicenter). By mid-January, 2005 official estimates placed the dead at ~140,000 but as the Indonesian government compared the accounted-for living with census figures, they have claimed by March that the total was much greater, approaching 300,000 (includes Thailand, India, Sri Lanka, and other countries around the Indian Ocean). The areas affected in and beyond the Indian Ocean are shown in red on this map

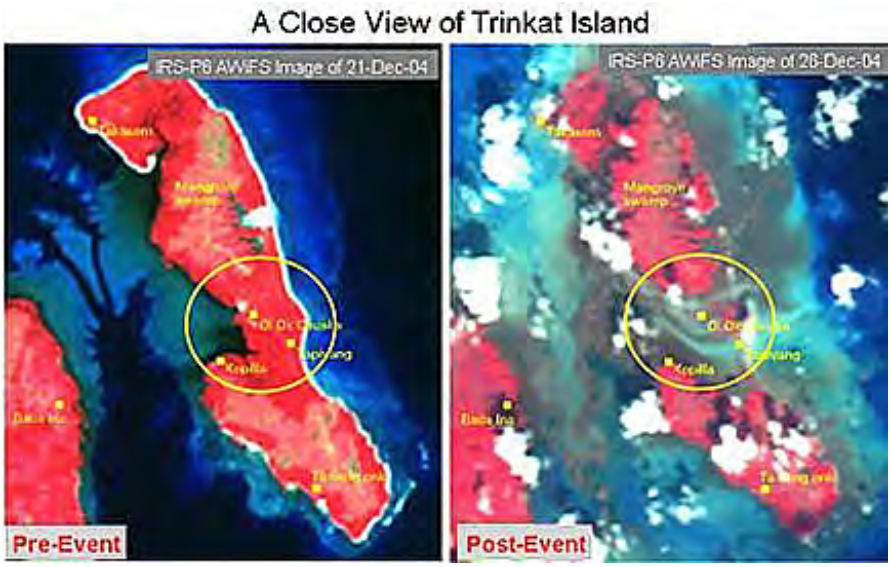
(note that there was a nearby smaller earthquake several days earlier).

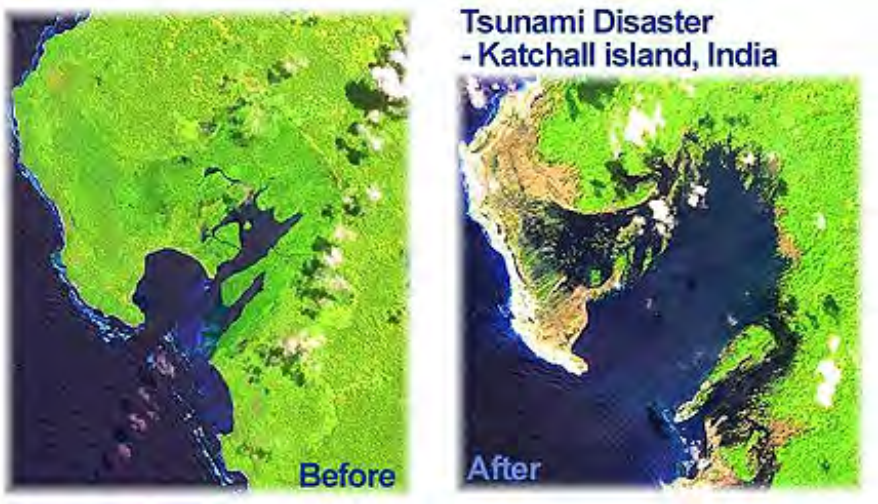


Radar aboard a NOAA satellite provided data that displayed upwards displacement of sea surface height during the tsunami:

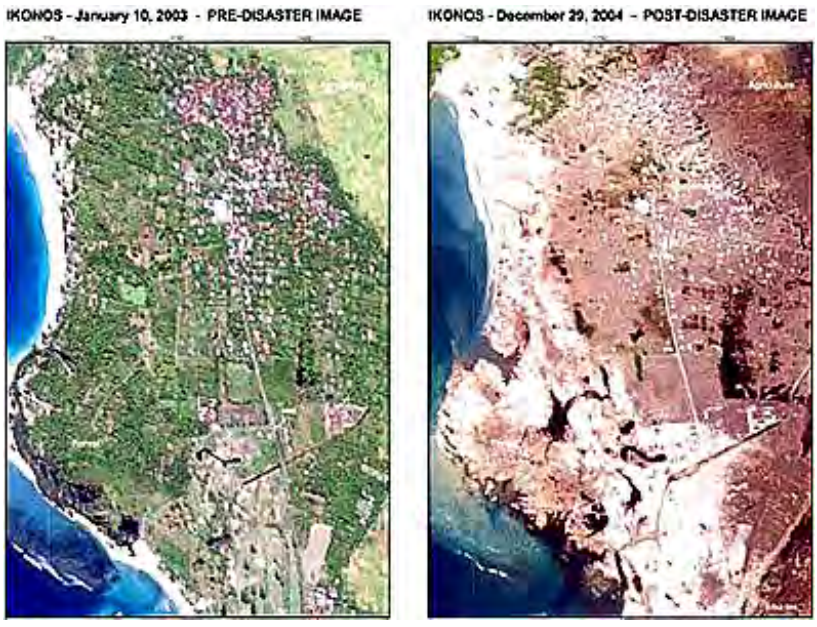


Some idea of the ability of the tsunami waves to destroy not only property but the land itself is evident in these two satellite images:



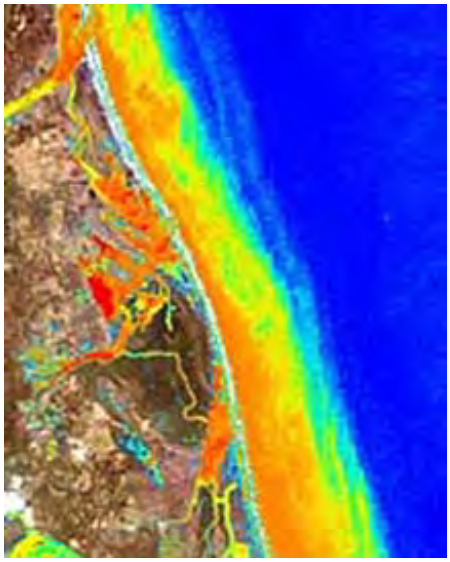


The southwest coastline of Sumatra suffered extraordinary levels of damage, with villages and vegetation being washed away:

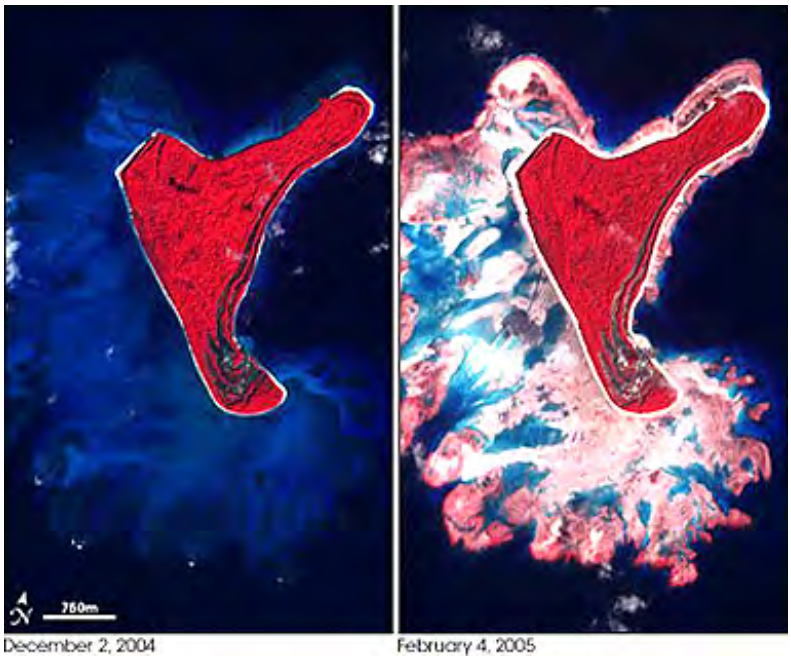


Hardest hit was the capital of the Aceh province of Sumatra, Banda Aceh. Here are Quickbird images of much of this city of several hundred thousand on June 23, 2004 before the tsunami struck and again after it obliterated much of the land as seen on December 28, 2004:

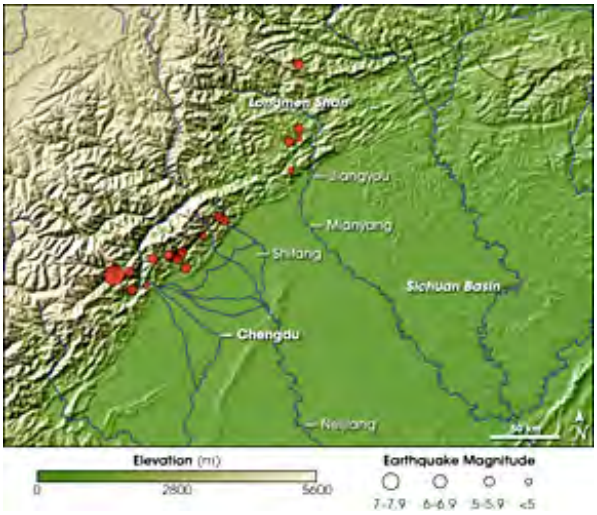
The tsunami generally carried for several hundred meters onto land beyond the shoreline. As the waves receded, they carried debris and soil well out into the ocean. The strong increase in the resulting sediment appears in this satellite image of the coastline near Cuddalore, in southeast India:



Earthquakes such as the December 2004 Sumatra event can cause permanent displacements of the ground - both on land and at the ocean floor. Movement along the low angle thrust that led to this earthquake extended for perhaps 800 km (500 miles) laterally. Parts of the ocean surface sank; other parts rose. In the Andoman Islands (Indian Ocean) this pair of ASTER images shows North Reef Island before and after the quake. It is obvious that the land rose here, killing off some of the now-exposed coral reef surrounding the island.



China experienced a major earthquake disaster just 3 months before the start of the International Olympics in Beijing. On May 12, 2008 a 7.9 quake whose epicenter was situated in the mountains in Sichuan Province; it was followed by several strong aftershocks:



Over the region, as many as 80000 lost their lives. Here is a before and after image pair of the town of Beichuan (about 90000 people); interestingly, it was taken by Formosasat-2, and supplied to mainland China by the Taiwanese:

The right image shows that there is a landslide. Landslides are frequent sources of destruction both when they occur during or immediately after the quake and because of increased instability when set off by rains. This Sichuan scene shows the devastation caused by the May 12th earthquake:

The Sichuan landslides have created a very dangerous situation in some areas. In the steep terrain, the landslides have blocked streams, creating natural dams. Water has rapidly built up lakes that in place threaten to break through, causing potentially catastrophic floods. Chinese engineers are blowing apart certain lakes, after tens of thousands were evacuated. These Formosasat images show one such threatening new lake as it built up:

Satellite data will likely play an important role in any future geologic "catastrophes". Images are especially valuable because they usually would show the extent of building damage. Ground displacement, landslides caused by earthquakes, volcanic eruptions, or heavy rains are another geologic phenomenon having ecological manifestations.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

*As a prime **case study** of how remote sensing is aiding a major land management and restoration project - the Everglades subtropical "swamp" - in southern Florida, this added-on page describes the nature of the natural ecosystems within the Everglades, the threats to its survival by the encroachment of civilization, and the plans recently devised to preserve and protect the natural habitats through expensive federal/state programs supposedly now underway.*

The Everglades: America's Most Threatened Ecosystem

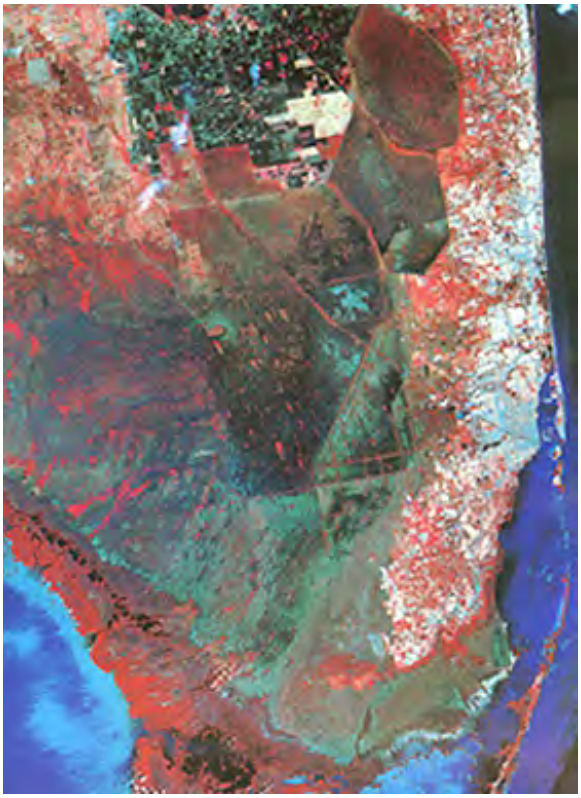
The Great State of Florida - the Sunshine State - with nearly 16 million residents, grows each winter as "snow birds" from the North pour into the state to enjoy often tropical-like weather and marvelous amenities for vacationers. Some, liking the lifestyle, come back to settle. (This invasion of resettlers assumes a fair tolerance for hurricanes.) But this influx, on top of the previous permanent residents, is steadily putting a strain on both people facilities and natural ecosystems. Nowhere is the long term outlook more troubling than that last refuge of the tropics in the U.S. - the Everglades. This huge area of water and vegetation, almost untouched by man, is the most fragile and susceptible ecosystem in this part of America. The problem is the insatiable desire of developers for more land to house more people and establish recreational areas as well as farmland taking advantage of fine soils. But a major decision has been made by federal and state governments, and environmental groups (the "Greens") to reverse adverse land management trends and restore the pristine state of the Everglades subtropical wilderness. This page will be a "case study" in examining these efforts and in showing how remote sensing plays a major role.

To see the Everglades as part of the entire Florida peninsula, look first at this Landsat mosaic:

The Everglades occupy much of the southern tip of the state. Miami (dark pink in image) lies near the southeastern tip of coastal Florida. The large body of inland water is Lake Okeechobee,

which in a sense was once the north end of the Everglades and still strongly influences the drainage patterns to the south. The entire peninsula is geologically a terrane that was added to the rest of North America in the last 50 million years. It is made up of ancient crystalline rocks topped by mainly limestones deposited in the Cenozoic Era.

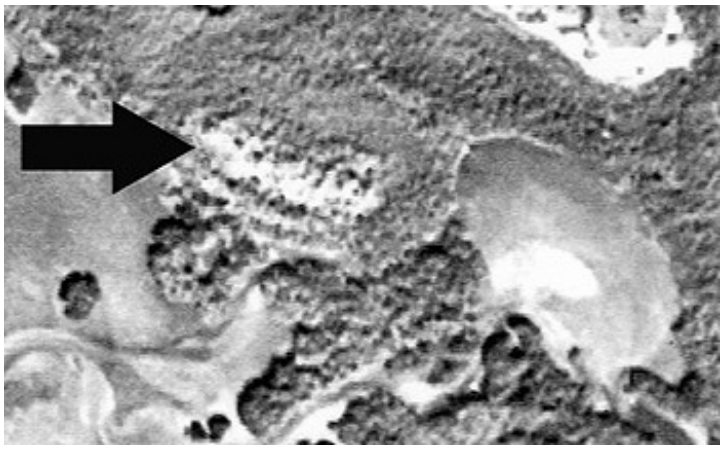
From that regional view, we next show a Landsat-3 subscene which covers part of the Everglades and the urban strip on the Atlantic Ocean.



Much of the Everglades is dark because most of its area is covered with water (adds black to the standard false color version). Areas in strong red at its south end are mostly mangrove swamps. Small elongate areas are "Tree Islands", slight rises above water that support hardwoods and other vegetation. Look for several red lines crossing the Everglades: these are drainage canals that host active vegetation. The eastern coastal area of this part of Florida is built on limestone bedrock, topped with soils and sands, and is now one continuous metropolitan area that includes Miami and Fort Lauderdale (see page 4-2). More features in this scene will be discussed elsewhere on the present page.

We now want to further familiarize you with what the Everglades are all about. You have a choice: open these online Everglades Summary and ParkVision sites and/or just keep reading here.

First, we will consider its pre-Columbian history. Archeaological evidence suggest humans may have inhabited the lands, shorelines, and wetlands as early as 10000 years ago. But the initial indications of a sophisticated tribal community goes back to about 100 A.D. with the presence of the Calusa (Caloosa) People (one of many populations in North America that are lumped together as Natives). They spread over much of present-day Florida but were concentrated mainly on the southwest coast around today's Fort Myers and Naples. They were builders of artificial islands and inland mounds made from oyster and other molluscan shells (their principal food). On these were established small villages. An aerial photo shows one elongate shell Mound



The population center for the Calusa was the Ten Thousand Islands area south of Naples, once part of the Everglades.

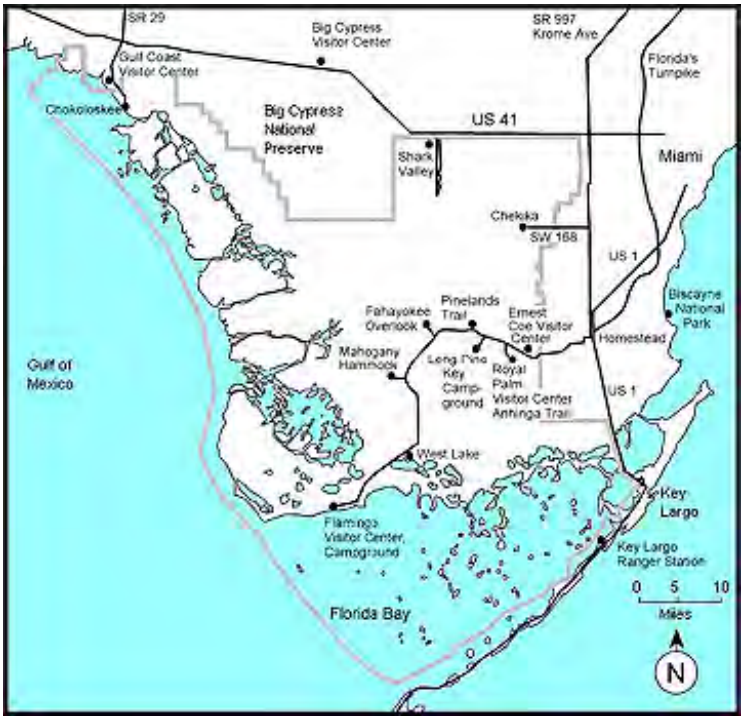
The most ambitious Calusa project was the building of a 125 acre island that rises from 25 to 40 feet above today's sealevel. The island had homes, religious buildings (the Calusa were in to human sacrifice), and water storage facilities. This was probably their capital "city", but today Mound Key is a National historical site, which shows only the foundations of some dwellings:

The Calusa culture was fairly advanced, comparable to the Mayans, as demonstrated by this rital mask used in religious ceremonies.

The Calusa persisted until the coming of the Europeans in the 16th Century. They shared Florida with other tribes such as the Tequesta. The Spaniard explorers eventually wiped them out, but not before a Calusa arrow had killed Ponce de Leon, thus cutting short his effort to find the Fountain of Youth in Florida.

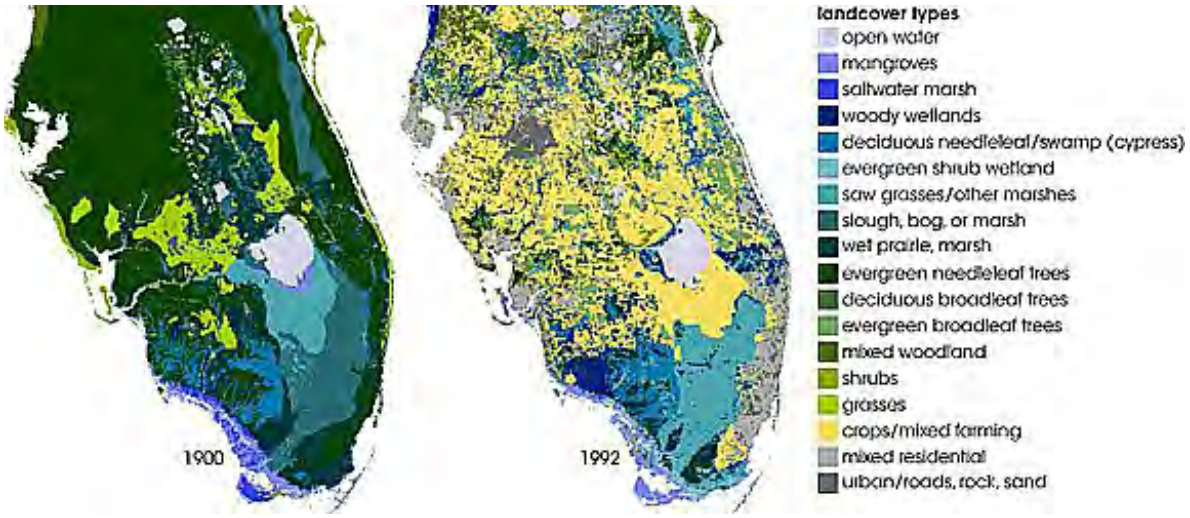
The Indian population of Florida went into sharp decline after the Spanish took over. The best known of modern tribes is the Seminoles. These originally lived in Georgia and elsewhere but were driven south all the way to the Everglades by Gen. Andrew Jackson and others. Today about 10000 still remain in the state. Some live in small settlements in and around the Everglades.

Next, to establish the geography of the Everglades, look at these three maps:

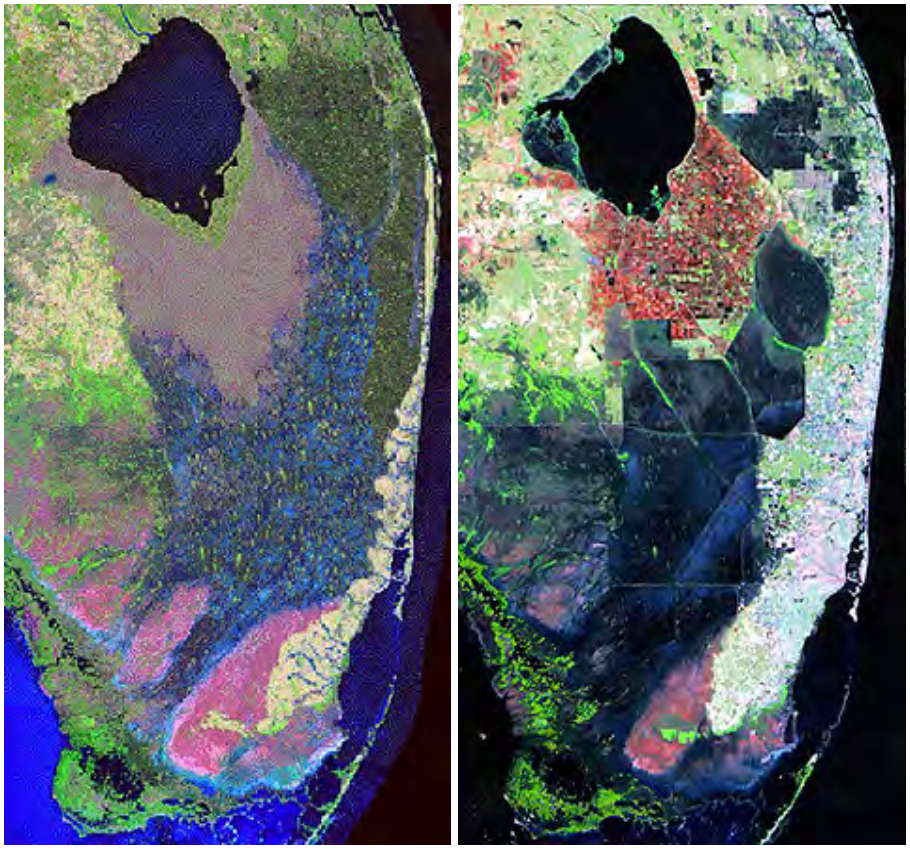


There is an even more detailed map that focuses on the Everglades National Park but to reduce it to fit this page would make the smaller print unreadable. So, you have the option of seeing the larger version on this extra page 8a.

From the above it is evident that the Everglades includes large areas outside the Everglades National Park. The Park was established in 1934, largely through the dedicated efforts of an architect, Ernest Coe, whose passion for this extraordinary wilderness made him an effective lobbyist. Since then the entire ecosystem has been in jeopardy from the development of Florida for its residents and visitors. Two maps indicate the trends.



Both maps are almost self-explanatory. Florida has altered much of its natural land cover types to farmland, grazing land, and urban/suburban units over the last 100 years. In the last 100 years the natural swamps of southern Florida that include the Everglades have shrunk by about 50%. This is a loss of about 2.9 million acres. There is another way to view these changes. The U.S. Geological Survey, one of the agencies involved in protecting the Everglades, has produced an "artificial" Landsat image of what the land cover was like in 1850. That image (left) is placed next to a recent (2000) actual Landsat image. Compare the differences:

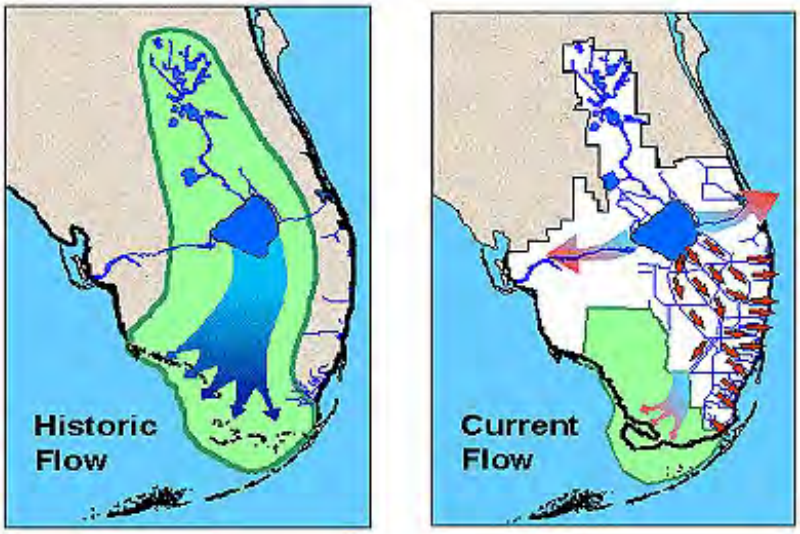


The reasons for the Everglades existence are largely tied to geology, climate, and drainage. Southern Florida is almost entirely capped with limestones that are more easily eroded than most other rock types. Sinkholes, drainage channels, and outliers making slightly higher ground are found in the Everglades and into central Florida. Here is a view in the Everglades showing

limestone outcropings.

The climate factor is the prevailing hot and humid conditions that promote vegetation typical of subtropical marshlands and forests. We will examine examples of this below. The dominant factor has been the pattern of drainage from pre-colonial times that has since been largely altered. Lake Okeechobee is a large depression (developed from dissolved limestone) that has been filled by streams such as the Kissammee River which begins south of Orlando. From this lake water for eons has flowed in a series of streams, some riverlike, others just sheetflow across the lowlands to the south. Water flow, at current rates southward of about 100 ft (30 meters) a day, extends **over most of the Everglades** to depths of a few inches up to six feet. So the Everglades is in effect a grandiose wetlands with 1-2 feet of peat below and limestone bedrock beneath that. In effect, the water flow is almost akin to a single, very broad and shallow river, which can be seen at the surface but is also hidden in large part by vegetation. This is a typical view of the channel system; most of the vegetation is sawgrass.

Before major development the general pattern of water flow in southern Florida was this ("current" refers to both flow direction and to "contemporary"):



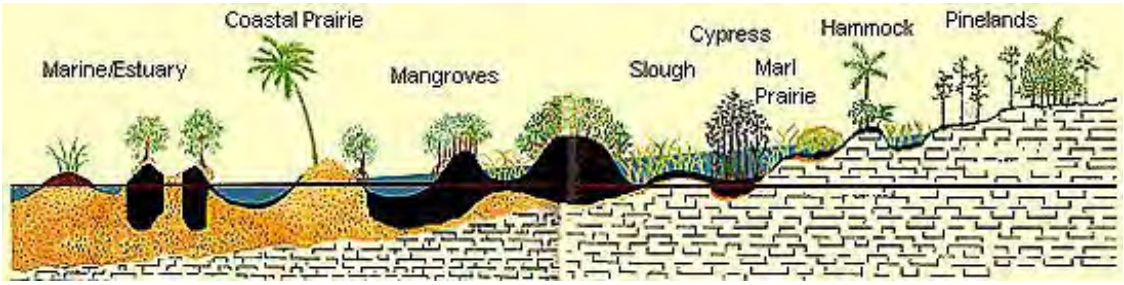
Lake Okeechobee has been the key to the changes in the regional drainage system. Suffice for now to say that the drainage system from Okeechobee south has undergone major alteration. The entire lake has a restraining wall or levee that allows control over the drainage outlets. This wall, 143 miles in length, was built by the U.S. Corps of Engineers in the early 1940s in response to flooding caused in a 1928 hurricane which carried excessive water into the Everglades and burst protective levees next to the coastal strip (several thousand people drowned). Made of earth and rock, the Okeechobee levee is 30-40 feet above normal lake level and 20 feet wide at the top. Several canals have been built in the 1960s (the water they receive passes through a lock system), trending southeast and south, from the lake country into the Everglades to aid in carrying off excess water. Now, about 1.7 billion gallons a day are released from Okeechobee in a controlled manner. As a planned by-product, this produced land for farming and metropolitan use. But, this re-engineering of the drainage has significantly modified the natural state of the Everglades so that it now receives much less water. And parts of this water are being polluted by the inhabitants surrounding the Everglades.

Now, instead of unrestricted flow southward into the Everglades, and thus fundamentally modifying its hydrology that preserves much of its natural ecosystem, Okeechobee waters are largely redirected to the Gulf Coast at Fort Myers via the west-flowing Caloosahatchee River and east to the Atlantic via the St. Lucie, West Palm Beach, Hillsboro, North New River, and Miami canals. The present system of natural and canal drainage is superimposed in blue on a Landsat mosaic:

Reclaimed land immediately south of Lake Okeechobee (see Landsat images) is used for agriculture - principally sugar cane and citrus groves.

Both in the past and now water released from the southern dikes at Okeechobee and surviving natural drainage flows through the Everglades as sheetflow and rivers. Most of the Everglades has water at its surface which will be moving at various rates towards the coast. The main flow routes are called "sloughs" (pronounced "slews"), similar to the bayous of Louisiana, that largely are covered in vegetation ("rivers of grass"). The largest, Shark River Slough, is shown here in a map and as an aerial view (water underlies the vegetation but is hard to see).

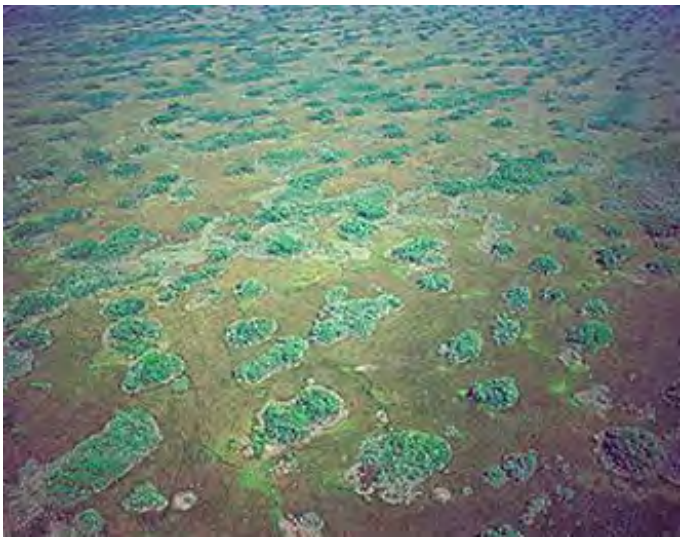
But despite these drainage perturbations in the original Everglades, most of the flora and fauna are still preserved. Let's look at a series of pictures that describe the typical members of this habitat. First, a cartoon summarizing the principal vegetation types in relation to sealevel.



The next two views show vistas in which the dominant vegetation - sawgrass - makes up large segments of the eastern Everglades. This underlies the appellation given to the Everglades as "The River of Grass":

These two photos show the grass and reeds in close-ups.

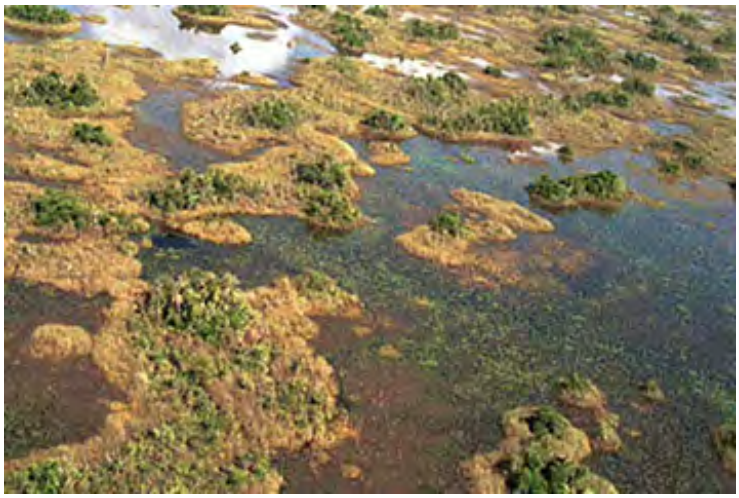
Note in the two images below the presence of isolated trees in clumps. Often referred to as hammocks, these "Tree Islands" are characteristic of much of the Everglades. On a grander scale, hammock groves are separated by water too deep to support much vegetation:



Sometimes the stream flow actually organizes into channels:

Stands of southern pines are found in areas above water level. They often are associated with palmettos and other bladed plants.

Cypress forests occur both scattered in the upper Everglades and as a dominant vegetation type in the Big Cypress National Preserve. The next pair of pictures show Cypress trees in summer, and the third, an aerial view of Tree Islands supporting Cypress stands.



One of the most distinctive tree families in the Everglades, found mostly in the coastal and southern parts, are the Mangroves. Mangroves, like Cypresses, grow in permanent water. These trees come in several varieties, as shown in the next two photos.

Continued on page 3-9

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Remote sensing tends to orient towards any applications that are of direct and immediate benefit to people and society. Map types that are designed to show both natural and manmade features on which data pertinent to those aims is recorded and from which up-to-date information is easily extracted are referred to as land cover and land use maps. A commonly used land cover classification is presented and an example of a classified remote sensing subimage that displays cover classes is given. Much of this section will concentrate on the appearance of cities and urban areas in space imagery; these images are good starting points for making land use/cover maps. Several urban areas in the U.S. Southwest are examined on this page.

URBAN AND LAND USE APPLICATIONS:

FROM LOS ANGELES TO BEIJING

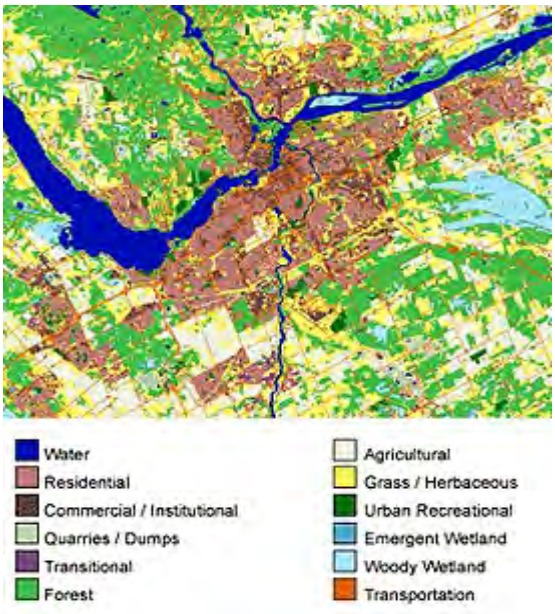
Some Basic Principles and Examples

In this Section we will consider the category Urban or Built-up Land. Since more than half of the American people (and likewise in many countries) now live in cities or metropolitan areas, one might expect, in land cover percentage, a signficant part of the continental land surface falls within this category. The U.S. Census Bureau places urban areas in their category "Developed Land". To qualify in this category, the Bureau requires that the land holds a minimum of 30 people as *residents* per square mile. It is surprising then to learn that as of the year 2000 the percentage of developed land in the continental U.S. is just 5.4%. The actual metropolitan area perentagewise is even less. But, since most people live in developed land, monitoring its distribution, its growth, and its characteristice becomes an important task amenable to vital input from earth-observing satellites.

As strongly alluded to in the first three Sections, space imagery (and aerial photography as well) is a powerful medium for mapping what's at the Earth's surface. We can identify and categorize the various natural and man-developed features present on the surface in terms of land cover. The term land useis almost a synonym, but refers specifically to how the land is used for human activities. Land cover and use maps are an essential input to Geographic Information Systems, described in detail in Section 15. Most of these maps follow some classification system. One in common use is that developed by James Anderson and his colleagues at the U.S. Geological Survey. This hierarchical system has four levels. The two higher ones (Levels I and II) are categories that we can usually identify and hence, map, using space images, whereas, aerial photos (for higher resolution) or IKONOS-type space images are needed for Levels III and IV. This classification in Levels I and II is shown here:

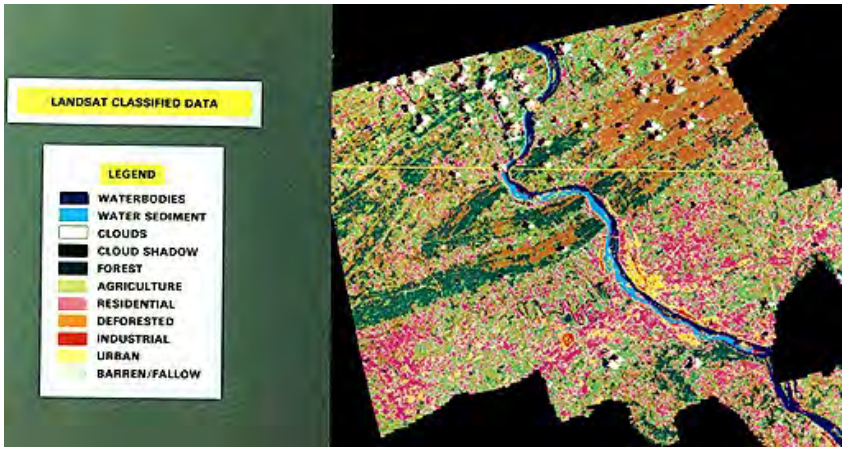
To illustrate the subdivisions at Level III (not shown), we pick Level I = Urban - and Level II = Residential, then the subdivisions under Level III are: units = Single-Family; Multi-Family; Mobile Homes; Hotels/Motels; and Other. Generally, the finer the breakdown, the more on-site ground truth information we need. Maps for Levels I and II can be colorized, and we can add numerical symbols (e.g., 115 = Residential Hotels) for individual features on maps drawn at a large enough scale to fit them.

As you saw in Section 1, a common map product now almost routinely made from satellite remote sensing data is the *scene classification*, done through computer processing. An example of a typical land cover and use map (some Level II categories) is this supervised classificaton of the city of Ottawa, Canada (the country's capital):



For comparison, here is a multi-theme classification of land use/cover derived from Landsat imagery that extends over an area in and surrounding Harrisburg, PA (you saw other Harrisburg examples in the Fiirst Exam, Section 1). The city is located next to the prominent yellow pattern

assigned to urban. Smaller yellow blotches to the east and west are indications of downtowns, such as Carlisle, Hershey, and Lebanon. This metropolitan area, with a population exceeding 300,000, is generally spread out beyond the Harrisburg city limits. The legend category called "deforestation" describes the same Gypsy Moth defoliation touched upon on page 3-6. Land use/cover mapping with space images such as Landsat, SPOT, JERS, IRS is capable of displaying with good (not excellent) accuracy some of the Level II categories. IKONOS and similar high resolution satellite imagery can pick up many of the Level III units.



Students at the University of Arizona have produced an excellent Web site illustrating in a series of well-documented steps the approach they took to classifying land use in a rural part of Cochise County, AZ. They compared Landsat subscenes for two dates, in 1973 and 1992, working towards a change detection map that picks out new features or different uses for some earlier features.

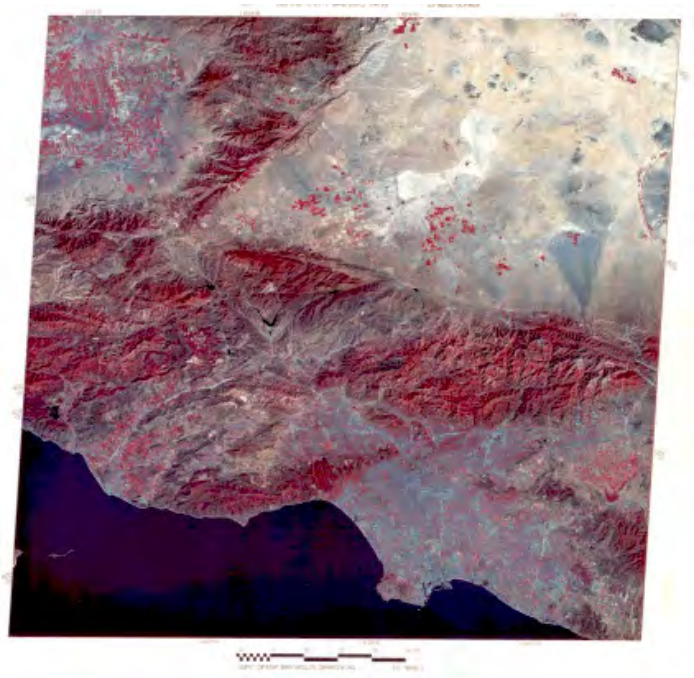
For the next three pages, we will concentrate on urban land use by looking at space imagery of a group of U.S. and foreign cities:

4-1: **As a warm-up and refresher, briefly review how a metropolitan area appears in a standard false color composite. It is sometimes hard to delineate the outskirts of a suburban land use class; suggest one way to roughly decide on its boundary.** **ANSWER**

The Urban Southwest: Los Angeles, San Diego, Tucson, Las Vegas, San Jose, and Mexico City

Because much of the U.S. and world populations are concentrated in and near metropolitan centers, we concentrate in this Section on urban land uses.

Major metropolitan areas are sprawling, often occupying a significant fraction of a full Landsat scene (although on a world scale, these areas comprise less than 10% of the land surface). This is certainly the case in one of the largest (areawise) urban regions in the world - the Los Angeles megalopolis as seen in a resampled (lowered resolution) MSS image that covers this part of southern California.



This Landsat image graces the cover of the NASA publication "Mission to Earth", which the writer conceived in 1974. This scene covers only a part of the total megalopolis, which extends well to the east and to the south. A MODIS image broadens the coverage to incorporate the full megalopolis:

Major geographic features in this image include the western Mohave Desert (containing Edwards Air Force Base), the southern tip of the Great (San Joaquin) Valley, the Tehachapi mountains on the north (extending from the southern tip of the Sierra Nevada), and the Transverse Ranges north of L.A. (bounded sharply on the north by the infamous San Andreas fault and on their south by the Santa Monica and San Gabriel Mountains). Los Angeles (lower right) lies within a structural basin, forming a lowlands that restricts air circulation and is a natural trap for pollutants (the famed L.A. smog).

4-2: **Consult an atlas; find the above features and others that you may know about from living in, having visited, or having heard about Southern California. For example, find Hollywood.** **ANSWER**

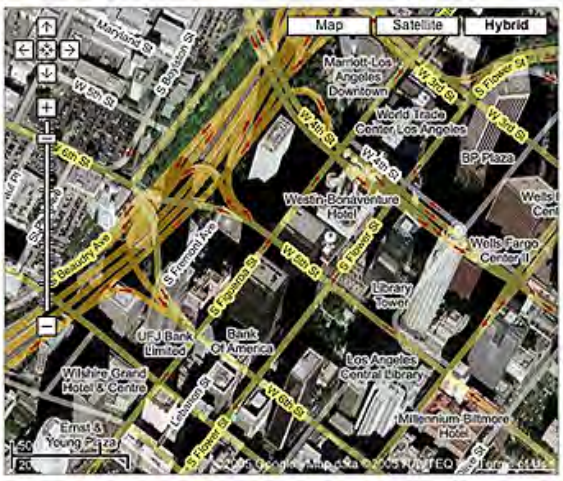
A more detailed look is given in this Landsat-7 subset:

The urban signatures in a false color composite are clearly evident in the segment of the MSS image containing a part of the numerous cities and suburbs that makes up the greater Los Angeles regional sphere of influence. The areas of higher population density show two dominant color themes: thin, usually blue to almost black, linear criss-crossing patterns representing streets and roadways; and areas in between that are also usually some shade of blue. The latter are the spectral expression of buildings, which tend to reflect brightest in the blue and green bands. Interspersed within these areas are patches of red, which correspond to city and town parks, cemeteries, golf courses, and pockets of intra-urban agricultural fields. Residential areas often have color signatures ranging from brown to pink to mild red, stemming from the mix of lawns and trees with houses and streets. In the Los Angeles hills and neighboring mountains, signs of homes and neighborhood centers are hard to recognize because the trees and brush swamp those areas with reds. The Palos Verdes Estates section on the ocean is a good example. However, bands and patches of blue in the valleys, such as in the Santa Clara valley east of Ventura, indicate commercial concentrations along highway strips.

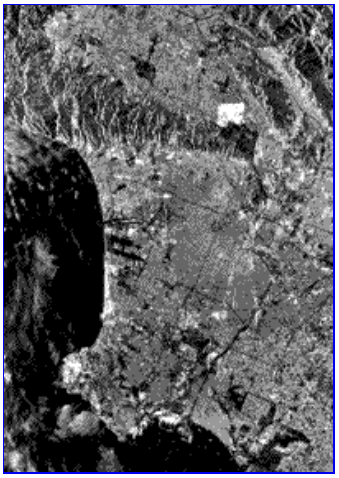
The Transverse Range is quite an impressive backdrop for the citizens of the Greater Los Angeles area to look at from various spots in the basin. This next pair of images are perspective views of these mountains and parts of the Los Angeles basin with its many metropolitan areas made by combining data from NASA JPL's SRTM radar mission (see Section 11) with an enhanced image acquired by Landsat-5.

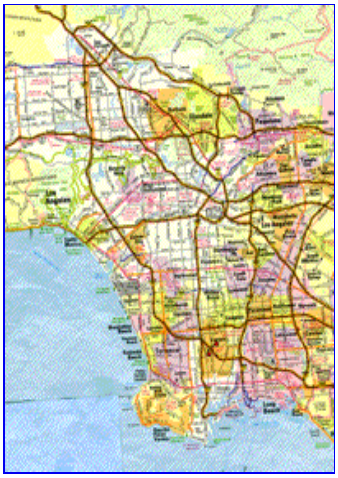
It also helps to show a ground photo of some part of a city or metropolitan area to tie in with the space imagery. We will do this for many of the cities (especially in the U.S.) in Section 4. Here is a view of downtown Los Angeles:

So, suppose you want to know what's where in downtown L.A. Google Earth, as we have seen before, provides a myriad of products online at their Internet site. This is an example of a high resolution Quickbird image labeled to show the main buildings in part of that downtown:



Urban areas are quite distinctive in radar imagery, as is clearly evidenced in this 1978 Seasat image (see Section 8 for a description of this satellite and a review of radar interpretive principles) that covers the west side of the Los Angeles metropolitan area. We include a segment of the 1989 Rand McNally atlas map that locates roadways, towns, and other features in the urban infrastructure (its writing is not legible; it is shown mainly to aid you in picking out major roadways in the radar image).





Many of the dark linear features in this radar image correspond to interstate highways and other major roadways. The two dark, narrow, east-west strips just on shore at Santa Monica Bay are runways at Los Angeles International Airport. These dark patterns come from specular (smooth) reflecting surfaces that deflect most of the radar beam away from the receiving platform.

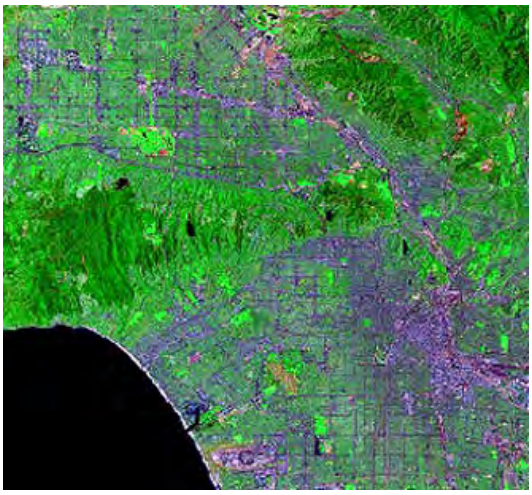
Generally, buildings return much of the radar signal, giving rise to an intermediate gray tonal signature (uniform here because most individual buildings are not resolved). Very light patterns are associated with high levels of backscatter caused, in some instances, by stands of vegetation. The light pattern along the west coast of the Palos Verdes peninsula, a luxuriant residential area, may come from dense vegetation. However, the very bright, squarish pattern at the east end of the Santa Monica Mountains that extends into the San Fernando Valley around Burbank-Glendale is rather mysterious. It approximately coincides on the map with Griffith Park, north of Hollywood, but its position relative to roadways that are discernible in the radar image, indicates an inexact fit (possibly an effect of non-rectification).

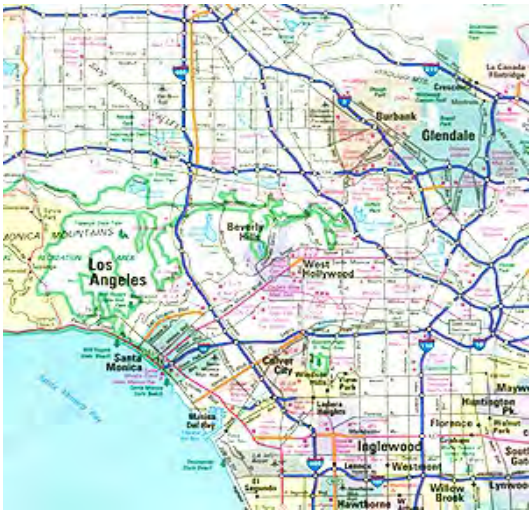
4-3:**Again, use an atlas and turn to the pages that show the Greater Los Angeles region. Try to locate the various Interstates and Freeways. Find such places as the Santa Monica Mountains, Glendale, central Los Angeles, Torrance, Long Beach. You will probably note that the road patterns in the radar image don't fit too well with the map - some main roads are not clearly singled in the image. Care to guess why this is so? What are the bright patterns protruding into the ocean at Long Beach?** **ANSWER**

The layout of metropolitan Los Angeles is made more evident when this region is viewed at night, as in this astronaut-taken photo from the International Space Station:

When "Los Angeles" is mentioned, people often think almost at once of "Hollywood". The IKONOS satellite succeeded quite well in finding Hollywood. In fact, it imaged the famous large white letters (look near top) that spell out this name against the backdrop of a steep slope on the south side of the Santa Monica Mountains near this most recognized city in the Los Angeles area:

Now that we have led you through this tour of the Los Angeles region, and prompted you with questions designed to familiarize you with its landmarks, we want to make sure you found the major cities and suburbs in an north of Los Angeles. This will be done using first this subscene made from a Landsat ETM+ image, and then with this Rand McNally map that should be legible enough to spell out localities:





A provocative question: what towns in the scene do you associate with the movies stars and other celebrities?

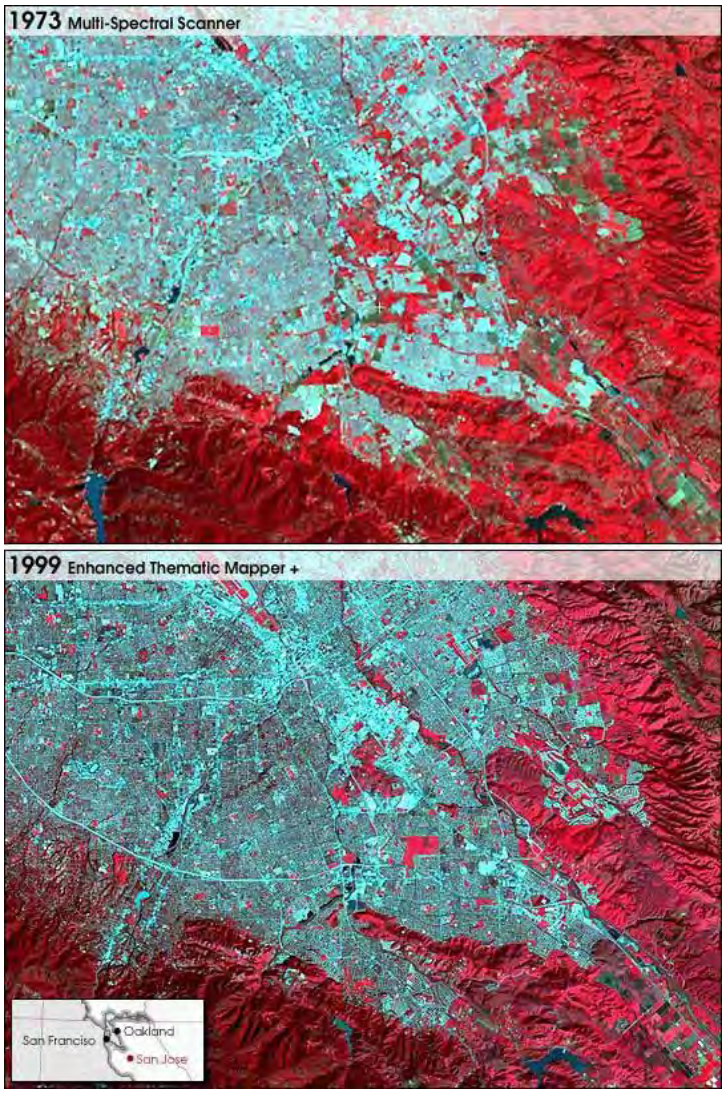
Let's leave the glitter of the L.A. world and journey south to what many believe to be the most livable city in the United States. Even the lower resolution subscenes created from MSS images are effective ways of manifesting the major features of urban areas, such as street patterns, large buildings, airports, recreational parks, and some industrial facilities. This effect is certainly evident in this March 30, 1975 MSS view of the central parts of San Diego in southernmost California. It shows the Bay harbors, military and civilian airports, the downtown, Balboa Park, Mission Bay, and Cabrillo Point.

But, the value of much higher resolution to discerning metropolitan features down to the individual building scale is convincingly revealed by this November 22, 1988, SPOT image of inner San Diego, in which the three multispectral HRV band images, at 20 m (66 ft) resolution, are combined and registered with the 10 m (33 ft) panchromatic image taken simultaneously. The details in this image are persuasive proof that civilian space imagery can match, at least, the smaller scale aerial photos in clarity and information content.

Many who live or visit San Diego consider it the ideal metropolitan area in the United States in which to live. Its climate is famed for its moderate temperatures and high frequency of sunny days. Its downtown waterfront skyline, shown below, defines its modern fast growth.

4-4: **First, with the help of your atlas, locate the San Diego landmarks mentioned above. Then, look at the U.S. Naval Air Station at the broad North Island at the end of the sea spit that makes up the Coronado area west of San Diego Bay. Using both the MSS and SPOT renditions, note any changes you can detect on this Island. Look for another change to the north. ANSWER**

America's Southwest and West Coast are among the fastest growing in population within this country. This is brought home by the next image, a change detection depiction of part of San Jose, California, a city that has more than doubled in size since 1973.



A dramatic example of building a city from "scratch" is found near the southern tip of the next state to the northwest of Arizona. Here in the midst of a flat desert basin surrounded by mountains lies the gambling heart of North America, Las Vegas, Nevada. This rapidly growing urban area is now approaching two and a half million in population. Water to nourish this "oasis" is drawn from Lake Mead, behind the Hoover Dam.

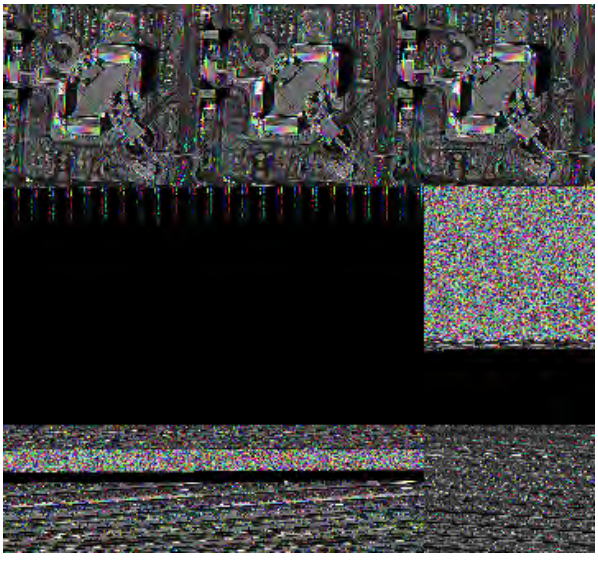
Las Vegas is built around a long avenue known as the "Strip", seen here during its prime time at night. You should be able to pick out that narrow area in the Landsat image.

Here it is in daylight as seen by a satellite:

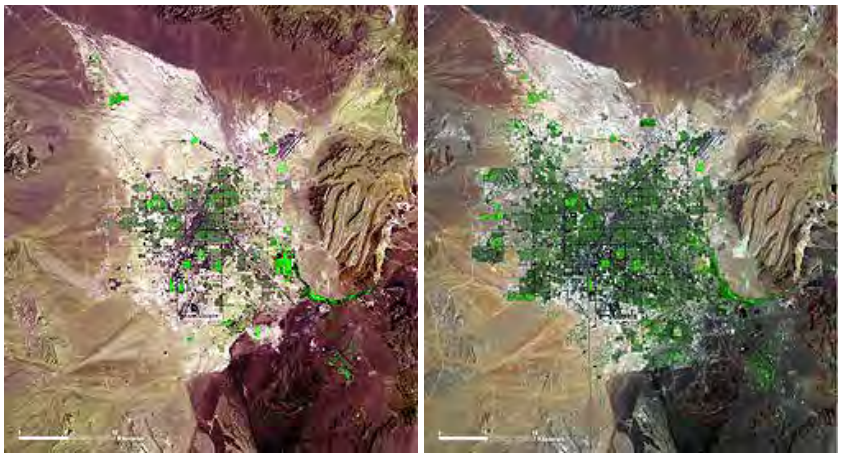
To see a very large but detailed image (Terra's ASTER) of Las Vegas, press here (this avoids an initial download that would have lengthened the time to complete this page; if you access the image, press BACK to return to this page).

Here is an IKONOS image of most of the Strip (near top of image); McCarron Field is in the lower left. We have deliberately left this oversized (for some screens) and have put a picturesque map below, so that you can scroll back and forth to try to locate some of the famous strip hotels in the IKONOS image. A cheap way to visit but not nearly as good as the real thing!

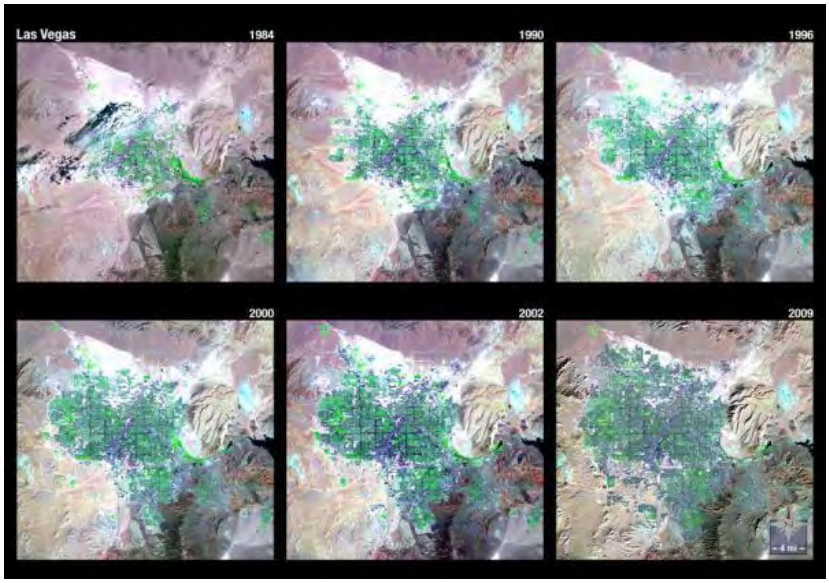
IKONOS can produce images with even higher resolution. But the one available online is made by Digital Globe's Quickbird-2, achieving a resolution of 0.68 m. Shown here is the 28-story twin towers of the Excalibur Hotel, on the Strip, and opened in June of 1990.



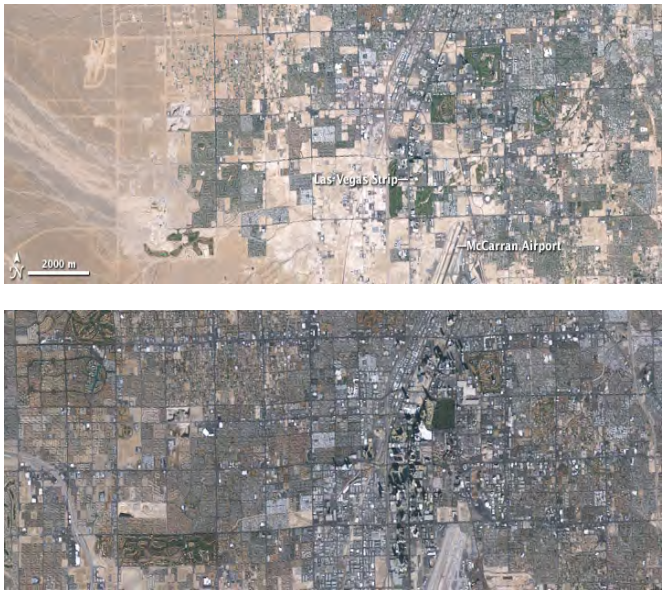
In 1973, Las Vegas had about 358,000 residents; by 1992 the city had grown to more than 937,000. Growth of the city is evident in comparing a 1973 Landsat subscene to one acquired in 1992.



This next illustration gives a more detailed account of Las Vegas' growth by showing 6 different years over a span of 25 during which Landsat-5 has been in operation:



In more detail, here is the Las Vegas strip area in 1984 and again in 2009:



Growth continued unabated through 2006, when the population topped 2,013,000. The city has been spreading south and east. But water availability problems have now slowed growth.

The main competition for Las Vegas as a gambling center is Reno, NV (once also known as the Divorce Capital of the U.S.). Here it is in a space image and from the ground;

We now look at another major western urban area in a somewhat different setting, Tucson, Arizona, nestled between mountains on three sides:

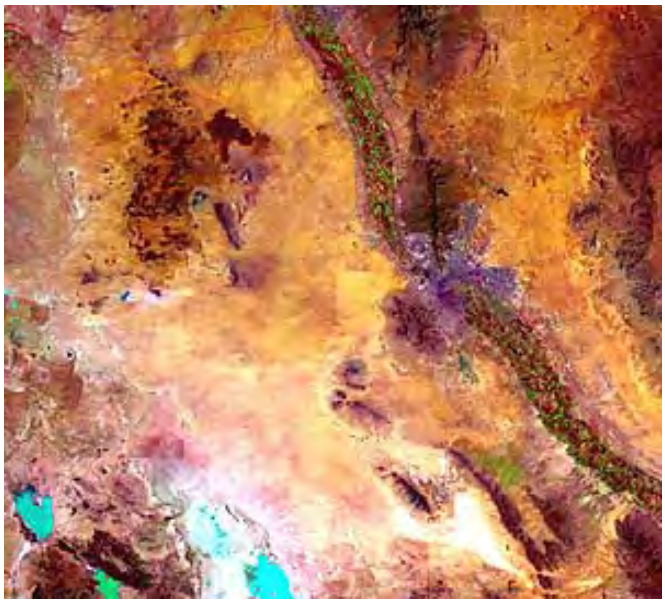
Much of the city, displayed here in a Landsat TM False Color Composite subscene extending about 58 km (36 mi) east-to-west, sits on nearly flat terrain, wedged in by several large alluvial fans from the southeast and north. The latter fan actually is becoming a major residential suburb, noted for fine homes amidst dissected canyons. The entire city area was built up from stream deposits coming off the surrounding mountains. The tip of the Rincon Mountains touches the eastern edge of the city, and the scattered hills of the Tucson Mountains extend to the west. Copper mining is a major regional industry, as evidenced by the dark fan-shaped lake and the bluish-white open-extraction pits of the Pima Mission Mines to the south.

The city itself is enlarged in this next view to bring out some of the details of the town; note the fan topography on the top.

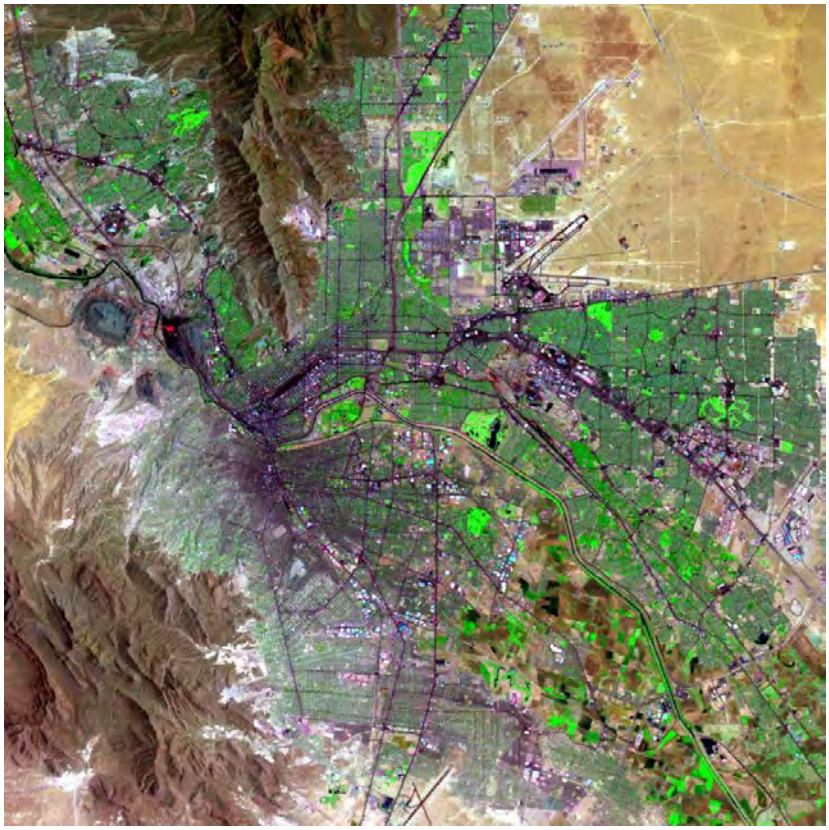
Many prominent, wide streets criss-cross the town, being laid out on the surveyor's section (one square mile) lines in the Township System. From the east, the Rio Rillito joins the Santa Cruz river (north-south) just north of the city's center. Tucson's International Airport sits to the south, about five miles from Davis-Monthan Air Force Base to its northeast. North of the city are the Santa Catalina Mountains, rising to 2800 m (9184 ft) at Mt. Lemmon, where citizens of this desert town in a basin around 825 m (2706 ft) can ski in the winter.

4-5: **Much of the greater Tucson area has a brownish tone. Why?** **ANSWER**

The writer (NMS) spent nearly all of his army career (1946-47) in El Paso, at Fort Bliss. In the next image, El Paso is the dark bluish-gray area near the bottom left in the upper right quadrant. Across from it is the Mexican town of Juarez (very popular nightlife for the soldiers). The Rio Grande, dividing New Mexico-Texas from our southern neighbor, is marked by the patchwork of farms using irrigation from that river. Volcanic flows are evident in the almost uninhabited desert to the west. The blue patches are normally dry (playa) lakes temporarily covered with a few inches of monsoon rain water. El Paso is sited at the bottom end of the Franklin Mountains (right of center); next up to the north is the Organ Mountains (top edge), near which was the firing range for the 90 mm Anti-aircraft guns the writer's unit was responsible for. The area covered by El Paso-Juarez has grown by at least 50% since 1947:



This image shows El Paso and Juarez in their regional context. Since the writer' extended stay there, El Paso has spread out in two directions - a broad area west of the Franklin Mountains, and downstream along the Rio Grande River, which appears as a narrow black ribbon in this satellite view.



As it so happened, the 100,000th photograph taken by astronauts aboard the International Space Station shows El Paso and Juarez in much more detail. In the scene below, Fort Bliss is outlined by a faint square, somewhat below the lettering for "El Paso". Note that the Rio Grande is lined with dikes on both sides.

Before we leave the western North American continent, lets look a one more example of pronounced growth. Mexico City has been forecast to enlarge to well above 20 million people by the mid-21st century (by then Tokyo will have gone above 30 million as the most populated urban area on Earth). Here is Mexico City in two satellite images, the left (or top) one taken in 1973 (population then at 9 million) and the right in 2000 (14+ million)

Let's divert this urban odyssey to other parts of the United States.

PREVIOUS NEXT

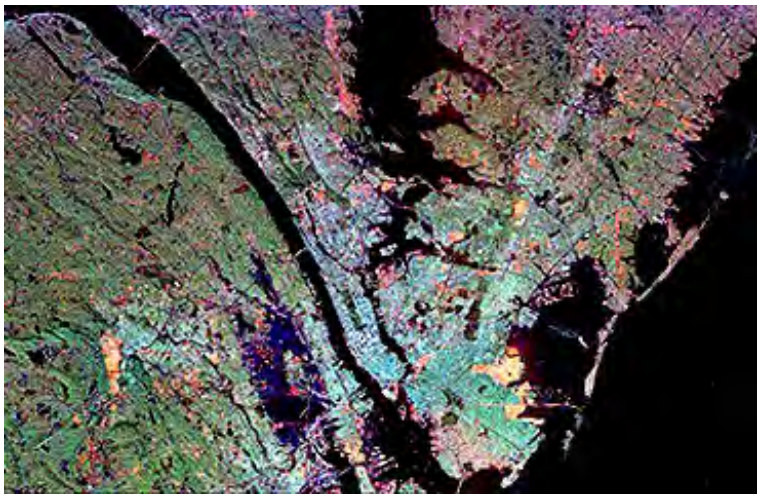
Primary Author: Nicholas M. Short, Sr.

The focus of attention now shifts to representative cities in the eastern half of the United States. Images from several satellites are used to highlight the appearance, at different resolutions, of New York, Miami, Atlanta, New Orleans, Houston, Dallas-Fort Worth, and St.Louis. The latest version of this page contains IKONOS images and various photos of the horrendous damage and loss of life at the World Trade Center Twin Towers in New York City, resulting from a terrorist attack on Sept. 11, 2001.

New York, Miami, Atlanta, New Orleans, Houston, Dallas-Fort Worth, St Louis, and Honolulu

The higher resolution of the Landsat TM and the commercial satellites (e.g., IKONOS) facilitates better recognition of many individual buildings, downtown layouts, shopping centers, industrial complexes, and other landmarks in a city. From these images we could draw a street map for most urban areas, if we enlarge the image to fill the screen. We illustrate this with four more examples: the first is New York City and surroundings, then Miami, the largest city on the Florida peninsula, followed by the great Mardi Gras city of New Orleans, Louisiana, and finally Atlanta, Georgia. They are all similar to Los Angeles and San Diego, being metropolitan areas near sea level.

To start, first look at this SIR-C radar color composite that covers the greater New York-New Jersey-Long Island region, which brings out some of the underlying topography. Note the Long Island Sound and Jamaica Bay:



Now, let's look at another Landsat image that focuses on New York City and the New Jersey cities across the Hudson River (you saw this area earlier in the "exam" at the close of Section 1).

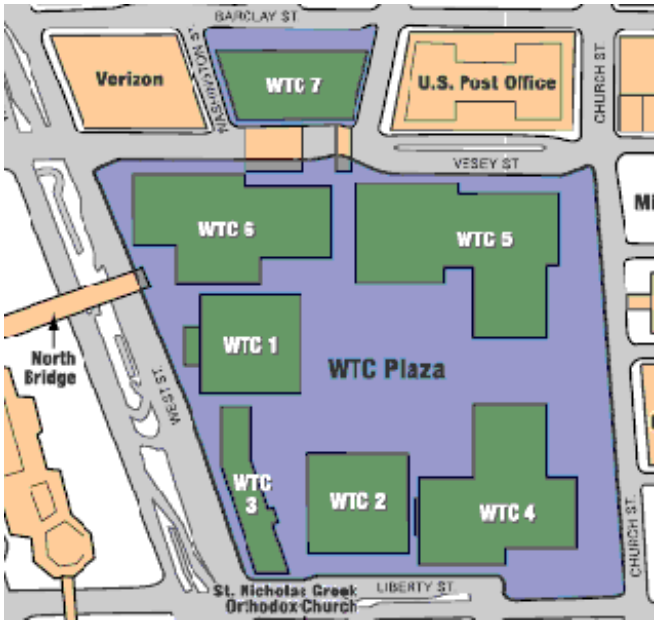
For reference, familiarize yourself with this map of New York City's five boroughs;

All five boroughs of New York, including Staten Island, are contained in the above subscene. Newark, Jersey City, and Paterson are also shown. Look for John F. Kennedy Airport and Central Park in Manhattan. Now, find that Park again in this Landsat-7 subscene in which the 20 meter panchromatic image has been combined with TM Bands 1, 2, and 3.

Now let's get down to Earth by zeroing in on Manhattan from the ground, first with a skyline photo of Midtown Manhattan and then a skyline photo of Lower Manhattan from the west (across the Hudson River) and an aerial oblique photo looking north at the skyscrapers in Lower Manhattan. (A little tidbit: from a distance one notices that there are two clusters of skyscrapers separated by a gap of lower buildings [mini-skyscrapers]. Why? Because of the geology underlying Manhattan Island - the tall skyscrapers are built on two outliers of hard, firm crystalline rock foundations but the land between contains a hundred feet or so of weak sediment unable to sustain the weight of tall buildings.)

You may recall that we introduced the topic of the infamous 9/11 attacks on the World Trade Cemter on page 2 of the Overview. Now it is appropriate to consider this historic event in detail, with emphasis on the role remote sensing has played.

The World Trade Center is actually a group of buildings, the two tallest - 110 stories high - being WTC 1 and 2:



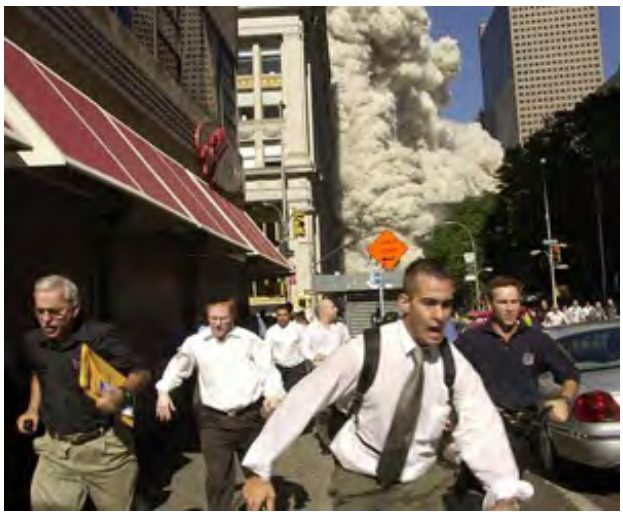
Approximately at 9 AM Tuesday on September 11, 2001 a massive terrorist attack on the United States began with the direct crashing of two highjacked commercial U.S. jetliners onto Towers 1 and 2 of the World Trade Center in lower Manhattan. This diagram presents the

timeline for the two crashes:

The first crash was unexpected - hence there is no photo of the actual collision. The second hit was captured even as it was happening since by then newspeople and photographers were on the scene:

Both buildings, about 1300 feet high, were so severely damaged that they collapsed within an hour. The South Tower went first, as shown in these two ground photos:

The collapse of each building caused a ground surge of dust (mostly fine concrete) and debris, similar to the base surge accompanying some volcanic eruptions. People fled from this rapidly advancing debris and some were engulfed by it (with no short term damage, since the dust was cool; many - firemen, police, and workers - experienced long term pulmonary problems afterwards).



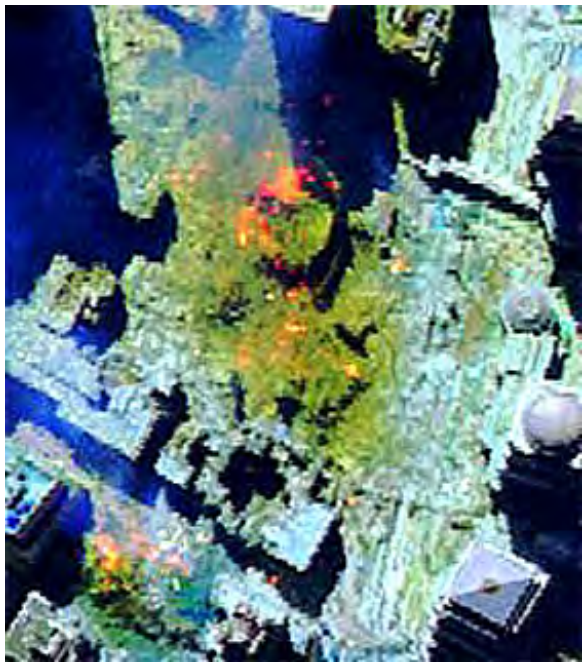
More than 2700 people lost their lives (and about 270 others elsewhere at the Pentagon near Washington, D.C. and 40+ in a rural area of southwest Pennsylvania, where the passengers on United Flight 93 heroically took over the plane as it headed for the U.S. Capitol) from airplanes deliberately used as "human-filled" rockets. The two ground photos below were taken from the same vantage point north of the twin Towers; the first shortly after the second plane had collided with the South Tower and the second soon after both towers had collapsed almost to their bases.

Space Imaging Inc. quickly provided two IKONOS images of the New York scene, one taken before the event and the second within a day after the horrific destruction of these, and several adjacent, buildings. Compare these two views, which illustrate the capability of space imagery to provide an overview of a disaster in an urban center:

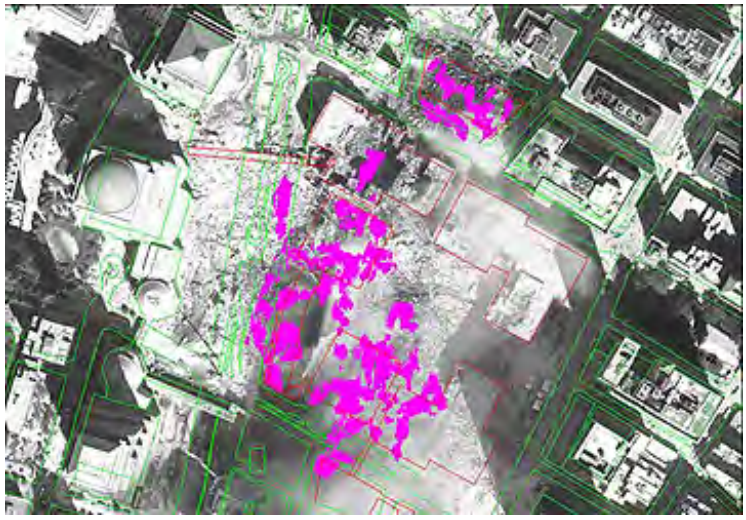
On Saturday, September 15, after much of the smoke and dust was cleared by a heavy rain, IKONOS captured this 1-meter resolution image (a bit degraded in this rendition) of the World Trade Center devastation and the avenues filled with rescue and recovery vehicles:

Seen from the ground, the extent of the devastation in the WTC complex produces horrific scenes such as this:

Also, several days later, the AVIRIS hyperspectral sensor, flown on an aircraft (see Section 13 for details on this instrument), was directed over the site. With one of its long wavelength bands as part of a color composite, AVIRIS was able to detect "hot spots" of smoldering flames, as seen in this image:

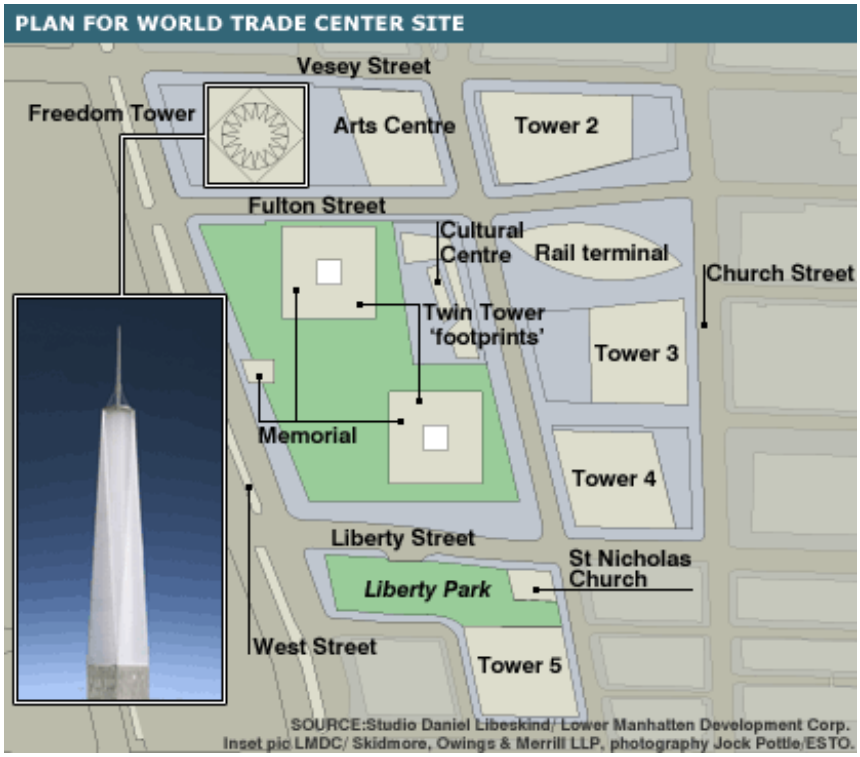


A thermal sensor mounted on an aircraft captured this map of hot spots on September 16:



Nearly two years later, on September 7, 2003, IKONOS imaged ground zero after the time that the destruction site was declared cleared of recovered bodies and debris.

Parts of the site are being prepared for a major rebuilding and Memorial to 9/11. A large Observation building temporarily straddles part of the open area:



There are those (wags) who think that the WTC attack could have been thwarted. The savior they select is none other than that famous gorilla - "you big Ape" says Ann Darrow - in the scifi classic "King Kong" (in the first [1934] version which puts Kong atop the Empire State Building, not the second [shown here] and third film versions):

Seriously, as we leave this topic, 9/11 brought about a rebirth and rededication in the United States that still pervades its society. This is epitomized by this photograph taken soon after the catastrophe:

">

But life went on during the early days of 9/11 and by the next year some sense of normalcy had returned. Sports played a large role. This Digital Globe Quickbird 2 m resolution image embraces Yankee Stadium, home of the Bronx Bombers - look with nostalgia as this venerable landmark will be abandoned at the end of the 2008 baseball season as the NY Yankees move to a new stadium:

Before we leave this fabulous city, and to break the thread to its recent tragedy, lets consider the region in terms of the marvels of its night life. Here is a night photograph taken from Space Shuttle (Mission STS-36) that shows the five boroughs, Long Island, and the New Jersey side of the Hudson:

But, let us next return to more halcyon days before the Terrorist War, and look at southern Florida. We establish a framework for the Miami metropolis by looking at this large Landsat mosaic (the subject of mosaics is treated in Section 7) constructed from several individual TM scenes, each rendered in quasi-natural color.

This mosaic extends from the Florida Keys at the bottom to the Orlando area in the center of state at the top. To the east is Cape Canaveral, site of all manned and many unmanned NASA space launches. Having set up your bearings, look for the following features in an atlas: the Tampa-St. Petersburg area around Tampa Bay; Fort Myers; Sanibel Island; Lake Okeechobee; the Everglades; Alligator Alley; Fort Lauderdale; the greater Miami region. The famed Everglades (examined in detail on page 3-8) is essentially a wide river of sluggish drainage, flowing over sink

depressions through a vast expanse of open water and land covered by sawgrass and hammocks - "tree islands", irregular, often elongated patches of hardwoods. The Everglades, a subtropical environment, is being ecologically threatened as building and drainage to convert to farmlands has seriously disturbed its natural state. Along the southern coast are mangrove swamps and marshes. An extensive farm region has been cultivated in reclaimed land between Lake Okeechobee and the northern edge of the present Everglades. The widely dispersed lake country in central Florida develops from sinkholes and other internal drainage features in the limestone bedrock.

4-6: Find each of the above features in the mosaic, drawing on an atlas as your guide. Where are the Florida Keys? ANSWER

Next, we view a false-color Landsat TM subscene that covers the Miami area and much of the Florida Gold Coast, from West Palm Beach to Fort Lauderdale.

The city and its suburbs generally hug the coastline, being prevented from spreading inland by the eastern edge of the Everglades. The elongated red features that seem to follow curvilinear tracks are hammocks, whose distribution follows water flow patterns. The linear red features are sloughs dredged by engineers to better control flow. The offshore bar that makes up Miami Beach is conspicuous. The slanted linear feature passing through Miami is the Miami Canal that parallels Route 27. Reclaimed swamplands to the north are converted to large farm tracts.

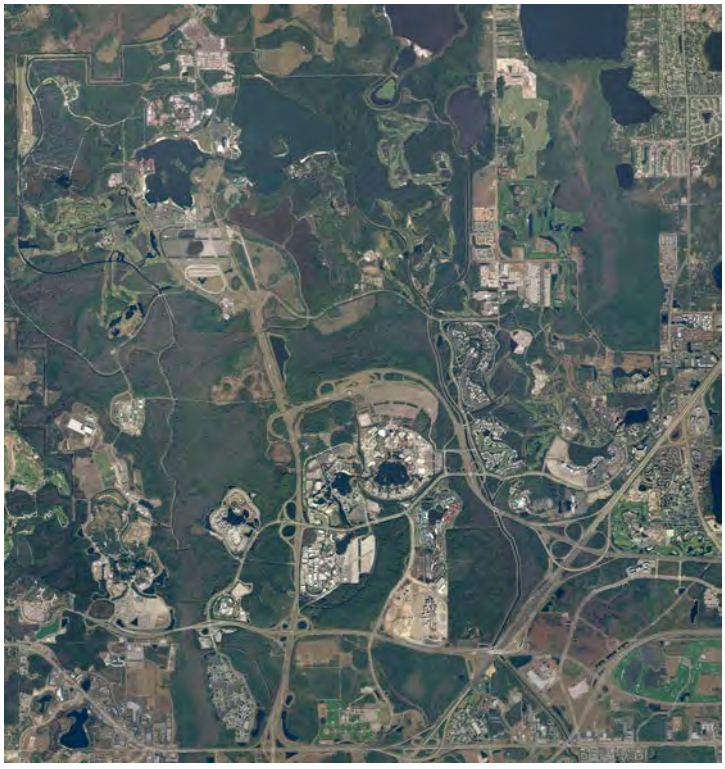
Like most large American cities today, Miami's downtown has an impressive collection of tall buildings:

4-7: **Along the coast, locate Coral Gables, Key Biscayne, Miami Beach, Hollywood. Is Fort Lauderdale in the scene? ANSWER**

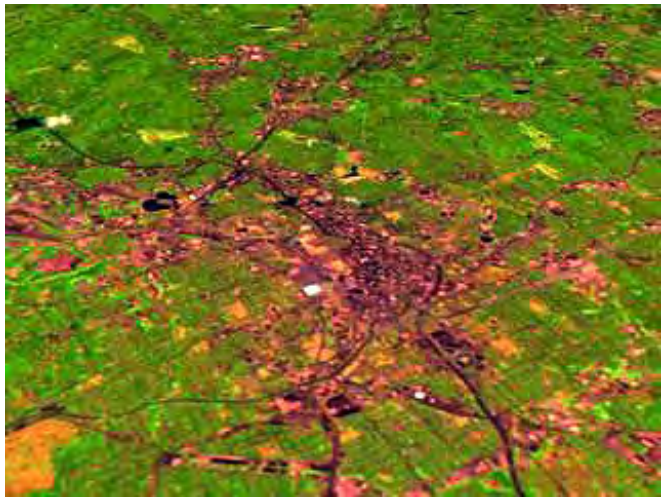
Like some other urban areas in this Section, Miami has been imaged at high resolution, here by IKONOS.

Jacksonville, Orlando, and Tampa-St.Petersburg are the other major population centers in Florida (now holding some 18 million people). Here is a space view of Tampa:

Probably the most visited area in Florida is in and around Orlando. Go east and one gets to the Kennedy Space Center. Go about 20 miles to the west and, 'lo and behold', there is DisneyWorld. These next three pictures tell the story (see their captions):



One of the fastest growing cities in the eastern U.S. in last 30 years is Atlanta, Georgia. Here is a Landsat-5 subscene taken on June 8, 1999.



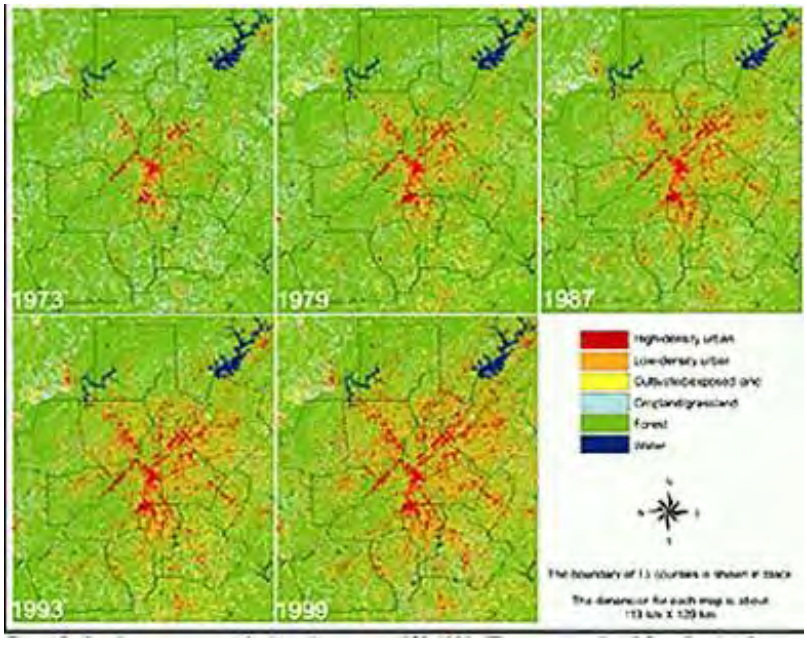
This next image was made by an airborne multispectral scanner; rather surprising is the low number of high rise buildings in the downtown area. Most growth has led to a great expansion of suburban areas.

Obviously, even more information about an urban center will result from higher resolution imagery. This next subscene shows part of central Atlanta, Georgia, as imaged by the Kosmo's 2-meter KVR-1000 film camera, which flew on the Russian MIR Space Station. This space photo has many of the qualities and details of aerial photos. Features of interest include the Atlanta-Fulton Co. Stadium, until 1997 home of the Braves baseball team, and the convoluted intersection of Interstates 20 (east-west) and 75-85 (north-south).

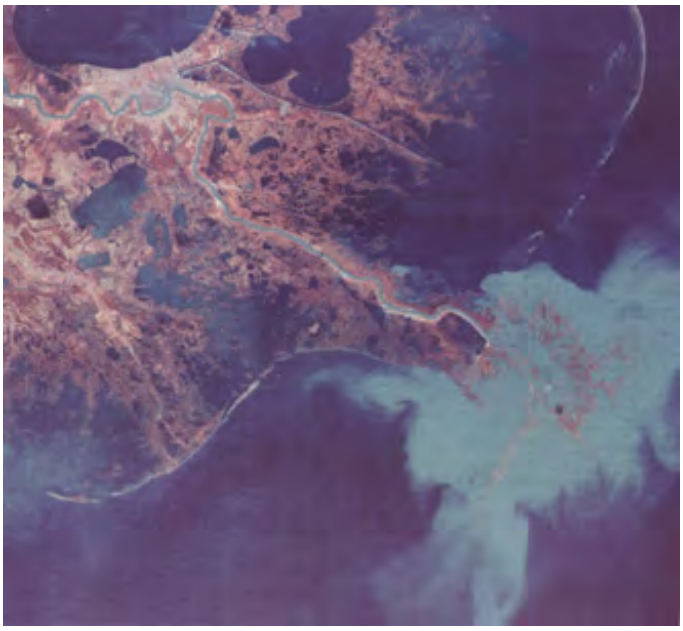
Now, this sports complex has expanded, as shown in this recent photo with the downtown in the background:

You should be able to match this photo to the Kosmos scene above.

Using Landsat data over the last 26 years, these subimages (reworked as classifications) really reveal the magnitude of urban growth in the Atlanta region:



Turning to New Orleans (treated in detail on page 14-10a), let's set the city into context with its surroundings by looking at two images. The image below is a full Landsat MSS image, cropped to remove the parallelogram effect, in which the city appears in the upper left between Lake Ponchartrain and meanders of the Mississippi River. This January 1973 scene is dominated by the "birds-foot" distributaries and bayous of the Mississippi Delta. They are set off by the subdued false-color reds of the swampy vegetation in moderate winter dormancy, and by the pronounced chalky blue of sediments from the river, as it empties at the currently-active distributary.



Zeroing in on the city itself, this Landsat-7 natural color subscene shows the entire city - siturated mostly on the east side of the Mississippi - and parts of Lake Ponchartrain (note the causeway bridge that was built from New Orleans north from the south edge of the lake).

Running through the city is part of the Gulf Intracoastal Waterway, a continuous 1725 km (1072 mi) long water route from Carrabelle, Florida to Brownsville, Texas. It also connects with the Mississippi River to provide a protected canal-like passageway for pleasure boats and freight barges. Where it doesn't use the open Gulf of Mexico, the waterway is cut inland from the shore to an average depth of 3.7 m (12 ft) and width of 38 m (125 ft). This waterway also merges in the scene with the Inner Harbor Navigation Canal that joins Lake Pontchartrain to the Mississippi River. Note the wharfs along it. Interstate 10 runs eastward here passing through the Gentilly section of New Orleans. Highway 46 follows the river through the suburb of Chalmette. To its

north is a large swampy area, or bayou, in St. Bernard Parish. Much of this area is heavily vegetated, but open water (blackish) takes up most of the interior. The New Orleans Lakefront Airport juts into Lake Pontchartrain. Green areas to its southwest are Pontchartrain Park and the University of New Orleans campus.

The Landsat subscene covers a part of the eastern half of the Crescent City, starting with just a bit of the world-famous French Quarter, at the bottom-left corner of the image (Bourbon Street near the Cathedral of St. Louis just enters there). The French Quarter extends from the docks along the Mississippi River, which has powder blue color, resulting from its high silt content that imparts increased reflectance in TM Band 2. Many of the city streets form true square blocks. Some of the larger buildings in this scene are resolvable at the TM's 30 m. resolution. The reds in the city proper are signs of the luxuriant vegetation that is profuse throughout this lovely city. The large body of water, at the top, is the south shore of Lake Pontchartrain, a 57 km (35 mi) wide (1619 sq. km; 625 sq. mi) inland, brackish (mildly salty) lake that connects by an eastward channel to Lake Borgne, which is a saltwater body joined by the Mississippi Sound to the Gulf of Mexico.

Below is a night view of central New Orleans - this city really comes to life after dark, when the humid heat cools and the activities that make this a Fun City commence. Still, New Orleans is a major port and center for the oil industry. The third image shows the downtown skyline.

The next image is a still closer look, consisting of a multiband (with RGB color assignments) SIR-C radar image, taken from the Space Shuttle on October 2, 1994. Compare this with the Landsat-7 image, noting similarities and differences.

Now look at another view of central New Orleans, lying mainly to the southwest of the Landsat subscene. This subscene, about 9 km (6 mi) on a side, shows the city at 10 meters resolution, as imaged by the multispectral scanner on JERS-1, a Japanese satellite launched in 1992. Note the radial pattern of the streets that act like spokes against 000the crescent bend of the Mississippi River. The large white round feature is the Superdome, site of several football Superbowls. Tall buildings downtown are just to its right. The dark patch nearby is astride the Vieux Carre or French Quarter.

4-8: **Locate the SIR-C, Landsat 5, and JERS-1 images in the above Landsat-7 MSS natural color subscene. ANSWER**

Below is a nightime ground photo of Bourbon Street within the French Quarter (see map shown):

Part of the French Quarter, and another New Orleans landmark, is Jackson Square (named after General [President] Andrew Jackson, the hero of the Battle of New Orleans between the British and the U.S. in the War of 1812); it took place a few weeks after hostilities in the eastern U.S. had ceased by armistice but word of this hadn't reached those in Louisiana.

The French Quarter, in the opinion of most who know 'haute cuisine', has the largest number of fine restaurants in a small space anywhere in the U.S., and rivals Paris for the quality and variety of eating places. Check out the most famous at this Restaurants website.

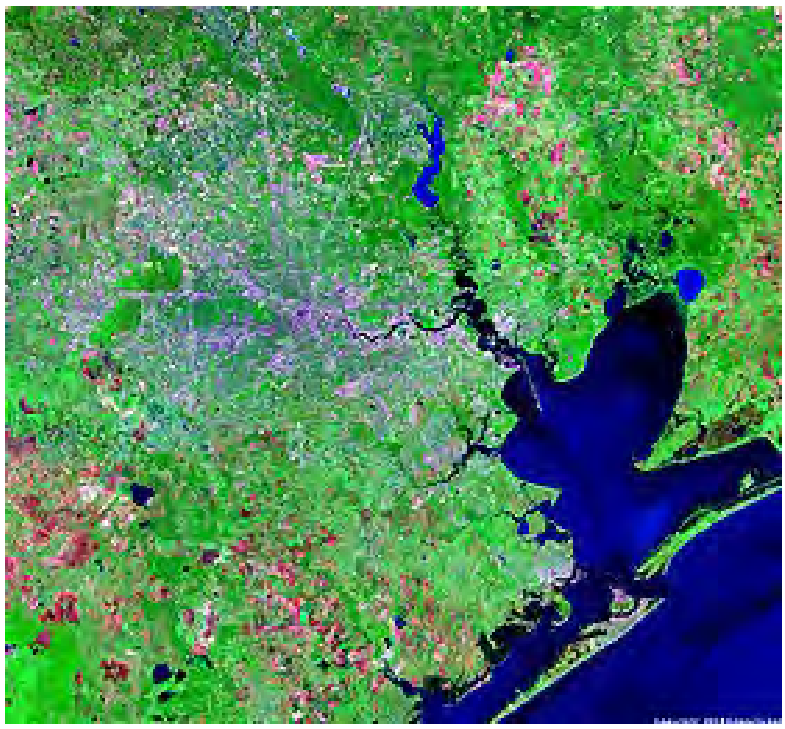
A tradition after a night touring French Quarter restaurants and bars is to end the evening at

the Cafe du Monde (off Jackson Square) for coffee and beignets. The writer (NMS) has done this 3 times; it remains his fondest memory of New Orleans.

The next big city west of New Orleans, and the country's fourth largest, is Houston. Let's look at this city using a variety of space imagery.

The urban area is the home of most of the astronauts, who are headquartered at the Johnson Space Center, Clear Lake. For decades now, almost whenever an astronaut was in a space craft orbiting Earth, he/she would look out as the crew passed over the Houston metropolitan area - and do what any real tourist does - **TAKE A PHOTO**. Here is a typical picture:

Houston is in the pinewoods of East Texas. In fact, it was carved out of cleared woodlands. Even in the metropolitan area, trees abound in every neighborhood. A sense of the "greenness" of the region is given by this quasi-true color Landsat-7 ETM+ image:

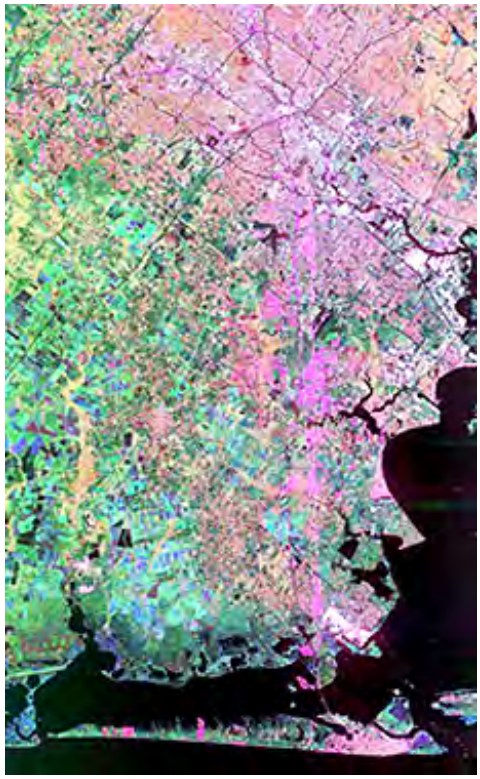


The greater Houston area is home to more than 5 million people. The extent of the metropolitan area stands out in this astronaut photo taken at night:

A better look at Houston and its environs is provided by MODIS (on Terra):

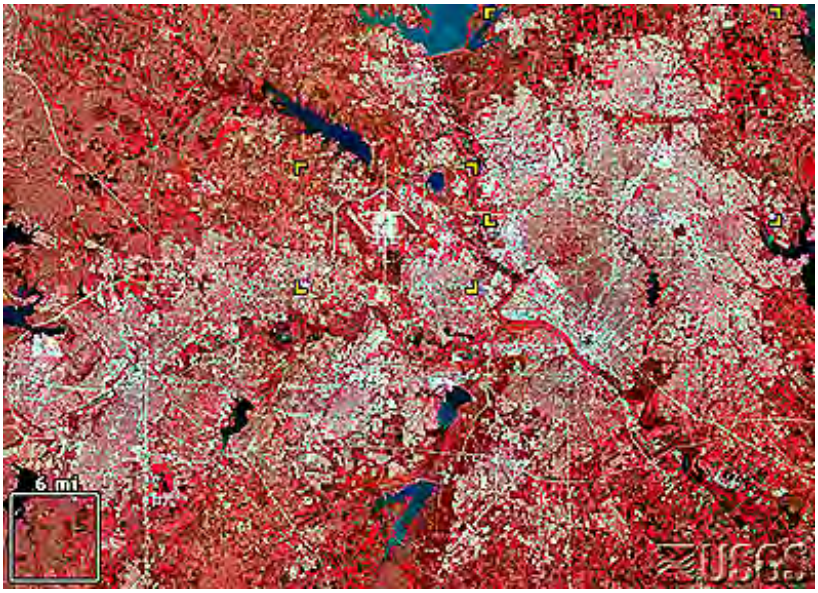
Here is a false color rendition made by the IRS satellite; at the higher resolution, the shadows of downtown buildings are visible:

Now look at this SIR-C X-Band radar image that extends from central Houston to the Gulf Coast at Galveston Island.



The photo below was taken by an astronaut aboard the International Space Station. The camera used is capable of much better resolution than has been acquired in most previous manned missions from space (see Section 12). The views of downtown Houston shows a part of the freeway system built through and around this city. Both Houston and Dallas-Fort Worth (below) have more major urban area road networks than most other American cities, owing to the foresight of transportation planners in recognizing that future growth requires such roadways even at the expense of replacing (removing) homes, etc. to provide the necessary right-of-ways.

Staying in Texas: One of the prime uses of space imagery is that aspect of change detection which examines an urban scene after intervals of years or even decades. This is an invaluable aid to regional planners and even to the Bureau of the Census, which has learned how to use these observations in estimating population changes (helpful to the U.S. Congress in re-allocating federal funds). The next two images are of the Dallas-Fort Worth, TX area (see also page 4-5), in 1974 and 1989.



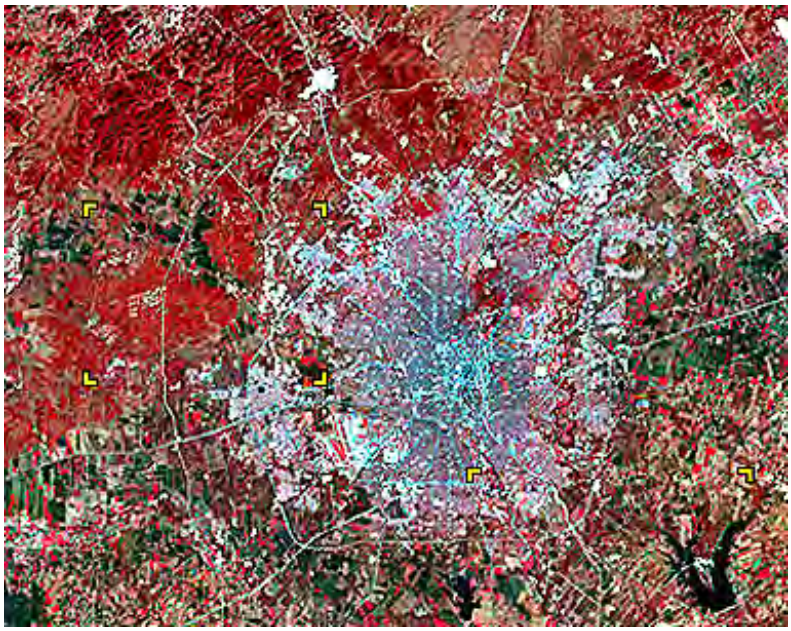
Both of these cities are now major urban centers - a conclusion one would draw from these views of their downtown skylines:

SKYLINE OF DOWNTOWN DALLAS

SKYLINE OF DOWNTOWN FORT WORTH

We will leave it as an exercise for you to evaluate the changes in the 15 years since the 1974 subscene. Are there any new lakes (reservoirs)?

Throughout the Tutorial, but mostly in this Section, we have a number of scenes of Texas. We have covered all Texas' major cities except San Antonio. To avoid possible negativity from its citizens, we show this Landsat image of the city built around the Alamo. Here:



The writer (NMS) would be remiss if he did not feature his home town in this Section. So, below is a Landsat 4 subscene showing St. Louis - the pride of eastern Missouri - the town where, just to its north, the Missouri and Mississippi Rivers meet.

A bit more detail is evident in this Google Earth-Quickbird image:

This is a photo of the St.Louis downtown area, including the stadium that Mark McGwire has made famous and the Gateway Arch, a monument symbolic of the opening of the American West to pioneer settlement:

The downtown section shown in the oblique view is rendered vertically in this March 2002 Quickbird (0.66 m; 2 ft) image:

Four landmarks are singled out: the Gateway Arch; Busch Stadium (St. Louis Cardinals); the TWA Convention Center, and (second from left) the Eads Bridge, first one built across the Mississippi River, considered an engineering marvel in the 1800s (but now closed to auto traffic, but kept intact for its historical significance).

And, for a different perspective, look at this SIR-C radar (from the Shuttle) composite (see Section 8) of the central city, with East St. Louis on the Illinois side (bottom).

Let's close this page with a quick look at our westernmost big city. This next image, made by the Landsat 7 ETM+, was added on December 7, 2001, the 60th anniversary of the Pearl Harbor attack. This is the southern part of the island of Oahu (see page 17-3 for image of Hawaiian Islands).

In the above image, the location of Pearl Harbor relative to the main section of Honolulu is evident. In the next image, taken by International Space Station astronauts, shows that section in more detail (higher resolution). It also better reveals the nature of Diamond Head - it is a small volcano (now extinct) with a wide central crater.

Now, let's head back to the eastern U.S.

PREVIOUS NEXT

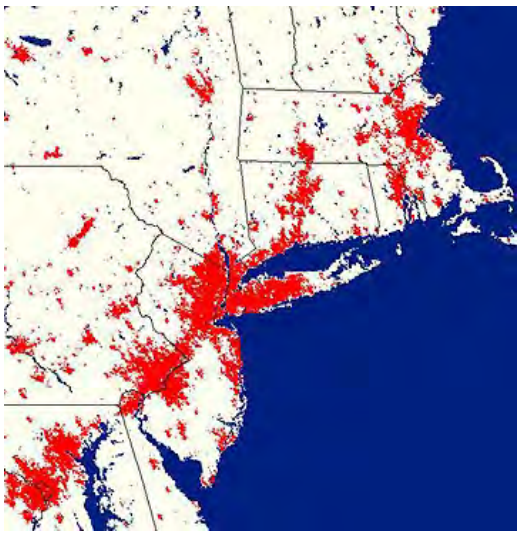
Primary Author: Nicholas M. Short, Sr.

Most United States citizens and many foreigners eventually visit Washington, D.C. to take in the wonders and sights of that Nation's Capital. This page takes one on just such a tour using a combination of ground scenes, aerial photography, and space images. You will be asked to identify a number of the famed monuments and federal buildings, using an atlas if necessary. You will also visit its neighbor city, Baltimore, and the City of Brotherly Love, Philadelphia.

The U.S. Capital: Washington, D.C.; Baltimore, MD; Philadelphia, PA

The cities visited on this page are part of the so-called Metropolitan Corridor that extends from Boston to Richmond. Much of this corridor was photographed by astronauts on Space Shuttle (STS-098) as seen by night, with urban areas in orange owing to the high density of light sources, beginning in Greater New York and extending south of Washington-Baltimore towards Richmond:

Using a variety of satellite imagery, this map shows in red the major metropolitan areas in the Northeast Corridor. Compare with the night view above.



4-9: **This general question, or activity, applies to this whole page. Each of the images represents a higher resolution version of the preceding image, generally with a different sensor/satellite system. Try to fit each successive one to its predecessor. Locate the landmarks and features that are described in the text. Your biggest challenge will be to relate features in the radar image to the others.** **ANSWER**

We've already examined New York City from space, and Boston is considered on page 6-2. Let's

move downcorridor to Baltimore, which was thriving during the Revolutionary War, and to Washington, an outgrowth of that war.

The first capital of the just emerged United States of America, during the Revolutionary War and for about 5 years thereafter, was the then largest city (about 30000) - Philadelphia (see bottom of this page). For part of the "Glorious Cause" it was occupied by the British under General Howe but later was evacuated by their Army. In 1790 Congress decided to move to an area where a new capital could be built from scratch. Both the northern and southern states vied for the privilege of hosting this seat of power. In 1791, President George Washington was given the task of selecting the exact site, somewhere on the Potomac River. He chose an area in Maryland that had been designated federal land, convenient to his Mt. Vernon home some 16 km (10 miles) downriver. The only settlements of significance near it at the time were Alexandria, VA and Georgetown, MD. The construction of the capital followed closely the master plan devised by Monsieur Pierre L'Enfant of France. Before Washington died in 1799 the building was well underway. To honor the "Father of the Country", the resulting city was named Washington, in the District of Columbia (DC).

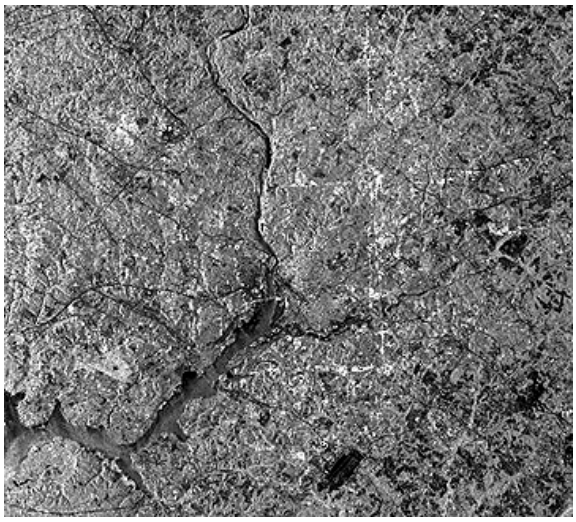
Let's start our space tour of the capital with a new millenium image (Spring, 2000) of parts of Maryland, Virginia, and New Jersey, and all of Delaware, as imaged by Terra's MISR, one of the instruments onboard Terra (see page 16-9). Most of the Chesapeake Bay and all of the Delmarva Peninsula appear. Washington and Baltimore in this rendition are bluish-white. At the head of Chesapeake Bay, the orange-brown patch denotes sediment carried into the Bay from the Susquehanna River.

We next show those two cities below in a famous subscene (about 113 km [70 mi] across) that was taken from the October 11, 1972 pass of Landsat-1 on a remarkable viewing day in which the air was crystal clear after a major storm passed.

The effects of that storm are evidenced by the light blue tones in the Potomac River, resulting from a heavy silt load dumped by runoff from the rainfall upstream. The increase in sediment did not affect the Chesapeake Bay to the east. The inner city areas of Baltimore and Washington are accentuated by the blue tones that signify numerous large buildings and a limited number of trees. More than five million people live in this urban region but much of the population dwells in

residential areas in which tree-dotted landscapes remain (recent surveys identified the Washington area, in particular, as the most heavily forested urban locale in the eastern U.S.). A large blotch of deep red (abundance of trees) in the northeast part of the Washington suburbs represents the largely undeveloped tract of land comprising the Beltsville Agricultural Experiment Station, which the U.S. Department of Agriculture operates. Interstate 95 between Washington and Baltimore - the heaviest traveled major road in the East - stands out in contrast to its vegetated edges (look closely for it - it is a thin blue line). Highway 50 past Annapolis, the capital of Maryland, and the high Bay Bridge over the Chesapeake Narrows, visible in the original print, are not seeable on screen.

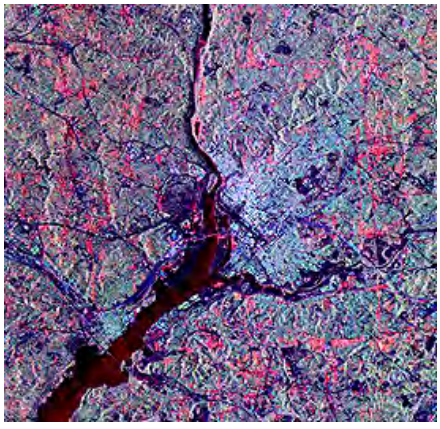
We use a SIR-C L-band radar image, which was obtained on April 18, 1994, from the Space Shuttle to look closer at the metropolitan area surrounding the U.S. capital.



In this image, north is to the upper right. No distinct tonal patterns marking the more densely populated and built-up sections of the area are evident, but several of the major roads leading in and out appear as thin dark lines. Two dark patches coincide with the National Airport (just below the center) and Andrews Air Force base (lower right). Farmlands (diminishing as outlying areas are developed into suburbs) make up many of the other darker patterns. We find most of these farms to the east in the Coastal Plains.

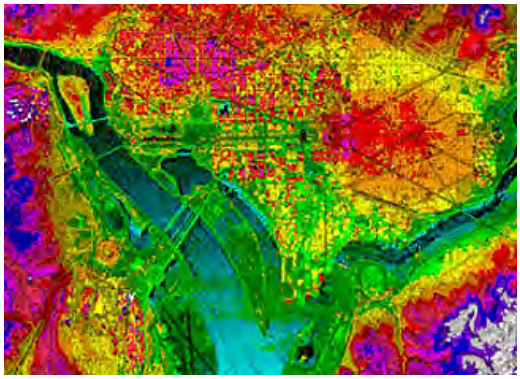
The Coastal Plains' precise juncture (at the Fall Line) with the hilly Piedmont (see map on page 6-1 and accompanying text) to the west is obscure but the rolling terrain on the left side of the image infers where the Piedmont begins. A remarkable feature of this image is the three light-toned lines forming part of a square. These lines coincide with the boundary between the District of Columbia and Maryland. Staff at the Jet Propulsion Laboratory (SIR-C's operators) interpreted this tonal phenomenon as expressing strong reflections from building corners that line avenues along the boundaries. But, this writer, who lived in the area for 21 years, is skeptical of the explanation, because there is no unusual concentration of large buildings lining the streets along the boundaries. Along the NE-SW edge many building are residences no different than other homes away from the boundary streets. The line next to the boundary running NW-SE (right side of square) that continues beyond the District is a power line.

Some of these features stand out better when the several radar bands on SIR-C are combined to make this color composite; note that the square boundaries still persist (the red blotches along them are probably not vegetation):



This radar image of part of D.C. was made by an aircraft system operated by JPL as part of its interferometry program to make topographic maps (see page 11-10). The IFSAR system uses two radar images, taken moments apart, that are thus staggered spatially to get a "stereo"

effect. In this image, the colors represent variations in elevation, which, although generally less than 30 meters (100 ft) in relief, can still be determined by the radar interferometric method:



The next two images, made by SPOT sensors, present much more detail in the central metropolitan area around the District of Columbia. The false color version at the top (about 24 km [15 mi] on a side) is made by the HRV multispectral sensor that provides 20 meter resolution. On the bottom is a 10 meter panchromatic image that extends about 13 km (8 mi) on a side, showing the central city and its many federal buildings.

The HRV image falls just within the Capitol Beltway, which rings the Washington area. Several vegetated areas are worth identifying: the wooded area known as Rock Creek Park; Roosevelt Island in the Potomac; the Arlington National Cemetery; and the National Arboretum. The circular Robert F. Kennedy (RFK) stadium is easily identified. The panchromatic SPOT image shows the layout of buildings around the Mall and the downtown business area, as well as homes (many are joined row houses) in the central District. Consult an atlas map to confirm the location of most of the well-known individual edifices. Use major streets such as Independence, Constitution, Pennsylvania, New York, and Massachusetts Avenues, and the Interstate 395 extension, so help locate sites.

The centerpoint in downtown Washington is the Mall that runs from the Lincoln Memorial (4, in

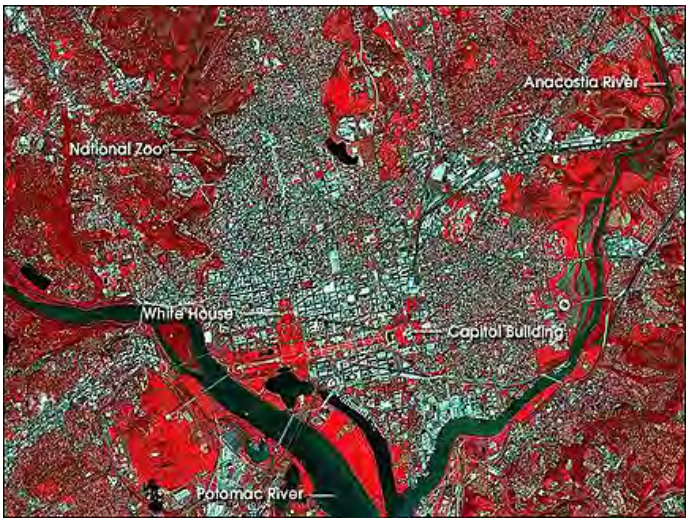
picture below) eastward to the U.S. Capitol (3) (Note: The spelling "Capitol" is used to specify the building and immediate surroundings and the hill [Capitol Hill] on which they are located to avoid possible confusion with other federal aspects of the nation's capital.). These two structures, along with the Kennedy Center for the Performing Arts (5), are examples of the graceful white marble edifices that house many of the government offices, monuments, and cultural centers, in what many cite as one of the most beautiful capital cities in the world.

This aerial oblique color photograph, which was taken from the Virginia side of the Potomac affords a "birds-eye" view of this central area.

Visitors to the capital city should recognize many of the landmarks in this picture. On the middle right edge is the Pentagon, and part of the National Cemetery is in the foreground. Beyond Roosevelt Island is the Kennedy Center. Nearby are the Lincoln Memorial and, on the Tidal Basin, the Jefferson Memorial. The White House is hidden by trees but the Ellipse (a near-circular drive and parkland) is just to its south. The Washington Monument is too thin to be seen clearly in this photo. But, many government office buildings lining the mall are evident. The U.S. Capitol building at the east (far) end of the Mall is barely distinguishable in this oblique aerial photo.

The Pentagon is shown in a high resolution Quickbird image taken in August 2002, about a year after the 9/11 terrorist attacks. See if you can determine which side is under repair at this time.

Because Washington, D.C. is the home base of NASA and it is often hailed as the "Capital of the Free World (or Democracy)", it is a favorite target for imaging by nearly all the Earth-observing satellites currently flying. Here is a Terra ASTER 15 m resolution image of Washington that, even though not the equal in detail of those shown below, is, together with imagery from Terra's other sensors, evidence of that versatility:



Compare what you can see in this high resolution color aerial photograph below with the scene just above. This and subsequent images should persuade you that space "cameras" are now competitive with those flown on airplanes. For reference, the U.S. Capitol Building and its grounds lie near right center in both scenes (above and below).

The image below, displayed in near full screen size, is a merge of a 5 m IRS panchromatic image with a 30 m natural color Landsat subscene. This visually attractive rendition covers much of the Virginia side of the Federal District, including the Arlington National Cemetery and the Pentagon. The Mall layout is well-displayed. The area east of the Capitol is all residential - row houses experiencing a "renaissance" of popularity as lobbyists and others jockey to be near the center of power.

Now to an identification challenge in this next image of downtown Washington, D.C. - a 2-meter (6.6 feet) resolution space photo made by cosmonauts using the Kosmos KVR-1000 camera onboard the Russian Mir space station that was co-registered with a Landsat image.

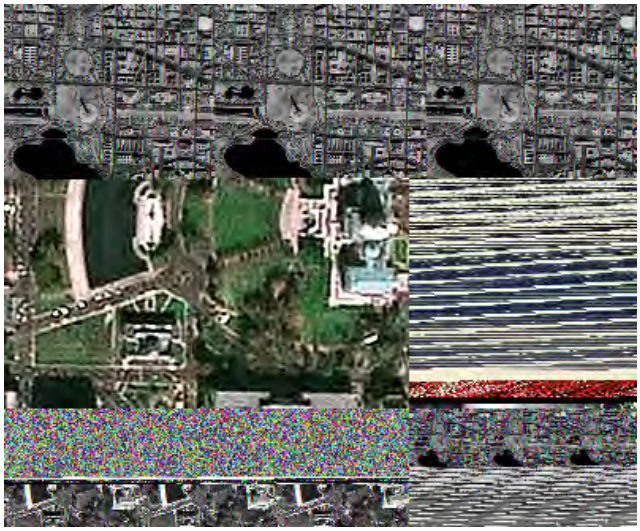
Use the map above to locate in the IRS and Kosmos images the following: Washington Monument (note its shadow in the Kosmos image); Jefferson Memorial; Lincoln Memorial; U.S. Capitol Building; Library of Congress; Union Station; National Air and Space Museum; Hirschorn Gallery; National Gallery of Art; the Kennedy Center for the Performing Arts, National Museum of Natural History; Smithsonian Administration Building; Federal Bureau of Investigation (FBI); The

White House; U.S. Dept. of State; NASA Headquarters, the U.S. Dept. of the Treasury; and Massachusetts and Constitution Avenues.

4-10: **There is something (a structure) in this image (related to the Kosmos part) that can give you the approximate time of day when Kosmos was passing over. Can you figure this out and estimate the hour. What important monument is out of the Kosmos picture? ANSWER**

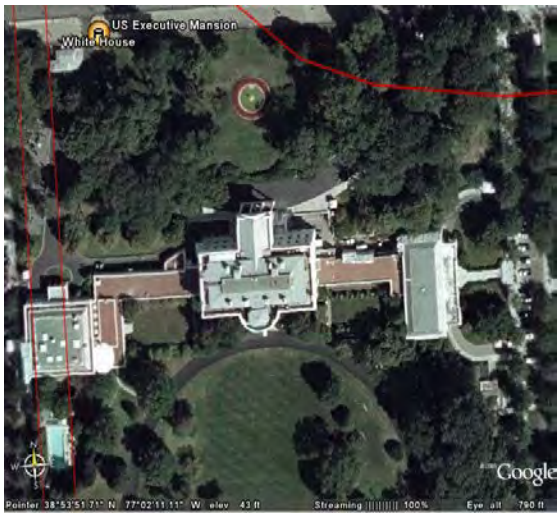
Now, we will take a closer look at the U.S. Capitol and its Reflecting Pool, as seen by SPOT-5's high resolution sensor (note the many changes in this area, compared with the 1965 image above:

Before we leave the U.S. Capitol section of Washington, let's look once again at this IKONOS-2 image showing the left half of the same area as in the above MIR KVR-1000 image. This was first introduced in the Overview to appraise you of the very significant development in applied remote sensing in which high resolution imagery (here, about 2 meters also) has come into the marketplace with the decision by the Russians to "go commercial", followed by the launching of several satellites operated by private companies. In that Overview is also the 1 meter IKONOS panchromatic black and white image of buildings and a bit of the Mall just west of the Capitol Building.

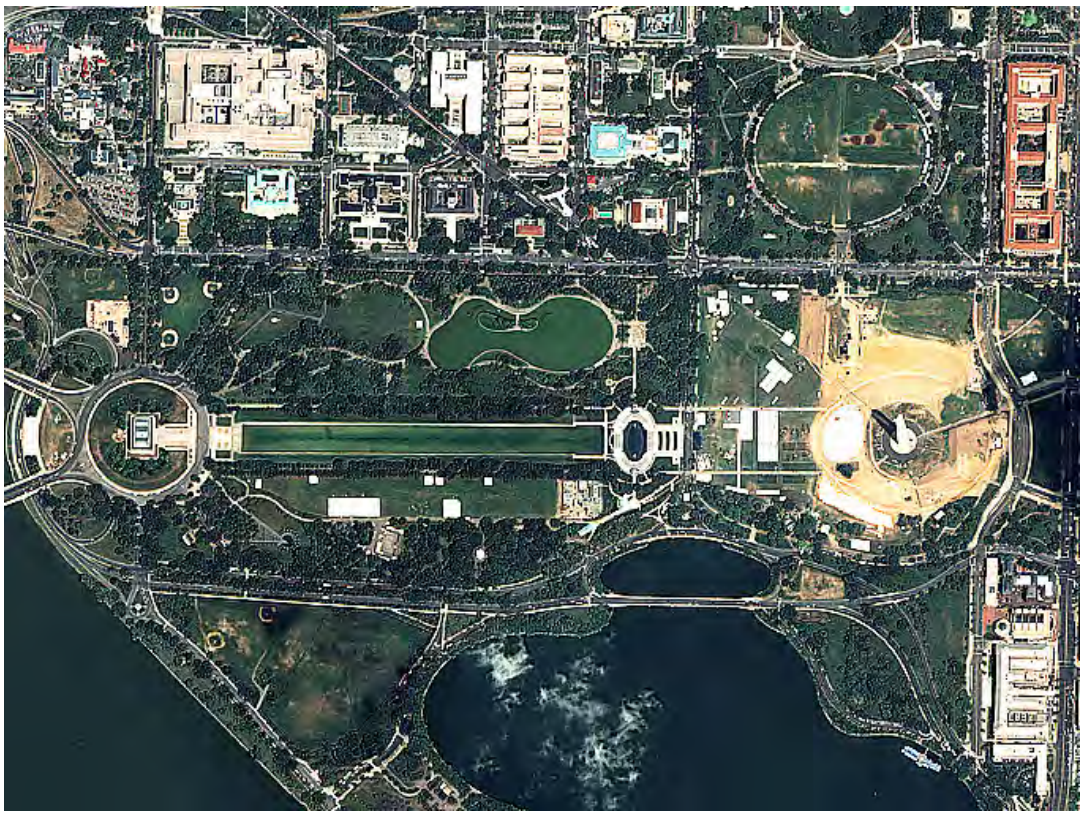


In January, 2002, an even higher resolution (<1 meter) image of the Washington Monument was provided among the first scenes released by the QuickBird satellite launched in later 2001:

Possibly the most famous building in America - perhaps in the World - is the White House, home of the President of the United States and seat of the Executive Branch. Here it is from space:



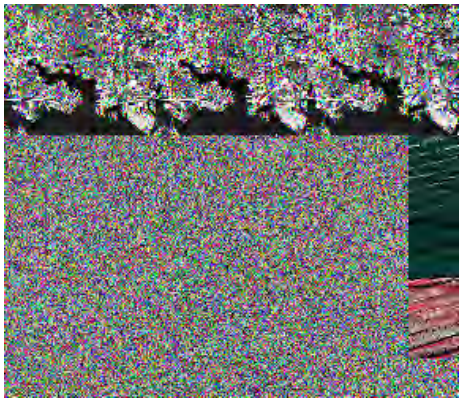
To the west of the Washington Monument, the largest new Mall structure in many decades was completed and opened in May 2004. This is the classic World War II Memorial whose location appears in this Quickbird image as a large yellow-brown (bare soil; this image shows the early phase of construction) patch near the Washington Monument (whose grounds were being renovated). During this war more than 490,000 American military were killed (second only to the Civil War in casualties).



You were informed that, in the Landsat scene, the blue area near the top of the Landsat image

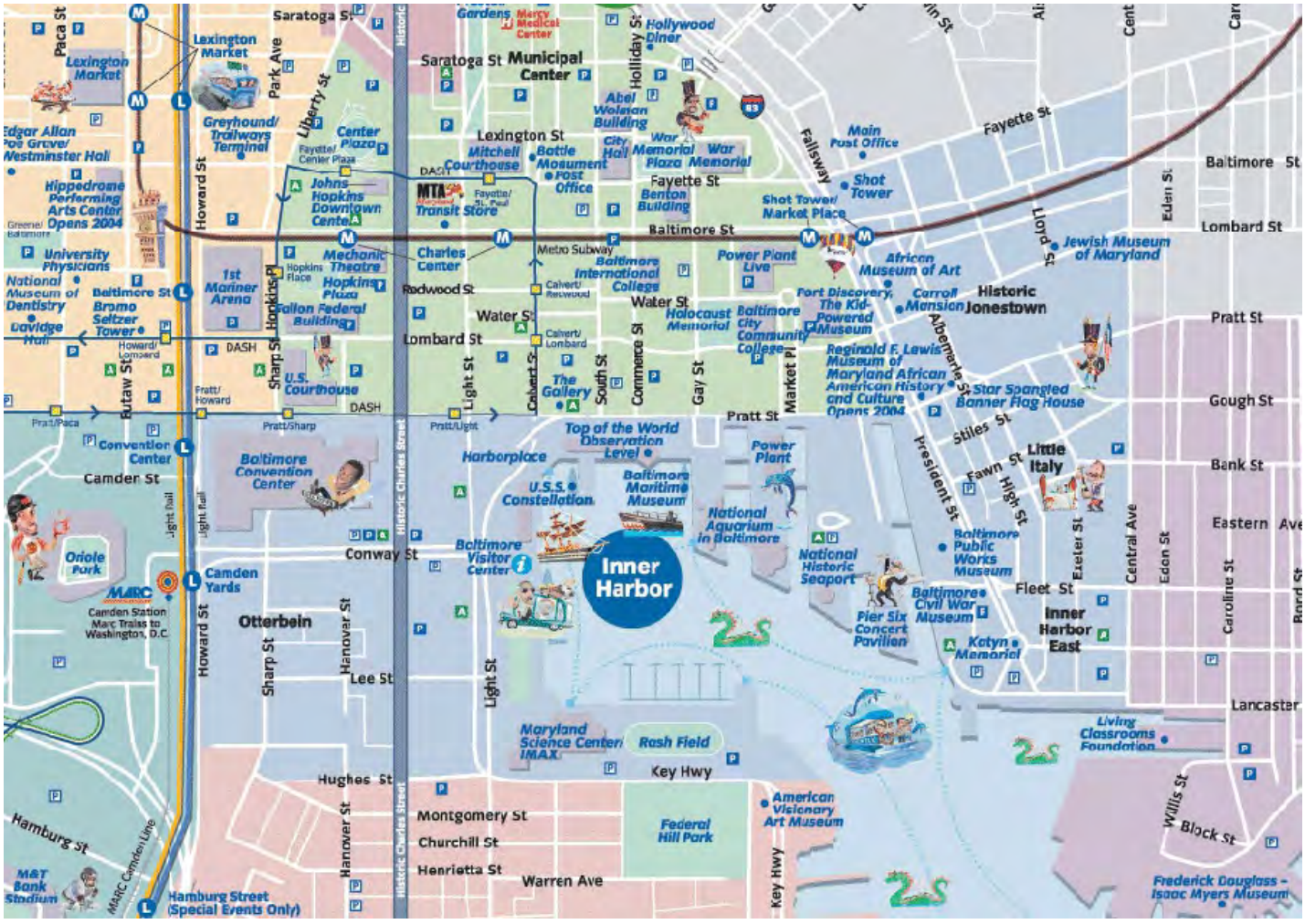
is the city of Baltimore, Maryland. Here are some scenes that capture the "flavor" of this town. The top one is a standard false color image made by the Landsat-7 ETM+. This shows most of Baltimore's central city, and its western, eastern, and southern suburbs (the northern part is not present). Conspicuous is the large inlet coming off Chesapeake Bay to the east, as rising post-glacial waters invaded the Patapsco Valley.

Second is a JERS-1 image of the downtown, with its wonderful Inner Harbor development. Below that is a color photo looking at the skyline. Unlike Washington, which has a restriction on the height of buildings to keep a balance in proportion, Baltimore, like most big U.S. cities, has an imposing set of tall buildings, as seen in the ground photo that shows much of the Inner Harbor.

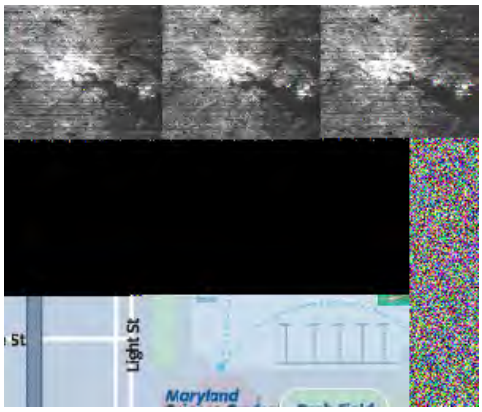


The next image was made by the IKONOS satellite. It is among several used in the movie "The Sum of All Fears", adapted from the book by Tom Clancy. In the book, the football field in Denver was the one blown up by a terrorist nuclear bomb. For the movie, the scene was shifted to Baltimore where hero Jack Ryan's wife worked as a medical doctor. In this image, details of the Inner Harbor stand out; the Camden Yards stadium, home of the Baltimore Orioles baseball team, appears near the bottom, along with part of the ("blown-up") stadium used by the Baltimore Ravens football team.

This map, produced by the Maryland Tourist Bureau, shows most of the landmarks and tourist attractions in central Baltimore:



Urban areas are normally warmer by several degrees or more relative to the suburbs and countryside. This is often referred to as the "heat island" effect. This Landsat-2 image of Baltimore shows a much warmer area (lighter tones) in the central city around the Inner Harbor:

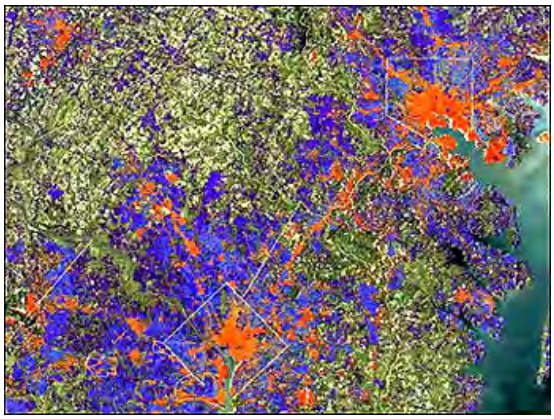


Baltimore has undergone a remarkable renaissance of the city starting in the 1970s. This is largely the result of one person - the dynamic Donald Schaefer, who was mayor for 16 years (later, Governor of Maryland). Mayor Schaefer has a passionate love for his hometown which prompted him to devise a master plan that would attract both industry and suburban citizens who were enticed to return to the central city.

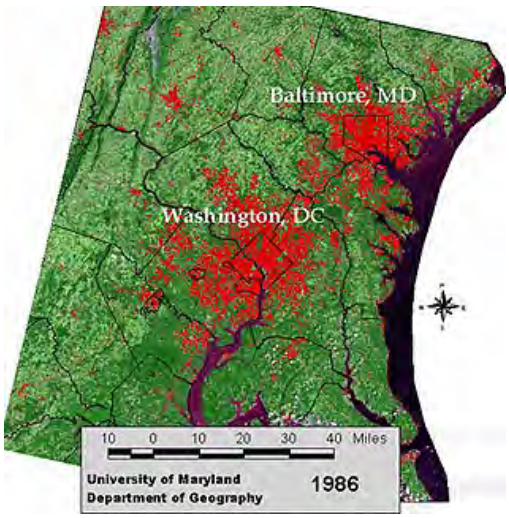
A renovated midtown with splendid buildings! Then, a neighborhood (Baltimore is famed for its ethnic neighborhood varieties) that had old granite stone cobbles in a street and beautifully restored rowhouses:

The view northeast from the Key Highway in the Federal Hill area was awesome to the writer, which included this vista:

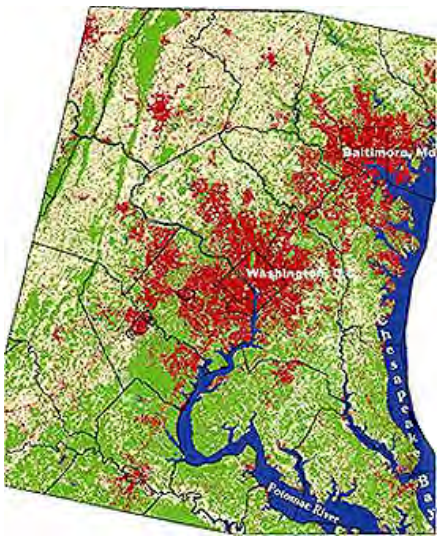
Before leaving this segment of the Metropolitan Corridor, lets look at two practical applications. A Landsat subscene that includes both Washington and Baltimore has been processed by a special algorithm to classify ground cover to determine the degree to which the surface acts as impervious (the term refers to ability of water to penetrate into the substrate) to rainfall (this is important in specifying how much water enters the ground and estimating the water that remains to cause possible flooding). Red is highly impermeable, i.e., strongly impervious; blue is moderate and green is low impervious. Reds associate with inner city dominance of buildings and roads; blues are typical of suburbs; and greens generally associate with forests and fields.



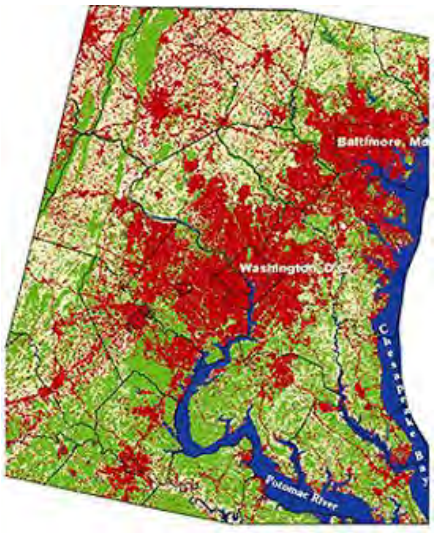
The Washington-Baltimore city environs and the "corridor" between these two metropolises that are centered about 60 km (40 miles) apart are experiencing rapid growth - especially the areas in Virginia near the Nation's capital. A visual overview of this growth is evident in this image which utilize Landsat imagery to broadly classify major urban land cover types in 1986. (This is part of a study of urban change, called SLEUTH, conducted by Claire Jantz of the University of Maryland and Scott Goetz of Woods Hole [MA] Oceanographic Institute; it is summarized on this Web Site.)



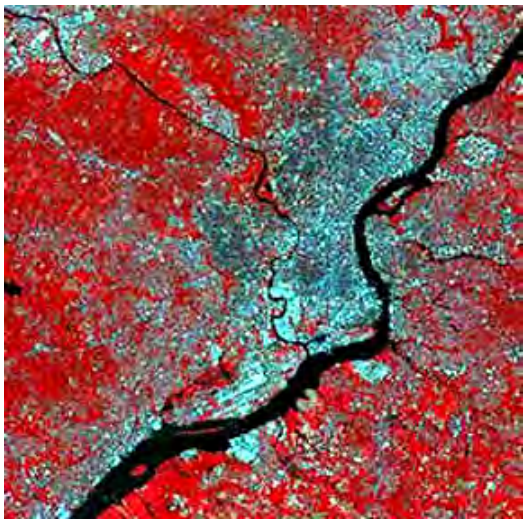
Compare the patterns of high density metropolitan population in 1986 with this SLEUTH map made from IKONOS images (used to provide more detail) obtained in the year 2000:



The SLEUTH workers have analyzed these change detection images to pick out patterns and trends in population dispersion. Using that information they have generated a forecast map indicating the population distribution in terms of land cover in the year 2030 if no major effort is made to guide and control the expansion. Here is their map in which growth is largely unplanned and follows a "free market" approach:



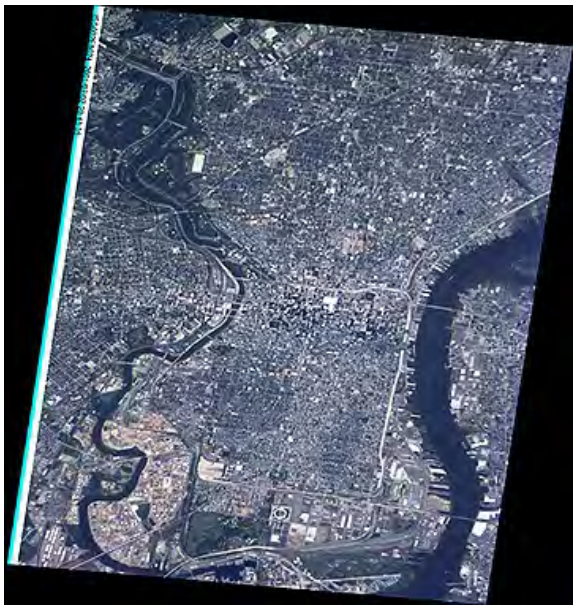
Finally, here are two Landsat images of Philadelphia, PA region - now grown to about 4 million (if the suburbs and Wilmington, Delaware are included) from the original 30000 people during the Revolutionary War:



The size of the city's environs is misleading, if one goes only by the blue areas. Most of the red areas are actually populated suburbs, which are noted for their preservation of natural and planted trees. Using the IDIMS processor at Goddard Space Flight Center, here is a density slice "map" using MSS Band 6. Six levels of grouped DNs were selected: Yellow; Red; Purple; Orange; Green; Blue. This was the result:

Green, of course, is associated with forest cover and blue with water. The denser populated areas in and around Philadelphia in yellows and reds correspond to surfaces with little or limited vegetation. The purples and oranges seem to denote built-up suburbs. Conclusion: as a first approximation, level slicing using Landsat imagery worked well in this case.

Today, Philadelphia is still among the largest cities in the U.S. This photograph was taken of the central city by astronauts onboard the International Space Station.



As one drives along the Schuylkill Expressway, next to the river of that name, the skyline of Philadelphia makes a strong impression:

In the midst of these leviathan buildings sits Independence Hall - Cradle of Democracy.

Primary Author: Nicholas M. Short, Sr.

We now leave the North American continent to look at city scenes in South America, Europe, and Asia. Images from several different space platforms are represented. The page ends with several ERTS curiosities.

Buenos Aires, Argentina; Brasilia, Brazil; Paris, France; Munich; Bavaria; Budapest; Hungary; Florence; Italy; Riyahd, Saudi Arabia; Beijing and Shanghai; China, Melbourne, Australia

We turn now to foreign cities. Let's swing south to the east coast of South America. One of the largest cities, both in areal size (1800+ square kilometers)and in population (13 million) is Buenos Aires, the capital of Argentina. Buenos Aires has been called the "Paris of South America" This aerial photo shows why this is appropriate - and we will visit Paris itself on this page.

Buenos Aires appears on the left side of the image below, a natural color photograph taken by the EarthKAM camera system mounted in the NASA International Space Station. The dimensions of this photograph are 121 km (76 miles) by 81 km (52 miles).

This great urban center was built along the west coast of the upper estuary of the Rio de la Plata. Into it flow the small Rio Pirana and large Rio Uruguay, both carrying a heavy load of silt which colors the water a reddish-brown. The agricultural fields beyond the eastern shore lie within the country of Uruguay. Similar fields abound in the lowlands around Buenos Aires. The head of the estuary is a delta covered with thick riparian vegetation.

Another South American capital, Brasilia, Brazil has a special history. It was built "from scratch" in the late 1950s and now has a population of 2 million. The layout of the city was carefully planned so as to have broad streets and open parks and clean, modern government buildings. Look first at this oblique view of the central area of this city, provided by Auguste Areal:

Digital Globe's Quickbird satellite has taken a look at the central part of Brasilia. Notice the large areas still kept as open space.

Crossing the Atlantic, the next city we visit is another nation's capital and is quite familiar to those who've traveled Europe. Paris to many is the most beautiful urban area in the world. Let's take our first look at the spread of the city around the Seine River, as imaged in false color by Envisat's ASAR radar, in which the white tones signify strong reflections of the radar beam owing to the many corners and other reflector structures from building in both Paris and its suburbs.

Here is an aerial overview looking northeast that captures the city's sprawl and centers on the famed Eiffel Tower, found on the Left Bank:

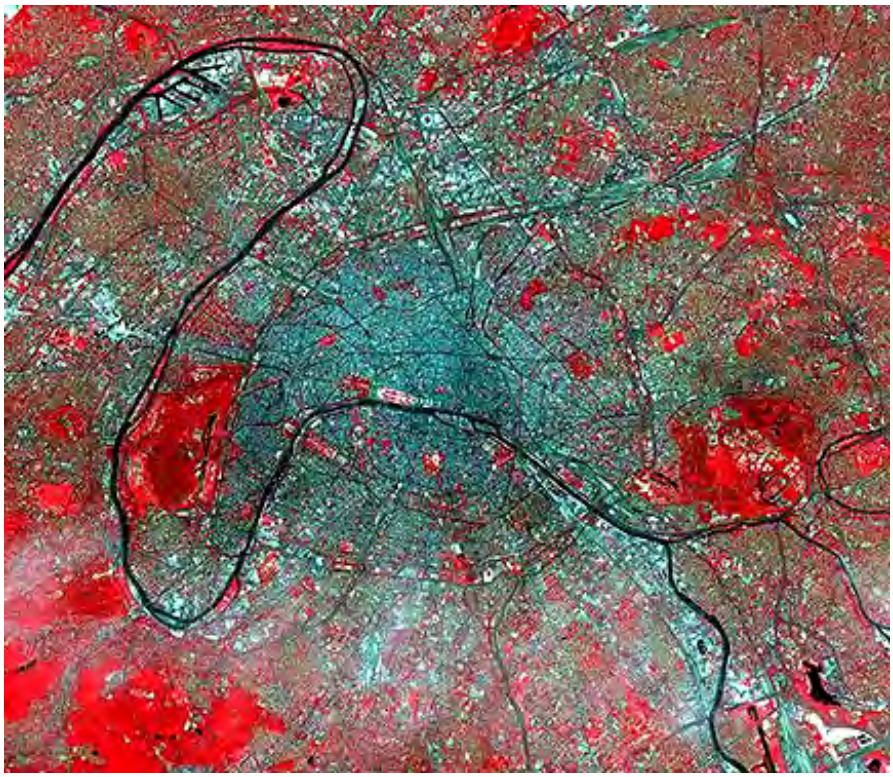
Courtesy; Carolina Map Distributors

The Eiffel Tower, the "trademark" of Paris, has been imaged at 4 meters from space by the IKONOS satellite (this image seems to have been taken either at night or when lights were on):

DigitalGlobe's Quickbird has also produced a high resolution image of the Eiffel Tower, this time showing its surroundings in color:

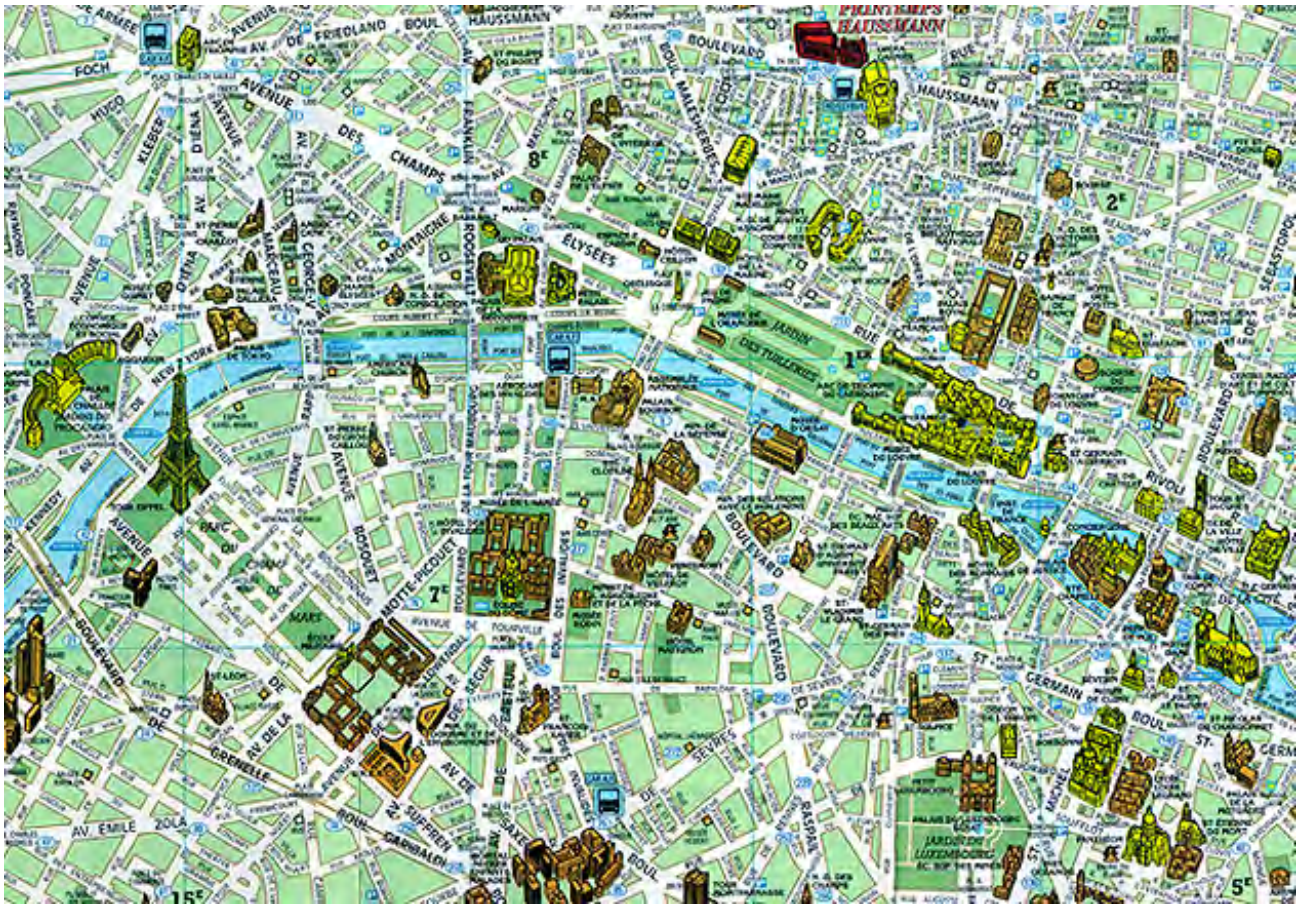
But both of these are "Johnny-come-lately's"! The U.S. military satellite KH-9 took this high resolution image in the late 1960s, more than 30 years before civilian satellites were allowed to achieve this quality of resolution.

Paris as depicted in a Spring 2000 Terra ASTER image looks like this:



Especially prominent is the strongly meandering (bending) River Seine; the small river joining from the south is the Orse. The large wooded area (red tones) within a river loop near the western limits of denser population is the Foret de St. Germaine. Central Paris, with its many famed monuments and buildings, lies just to the west of the juncture of the Marne (of World War I fame; in the lower right quadrant) with the Seine. Note the wooded areas within Paris and its suburbs (bright red tones).

Use this map to locate some of the main streets and places of interest in central Paris - the area best known to tourists. You should be able to find the equivalent area in the ASTER image. And, as a challenge, locate the IKONOS Eiffel Tower scene in the ASTER image.



Almost the same map area is seen in this astronaut photo taken from the International Space Station:

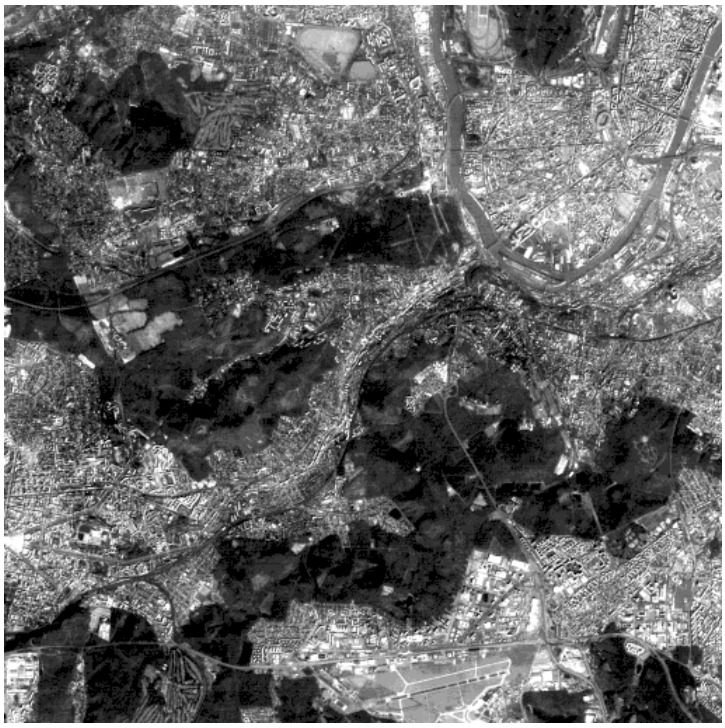
Compare this scene with that which appears in this GoogleEarth/Quickbird image :

Find the Louvre - the most famous museum in the world. It is just east of the Jardin de Tuilleries. At the southwest corner is the building known as L'Orangeries - it houses many of the best French Impressionist paintings, including 8 Van Gogh's at the top of the stairs to the second floor, which stopped the writer (NMS) "dead in his tracks" when first seen (in 1960).

Locate the Hotel des Invalides in the Google version (it is just to the right of the words "Motte Piccoute" near the center of the map above). This building is a major tourist stop that contains Napoleon's tomb. Look also for the Luxembourg Palace.

Almost as famed as the Eiffel Tower as a symbol of Paris is this Notre Dame Cathedral. Here are two views, one from space (Quickbird/Google Earth), the other from across the Seine:

Look now at a small part of this great metropolis as seen at 10 m resolution by the panchromatic sensor on SPOT-3; blackish areas are intracity forests.



4-11:**Try to fit the SPOT subscene in the ASTER image.** **ANSWER**

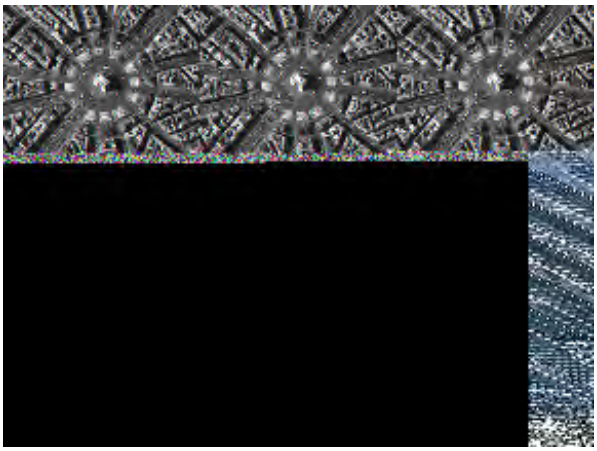
The upper right corner is about 1.2 km (0.75 mi) southwest of the Eiffel Tower, the symbol of this city. The panorama seen in the next picture is a view which I (NMS) took from this tower looking generally southwest towards the SPOT scene. Prominent is the stadium known as the Hippodrome de St. Cloud.

4-12: **Is any of the SPOT scene in the photograph from the Eiffel Tower? ANSWER**

The south end of a great meander loop of the Seine River that passes across this section of the city is prominent in the SPOT image. Within the loop is a community known as Boulogne Billancourt. The southern end of the Bois de Boulogne, the city's largest park, appears as a dark area at the top right. This aerial photo shows the wooded park and a group of tall buildings to its west; the foreground in this picture corresponds to the low height buildings in the photo from the Eiffel Tower; in the background are tall buildings in the Hauts de Seine.

The major road crossing the SPOT image center is a large expressway, the AutoRoute de l'Ouest (A13) to Rouen and then Le Havre on the English Channel. Adjacent is the Parc de St. Cloud, within the inner suburb by that name. The lower half of the image shows Highway A86 just above the Villacoublay airport. Route 10 winds northeast between the two irregular wooded areas bounded by Meudon on the east and the center of Versailles on the west. Route 118 passes northward through Velizy-Villacoubay, joining 10 near the Seine loop.

Paris has many circular roads (roundabouts) with major converging streets. The most famous, perhaps in the world, is the 8-lane circular road around the Arc de Triomphe, seen here in this Quickbird-2 image (about 2 meters resolution).

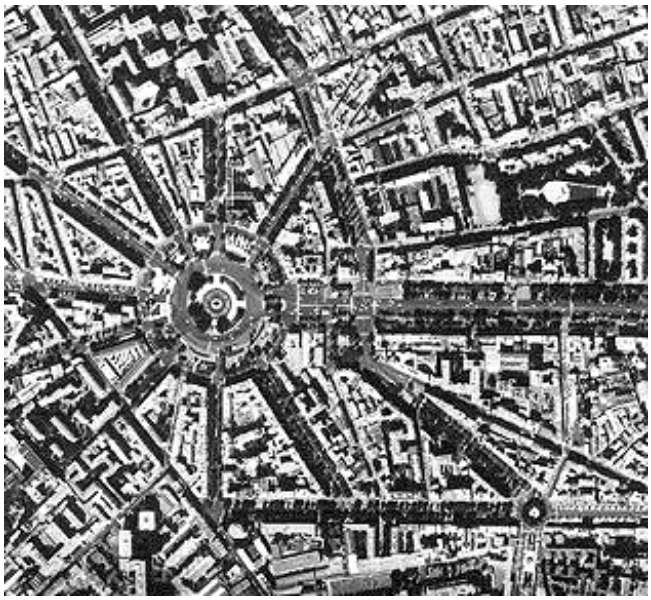


SPOT-5 has also looked at this famous area of Paris, with its HRG camera that produces images in the 2 meter range. Fit it into the ASTER image above (not easy; there is something "wrong" that may trick you).

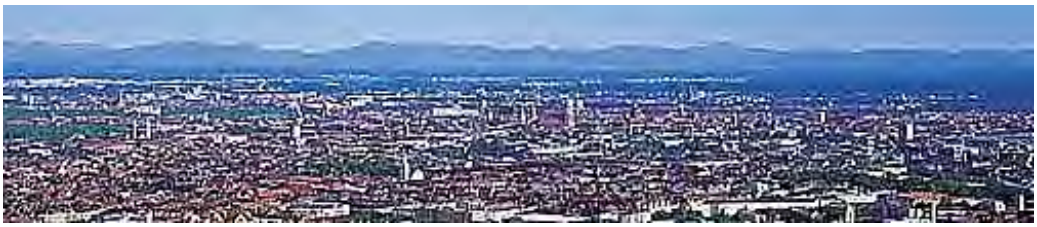
4-13: **Inspect the SPOT-5 image and orient its scene relative to the Quickbird image. Clue: something is amiss with the Quickbird image. To assist you, we insert a map of the Arc de Triomphe area below this question.** **ANSWER**

Another roundabout is the Place de la Nation in the eastern part of the city. This is what it looks like at 2 meter resolution using the panchromatic digital camera KVR-1000 on the Russian SPIN-

2 satellite launched in 1998; the wide roadway to the east is the Cours de Vincennes.



With nostalgia, we must say au revoir to the City of Lights. Next, we cross the Rhine into Germany's Bavaria to visit the city of Munich. To realize its relation to the Bavarian Alps to the south, here is the city's setting seen in a panoramic photo:



Courtesy: Carolina Map Distributors

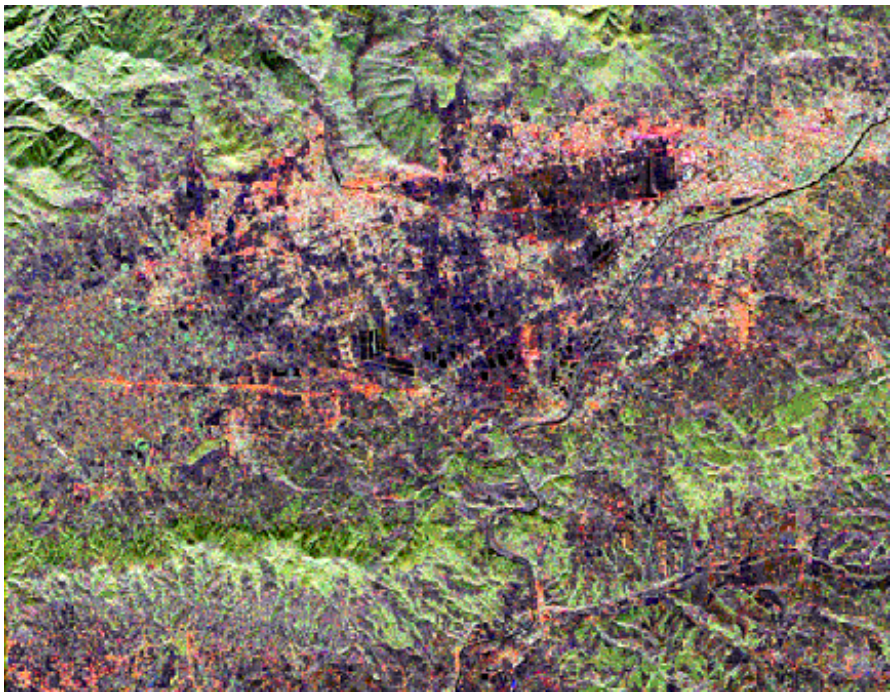
Now, to glance at this city (as some would say, of "Beer Halls"), first look at the color composite made from C- and X-band images acquired by the SIR-C radar system:

This natural color image of the northeast edge of Munich was made by India's IRS-1. Note the widespread distribution of trees and other vegetation.

IKONOS obtained a high resolution color image of the old sections of central Munich (note the red roofs, a trademark of many German towns, composed of baked clay tiles).

About 215 km (135 miles) east-southeast of Vienna (which city is examined in Section 6) is Budapest, capital of Hungary. This Landsat-5 natural color image shows the city, divided by the Danube River, so that Buda lies on the west side and its other section, Pest, is to the east. Below this image is a ground photo of these two segments.

The last stop on the European urban tour is in Italy. We look at the beautiful smaller (about a half million people) city of Firenza or Florence, nestled in the Tuscany hills several hundred kilometers north of Rome in Italy. Florence is known as the cradle of the Renaissance, the home of Michelangelo, and a treasure house of art and culture. Here it is in a radar color composite (L-HH = red; L-HV = green; C-HV = blue (see Section 8 for an explanation of these bands and polarization modes) made by the SIR-C radar from the Space Shuttle on April 14, 1994:



For many American travelers Florence is one of the five top cities in Europe that warrant a "must-see" status. It is for many the place most identified with the blossoming of the visual arts in the late Middle Ages. Here is an overview along the River Arno:

Courtesy: Carolina Map Distributors

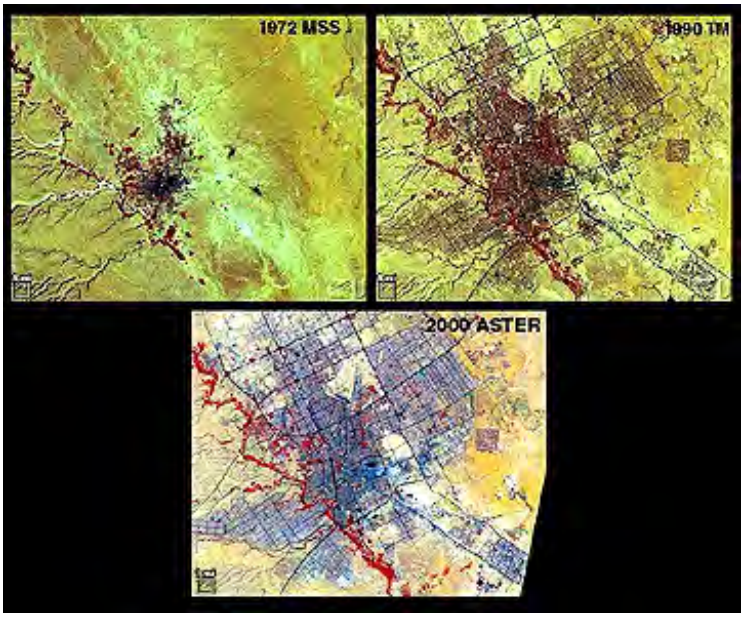
This next view is at an odd orientation but north is towards the top. The Arno passes through the photo as a narrow, winding black band. That same river runs through the ground scene (below) taken from the Piazza Michelangelo on the south side of the city. The famed Ponte Vecchio (a 15th century covered bridge with shops and stalls still in business today) is one

landmark. The tower in the town hall at the Piazza Vecchio lies near the photo center and part of the Duomo (cathedral) with Giotto's Tower is at the right edge.

In the radar composite (above), much of the city stands out in very dark gray to black tones, signifying low radar beam returns. The large, dark, partial "V" in the central city is the railroad station, which has a flat roof.

4-14: **Speculate on a possible cause of the black tones in the radar image. Check the ground photo again. ANSWER**

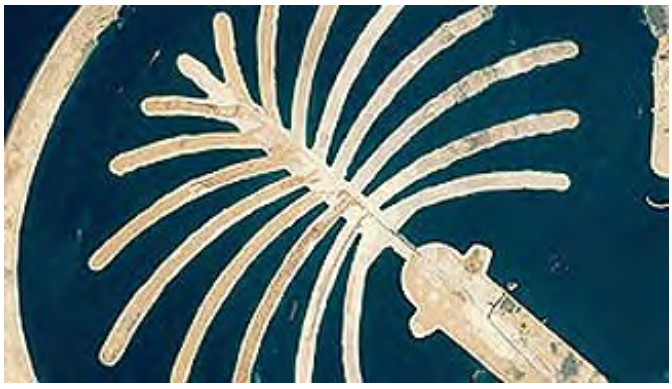
Reluctantly, we must leave the charms of Europe for a look at cities in the Middle East and Asia. The capital of the vast country of Saudi Arabia is Riyahd. In 1970, this city was still relatively small. But the Royal House of the Saudi's embarked on a huge development program thereafter, financed in large part by their income from their huge petroleum reserves that became the principal source of oil for much of the world in the second half of the 20th Century. From the illustration below, one can see in two Landsat scenes the extent of expansion between 1970 and 1990. The 2000 ASTER view of Riyahd shows that enlargement has thereafter slowed considerably.



The second city of Saudi Arabia is Mecca which contains the holiest shrine in Islam, a great plaza with a rectangular structure containing a meteorite that is considered to have fallen into the Muslim world directly from Heaven. Devout Muslims make the Hajj, a pilgrimage to this shrine

required at least once in a lifetime. Here it is as seen in a Quickbird image:

With their vast wealth stemming from huge reserves of oil and gas, countries in the Arabian MegaPeninsula have been modernizing their cities and building new facilities. Among the most striking is this array of arcuate land sites for residential homes built out from Dubai, United Arab Republic, into the Arabian Sea:



Let's close this visual excursion to modern foreign cities by looking at two in China. Here is central Beijing, including the Inner City and Tiananmen Square, as seen by IKONOS:

Try to fit this next picture, an aerial oblique photo showing Tiananmen Square and its surroundings, into the IKONOS image (hint: there are two squares in the image).

GoogleEarth/Quickbird has produced a high resolution of Tiananmen Square, with features easily matched with the above oblique aerial view.

Prudently preserved, the old city of Peking fits within a now much more expansive modern city with many high rise buildings:

In the summer of 2008, Beijing was much in the news, as the central site of the Summer Olympics. The centerpiece of the Olympic complex is the Bird's Nest stadium, seen in this IKONOS view from space and from a ground photo:

One of the largest cities in Asia, and its fastest-growing, is Shanghai, on the east coast of China. Here is part of this now very modern city, as rendered from a Landsat-7 color composite registered with a SPOT 10 meter resolution image (these images were received by at the Chinese Remote Sensing Ground Station, which is operated by the Chinese government under contract with the U.S. and France).

SPOT-5 has provided this 2.5 m resolution image of central Shanghai (not in the above scene). This specially processed image is a contrast stretch that makes it resemble a radar image.

Within this image is an intersection of many roads that cross each other in a multi-tiered "spaghetti bowl". Try to count the different roads involved; this may be a record.

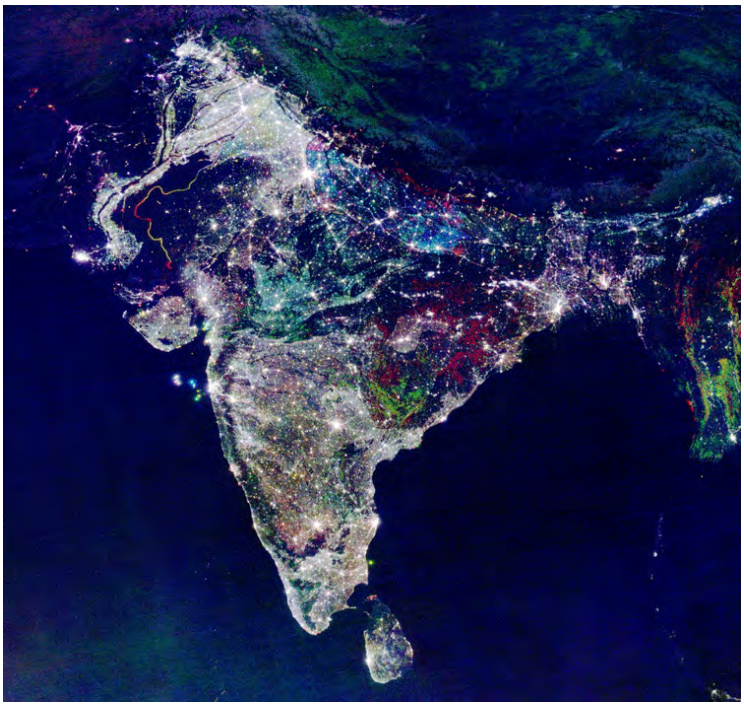
These two ground photos, 1) looking west (from right to left in the Landsat-7 image) and 2) looking south along the river, show the many new skyscrapers that have been built in the last quarter century in Shanghai. The view is from across the Huangpu River, with the Pearl of the Orient Tower shown in the right foreground (present in the Landsat-7 image right of the bend in the land at the extreme right but hard to pick out).

The Pearl Tower is part of the scene in this IKONOS image of Shanghai:

So, now Asia has a structure to rival the Eiffel Tower in France.

Before we leave Asia let us look at a novelty that actually can be helpful in assessing land use. On the first page of this Section we saw a nighttime image of Los Angeles. Night images tend to pinpoint metropolitan areas, and regions undergoing modernization (electrification). Witness this Defense Meteorological Satellite Program (DMSP) portrayal of the subcontinent of India, which

superimposes night lights on a daytime background image; the cities show up in light yellow:



With the advent of high resolution commercial satellites (IKONOS; Quickbird) and a diversity of earth-observing governmental satellites, images of urban areas have proliferated, both for purchase and for access on the Internet. One recent site is the Socioeconomic Data and Application Center (SEDAC) at Columbia University. There, many images of the world's major cities, as imaged by Landsat, are on display. Here is one example: Melbourne, Australia.

The First ERTS Images: Dallas-Fort Worth and California

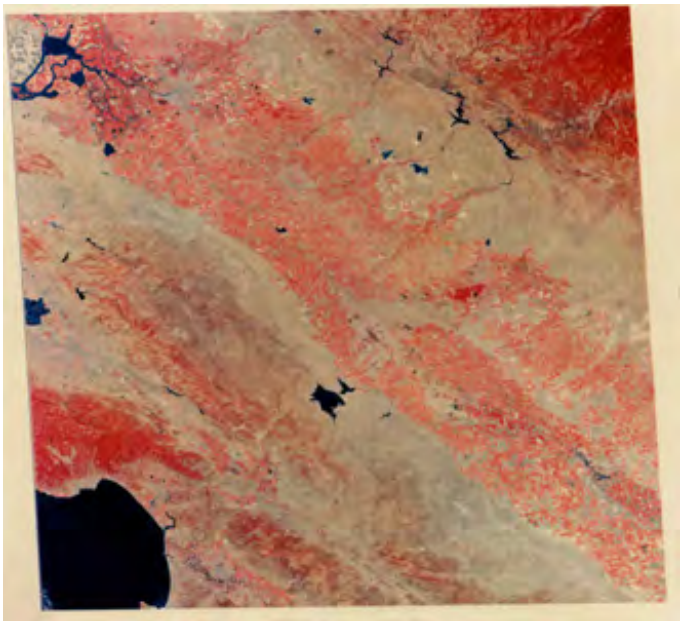
We close this page on an historical note. The first scene processed by Landsat-1 (ERTS-1) was presented to an expectant and excited group of scientists, engineers, and technicians at the Operations Center of Goddard Space Flight Center, two days after its July 23, 1973 launch. The first product was an RBV image made as a low resolution Polaroid of a subscene along an orbital path crossing through Kansas, Oklahoma, and Texas. Only several levels of gray tones were recorded since the image had not been adequately "stretched" to its maximum levels and hence details were lacking. As everyone tried for more than an hour to identify this puzzling area, not one of the experts present could locate this subscene based on its pattern. (Regrettably, this

historical product was not saved by anyone in the room.) When the corresponding MSS was fully processed a day later, it looked like this.

This is Dallas, Texas, near the center and a bit of Fort Worth along the left edge. The Trinity River runs through downtown Dallas towards the lower right. Interstate 20 passes E-W near the bottom of the scene and Interstate 30 follows parallel to it in the left center. I-635 is joined by I-20 on the east but in this scene, it is apparent that the former was not yet complete in '73. Also, the white area in left center is the partially completed Dallas-Forth Worth International Airport.

The prime problem in identifying the location was that the atlas was about 5 years out-of-date. Several reservoirs and lakes clustered around Dallas -Fort Worth, which were built since its publication, were not depicted in the atlas, although several appear in this subscene (elongated black patterns). Thus, no one initially could match the broader Polaroid pattern to this incomplete map pattern with satisfactory certainty. Eventually, some "wise men" deduced that this was the cause of the misfit, and Dallas-Forth Worth officially was recognized as the inaugural ERTS scene. However, in fact, the first full and processed black and white scene that they viewed later that day was a stunning cloudfree image of the Ouachita Mountains in eastern Oklahoma (bottom of page 6-3 and the first color composite covered the Monterey Bay to San Joaquin Valley section of California, shown in the middle of page 2-8.

We reproduce that false-color composite here, for its historical significance and because the first version, seen here, serves as a collector's item because of a production quirk.



Before reading on, study this image to see the obvious anomaly. You may want to look again at the full scene images of New York City to refresh your awareness of image shape.

Remember that this shape (New York City image) is a rhombus with the upper part tilted to the right. This inclination is the consequence of the Earth turning (counter-clockwise or W to E) under the spacecraft as it passes from north to south. Each successive scan sequence shifts slightly to the west. When processed, the image must express this by its vertical boundaries being moved progressively to the left. Somehow, this was not done correctly as the technicians rushed to get the first color composite to the waiting ERTS specialists, so that the upper part is tilted left instead of the usual right.

These ERTS images are themselves of historic significance. Before transitioning to the next topic, we insert here another example of how space imagery can record a geographically historical site. The Quickbird view below is of the cemetery and memorial building at Omaha Beach in Normandy, France, where well over 100000 military landed on June 6, 1944.

On the next page we will look at some truly historical (pre-modern times) imagery.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

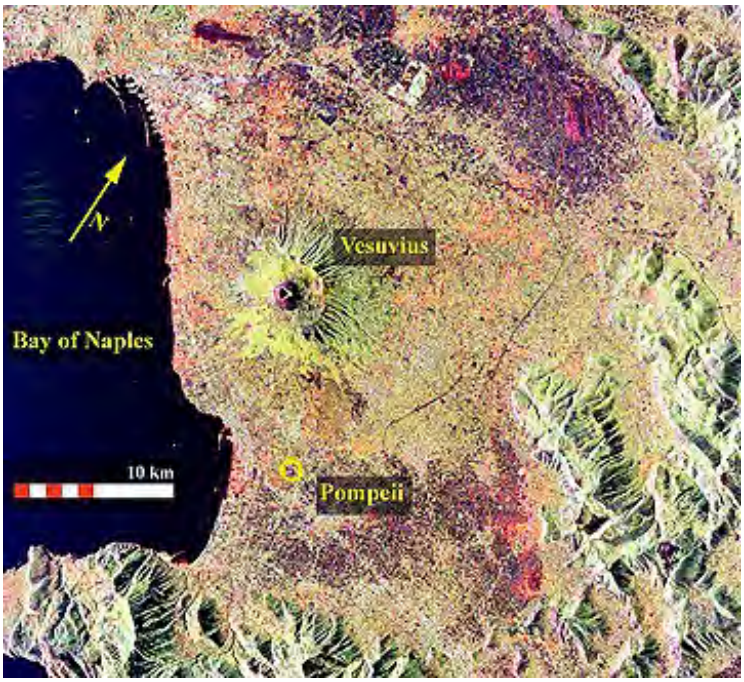
The Section ends with several archaeological wonders as seen from space: an Indian Pueblo in the U.S.; a fortress in Europe; a disinterred city buried by a volcanic disaster; a view of an ancient city's inner walls - a true throwback to the early saga of the Greeks; a discovery using radar to reveal road patterns that led to the finding of an ancient trading city in the mid-East that was since buried under sanda possible site for Noah's Ark; the two most famous ancient sites in South America, and finally a somewhat detailed look at one of Asia's best known archaeological settings.

Throughout this Section we have taken looks from space and the ground at modern cities, although some have old roots. As a finale, let's look at several striking examples of how remote sensing participates in archaeology, both in terms of ancient cities and of regions with no evidence of buildings but signs of human activity. We start with a look at one of the famed cliff cities in the U.S. Southwest, known as Chaco Canyon, in New Mexico. It was occupied by Pueblo and Hopi Indians between 850 and 1250 A.D., then largely abandoned (about the same time the the Anastasi Indians disappeared (experts think that a prolonged drought caused abandoment of many sites and migration to better climes). Here is a view from IKONOS and a ground photo of Pueblo Bonito:

Below is a more modern "relic" - the star-shaped Bourtange Fortress, in the Netherlands, about 42 km (26 miles) southwest of Groningen. Built in the 17th Century, it withstood a number of attacks during the Eighty Years war. It occupies a sandy ridge midst now heavily-farmed "moors" underlain by peat. After this star within a star redoubt fell into disrepair, it has been restored to almost its original state, as a major tourist atrraction. This IKONOS image shows it at 1 meter

resolution, first in its setting within the town and then as an enlargement:

The most famed excavation in Europe is that of Pompeii, a small city near modern Naples, that along with its sister town of Herculaneum, was totally buried by an glowing ash avalanche released during the 79 A.D. eruption of Vesuvius. Here is its location as displayed by a yellow circle on a SIR-C radar image. Beneath is an aerial oblique photo of the present-day ruins.



The highlight of any visit to Pompeii is found in a room in which human body shapes were preserved as casts - made by backfilling the spaces in the ash that survived even as the bodies disappeared through decay. This is what you should see:

(Note to male readers: Be sure you see the room with the scatological graffiti.)

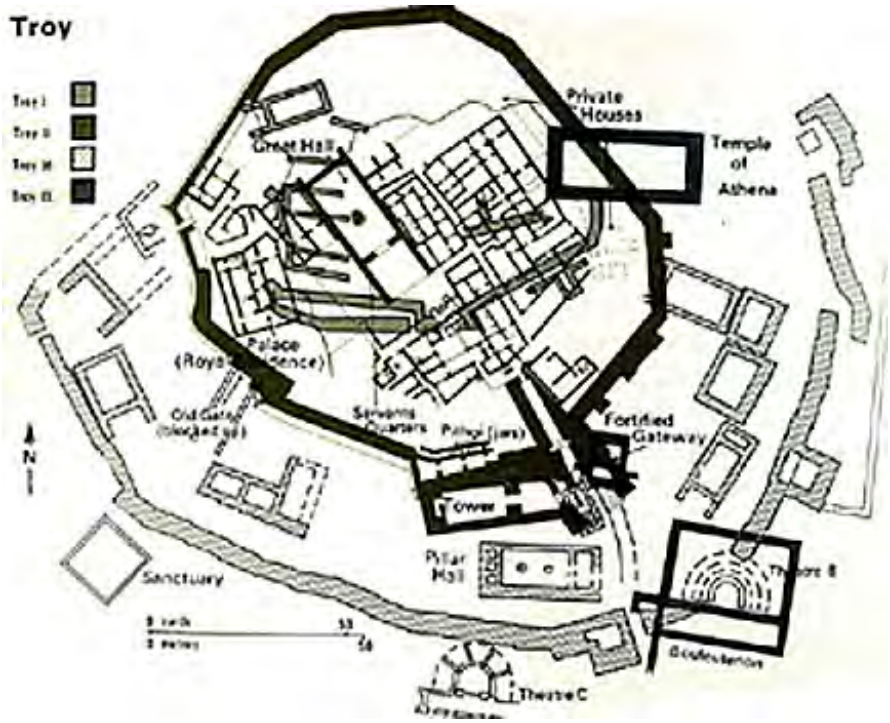
In modern Italy, Vesuvius is close to the large port city of Naples. As evident here in an aerial oblique photo, another Pompeii-like eruption could dangerously affect that city and the towns around it. So, Vesuvius' activity is closely monitored:

This next scene was taken by the IKONOS sensor. The resolution here is 4 meters. Before reading the caption or the next paragraph, take a wild guess as to its identity (clue: it is in modern Turkey not too far from Istanbul, and near the Hellespont - gateway to the Bosporus).

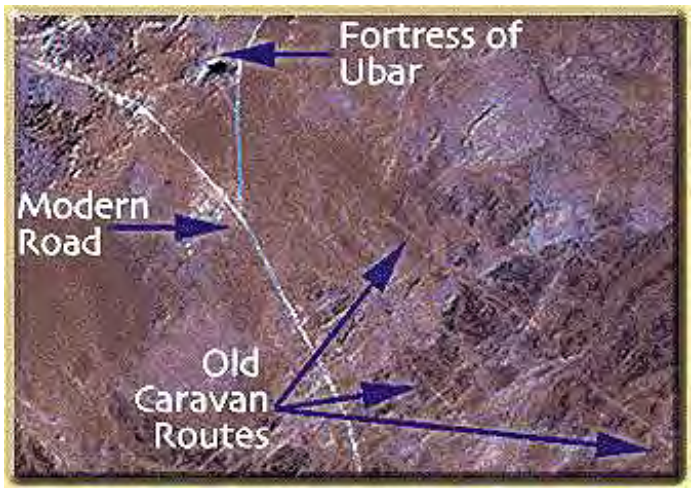
Next to Pompeii, this is probably the most famous rediscovery and excavation in modern times. The city is the fabled Troy written about by Homer. It was settled by the Aecheans of early Greek history and was laid siege around 1200 BCE by the Greeks; it is where the Spartan warrior Achilles defeated the Trojan Prince Hector (who in turn was slain by Paris, the lover of the reknown Helen) and where the myth(?) of the Trojan Horse led to the downfall of King Priam as Troy was overrun when soldiers inside the huge horse debarked and opened the gates to allow the Greek hordes to enter the walled city. Ultimately, it was rebuilt and buried several times, disappearing from sight so that its very existence was questioned. But, the German archaeologist Heinrich Schliemann firmly believed it was real and out there somewhere. Beginning his digging in 1870, he eventually proved to the world that this lost city was an historical fact (but the actual site to the north, near the entry to the Bosporus was unearthed later).

At least 9 levels of urban construction at the Troy site were found during the diggings. These

represent hundreds of years of time, going back perhaps to 2500 B.C. A schematic showing reconstruction of Troy at different times appears below, followed by a map that highlights Troy VI, the purported time of the Trojan War.



Remote sensing has had several notable successes in finding evidence of buried cities. One is the lost "Arabian Nights" city of Ubar, which was an important trading center in the eastern Arabian Peninsula (in today's Oman) for several thousand years until it disappeared perhaps around 300 A.D. Using remote sensing, Dr. R. Blom and others found evidence in this Landsat image of ancient trails that led to a spot on the side of a mountain.



SIR-C radar (see page 8-7), operated from the Space Shuttle, also could pick out these caravan roads (in pinkish-red) that seemed to converge near a dry stream or wadi.

When these archaeologists visited the site area singled out in the imagery, and carried out a "dig", they found evidence (including remnants of walls at the surface) of a large fortress that guarded this city, which apparently was destroyed by invaders after capture.

Eygpt is a land noted for its ancient sites connected with the reigns of numerous Pharoahs who built great temples and pyramids during their dynastys. Satellite imagery is being used effectively to look beneath desert sands to locate buried ruins. The Great Temple of Aten, located on the east bank of the Nile about 300 km (200 miles) south of Cairo close to the modern town of Amarna, was known since 1935. Prof. Sarah Parcak of the University of Alabama-Birmingham uses Quickbird and other satellites to search for signs of hidden antiquities. She has found much more of the Aten temple, which has now been excavated. Here is a vertical view from space and

a perspective view that shows the temple layout:

Satellite imagery is being used in an exotic hunt for a biblical treasure: Noah's Ark. Claims of it being in glacial ice on Mount Ararat in Turkey (location chosen from bibilical interpretation) has been reported for six decades now. This illustration documents the first attempt in 1948 to pinpoint something (initially just called "the anomaly" on the mountain that could be a ship:

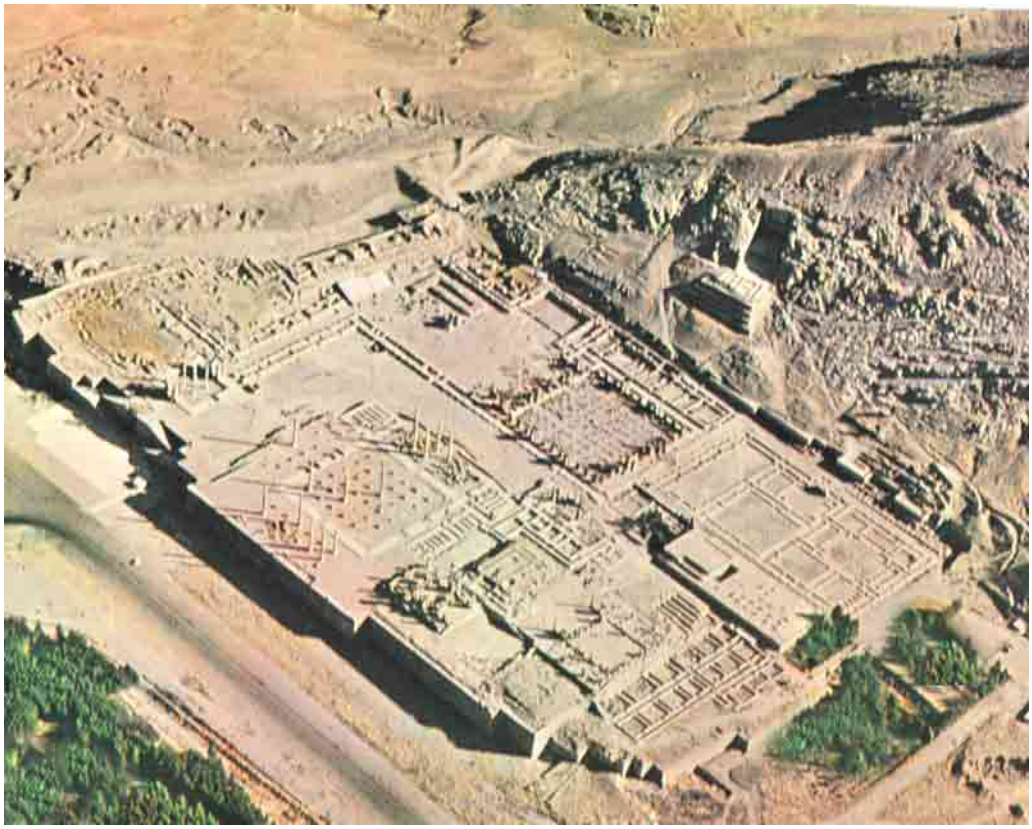
This image, from Digital Globe's Quickbird, shows a provocative dark area in ice which the imagination can conjure up as part of a vessel of some kind caught in the moving ice mass. A group called the Trinity Corporation is planning an expedition to check it out. Stay tuned!

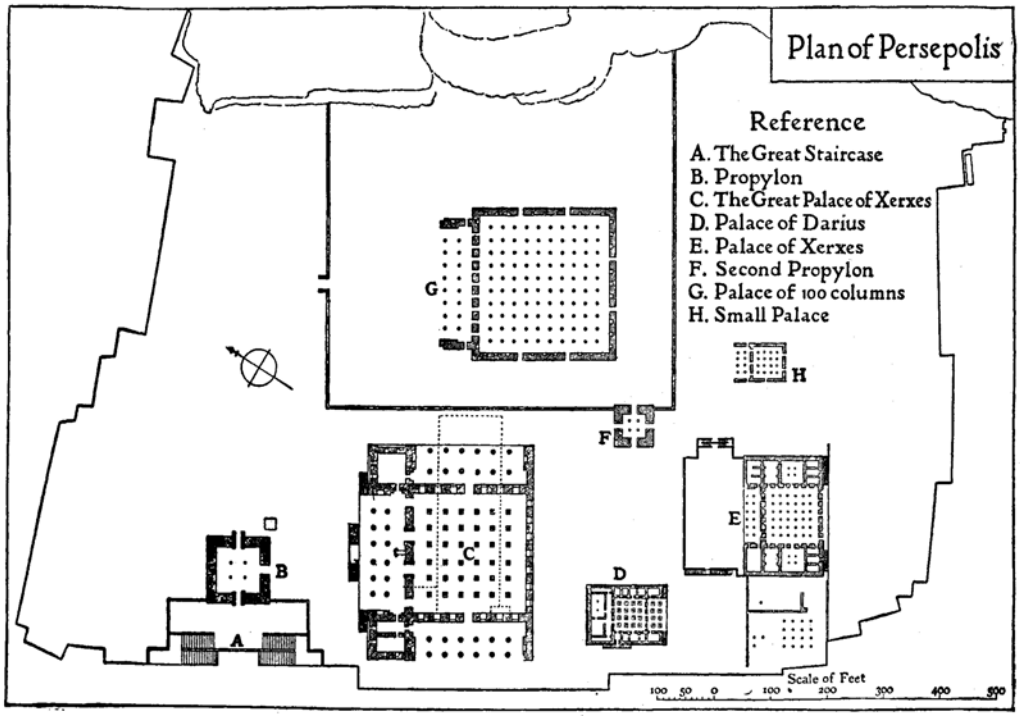
Mount Ararat itself is an imposing volcanic stratocone, whose appearance from the ground is captured in this SRTM-Landsat color composite:

The anomaly is buried in ice. Various satellites have taken additional images. Here are two

Something unusual is there. The site has been visited several times on foot, with (wild??) claims of discovery, but nothing really convincing. The reader is invited to surf the Internet under the headings of "Mt. Ararat" and "Noah's Ark", but be prepared for fanciful speculation.

One of the great ruins in the Middle East is Persepolis, in modern Iran. For nearly two centuries, starting in 515 BCE, it was the capital of the Achaemerid (Persian) Empire. Names associated with it are the Persian emperors Darius and Xerxes and its conqueror Alexander the Great. Below it is imaged by the ALI on EO-1, then from an aerial photo, a ground scene, and a plan map of its layout in its heyday:





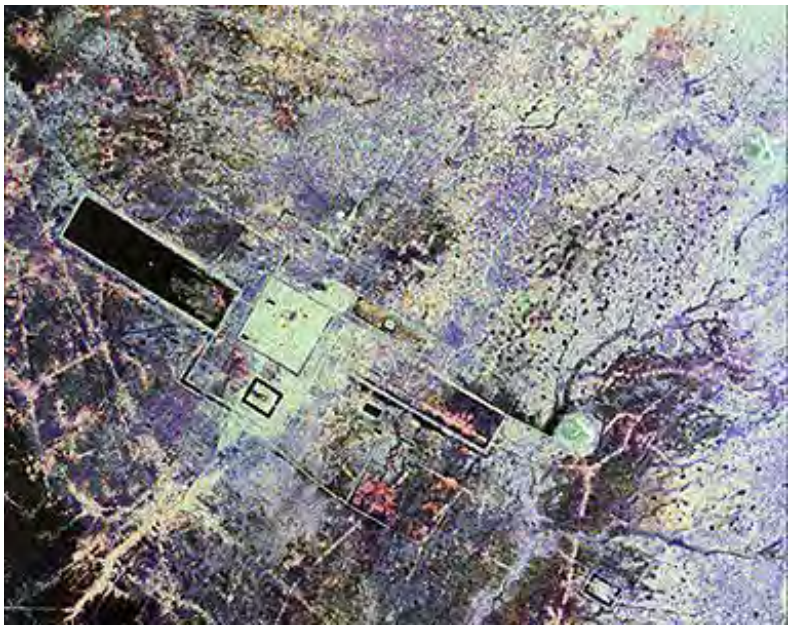
Remote sensing has had a number of other successes in helping archaeologists to search for new ancient sites or extend their knowledge of ones being worked on. On page 15-10 this subject is reviewed in depth, using an example of Celtic-Roman-Medieval ruins in Burgundy, France, analyzed both by SPOT and radar and by the Geographic Information System (GIS) methodology.

In South America, one of the most famous - and mysterious - sites of some, still not adequately explained, ancient activity is the Nasca (or Nazca) lines in the dry low hills and plains in Peru. This Terra ASTER image shows these features - remarkably straight lines as much as 10 km long, some of which cross. Imaginative interpreters postulate these to be landing strips for extraterrestrial vehicles (space ships). Other opinions ascribe them to some still mysterious project conducted by pre-Columbian South American natives. Below the ASTER view is a higher resolution IKONOS image:

Elsewhere in Peru is perhaps the most famed of pre-Columbian archeological sites in South America. This is the ruins of the Inca city, Machu Picctu, built in the 15th Century and abandoned a century later. Situated in the Andes, on a small flat plain near two "sugarloaf" mountains, at an elevation exceeding 2800 meters (9250 ft), multitudes of tourists journey by rail from Lima nearly 400 km to the west-northwest, to witness this fascinating site set against a beautiful background of rugged mountains (a small town nearby at the railhead houses some of these visitors). Quickbird has imaged Machu Picctu at 1 meter resolution, shown below, with an aerial view beneath that gives an overview of the surviving ruins.

In Asia, Cambodia and neighboring modern states were once home to the Khmer civilization which flourished between 800 and 1400 AD. These were Hindu-Buddhist worshipers who developed a distinctive art and were skilled engineers. By 1100 A.D. a complex urban center called Angkor Thom, with up to a million people, had developed in west central Cambodia. As seen from space in this ASTER image, a large manmade lake (the Western Barat Reservoir), a water supply, a walled city (larger square), and a temple (smaller square with moat) appear in this area north of a natural lake, Tonje San.

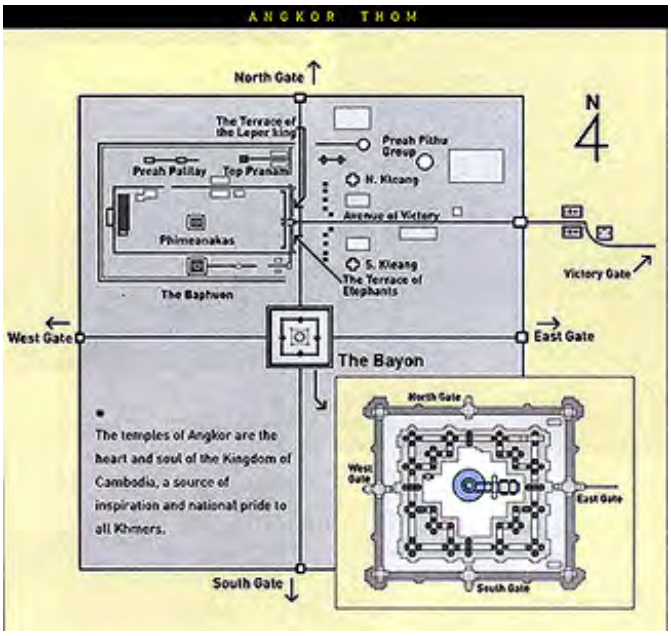
Much detail of the various building in Angkor Thom is shown in the SIR-C color composite



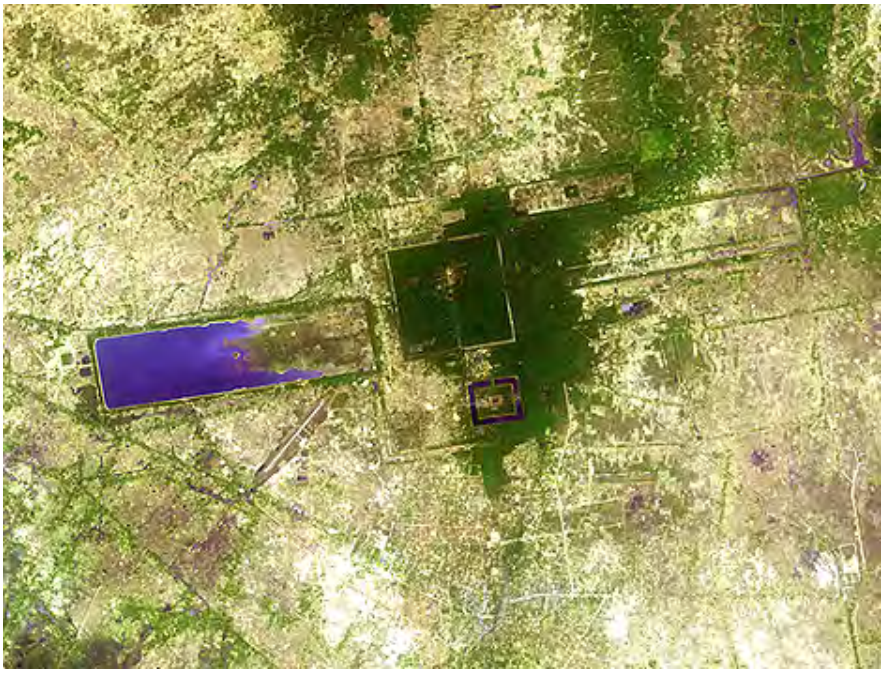
Until the 20th century, this region was heavily forested so that the Angkor ruins were poorly explosed. Much of the land has now been cleared. This map shows the spread of central buildings area:

By far the centerpiece of this complex is the famed Angkor Wat temple, which has been freed from jungle vegetation and restored. It is now the main tourist attraction in Cambodia, reached by road from Phnom Peng, the capital. Here are two views:

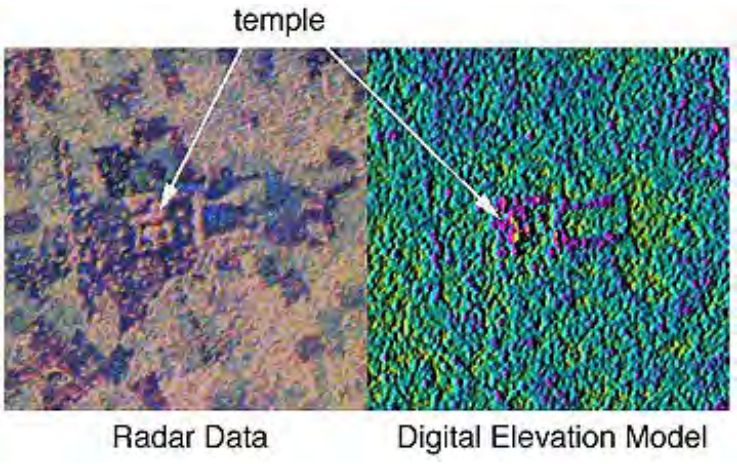
Built around 1140 A.D. by King Suryavarman II, the interior has been studied leading to this functional map:



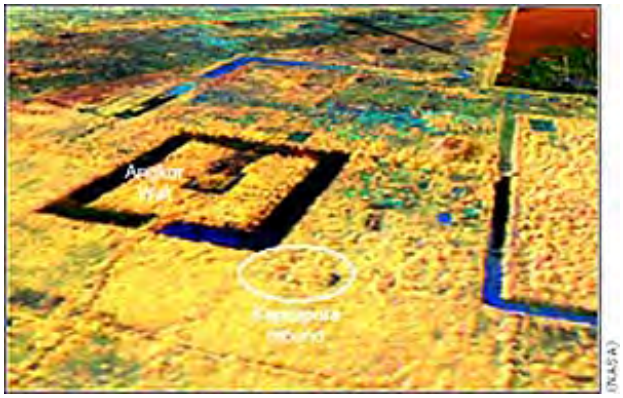
A Landsat-7 subscene shows some of the details of the area, including roads and a modern air strip.



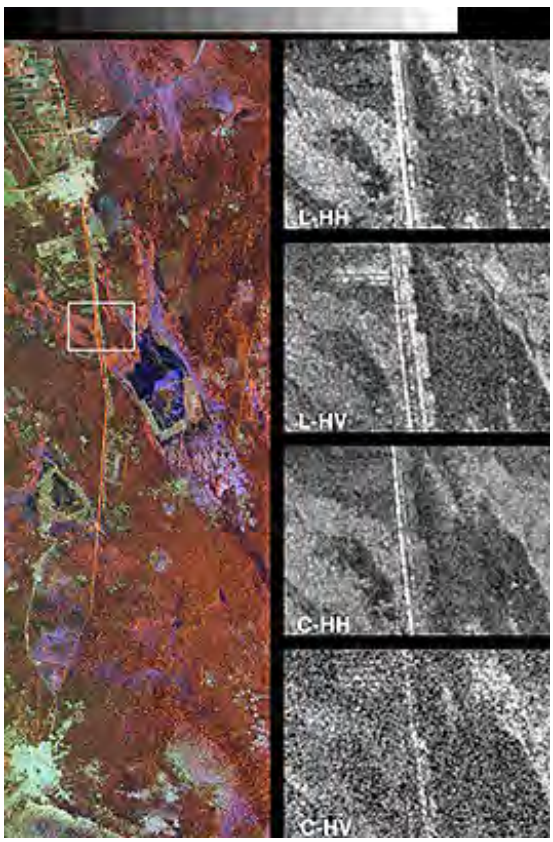
Prior to clearing the temple grounds, an early radar image showed the extent of trees and the ability to penetrate the vegetation to see parts of the Angkor Wat group.



NASA JPL flew a radar mission of Angkor, leading to this perspective image:



In Section 6, on page 6-14a showing Beijing, China, the famed Great Wall built nearly 2000 years ago is visible. But it really stands out in the SIR-C images shown here:



The Wall, constructed for defense against the Mongols and others, is 6700 km (4160 miles) in total length. Large parts of it are located in mountaineous terrain:

Now, let's leave these once gloried ruins to journey again into the vast countryside, this time back to the U.S. by considering in Section 5 locations in southwest Utah and the Plains of Oklahoma, as examples of how remote sensing contributes to mineral and oil/gas exploration.

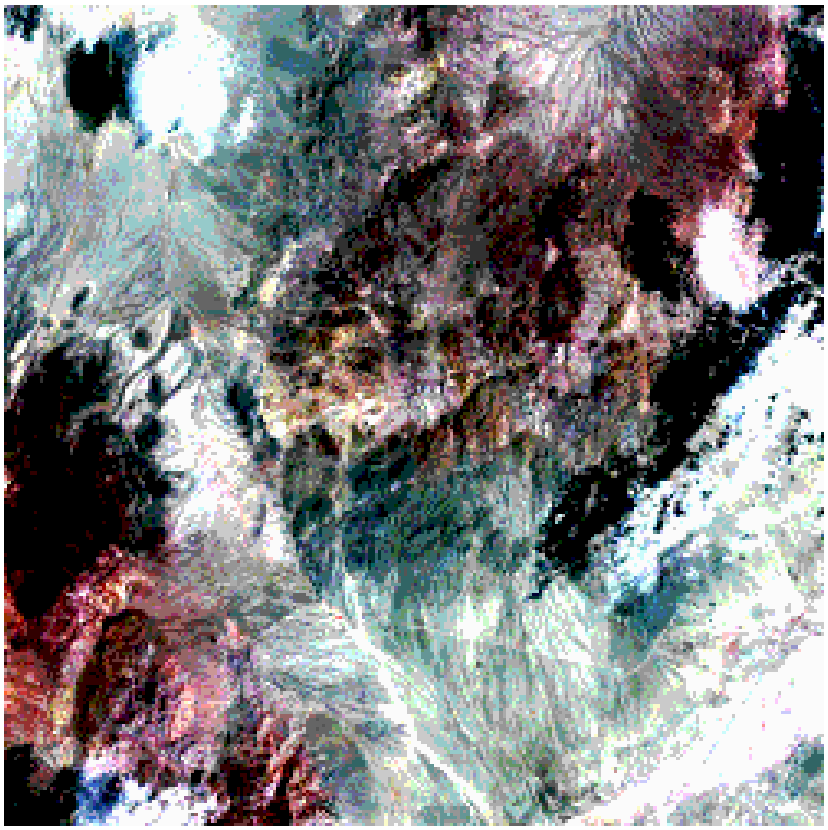
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The test case to exemplify the value of remote sensing in picking out diagnostic alteration is an area in southwest Utah, known as White Mountain, studied by the writer (NMS) during his research days at NASA. Both iron and clay minerals are conspicuous at the site owing to decomposition of minralized volcanic rocks (the ore minerals are of insufficient concentration to be viably minable). The area was overflown by the 24-channel Bendix multispectral scanner and numerous Landsat TM images also cover it. With the aid of a published geologic map and ground photos, the alteration and other lithologies at White Mountain are described.

Geological Setting at White Mountain, Utah

We now investigate a superb illustration of this last situation. Here is a Landsat Thematic Mapper (TM) subscene of an area in southwestern Utah:

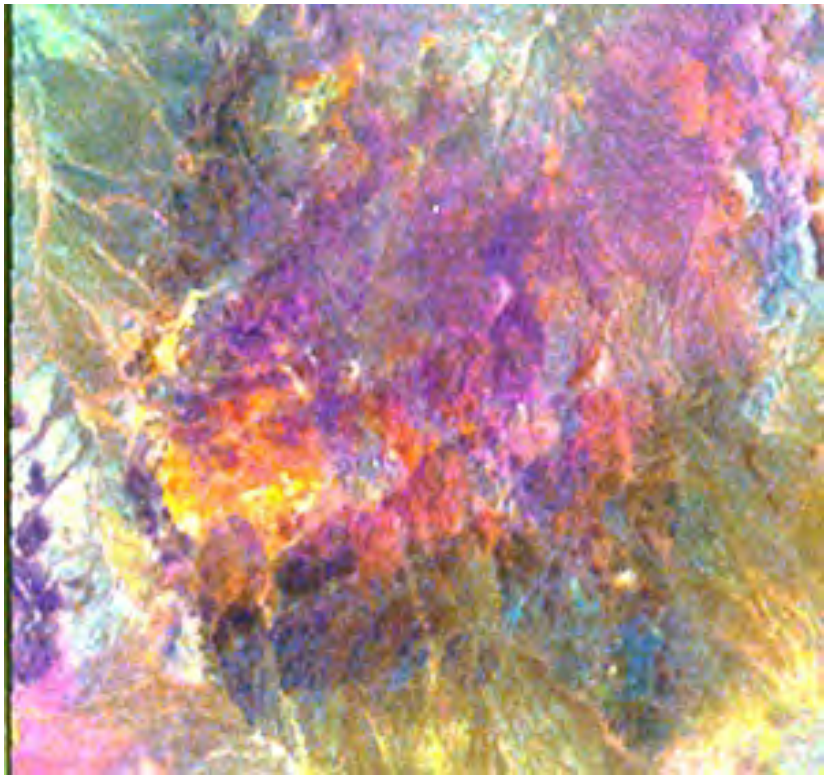


The subscene, about 35 km (22 mi) on a side, resides within Beaver Co., Utah. It is about 24 km (15 mi) west of the small town of Milford and 64 km (40 mi) north of Cedar City, gateway to Zion National Park to its south. The physiographic setting places this area well within the Basin and Range Province, which, geologically, is a series of generally north-south trending mountain blocks that are upfaulted between downdropped valleys or basins. The area lies within the High Desert country of the U.S. Southwest. On the left side of the image is a part of the Wah Wah Range, a typical block-fault mountain system. To the right are a group of isolated hills composed of eroded volcanic flows and small outliers, such as White Mountain (gray roundish landform near the image center), of Paleozoic sedimentary rocks. Lowlands within the scene extend into basins filled with eroded debris from the uplands, including alluvial wash that form sloping aprons. Distinguish this terrain from the clouds and their shadows. Just to the left of White Mountain is an elongated (east-west) area, about 8 km (5 mi) long, characterized in this false color version

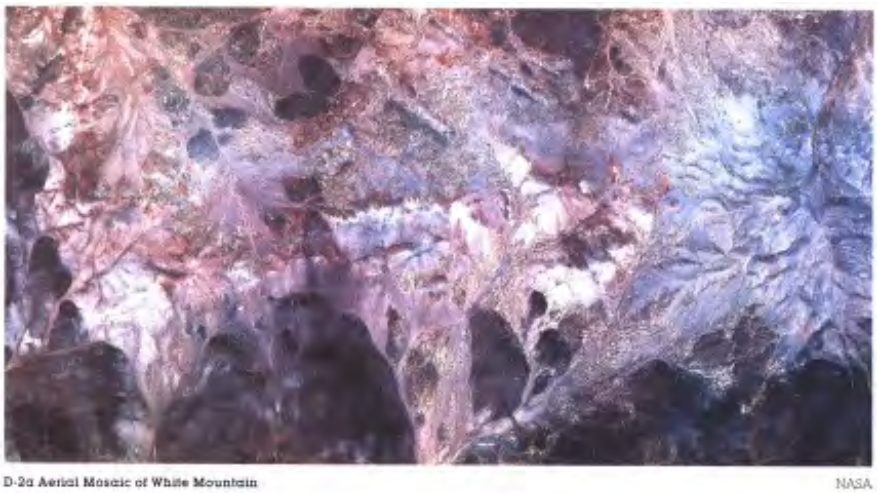
(Bands 2,3,4) by blotchy orange-brown to tan tones. These colors are a classic expression of gossan staining associated with a mineralized zone.

5-3: Make sure you find this mineralized area with its distinctive staining; is there evidence for similar gossan elsewhere in the image? ANSWER

This is an enlargement of the scene above, but produced as a color ratio composite (the altered zone still has a brownish-yellow tone):



Here, we show a part of this area in a natural color aerial photo (somewhat overly red-brown in its printing) obtained during the NASA Aircraft Mission 340 flown for the writer (NMS) in 1976 (see page 2-3 for details).



D-2a Aerial Mosaic of White Mountain NASA

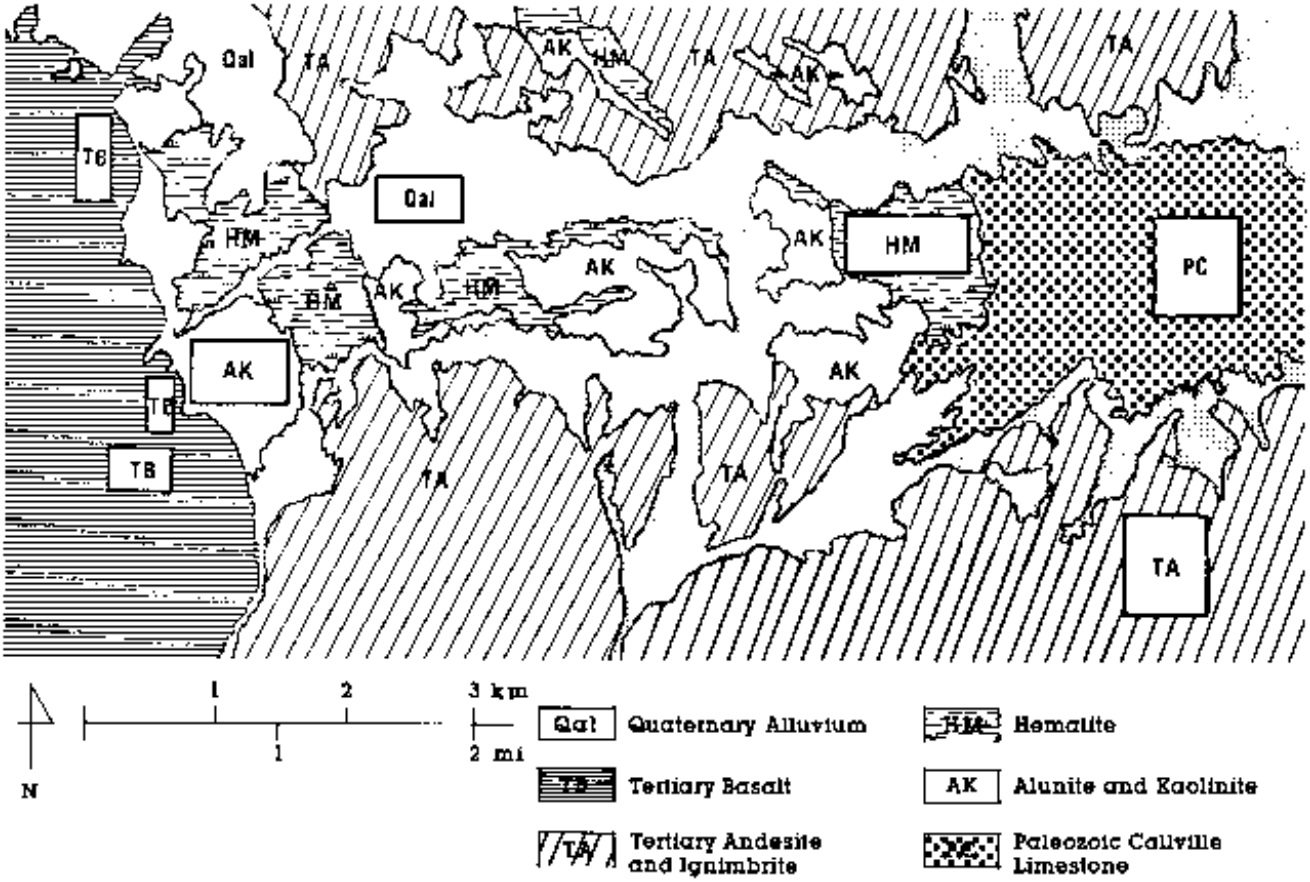
In a ground photo, we see the same area consists of gentle rolling terrain covered sporadically by cedar and sagebrush.

D-1b Ground Scene of White Mountain Alteration Zones NASA

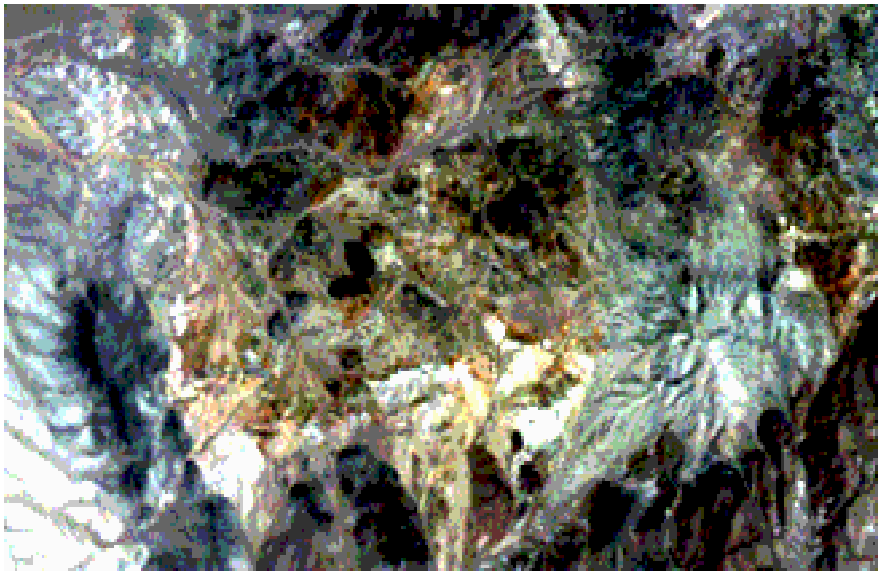
In the satellite image, this vegetation does not stand out as blotches of red, suggesting that cedar has a weak reflectance response in the near-IR.

5-4: How many different colors, that may relate to the alteration, can you discern in this ground photo? ANSWER

A published geologic map (Stringham and Brooke, 1962) of this area indicates a wide variety of alteration types have developed here. Their report states that mineralization consists of pyrite and chalcopyrite (copper-bearing). The latter is not of sufficient grade (amounts) for mining. A group of volcanic rocks called andesites show the most alteration. After they flowed into this region and cooled, volcanic gases and hot solutions modified some of them. One alteration type is kaolinite, a common clay mineral used in ceramics. It is light brownish-white at this location. A second, light-gray mineral is alunite, a hydrous potassium aluminum sulphate. The principal alteration mineral is hematite, the anhydrous iron oxide, which is medium grayish-red. Minor amounts of limonite are present, but the yellow-brown color of the space image corresponds mainly to altered hematite rather than the similar gossan. Very white areas of siliceous sinter are scattered about and a few areas contain natural sulphur, which comes from the breakdown of the original sulphides. In the ground-based photograph above, the light patch coincides with mainly a kaolinite-alunite (k/a) mix, and the reddish area is hematitic staining on weathered volcanics. To appreciate their general distribution in this subscene, examine the map below, a simplified version of the published map, which didn't reproduce well on this Web Page. This map can be printed for ease of reference in studying the images below.



We now consider a set of Landsat images we processed to illustrate how to apply remote sensing data effectively in mineral exploration. Suppose, to start, that we are modern prospectors, searching for some metal (gold is the obvious choice, but copper is often the more likely find). If we had chosen a large region in which to hunt, we would be grateful for Landsat imagery that encompasses many thousands of square miles. Even in a full scene, the anomalous color in the White Mountain area (as this alteration district is known) would grab our attention because it is typical of a gossan signature. Our first logical step would be to zoom in on our image processor to the immediate area showing this signature in a natural color (TM Bands 1-3) rendition.



Dark areas in the subscene relate mainly to the volcanics. On the left side, the map shows that these are basalts (an almost blackish rock, typical of Hawaiian volcanoes and of the Snake River volcanic plains of Idaho). Those dark areas at the top and bottom are somewhat lighter (but still dark grayish-brown) andesites, which are typical of volcanoes in the western U.S., such as Mt. St.

Helens. White Mountain stands out on the right in light bluish-gray tones. It contains a radial pattern that corresponds to furrowed gullies draining down from its central peak. The main alteration zones make up a pattern with an east-west branch that meets a north-south segment on the left. Most of this alteration is displayed in the yellow-browns seen in the regional subscene. Other zones are much lighter (sort of tan) in this image. These two principal zonal types broadly match the hematite and k/a areas on the generalized map shown above. Areas on that map that we identified as alluvium (soil and loose surface debris) show up in several variant colors in the subscene. Those with a color similar to the hematite are alluvial deposits, derived from the altered hematitic zones. An equivalent subscene made from Bands 2,3,4 (not reproduced here) contains very little discrete red patches, confirming the sparsity of active, reflectant vegetation.

5-5: Granted that it is troublesome to scroll up and down on this page, we still think it worth your time to compare the units in the geologic map with their approximate counterparts in the aerial photo at page top and the Landsat subscene near the bottom. You may conclude that the match isn't perfect. Why might this be so? ANSWER

PREVIOUS RST NEXT

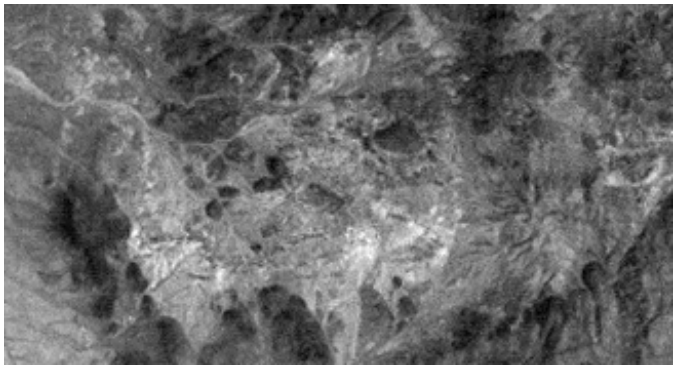
Primary Author: Nicholas M. Short, Sr.

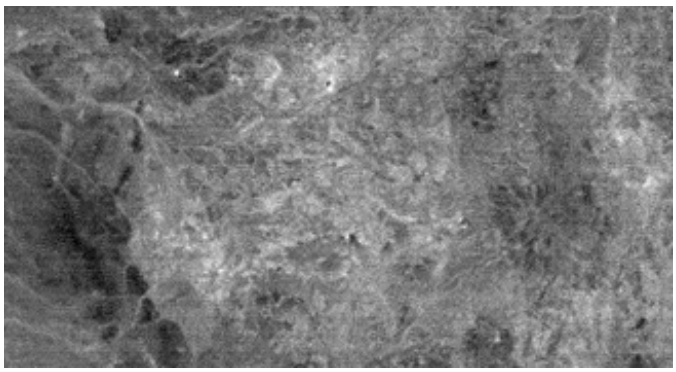
Very believable portrayals of the different alteration types at White Mountain were achieved by ratioing, principal components analysis, and maximum likelihood supervised classification, acting on both Landsat TM and Bendix scanner data. Also, a preview of the power of hyperspectral remote sensing is introduced on this page.

Ratio, PCA, and Maximum Likelihood Analysis of the Utah Site

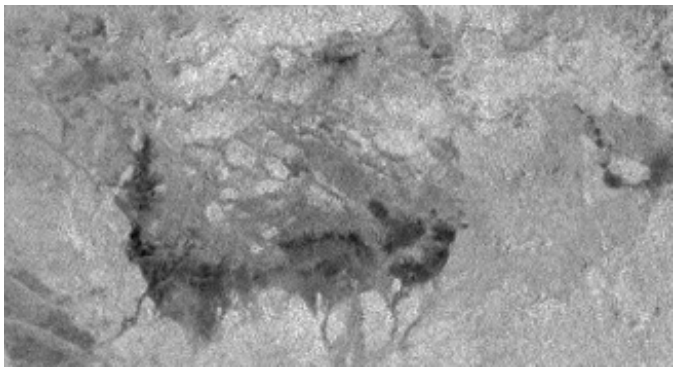
The question now becomes: what special processing will accentuate or enhance the alteration types? Using this summer, 1984, Landsat 4 TM image, we try three types: ratioing, Principal Components Analysis (PCA), and Supervised Classification.

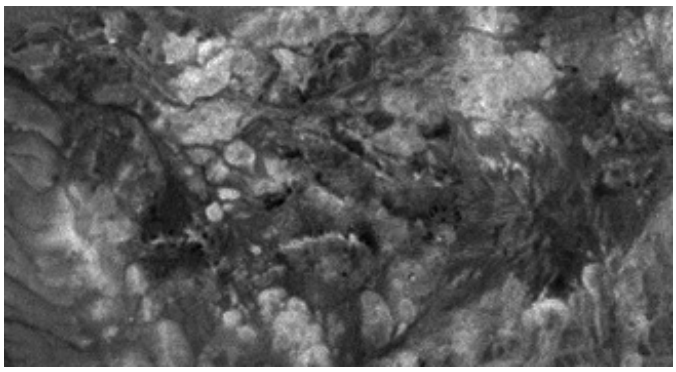
Examine in succession, four ratio images, made with the Idrisi processing system, about which we will briefly comment. The first two are ratios 3/1 and 4/2.





The ratio of TM Band 3 to Band 1 (3/1) renders most of the area in rather dark grays, but several areas are whitish (brighter). These probably correspond to zones of strong hematitic alteration (very reflective in band 3 but dark in band 1). The ratio 4/2 is similar but the bright areas appear displaced. These may mark local areas of denser vegetation. Now, look at 7/5 and 1/7 ratios.

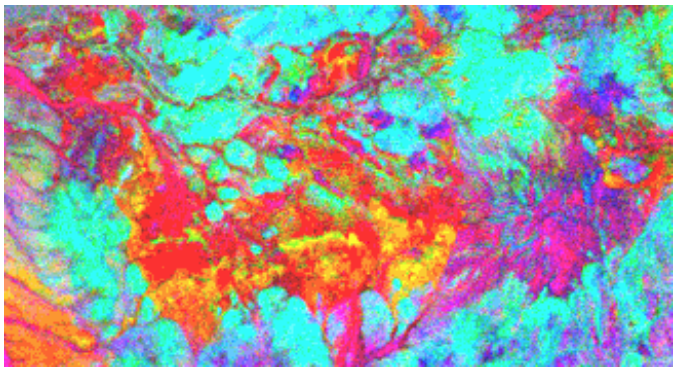




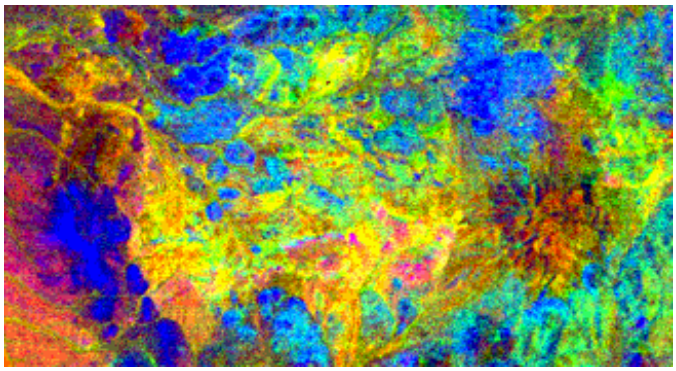
The 7/5 image has a unique pattern, in which a hook-shaped dark area within a scene otherwise light-toned coincides closely to the general altered zone. Band 7 is an excellent detector of hydrous minerals such as the clays, alunite, etc., because these absorb radiation (hence significantly reduce reflectance). Ratio 1/7 shows bright areas that are approximately the same as the basalts and some of the andesites.

5-6: If you had to choose a single ratio image as the one to take with you into the field to check out alteration, which would you pick and why? ANSWER

A ratio color composite made up of Blue = 7/5, Green = 1/7, and Red = 3/1 does not separate the two volcanic rock types (both are blue-green) but shows White Mountain as purple and renders some of the k/a (kaolinite/alunite) areas yellow.

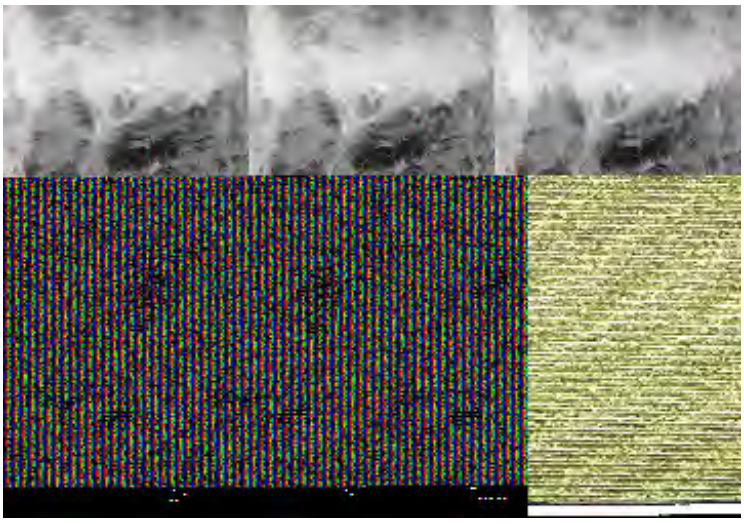


A second ratio composite (second image above), with Blue = 1/7, Green = 4/2, and Red = 3/1, produces a much different result. The deep blue closely expresses the basalt outcrops with the andesites now in a different shade of blue. White Mountain is a distinct orange-brown, but note that the same color appears north of the basalt. The k/a zones are a purplish-red, and quite distinct from a different red for the more iron-rich zones. So, ratio images seem to improve on standard composites in terms of alteration detection.

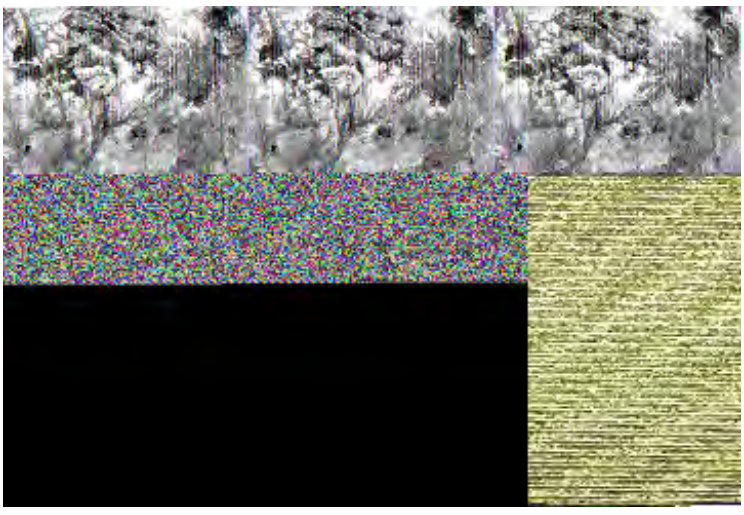


5-7: Which of the two ratio composites would you take in the field? ANSWER

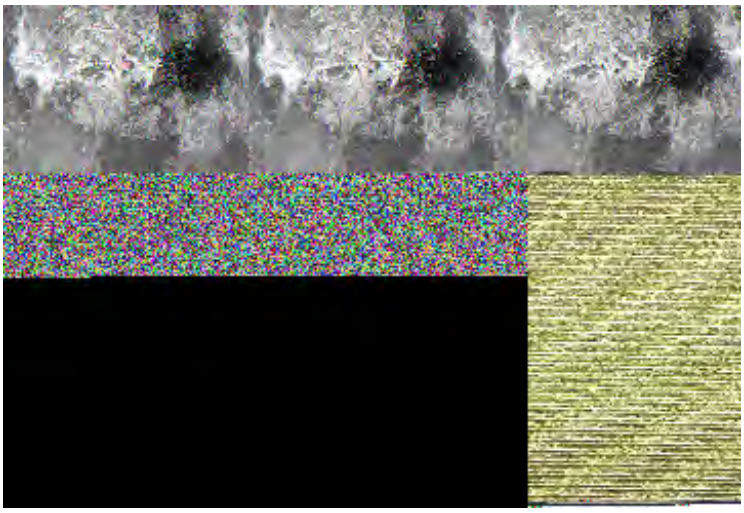
Let's evaluate the utility of Principal Components Analysis (PCA). Here, we use the IDIMS system at Goddard to make these products. The imputs were channels 3, 4, 5, 6, 7, 8, 9, 12, 13 from the Bendix 24 Channel Scanner The first principal component provides, as we've seen before, a view much like a black and white aerial photo. Lighter tones mark k/a alteration areas. There are several very bright, small spots. These are probably pits dug by prospectors (one pit exposing exceptional quality alunite was visited by the writer but was not entered because about 30 rattlesnakes were slithering about - there are limits to one's ambitions as a remote senser! - and it is not being senseless).



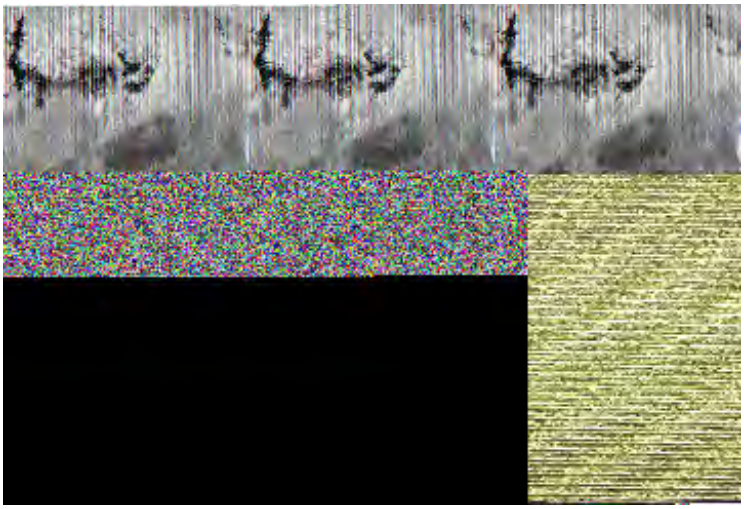
The second principal component is darker overall, with some alteration especially dark. The image again shows the bright spots, some correlative with local alterations, other associated with more vegetation.



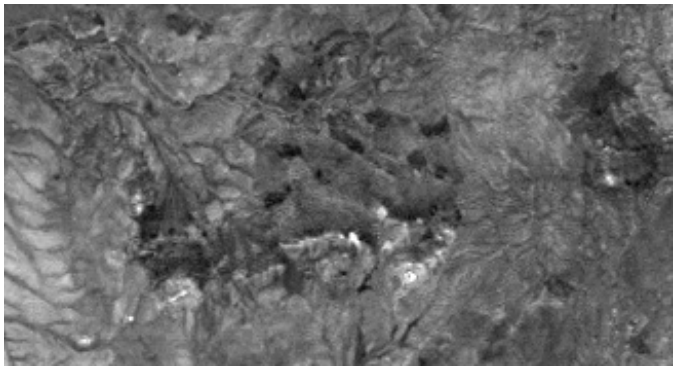
The third component shows White Mountain itself in black. Much of the hematite-rich zones tend towards white.

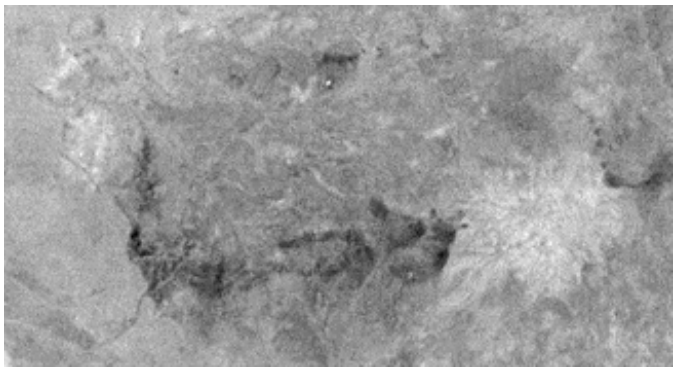


A glance at the fourth principal component shows the same dark hook-like pattern that was observed in the ratio image 7/5. White Mountain completely disappears as a tonal entity.

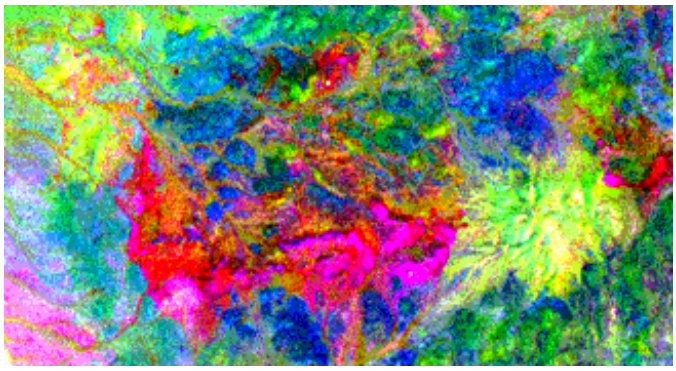


As an experiment, a PC analysis was done on the TM data set using Idrisi. A lesser number of bands were involved. The result was a quite different set of PC images from Components 1, 2, and 3. PC 2 appears here. However, PC 4 from Idrisi was much like the IDIMS PC 4 image; it also distinguishes White Mountain.

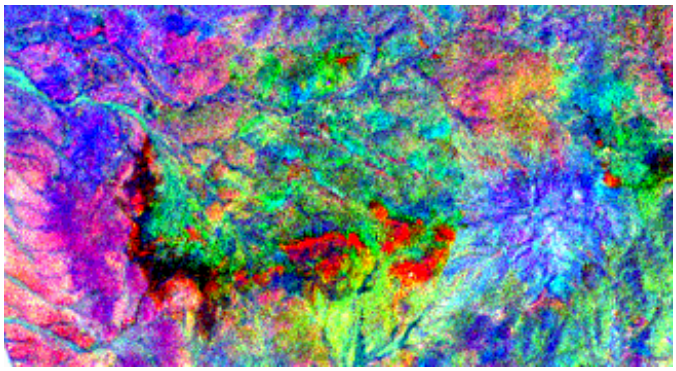




Using the TM data, a principal component color composite, consisting of blue = PC2, green = PC4, and red = PC1 is resplendent with information. The basalt rocks appear blue-green, whereas the andesites tend to be dark blues. White Mountain is a distinct yellow, as is the area above the basalts. This strongly implies that these are outcrops of limestone similar to those at White Mountain, however, no geologic map to prove that supposition is available. The k/a zones appear in wine purple color. The hematitic zones are deep reds and yellows. The areas covered by alluvium tend to be multi-colored, with uncertain boundaries.



A second PCA composite where Blue = PC4, Green = PC5, Red = PC2 is less definitive. Basalts are purple and andesites may be green and/or yellow. White Mountain is not distinct. The k/a zones are bright red, but part of these is bounded by a black pattern, whose nature is puzzling. Nothing like it is evident in the individual PCA images, but the nebulous PC5 may be contributing.

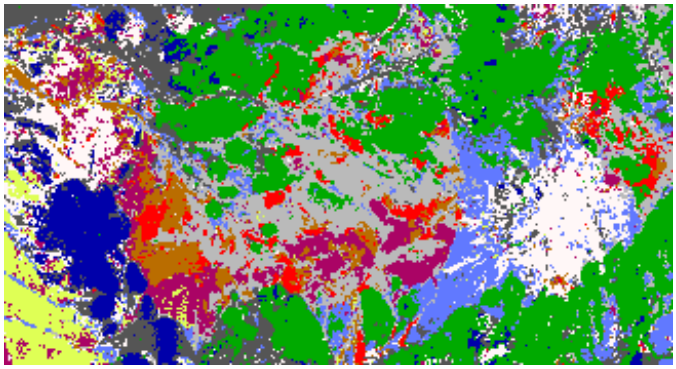


A PCA color composite (shown below) was made from a 24-channel Bendix aircraft flight over White Mountain. For this, they included eight non-thermal channels and combined components 1, 2, and 4 into the composite shown below. Most of the rock units mapped in the field and identified in the Landsat images seem to show up but in some instances occupy somewhat different areas and have dissimilar sizes of outcrop. But, we safely conclude that, using the Bendix image as a standard, the Thematic Mapper (TM) PCA composites match fairly well. At least, they are good enough to stand alone as successful guides to the principal rock and alteration types defined by field work.

Again, as an experiment, a PCA color composite was made from the first three components derived from the TM data, using the same color assignments, to see how the space sensor data compare with the aircraft scanner data. The result below shows a rather good match:

5-8: All the above PCA products each seem to have useful information. Evaluate them beyond the information already offered in the preceding paragraphs. What in particular is separated rather well in the two TM images? ANSWER

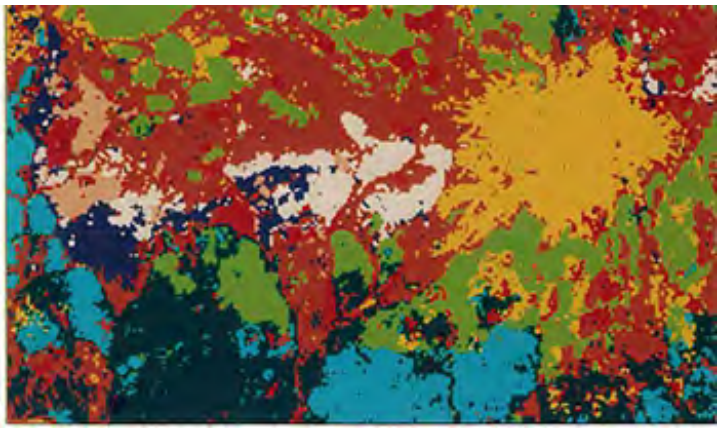
We come now to a highlight of Section 5, a Supervised Classification of the Landsat TM data made by IDRISI with training sites based on the maps and on field observations by the author. Ten classes were established and then identified in training sites used to run a Maximum Likelihood classifier. Here is the end result:



This is a colorful and a **believable** product. The classes designated *Basalt* (dark blue) and **Andesite** (green) are largely where they should be in field terms. White Mountain is well separated but its legend color (whitish) also is found where additional limestone outcrops are postulated north of the basalts. The *Kaolinite/Alunite* zone (purple) coincides well with the map

information. The class designated as Ironrich (brown) is broadly equivalent to the geologic map unit called *Moderately Hematized*, whereas the class called *Hematite* (red) matches at least some of the map units called *Strongly Hematized*. Arbitrarily, four different classes of alluvium were set apart, based on photointerpretative and geologic reasoning. The class *MixAlluv* (gray) is partly within the altered zone and we assume it is a mix of altered rock and volcanic rock debris. *DrkAlluv* (dark gray; found in top left and in from bottom left corner) is a differentiable deposit consisting mostly of weathered volcanic residue. The class *LsAlluv* (light blue), we presume contains considerable contributions from White Mountain and other limestone sources. *BrtAlluv* (yellow) refers to alluvium west of the western basalt hills, which probably received much of its input from the Wah Wah Mountains. Its brightness (in individual bands and color composites) implies a variety of light-colored detritus (fragmented debris) and clays.

We also classified the White Mountain data collected by the Bendix scanner, using 7 of its channels. We consulted the same map to pick training sites, but these sites were not necessarily the same as we selected for the above Supervised Classification. Plus, we employed the IDIMS program for the processing algorithm, once again applying the maximum likelihood classifier. In the image above is the resulting classification and a color code for the selected classes (note that these are not the same as for the first classification). The major differences between the two classifications are: 1) we subdivided the volcanic units in part by degree of vegetative cover, and 2) we treated the unconsolidated cover (alluvium) as a single unit. In general, the correspondence between the two classification images and between this classification and the published map is good in both instances.



5-9: Comparing the two classifications, how does the Bendix classification differ from the IDRISI one? Account for this. ANSWER

This case study with its accompanying images should convince you that satellite remote sensing has practical value. We already insinuated this idea in the Overview and the content of

the first four Sections. Besides, remote sensing can lead to possibly sensational payoffs. We grant that the White Mountain example is almost a sure thing. The major types of alteration are distinctly different, so much so, that the color aerial photo is almost sufficient to produce an accurate map (remember, we said that the images seem superior to the pre-Landsat field map). But, the special processing products make these differences even more obvious.

Imagine that we have a one-time opportunity to stake several claims anywhere in Utah but must do so in 30 days. Its a big state! But, with Landsat and other space imagery, properly processed, we can shrink this huge area to just those small patches that apparently display abundant **gossan**. We can visit the most promising of these patches in brief trips, using rapid reconnaissance to seek signs of minerals. We can take samples for quick assay to determine grade or concentration (amounts of useful metals present per unit volume). Favorable results mean that we ought to file a claim by the deadline. Then, we must drill and map in detail to determine whether any mineralization we have detected is in enough gross volume to warrant developing and mining. With any luck, we'll learn to fully realize the merits of prospecting from space.

Our chances of finding promising signs of mineralization will improve sharply if we use a spectrometer rather than multiband imagery. This improvement comes from many subtle variations in composition, as well as key information that helps to identify individual mineral species, that are present in detailed spectral curves but may be lost in undersampled multiband spectral data. We can now fly spectrometers on air and space platforms, providing hyperspectral (very narrow bands) rather than multispectral (bands consisting of broader wavelength intervals) images and plots.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

The White Mountain demonstration of the ability of multispectral data to discriminate mineralogy and rock types depended exclusively on Landsat and aerial multiband data. On this page we describe the use of Landsat and AVIRIS in the study of alteration in the Goldfield, and briefly the Cuprite, mining districts in Nevada. Goldfield was the first in-depth study of the potential of multispectral imagery for singling out several of the diagnostic alteration minerals associated with the surface manifestation of certain types of mineral deposits lying near the Earth's surface

The Goldfield, Nevada Study, and Other Sites

In the early ERTS (Landsat-1) days, several geologic investigations were aimed at determining whether the MSS (and later, the TM) sensor could produce images (any mode: natural or false color; ratio; PCA; Unsupervised Classification; others) in which tell-tale signs of alteration of minerals in near surface deposits could be detected. Flip back to page 5-1 for the discussion of the principles involved. Most commonly sought were indications of hydrated iron oxides (gossans).

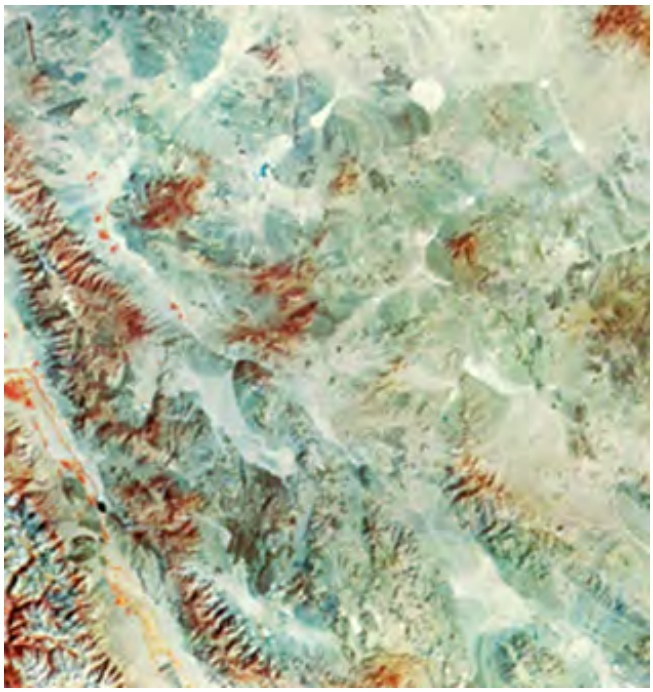
The best known and ballyhooed study was that of Goldfield, an active mining district in south-western Nevada, about 30 km (20 miles) south of Tonapah (a major mining area) on U.S. Highway 95. It was one of many small to medium active or abandoned mines (including the Cuprite district, discussed later) from which gold, silver, and copper have been extracted. The study was directed by Dr. Lawrence C. Rowan of the U.S. Geological Survey, with assistance from Pamela Wetlaufer of the U.S.G.S. and Alexander F.H. Goetz of JPL, and others. **Their strategy was simple: Test the Landsat capabilities for gossan detection by using areas known to have this alteration widespread and well displayed.**

The gossan can be observed easily when driving the highway that passes through the Goldfield site. On either side are natural discolorations in browns, yellows, and infrequently orange in the surface soil above the underlying volcanic rocks. This landscape also displayed prospecting pits exposing even fresher gossan alteration. This is a typical view:

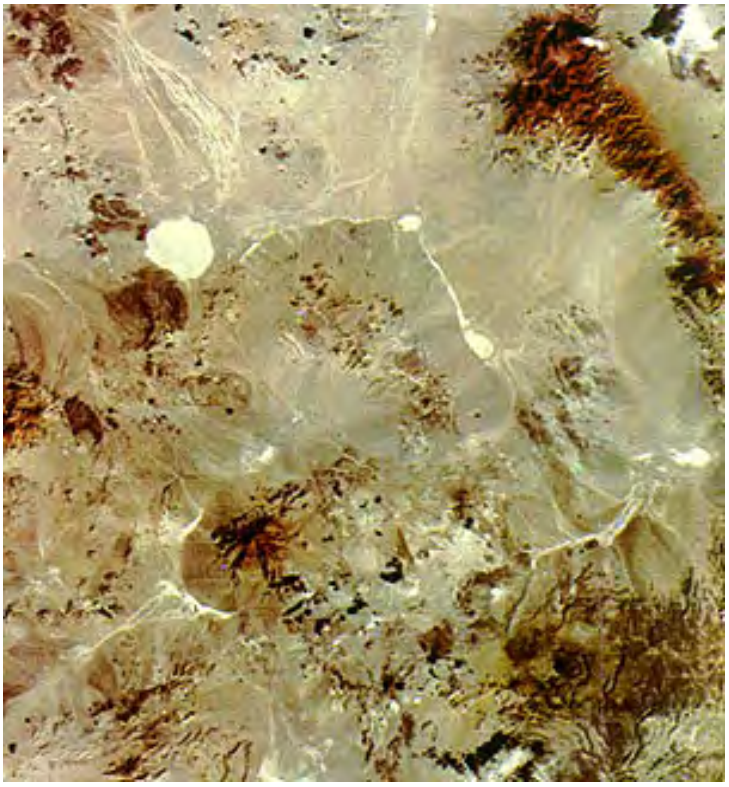
Photo courtesy Phoebe Hauff, Spectral International, Inc.

Here is the original ERTS-1 scene in color which contains the Goldfield alteration anomalies. It is difficult to find Goldfield in this rendition - it is close to the white circular patch in the upper right -

but as you become familiar with it in subsequent views on this page you can return to this image and probably will find Goldfield. As rendered here, it is hard to spot anything anomalous around Goldfield.



Here is a large part of a Landsat-1 image, in quasi-false color, that includes the Goldfield district, to the SSW of the white patch:



The prominent mountain belt near the upper right is the Kawich Range. The near circular white patch is Mud Lake (normally dry - a playa - but may contain a thin water cover after heavy rains). Goldfield lies just below the brown patch that touches Mud Lake at its south end. The Cuprite Mining district is hard to see but occurs west of the peak-like Stonewall Mountain in the lower left of the image.

A perspective view, made by combining ASTER with DEM data shows this area as though seen from an airplane. The alteration zones have been colored in with blues, reds, and greens:

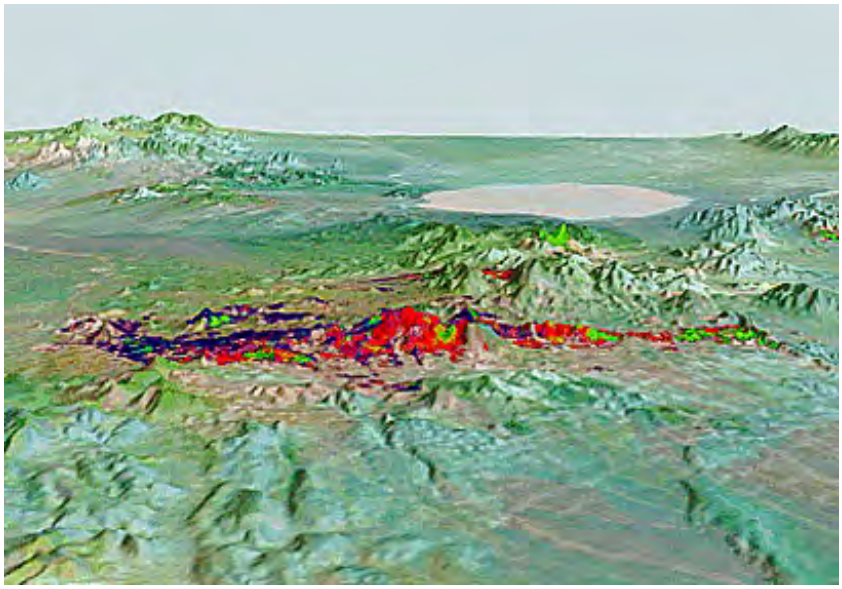
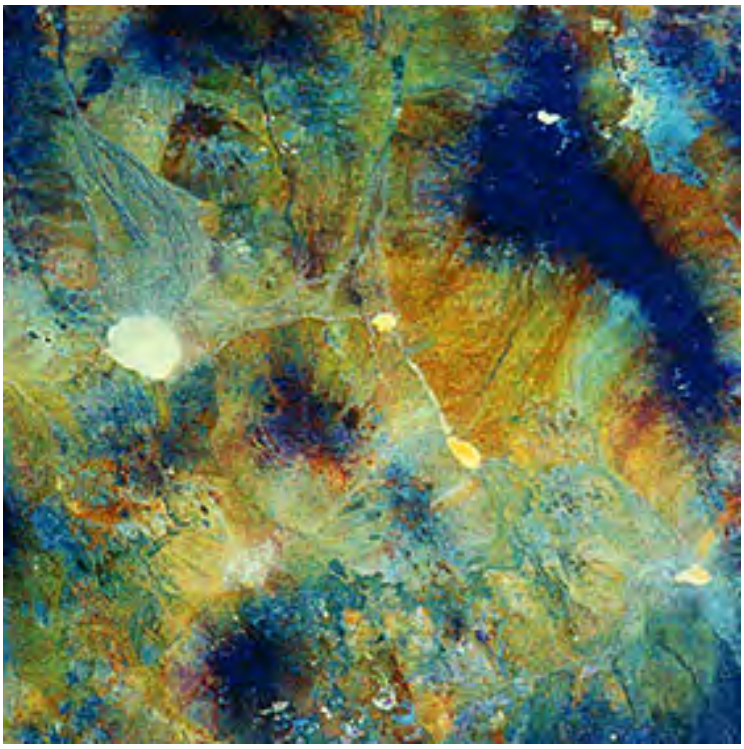


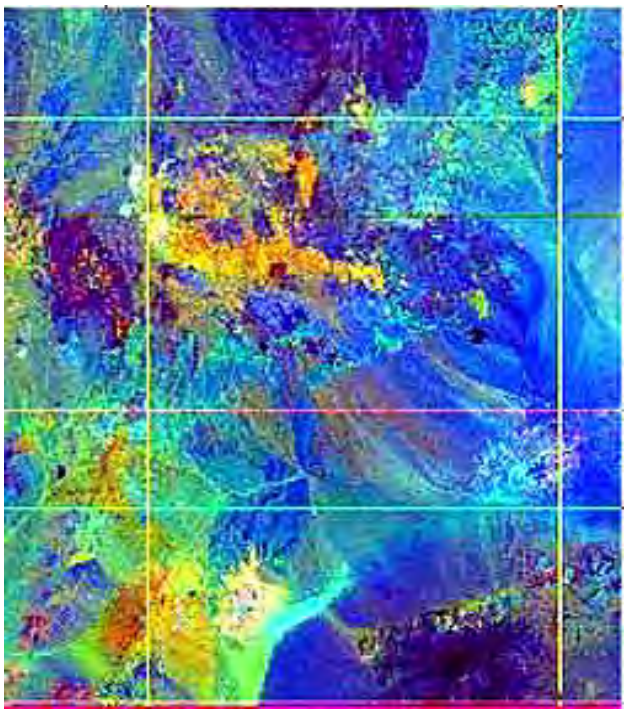
Photo courtesy Spectral International, Inc.

The Investigation Team produced a ratio version of this image, in which MSS Bands 4/5 appear in blue; 5/6 in yellow; and 6/7 in magenta. Vegetation is shown in orange; basalts in gray; silicic extrusives in pinkish-orange; the playa in blue. Goldfield is below A (look for arrows); Cuprite is near O and D; G locates the west end of Pahute Mesa, whick lies in an extension of the Nevada Test Site. The large mountain block near upper right (in orangish-red) is the Kawich Range.

The writer (NMS) tried his hand at bringing out gossan anomalies using ratioing. In this next image, MSS Bands 4/5 = blue; 5/6 = green; and 6/7 = red. The anomaly at Goldfield showed up as a distinct and separable orange-brown (near and below center left). Its shape is an elongate E-W patch, with a tendency to form a hook at its east end, which in different versions is a "trademark" of the Goldfield anomaly. That color occurs elsewhere, particularly in desert slopes to the east, and also around Cuprite.

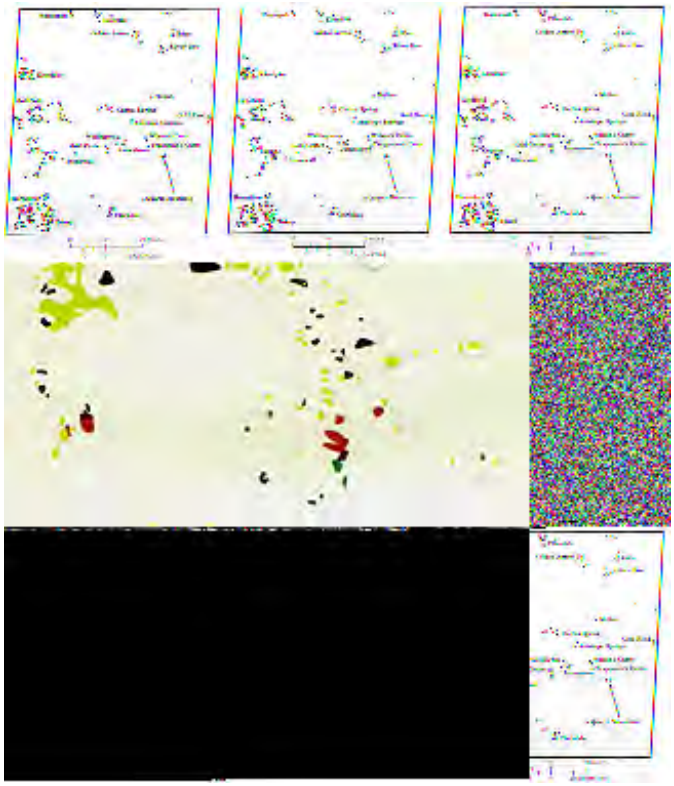


Using Landsat TM data, this hook pattern has been more emphatically defined in this ratio composite (5/7, 3/1; 4/5 as RGB); the Cuprite alteration zone also is identified at the red-brown patch near bottom left:



Larry Rowan's group extracted what they believed to be anomalies relating to several varieties of alteration. Check their map (the form of the Goldfield anomaly [large green patch on map] is sketched in the lower right):

In the map below are X's that represent individual mining areas and prospects in a region corresponding to the Anomalies map just above. Switch back and forth between the two; you will note that most of the X's coincide with the color patches. The correlation appears good enough to be reasonably convincing.



TM data have been processed to extract more information on the alteration present. This is a ratio image, with TM Bands 3/1 = red, indicating iron; 5/4 = green, denoting silica; and 5/7 = blue, picking out clay minerals.

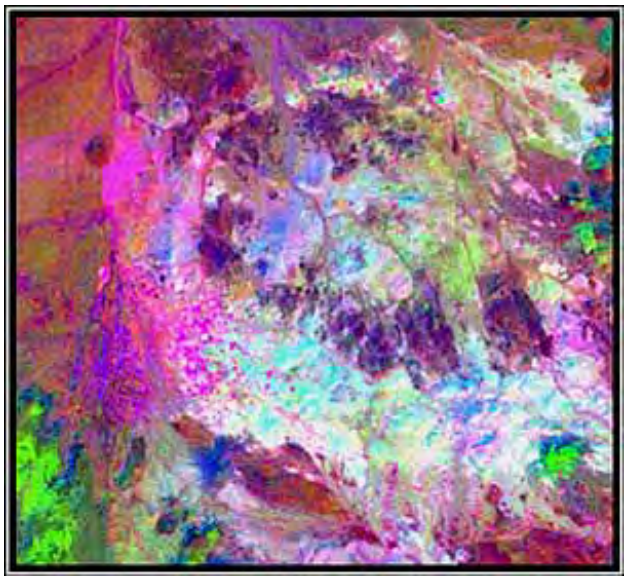


Image courtesy Sandra Perry, Spectral International, Inc.

A Principal Components image (subjected to a decorrelation stretch) of the Goldfield anomaly using data obtained from an AVIRIS mission yields this result:

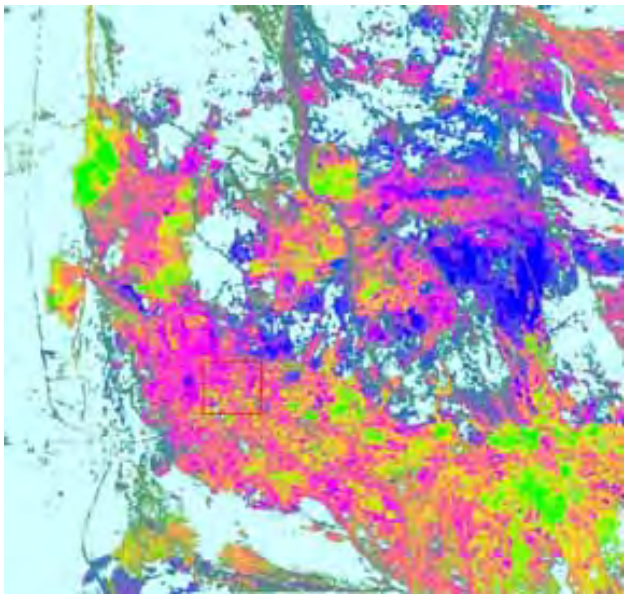
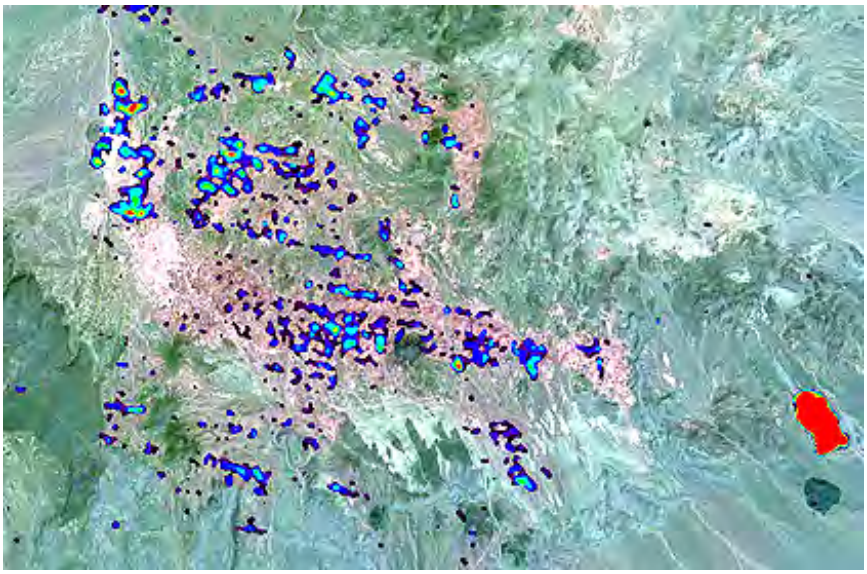
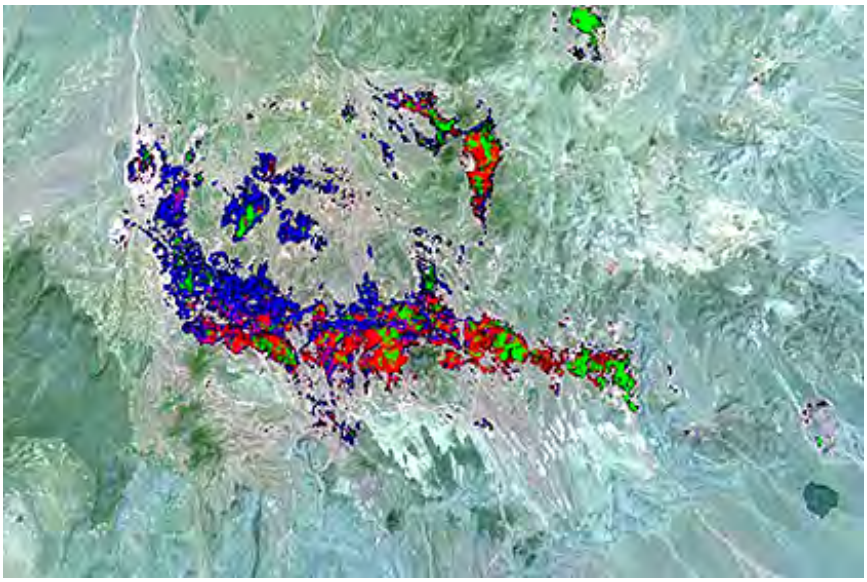


Image courtesy William Peppin, Spectral International, Inc.

Specific minerals are identified using this hyperspectral data set. For clays, pink = Kaolinite; red = Dickite; dark blue = Illite; for other secondary minerals, green = Alunite;

This group has also tested the ASTER sensor on Terra (page 16-10) for its capability in recognizing specific types of chemical alteration. The next two images show a color composite of the Goldfield area with, in the top image, Silicon dioxide (silica) colored blue, and the bottom one, with the clay mineral illite in blue, the clay minerals dickite and kaolinite in green, and alunite and pyrophyllite in red:

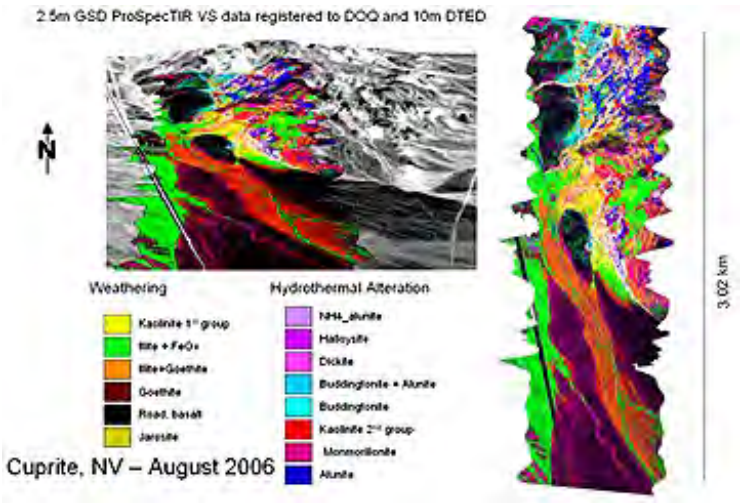




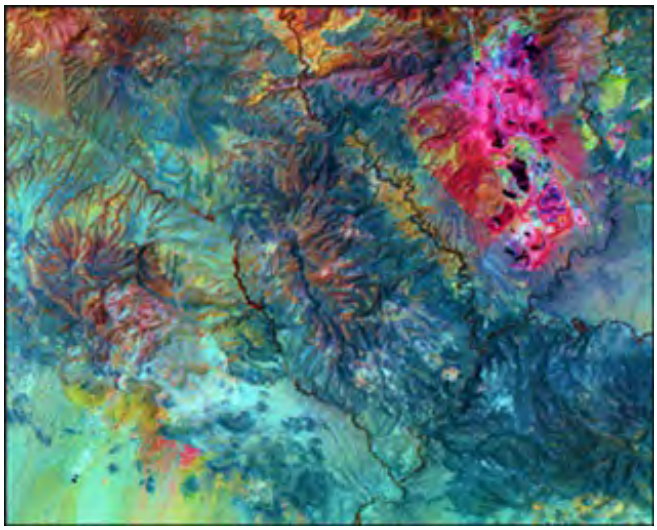
Additional information on the above four images, and on other applications of hyperspectral remote sensing can be found by visiting Spectral International, Inc's Web Site.

Turning to the nearby Cuprite mining district (just south of Goldfield), take a look at a tantalizing look at a powerful remote sensing product - hyperspectral images made with the AVIRIS (Airborne Visible-InfraRed Imaging Spectrometer) now being flown by JPL and the U.S. Geological Survey - which will be covered in detail in Section 13. Shown below (and again, in context, on page 13-10) is an image that delineates iron-bearing minerals and clay minerals around the Cuprite, Nevada mining district. This kind of detailed plot of the distribution of ore-guide mineralogy represents the current state-of-the-art capability of sensors suited for mineral exploration (along with many other uses outside of geology).

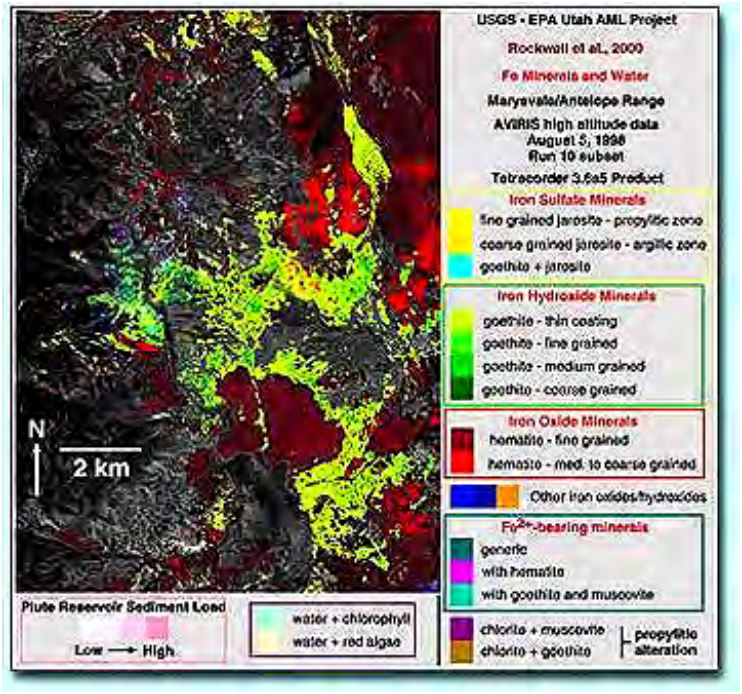
Cuprite is used as a prime test site by private companies that fly hyperspectral missions for mineral and petroleum exploration. Here is an image made by the TeraElement Corp. of this site:



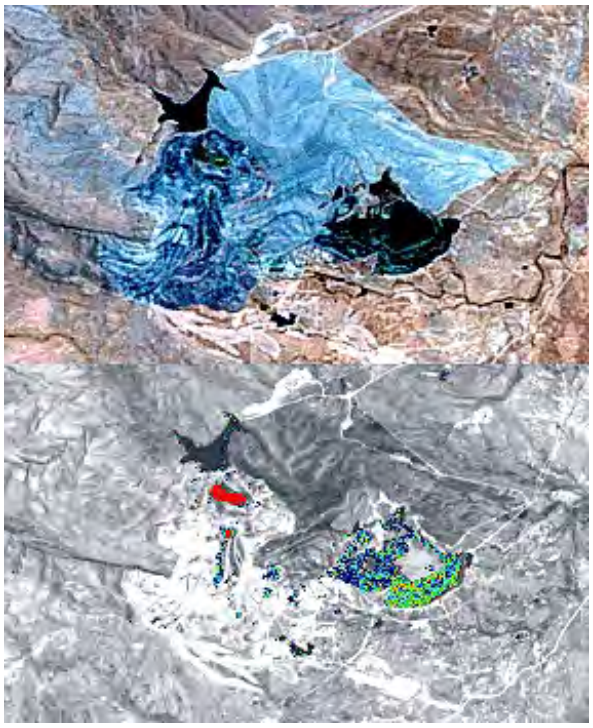
Another example of ASTER imagery's value for discerning alteration is found at the Morenci mining district, in Arizona. This image shows a bright, broad reddish area around the active open pit mine, which is an expression of gossan alteration:



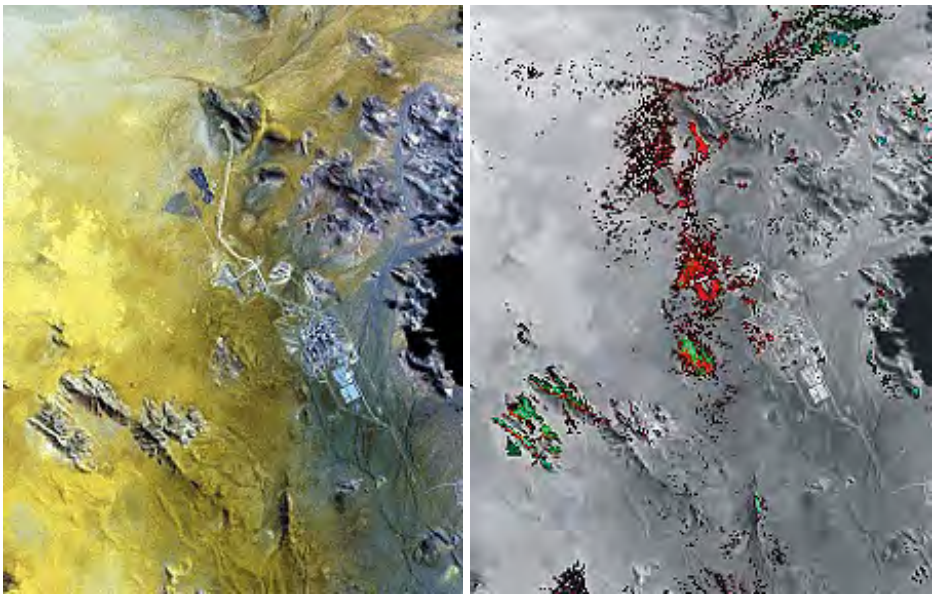
Many minerals in an alteration zone can be identified specifically, as indicated from the legend for this AVIRIS image of the Marysvale mining district in Utah.



Spectral International provides services worldwide for mineral exploration using remote sensing. Keeping with the theme of precious materials, here is an ASTER image of the Kimberley District in Australia, where diamonds are mined from an ultrabasic igneous rock type called Kimberlite. Here at the Argyle mine, Kimberlite veins in the open pit and surroundings are shown in two views. The veins are in green in the upper view; the red in the lower view is a kimberlite pipe (a cylindrical-like intrusion).



Spectral International has also surveyed a zinc metal mining district in Namibia, Africa. Two views based on ASTER data (with an image background) of the Skorpion Mine shows left to right 1) a false color image; and 2) a plot of alteration dominated by variations in illite in reds and greens:

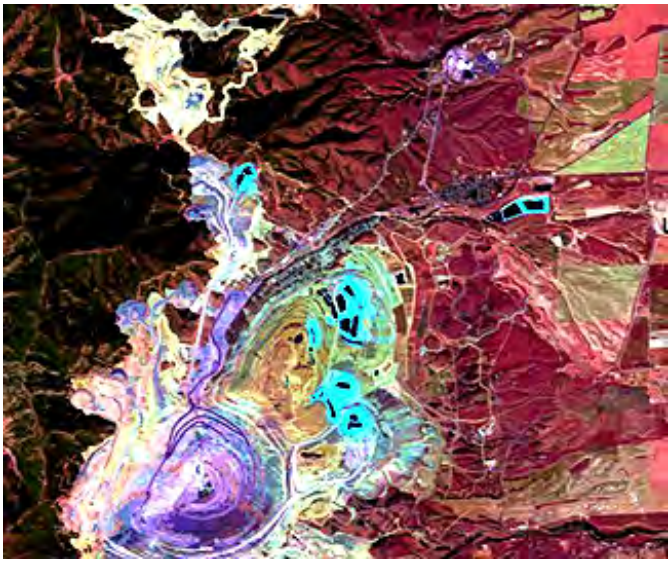


These last images above, of the alteration minerals at severak well-known mining districts, all support the claim that remote sensing - especially that using the hyperspectral approach - is fast becoming a major tool in exploring for mineral deposits of commercial value from air and space platforms. For geologists, this approach is almost "revolutionary".

Remote sensing has also been applied to the hunt for uranium minerals. Some of these are associated with mineral deposits that oxidize or "rust" giving telltale alteration signatures. But prospecting for uranium minerals is usually done with handheld Geiger counters or scintillometers. Many uranium minerals are sought in certain rock types - often shale, but also sandstones and limestones. This next image is a specially enhanced subscene of a Landsat image in which different colors associate with different lithologies. The area shown is the San Rafael Swell, a broad dome with gentle outward dipping strata. It lies just north of the Waterpocket Fold, Utah, that we studied in Section 2. It, and surrounding regions in Utah, were prime prospecting targets of the "uranium boom" of the 1950s.

As a peripheral observation, but again in Utah, we show here a space image of the famed Bingham Copper Mine, which is the world's largest open pit mine, in the Oquirhh Mountains west of Salt Lake City.

There is an alteration map of the natural alteration and that brought about by mining operations at the Bingham Mine which was made during an AVIRIS overpass. Unfortunately, the colors were not identified as to mineralogy in the online source.



Mining waste is a sure sign that some ore deposit of commercial value is being exploited. After the first California gold rush of 1849, placer deposits within the Sierra Nevada eventually "played out". Some wise prospectors reasoned that gold may also have been carried by the rivers into the Great Valley at the Sierra foothills. Techniques for dredging the river sediment soon proved fruitful, as more gold was discovered. This dredging continued well into the 20th Century and is still active on a limited basis. The visual signatures of this dredging are evident from space, as seen in the patterns formed along the Yuba River in an ASTER image:

Metals exploration is clearly benefiting from remote sensing as a powerful technique for prospecting in isolated regions of the Earth. But the really big payoff could come from successful discoveries of new sources of energy - oil and gas primarily, but coal and uranium as well. That is next in this Section.

PREVIOUS RST NEXT

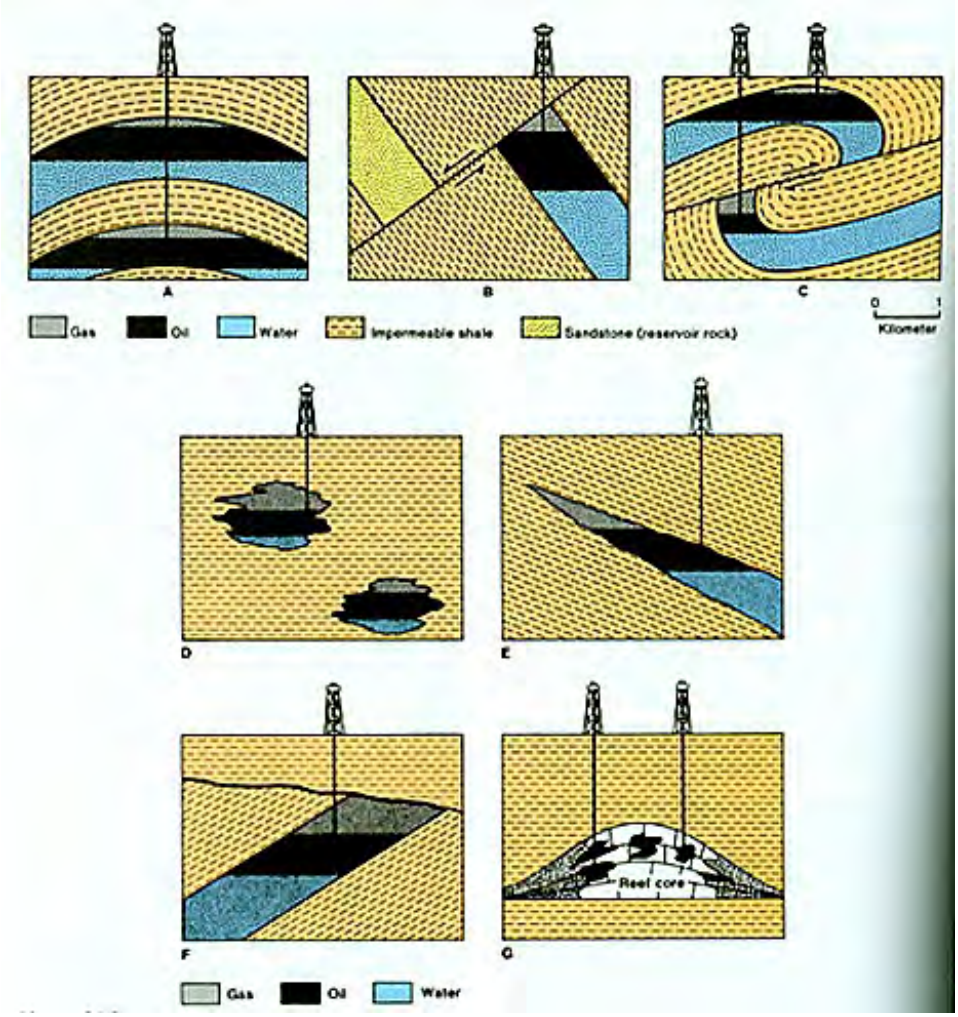
Primary Author: Nicholas M. Short, Sr.

Landsat and other remote sensor systems (SPOT; JERS; radar, etc.) have been heavily used in searching for surface indicators of "leaking" subsurface oil and gas. The general approach to petroleum exploration is described. One line of investigation looks at structural analysis of space imagery in search of subsurface traps. Another, of infrequent success, seeks alteration at the surface caused by chemical changes related to surface-reaching oil or gas. In the early days of Landsat-1, a study in the Anadarko Basin of Oklahoma sought to demonstrate how alteration anomalies and lineaments analysis can aid in finding new petroleum by showing a relationship to already known fields. This pioneering program led to ambiguous but interesting results. Other examples are also considered on this page. The importance of Canadian oil sands as a major source of petroleum (oil) in the future receives special attention.

Finding Oil and Gas from Space

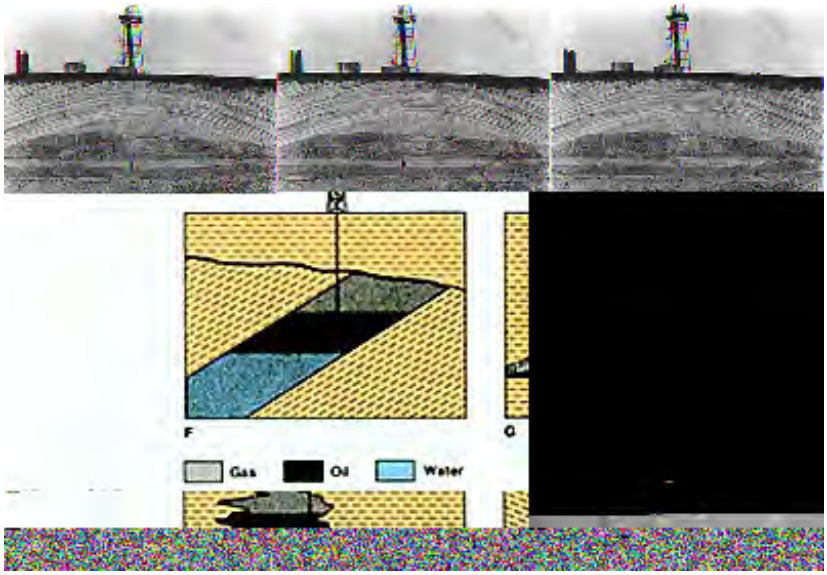
If precious metals are not your forte, then try the petroleum industry. Exploration for oil and gas has always depended on surface maps of rock types and structures that point directly to, or at least hint at, subsurface conditions favorable to accumulating oil and gas. Thus, looking at surfaces from satellites is a practical, cost-effective way to produce appropriate maps. But verifying the presence of hydrocarbons below surface requires two essential steps: 1) doing geophysical surveys; and 2) drilling into the subsurface to actually detect and extract oil or gas or both. This Tutorial website sponsored by the Society of Exploration Geophysicists is a simplified summary of the basics of hydrocarbon exploration.

Oil and gas result from the decay of organisms - mostly marine plants (especially microscopic algae and similar free-floating vegetation) and small animals such as fish - that are buried in muds that convert to shale. Heating through burial and pressure from the overlying later sediments help in the process. (Coal forms from decay of buried plants that occur mainly in swamps and lagoons which are eventually buried by younger sediments.). The decaying liquids and gases from petroleum source beds, dominantly shales after muds convert to hard rock, migrate from their sources to become trapped in a variety of structural or stratigraphic conditions shown in this illustration:



From Physical Geology: Earth Revealed by McGeary and Plummer, First Ed., W.C. Brown Publ.

The anticlinal trap, among the most common, is nicely revealed in a real world setting in this old photograph:



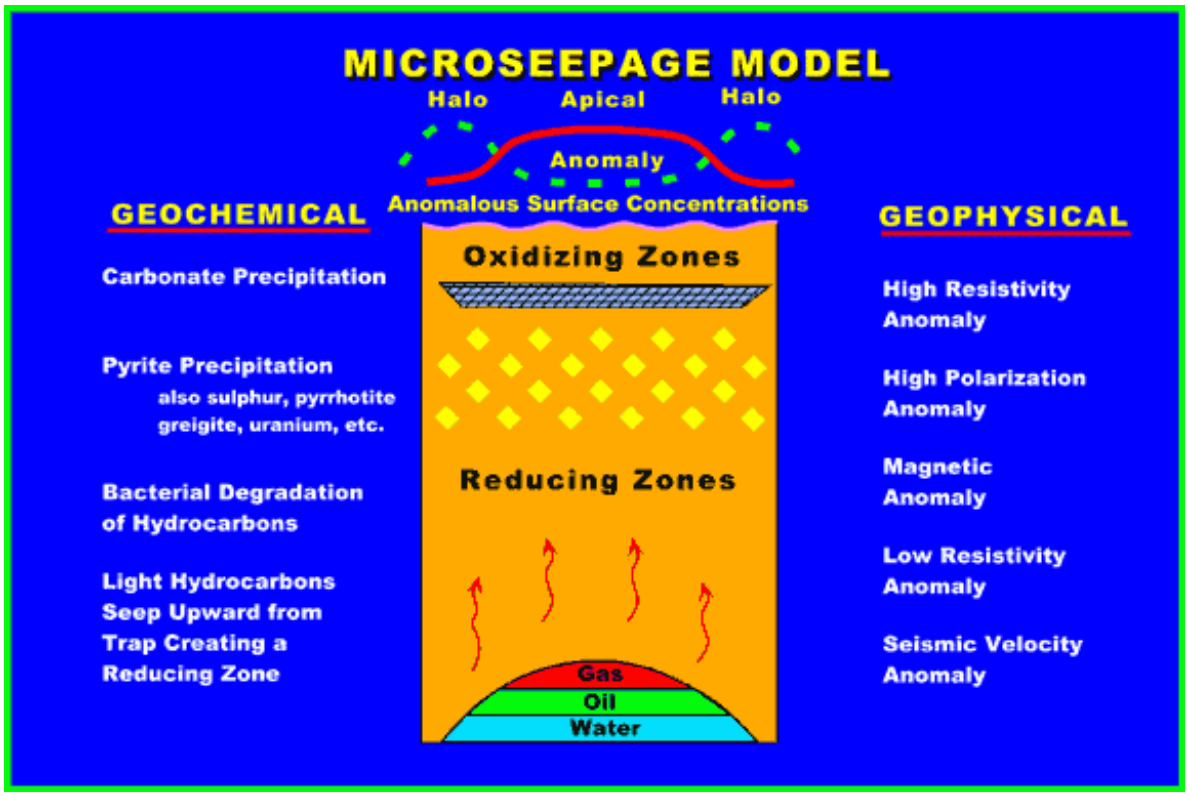
The oil and gas must migrate from deeper source beds into suitable reservoir rocks. These are usually porous sandstones, but limestones with solution cavities and even fractured igneous or metamorphic rocks can contain openings into which the petroleum products accumulate. An essential condition: the reservoir rocks must be surrounded (at least above) by impermeable (refers to minimal ability to allow flow through any openings - pores or fractures) rock, most commonly shales. The oil and gas, generally confined under some pressure, will escape to the surface - either naturally when the trap is intersected by downward moving erosional surfaces or by being penetrated by a drill. If pressure is high the oil and/or gas moves of its own accord to the

surface but if pressure is initially low or drops over time, pumping is required.

Exploration for new petroleum sources begins with a search for surface manifestations of suitable traps (but many times these are hidden by burial and other factors govern the decision to explore). Mapping of surface conditions begins with reconnaissance, and if that indicates the presence of hydrocarbons, then detailed mapping begins. Originally, both of these maps required field work. Often, the mapping job became easier by using aerial photos.

After the mapping, much of the more intensive exploration depends on geophysical methods (principally, seismic) that can give 3-D constructions of subsurface structural and stratigraphic traps for the hydrocarbons. Then, the potential traps are sampled by exploratory drilling and their properties measured.

Remote sensing from satellites or aircraft strives to find one or more indicators of surface anomalies. This diagram sets the framework for the approach used; this is the so-called microseepage model, which leads to specific geochemical anomalies:



The surface geochemical expression of petroleum seepage can take many forms: (1) anomalous hydrocarbon concentrations in sediment, soil, water, and even atmosphere (2) microbiological anomalies and the formation of "paraffin dirt" (3) anomalous non-hydrocarbon gases such as helium and radon (4) mineralogical changes such as the formation of calcite, pyrite, uranium, elemental sulfur, and certain magnetic iron oxides and sulfides (5) clay mineral alterations (6) radiation anomalies (7) geothermal and hydrologic anomalies (8) bleaching of redbeds (9) geobotanical anomalies (10) altered acoustical, electrical, and magnetic properties of soils and sediments.

Landsat, and other space imaging systems, serve as mega-photos that depict large areas, within which clues to subsurface conditions may be evident. In general, most of the obvious structures that have surface expression had been discovered and mapped (to varying extents) over much of the world. Some regions, however, were not adequately mapped even in the 1970s, so that the advent of higher-resolution space imagery proved a boon to energy companies seeking new sources of fossil fuels. Sometimes the imagery proved especially sensitive to subtle indications of interior structures. For instance, fractures around structures in known oil/gas fields may extend further, as seen in the coherent space images, than suspected from ground work. Also, drainage patterns at broader scales may reflect control by underlying rocks involved in

suitable traps. And even vegetation distribution may disclose signs of structure. These and other indicators discernible in space imagery appealed to exploration geologists as another means to survey large areas.

The two most useful indicators discernible in airborne or spacecraft remote sensors data are fracture systems (mainly lineaments) which can control or affect the migration of gas and oil to the surface and geochemical alterations of surficial rocks by hydrocarbons which lead to compositional and color changes. This second effect is reviewed on a website that deals with hydrocarbon detection.

We will now illustrate these ideas by examining and evaluating one of the first case studies using Landsat-1 to demonstrate the feasibility of direct exploration from space. This pilot study, conducted jointly by the Eason Oil Corp. and the Earth Satellite Corp. of Rockville, MD, sheds considerable light on effective criteria for recognizing conditions that might relate to buried hydrocarbons. In addition, some of the pitfalls associated with the space approach were also discovered by carefully assessing the results reported by these investigators.

The strategy behind the study was to look at Landsat imagery of a region already established as a petroleum province, giving special attention to telltale surface indications of the presence of known underlying fields. The investigators used standard-processed and computer-enhanced versions. Rather than test capabilities in a region where there is obvious structural control and other clear-cut evidence, they selected producing areas where the surface does not give clear indication of subsurface conditions. If they could succeed in detecting hydrocarbons under such difficult circumstances, then Landsat would increase in stature as an oil/gas discriminator .

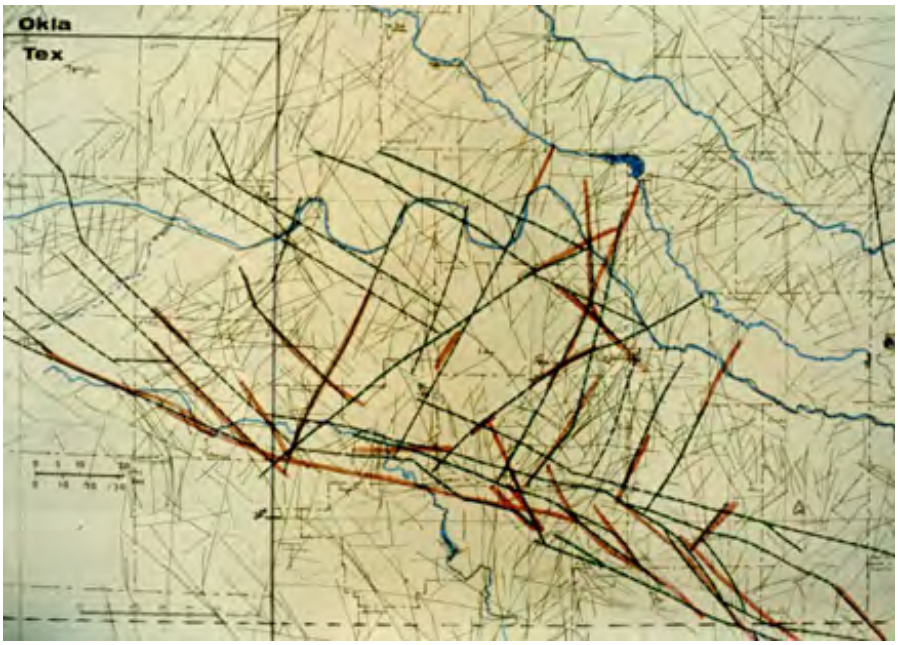
The Anadarko Basin of south-central Oklahoma fits this requirement well. Located in the eastern Great Plains, with most of the land used for farming and ranching, the Basin is one of the great producers of the mid-continent petroleum province, which also includes much of Texas, as well.

The Basin is a down-sag in the crust that has allowed up to 15,200 m (50,000 ft) of Paleozoic sedimentary rock to accumulate. Structurally, the Basin is an asymmetrical geosyncline (a regional-scale downfold), with the deepest part near the south edge. Oil and gas are present in porous rocks associated with structural (anticlines; fault blocks) and stratigraphic traps. Large gas fields occur mainly along the Basin's western half, whereas oil is more common in the eastern half. Wells as deep as 7,600 m (25,000 ft) have recovered both hydrocarbons, although most pay zones are between 2,750-5,250 m (9,000-15,000 ft).

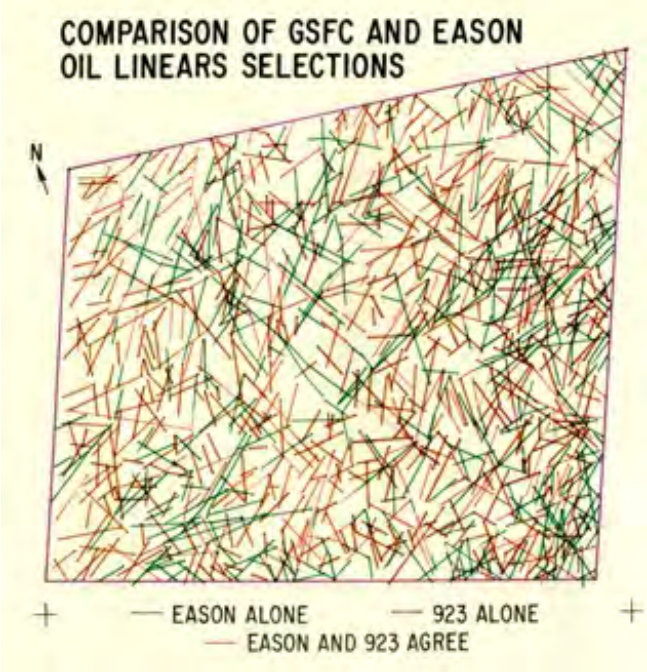
Generally, surface expression of underlying oil or gas traps in the Basin is meager, because first, there are few structural indicators in the flat-lying sediments atop older folded units and

second,there is overprinting of geologic features by vegetation and land use (grasslands; hilly sage-covered terrain; and wheat farmlands). The Eason Oil/Earthsat investigators decided to focus on two search elements: previously undiscovered fractures and subtle chemical alterations of surface rocks by escaping hydrocarbons.

Lineaments analysis was conducted by Eason Oil using Landsat image transparencies backlighted on a light table. The linear features they picked are shown by lightweight black lines on the map below. Superposed as brown and green-black heavier lines are faults that had previously been discovered and mapped. As a geographic reference, note the meander bends (curved segments) of the Canadian River, traced in blue. The majority of the Landsat-mapped linear features are inconspicuous in the imagery. Many of them are suspect, i.e., they could be non-geological or some type of lighting artifacts.



As first mentioned in Section 2, a group of four geologists, including this writer (NMS), at Goddard Space Flight Center, decided to check on the reproducibility of these map results, using the same April, 1973 Landsat MSS full scene (see below). Each person used the same transparencies (mostly winter images) as Eason Oil and worked independently of one another to minimize bias. When done, we registered the tracings to a base map, on which the Eason Oil lineaments were also plotted, as seen below. The comparison disclosed rather startling discrepancies in terms of variance between the two groups. We found only about 20% of the total linear features in common. Eason Oil chose approximately 35% of the questionable features, exclusively, while Goddard geologists chose the remaining 45%, which represented those "missed" by Eason Oil. We immediately suspected that this kind of result is partially due to considerable subjectivity in deciding whether a given linear feature a) really exists, b) is geological in nature, and c) means anything.



This suspicion was reinforced by comparing the linear features selected by the four Goddard geologists. Here are the results - a mishmash that requires the following interpretation:

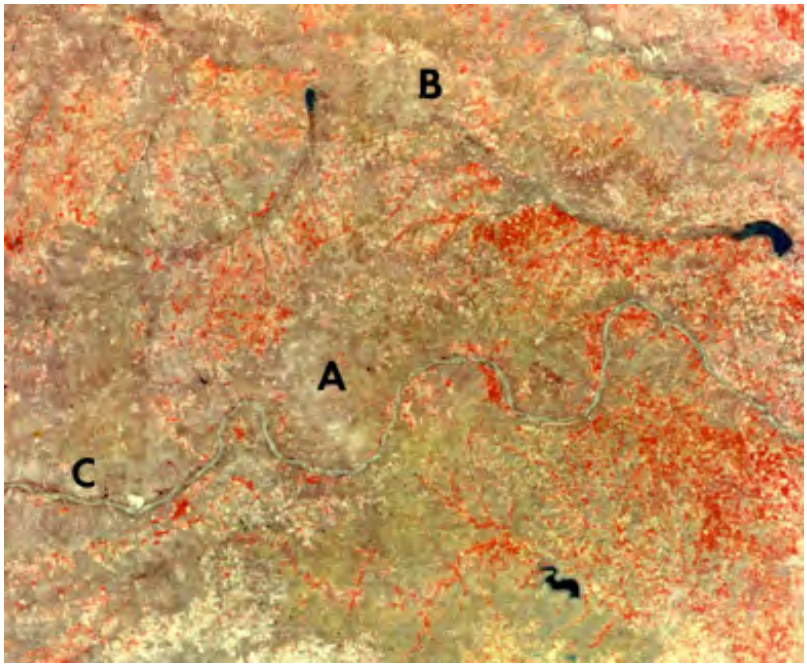
Of the 785 linear features identified by all four combined, only 4 (0.5%) were noted by every operator. From the remainder, 3 operators mutually selected 37 (4.7%), two operators agreed on 140 (17.8%), and the rest, 604 (77%), each operator found exclusively. This type of result has been reported in similar studies, although the above scores were particularly discouraging. Each geologist had ample experience in photointerpretation and special skills in analyzing Landsat imagery. Their choices were justifiable but overall, our results were questionable.

5-10: In this experiment, and in the technique of picking linear features in space imagery, what do you think was really going on behind the end result of some many linears being found but not consistently by multiple interpreters? ANSWER

The bottom line here is that there often is a strong tendency towards overkill in choosing features that appear to be meaningful lineaments. So many are drawn that it would take a monumental field effort to check them out. If plotted as rose diagrams (see page 2-9), they may

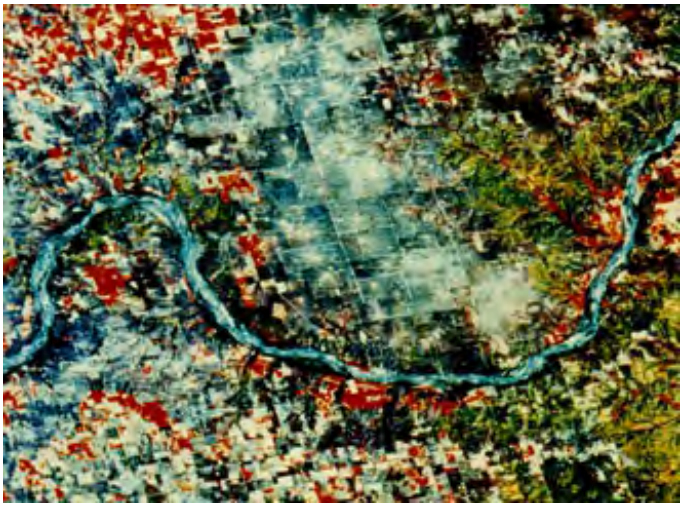
reveal valid trends for the orientations of regional fractures, because statistically lineaments of non-geological nature should be in the minority. (A study of obvious lineaments in the Adirondacks confirmed this result.) Of the 200+ prominent ones in the Anadarko Basin that were field-checked, geological fractures directly or indirectly controlled most of them, but about 20% related to human factors, such as fence lines, roads, etc. Thus, we conclude that we should combine lineaments analysis with other indicators of mineralization or hydrocarbons. This combining would encourage geologists to field-check particular sites to verify the lineament presence and nature and their possible correlation with these indicators.

The Eason Oil study sought to recognize such indicators. Their interpreters delineated certain geomorphic anomalies, such as circular patterns and unusual drainage. In the course of their image appraisals, they noticed unexpected tonal patterns that looked a bit like light-colored smudges on the images, such as evident in the April, 1973 full Landsat MSS scene that became the reference base for the study. These they called "hazy" features, as seen here:



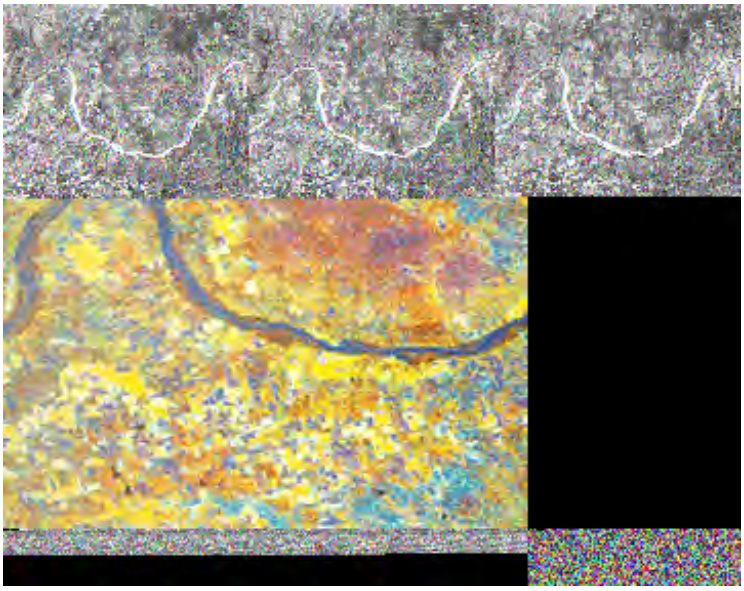
We labeled three typical hazy patterns A, B, and C. The one at A, at a bend in the Canadian River, is especially prominent, and occurs over a known oil field.

A standard false color subscene (computer-enhanced) around A shows the hazy to have a bluish-white color similar to soils in barren fields. Note the road pattern and white blotches which are accesses to producing wellheads. The yellowish areas coincide with unaltered Permian (late Paleozoic) red beds.



When we process this April Multispectral Scanner image into three ratio bands that we then combine into a color image (4/5 = Blue; 5/6 = Green; 7/5 = Red), the hazy feature at A takes on a unique yellow-green, and the red beds become orangish.

The signature for the hazy area, seen in darker orange-brown, is conspicuously different from its surroundings. It is associated with a small oil field that was developed after the Oil and Gas Map of the U.S. was published (see below). An aerial photo shows the roads that cross the hazy patch (inside the large meander loop of the Canadian River):



One might argue that the activities from the drilling had somehow lightened the whole immediate area, accounting for the "hazy". But, this is unlikely, especially since other hazies usually do not have active oil fields associated with them. The lighter tone is more likely to be a condition within the soil.

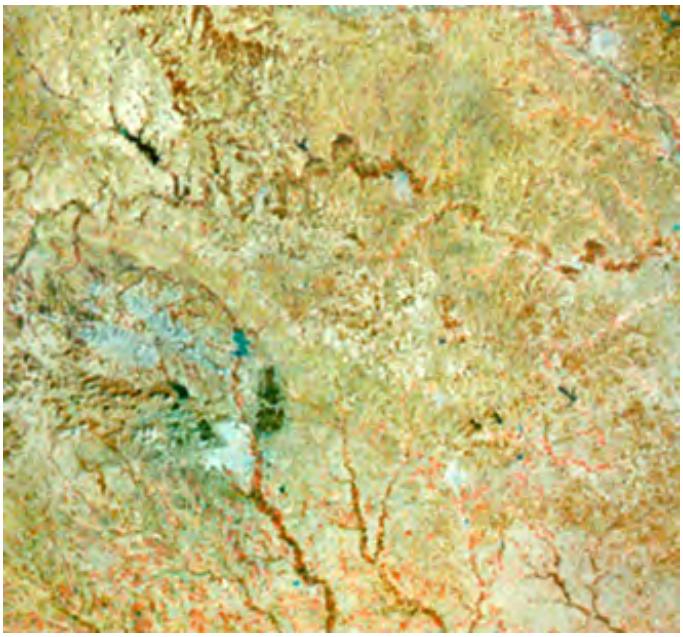
From the multiseasonal data sets, only those scenes imaged in late winter to early spring show hazies. At other times of the year, vegetation masks the phenomenon. To understand their explanation of the features, we look now at this photograph of two rock types:

The rock on the far left is a sample from the red beds (sandstones) of Permian age. Next to it is the same material that has been color bleached to yellow-brown by converting iron oxide cement into hydrated iron oxides (analogous to rust). The gray rock on the far right is a limestone (calcium carbonate). To its left is a gypsum rock (hydrated calcium sulphate). Both interior rocks appear to be altered equivalents of the primary exterior rocks. In the field, comparable altered rocks can occupy many square miles.

To account for these hazy features, the Eason Oil people postulated that chemical reactions affected the iron cement, bleaching it out, and/or transformed the carbonates into sulphates. This, they surmise, happens when sulphur-laden gases or fluids leaked out of petroleum traps and rose towards the surface, interacted with susceptible rocks, and brought about compositional changes. Microseeps along lineament would be particularly effective.

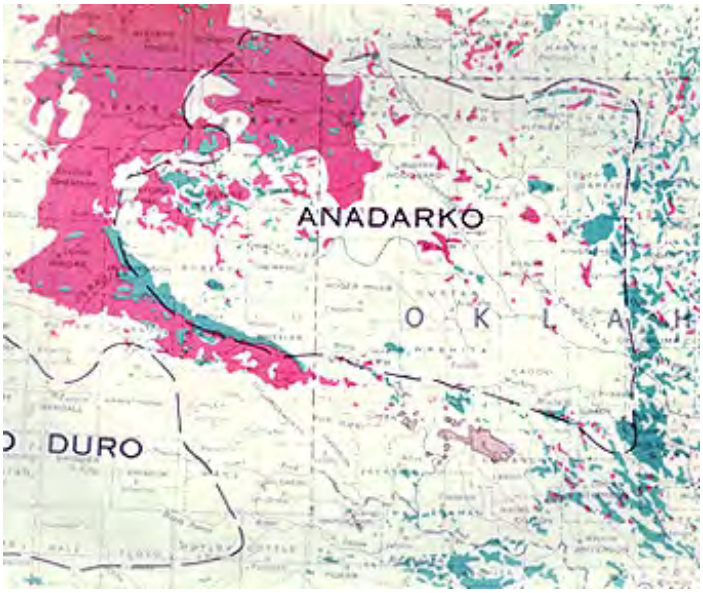
About the time of their conclusion, evidence for such changes was reported as the Doctoral thesis of Terrence Donovan (later of the U.S.G.S.), in which escaping hydrocarbons drastically altered rocks above the Cement Field, at the southeast edge of the Anadarko Basin. Dr. Donovan found a pronounced set of anomalous values of the ratio of C^{13} to C^{12} in samples collected over both producing zones in the field, shown as contoured areas below:

These values represent some of the highest departures from normal ratios known anywhere in the world. He attributed them to the effects of chemical action by carbon-rich fluids on the rocks which, as a consequence, appear bleached. The Cement Field does not show any evident hazy-type anomaly in the imagery Eason Oil used. But an image processed by EarthSat did show a whitening about where the Cement Field is located, seen as a lighter tone near the center of the image (this is a winter scene, and a trace of snow is found around a reservoir to the northwest, but appears absent in the vicinity of the Cement Field).

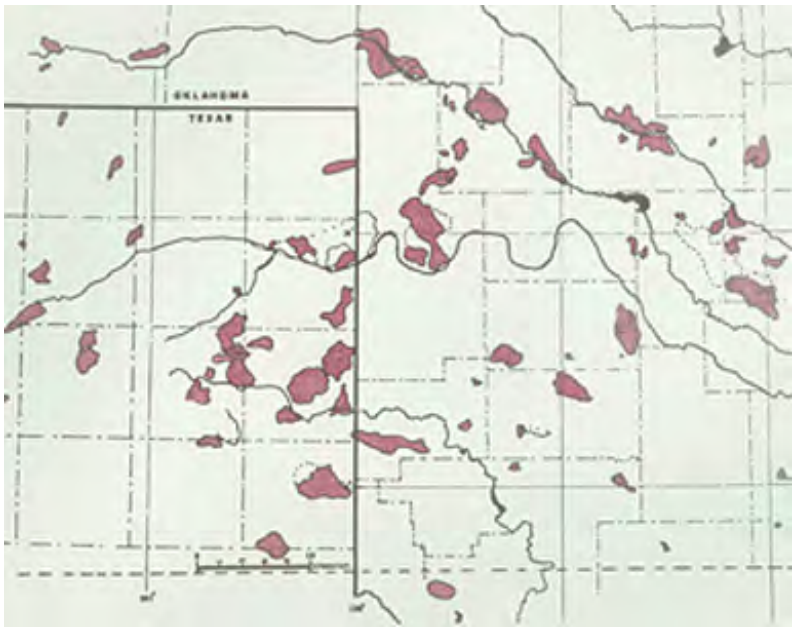


Accepting this alteration hypothesis, the Eason Oil group looked for at least partial coincidence between these hazies and the surface projections of subsurface oil or gas fields. Of the 57 anomalies they mapped in a control segment of the imagery, they claimed an association with 42 producing fields. Another six occurred above or near non-producing structures, and only 9 showed no coincidence. If this observation remained true, then detecting hazies, sometimes correlative with lineament concentrations, could promise a powerful new way to hunt for oil and gas using space imagery.

The present writer (NMS), being skeptical in habit, decided to challenge these findings. The begging question: To what extent do the hazy anomalies correspond to known oil and gas field distributions. Here is a part of the Oil and Gas Map of the U.S. published by the American Association of Petroleum Geologists (AAPG):



On this map, the Cement Field is shown in a unique color - a purplish-brown. Next, here is the Eason Oil Map of the hazies shown in purple-blue:



It's hard to check the degree of correspondence by shifting between the two maps. So, I traced the outlines of the Eason Oil hazy features (in a hachured pattern) on a transparency and then overlaid and registered it to the oil (greens) and gas (reds) AAPG map of Oklahoma. The resulting combination is shown here:

Visually, the coincidence between hazies and fields does not appear strong. This was supported by a spatial correlation analysis, which demonstrated there is no statistical significance to the pattern distribution, i.e., the coincidence is random rather than associative. In practical terms, there would be at least as much chance of striking oil by drilling into points selected by throwing darts at the map, as there would be in drilling into the centers of hazies. (That is not facetious: I did drop the overlay randomly several times onto the AAPG map - a few hazies always landed on a few oil fields.) Based on a quick field trip to the A hazy, the writer (NMS) believes hazy features are areas where wind has blown away much of the soil fines, leaving reflective quartz grains behind. Of course, if escaping hydrocarbons affect the soil, that may be degraded enough to foster the wind removal.

However, at one locality designated as a hazy feature, the writer did find convincing evidence of

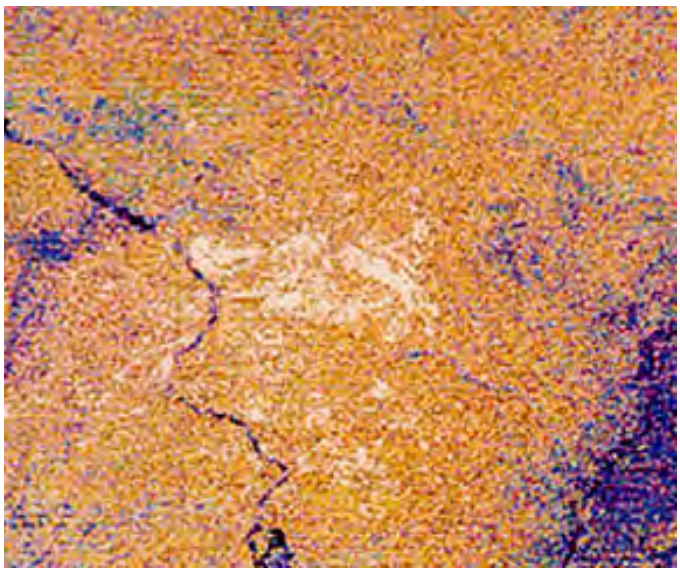
what appears to be distinct color difference attributable to hydrocarbon alteration of red beds. In a dirt road, the reddish-orange of unaltered Permian rocks gives way to a yellow-white color representing hydrocarbon "bleaching" as proposed by EarthSat/Eason Oil. Here is a photo taken at that point:

The Goddard geologists under my direction didn't perform these studies to discredit the Eason Oil study, which provided some valuable insights into the discerning power of space imagery for petroleum exploration and the potential shortcomings of the apparent results. We did them to independently evaluate this approach and to inject caution into any beliefs that this technique might become a panacea for finding petroleum.

5-11: Critique the Eason Oil study, devising if appropriate a defense of their approach. In general, what do you believe to be the most effective use of remote sensing in exploring for hydrocarbons. ANSWER

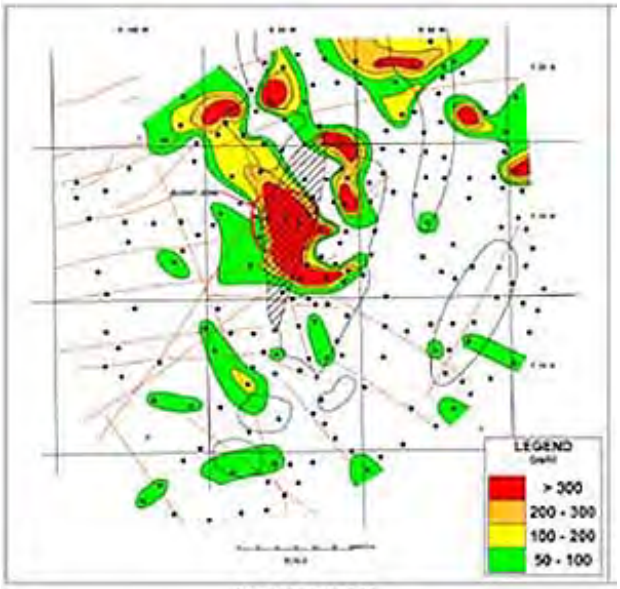
Our bottom line: "The Jury is still out" on making positive claims about oil and gas exploration if based solely on the Eason Oil/EarthSat report.

At the time of the Anadarko study, several other investigators claimed to have found similar evidence that appeared to indicate that leaking oil and gas reservoirs could indeed be altering surface rock and soil. One that seemed to confirm this was the Beaver Creek Oil Field in the Wind River Basin of Wyoming. Dr. Robert Vincent presented this evidence, an MSS Band 5 (red) to 4 (green) ratio image in which a prominent oval shaped anomaly (shown here in tan) coincided very closely with the outline of the field as determined by subsurface drilling:



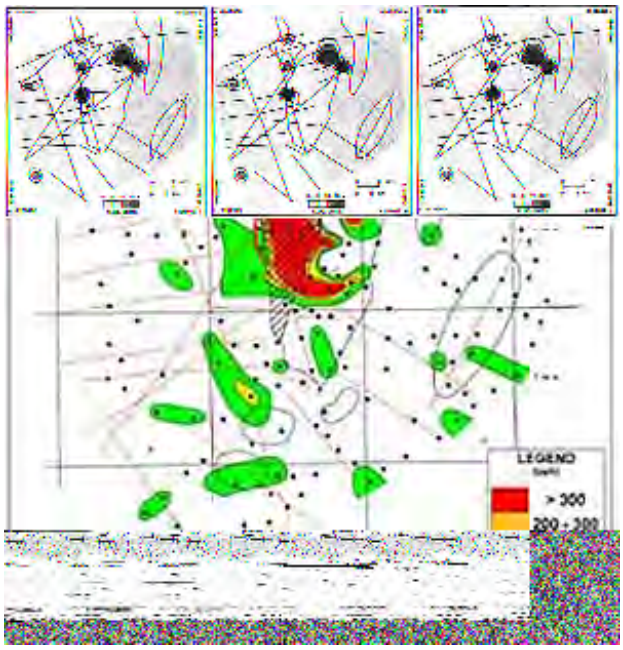
The writer (NMS) visited this field while engaged in his Wyoming investigation work. The area consisted of Lower Tertiary sedimentary rocks that were strongly dissected into gullies. Many of these beds were reddish and could in themselves account for some of the anomaly. A rather quick search for obvious signs of alteration by escaping gases or fluids failed to find any convincing evidence. But the remarkable co-incidence of the 5/4 anomaly with the outline of the Beaver Creek field suggest that this may be a valid example of the concept of alteration by petroleum compounds.

Landsat results in geological applications excited many in the petroleum and mining industries. Various companies banded together as a consortium, starting in 1976, in what became known as The Geosat Committee. Their avowed aims were along three lines: 1) to share information and conduct studies using space imagery to search for petroleum and minerals (mainly metallic ores); 2) to "lobby" NASA and Congress for a continuation and expansion of the Earth-Observing Satellite program; and 3) to provide inputs in determining and improving sensors in future satellites. One of their principal study sites was the Patrick Draw oil field near the Beaver Creek field in Wyoming. (see summary online at this website: Patrick Draw oil field). Hydrocarbons appear to be leaking as gases at various points above the oil field. This map shows the results of a field study (ground cored typically to depths of 3-4 m) that retrieved samples analyzed for propane:



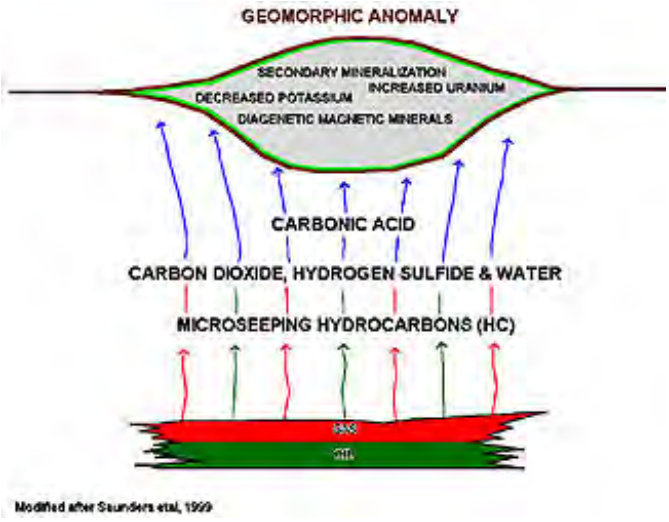
Soil Gas Propane
4' Probes

When the Patrick Draw field was overflown by an airborne UV sensor, this map of fluorescence anomalies was constructed; these results seem to confirm detectability of hydrocarbon-related gases at or above surfaces where leakage of the gases occurs:

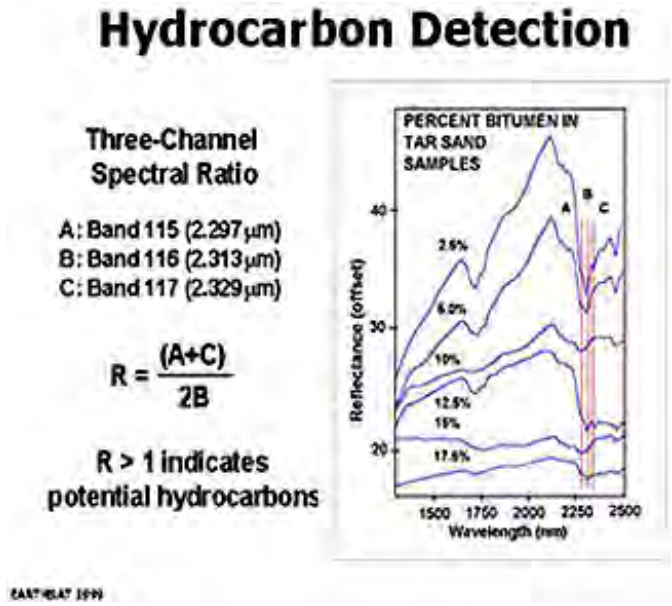


Two discoveries stemming from the Geosat study of Patrick Draw are significant: 1) a map of lineaments shows microseeps at several intersections, and 2) there is a distinct geobotanical anomaly in and near Patrick Draw - sage plants are damaged, presumably by escaping hydrocarbons, and this is detectable in hyperspectral imagery. Unfortunately, key illustrations supporting this have not been made public.

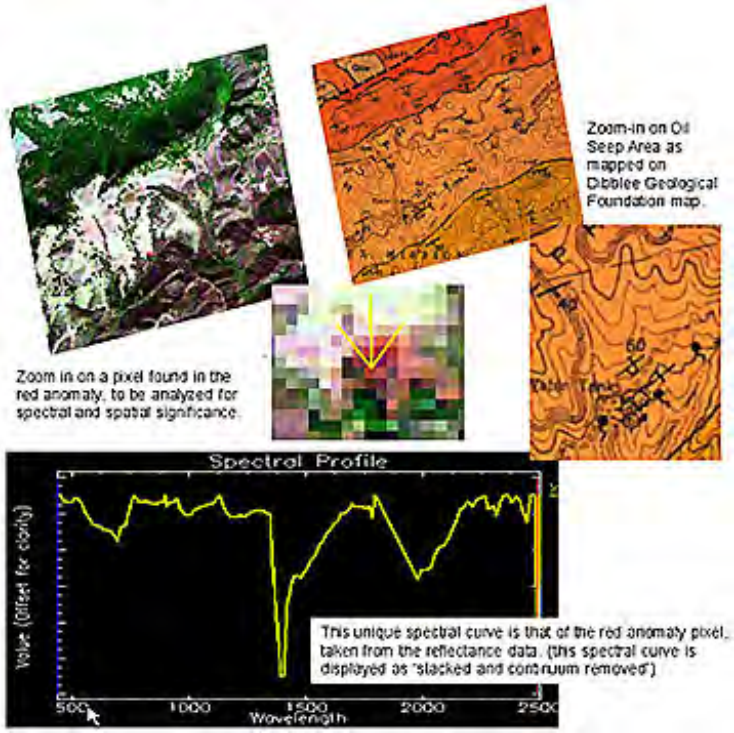
Earth Satellite Corp. (now renamed MDA Federal, Inc), and another group, Earth Search Sciences, have continued to validate data obtained from sensors on satellites and aircraft as potentially decisive indicators of subsurface oil/gas fields. This next diagram summarizes recent thinking:



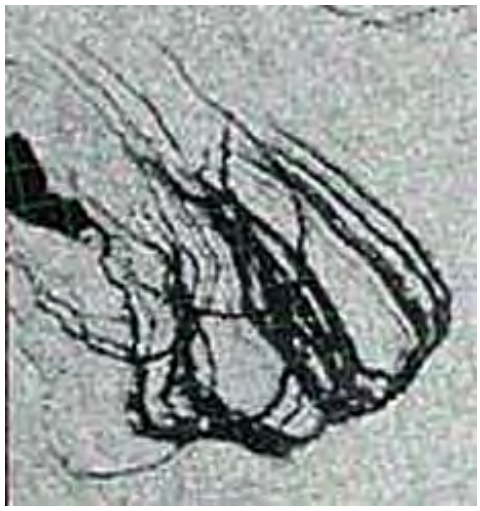
Airborne hyperspectral sensors that were flown over known hydrocarbon leaks (in some settings, called microseeps) have found that an absorption feature near 2.31 µm (one of several in the near IR) is very sensitive to the amount of a specific component of the hydrocarbons. A ratio of two reflectance values on either side of that absorption feature divided by the value of the decreased reflectance in the spectral curve at the feature's low point enhances the detectability of the hydrocarbon and quantifies its magnitude.



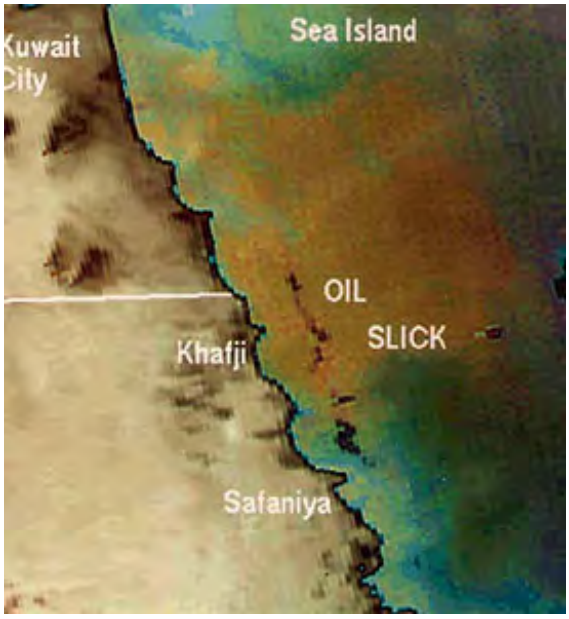
This next image display shows an actual field case conducted jointly by the HJW GeoSpatial, Inc, the Geosat Committee and Earth Search Sciences in which an oil seep that corresponds to a specific pixel (red) in the Probe-1 hyperspectral scanner image shows the 2.31 µm diagnostic anomaly (strong absorption bands at 1.4 and 2.0 µm are related to other materials):



Leaks of oil from fields below the ocean can serve both as an exploration indicator and as a source of environmental damage. Prospecting for oil beneath the open ocean requires some different techniques as well as use of some of the conventional land methods. Oil seeps and slicks can remain intact on the surface and may be detectable in Vis/NIR and radar imagery. The Earth Satellite Corporation has developed SEP - the Seep Enhancement Algorithm - to bring out an oil signature using radar imagery. Here are two examples:

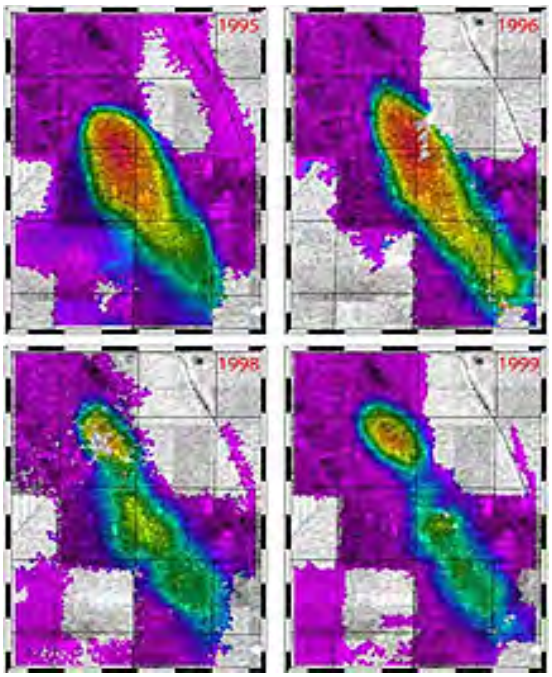


Oil slicks can be both natural or due to manmade oil spills. This EarthSat image shows a slick off the coast from Kuwait as rendered in a natural color Landsat image.



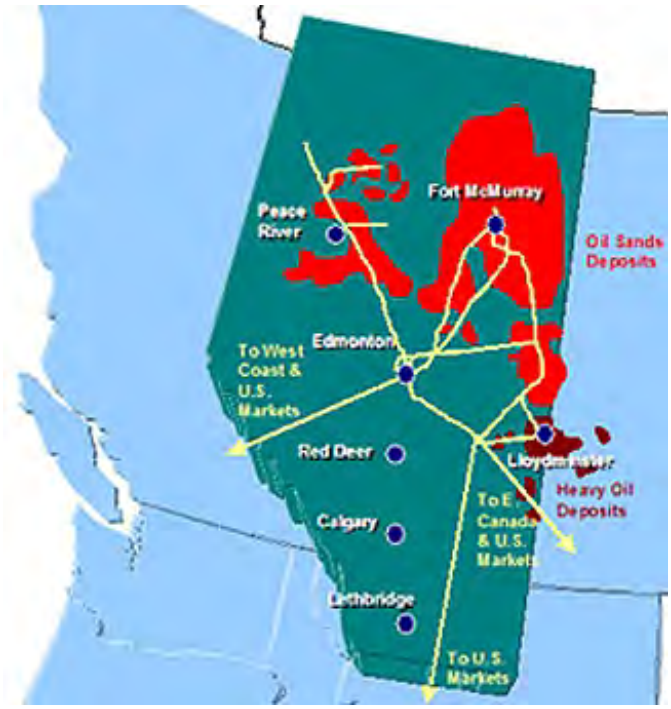
Specialized remote sensing can monitor another aspect of petroleum withdrawal not necessarily expressed as leaks. In time, as the oil is removed from pores leaving a partially filled void, the rock units bearing the oil start to contract or crush inward into the voids as support diminishes. This is commonly expressed by all the overlying units pushing downward on the now compressed reservoir rocks, giving rise to progressive surface subsidence. This lowering of elevation can be

monitored by radar interferometry (see page 11-10 This next illustration, made from ESA radar data, shows interferometry rings, which can be computed into elevations, at the Lost Hills oil field in the San Joaquin Valley of California. The field is subsiding now at a rate of about 3 cm (1.2 inch) per month, with a cumulative drop since 1989 of 3 meters (10 ft). Subsidence is greater at the two ends of this 1.5 x 6 km (~1 x 4 miles) elongate field.



At the time of this writing (February, 2007) the intensity of the economic and political aspects of the availability and costs of oil and gas as still the principal energy sources for such multiple uses as transportation, heating, and plastics is at a "dangerously" high level. Prices are rising everywhere because of OPEC decisions, rapidly growing markets (e.g., China), and threats of cutting production (Iran's nuclear program). Alternate sources of energy, including oil in non-conventional modes of recovery, are being pushed. Two huge potential suppliers are Canada (tar sands) and Venezuela (heavy oil; requires pumping in hot water to release the oil from its host rock). Estimates of available oil from these types of deposits in Alberta, Canada approach, or may exceed, two trillion barrels (Venezuelan heavy oil is at least one trillion barrels).

The Canadian oil sands were first discovered in the late 1700s. The sand units outcrop at the surface in the northeast part of the province of Alberta but have a wider distribution subsurface, as seen in this map:.



The Cretaceous sandstones that contain sticky, near-solid bitumens (up to 20%) filling interstitial pores have been called Athabasca Tar Sands or now more commonly Alberta Oil Sands. Here is a surface photo of an outcrop rich in the blackish tar that pervades the rock.

As seen by the Space Shuttle astronauts in 1989, the area along the Athabasca River where surface stripping of the oil sands is active, is shown in the middle. The town of Fort McMurray, which has since grown considerably in anticipation of greatly increased production, appears to its south:

Here are two aerial views of this main strip mine complex;

The oil sands after surface removal are further broken up and then extracted from the rock pores by subjecting the material to hot water and other chemicals. A barrel of thick oil requires processing of about a ton of the oil sand. Here is a processing plant where this is accomplished:

For decades the cost of obtaining liquid products from the sands for further refining had been too high to turn a profit. But that has now rapidly changed as prices of oil and gasoline in the U.S. and worldwide climb in response to demand. This has convinced various companies to set up greatly expanded operations. If stripping becomes impractical, the cost of subsurface mining

must be weighed against expected increases in price and demand. In time recovery is likely to include underground mining. One plan is to confine stripping recovery to the summer months and go underground when the snow cover impedes surface mining. The reserves in Alberta are huge - comparable to that known in the Arabian Peninsula. There may be enough tar sands in Alberta to place Canada in "the driver's seat" in the 21st Century; for North America alone there could be sufficient oil sand reserves to last a century. As of 2009, the U.S. gets 22% of its oil (about 1.5 million barrels a day) from these Canadian deposits; China is becoming another major customer.

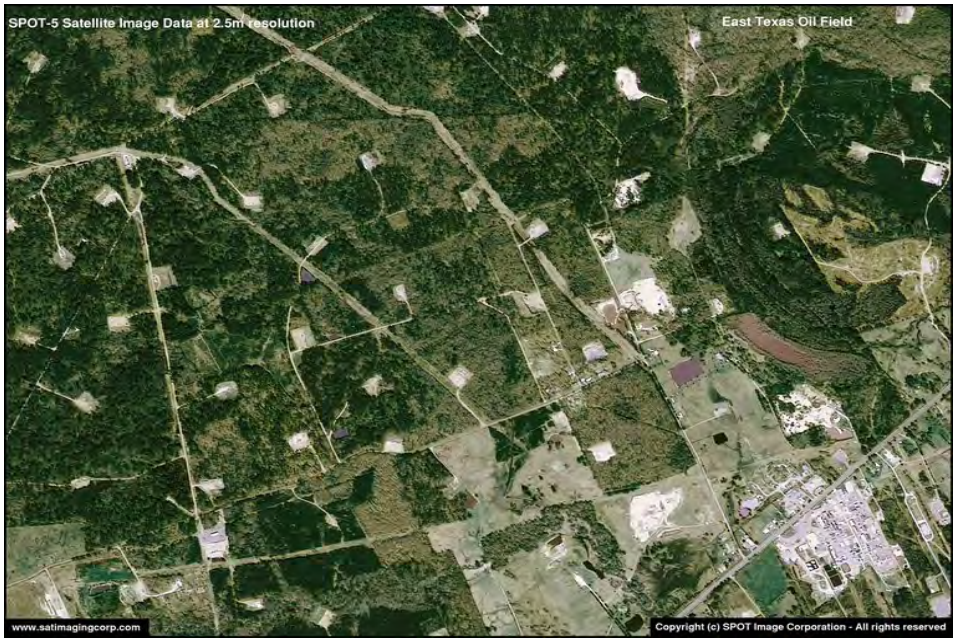
The Canadian government is carefully monitoring and controlling the expansion of the oil sand industry. Waste at the surface, as seen below, must in a reasonable time be reworked to form a smooth surface and then replanted with trees and grass. Here is where space imagery will play a leading role - determining that the reclamation requirements are being met.

Among other uses of remote sensing for applications related to oil and gas exploration and production include: 1) monitoring pipeline location and possible breaks (leaks), 2) monitoring environmental damage from drilling for oil/gas, 3) monitoring recovery of the natural terrain after a field is no longer producing, and 4) producing land use/land cover maps of a region where new or increased development is anticipated.

An example of item 2 is offered by this photo taken from the International Space Station. It shows the barren ground patches around individual drilling sites and developed wells in the West Texas Permian Basin - a major producer in the U.S.:

Astronaut photography is occasional and target-selective. Environmental effects in areas immediately around drilling sites need more frequent monitoring. This SPOT-5 image shows

multiple sites near Carthage in west Texas. There appears to be almost no adverse impact on areas surrounding the white scars that result from land cover clearing at each site.



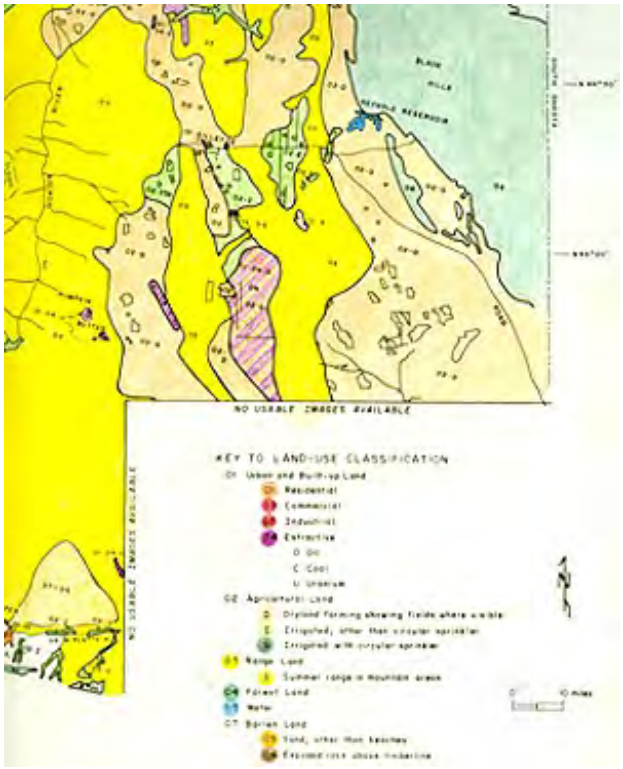
This EO-1 satellite image also has environmental significance. It shows the markings saturating the hills of southern California's Coast ranges at the Elk Hills oil field, which first started producing in 1912 and is still active. It has yielded more than 1.3 billion barrels of oil and some natural gas. For many years it was the kingpin of the Naval Petroleum Reserve before being sold to Occidental Petroleum Co. in 1997.

Regarding item 4 - land use mapping of a developed or developing natural energy-rich region - is

done in compliance with regulations for managing all of the resources. The writer (NMS) participated in a singular example of this requirement, which engendered an outcome of some notoreity. Here is the story:

Landsat-1 started to send back images in late July of 1972. The writer (NMS) was then a co-investigator with Dr. Robert S. Houston and Dr. Ronald Marrs of the Geology Dept., University of Wyoming (Laramie) in the NASA-funded Wyoming Geology program. I received the first images of Wyoming in mid-August and left at once for the field there. I spent a week roaming the state to check out what I could relate between image features and ground truth. Upon return to Laramie, Drs. Houston and Marrs and myself drove to the state capital, Cheyenne, to meet with state officials about possible uses of Landsat imagery in environmental and land use projects. A big one was in the offing - 8 million dollars to prepare land use maps of the Powder River Basin. That basin is one of the richest energy sources in all the U.S. - huge (and thick) deposits of coal, active uranium mining, and some oil and gas productions. The maps were needed within 3 years.

The state officials decided to gamble and earmark $60000 for the University of Wyoming to "try" to produce some preliminary maps using Landsat. The project began in September. On January 13, 1973 the faculty and students who worked on Landsat imagery presented a large (30 inches by 16 inches) folio with a series of maps - all patiently colored by hand using student help - that addressed several themes. Most important was the land use map, a portion of which is reproduced here:



The State was so impressed, it accepted these folio maps as adequate in meeting its multi-million dollar objectives. No further work was done.

A copy of the folio was soon sent to me. I too was impressed. I brought this copy down to the Earth Observations program office at NASA Headquarter for the managers to see. Their reaction was almost boisterous. This is just what they were looking for when they went to the U.S. Congress in just a few weeks to an Appropriations Committee hearing. The NASA chief, Dr. Len Jaffe, just commandeered (syn., confiscated) the folio and told its story to the congresspersons. (Never got the folio back; students had to color another one for me.) But a small price to pay to get the overall program in high gear.

Suffice to close this page with the remark that since the launch of ERTS-1, the petroleum industry has found new oil and gas fields with the aid of space data and has developed criteria

from the images that continue to prove worthwhile in planning and conducting exploration programs, which are leading to payoffs. Most of the successes have come by using space imagery (as has been done before with aerial photography) in the tried-and-true (conventional) way of using the pictures as base maps on which to analyze and plot structural patterns and trends, often supplemented by recognition of stratigraphic units. (Detection of surface alteration, while it happens sometimes, remains a rather rare event.

In sum, remote sensing aids in exploration for oil and gas by 1) providing overviews of the regional geologic setting in which oil and gas is being sought; 2) helping to define existing fold/fault structures; 3) demarcating linear features that are usually fractures along which hydrocarbons migrate; 4) detecting alteration of rocks by escaping hydrocarbons; 5) finding other signatures indicated by fluorescent anomalies in the UV and compositional anomalies in the IR; 6) noting oil directly as leaks, spills, and seepage in the oceans/lakes or on land; and 7) observing environmental damage associated with drilling, pumping, pipeline transfer, and refining of hydrocarbons.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Before we begin the LANDSAT GAME (below), we will now demonstrate for you a suggested reasoning strategy involved in identifying this Landsat scene.

Hint: Somewhere in Europe.

The first clue is that there appears to be a large body of water on the right (east) side. Most likely this is an ocean rather than a lake. This alone narrows down the possibilities. The only parts of Europe (stated above as a hint; the images in the game will have similar hints) that would have marine waters to the east would be several areas in Scandinavia (for example, Denmark), Italy' east coast, Spain's east coast, part of Greece, and England, Scotland, and Ireland. There is no obvious variation in topography, thus the land is probably flat, ruling out Greece, and most of Italy. The countryside seems to have mostly farms and some woods. There appears to be a major east-west river that ends in a large estuary. On that river is a large area that looks like a purplish-red blotch (normally the color would be blue in a standard false color composite, but this TM image uses one of the longer wavelength bands that then renders the area in the blotch color). Its relative uniformity suggests a major city. The metropolitan area represented by the blotch is quite large. Too big to be Edinburgh in Scotand (which is on an estuary). Dublin is not inland, but on the Irish Sea, so that is eliminated. In Spain only Barcelona comes to mind as fairly large, but again that city is directly on the Mediterranean. Venice might be a possibility (the river would then seem to be the Po) but it too is much closer to the Adriatic Sea.

So, by progressive narrowing of possibilities, I have been left with very few likely candidates for identifying the region making up the scene. But my knowledge of geography provides me with a plausible answer. The estuary is the key. I happen to know that London is on an east-west river - the Thames - that ends as a large mouth, the Thames estuary. The countryside in southeast England is flat and well-developed - farms and smaller towns. This, then, is my choice, and it is a correct one.

This, then, is the procedure I used to zero in on the location of the Landsat scene. It, or variations, can now be applied to the Landsat scenes that follow (these are scaled to about 100

miles on a side).

THE LANDSAT GAME

Your challenge is to identify where each of the (mainly) Landsat images below is located. Hints are given. No cheating - make your guesses first and then go look at the answers.

A

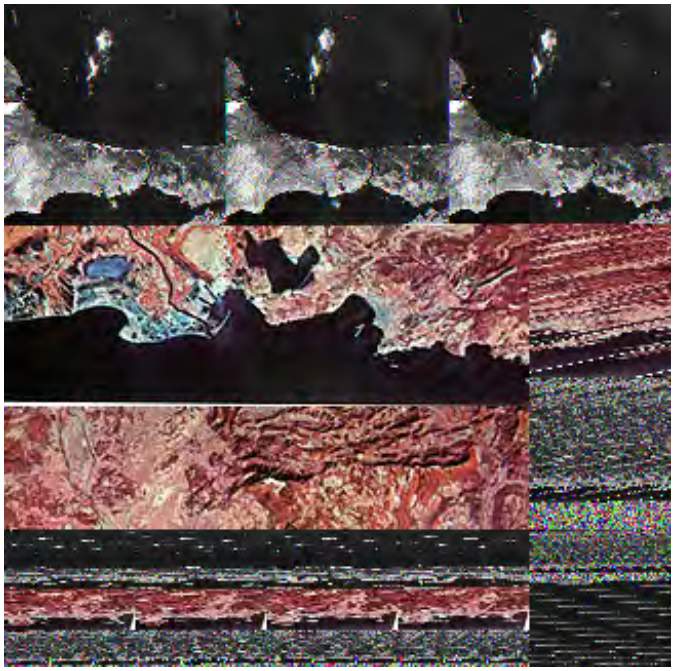
Home of Ancient Philosophers and Big Games

Check the answer to this especially for some great information

B

Allied Invasion from the South; A Coastal Playground

C

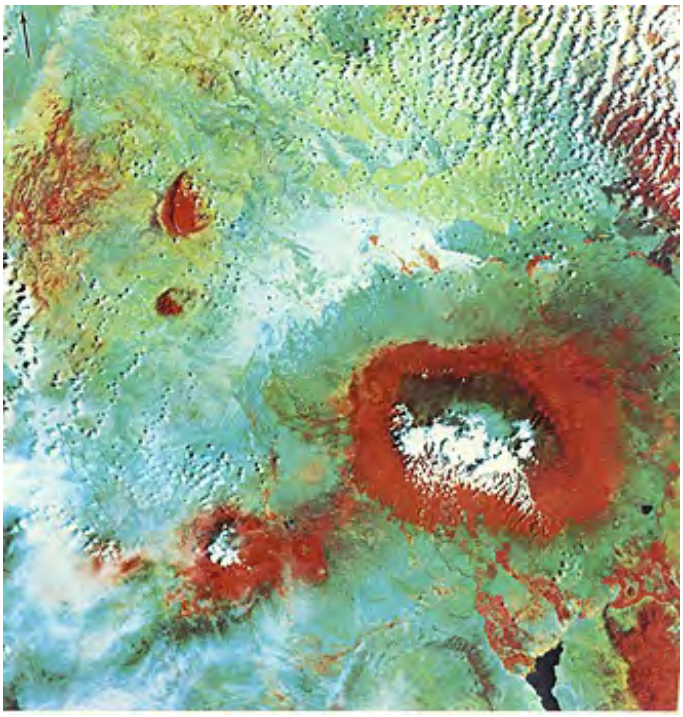


The Meeting of Two Continents (Geographers' arbitrary decision)

Perhaps this scene's identity may become more obvious when it is placed in a larger context, as is done with this computer merge between a Landsat natural color image and topographic data from the SRTM mission.

Or going to the other extreme by looking at the area around the "riverlike" body of water separating the two large "sealike" water bodies. At this meeting of the continents, there is a great city, shown below - famed as a center of Christian Orthodoxy and later important in the Muslim Faith. This is an astronaut photo taken from the International Space Station:

D



Tallest Mountain on the Continent; Big-Game Country

E

Longest Manmade Structure in the World; Note large Square (right)

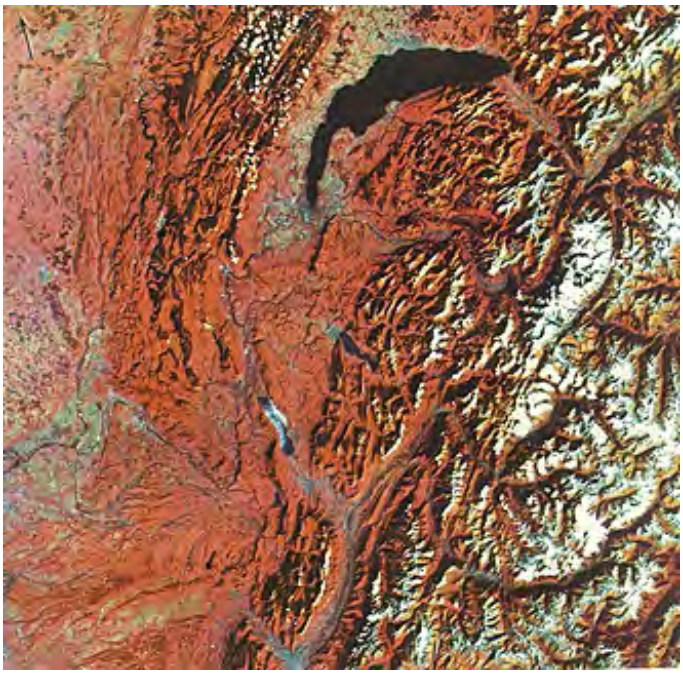
Check out the Answer to see this Longest Structure.

F

Asia's famed financial city is at the lower center in blue

(see SPOT 3 closeup below)

G



UN Home City here; Land of the Yodelers

H

The War of the 1960s

When you have made your guesses, click here for the answers.

[BACK]

Primary Author: Nicholas M. Short, Sr.

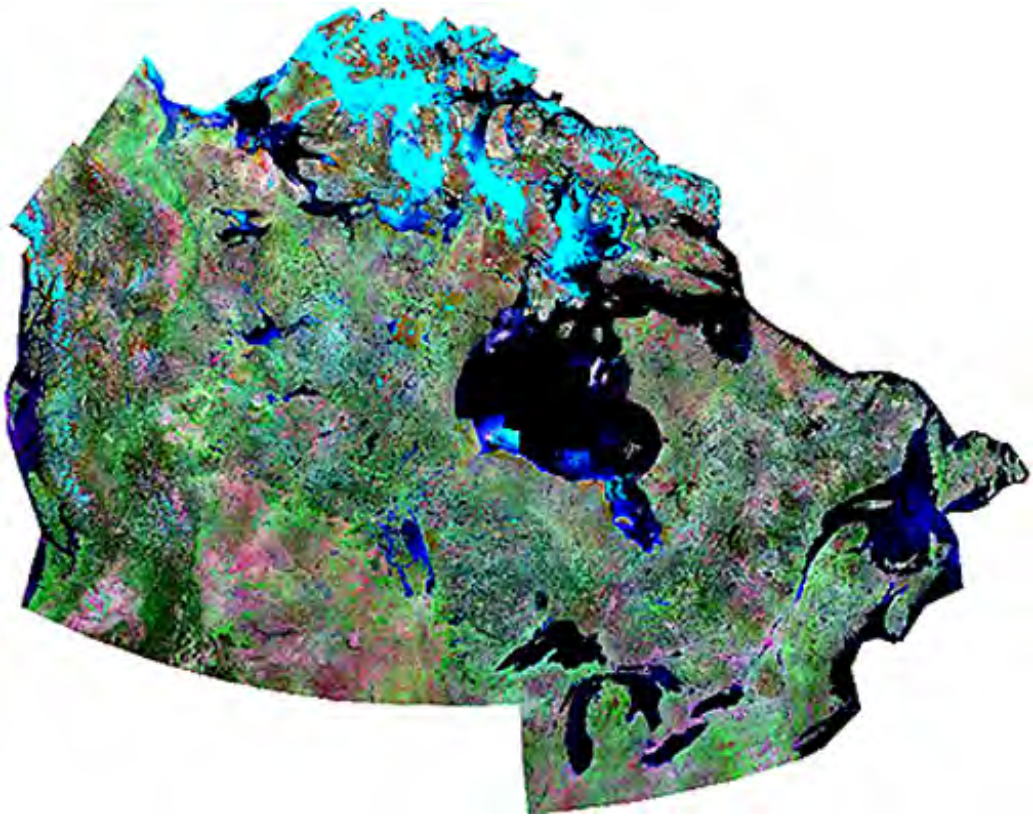
The tour from space now goes international, as in the next six pages we will look at scenes from five different continents to get a feel for the diversity of landscapes worldwide. On this page, various images of parts of Canada will be examined.

AN EXCURSION ACROSS THE CONTINENTS

The many Landsat images spread over the next 6 pages that comprise Part 2 of this Section have mostly been selected from the master set used in developing the 1976 NASA publication *Mission to Planet Earth: Landsat Views the World*, using imagery acquired by Landsat 1. Scenes made by other earth-observing satellites and as photos on the ground are also used in this excursion The description associated with each Landsat image is extracted from the captions in that book; consult it, NASA SP-360 (residing in many libraries), for more extensive descriptions of the scenes reproduced in the following Tutorial pages. Recall, too, that you have already seen a number of foreign images, e.g., specifically, London, Paris, Florence, Peking, Tokyo and other cities on page 4-4 and the Game preceding this page in this Section.

Canada

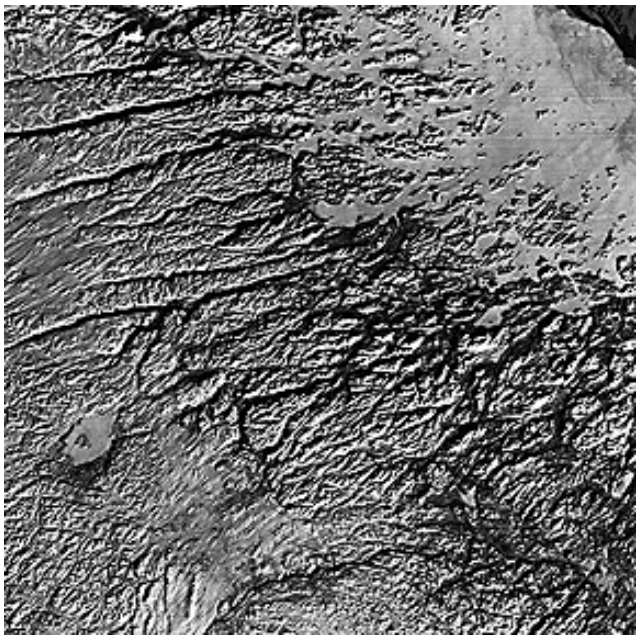
Our trip to view examples of the Rest of the World begins in Canada. Canada is the third largest - areawise - country on Earth, although its population is around 40,000,000. Here is a mosaic of Canada, made by Earthsat from Landsat images, and below it a map of this country with its principal cities (note: Canada has Provinces, not States):



We'll begin with Newfoundland. A popular attraction on its north coast is L'Anse aux Meadows, now considered to be the first place in North America visited by Europeans. The Vikings landed there and stayed for a while, building sod-covered dwellings to protect from the harsh winter. Carbon-dating of fire embers places this around 1000 A.D., thus preceding Columbus by almost 500 years. Here is an ASTER view of the area, at the northern tip of Newfoundland, and a ground photo:

Off the east coast of Newfoundland is a curiosity that almost no American knows anything about. There are three islands that are still the property of France - a vestige of the days when Canada was a French possession. They collectively are known as St. Pierre and Miguelon. Comprising just 93 square miles, they host only about 7000 inhabitants. In the ASTER image below note that the two larger islands are connected by a natural sand spit called a tombolo.

Labrador is one of the premium regions in Canada to see the styles of deformation in ancient (Precambrian) terrains. These next three images show 1) another space image in which the black and white rendition brings out fractures and other structures, 2) an oblique aerial view of the bleak scenery, and 3) a detailed aerial photo showing style of folding.



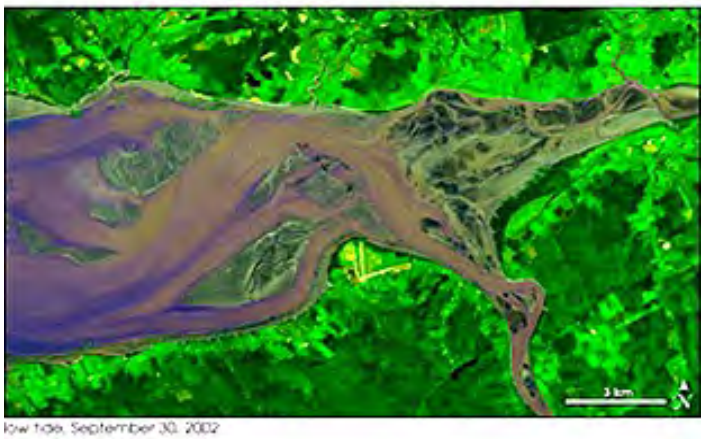
In eastern Canada, Nova Scotia is a Province in the Maritimes. This image shows the Cape

Breton Island (this becomes an "island" because the narrow Strait of Canto cuts across its southern boundary), and other parts of northern Nova Scotia and the tip of Prince Edward Island.

Cape Breton boasts of the most rugged coastal scenery in the accessible eastern part of the North American continent. Its Cabot Trail is famed for its scenery.

Between Nova Scotia and New Brunswick is the great indentation known as the Bay of Fundy. The Bay is noted for its extremes in water level fluctuations. The Bay of Fundy has the largest variations in tide levels in the world - as great as 17 m (56 ft), resulting from the tidal bore effect caused by the tapering and shallowing of the Bay floor. Timing of the tides could contribute by helping the in and out surging waters to "resonate". This pair of Terra ASTER images shows the tidal extremes around Cobequit Bay in the Minas Basin:

high tide, April 23, 2001



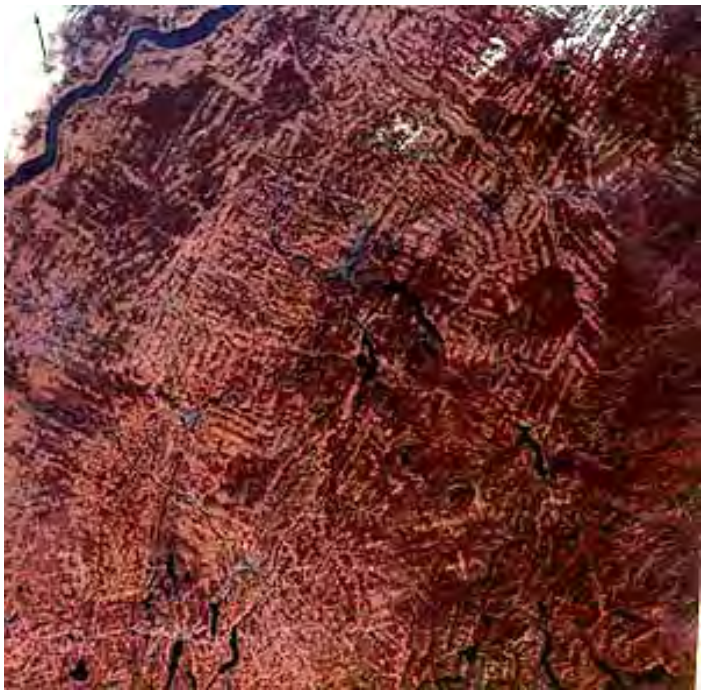
low tide, September 30, 2002

Moving northward into an area where eastern Quebec meets the western boundary of the Labrador subprovince of Newfoundland, we see next a segment of the Canadian Shield. This is the Labrador Fold Belt, made up of tightly folded sedimentary rocks wedged between two granite complexes. The structures represent an ancient (Precambrian) tectonic zone formed when one irregular continental mass collided with a sedimentary basin; the entire region has been much eroded to form "roots of mountains". The present surface is strongly glaciated, leaving low hills and lakes and a paucity of trees so that rock exposures are widespread. The high northern latitudes favor an arctic-like vegetation assemblage called tundra. This inhospitable landscape is very sparsely populated, with a few small towns built near iron mines.

Less well traveled by tourists, but still replete with interesting places, is the Gaspe Peninsula,

south of the St. Lawrence River:

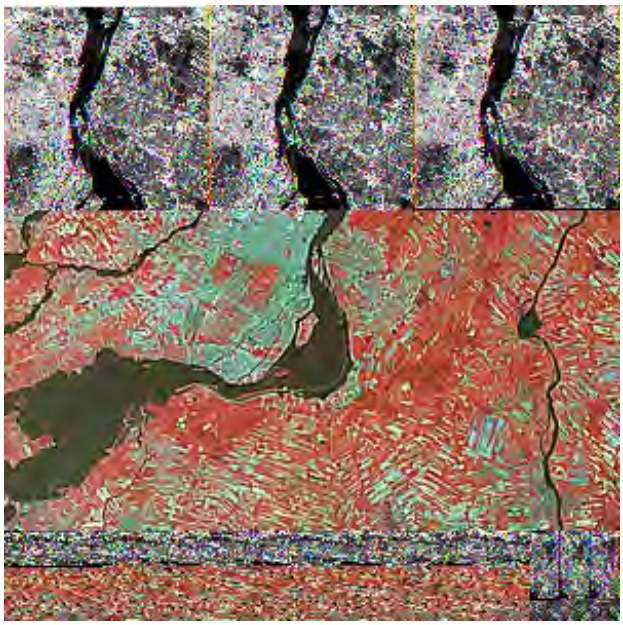
The next scene covers some of the Eastern Townships in heavily forested (spruce; fir; pine; tamarack; birch; maple) hills in the Laurentian Lowlands of the Province of Quebec. The St. Lawrence River appears in the upper left and glacially-elongated lakes are scattered throughout the scene. The most striking feature throughout the image is a distinctive pattern of land-clearing by deforestation that stands out as elongate bars with jagged edges. Their lighter red color shows them to be grasslands. These long strips tend to be perpendicular to roads and small streams. They are a land use pattern brought over from France and adopted by the early French settlers (seigneurs) who colonized this region in the 1600s; the Cajuns around New Orleans used this style.



The capital of the Province of Quebec is Quebec City, one of the oldest settlements in North America, founded by the Frenchman Samuel de Champlain in 1609. It was here that the battle of the Plains of Abraham in the Seven Years War in 1759 that the British General Wolfe defeated the French General Montcalm (although both were killed during the conflict), which led to France ceding Canada to England in 1763. We see the city in a SPOT image and in a ground photo which shows the bluffs along the St. Lawrence River and the Chateau de Frontenac, a plush hotel that is part of the Canadian Pacific rail system.

Most of Canada's cities lie in regions near the U.S-Canadian border. Here is a Landsat image that shows Montreal, Canada's second largest city, and its surroundings. Note the elongated farms which follow the "long lot" style characteristic of France (and seen again around New Orleans, once settled by French Canadians).

This SPOT-4 panchromatic image shows much of the central part of Montreal.



And here is this city as seen from the ground, in the early evening:

Here is a Landsat 7 image of the western half of Canada's capital city, Ottawa:

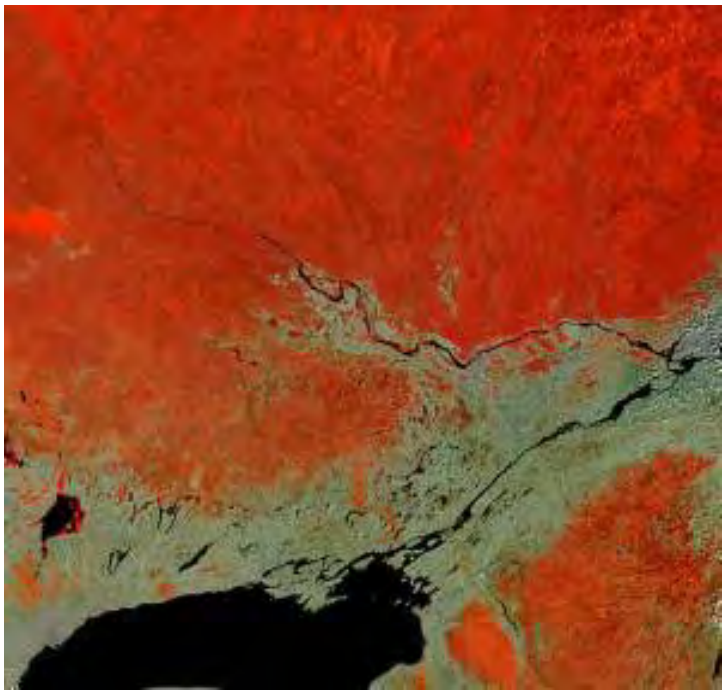
The red rectangles are part of the government's experimental farms located within the city.

The largest city in Canada is Toronto, at the northwest end of Lake Ontario. Here are two views of the city from space:

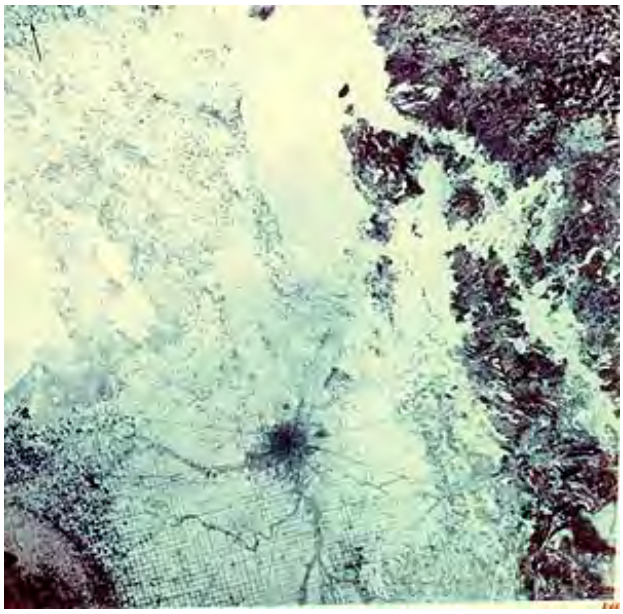
The Quickbird satellite has produced this high resolution of downtown Toronto's tall buildings:

Compare this scene with the ground photo showing Toronto's skyline. The tall needlelike structure on the left is the CN Tower, the city's best known landmark.

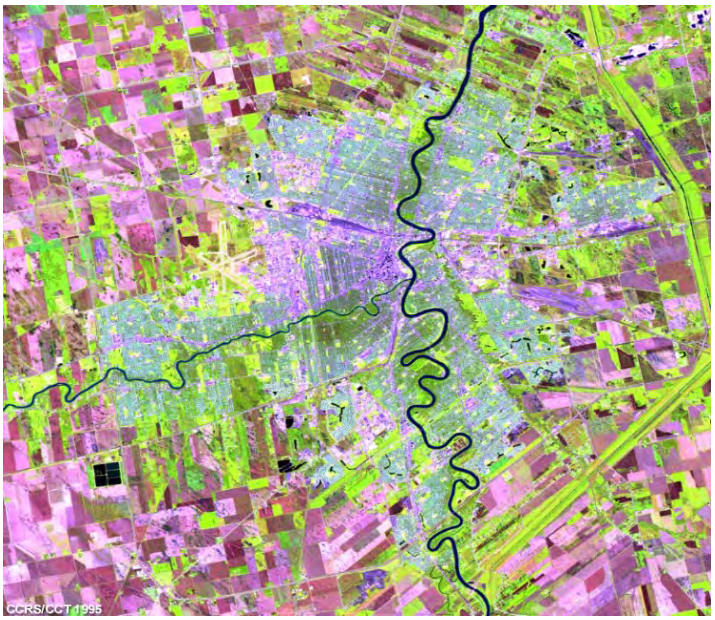
Most of the Canadian Shield in the eastern two-thirds of Canada is undeveloped, with few towns. Forests - both coniferous and deciduous - are still largely pristine. In this satellite image, the red denotes such forests; these are almost as thick as those in the Amazon and Congo.



Winter still remains in this March scene in the Province of Manitoba just north of North Dakota. Dominating the scene is Winnipeg, appearing black against the light snow because of snow removal and melting in that large city. In the countryside, dark lines (cleared farm roads) block out squares that are one-mile sections (part of the Township-Range system of surveying) similar to those so prevalent in the Great Plains (the lowlands here are an extension of that physiographic unit into Canada). Lakes Manitoba (left-center edge) and Winnipeg (top center) are remnants of a once larger lake system that covered wide areas of the plains at the end of the Pleistocene ice age. The dark areas that extend over the right third of the scene are the western edge of the Precambrian Canadian Shield. Pine and spruce are the prevalent trees in this part of the Shield.



Winnipeg looks less foreboding in summer, as evident in this image which brings out the city's location within extensive farmlands:



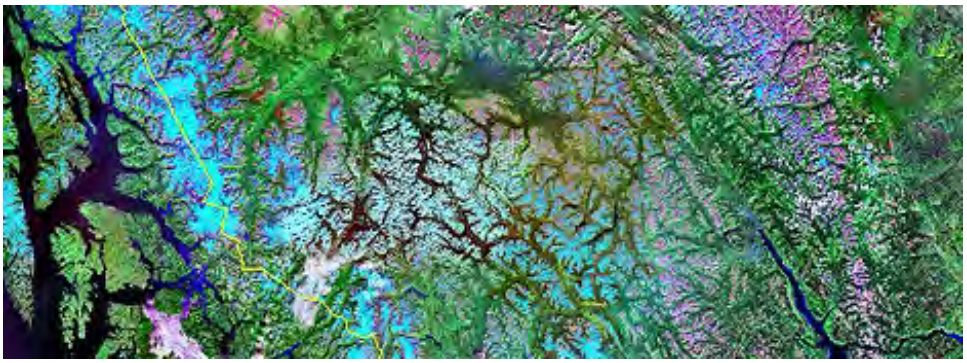
Now, a look at the border between the U.S. and Canada in eastern Montana. The borderline is sharply defined. Why? Not because of a fence but the result of different land use practices. The Americans have opened up the high plains to farming; the Canadians have left most of their side in its original state - grass-covered rolling swales.

A strange feature showed up in a Digital Globe image of "badlands" in Alberta. Erosion has carved into clay-rich sedimentary rocks leaving a bas-relief figure that remarkably resembles an Indian, complete with a warbonnet. See for yourself and judge its authenticity: is it manmade(?):

The plains of Alberta contain two large cities - Calgary and Edmonton. The former is the oil capital of Canada. It lies less than 160 km (100 miles) from the east front of the Canadian Rocky Mountains (Rockies) which have constricted to a tectonic zone about 250 km wide. Banff and Lake Louise are famed resort towns in these mountains.

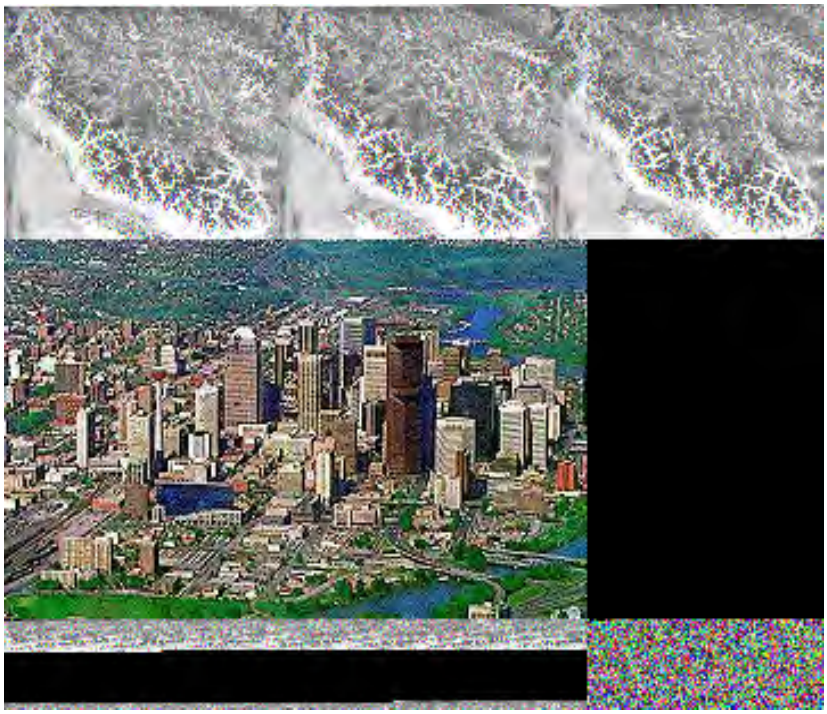
This satellite image shows parts of Alberta and British Columbia, including both Edmonton and Calgary. Lake Louise is a world-famed resort (located to the upper left of the Highway marker for Route 1 [lower center]). The Chateau Lake Louise and the sheer cliffs to its south are shown in the aerial view below:

The rugged Canadian Rockies appear in this satellite image - transect across British Columbia, from the Great Plains to the Pacific Ocean.



To one traveling on the Canadian Pacific scenic railroad route, the mountains of western Canada appear continuous. But the Rockies blend in with the wide expanse of the Coast Ranges. This Radarsat image shows these mountains, here traversed by the Columbia River just north of the U.S. border.

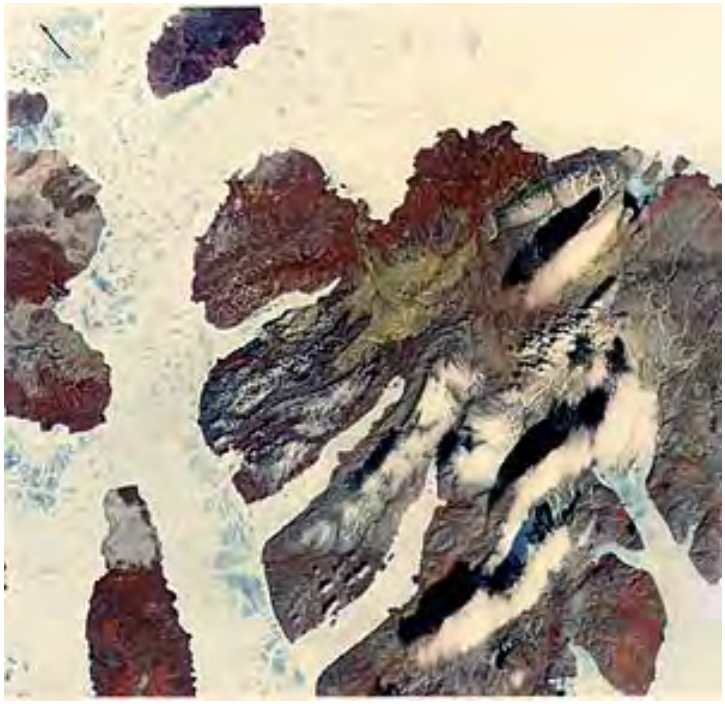
This next image is a HCMM nighttime image covering much of southwestern Canada (swath width = 715 km [447 miles]). Starting in the upper right with part of the Rocky Mountain Trench, the terrain passes through inland hills and the mountains of the Coast Ranges (the black is cold nighttime snow), part of the Inland Passageway, and the northern half of Vancouver Island. Underneath this night expression of topograpy lie many tectonic terranes (see Section 17).



The Jewel of western Canada is Vancouver, BC, now a full-fledged city built on the Delta of the Fraser River. Across from it is Vancouver Island, with its tourist-popular small city of Victoria. South are the Georgian Bay Islands that extend to the San Juan Islands of Washington State.

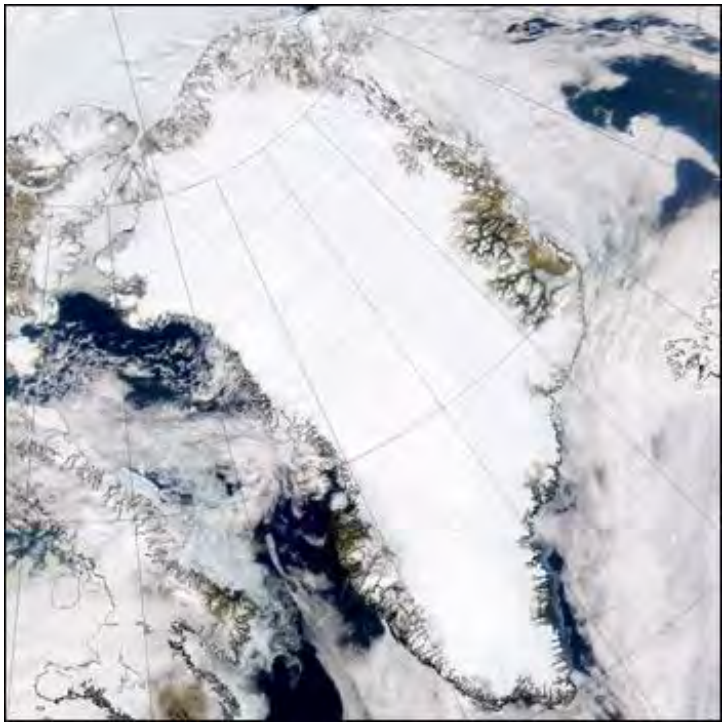
This is an aerial view of Vancouver:

Tundra again is typical of Canadian island surfaces well above the Arctic Circle. In this scene, the major land body is Melville Island, made up of Paleozoic sedimentary rocks and characterized by deep embayments that are valleys gouged out by continental glaciation of the Pleistocene, now occupied by the Arctic Sea which even in this July view is still completely frozen. Human life in this part of Canada's Northwest Territories is limited to a few native villages (Eskimo types).

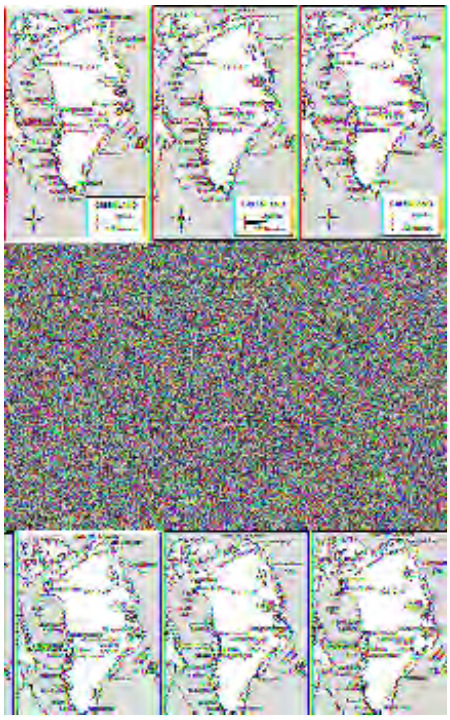


Now that we are in the Arctic, let's face a dilemma: Where to place Greenland in the grand scheme of the world's geography? Greenland is the largest island in the world, 2,176,000 square kilometers (the next largest, New Guinea, is only about 40 as large: the much larger Australia and Antarctica, while they look like huge islands, are arbitrarily designated "continents"). Greenland is considered part of North America and indeed it lies on the North American tectonic plate. Here is a map of Greenland in context with Canada (and Russia):

As seen from space by MODIS, Greenland is revealed to be almost entirely covered by an ice sheet or cap several kilometers thick:

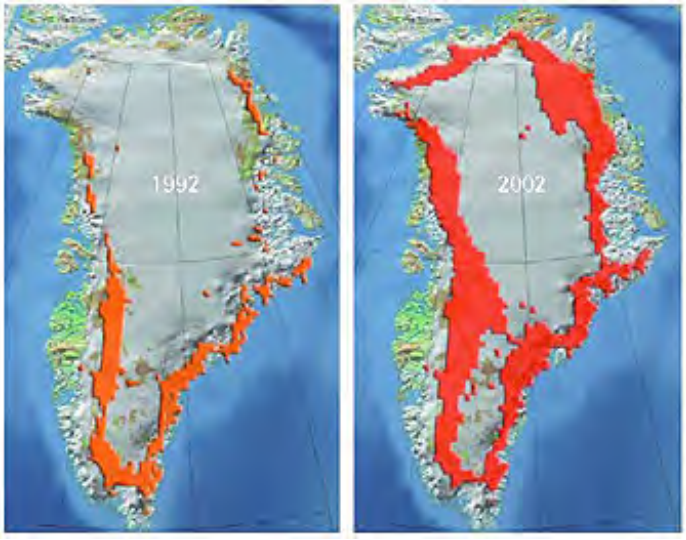


Actually, there are small towns and settlements along the ice-free fringes of Greenland, as shown in this map:



The total population of Greenland is about 73000. Here is a typical village, whose inhabitants live mainly by fishing:

The ice on Greenland contains about 6% of the freshwater on the Earth's land surface. But, as global warming proceeds, the amount of renewed ice has been diminishing as the cap also shrinks, exposing more land.



The ice-free parts of much of Greenland are mountainous. Here is the northernmost mountain

group (highest point nearly 2 kilometers above sealevel) on Earth, in a region of Greenland known as Peary Land (after the Arctic explorer, Richard Peary):

Now, we follow the birds in winter and journey southward into the Caribbean and thence to South America.

PREVIOUS NEXT

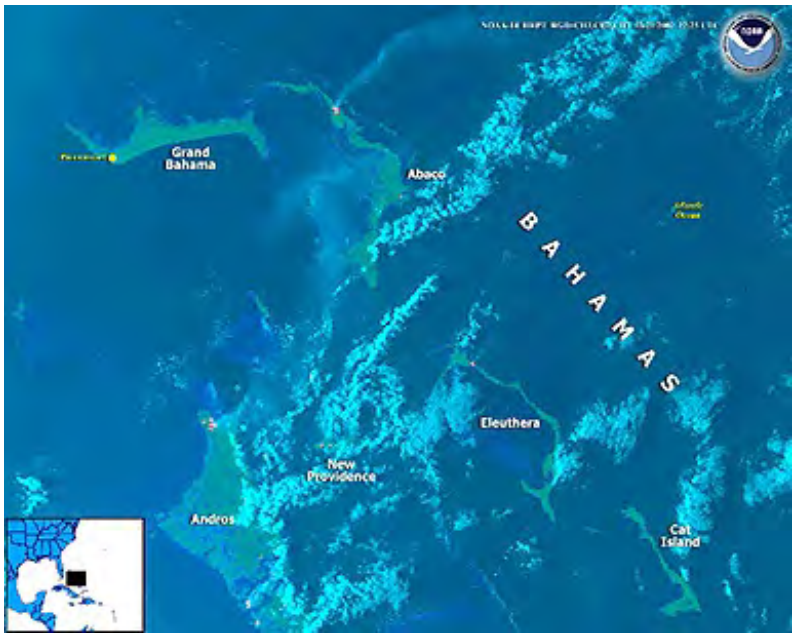
Primary Author: Nicholas M. Short, Sr.

This page is devoted to islands in the Caribbean and to the Latin American countries of Central and South America. Vegetation cover ranges from desert to forested mountains to subtropical lowlands.

PART 2: TRAVERSES OVER THE OTHER CONTINENTS

The Caribbean and Central America

The Caribbean is a broad geographic area that extends south and east of the Gulf of Mexico, and includes islands such as Cuba and Puerto Rico, a string of islands along a plate boundary called the Antilles (the West Indies reached by Columbus), and islands off the northern shores of South America. The Bahamas are generally included in this group. Here they can be seen from space in this NOAA weather satellite image:



The Bahamas lie off the east coast of Florida and are also near Cuba. These coral atolls are evident from the light blue-green waters (caused by reflection of sunlight off suspended sediments of limestone) that surround them, as seen in this astronaut view looking west over the Atlantic Ocean:

For a moment we will leave the Caribbean and look at another coral island in the Mid-Atlantic. Bermuda is a favorite destination of ocean liner cruises:

This next view from space (mosaic) shows the eastern and southern Caribbean. The map covers the entire Caribbean from Florida to Venezuela:

The largest of the Caribbean islands is the nation of Cuba, seen in this MODIS image:

The capital of Cuba, Havana, appears in this Landsat subscene

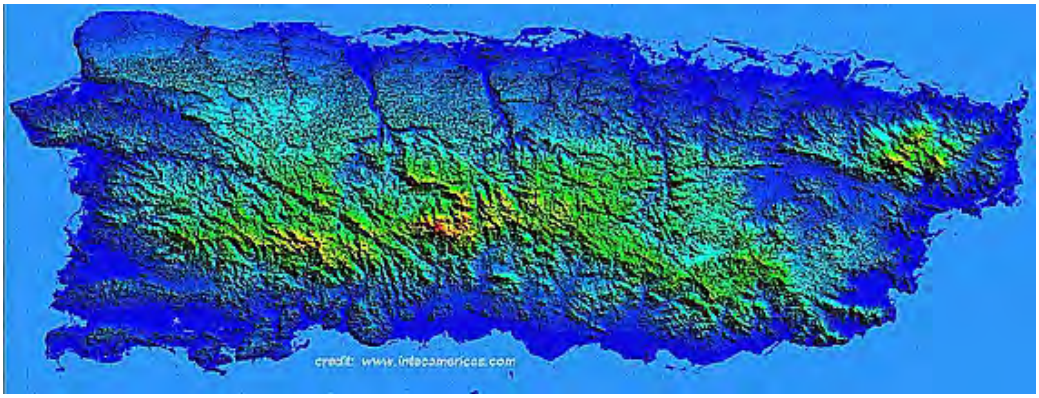
More details of this city, once a popular tourist destination for Americans in pre-Castro years,

appear in this SPOT image:

At the bottom of the first Cuban image is Jamaica. Here is a closer look contained in an astronaut photo:

Typical of the northern islands are Puerto Rico and the Virgin Islands. The first image is an STS astronaut-taken photo of Puerto Rico, which by treaty is allied to the United States as a Territory:

It is not obvious from this photo that the interior of Puerto Rico is mountainous. A DEM topographic image reveals the extent of these mountains:



Puerto Rico contains the only true tropical rain forest in territories tied to the U.S. This view is from the El Yunque National Park:

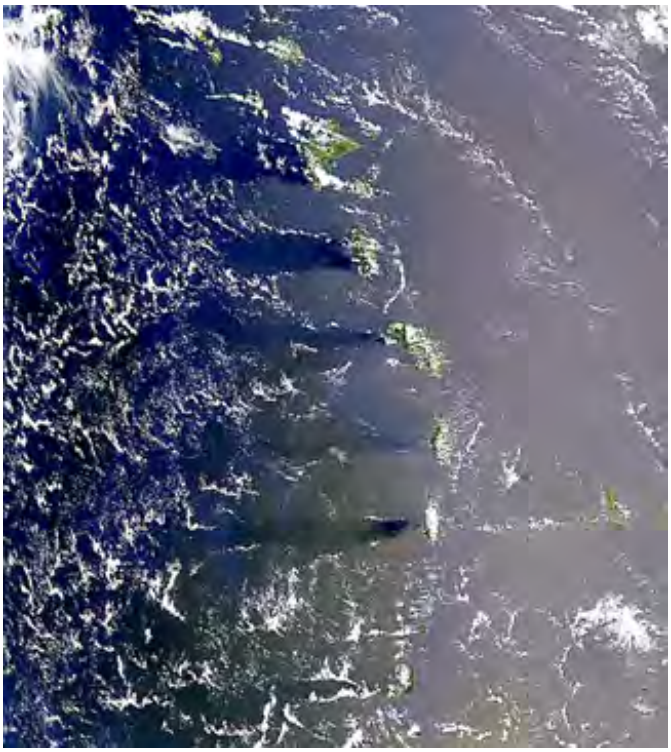
San Juan is one of the oldest cities in the western hemisphere (Columbus discovered Puerto Rico in 1493). Here is a view of the old city in the foreground, the new city further back, and the western mountains in the distance.

One of San Juan's landmarks is the El Morro fortress:

A short distance to the east of Puerto Rico are the Virgin Islands, seen in this astronaut photo:

The left (west) two islands are U.S. territories, being the islands of St. Thomas and St. John (to its east). The large island in the middle is the British Virgin Island of Tortola.

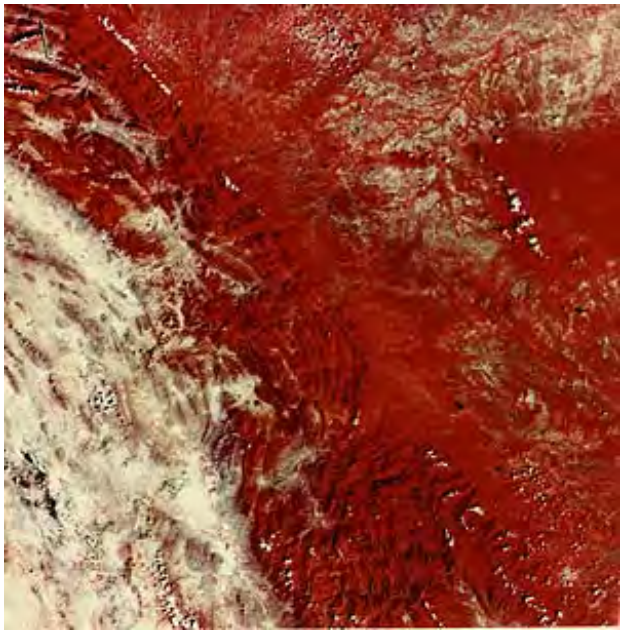
The Lesser Antilles is a long chain of island that run from the Virgin Islands south to Grenada, then east through Trinidad to the Dutch islands of Curacao and Aruba. The Outer Antilles face the Atlantic Ocean, as seen here in this MODIS view:



Typical of this eastern group, popular in winter with tourists looking for warm climes and dreamy beaches, in the Windward Islands is the volcanic island of Martinique; note the wind shadows:

The small island of Aruba, off the Venezuelan coast marks the end of the Antilles. This island, a semi-desert, has been photographed by the Shuttle astronauts:

We now head westward across the Gulf of Mexico and move south of the U.S. border into northeast Mexico, in the States of Nuevo Laredo and San Luis Potosi (left two-thirds of the image below) and Tamaulipas (right). The strongly folded sedimentary rocks of the Sierra Madre Orientale run through the center of the scene. Coastal plains make up the area to the east and semi-desert at high elevations occupies the land to the west. The reds denote regions that can experience 30 inches (75 cm) of rainfall each year owing to moisture moving inland from the Gulf of Mexico. This is a monsoonal climate, with wet, hot summers and dry cold winters. The plains supports typical subtropical savannah vegetation and the mountains are havens of both broadleaf and coniferous trees. The arid interior to the west is host to brushy plants and cactus.



We saw another part of the Sierra Madre Orientale on page 2-6. Here is a different Landsat image which shows both the mountains and the city of Monterey:

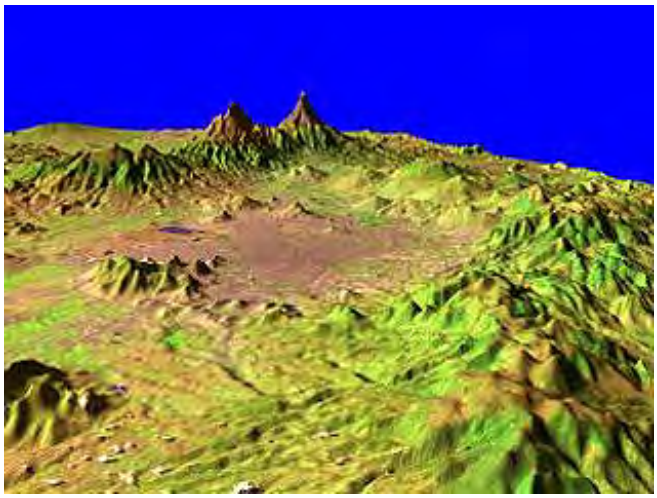
The Sierra Madre Oriental provide a spectacular backdrop for Monterey, as is evident in this photo (the architectural feature is the Puente de la Unidad, a suspension bridge):

On to central Mexico and a peak at the oldest city in the Western hemisphere, the present day Mexico City, built on a site where in the 12th century stood Tenochitlan, inhabited continuously since then. Mexico City now is seen as a blue area in the upper left part of the image. In 1973, M.C. had just over 7 million but has grown so rapidly that it will approach 30 million early in the 21st Century. Parts of the city stand on swampy ground and lake beds, particularly susceptible to failure and building collapse during the strong earthquakes that frequent the region. To its east are a line of active to dormant volcanoes, marking the surface expression of subduction of the Pacific plate under the North American plate.The two biggest volcanoes are Istaccihuatl and Popocatepetl, both snow-capped in this May scene, that lie to the southeast of M.C. Their slopes, and highlands elsewhere, are forested (reds) but most of the land is arid and sparsely vegetated. Most of the region shown is elevated - at 2800 m (7500 ft) at M.C. and higher.

The sprawl of Mexico City is evident in this satellite image:

Mexico City is now the largest city in the western hemisphere. It has grown from less than 10 million in the 1970s to more than 21 million in 2010. It also a very old city, having been started by the Aztecs in the 1300s. These two Landsat images show the spreading of the city in the last 40 years.

Combining Landsat and DEM data yields this perspective view of Mexico City. The two volcanoes appear to its south-southeast. The ring of hills around much of Mexico City indicates why it is often a smog-drenched city since winds may be too blocked to drive the gases away.



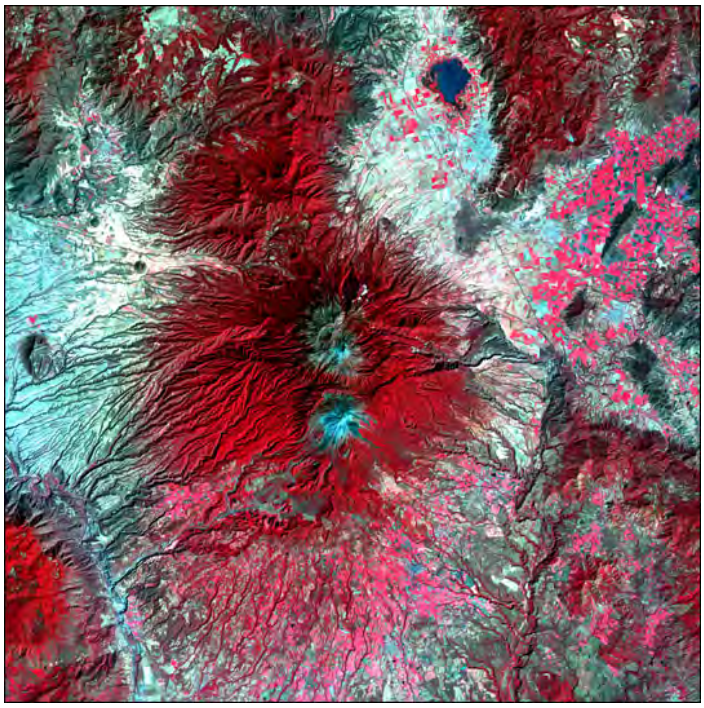
Here are two ground views of part of Mexico City. The first shows the central high rise area in the background with the church of Our Lady of Guadalupe in the foreground. The second shows part of the famed Avenue de la Reforma:

This aerial view shows the city park and the Chapultepec palace, with part of the central city.

Mexico City is situated at the edge of a volcanic belt. Here are the volcanic peaks of Iztaccihuatl and Popocatepetl in the background, with the southeastern edge of the city in the foreground.:

Popocatepetl is one of several dormant volcanoes in this part of Mexico. Here it is as seen by SPOT:

Popocatepetl is not the most active volcano in Mexico. That 'honor' belongs to Colima, seen in this ASTER image:

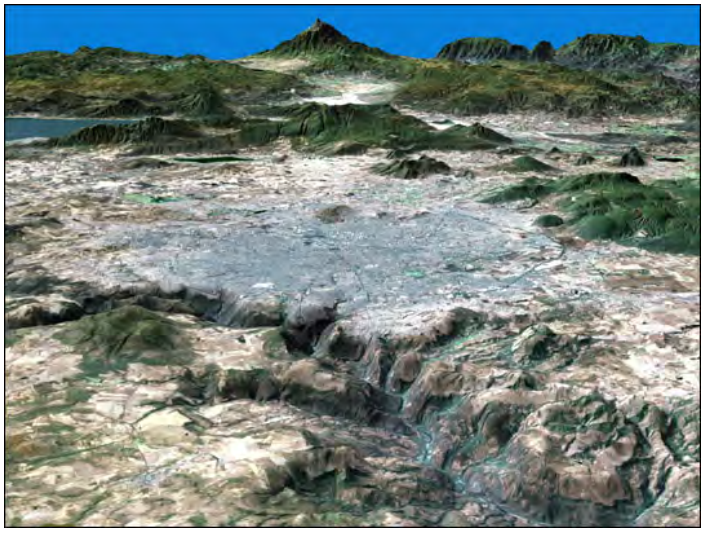


There is a fair number of volcanoes, many dormant and some extinct in Mexico. Consider this map:

The map reveals two different lines of volcanoes. The major one crosses the country in a trend that does not quite follow the present day subduction zone where the Cocos plate is diving below the North American plate. One model postulates a hot spot that has led to volcanoes of

progressively younger age to the southeast as the North American and Pacific plates interact. The second, to the south, is along an active zone.

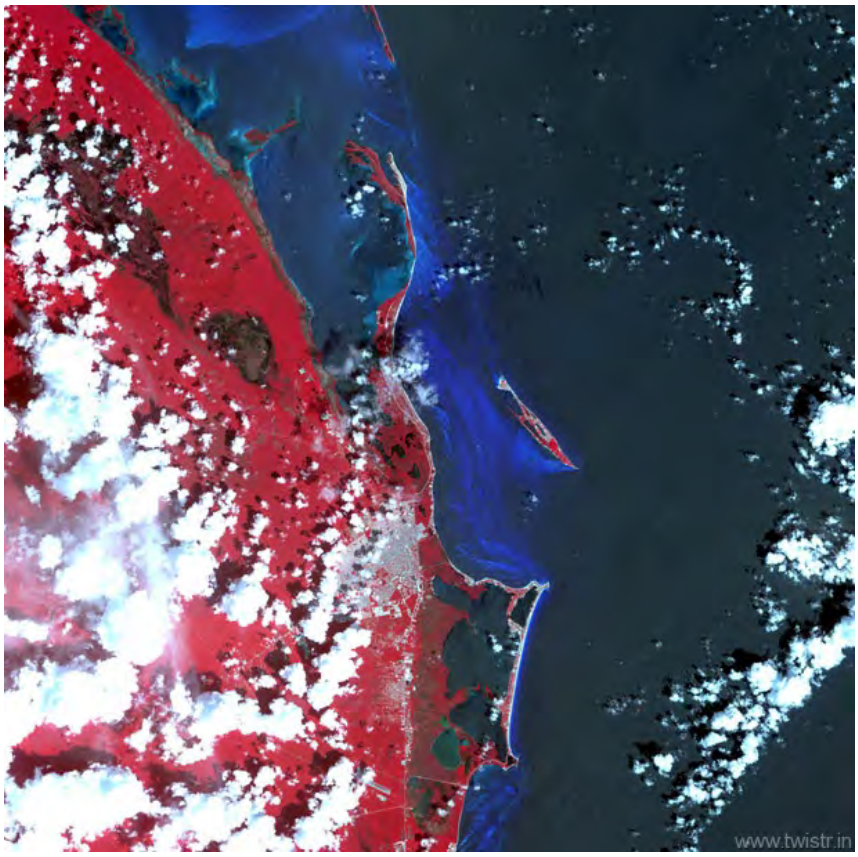
Mexico's second city is Guadalajara. Here is a satellite perspective of that city made from Landsat imagery:



A must-see when visiting Mexico are the ruins at Teotihuacan, northeast of Mexico City. This ancient metropolis was built by progenitors of the Aztecs, perhaps the Tototacs, about 100 BCE. Included were temples for worship of their gods. Several of these were pyramids, including the Pyramid of the Sun - third largest in the world. Here is an IKONOS image of that Pyramid, and a ground scene:

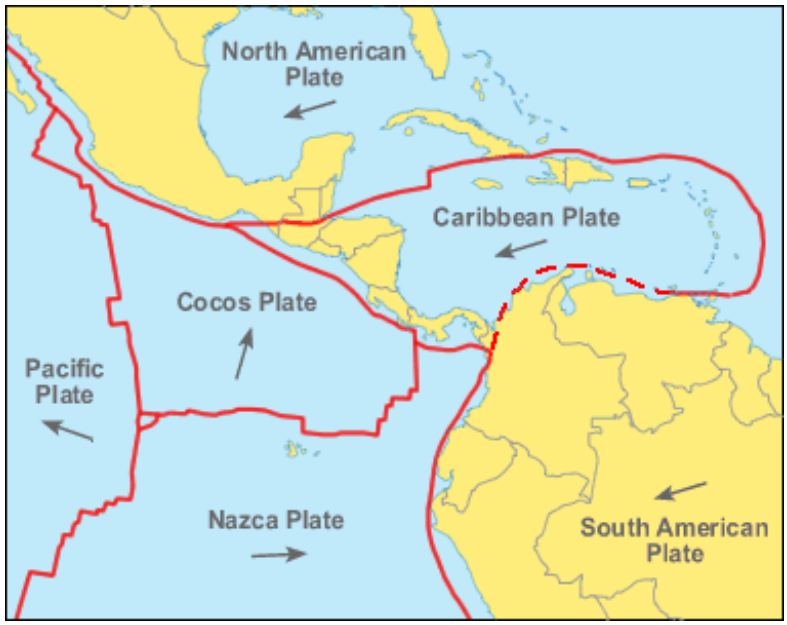
Mexico has become a popular destination for tourists, especially in the winter. Among ocean resorts are Acapulco on the west coast and Cancun and Cozumel on the Caribbean (in the Yucatan). This Landsat image of Cozumel reveals that it is an island and the main town is on the sheltered west side:

Cancun also is an island, as seen in this Landsat image:



We move south into Central America through Guatemala into Nicaragua. The countries of Central America are shown in this map:

Why is Central America where it is and is there in the first place? The answer is suggested by this geologic diagram:



Only recently geologically, in the last 4 million years, has Central America existed as a land bridge. After the North and South American plates split from Pangaea, they traveled westward separately as they moved against the Pacific plate. Nestled between them are smaller plates. The Cocos tectonic plate is being squeezed by the Pacific plate on the west and the Caribbean plate that is caught between the North and South American plates to the east. As a result volcanoes have developed in the sedimentary rock units shoved up above a subduction zone - this has resulted in the land bridge that joins North and South America, namely Central America.

The Volcano Arenal, in Costa Rica, is typical of the stratocones that extend along most of Central America:

From a space perspective, we look first at Guatemala, shown in a MODIS natural color image. The west coast is relatively arid; note the line of volcanoes indicated by snow capping. The interior is a tectonic belt and is higher, accounting for the vegetative cover that denotes a jungle.

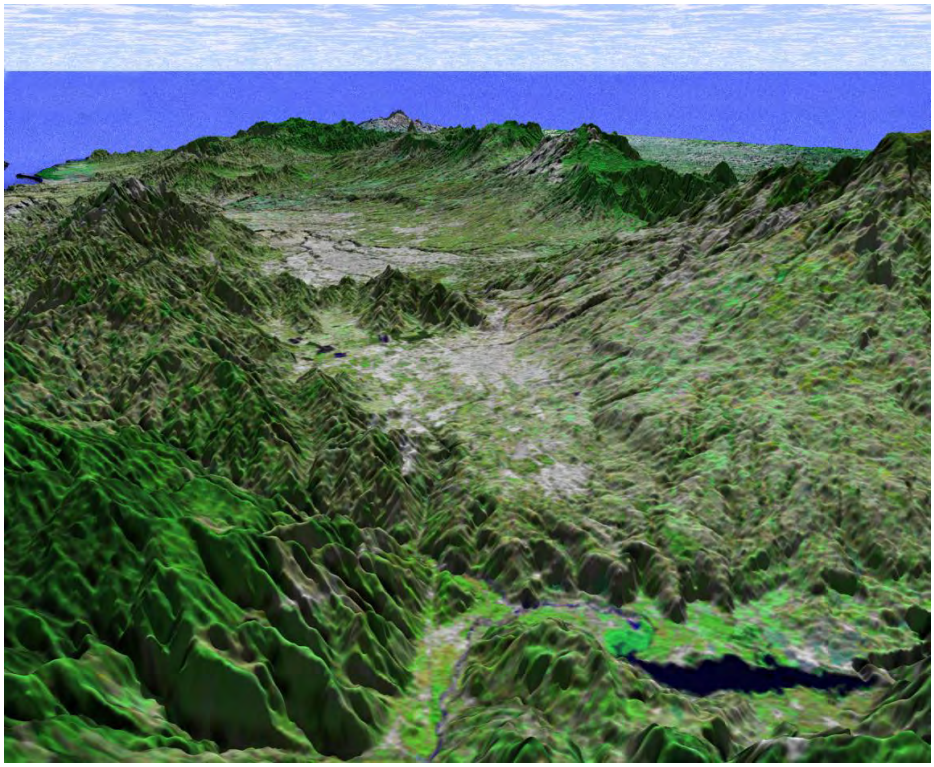
Nicaragua's western region, against the Pacific Ocean, appears below. The larger body of water inland is Lake Nicaragua, with several volcanoes on islands within. In the upper left is Lake Managua, with the capital city, Managua, along its south shore. Circular lakes are fillings of central vents or calderas. The lower elevations consist of semi-arid vegetation, with some farmlands. Areas of red represent uplands with forest cover, with more mountainous highlands at the upper right corner.

This part of Central America, which includes Guatemala, Honduras, and El Salvador, is especially mountaineous, with its ranges marking the upthrown crustal blocks at a plate boundary. The relief map made from SRTM (radar) data bring out the rugged topography (note the fault lines that are etched out from the terrain):

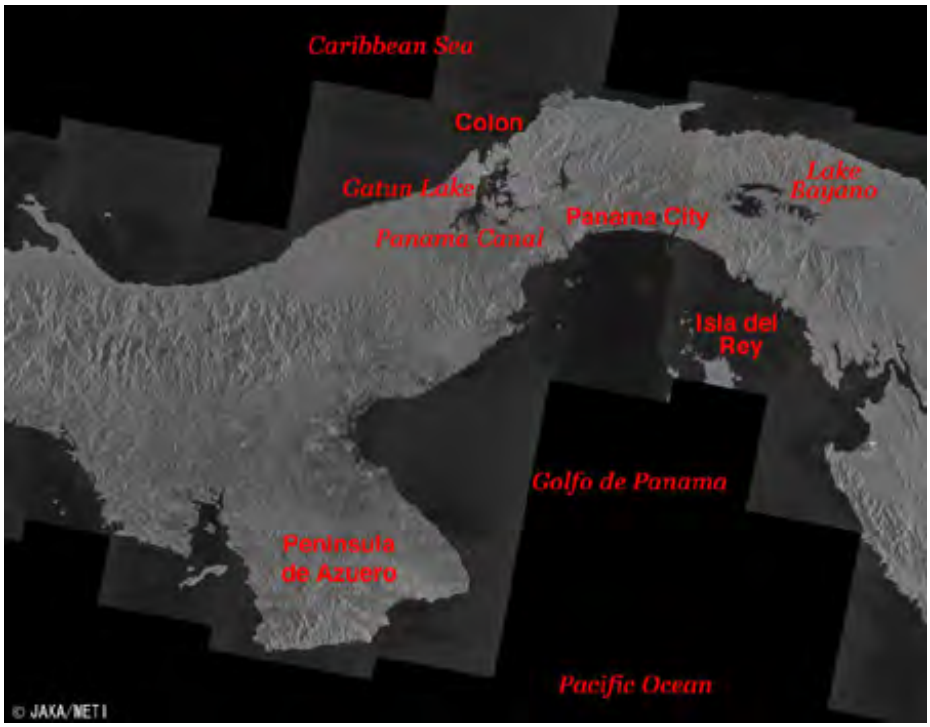
Costa Rica has, in recent years, become popular as a tourist attraction and for some Americans a splendid retirement place. It is largely forested as is evident in this satellite view:

It is also mountainous, as evident in this ground scene

The capital of Costa Rica is San Jose, shown next in a satellite image looking straight down and then in a perspective view made by combining SRTM topographic data with a Landsat image:

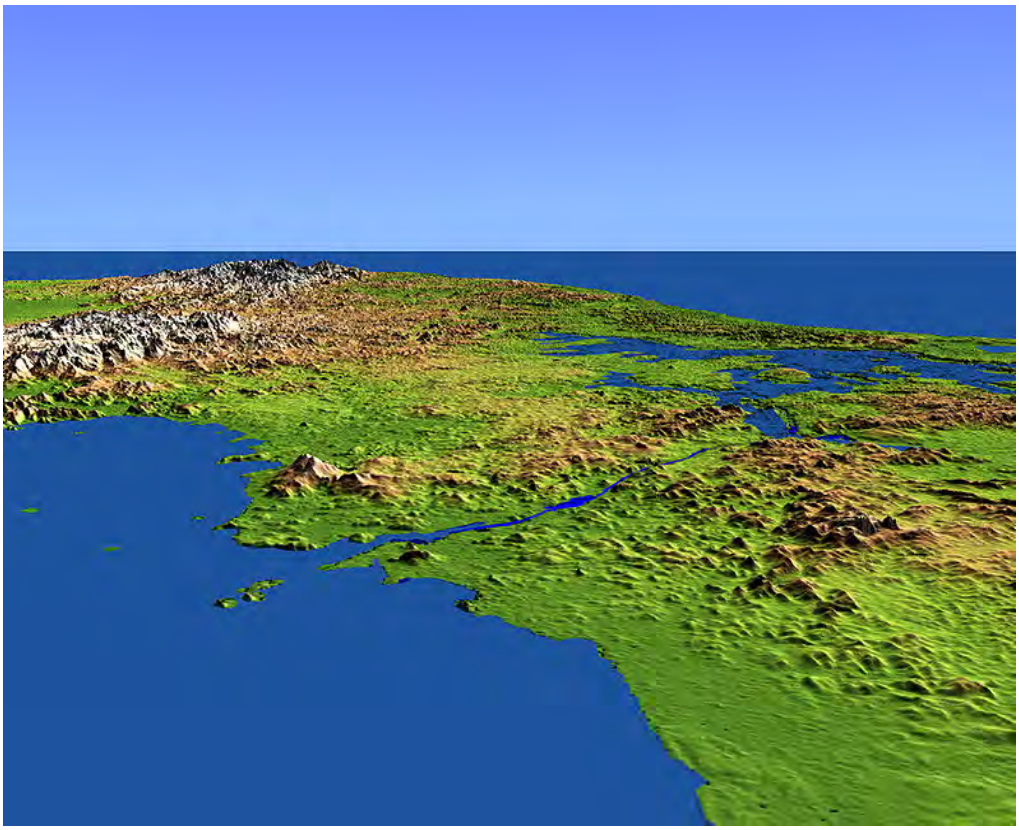


As we approach the South American continent, we move from Costa Rica into Panama, shown first in another relief map created from SRTM data and then using a SAR radar image made by the Japanese Fuyo-1 satellite:

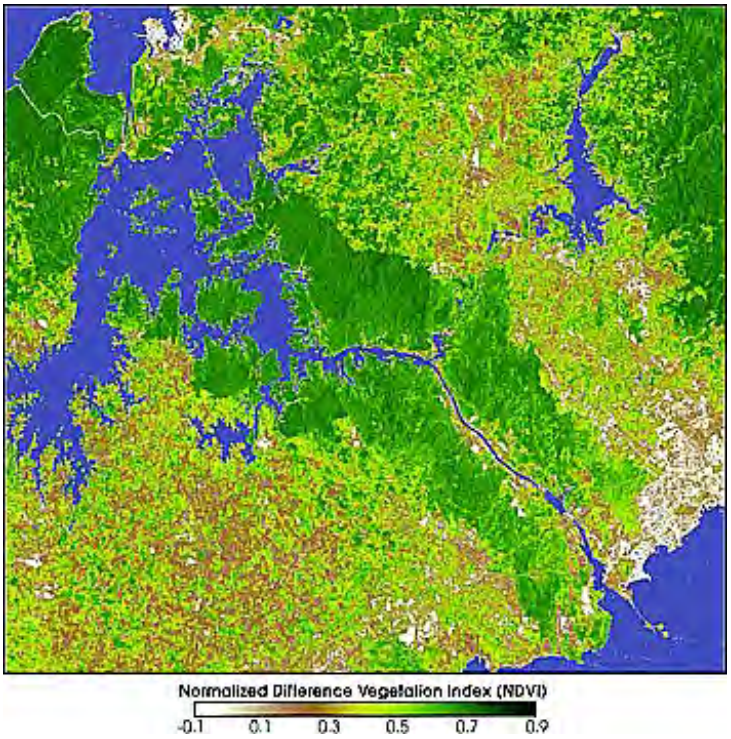


In the scene below, a Landsat-DEM perspective image we see a perspective view of the "Big Ditch" that was one of President Teddy Roosevelt's greatest accomplishments - the Panama Canal, which rivals the Suez Canal (see page 6-13) as the greatest aid in oceanic navigation. A French company, which had dug the Suez Canal, made an attempt in the 1880s to excavate a Panama Canal, but had to desist because of insufficient funding. The present canal was started in 1903 and finished in 1914. Approximately 83 km (52 miles) of canal was actually dug, but much of the passageway consists of natural lakes.

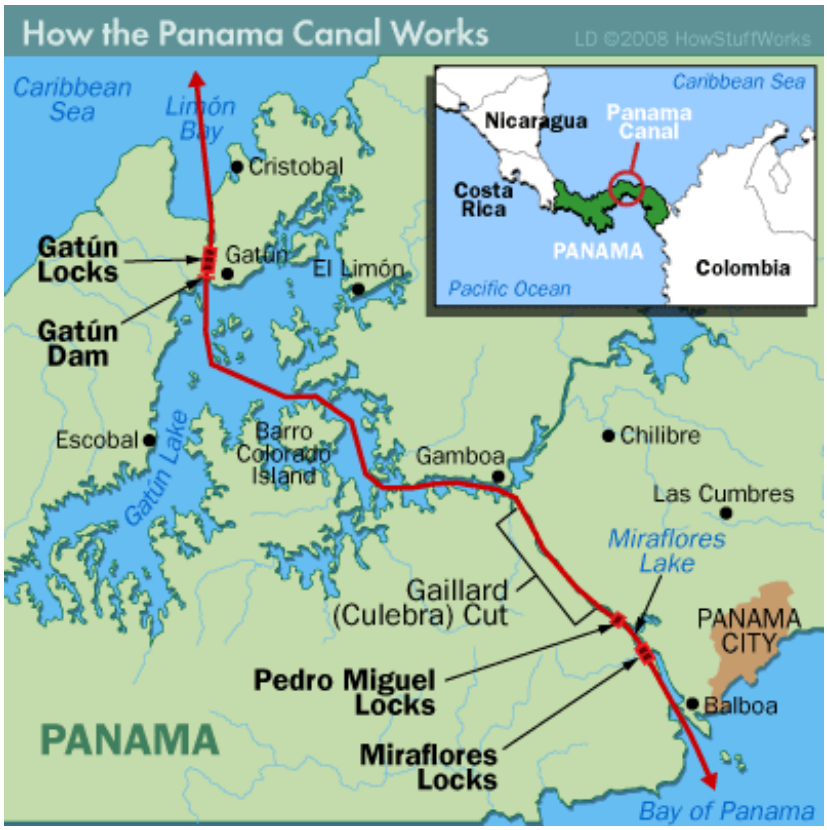
Just as the Suez Canal obviates the need to sail around Africa for Europeans to get to Asia, so does the Panama Canal allow ships from the east to get to Asia without having to sail around South America. The canal is shown in this perspective, made from Landsat and SRTM inputs.



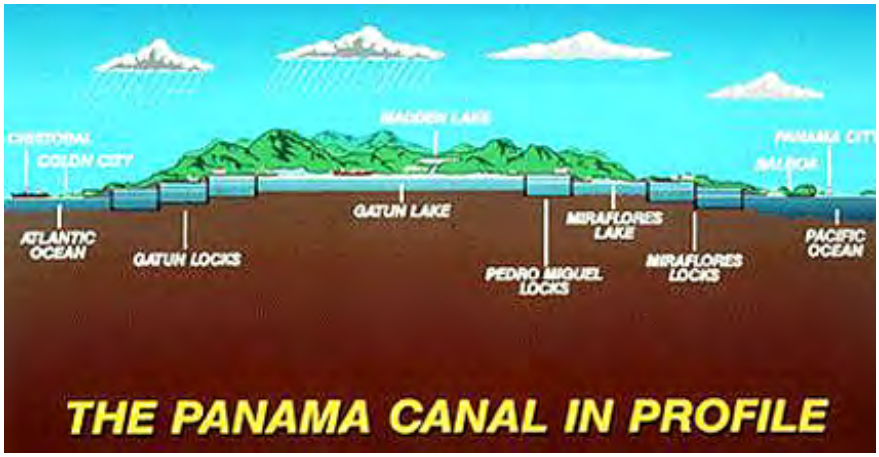
In this next image, Landsat data are displayed in colors that represent the calculated NDVI (a measure of vegetation density, see Section 3) for the area that include the Canal and Lake Gatun. Most of the land on either side has a strong NDVI rating, expected since this is heavily forested tropical "Jungle".

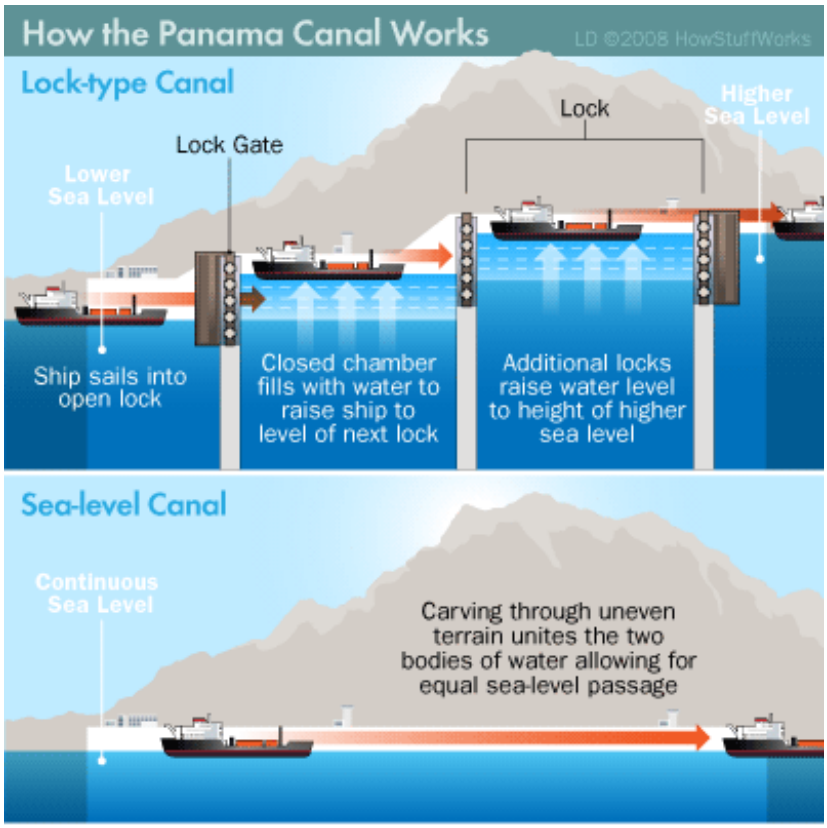


Here is a map showing the main landmarks along the Panama Canal.

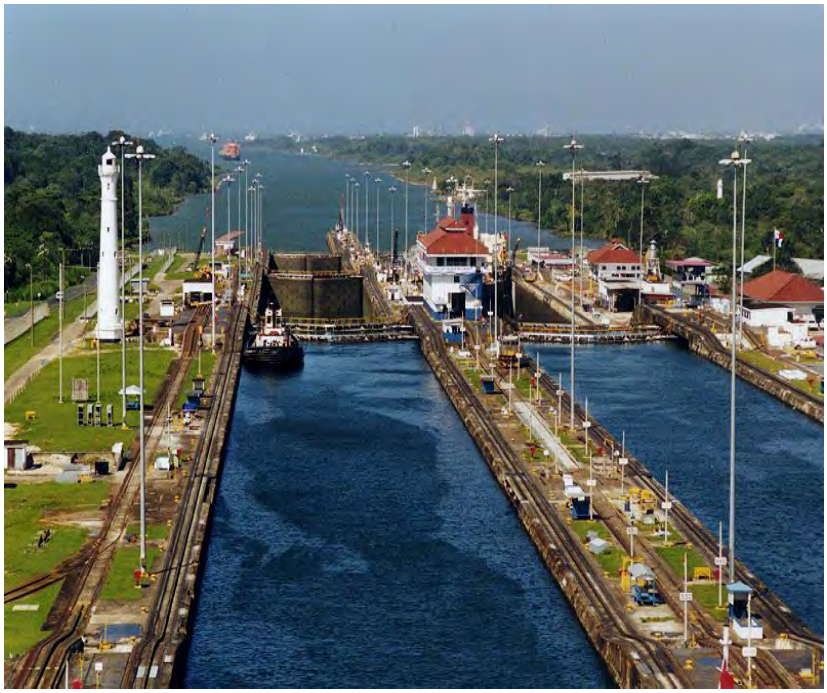


Unlike the earlier Suez Canal, the Panama Canal requires a series of locks to acommodate the difference in sea level heights between the Atlantic (Gulf of Mexico) Ocean and the Pacific Ocean (the Pacific is about 5 meters higher). If going westward, a boat enters a lock, behind it another lock gate swings shut, and water is pumped in to raise the boat level to that of the next lock entered. Here are two diagrams of this lock system:

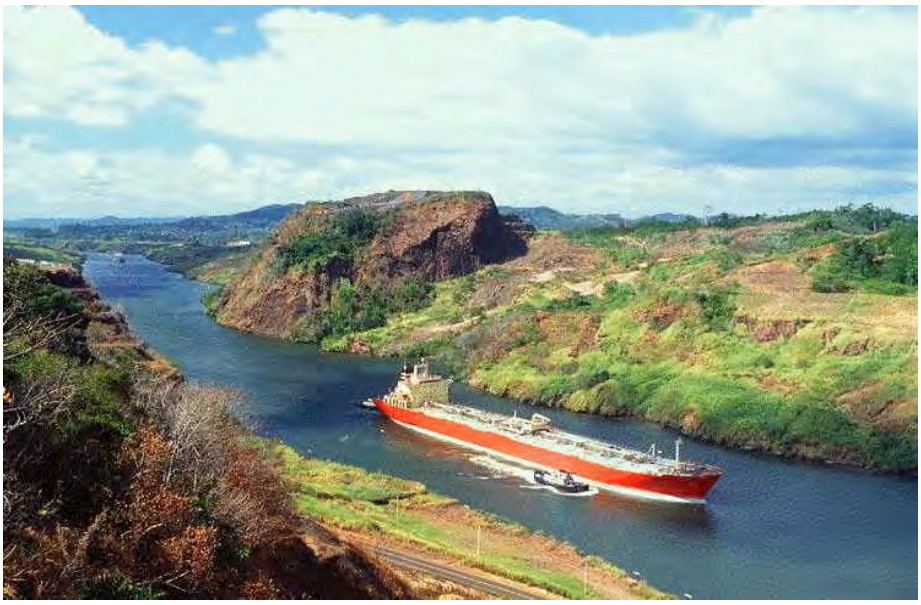




Here is an aerial view of the Gatun locks:



Most of the canal is without obvious artificial sides; the canal where dug appears almost like a stream bank:



The western end of the canal system is a natural embayment leading to the Gulf of Panama; the bridge is part of the PanAm highway system that starts in Alaska, proceeds through the western U.S., down all of Central America, and then along western South America into Chile:

Nearby, on the Gulf and the Pacific Ocean, is Panama City. It is now surprisingly modern, with many skyscrapers (the writer [NMS] has been astounded at how many large cities outside the U.S. have such high rise buildings; these will be evident throughout the remainder of this Section.

South America

Our travels next take us through representative parts of South America. This map shows the countries making up this continent and the major geographic land units.

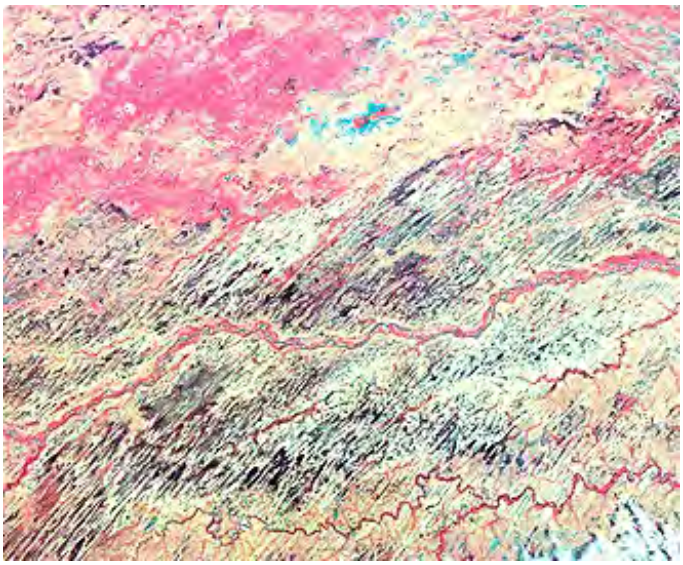
Staying in the Caribbean but now touching the northern coast of South America, this next scene is an Envisat-MERIS image of parts of Venezuela and Columbia to its west:

The greenish water off the Caribbean is the Gulf of Venezuela. A narrow strait passes south into Lake Maracaibo, within and around which ar most of the oil fields that make Venezuela the fifth largest petroleum producer in the world. On its west side is the Sierra de Perija whose crest forms the border with Columbia. Along the eastern shore of Maracaibo is the Cordillera de Merida. Both mountain systems are the northern extension of the Andes. The west mouth of the Gulf of Venezuela is formed by the Guajira Peninsula. An island like body on the east side is

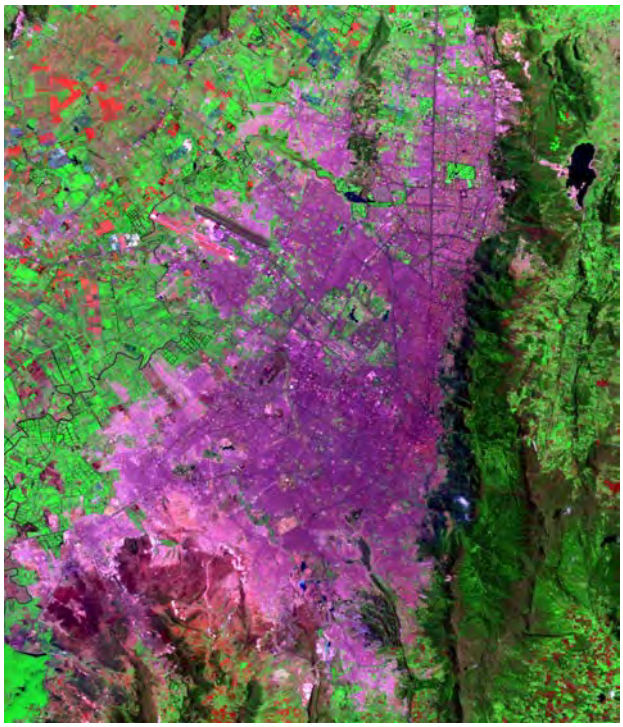
the Paraguana Peninsula, joined to the mainland by such a narrow strip of land as to be almost invisible in this image. Just above the peninsula in the center is Aruba, shown above. The island of Curacao (famed for its liqueur of that name, made from sour orange peels, and as a stop for Caribbean cruise ships) lies to its east. Caracas, the Venezuelan capital, is just off the image on the upper right side.

Caracas, Venezuela's capital, is nestled between mountains just inland from the Caribbean coast, as shown in this Landsat-7 image:

While in Venezuela, let's look at a remarkable scence in its southern Orinoco Basin. The Orinoco River in this Landsat-1 subscene cuts across a low terrain that is an extension of the Guyana Highlands to the northeast. Look carefully at the image to see a pronouced zig-zag pattern (emphasized by black lakes). The writer (NMS) has not found a reference to this feature; he interprets the pattern to be tight chevron folds in the crystalline rocks that outcrop also in the Highlands.



West of Venezuela are the countries of Columbia and Ecuador. The capital of Columbia, Bogota, is shown in this Geology.com Landsat rendition:



The Andes mountains have their narrowest extent in these countries. Here is a composite image of Ecuador:

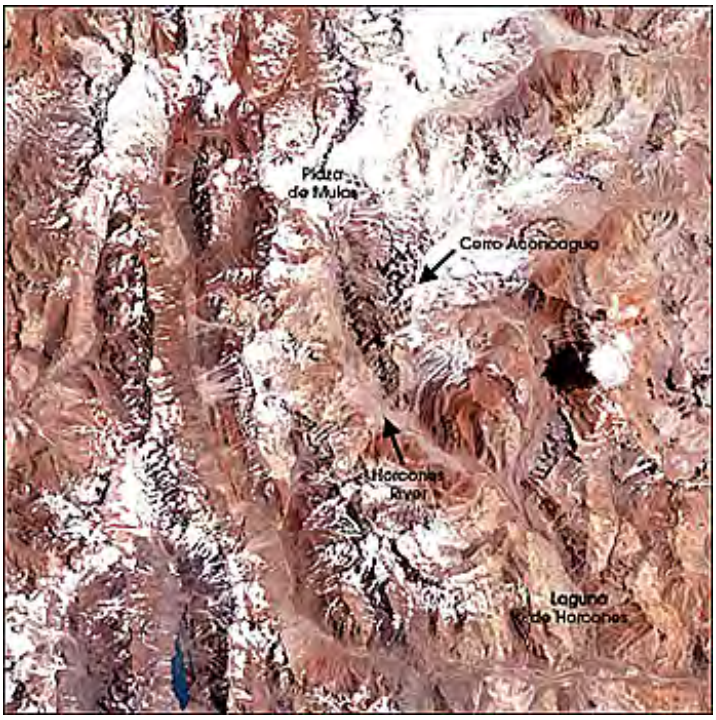
In a loose sense South America extends some 500 miles (800 km) into the Pacific Ocean (in reality, beyond the western edge of this continental plate). Ecuador "owns" the famous Galapagos Islands, now a popular tourist attraction. Look first at this image from space in which the islands are labeled and then at an astronaut photo from the International Space Station of part of the Galapagos - volcanic islands all - covering most of Isla Isabela and Isla Fernandino:

The dominant landform features in western South America are the mighty Andes Mountains. These are part of the general cordillera that lines the western regions in both Americas and Central America. The Andes are the longest continuous mountain chain on Earth; they are less than 40 million years old. The Andes are uplifted mountains, involving intense folding, faulting, and volcanism on the continental tectonic plate as it is subducted by the east moving, incoming Pacific tectonic plate. JPL has prepared a simulated journey along the Andes as a movie using radar data. Access it through the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes "Andes Mountains Flyover", June 20, 2003. To start it, once found, click on the blue RealVideo link.

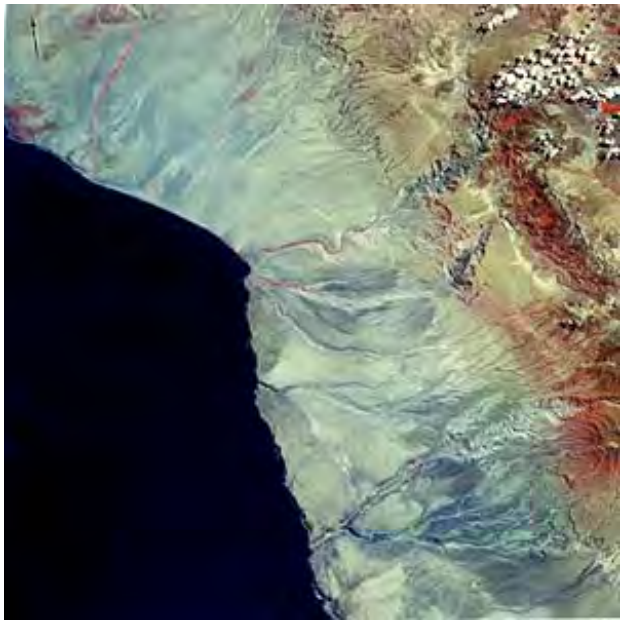
As we saw above, the Andes tend to be relatively narrow over much of their length. This is evident in this astronaut photo looking eastward across the vegetated coastal plains, then the high Andes and the moderate elevation Pampas of Argentina, which in this wintertime scene is brown because of dry grasslands:

In this Landsat-1 image in southern Peru, a dry desert appears blue. The land rises abruptly eastward in dissected mountains whose elevations range to altitudes from 4300 to 5500 meters (14000 to 18000 ft; the highest peaks in the Andes approach 23000 ft). Progressing to the northeast, the terrain is first dissected, then gives way to a broad, flatter Altiplano, and ends (upper right) in the High Andes.

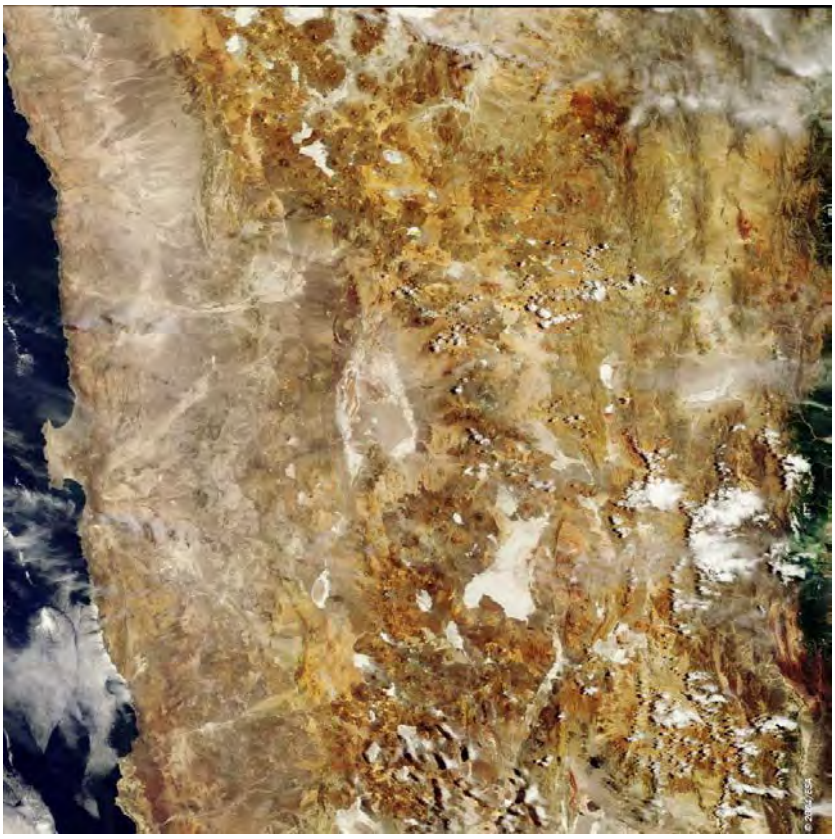
The highest peak in the South American Andes is Cerro (mountain) Aconcagua (6982 m or 22840 ft) in Argentina, shown in this Landsat-5 image:



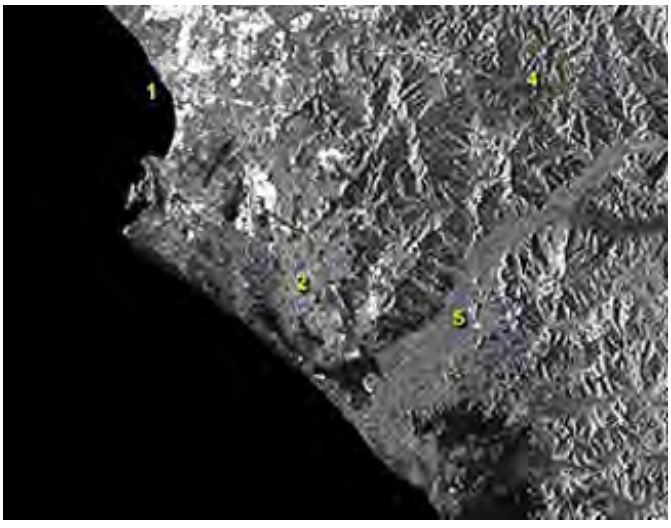
The next view covers a very sparsely populated segment of southern Peru (top) and northern Chile. The bluish-gray stretch of lands from the coast inward in part of the Atacama Desert, notorious as one of the driest regions on Earth. The desert results from the "drying out" of moisture in air masses crossing the Andes, which leds to rains and heavy snows. Some places in this desert, which continues well to the south, have seen as little as 1 inch of rain in five years. But, note that a few river valleys have ribbons of red indicating some vegetation (in oases where small villages can subsist), fed by occasional water coming from the better water uplands of the volcanic Cordillera Occidentale that comprises the western extent of the Andes. Note the landforms against the Andes flanks which are yellowish-brown - these are huge, coalesced alluvial fans now being dissected.



The bluish tones in the Atacama Desert are not true color. The dominant colors are actually browns to reds. We show a more realistic rendition in this ESA MER image; a ground photo also is typical of this landscape:

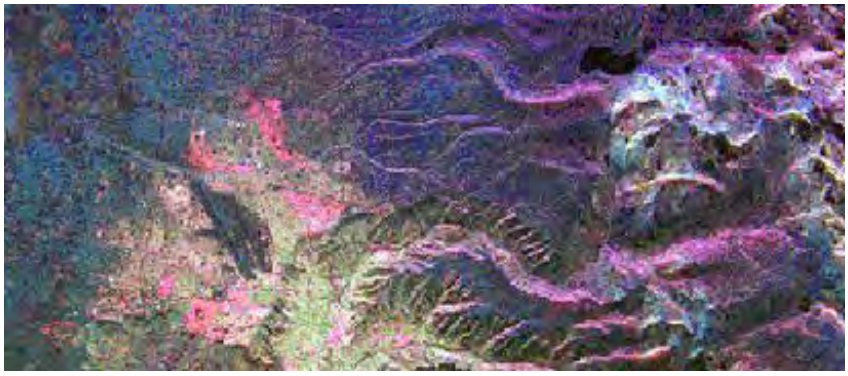


The capital of Peru is Lima, set along the Pacific coastline, as seen in this Radarsat image:

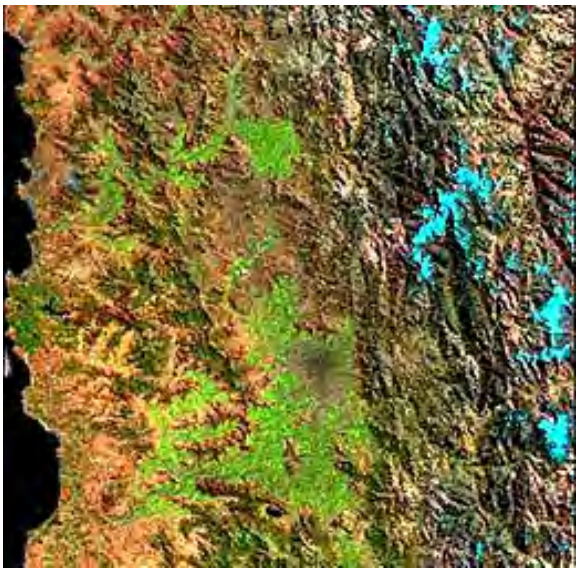


Between Peru and Chile is Bolivia. Its capital, La Paz, is inland, surrounded by low mountains. Here is a ground scene:

The next two scenes of La Paz are from space, the first a SIR-C radar image, the second a Landsat image. Note that each one contains a long, straight strip. This is the main runway for the airport (named Kennedy, because the United States provided significant funding to build the facility during the Kennedy administration). This lies amidst newer, more suburban parts of La Paz - the older sections are located within the valleys framed by the Andes foothills.



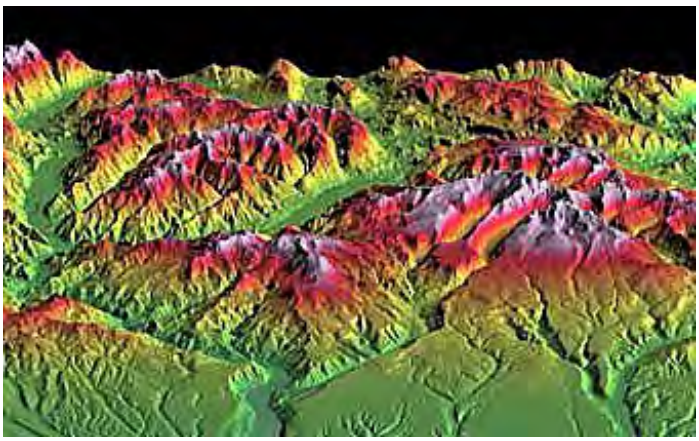
The desert is replaced by vegetation (some cultivated) in the lowlands extending inward from the Pacific. In this Landsat-7 ETM+ image, Chile's capital of Santiago appears as a dark-gray area in the green valley surrounded by coastal ranges to the west and the Andes to the east.



A ground view of Santiago shows the high Andes in the background:

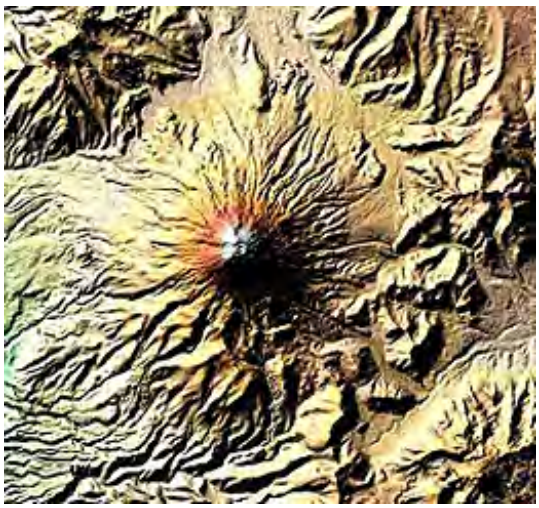
Moving across the Andes, Western Argentina occupies the scene below. Along the left margin is the eastern terminus of the High Andes, with ridges above 4800 meters (16000 ft), and surface with few extended forests but with some brushy vegetation. A large alluvial fan, in blue, appears near the upper left; at its eastern (right) margin is a conspicuous area of red-rendered vegetation which marks a zone where subsurface and surface waters from snow melt in the Andes passes onto the high plains. Lake Ilancanedo is seen to its southeast, a body of water that varies considerably with the seasons (in March, for this scene, the southern Fall is drier and the lake has shrunk). The landscape at this time of year shows minimum active plant growth. The volcano Cerro Nevado, with its snow cover giving the clue that it is high (3700 m; 12000 ft), seems isolated from the Andes. The caldera topping Cerro Payun (near bottom center) lies east of a broad field of basaltic volcanism.

Using a combination of a Landsat image and SRTM elevation data, a perspective of the Andes in Argentina gives an impression of their great heights:

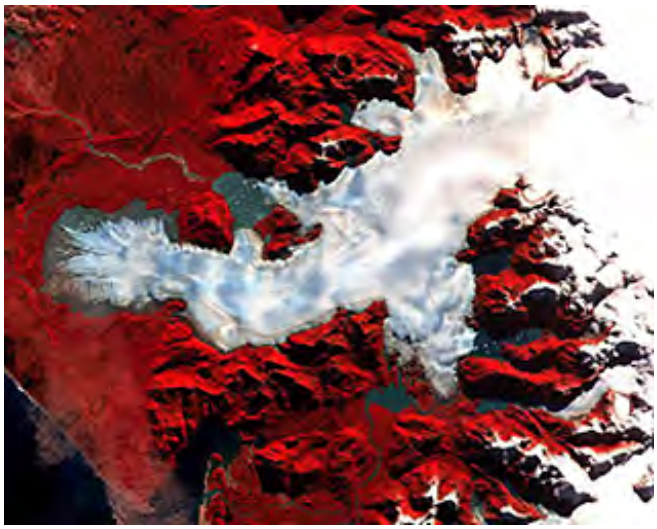


The Andes are a relatively young mountain system. The Chilean segment of the Andes began its major upheaval about 15 million years ago. The eastward subduction of the Pacific plate leads to extensive melting that produces numerous volcanoes (as reported on page 17-3). These lie along the Pacific Ring of Fire. This SIR-C image shows several in northern Ecuador, including the trio Cuan, Mojande, and Imabua, and to their south, Cayambe.

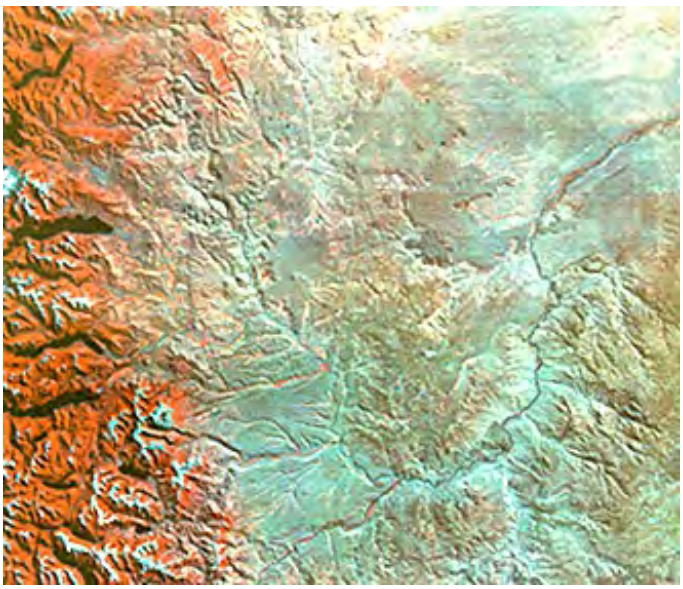
One of the most magnificent volcanoes in South America is Cotopaxi, in Ecuador, not far from the capital, Quito. We look at from the ground first and then in a radar image obtained during the SRTM Shuttle mission which used radar data to reconstruct topography:



The Andes mountains support numerous mountain glaciers. This ASTER scene shows the San Quintin glacier and ice field on the east side of the Andes.



In Argentina, the Andes grade eastward into the 800000 square mile expanse of plateaus and plains of the Patagonia. Here is a Landsat-1 subscene that shows the gradation from the High Andes into Patagonia:



North of Patagonia the land is lower, forming a vast plains known as the Pampas. Here is one of the great cattle ranching regions of the world. This space image shows a typical landscape in winter:

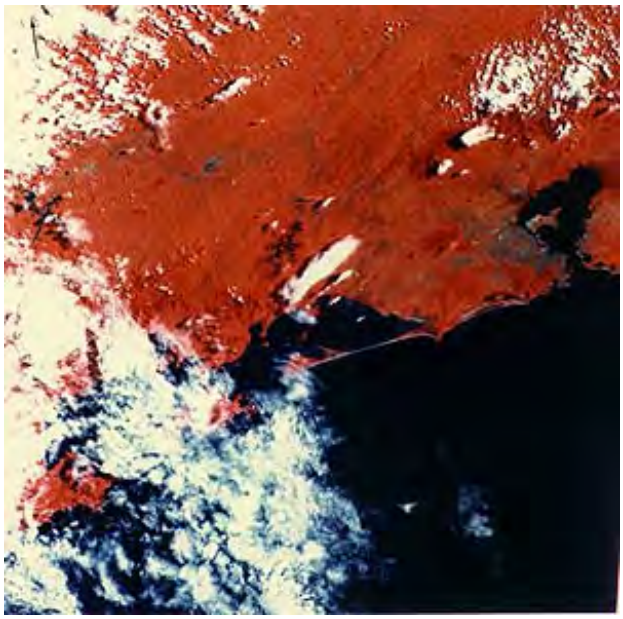
Buenos Aires, Argentina's capital, has been called the "Paris of South America" because of its bright, clean buildings (despite a Shanty Town within its limits) and planned layout. This city of 11 million, at the edge of the Pampas, lies on the Rio Plata's estuary; to its north is Uruguay. Let's first look at its skyline:

Here is a 15 m Landsat-7 ETM image of the region, including the north end of the Pampas (plains land noted for its favored grazing of beef cattle) and the delta and estuary of the Rio Plata. Below it is an astronaut photo (Mission STS-56) that shows some of the details within the city.

Moving north into Brazil, the second largest city is Sao Paulo - more than 10 million citizens. Here is an overhead view of this coastal area:

Like other South American cities, in recent years the central urban area is replete with skyscrapers:

Rio de Janeiro, the Queen city of Brazil, spreads out from the western shore of Guanabara Bay, along the coast (the city, in black, lies above center right) as seen in this Landsat-2 image. The low Serra do Mar passes through this area. The Serra de Orgaos, up to 1000 m (3000 ft) lies to the north. Similar low mountain terrain extends to the west. An evergreen rain forest lines the coast but mixes with semideciduous and mountain vegation further inland. Note the narrow strip of land (near image center) made up of marine deposits that encloses the Baia de Sepetiba.



This Landsat-7 ETM+ image shows much more detail within Rio de Janeiro:

This astronaut photo from STS61 Shuttle mission shows much the same area but with several landmarks labeled on the scene:

This city is the jewel of South America (although many who visit Buenos Aires would quibble with this statement).

This ground photo shows only part of the area of high rises set against the background of the Bahia Guanabara and the famed Sugarloaf mountain. Nearby, on another hill, is the hallmark of Rio, the statue of the Christ, with outstretched arms:

We have shown the inland pampas of Brazil as seen at Brasilia, the capital of Brazil built from

"scratch" in the 1950s, on page 4-4. We repeat its coverage, using an astronaut photo from mission STS-38.:

A more detailed view of the central city is evident in this ALI (EO-1) view which includes the artificial lake and the Brasilia National Park (a savannah):

The government buildings in Brazilia are laid out in a broad expanse of open space:

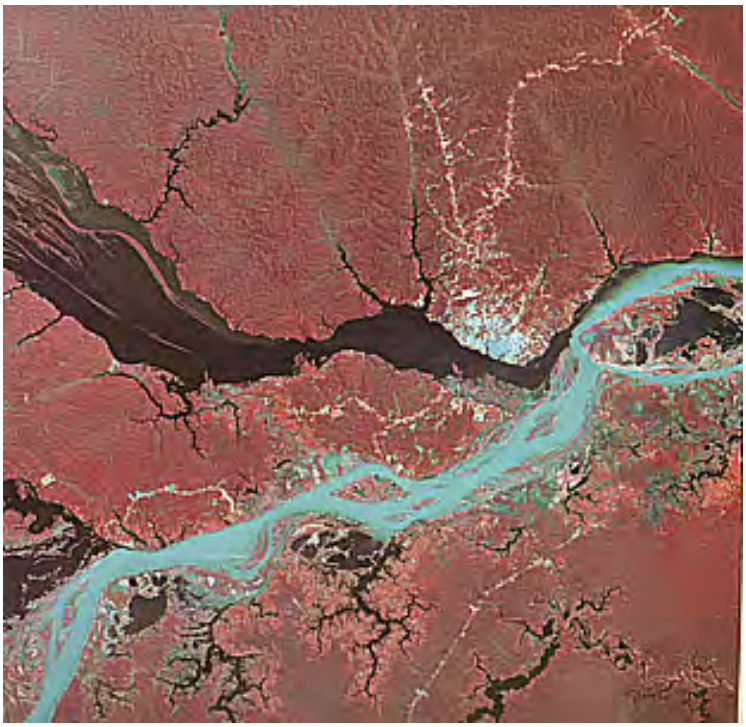
Brazilia lies near the edge of the great drainage basin controlled by the Amazon River. Much of the basin is lowlands and is heavily forested but many of its rivers start on the eastern flank of the Andes and other areas of higher elevation. This map zeroes in on the main drainage elements of the Amazon Basin.

If one moves further north into the tropical forest of the Amazon basin, the denseness of the jungle where not clearcut is truly amazing (see page 3-5). Here is another aerial oblique photo of the Amazon as it flows through the eastern lowlands. Beneath it is an example of what usually comes first to one's mind when the Amazon jungle is mentioned.

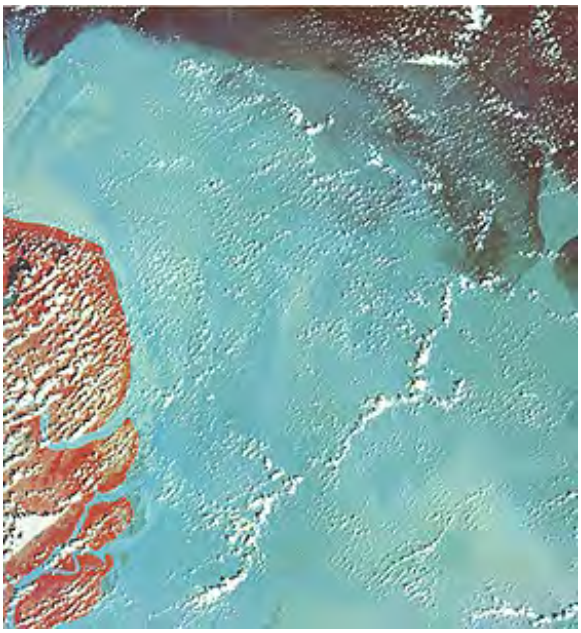
Considering the vast density of jungle vegetation in the Amazon, it seems astounding that a major city of 1.2 million people has been carved out of the wilderness at the junction of the Amazon and Rio Negro Rivers. This is Manaus, shown in an astronaut photo.

The modernity of this expanding urban area is suggested in this aerial oblique view:

The sharp difference in the color of each river is brought out in this false color Landsat subscene; the blue marks strong reflectance in that spectral range by the great load of sediment in the ' Amazon:



The Amazon is one of the two longest rivers on Earth. It carries a tremendous load of sediment - largely contributed from the higher elevations in the basin. Three million tons of particulates are carried into the Atlantic Ocean each day. The mouth of the Amazon is some 380 km (207 miles) wide; note in the map of the basin above that the drainage area has actually constricted to a narrow zone on the east. Below are a Terra MISR image of the mouth region, with the sediment in brown, and a Landsat-1 image (fit it into the upper image) that shows just a part of the vast spread of sediment through the mouth. Much of the land around the mouth is actually part of the delta being constructed by the Amazon.

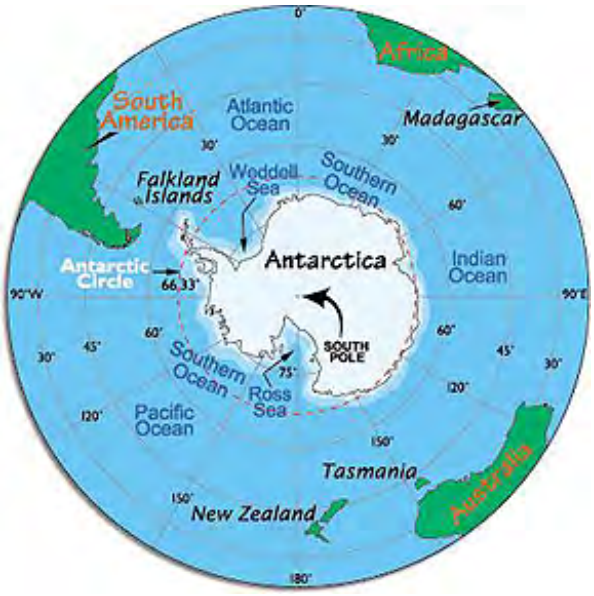


Before we leave the Amazon Basin and its rainforest, let's call attention to a sylvan variant, the cloudforest. This ecological biome refers to trees that form dense woods at elevations well above sealevel. The eastern tropical slopes of Ecuador support a cloudforest that transition into the rainforest. Rainfall is moderate to high but less than in the Basin interior. The cloudforest is the green area on the right in this Quickbird image:

Let's end our tour of South America by going to its southernmost point. Much of Chile and Argentina to the east that includes the Andes is called Patagonia. Here is an Envisat view of this region that makes up the southern fifth of South America:

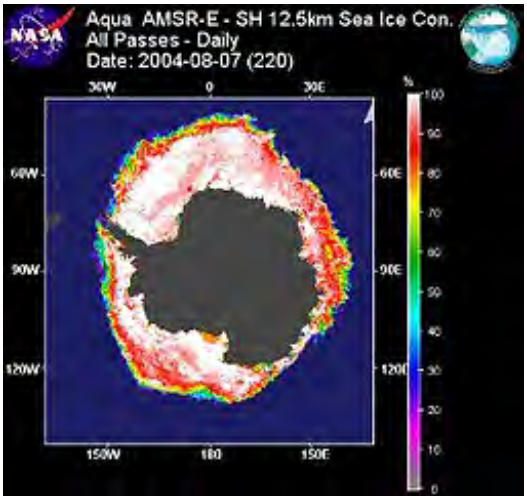
The Andes continue to the tip of South America, a rugged region known as Tierra del Fuego. A Terra image of this region appears below. The scene is unusual in that it is mostly cloudfree; the passage of ocean cruise liners through the Straits of Magellan or around Cape Horn is always an adventure since the weather is capricious and the sea can be ferocious.

As you saw at the bottom of page 6-10, when we bundled Greenland with Canada, we again face an apparent *non sequitur* as to where to place the continent of the Antarctic during this world tour. This map illustrates the "problem".



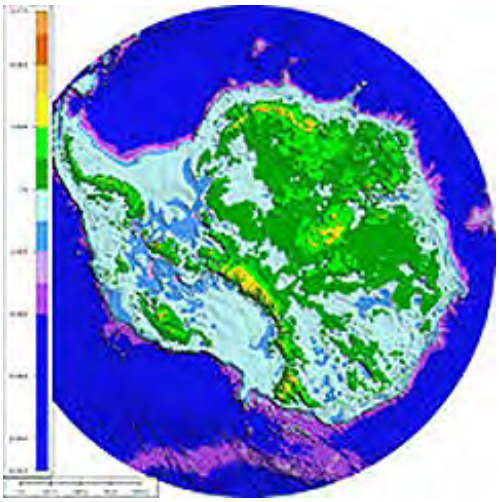
Parts of Anarctica are closest to the Tierra del Fuego in South America, which you just saw above, and are about twice as far from southern Australia and South Africa. We will examine the Antarctic again on page 7-3 but for now there is this overview. First a mosaic of the continent, whose ice cap extends over about 13000000 square kilometers (geographic features of the Antarctic on imposed on another mosaic on page 7-3).

In the winter (during the summer in the northern hemisphere), sea ice forms an extensive shelf that greatly expands the Antarctic's size:



It is surprising to learn that there is really no one continuous rock continent below the ice cap comprising the Antarctic. Instead, its bedrock is a series of island archipelagoes that formed a

base on which ice built up and extended from one island to its neighbors.



But now let's leave the western hemisphere to head east and cross the Atlantic Ocean to Europe.

PREVIOUS NEXT

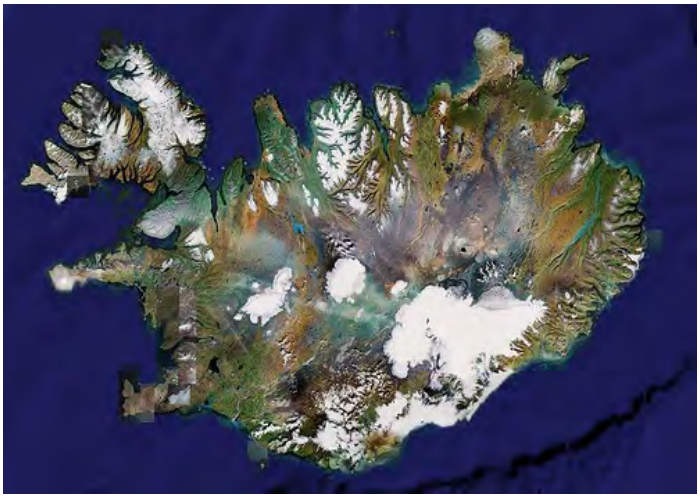
Primary Author: Nicholas M. Short, Sr.

Europe, the Mediterranean, and North Africa have been the cultural center of Western Civilization since the eras of the Sumerians, Egyptians, Greeks, the Romans, and various empires and lands up to and after the Renaissance. Iceland is considered to be within the sphere of Europe, so we begin there. For most traveling from the U.S. to European centers of our Western roots and culture, a full tour begins in Ireland and ends to the east at Moscow. The majority of the continent is moderately vegetated - much land has been cultivated now but forests remain.

Europe I

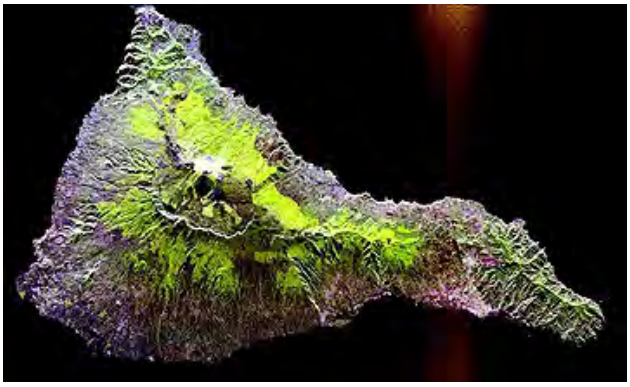
(Note: Some of the images below were provided by ESA's Eduspace. Additional images can be found at their Eduspace links site.

It may come as a surprise to learn that many geographers consider Iceland to be a part of an extended Europe. It was settled by Europeans centuries ago and still has Danish roots. But, geologically, Iceland lies at the boundary between the Eurasian and North American tectonic plates. It sits atop one of the few places where the Mid-Atlantic Ridge, a divergent spreading zone of volcanic origin, extends over a wide area above the sea surface. Take a look at this satellite image taken in summer when the winter snow have melted (but some of the white is a permanent icecap, shown again on page 17-5).

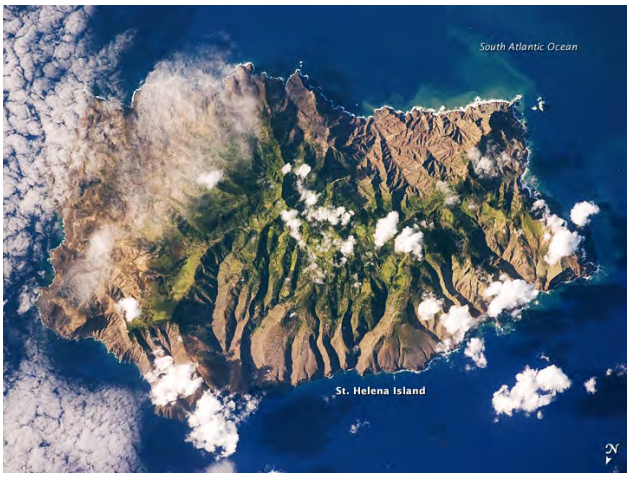


Iceland has become a popular place for tourist visits. The Icelandic nation has a population of about 330,000 people. Its capital is Reykjavik, on its southwestern shores. Here is an ASTER image along with an aerial photo:

In a sense, this next image also belongs to Europe. The Canary Islands are a part of Spain but lie in the Atlantic off the coast of Morocco. The island of Tenerife is built as a volcano (called Teide) rising as an offshoot of the Mid-Atlantic Ridge. It is seen here as a SIR-C image:



Another volcanic island on the Mid-Atlantic Ridge is St. Helena, shown here in an astronaut photo from the International Space Station. St. Helena was the British island where Emperor Napoleon Bonaparte I was sent by the European allies into exile following his defeat in 1815 at the Battle of Waterloo. He remained on this bleak landscape until his death in 1821.



Now to Europe proper.

Most of us are familiar with the shape of Europe and its surrounding regions from our geography lessons since our grade school days. Now here is what that continent (continuous with, but arbitrarily separated from, Asia) looks like as seen in a physiographic map (greens imply significant vegetation; browns could either mean less vegetation, as in Russia, or, in Africa, desertlike conditions):

Here is an up to date map of Europe that shows the countries from the British Isles to Russia.

To assist those who are not too familiar with the locations of Europe's major cities, we provide this map:

As we might cross the Atlantic Ocean in a spaceship, the first land we might see if enroute to, say, London is the British Isles. Here is a satellite view, looking east (top), of these great islands (Britain-Scotland-Wales and Ireland), with the continental coastline from about Normandy to southern Norway:

Here is a map showing the major cities in the British Isles:

Ireland, the "Emerald Isle" so dear to every Irishman, has been colored green in this next scene by projecting the TM Band 4 image through a green filter. The upper eastern part of the island is Northern Ireland, at the end of the Belfast Lough (actually, a marine bay). The large inland lake to the west is Lough Neagh. In this June image, late spring, the Irish countryside is truly a rich green almost everywhere. The Republic of Ireland occupies the larger part of the land. Its capital, Dublin is on the coast just below center right. The large island in the Irish Sea is the Isle of Man. The jutting land peninsula in top right is the tip of the Scottish mainland, at the Mull of Galloway.

The above image is much like one produced from SRTM radar data used to determine elevation. Low areas (central valleys) appear in green; higher, more hilly to mountainous terrain are in shades of brown:

A bit of the charm of Ireland appears in this photo of Clifden, a resort town on the west coast in County Galway:

Dublin on the Irish east coast is quite familiar to American Irish who journey back to the land of their forekin. Here is Dublin as seen by SIR-C radar (Section 8):

And the center city as seen by SPOT:

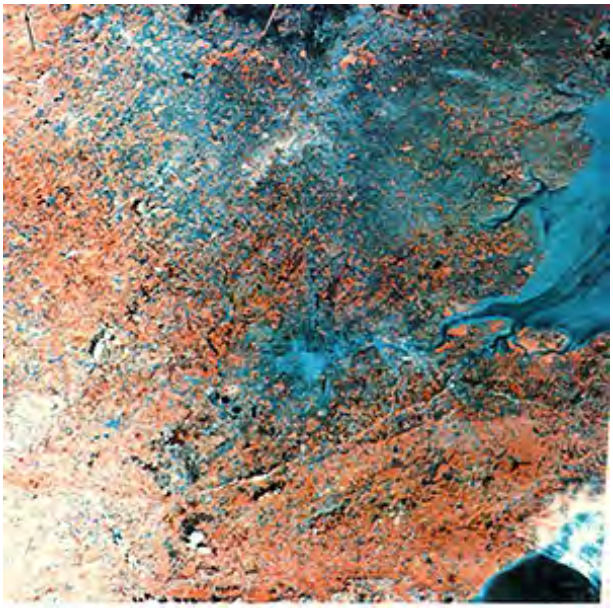
And here it is "up close and personal", looking down O'Connell Street, the city's main thoroughfare:

Ireland has always been a favorite destination for Americans, many of Irish ancestry and others who seek its enchantment as tourists. A must-see is the rugged west coast of the Republic of Ireland. A favorite of most visitors is the Ring of Kerry, which follows the coastline around the Iveragh Peninsula, in County Kerry. This SIR-C radar image shows the entire Peninsula, with a hint of the hilly terrain that marks its interior; below it is a typical scene along the Ring:

The British Isles have a special meaning to many Americans. From an historical standpoint, England especially is cherished, since it can rightly be called the "Mother Country". We look at southern England next in an unusual situation: a thick cover of snow blankets most of the landscape, a relatively rare event owing to the normally milder climate afforded by warm ocean currents.

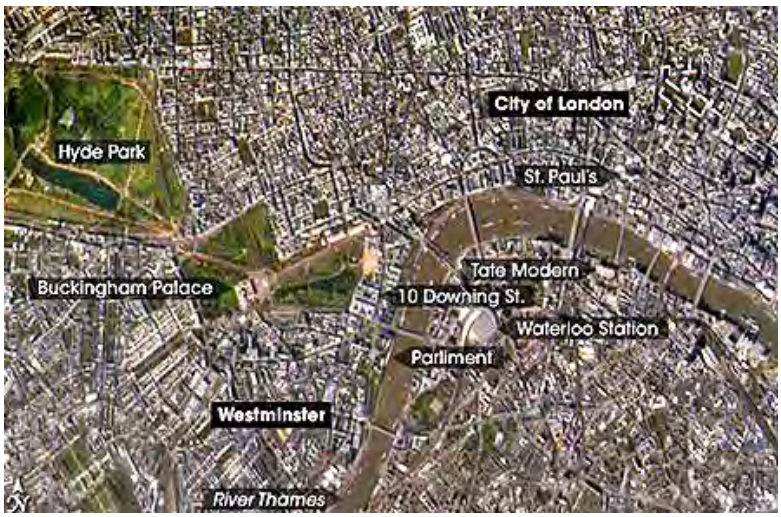
The demographic structure of the main cities of the southern British isles is well-defined by seeing it at night, as it is displayed by its lights:

One of the great cities of the world is London, shown here in this scene that covers part of southeast England. The Thames Estuary is conspicuous. The countryside, in this March 1973 Landsat image, has just begun to "green up" as field crops and grasslands resume growth. Besides greater London (blue, near center), other well known smaller cities, among them Oxford and Reading, are hard to discern. The dark, vegetation-poor area in the upper right is East Anglia, whose rocky soils inhibit extensive farming. Through London itself passes the Prime Meridian (0°) for the global latitude-longitude coordinate system.



This ASTER image zeroes in on Greater London and surrounding countryside. The Parliament area shown two images down is near the center. The green areas to the west are St. James Park and Hyde Park.

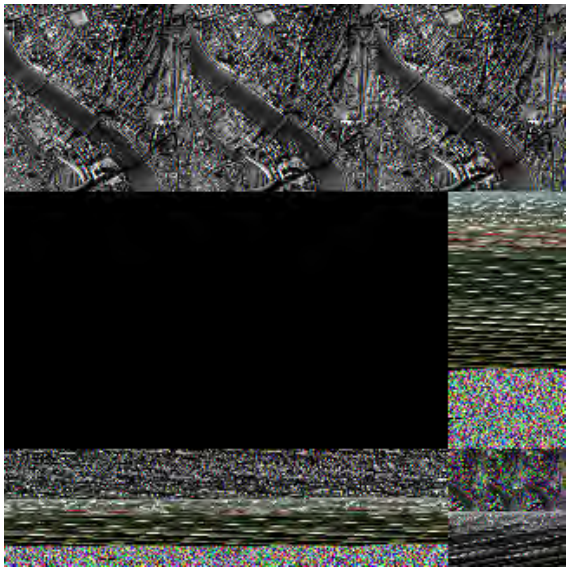
This annotated astronaut photo taken from the International Space Station shows many of the famed landmarks within the central city.



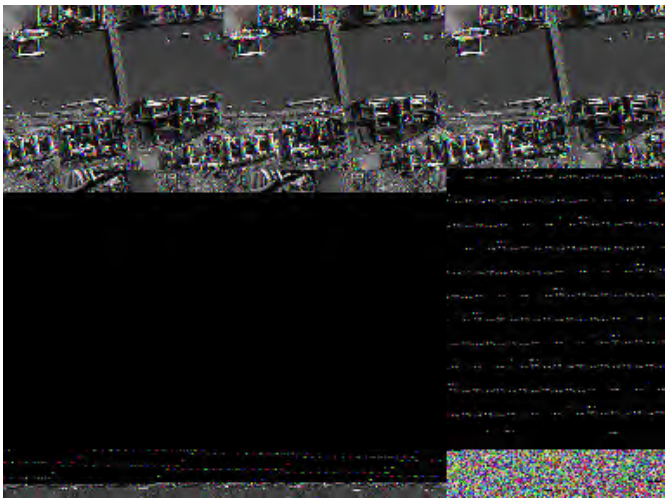
Americans feel a special kinship with England, despite our revolt against our "Motherland", its monarchy system, and unrepresented TAXES. But England's Kings and Queens remain through today. Here is the London home of royalty, Buckingham Palace, as seen by an OrbImage satellite:

London's fame and its tie to Americans' heritage will prompt us to spend a bit more time in looking at some of its landmarks. Greater London's area is among the largest in Europe. The image most of us conjure up when asked to visualize London will include the Parliament Building and Big Ben.

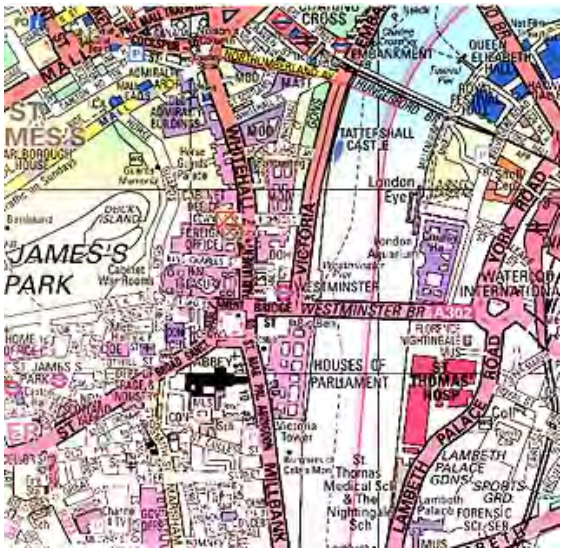
The Quickbird satellite has taken high resolution images of the Parliament buildings, the nearby bridge across the Thames, and the Jubilee Park rededicated in 2000 (Millenium Park) with its huge Ferris Wheel (400 ft high) known as the London Eye.



An enlargement, taken at another time, shows both the Parliament buildings and the great Wheel in a different orientation than above.



Much of the same area seen in these two space images is identified in this map:

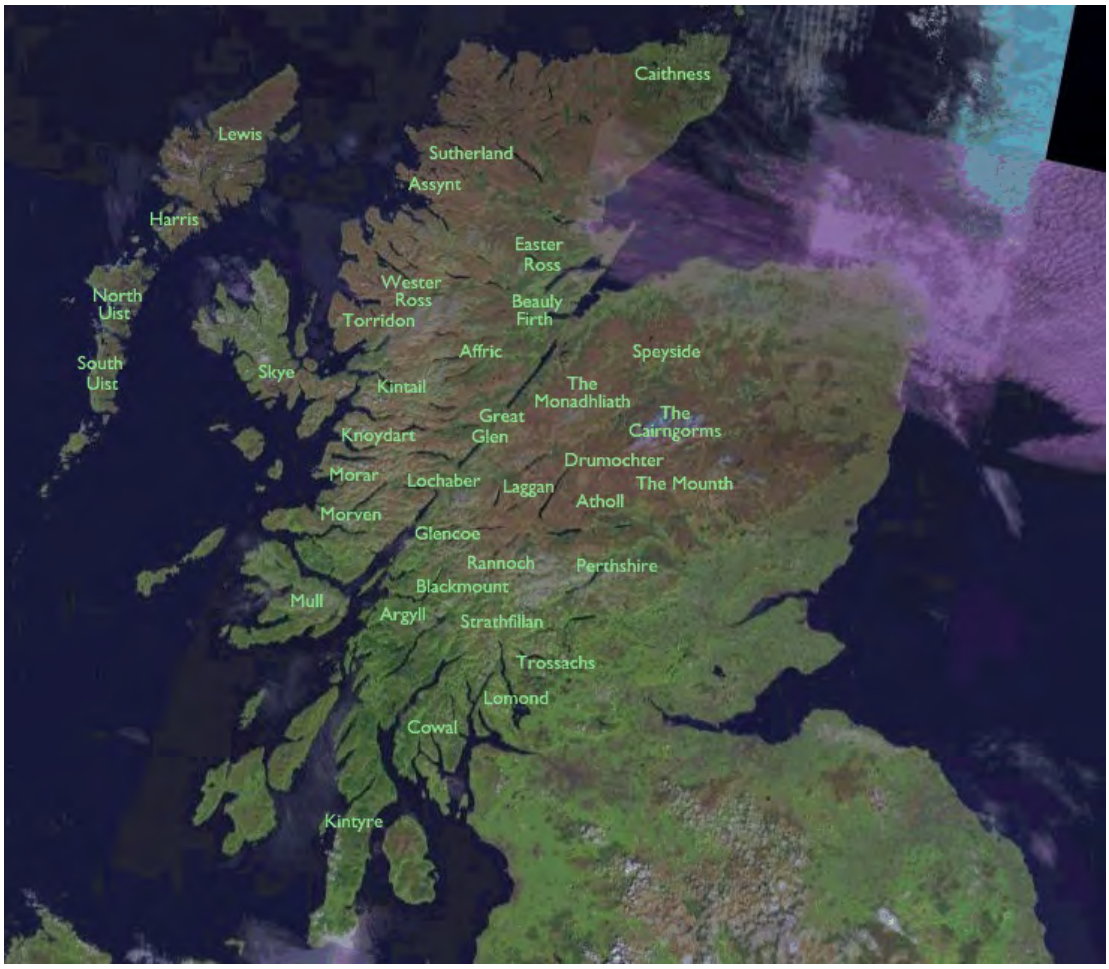


See if you can find the following in the images: Parliament buildings and Westminster Abbey; St. James Park; Waterloo Station; Charing Cross Station; Her Majesty's Treasury; St. Thomas Hospital. We show below a photo of perhaps the visual symbol of London, Big Ben and the Parliament buildings:

Beyond the Thames and the government buildings, present day London is acquiring a modern look, with tall buildings (but few skyscrapers). Some tall buildings noted on the north side of the Thames in the space imagery take on a somewhat different perspective when seen in this color aerial oblique photo:

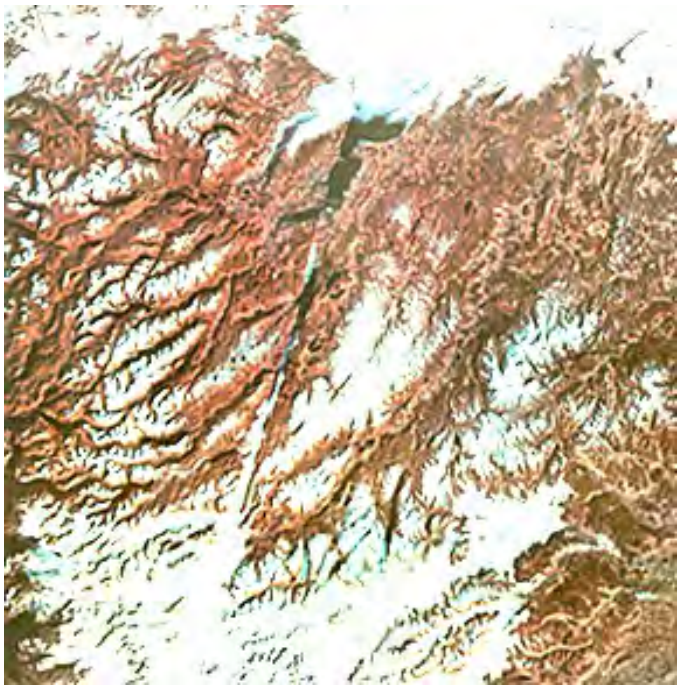
Two of the largest cities in England are Manchester and Liverpool, seen here in big versions (hopefully, to allow labeling to be read) of Google Earth images:

In much of Scotland are the Scottish Highlands, low mountains that experience glaciation. They are a large part of this scene from space, with well known towns indicated by the annotation.



The northern and western parts of Scotland are mountainous (reaching altitudes above 1300 meters [4000 ft]). The Scottish Highlands (part of the great alpine chain that includes the Appalachians) are a popular tourist destination.

In the history of modern Geology, this is a famous region where many of the 19th Century concepts about metamorphism were developed. The Highlands are divided by the Great Glen Fault, a dextral strike-slip fault (west side has moved southward) around which the west block moved about 100 km southward. The fault localizes valleys and lakes, the most famed of which is Loch Ness (guess who lives in it?) at the north end. Here is a Landsat subscene showing the fault and the Highlands (Grampian Mts. at bottom center and Northern Highlands in upper left):

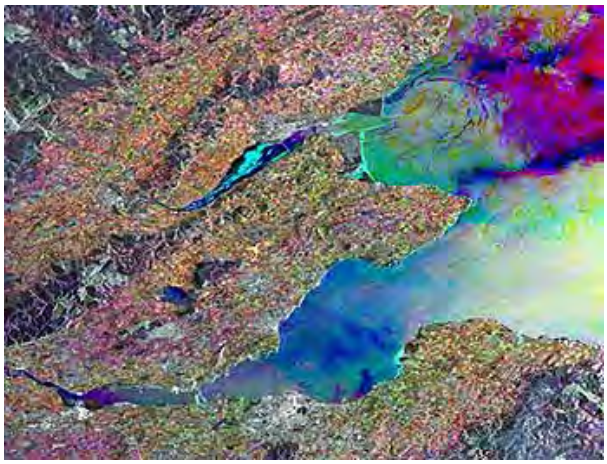


The mountains west of Loch Ness were once part of the present North American crustal block whereas those to the east belong to the Eurasian block. These blocks came together about 400,000,000 years ago as a supercontinent that has since split apart. The Great Glen fault is of that age. The fault itself is a zone of weakness where a river carved out a valley that was occupied by glaciers as late as 10000 years ago. Glacial deposits blocked the northern outlet causing the post-glacial runoff to collect into the deep Loch Ness. Here is another view from space together with an aerial view of the south end:

In the western Highlands is the large Ardnamurchan Peninula, where the concept of a "ring dike

complex" was first developed. Here is the area from space:

The capital city of Scotland, the upper half of the British Isles, is Edinburgh (pronounced "Ed-in-bruh locally but "Ed-in-buro" by Americans). This is an ERS-SAR view, somewhat enlarged. Note the blackish oval feature in from the right center. That is the famed "Arthur's Seat", which was a key aid to the Scot James Hutton who is credited as the one who laid the foundation of modern Geology.



Less colorful but also informative is this Landsat image of Edinburgh and the Firth of Forth:

The old city of Edinburgh has a low skyline. Its wide streets help to make it an especially attractive city. Here is a view:

Across the North Sea from England and Scotland lies the northwestern part of the Netherlands, or Holland. The scene below reveals one of the great engineering feats by any nation. Much of the land was once under water or has been created by backfill. The great inland bodies of water were known as the Zuider Zee after they enlarged following breaches in the 13th century of the line of present day barrier islands known as the West Frisian Islands. In 1932 a 25 km (16 mile) barrier dam was completed to divide this water into the Wadden Zee (north) and Ijssell Meer (south). The change in flow facilitated drainage that reclaimed landmasses such as Flevoland, known as polders; these show up a bluish-gray areas. Some of the many Dutch canals are visible. Amsterdam, Holland's largest city (in blue tones, near bottom center), lies along a major canal to the sea at the bottom of the silty Meer.

This Envisat image shows Amsterdam near the top and Europe's largest port, Rotterdam, in the lower left:

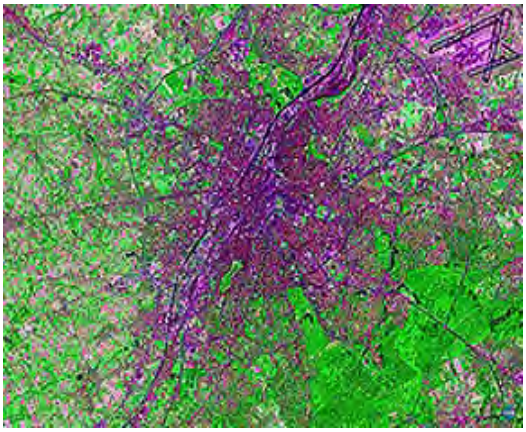
This Landsat subscene provides a closer look at Amsterdam and the surrounding countryside

Look at the inner city, shown as a series of semi-circles. The map reveals their identity

Amsterdam is a city of canals and travel by boat is common, but bicycles as well as cars are also much used.

Amsterdam and some other parts of Holland are on reclaimed land. Seawall dikes protect these areas, some being below sealevel. Two other cities in Holland are well known - Rotterdam (on one of the distributaries at the mouth of the Rhein) and the Hague (Gravenhage):

The capital of Belgium (and Headquarters of NATO) is Brussels, seen first in a Landsat image and then by SPOT:



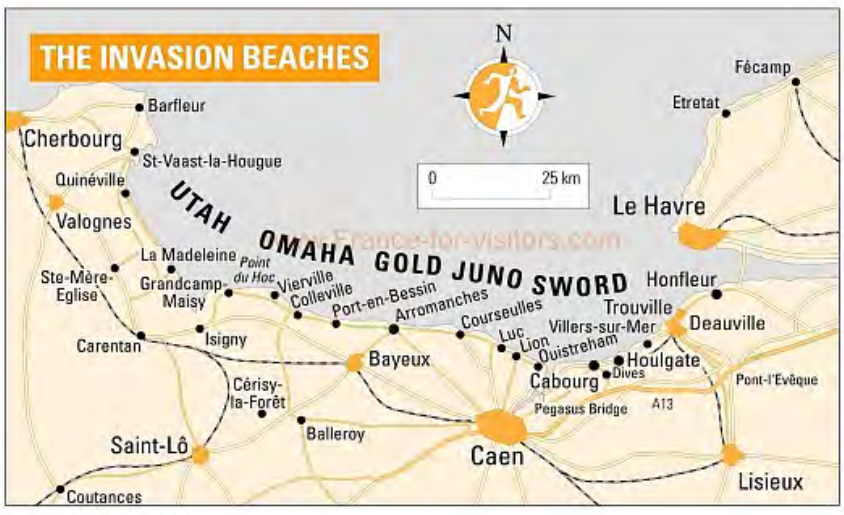
Nestled among Belgium, Germany, and France is the Duchy of Luxembourg, a small independent country. Here is a view obtained by the SPOT satellite:

France has long been a major, generally stable part of Europe. Here is the entire country as seen in a satellite mosaic; below it is a map of the provinces of France (equivalent to the states in America):

The queen city of Europe - in terms of beauty - is Paris, France. We have already taken a thorough visual tour of Paris on page 4-4, which you can check out as a refresher. As a reminder, here is a SPOT image of the City of Lights. Then we will show some other images of France:

First we examine a very famous area off the English Channel - the beaches of Normandy, the scene of the Allied invasion of France on June 6, 1944 (see the movie "The Longest Day"), and part of the Cherbourg Peninsula:

This is a map of the five landing beaches and a satellite strip showing Omaha and Utah beaches (both attacked by American troops):



from www.france-for-visitors.com

This next photo can be labeled "What was Eisenhower thinking when this was chosen?" It shows the cliffs at Omaha Beach and remnants of fortifications on the flat land above.

Aside from Paris, there are few really large cities in the French interior. One of the smaller is Clermont-Ferrand on the Allier River which passes through the Auvergne:

One of France's largest cities is Bordeaux near the Atlantic in the southwestern region. The city lies near the Garonne estuary, long and sheltered, which makes up an ideal protected area that gives this region port status:

Lyon is another major French city, shown here in a SPOT image:

Part of the Alps extends into France. The alpine city of Grenoble is seen in this Landsat image:

Next, we traipse southward to the Mediterranean. Here is much of the French Riviera, with Toulon on the left and St. Tropez and Cannes further east:

A closer look shows Toulon in the center of the shoreline in this image and Marseilles at the left edge.

Then, we go offshore to the French island of Corsica, birthplace of Napoleon Bonaparte:

Heading next into Scandinavia, the island of Zealand (Sjaelland) in Denmark dominates the next image. Copenhagen, that country's capital, is the dark blackish area at it upper right (east) end. A narrow strait, the Ore Sund, separates this part of Denmark from southern Sweden; Malmo, one of that country's larger cities, appears as another black spot. Farming dominates the land use in these sections of both nations, principal crops being wheat, potatoes, and animal fodder. In Sweden, the reddish area is a forested region known as the Smaland Plateau.

From space Copenhagen, Denmark's capital, looks like this:

While we are presently in Scandinavia, lets look at three more capital cities. First is Oslo, Norway, seen here as a small blue patch at the end of a Baltic Sea inlet in this Landsat-2 subscene:

This is Oslo, as approached from the Baltic Sea; the city has a distinct atmosphere of being a large, friendly "town".

This ESA's ERS-SAR close-up of Oslo gives a better sense of the low hilly topography around the city (white):

Oslo is a relatively small city (population about 750000) nestled in low mountains (the high mountains of Norway, with their fjords, are to the west). This view of the town was taken looking west from the Eckeberg Restaurant where I had lunch in 1960 with the famed Russian geologist V.V. Beluossov.

But the wonder of Oslo is not its views. Near the heart of the city is a park containing dozens of statues made of bronze or granite. These were all cast or carved by one man - Gustav Vigeland - in the first half of the 20th Century. The theme is "Man - from the Cradle to the Grave". Here is just one example of these wonderous sculptures:

The west coast of Norway is famed for its rugged scenery and jagged coastline. Much of this is due to Pleistocene sculpturing of stream valleys coming off mountains by glaciers. The glaciers deepened the valleys, which flooded as sea level rose in the last 10000 years. The submerged valleys, with steep-walled sides, produce the fjords, some of which can be 100s of meters deep. This image of Norway's coastline was made by the MISR sensor on Terra:

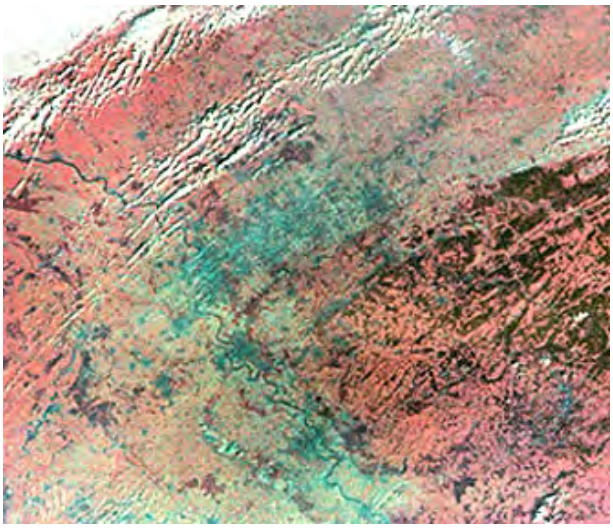
Let's go east for a look at the inner city harbor of Stockholm, Sweden as seen by IKONOS, and then the equivalent area from an aerial oblique photo.

The fourth capital - Helsinki, Finland - lies along the eastern Baltic coast:

One thing the four Scandinavian capitals seem to share: They have few tall buildings, perhaps an attempt to keep their skylines more medieval.

Germany has been a pivotal European nation (although that status only happened in the 19th century [1871]; prior to that, especially during the 900 or so years of the Holy Roman Empire [962 A.D.] it was a loose collection of Duchys, Kingdoms, Principalities, and Free Cities) for much of modern history. Below is a satellite mosaic of Germany as reconstituted in 1992 when East and West Germany reunited; the principal states of Germany are shown in a map beneath the mosaic:

The industrial heartland of Germany, so vital to Deutschland in World War II, dominates this next image. The famed Rhine River runs through the State of North Rhine-Westphalia. Numerous blotches of dark bluish-black mark the locations of many large cities (use an Atlas to pin these down). Along the Rhine are Bonn, Cologne, and Dusseldorf. The smaller Ruhr River, coming from the East, makes up the Ruhr Valley, with such cities as Mulheim, Essen, Bochum, and Dortmund whose steel and industrial works were the targets of many bombing raids by the Allies during World War 2. Much of the countryside is part of the North German Plains, where farming is the principal activity. The forested Central Uplands to the east blends into the Rothaar Gebirge. Small parts of Belgium and Holland lie along the left edge of the image.



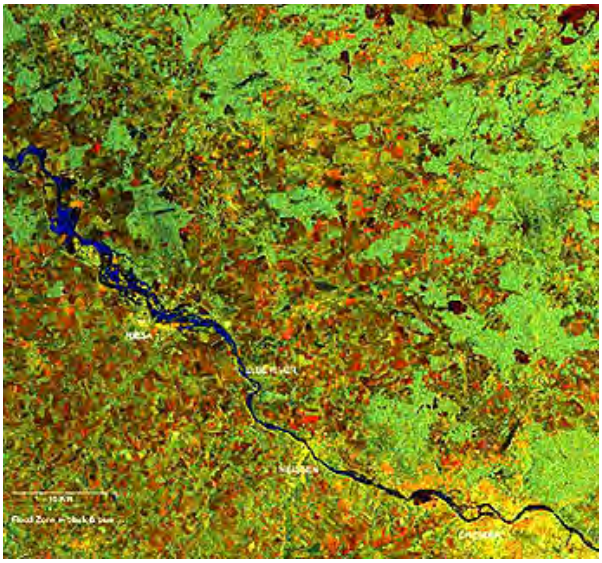
The four principal cities of the Ruhr are located on the map below, along with a ground photo that hints at the widespread industrialization of this Valley:

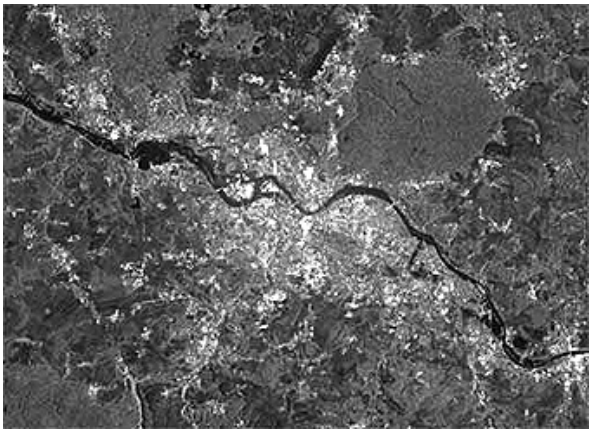
Upstream - to the south - on the Rhine are important cities such as Bonn, Frankfurt, and Wiesbaden. Another is Cologne (Köln), a very old city - it was founded by Emperor Claudius' wife Agrippa in the 3rd century - that has been an important center for shipping and industry in modern Germany. Here is a SPOT-1 pan cam 10 meter resolution image of Cologne (yes, the perfume originated there), and an aerial view beneath it.

Dominating Cologne is its great Gothic cathedral (Dom), started in 1248 A.D., that is the largest in Germany. Here is an IKONOS view that includes the cathedral:

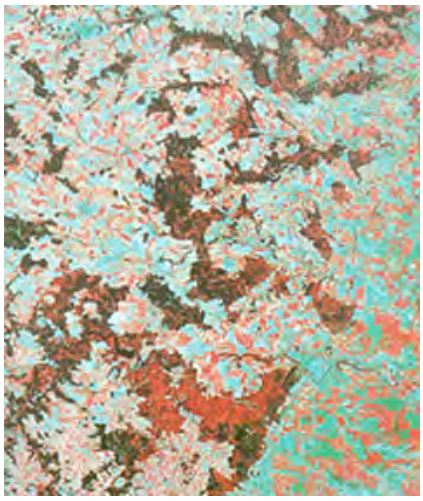
Cologne was destroyed by Allied bombers in World War II, but U.S. airmen took special pains to avoid hitting the cathedral. This picture shows the devastation:

Even more destruction leveled the city of Dresden, on the Elbe River. One terrible night saw massive firebombing that incinerated most of the city. But, like most such calamities in the cities of Germany, Dresden has been almost completely rebuilt (both with modern new buildings and with reconstruction of ancient and medieval landmarks). Here is an Envisat image of Dresden and other cities on the Elbe River; below it is a Radarsat image of today's Dresden; then an IKONOS image of central Dresden:





Both the Dresden scene and that of Berlin, to be looked at next, point to something considered to be a land cover signature of Germany, and to a lesser degree, several other European countries. Look again at the Envisat image and note the many irregular polygons that have a medium-dark brownish color. Early in my experience with Landsat I noted the same distinctive signature and attributed it then to large stands of natural forest land. This false color subscene shows such a feature in more detail:



Here the polygons appear to be a mix of deciduous trees, possible small lakes, and whatever else is contributing to the "blackness". This last may be dark evergreens, but the interpretation is inconclusive. It seems surprising that so much of the land would still be forested inasmuch as the Central European countries have large populations. But many decades ago, the Germans undertook a massive reforestation program to counter the downcutting of most of the medieval forests that served to provide firewood. I'm guessing, but think I'm right; on my first day in Germany in 1960 I took a lengthy walk through such a woodlands (whose pathways were frequented by cyclists and native pedestrians).

So, we would like to add Berlin to our display of European capitals. An Envisat image does include this famous city, as a pink patch in the lower right, in a wide field image that covers much of northern Germany, eastern Denmark, and southern Sweden.

An ERS SAR image affords a closer look:

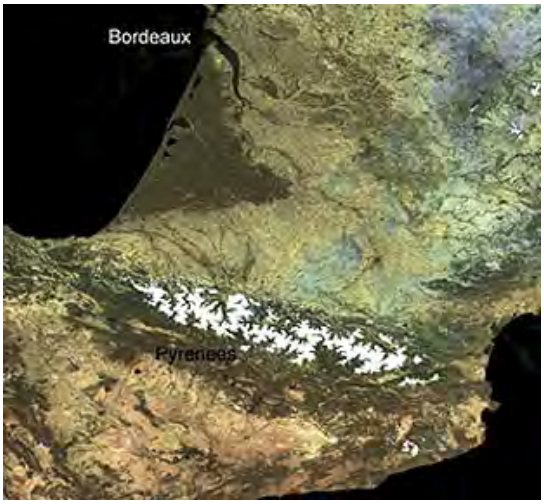
This natural color ASTER image establishes the setting of Berlin within its countryside.

Another ASTER image shows the heart of Berlin:

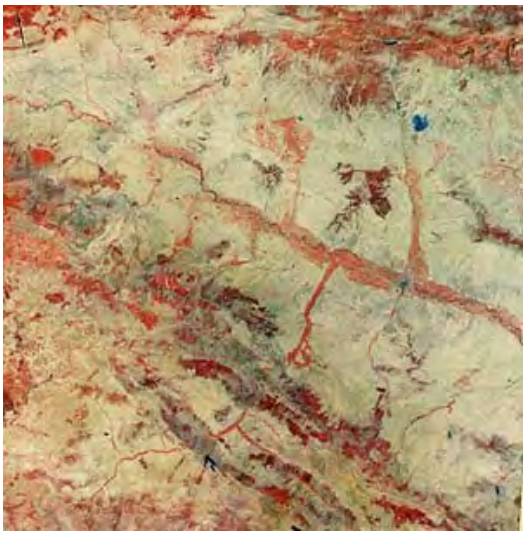
But Germany is also noted for its picturesque towns and small cities. Among the most famed is Wittenberg, on the Elbe in north central Germany. It was here that Martin Luther on October 31, 1517 nailed his 95 theses to the castle church of All Saints, setting into motion what became known as the Protestant Reformation. As seen from space and on the ground:

Having looked at western and northern Europe, we now make a big leap to the southwestern part of the continent, specifically for a tour of the Ibernian Peninsula, shown here first in an ERS-1 image and a corresponding map of Spain.

The Iberian Peninsula joins the rest of Europe in southwestern France. The peninsular landmass crashed into the European block as part of a plate tectonics collision starting about 150 million years ago. The result was a mini-version of the crumpling of sedimentary rocks into the folded mountains of the Pyrenees:

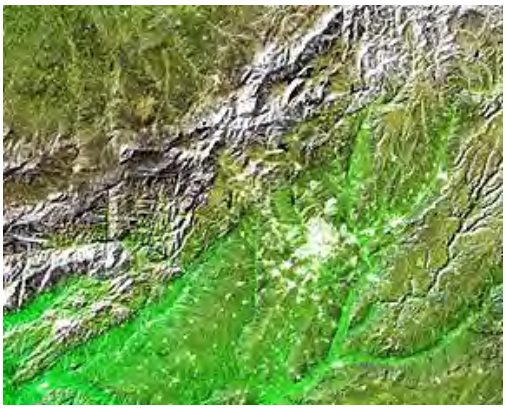


Parts of the peninsula are semi-arid (light buff) whereas other parts are covered with more green vegetation, particularly those regions with higher elevations. Much of northern Spain is dry with rainfall commonly less than 30 cm (12 inches). In the next image, the wide Ebro Basin, drained by the Ebro River, forms arid lowlands in which most of the agriculture (cereals; sugar beets) is concentrated along that river, tributaries, and irrigation ditches. Away from these farms, sheep grazing is conducted where grasses grow. The city of Zaragosa is evidenced by a blue patch in from the right center edge. Low blocks of hills and mountains run northwest through the picture. The southern edge of the Sierra de Guara, which rises to more than 2000 m (6000 ft), is at the top.



Barcelona is the biggest city in eastern Spain. It sits astride the Mediterranean, as evident in this Landsat image:

Vegetation increases somewhat as one moves into central Spain in the region known as the Meseta (a plateau). This C-Band SRTM radar images shows Madrid and its surroundings:



This March 7, 1973 Landsat-1 image of central Spain shows the capital, Madrid (dark blue-black) to the west of the Tagus River. To its west is the Sierra de Guadarrama, wooded and high enough to have snow in the part that reaches 2400 m (8000 ft).

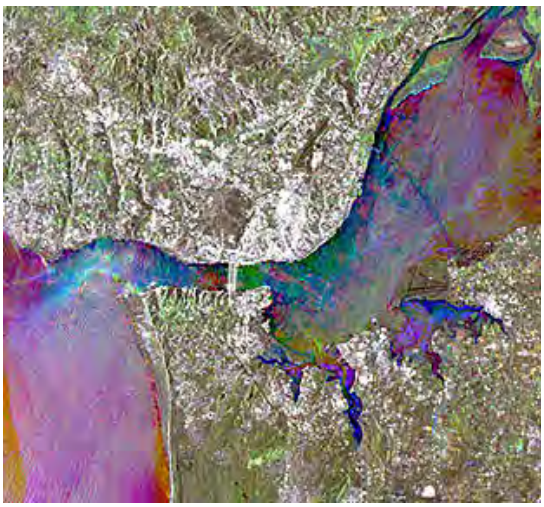
Madrid itself is shown in more detail in this Landsat-7 subscene.

The Palacio Real, the national palace now occupied by King Juan Carlos, is shown first in an IKONOS 2 m black and white image, and then in an aerial oblique view.

Granada, in Andalusia, is one of Europe's most charming cities. It lies against the snow-capped Sierra Nevada (mountains). Expanded during the Moorish occupation, the city is dominated by the Alhambra, built by the Moors, seen along the skyline in the aerial view:

Southwest of Madrid is the city of Seville (locale for George Bizet's opera "Carmen"). Here is part of Seville as displayed in this Google Earth image:

On the western side of the Iberian Peninsula is Portugal, whose capital is Lisbon. It lies on the northwest side of a bay created by the Rio de Tejo (Tagus River). This area is shown first as a Landsat-4 subscene and then as an ERS SAR scene.



One of the writer's strongest recollections of Lisbon is entering it from the east over the Vasco da Gama bridge:

The tip of Europe at the western end of the Mediterranean Sea is the English Colony at Gibraltar, next to Spain. Across the Strait of Gibraltar lies Tangiers in Morocco. This astronaut photo looks eastward.

At the tip of the Iberian Peninsula is the famed Rock of Gilbralter (an outlier made of limestone):

Most people do not know that 5.3 million years ago, the Mediterranean Basin was largely dry (some lakes). The Atlantic Ocean then was lower than today. As sealevel rose, the Mediterranean filled up rapidly to its present extent.

(Europe continues on the next page.)

PREVIOUS NEXT

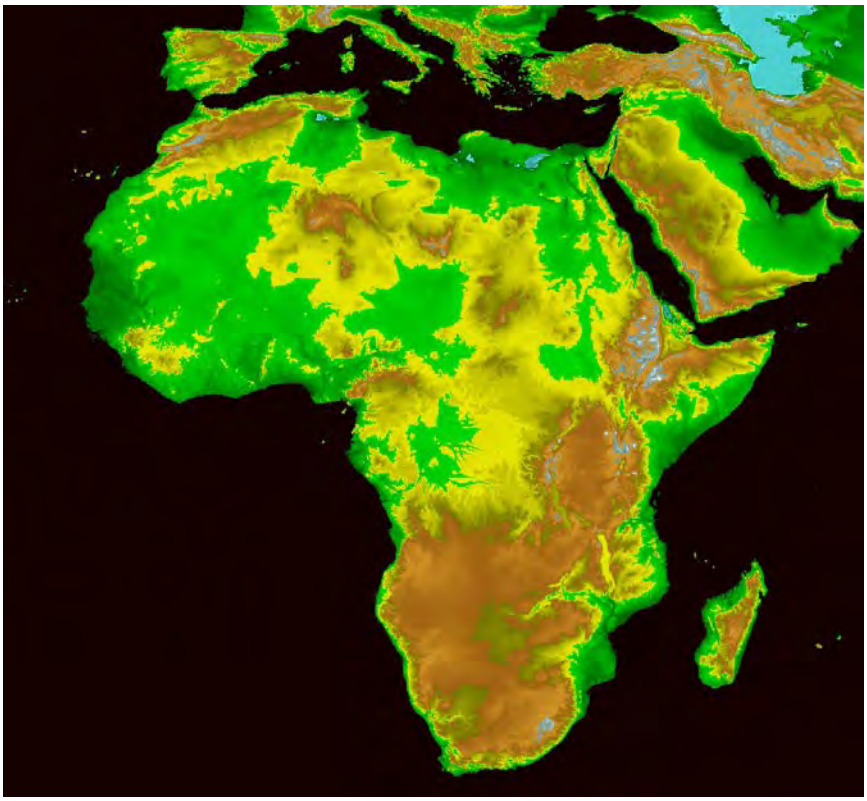
Primary Author: Nicholas M. Short, Sr.

Most of the northern third of Africa and the adjacent Middle East has landscapes that are the product of past and current arid to semi-arid conditions. Desert dominates most of northern Africa. The central third of Africa in many places is much more heavily vegetated, including tropical rainforests. The southern third of that continent again reverts to semi-arid conditions in which savannah vegetation is particularly common. Much of Asia Minor and the Middle East bears resemblance to landscapes similar to northern Africa.

Africa and the Middle East

Africa is a continent noted for its stark varieties of diverse landscapes. While there are relatively few large cities, and many nations are still in the third world underdeveloped category, populations of pluralistic ethnic groups are distinctive. We will establish a framework for viewing Africa using the next four illustrations. The first sets the geographic boundaries of its many nations, and shows its relief:

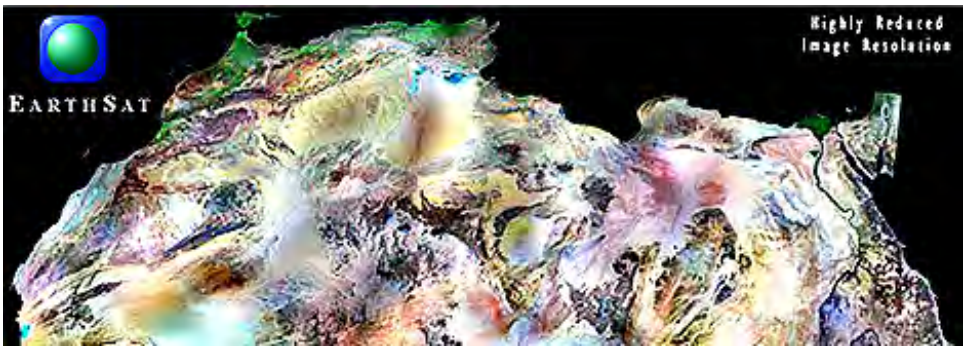
This relief can be emphasized to better show the distribution of mountains and highlands:



Prominent are the Atlas/Anti-Atlas group of mountains in Morocco, the Ahaggar, the Tibesti, the Ethiopian Highlands, the East African Rift zone, and the high plateaus of southern Africa. These are labeled and located in this map:

The next pair show a construction of the main surface cover types (desert, forest, savannah, etc.) and a classification of the principal types.

The yellow in the land cover map is listed as desert. Indeed, the northern part of Africa is mostly desert except for vegetated coastal areas. Much of the desert is rock rather than sand. This Landsat mosaic of northern Africa shows the actual variety of land cover types (rocks are darker; sand is yellowish):



We will begin our journey through Africa in this northern sector. Shown first is the western gateway to Africa, across the Mediterranean from Gibraltar to Tangiers, Morocco:

Gibraltar is a British colony, won from Spain after a war. The best known landmark is the Rock of Gibraltar (the symbol of Prudential Insurance). Here is a ground photo:

This limestone massif is actually a sea stack now emerged as sealevel has readjusted. This view shows the Rock as part of a spit:

Further south along the Moroccan coast is the city of Casablanca, which gained fame in the West as the locale of the movie "Casablanca", starring Ingrid Bergman and Humphrey Bogart:

To the east are the Atlas and Anti-Atlas mountains of southern Morocco. These are known locally as the Jebel Bani. The Atlas and Anti-Atlas Mountains of northwestern Africa, seen earlier on page 2-6, are a collision-fold belt similar (and related) to the Appalachians. The region is sparsely populated and experiences low rainfall. The next three images show these mountains in an Envisat, a Landsat, and a Digital Globe perspective view:

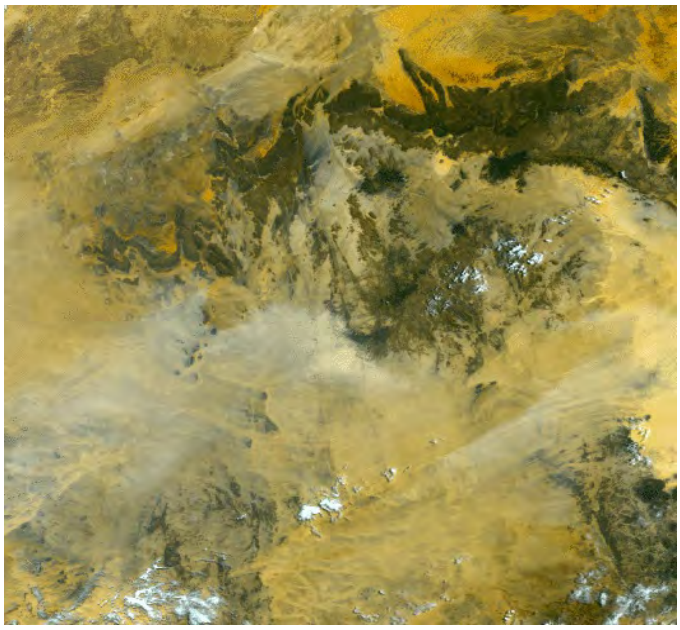
It was in the eastern end of the Atlas that the U.S. Army had several fierce battles with the German Army in 1943 during World War II.

Algiers is one of the larger African cities on the Mediterranean coast. This capital of Algeria is seen in a Landsat subscene:

Tunis is the capital of Tunisia. It is located on the Mediterranean Sea coast, not far from the ancient city of Carthage (a city founded by the Phoenicians [whose home land was just north of modern Israel]).

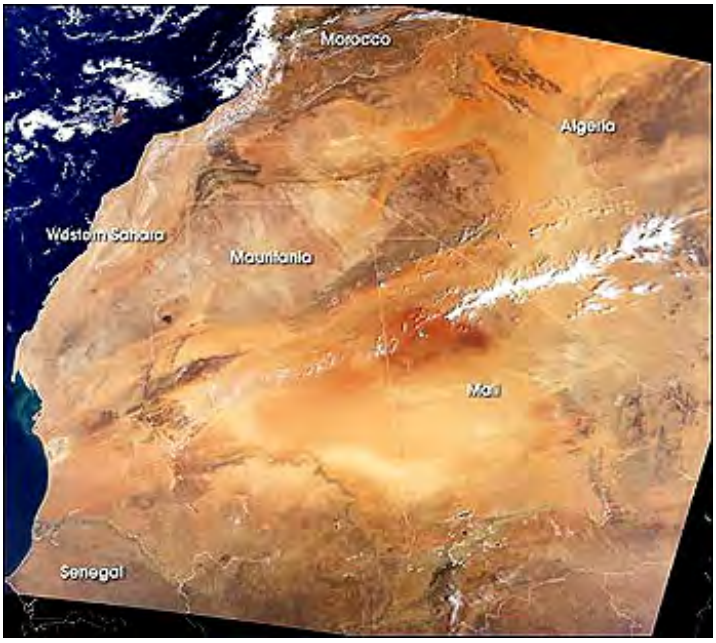
Many people, influenced by movies, visualize the African desert as never-ending sand dunes. This next scene, in south-central Algeria, is in fact a more typical expression of the Sahara Desert. In the lower left third is the Tademait Plateau, underlain by sedimentary rocks, and crossed by dendritic intermittent streams whose channels were cut during a wetter epoch in the Pleistocene. The land in the lower right is the Tidikelt, a barren gravel-surfaced plains. The upper part of the image is a segment of the Grand Erg Orientale, whose most prominent feature is the series of longitudinal dunes first established when the wind blew more from west to east than today. This whole region is almost uninhabited.

The Ahaggar Mountains (also known as the Hoggar Mountains) of southern Algeria are shown in this image, together with a ground scene:

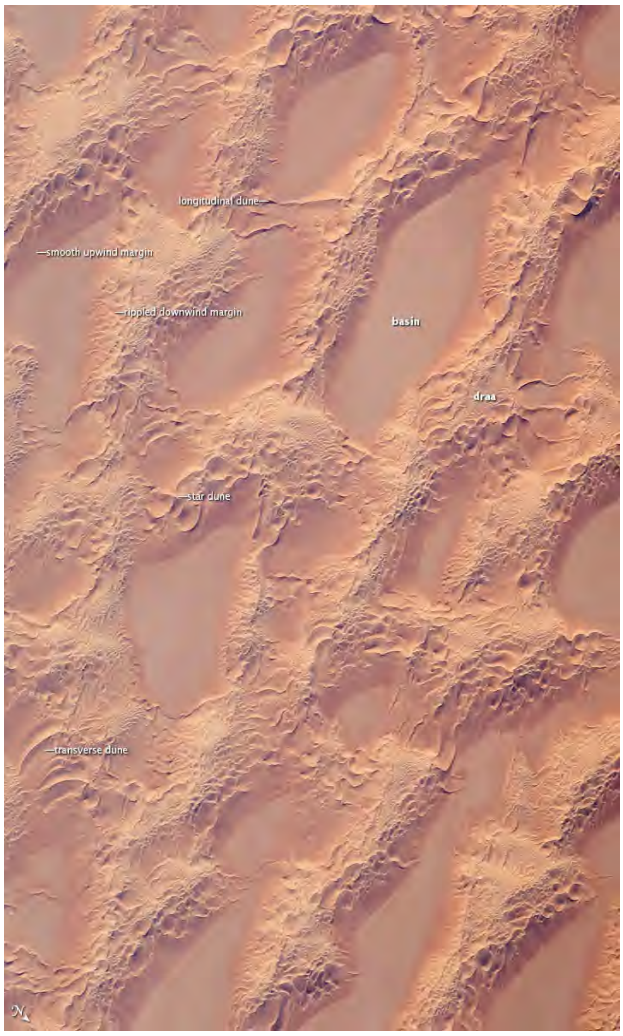


Because of the light colored background afforded by the sand, rocks such as basalt stand out in sharp contrast. Astronauts generally comment about one feature they see almost as the "African landmark" - the fresh basaltic flow from the Pic Tousside volcano in the Tibesti Mountains of northern Chad. Here is a Landsat-7 ETM+ image of this feature within the Tibesti Mountains, first in a fuller setting and then in close-up:

As said above, the hallmark of the Sahara is the absence of notable vegetation, so that the desert from a distance looks like a great expanse of sand and soil having a characteristic yellow to orangish color overall, without much diversity of features when an image is low resolution. This Envisat image shows these characteristics; the scene extends from the Anti-Atlas Mountains on the north to the Inland Delta of the Niger to the south (individual Landsat images show these areas on this page).



When northern Africa is mentioned, many people get images of deserts. The Sahara is the greatest expanse of sand and dunes in the world. This next scene is typical:



Desert people make up much of Mali's population. Mali is one of the countries of Sub-Saharan Africa (desert interspersed with vegetation owing to greater rainfall). The nation is crossed by the Niger River on its way to the bottom of the western hump of Africa. During the rainy season water spreads over the flat lands producing an "inland sea" of sorts. This feature is actually referred to as the Inland Delta of Mali. The water in this scene is black, with thick wetlands vegetation in red. The Bambara and Peuhl peoples farm the area and raise Zebu cattle.

Near Mali is the small country of Burkim Faso (chances are you've never heard of it but check it

on the map of Africa above). Its capital - Ouagadougou - is a challenge to pronounce but is a thriving city of more than a million:

Africa conjures up an image to many as a vast dense jungle. However, most of the continent is ecologically either a savannah (brush and grasslands) or a desert similar to the type in the image just above. But, in the "Congo", much of that region being in today's Democratic Republic of the Congo (DRC; known for a while as Zaire), a true jungle (Tarzan type) - much like that in the Amazon - of continuous rain forest made up of teak, ebony, copal, palm, and cedar trees remains one of the pristine timber regions of the world, although deforestation is now underway. Here in this Landsat MSS false color scene, this status is disclosed by the solid expanse of red (= NearIR) vegetation; usually the region is largely cloud-covered but this clear March 1973 scene is unusual. The large river crossing the scene is the Congo (renamed the Zaire) River; the large tributary is the Aruwimi. The Congo River is 4370 km (2718 miles) in length, making it one of the longest on Earth; it is navigable far upstream.

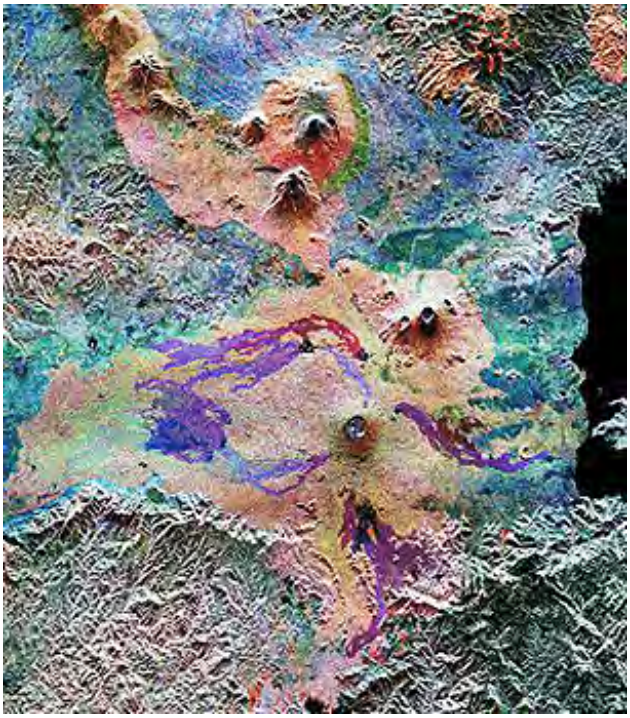
As seen from space in natural color, the Congo jungle looks like this:

The jungle is spread over both mountains and flatlands, as shown in this ground photos:

Two of the largest cities in the DRC are Brazzaville and Kinshasa (the latter is approaching 8 million); both are seen in this photo taken from the International Space Station; an aerial view of Kinshasa is below it.

However, the tropical forests are shrinking and desertification from the Sahara is moving south. Some of this loss is natural but much has resulted from land clearing by the indigenous tribes. Here is a Landsat view of part of Ghana and Togo, in which once dominant rain forest has been removed by cutting and burning, converting the land to savannah. The dark areas are probably burn scars from recent burns.

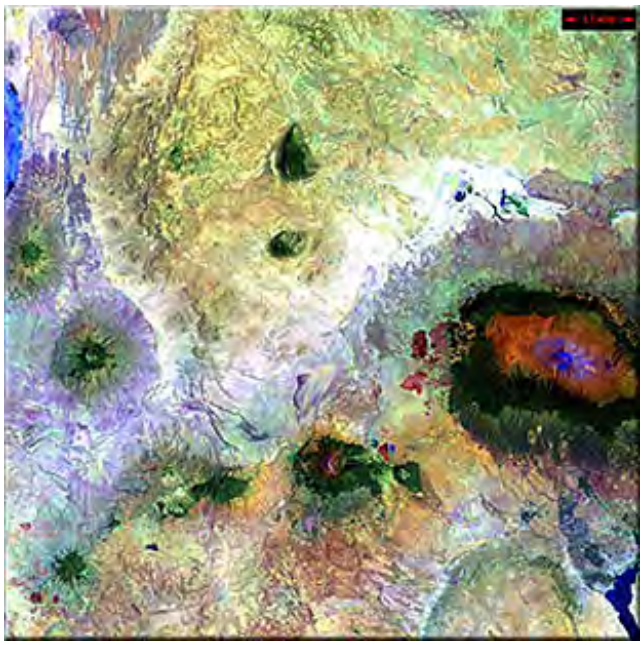
But, the still extensive forestlands offer shelter to many of the indigenous (but to zoo-goers, exotic) animals, such as the African gorilla. Radar imagery helps to penetrate the tropical growth, as shown in this SIR-C image of Lake Kiva and parts of Rwanda, Uganda, and Zaire. One prominent volcano is Mt. Karisembi:



The next Landsat scene (second below) covers part of northern Kenya in East Africa. The dominant feature running north-south through the right half of the Landsat image is the African Rift Valley, shown here in dark bluish-gray representing basaltic rocks. Notable are the many scarps caused by normal faulting as the rift develops and pulls apart. (See page 2-9 for another example of this rift terrain.) Small volcanoes dot the Valley. The lake at the bottom is Lake Natron, whose level varies seasonally, and as it dries leaves sodium carbonate deposits. The lake at the top of the Valley is Lake Naivasha. Relief in the region exceeds 1000 meters (several thousand feet), so that higher elevations remain forested whereas lower areas are near-desert. Rain in the region can be up to 100 cm (40 inches) gives that area enough moisture to support crop farming, including coffee planation.

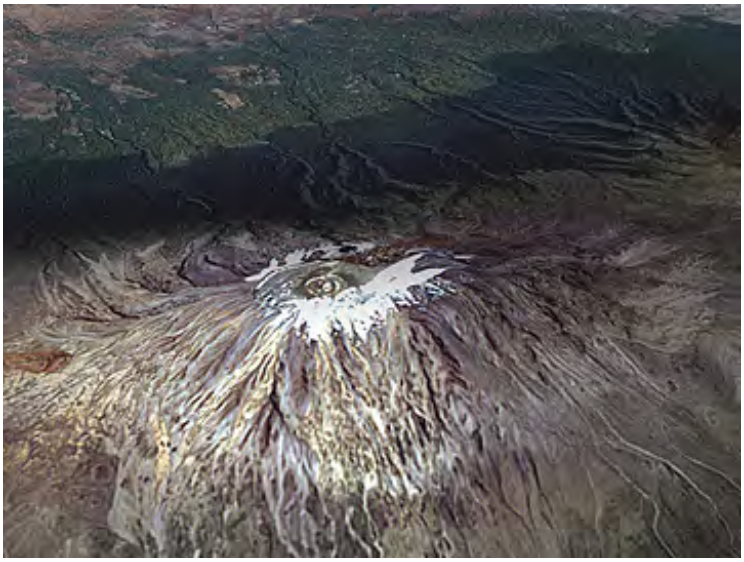
The capital of Kenya, Nairobi, lies at the right margin about one-third down from the top (it is a bluish area midst the red). A SPOT image shows the city; below it is a ground photo that indicates the modern development of Nairobi:

Like all continents, parts of Africa are mountainous as well. You saw the highest peak in Africa, Mount Kilimanjaro, a huge volcano, on the Game page of this Section. Because the coloring is so attractive, we show it (framed by IR red vegetation; near center right edge) again in this Landsat image which covers a larger area:



Kilimanjaro is the tallest peak in Africa (5895 m; almost 20000 ft). It is quite photogenic, as evident from this ground view:

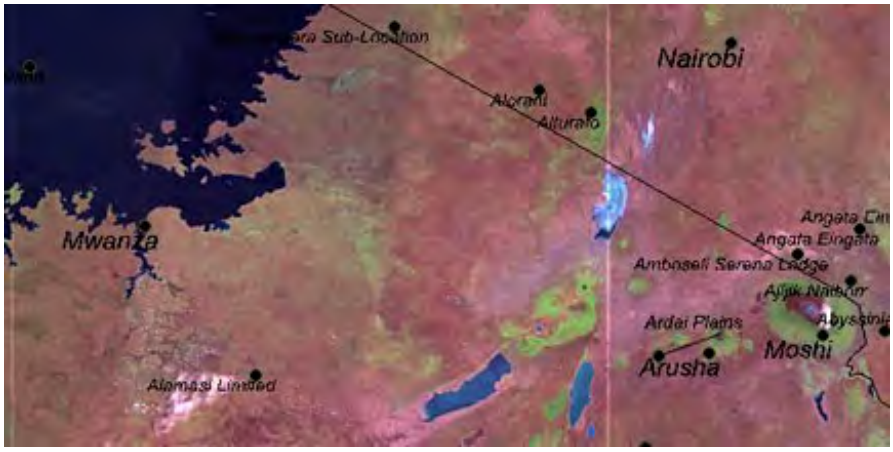
Landsat data have been used to portray this volcano, as a perspective, as though viewed obliquely from above;



Landsat 10 meter ETM+ imagery shows the top of Kilimanjaro from above. Using earlier imagery, the extent to which the summit glaciers (20 individuals) has wasted away since 1962 has been plotted on the Landsat image. Kilimanjaro is an "indicator glacier" used by advocates of global warming to show the effects of temperature rise in the last half century; the famed "Snows of Kilimanjaro" are fast disappearing.

In Tanzania, between Lake Victoria (next page) and the volcanoes west of Kilimanjaro lies the Serengeti National Preserve, one of the most stupendous game lands in the world. The park is largely flat plains, as depicted in this SRTM topographic image; the volcano is Ngorongoro:

This Landsat subimage gives a further impression of the grasslands terrain:



This is a typical photo of the Serengeti landscape.

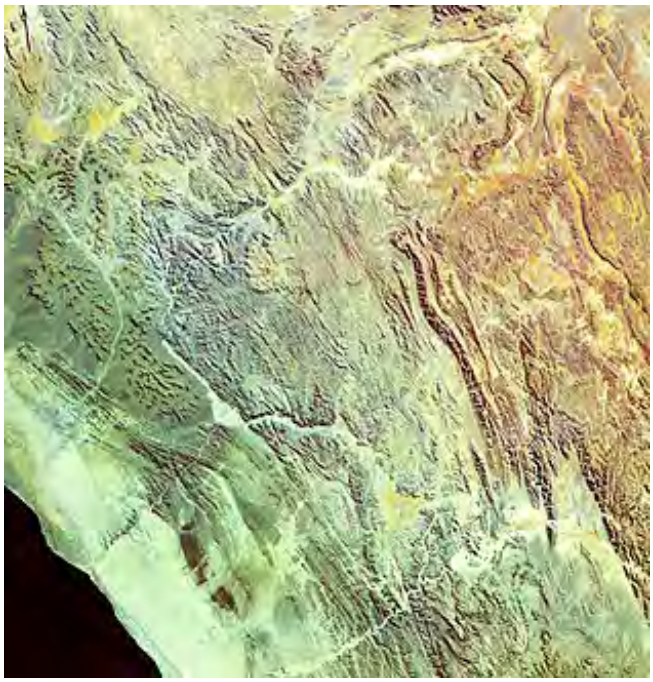
The wildlife is diverse, plentiful, and often spectacular. Lions, elephants, giraffes, hippos, and many other varieties of African animals we have associated with this continent abound. As shown in the next photo, herds of wildebeests, and large white birds (egrets?)attest to the fauna people come to this part of Africa to observe on Safari; the photo below this shows greater diversity.

Within the Serengeti lies an ephemeral river that has carved out the Olduvai Gorge from Pliocene-Pleistocene sediments. Here is the famous location where the Leakey family and their associate anthropologists have made some of the great discoveries of hominids dating back more than a million years. This ASTER space image shows some details of the immediate area:

Small parts of four countries - Zambia, Botswana, Zaire, and Namibia - appear in this Landsat image. The Zambesi River flows through the Kalahari Basin whose vegetation is a savannah grasslands and some swamps.

Typical savannah land cover (here called the Veld) is dominant in this Landsat image of southwestern Botswana. The mottled pattern is largely the result of the annual controlled brush and grass fires set to prepare the land for later planting. The Kalahari Bushmen inhabit the region.

Further to the southwest along the Atlantic is Namibia, a barren land of deserts and treeless mountains. This is the Kaoko Veld region. The eroded fold belt mountains contain rocks of the Damara and Karoo systems:



The capital of Namibia (once known as South West Africa) is Windhoek, seen in this SPOT image:

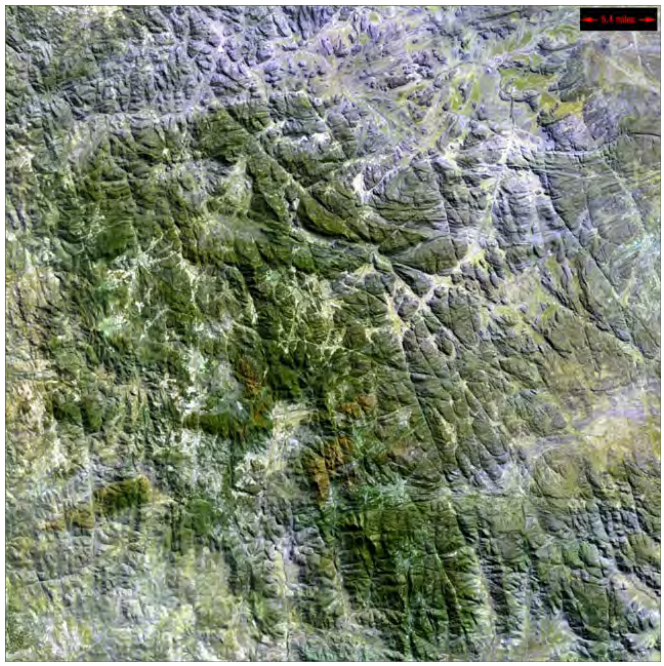
Like so many other African cities, Windhoek would strike the viewer as modern:

South Africa is one of the more advanced countries on the continent. Here is a map of its main geographic features; most of interior South Africa is a plateau surrounded by a long continuous

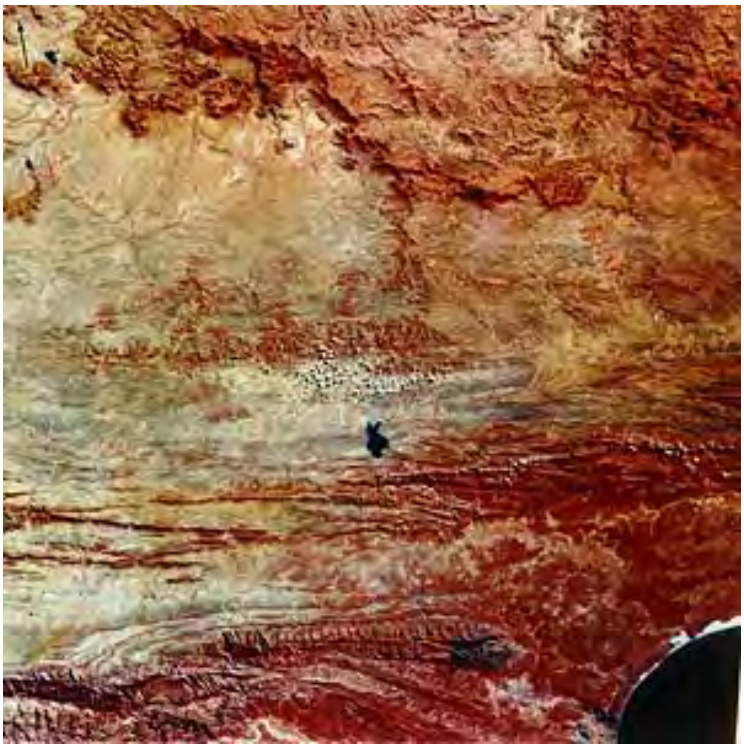
escarpment:

The northwest region of South Africa is a continuation of the Namibian Desert as seen in this image:

In the above scene are a cluster of low mountains in the eastern part of Namaqualand, seen here in this ASTER image; they consist of crystalline rocks, a mix of granitic and metamorphic rocks and basalts, mostly dark in color:

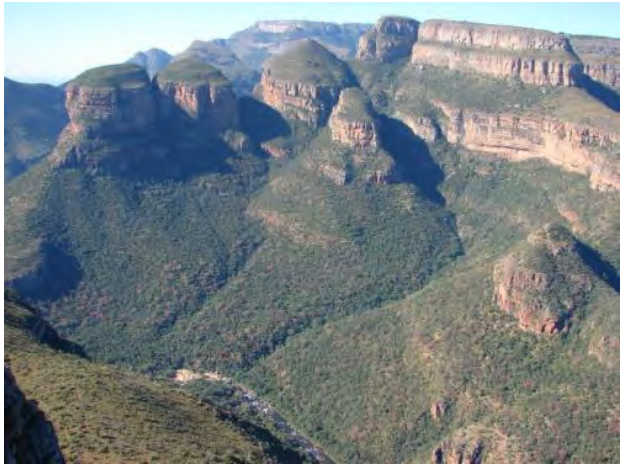


In this next scene, near the coast in southeastern South Africa, is another fold belt known as the Cape Ranges, a collection of anticlines and synclines running about east-west, and made up of Paleozoic rocks. In places, their elevation exceeds 1500 m (5000 ft). The dark reds associated these mountains imply forests that include evergreens. They block winds coming from the Indian Ocean (lower right) so that precipitation is "wrung out" over them (up 130 cm or 50 inches annually), leaving the interior (left) in a rain shadow that fosters a semi-arid desert flora. Along the top of the scene is the Great (or Drakensberg) Escarpment, a steep cliff nearly 1.5 km (1 mile) high on top of which is a vast peneplain referred to as the Post-Gondwana Surface, dating from the Cretaceous Period. Elevations above the escarpment reach to 2400 m (8000 ft). The region is harsh for living, with low populations.



The Drakensberg Escarpment in many places is topped by basalt flows which are also jointed, as indicated in this SPOT image:

Here is a ground photo of the Escarpment:



A continuation of the Cape Ranges carries one to Cape Town and the southernmost tip of Africa - the Cape of Good Hope - as seen in this Landsat view:

Compare this overhead scene with a close-up ground photo of Cape Town

Nearby are these imposing cliffs composed of basaltic rocks:

Next we show part of a MISR image that features the largest city in South Africa, Johannesburg, which lies in the heart of the Witwatersand - which contains the main gold-producing fields (numerous underground mines) on the continent:

Compare this with this photo taken by astronauts aboard the International Space Station:

Johannesburg, the largest in southern Africa, is a distinctly modern city:

Just to the northeast of Johannesburg is the South African city of Pretoria, capital of the country, as seen in this Landsat-7 image.

This Proba CHRIS image provides a more detailed look at Pretoria from space. The city is also involved in gold mining. Beneath it is a panoramic view of the city (the purple tree is the Jacaranda):

Now, to the Indian Ocean side of Africa.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Asia's landmass area is the largest of the continents (its western end continues into the European continent); this combined land is arbitrarily divided from Europe along a line set just east of the Ural Mountains, runs south to Magnetogorsk, then moves west until it turns south again to the Caspian Sea, and passes Europe at Turkey (and Africa at the Middle East). Its landscapes are also extremely diverse as its climates range from arctic through arid to tropical humid. Asia contains the most dynamic mountain system on any continent.

Asia

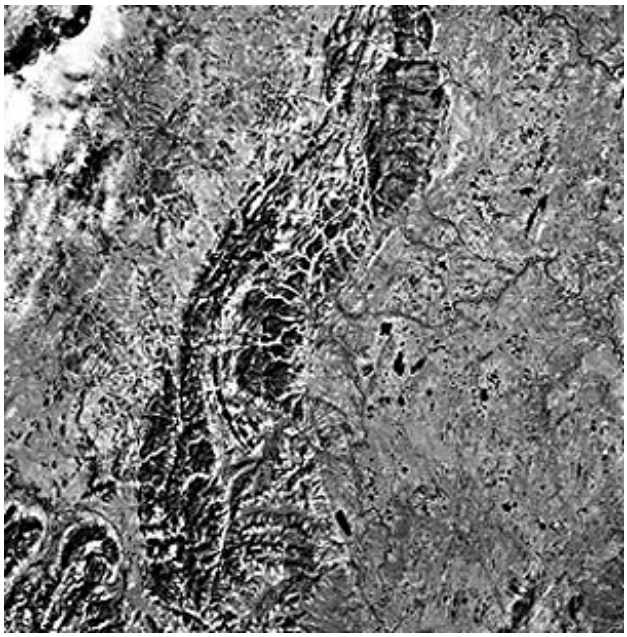
Parts of Asia have been imaged prior to this Section, namely Tokyo (page Intro-23), Beijing and Shanghai (page 4-4), Iran (page 2-6), China (page 2-7), Iran-Pakistan-India (page 7-3), and elsewhere.

An idea of just how vast Asia is as a landmass becomes evident in this relief map that shows its extent and the major mountain systems in various regions within it:

The current geopolitical units (countries) in Asia are named in this map:

By convention the land designated now as Russia consists of European Russia (to the Urals) and Asiatic Russia. The largest expanse of land in the world that is a single political unit is Siberia, the huge tract that makes up most of Russia. Siberia is sparsely populated relative to European Russia, owing mainly to its harsh winter climate. It is one of the richer sections of the Earth's crust in mineral wealth, oil and gas, and timber. Three types of landscapes/ecosystems are predominant: the lowlands steppes, the subarctic tundra, and vast stretches of mountains in its east and along the south. Because of its size, we show only a few examples. First, take a look at a map of Siberia and note the main cities:

As said, the boundary between European Russia and Asiatic Russia (Siberia) is arbitrarily set at the western edge of Ural Mountains. This north-south range of folded Paleozoic rocks rises abruptly from flatlands on both sides:

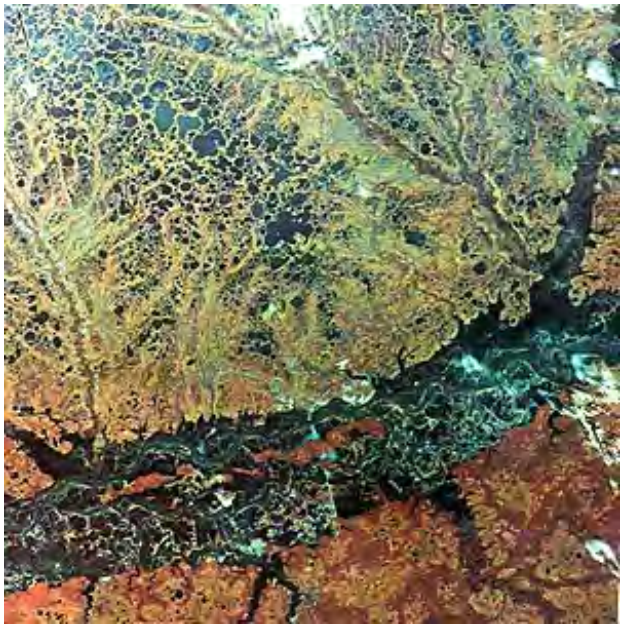


The steppes of Siberia are vast grass-covered plains with low relief. This is a characteristic ground scene:

Typical flat and largely barren steppes are found around the city of Kurgan, east and south of the Urals. In this May 1972 Landsat image, the dark gray areas have yet to be covered with vegetation since snow cover had only recently been melted. But some areas have begun to produce spring vegetation.

The second Landsat scene shows the westward flowing stretch of the Ob River, in the western Siberian Lowlands, about 500 km (300 miles) east of the Urals; as it moves further west it will turn north into the Kara Sea above the Arctic Circle. Its overall length is more than 4000 km (2500 miles). This meandering river now is in flood (June) after spring snow melt. The myriad of lakes in the upper half are formed as sinks owing to poor drainage in the underlying glacial tills.

The entire region lies within the taiga forest zone, consisting of Siberian fir, stone pine, larch, and spruce. Surgut is the only town of any size in the image.

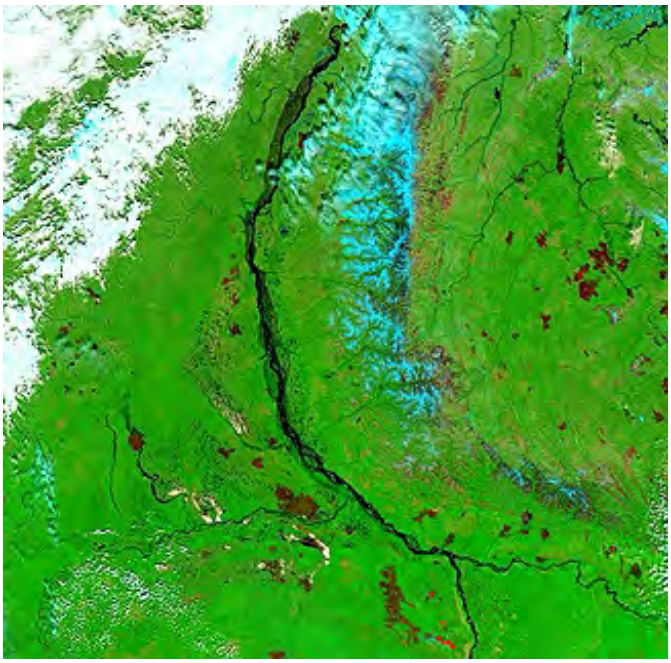


Two major Siberian cities are situated along the banks of the Ob. The first is Novosibirsk, the third largest city in Russia. To its west is Omsk. Both are near the border with Kazakstan - thus in southern Siberia.

The Yenisey River east of the Ob appears here in this Envisat MERIS image. Below it is the small city of Noril'sk, in the far north, which is a center for platinum and nickel mining.

Much of Siberia in its northeastern extensions is heavily forested. This boreal forest is also known as a Taiga biome. Here is a photo showing this forest:

From space the boreal forest in much of Siberia has the same uniform density - showing as continuous green in natural color - as do forests in the Amazon and Congo. Here is part of the eastern end of the Central Siberian Plateau, drained here by the Lena River, near the city of Yakutsk. The dark red patches are forest fire scars.



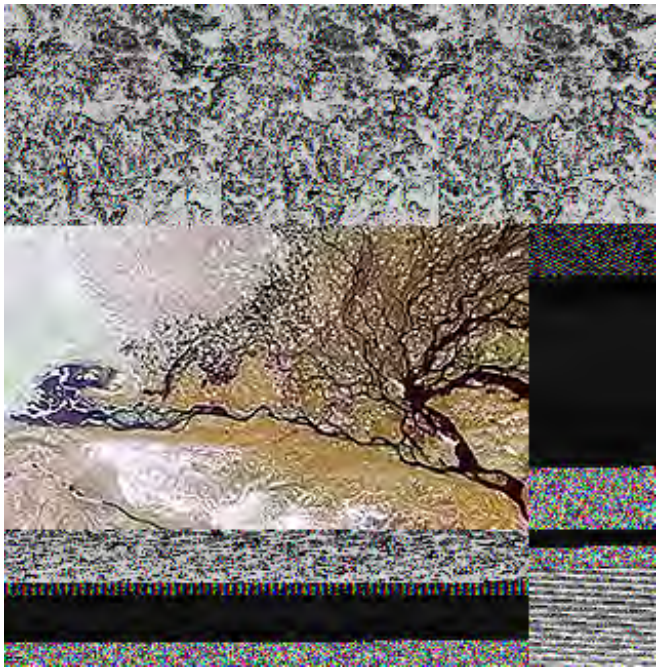
Here is a view of Yakutsk itself, the largest city in mid-eastern Siberia, as made by the AVNIR sensor on the Japanese ALOS satellite:

The Lena River, like the Ob and Yenisey Rivers, drains into the Arctic oceanic waters (the Lena into the Laptev Sea; the other rivers into the Kara Sea). The Lena forms one of the most spectacular, from space, deltas in the world, as shown here (another satellite image appears on page 17-4):

Parts of northern Siberia are treeless, being similar in landscape to the tundra of Alaska. This is the permafrost ecosystem of the Arctic:

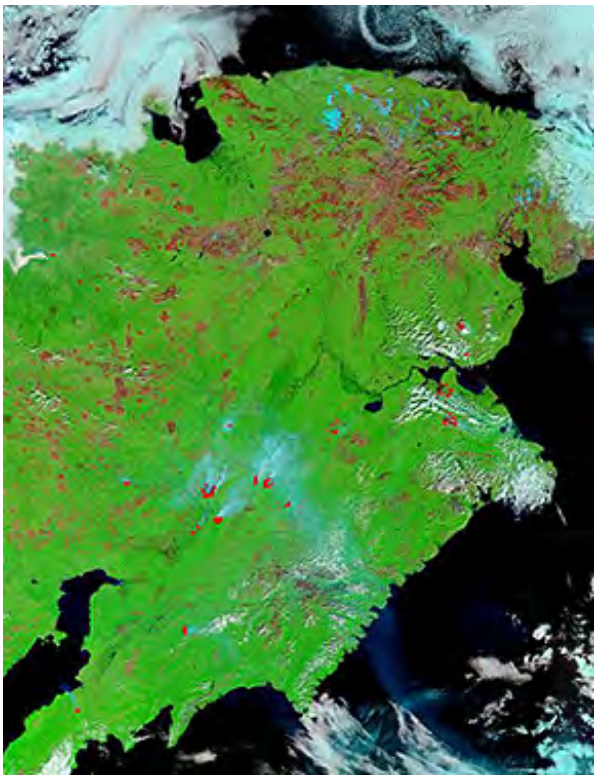
The aforementioned Central Siberian Plateau is a wide platform of uplifted rocks including some

that are folded; volcanic lava flows are also present. This Landsat black & white image shows a region of dissected rocks that form low broad hills, with maximum elevations around 700 meters (2300 ft). These rocks are mainly basalts, the so-called Siberian Traps of late Paleozoic age. The left-right river (white here because it is ice and- snow-covered) is the Nizhnyaya Tunguska, which flows into the Yenisey.



In the far eastern reaches of Siberia, mountainous terrain predominates. Here are mountain-like hills and divides on a rolling plateau surface etched by past glaciation and current stream erosion; these extend from the Chersogo Mtns just to the south. Already, by this October 28th, 1972 date, the entire region is snow covered. The main drainage path is the Indigirka River (lower left), into which flow the Ulakhan (mid-left) and Nera (upper left) rivers.

The northeastern end of Siberia is largely a wilderness, with only a few towns along its coast. This image shows it to be, like the interior, mostly forest-covered. The reddish patches are fire scars; several active fires are indicated by smoke:



In southern Siberia is Lake Baikal, the largest freshwater body in the world - 23000 cubic kilometers (5520 cubic miles), holding 20% of all such water on the continents (other than the Antarctic which holds more than half the world's drinkable water). Its depth reaches 1620 meters (5310 ft), also a world record. The city of Irkutsk (next paragraph) is to its west; it shows in the image below as at the end of the wide part of the Angara River near center top. Only the southern half of Lake Baikal is shown here:

Most westerners think of Siberia as desolate and in places almost uninhabitable. While there are vast stretches that are sparsely populated, some larger towns and a few cities are growing as the Russian people are encouraged to resettle in Siberia in an effort to better exploit its natural resources. About 100 km (62 miles) north of Lake Baikal is the city of Irkutsk, on the Trans-Siberian Railway. This astronaut photo shows its extent. It is the source of much of the electric power in that part of Siberia, generated at the Dam shown:

In the southeasternmost part of Siberia is the city of Vladivastok, home of the Russian Pacific fleet and a major shipping port. It was lost to Russia after the Japanese-Russian War of 1905. Today it is the eastern terminus of the world's longest (more than 8200 km [5000 miles]) railway - the Siberian Express.

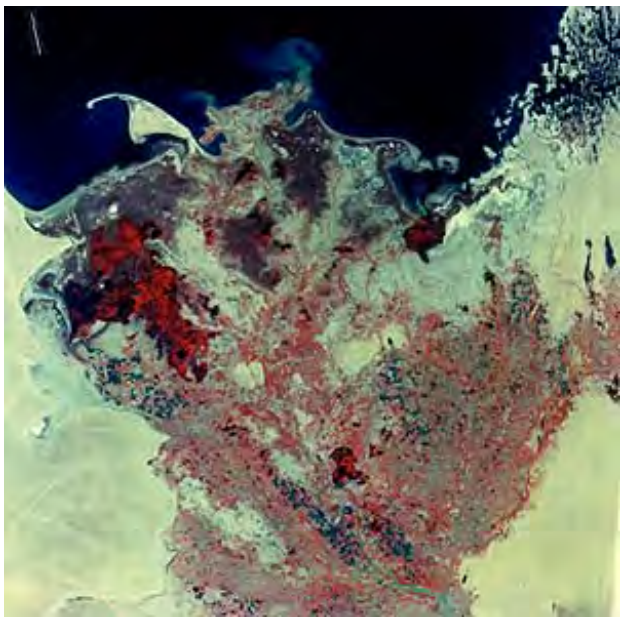
The eastern end of Siberia includes a long land extension known as the Kamchatka Peninsula. Here we see it as rendered in a topographic image as made by the SRTM (Shuttle Radar Topography Mission) software that utilizes radar data to extract elevations (see page 11-10; that page has other Kamchatka SRTM images).

The Kamchatka Peninsula has more than 150 volcanoes, many erupting frequently, making this one of the most active volcanic regions worldwide. This STS 53 astronaut photo shows several that have erupted in the 20th Century, including Kronotskaya (which rises 3528 meters [11570 ft])

Southern Siberia joins many countries owing to its wide E-W length. Steppes give way to semi-desert lands as seen in this image.

Now we move from Siberia to South Asia. Parts if the old Soviet Union became independent countries in the early 1990s. These are the "stan" nations, shown in this map (Afghanistan and Pakistan were not part of the Union):

The image below is part of the Kyzul Kum desert region of Uzbekistan, one of the Muslim countries loosely tied to Russia. The region is north of Afghanistan, in southwest Asia. The major river, running up through the center of the image, is the Amu Dar'ya, the longest (2300 km; 1440 miles) in this part of Asia, and is noted for carrying the heaviest sediment load (derived from the Tian Shan mountains) of any major river in the world. This load is carried into the Aral Sea (top) which actually is a large lake, slowly evaporating so that its maximum depth now is about 25 m (80 ft). The river has built a very large delta on which cane thickets and woody brushlands are widespread. Marshlands are indicated by the deep reds. Some farming occurs on the delta but is isolated owing to frequent flooding and is most prevalent where irrigation ditches have been dug. The swarm of sandy "islands" in the upper right are dunes now dissected and submerged by locally rising waters.



The capital of Kazakhstan, in one of the breakaway Republics from the former Soviet Union, is Almaty (formerly called Alma-Ata). This is a city of more than a million. It lies between the southern edge of steppe agriculture and a small mountain range:

In the city of Komsomolets, in Kazakhstan, is an interesting urban land use pattern. This high resolution image shows that there are walls around various blocks:

Another former Republic is now Kyrgyzstan. Its principal city is Bishkek. Here is a Quickbird image of central Bishkek. This shows that most of the buildings are apartment complexes. During the Soviet era the housing policy favored putting most of the citizens in such complexes.

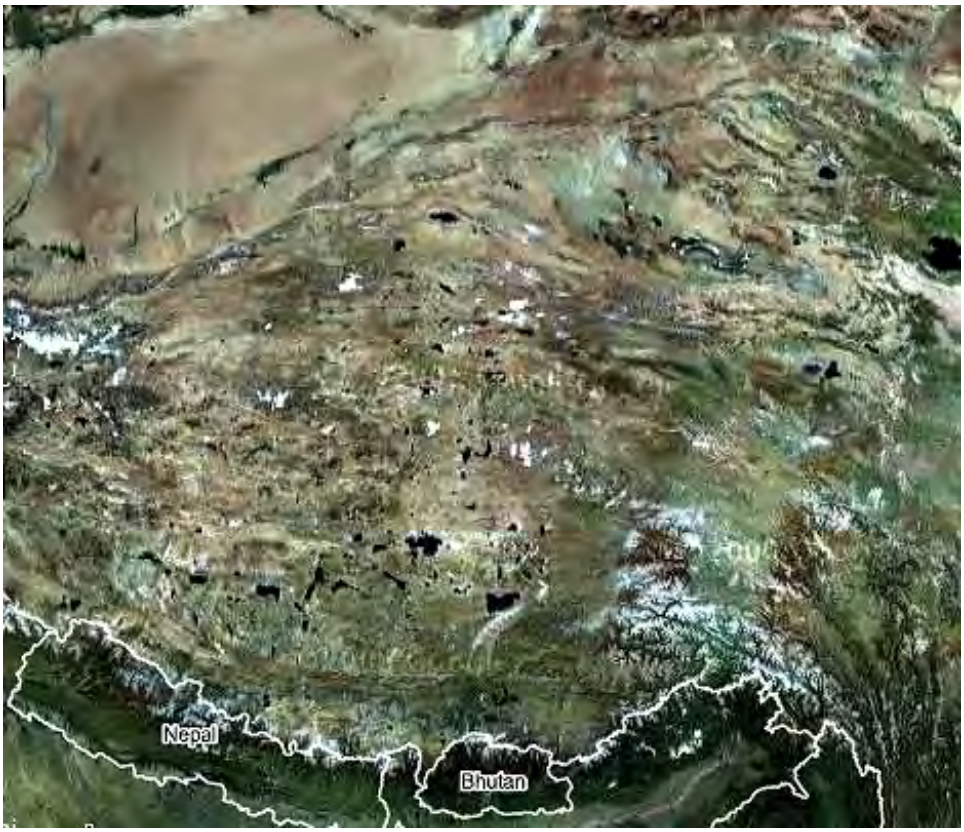
Here is a satellite image of Turkmenistan, on the Caspian Sea, and surrounding countries. Although largely arid, this country owes its limited prosperity to extensive reserves of oil and natural gas.

As the writer (NMS) surfs the Internet daily in search of more material to add to the Tutorial, every once in a while an image is found that is intriguing by itself and, while not always germane to the topics, worthy of putting online with comment. Such is the next image, a Landsat-7 view of the capital of Turkmenistan, Ashgabat, a SPOT image beneath it, and a view of the city from the ground.

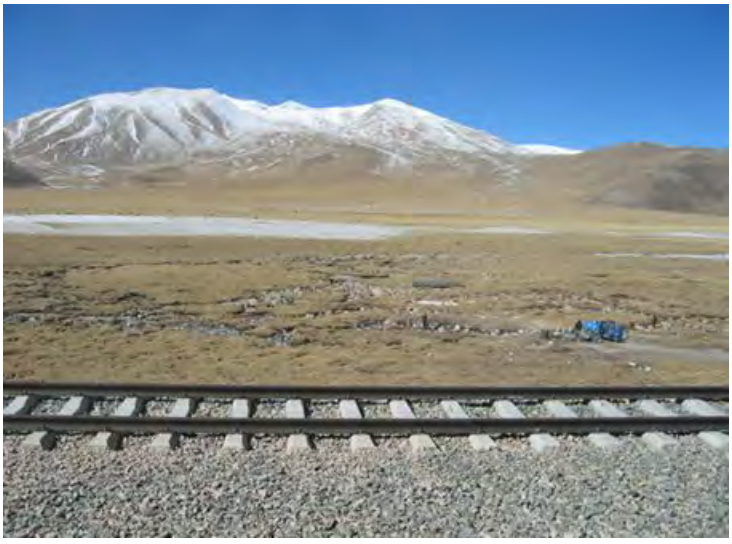
Until the day the above was put on line, the writer (NMS) had never heard of this city of more than 650,000. But I learned a lot by accessing this Wikipedia website. The city was first established in the 19th Century. In 1948 it experienced a 7.6 magnitude earthquake which killed upwards of 100,000 people.

We have seen images of Afghanistan before, in the Overview and elsewhere. Here is another, a photo taken by an astronaut during the Apollo 7 mission. The extreme ruggedness of the terrain offers strong evidence for why it has so far proved so difficult to find and capture Osama bin Laden, the terrorist behind 9/11.

The next two scenes begin our introduction to what is literally "The Top of the World": The Himalaya Mountains, highest on Earth, and the highest flatlands (up to 5150 m; 17000 ft) on our planet making up the Tibetan Plateau to the north. The top scene covers these two features plus parts of China (Sinkiang Desert) and the Indian subcontinent. The lower scene Terra image shows much of both high altitude topographic features:



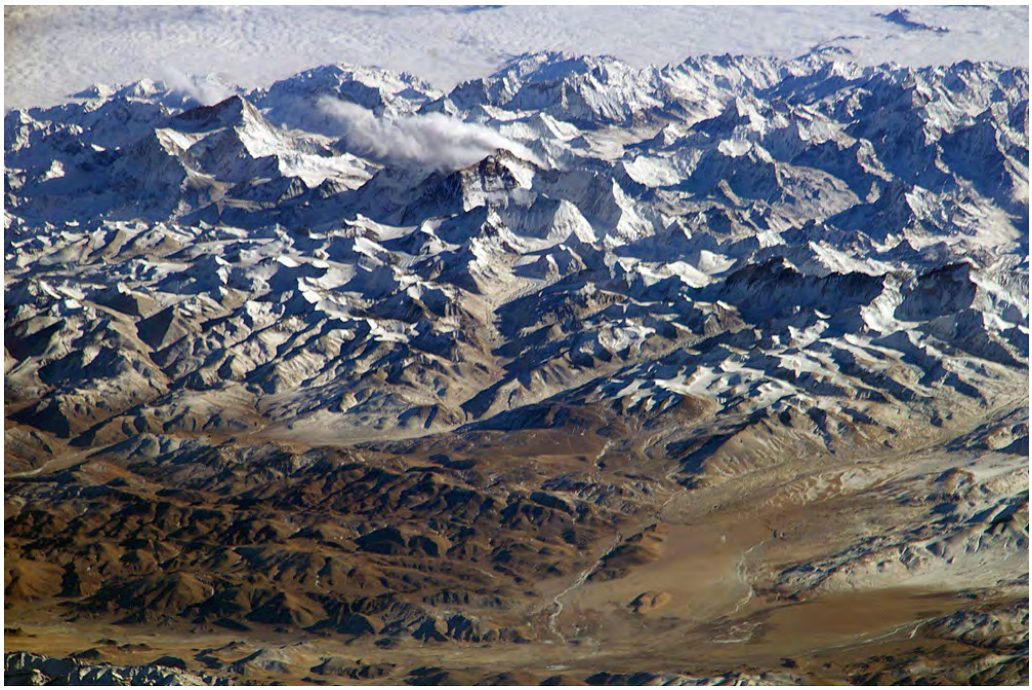
Vegetation is notably sparse over most of the Tibetan Plateau, as the Indian monsoonal rains have fallen when the air rose over the Himalayas, leaving the plateau with little available moisture (but some lakes survive). The bleak landscape is broken by occasional mountain blocks most of which are not particularly high (they're hard to find in the MODIS image). Over this landscape passes the highest railroad in the world (more than 80% is above 4000 meters; oxygen is available on the train). The Qingzang Line begins in Xining, passes through Golmud, and ends in Lhasa, the main city in Tibet. This is typical scenery:



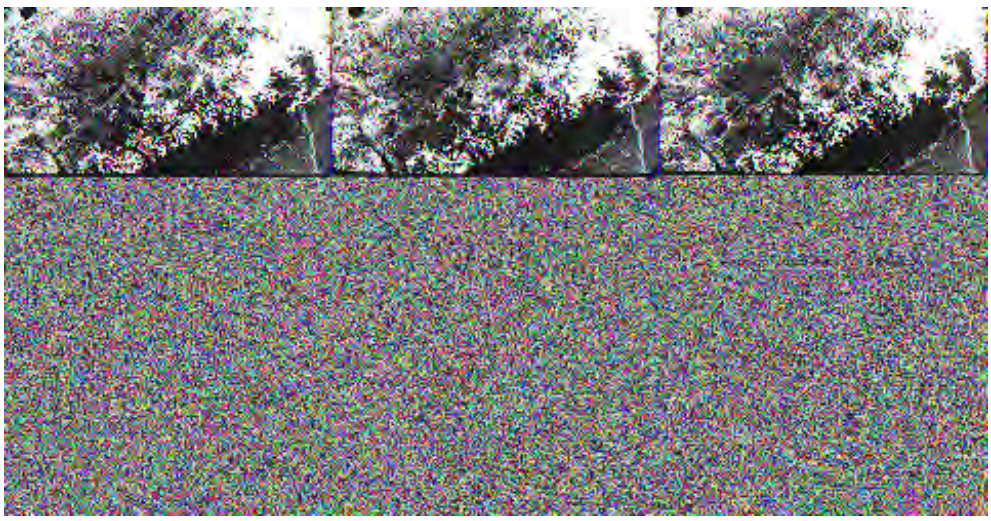
The capital of Tibet is Lhasa, located within the Himalayas, seen here from space in an ASTER image restructed into a perspective view:

We remind you with this illustration (seen before in Section 2) that the Himalayas are the result of a massive collision by the northward migrating Indian subcontinent against the offshore sediments and land rocks in the "underbelly" of Eurasian plate along South Asia. The result has been a mix of crumpling and uplift of these rocks within the Himalayas and a rise of the Tibetan Plateau beyond as the front edge of the Indian plate subducted forcing the plateau rocks upward.

The grandeur of these mountains has been captured in these two photos taken from the International Space Station from a height of 120 miles. The first looks to the north, over the foothills, and includes Mt. Everest on the far right top. The second looks from the south with the seventh tallest mountain in the Himalayas, Dhoulogiri, (labeled); the Tibet Plateau is in the middle ground.

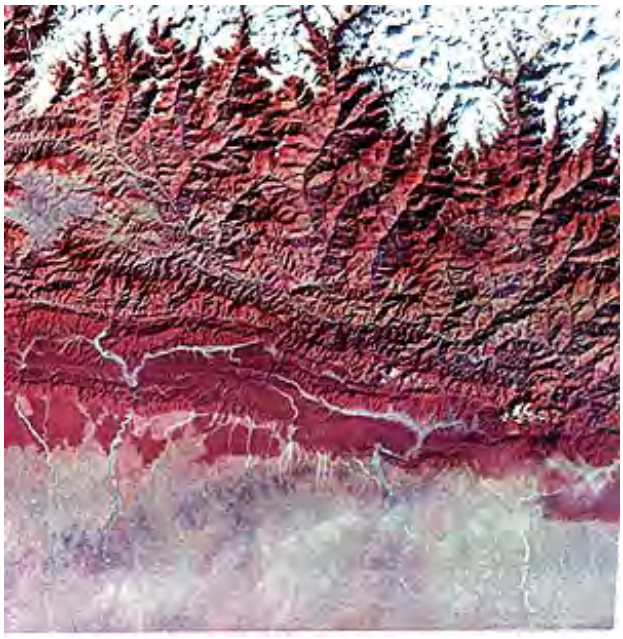


This is a good time to introduce an odd-shaped image, made by the Large Format Camera (LFC) flown on one of the Space Shuttle missions (the camera and mission will be reviewed in more detail on page 12-4). What you see below includes an area farther west and north of the above Landsat scene.

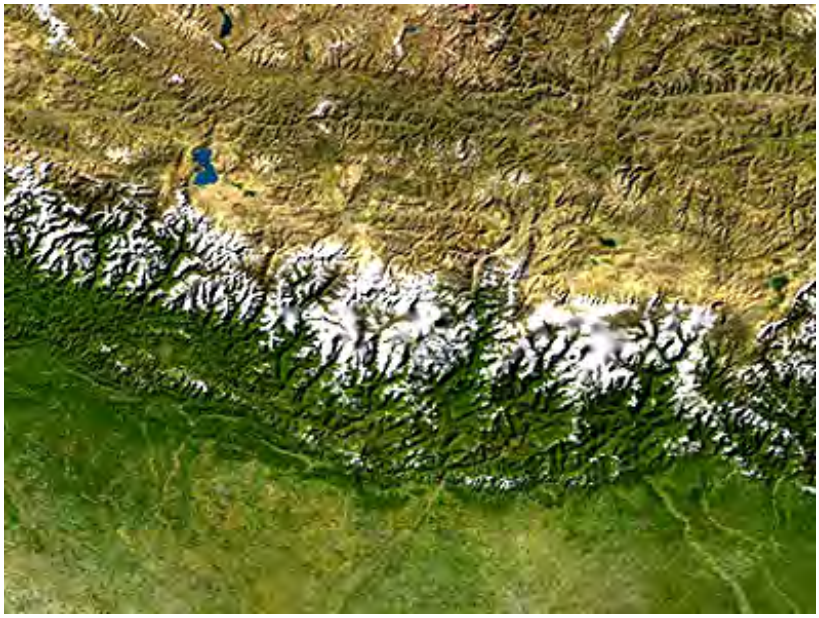


The Himalayas occupy nearly all of the country of Nepal (a bit of northern India is present at the bottom of this next Landsat image). Its capital, Kathmandu, is visible in the valley above the left

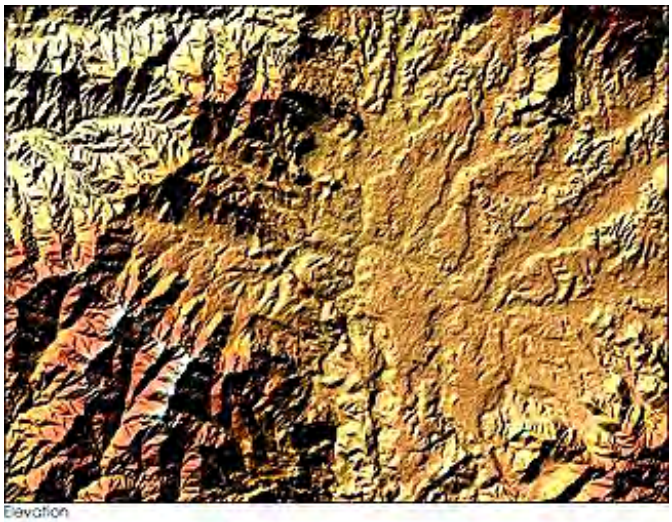
center edge. Going northward from the bottom, one passes across the High Plains of the Ganges to a line of dissected gravel deposits, known as the Siwalik Hills (elevations up to 1300 m [4300 ft], carried down from the high mountains during active uplifts in the later Tertiary. Their deeper red color indicate subtropical forests of bamboo and other vegetation. The relief becomes strikingly rugged in the Lesser Himalayas (3000 m [10000 ft]), that continue to rise towards the crest region of the High Himalayas (6000-8800 m [20000 to 29000 ft) marked by snow cover in this December image. Mt. Everest (8848 m [29030 ft]) does not stand out from neighboring peaks; it is near the upper right corner. Surprisingly, snow is largely absent from the intermediate heights, owing to the drying out of monsoonal rain clouds that have crossed the Indian subcontinent.



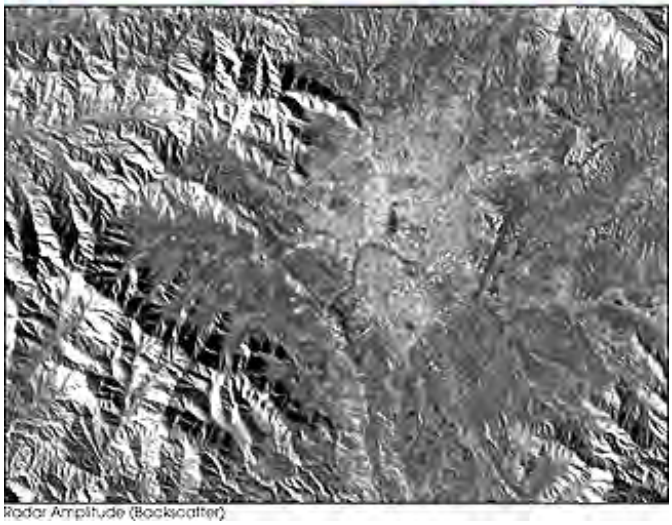
This next Landsat mosaic is a natural color rendition of much the same region as seen above, but extending over a wider area:



Most expeditions to Mt. Everest begin at Kathmandu, the capital of Nepal. Here are a pair of images made from Shuttle Radar Topographic Mission (SRTM) data in which the bottom is a black & white radar image in which Kathmandu is seen as a darker area in a valley and the top shows the topography of its surroundings, as produced using several SRTM bands and a Landsat subscene:



Elevation



Radar Amplitude (Backscatter)

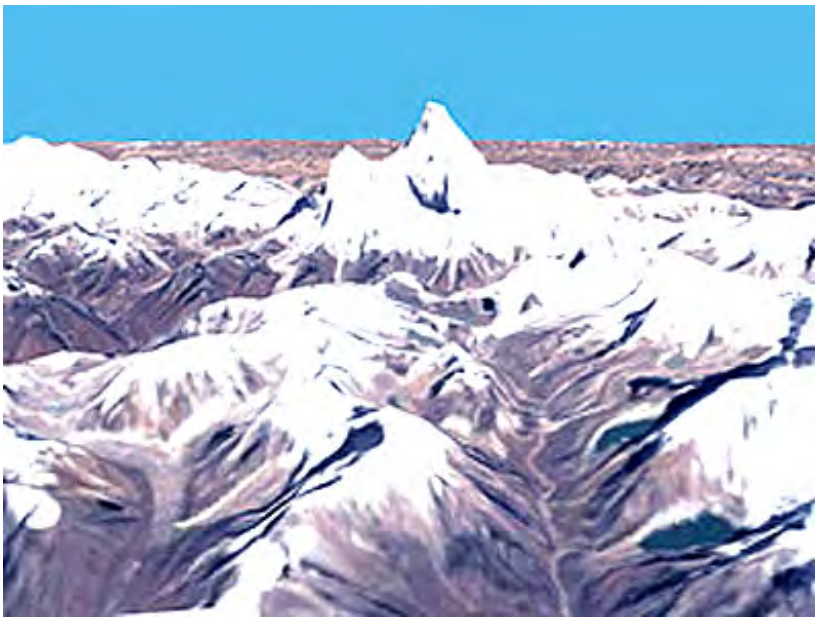
Kathmandu is an exotic place, in recent years becoming a popular tourist destination. Here is a closer look, in an astronaut photo:

Kathmandu is a city containing numerous apartments. It also has temples, both Buddhist and

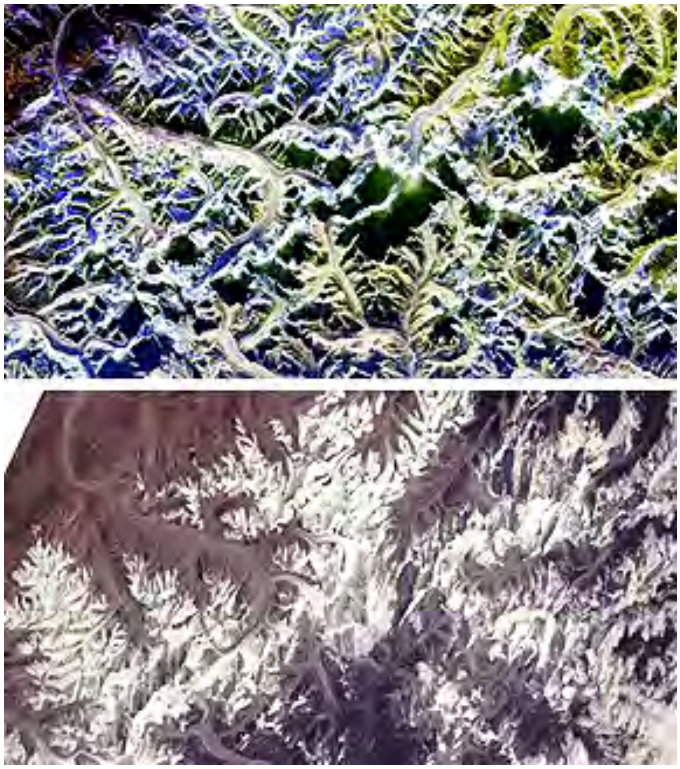
Hindu:

When one thinks of the Himalayas, often the name "Mount Everest" flashes through the mind. This tallest (28030 ft) peak on Earth is the ultimate goal of intrepid mountain climbers (more than 200 have died in their attempt to reach its summit). It lies on the border between eastern Nepal and Tibet (annexed by China). Here is a photograph of Everest, looking southward at the North Face which shows the mountain at its most challenging:

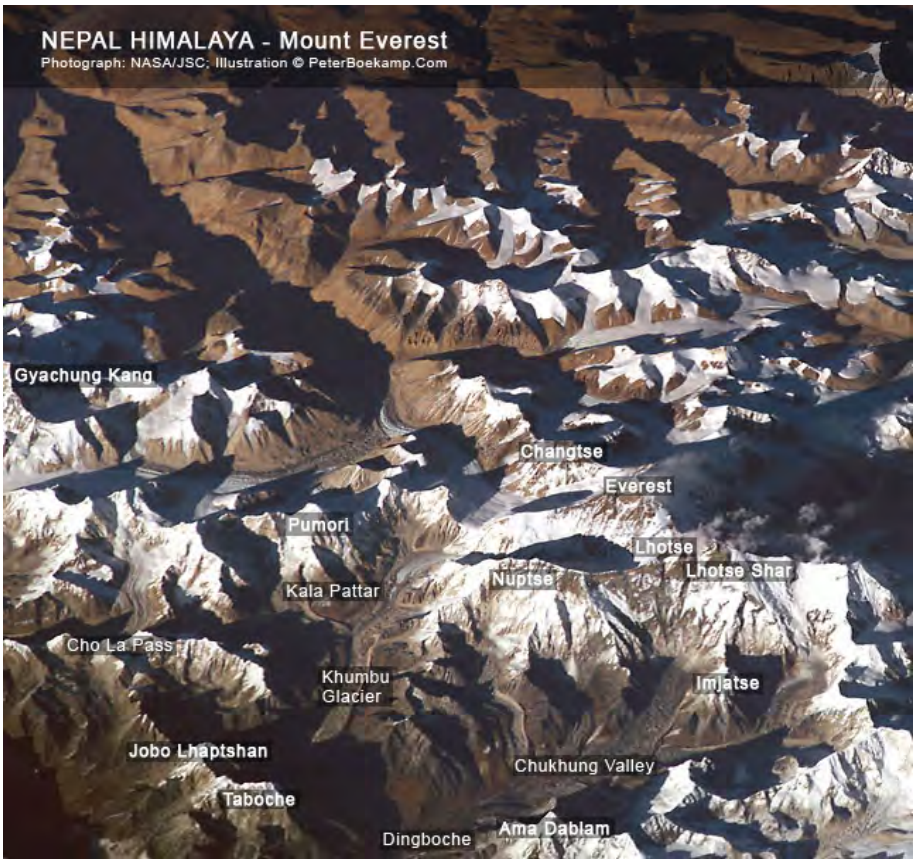
Compare this photo with a view of Everest made by combining SRTM elevation data with a Landsat subscene, in which Everest is even more prominent because of vertical exaggeration:



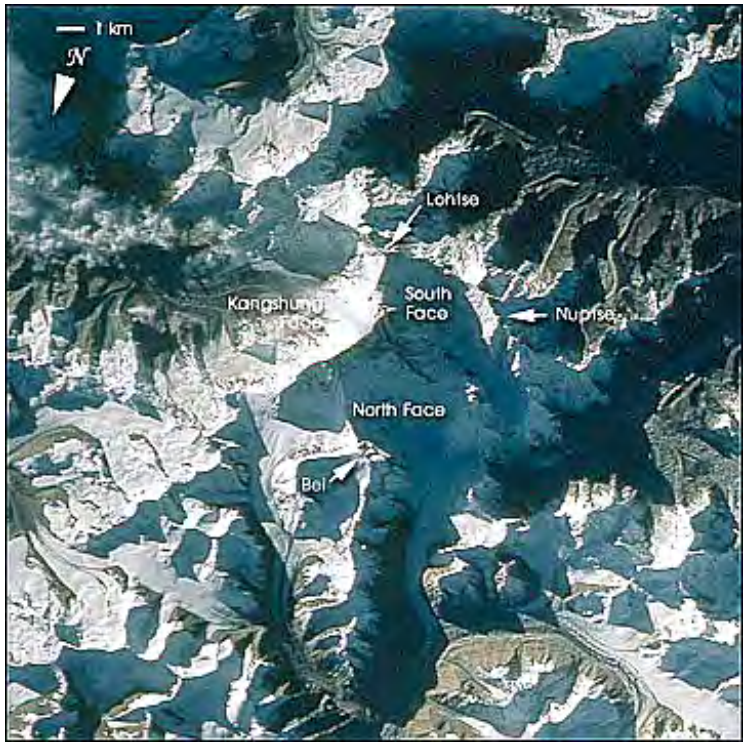
The best reason for trying this daring feat was given by George Mallory's famous (and profoundly simple) dictum: "Because it is there". (Mallory, a Brit, failed to reach Everest's top in a 1924 climb, dying somewhere on the upper slopes; his frozen body was discovered in 1999.) More than 1400 have since scaled it following the first successful try by Sir Edmund Hillary (July 3, 1953) and his Sherpa, Tenzing. Here are two views: the top a SIR-C radar image that brings out the rugged topography; the bottom a Landsat view:



This astronaut photo shows Everest and other named peaks:

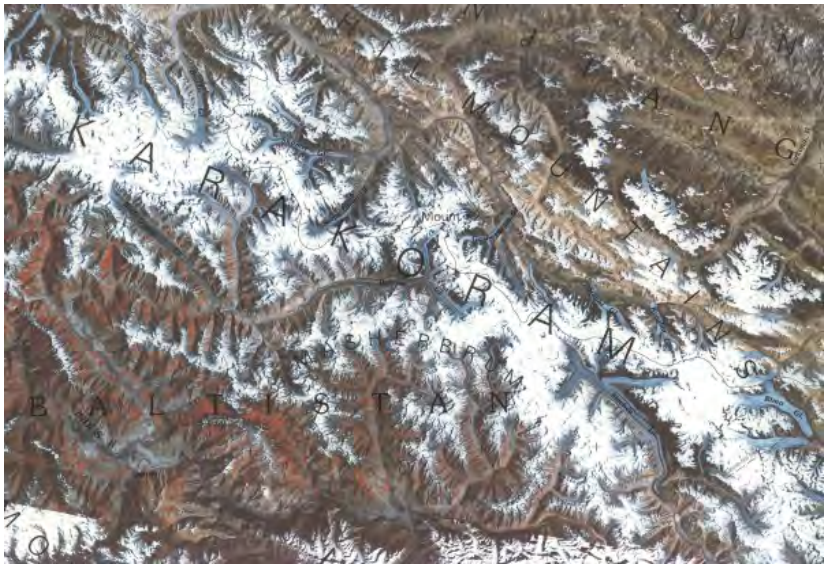


Recently, the IKONOS multispectral sensor made a notable image at 4 meter resolution that includes Mount Everest. It is the triangular-shaped feature just above the center; with this pattern, go back to the previous figure to locate the peak using Lhotse as a guide, keeping in mind that the IKONOS image is "upside-down" relative to the SIR-C and Landsat images.

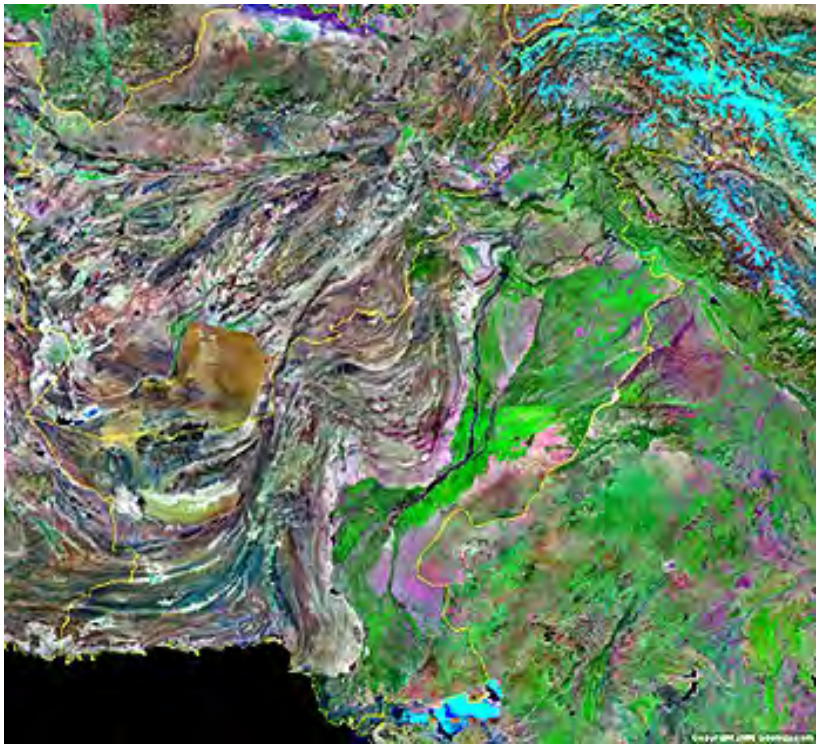


The Himalayan chain beckons mountain climbers from across the world. Some think that K2 (also known as Mount Godwin-Austen, named after a British Explorer) is the most difficult peak in the world to climb. The top was not reached until 1954. The mountain is the second highest in the world (8611 m; 28251 ft). It lies just inside Pakistan in the Karakorum extension of the Himalayas. These illustrations tell the story in their captions:

Spot elevations in metres
Intermittent streams
Dams
Glaciers
Sand areas
Elevation
metres
feet
5,000
16,405
3,000
9,843
2,000
6,562
1,000
3,281
500
1,640
0
0
TURKESTAN MTS.
TRANS-ALAY RANGE
ZERAVSHAN RA.
GISSAR RANGE
Lenin Peak 7134
Communism Peak 7495
SARYKOL RA.
Dushanbe
TAJIKISTAN
PAMIRS
Revolution Peak 6974
Vakhsh
Panj
Karl Marx Peak 6726
Mayakovsky Peak 6095
Mintaka Pass
Khunjerab Pass
Shimshal Pass
CHINA
Mazar-e Sharif
Kondūz
Balkh
Qondūz
KHVĀJEH MOHAMMAD MTS.
Darkot Pass
Mt. Tirich Mir 7690
Ghizar
KARAKORAM RANGE
K2 8611
Gasherbrum 8068
Mt. K'o-li-ya 7120
Kasakoram Pass
35°
TORKESTAN MTS.
Sālang Pass
Dorah Pass
Shandūr Pass
GHIZAR VALLEY
BATURA MTS.
Nānga Parbat 8126
JAMMU AND KASHMIR
SAFĪD MOUNTAIN RANGE
BĀBĀ MTS.
Amu
Mt. Fulādī 5135
HINDU KUSH
Kabul
Konar
Chitral Pass
Indus
Tarbela Reservoir
Srīnagar
LADAKH RA.
CH'IANG-CH'EN-MO
TOWR MOUNTAINS
BĀYAN MTS.
Farāh
Kermū Pass
Jalālābād
Peshāwar
Khyber Pass
Islāmābād
Rāwalpindi
VALE OF KASHMIR
AFGHANISTAN
PAKISTAN
Jammu
Lake Īstādeh-ye Moqor
Indus
THAL DESERT
INDIA
Lahore
Qandahar
0 100 200 mi
0 100 200 300 km
©1997, Encyclopædia Britannica, Inc.
70°
75°
Controlled by Pakistan; claimed by India.
Claimed and controlled by India; status disputed by Pakistan.
Controlled by China; claimed by India.


KARAKORAM
BALTISTAN

Before entering India again, lets first look at Pakistan in this color mosaic:



The mosaic indicates the eastern half of Pakistan is a broad plains made by the Indus river. The western half is mountainous and sparsely populated. Here is a Landsat image that expresses the terrain dominated by the mountains:

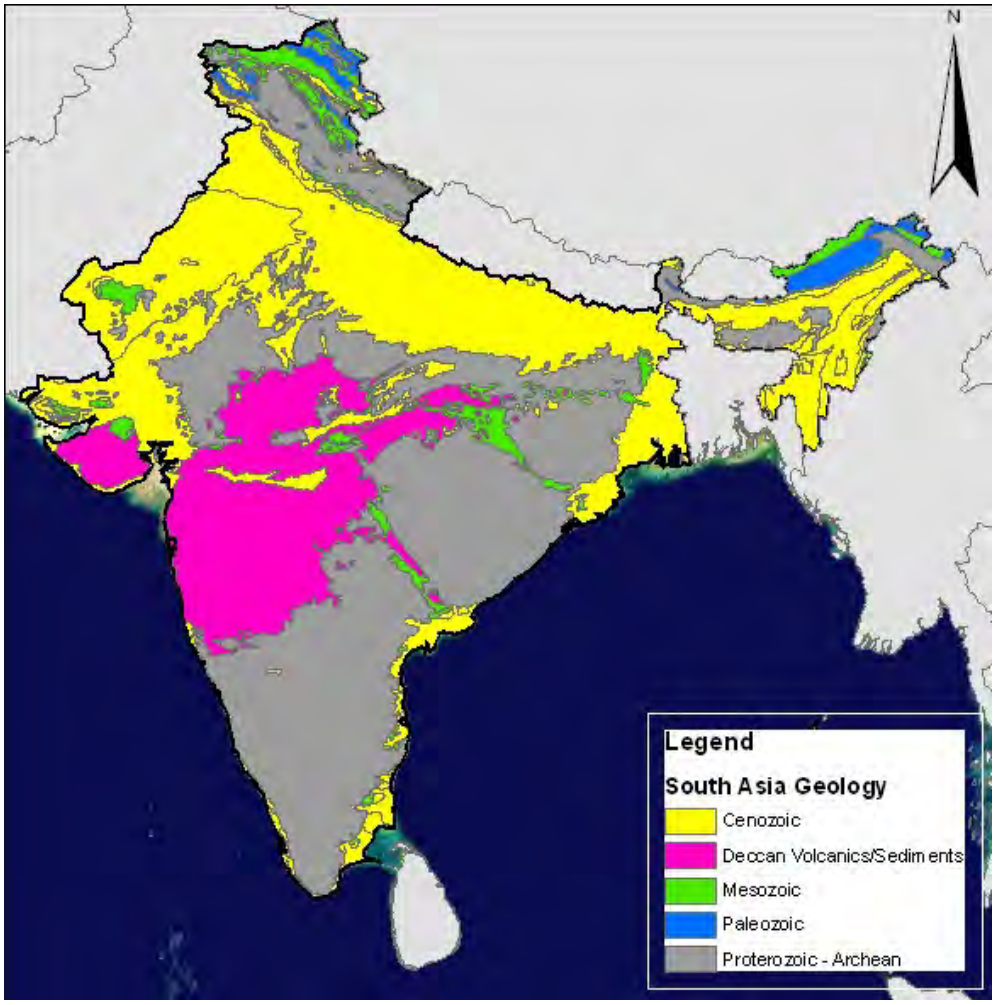
The capital of Pakistan is Islamabad, which is next to the older Rawalpindi:

Karachi, Pakistan is on the Indian Ocean. This SPOT image shows Karachi from space:

Now, to India, the world's third most populous country (over a billion people). It is often referred to as a "subcontinent". Here is a map that shows its largest cities:

A map of the geology of India shows that a large part is a Precambrian Shield made up of igneous and metamorphic rocks (gray). Below the Himalayas, the rocks are Cenozoic and

younger (yellow). Volcanic extrusions (red) are carved into valleys and canyons/mountains of the Deccan Plateau.



Much of western and central India is dominated by the basalts of the Deccan Traps (a series of flows extruded between 68 and 60 million years ago). Here is a space image showing part of the plateau, along with an aerial view:

The Gulf of Kutch (lower left) in western India (State of Gujarat) is surrounded by the Kutch lowlands on the north and the Kathiawar Peninsula on the south. The region is also known as the Rann of Kutch. Vast tidal and saline marshes, with little vegetation, are distributed both in the upper left (the Great Rann) and at the head of the Gulf. (the Little Rann). These mudflats are superposed on alluvial plains, possibly developed when the Indus River to west once emptied further east into the Indian Ocean. The dark brown areas on land are low rises capped by part of the Deccan basalt flows that extend over much of western India. Only the area in the lower right is notably populated.

South of the Rann, on India's west coast, is one of its famed cities - Bombay (now renamed Mumbai, to detach it from its English colonial history). Here it is in a Landsat-7 ETM+ image.

Moving north, then east we see this ASTER image of India's capital, New Dehli.

A more detailed view is afforded by this SPOT image of part of the Delhi group:

New Delhi is a comparatively young city, being built largely during and after the last years of British rule in India. Below is a skyline view of the modern part, and finally some government building in the style typical of Indian architecture:

About 160 km (100 miles) southeast of New Delhi is the most famous landmark in India - the Taj Mahal, built in the first half of the 17th Century by the Mughal Emporer as a promise to his wife, the Empress Muntaz Mahal, upon her death. Below is an IKONOS image of the Taj Mahal, just outside the city of Agra, and beneath it are two ground views:

The largest city in south-central India is Hyderabad, outside of which is the Indian space program's National Remote Sensing Agency. It appears here in an LISS-3 false color image. The center of this city is home to India's "Arc de Triomphe", the Charminar, constructed as a monument in 1591 by Muhammed Quli Qutab Shah.

Still further south is the fast-growing city of Bangalore. Chances are you have talked with someone there - Bangalore has become the Information Technology (IT) capital of India. When you telephone an American business number, for example, you may be switched to a specialist in India who has been trained to provide the needed service. The Indian accents are distinctive, yet easily understood. But many software companies have relocated major facilities in this town. And the center of the Indian Space program is in Bangalore. Here is a Google Earth image of the city and an enlarged view of the Philips Electronic plant, one of the many foreign operations using the skilled but inexpensive labor force that is now becoming common in Asia as part of the "Flat Earth" - a term now applied to the process of globalization of the world's economy:

A Landsat-7 image shows Calcutta, in southeast India; this is still the most crowded city in the country, and was home to Mother Theresa for decades. First, look at it in context with its surroundings, in this Landsat-7 image:

A more detailed rendition of Calcutta is this SPOT image:

East of Calcutta is Bangledesh, a Muslim country through which flows the Ganges River. Its delta has many distributaries; the coastal zone is sparsely populated but contains vast mangrove swamps, dark green in this natural color astronaut photo. The region is subject to fierce

typhoons and monsoonal rains that can severely affect the lives and safety of the densely populated areas (brownish in the upper part of the photo).

The Brahmaputra River northeast of Calcutta, noted for its huge load of sediments, begins where several feeder rivers from the mountains of Tibet and the easternmost State of India, Assam meet. The river flows through Assam (right half of image) past the Shillong Plateau and the Garo Hills (a Precambrian crystalline complex) and upper Bangladesh (left), and finally into the Lower Ganges at Dacca (below this scene). In this view, it is joined by several rivers from the foothills of Sikkim and Bhutan. The Brahmaputra during rainy season can be greater than 8 km (5 miles) wide. Here, more than a month after the end of the monsoons, most of the water has flowed on, leaving choked stream beds and numerous small channels, a condition known as a braided stream.

Below the south tip of India is the former Ceylon, now an independent nation called Sri Lanka. Here is a mosaic of this country:

East of India and Bangladesh is the country of Myanmar. In World War II, it was known as Burma, and its chief city was then called Rangoon (now, Vangon). For several years, the Japanese Army occupied Burma and threatened to invade India.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Australia is one of two "island continents" (the other, of course, is the Antarctic) - so big that it is misleading to think of it as an island per se. Most inhabitants live along or near the coasts where nearly all its large cities are situated. Vegetation sufficient to stand out in the landscape is confined mainly to the southeastern corner and tropical areas in the northern sectors near the equator. Semi-arid and desert conditions prevail in most of the interior. High mountains are scarce; plains dominate. To the east lie the two islands of New Zealand. North Island has active volcanoes. South Island has a great mountain range, the Alps. The vast Pacific ocean is home to many islands, some volcanic and others atolls formed from coral reefs.

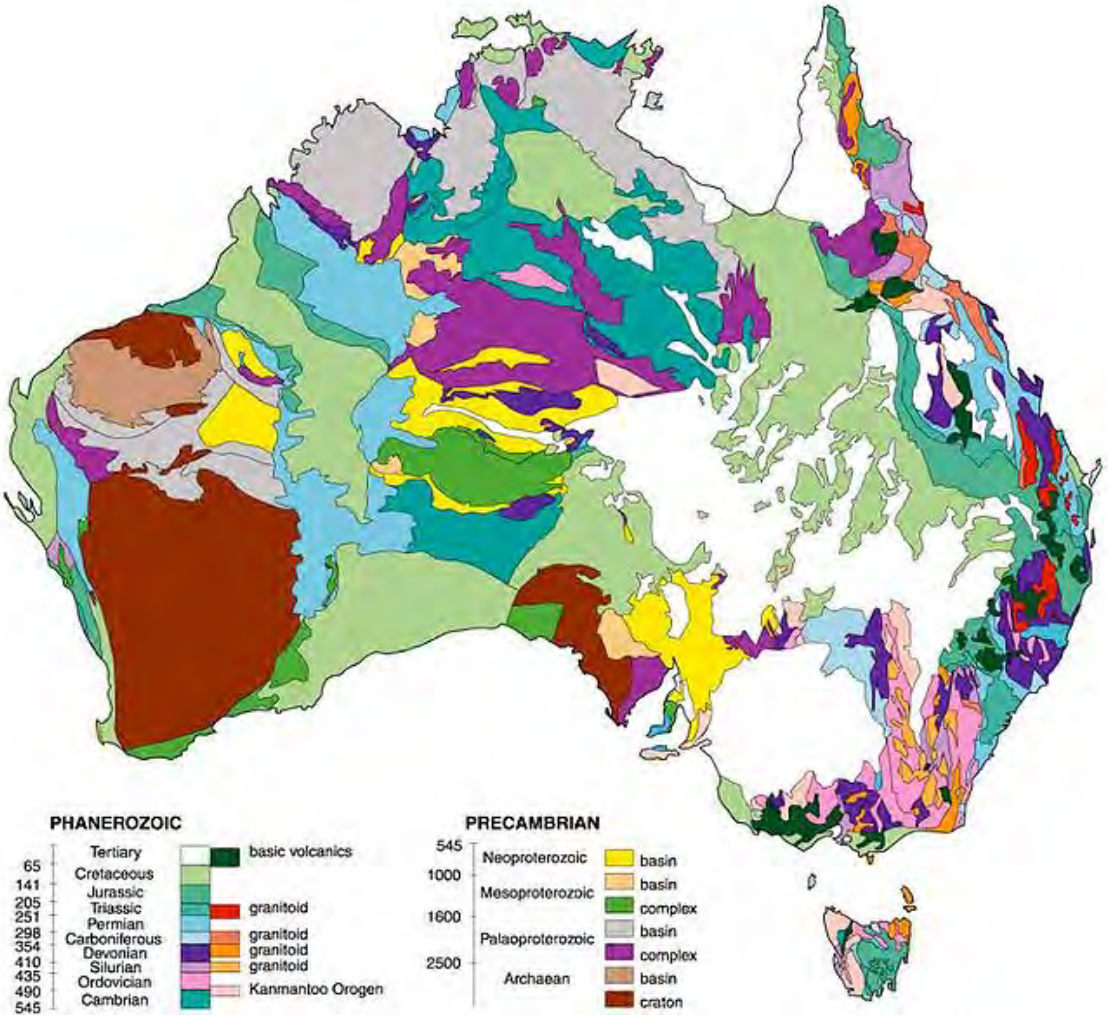
Australia and New Zealand and the Pacific Islands

Ausralia is a vast continent embedded in the Indo-Australian tectonic plate that is moving northward against the Indonesian Islands. Most of its citizens live near the coast. This map names the states of Australia and shows its major road network:

All of Australia is pictured in this Landsat mosaic (a larger and somewhat different version appears near the bottom of page 7-3).

Most of Australia is semi-desert. Mountains are infrequent and relatively low. The geology of the continent is summarized in this simplified general map:

This map is similar but breaks down the structural and stratigraphic units more specifically:

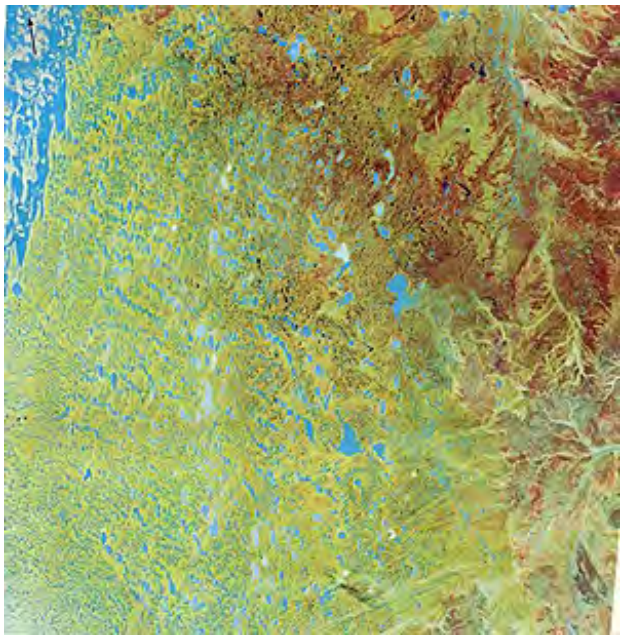


We saw part of northwestern Australia when we examined the Pilbara district on page 2-6. We will look again here we will look more closely at one of the images making up that mosaic. This is the so-called Pilbara district of northwest Australia, near the coast, made up of Archean and Proterozoic rocks. Near the batholiths are Proterozic rocks making up the Hamersley Group. Some of the units are banded ironstones. This ASTER image shows off this spectacular geology:

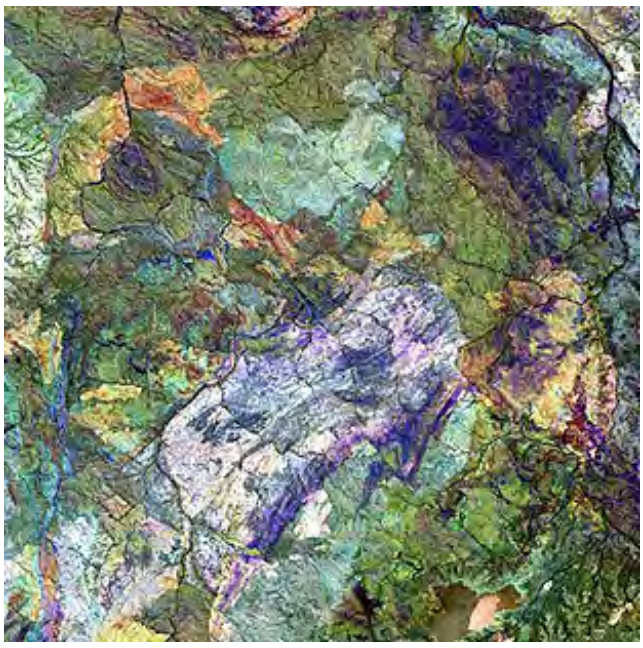
This ground photo is typical of the country comprising the Pilbara district

In the Northern Territories is Arnhem Land, populated mostly by aborigines. This scene is about 200 km east of Darwin, the only large town in the region. The scene below was taken during a dry period, but rain forests (dark red) occur, and vegetation blooms after rains. This land contains animals such as portrayed in the movie "Crocodile Dundee".

Much of the Australian Interior is dry, harsh, inhospitable, and sparsely developed. This next scene lies in the Sturt Desert where New South Wales, South Australia, and Queensland meet. Numerous sand dunes aligned with long dimension running NW-SE blanket the landscape. Normally rainfall is less than 12 inches (30 cm) a year. But in this scene there had been a heavy period of rain a bit earlier such that water has collected (temporarily) in the low areas between dunes.



Approximately 80% of Australia is arid lands (the other 20% includes tropical forests in the north and temperate forests along some of the coasts). Only a fraction of the drier terrains have numerous sand dunes like those in the Sturt Desert above. Here are two views of the Great Sandy Desert in northwest Australia, the first showing some broad sandy cover, the second showing mostly bedrock with sparse vegetation:

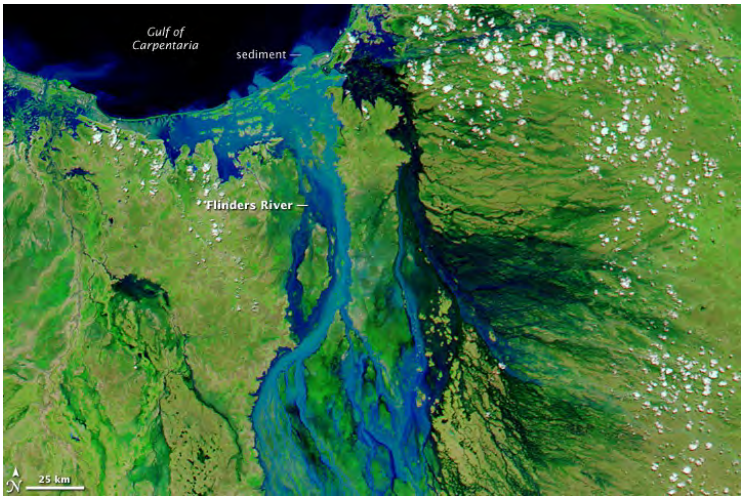


A group of low plateaus appear in desert lands in interior Queensland. In this unusual satellite image of the Porcupine Gorge region, the individual plateaus stand out because of an interbedded basalt lava flow unit which traces their outlines as though with a dark pen:

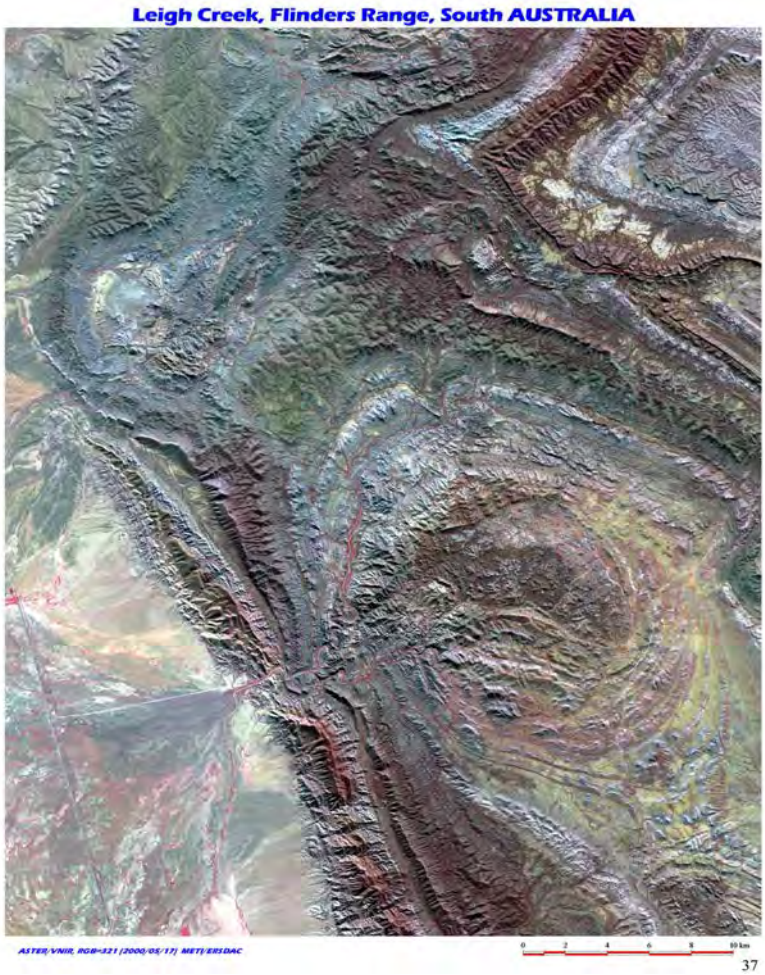
The influence of a prolific, if short-lived, rainy season in the interior is evidenced in these next two Landsat scenes within the Great Artesian Basin, located at the border of Queensland and New South Wales. Top image: After especially heavy rains, the Barcoo River fills a depression with water that extends as a large lake, Lake Cuddapan. The surrounding land is also now

displaying grass and marshy vegetation, the former suitable for cattle grazing. Bottom image: After the wet season, the land reverts to a barren state but the lake beds remain wet enough to support extensive wetlands vegetation.

Because so much of Australia is relatively flat lowlands, floods such as the one shown above spread over wide areas. This happened in January of 2009 in northern Australia when the Flinders River - normally a stream that flows over a semi-desert landscape - received copious water from heavy 'southern summer' rains and spilled over wide areas, as seen in the lower MODIS (on Aqua) image below:



The dry climate and lack of vegetation in much of Australia's interior favor excellent exposure and expression of geologic structures. Mountains in the Flinders Range are strikingly exposed in this ASTER image:



The main destination of tourists exploring the Australian Outback (interior) is the town of Alice Springs. Nearby is the MacDonnell Range, fold mountains that are expressed as ridges. In this next image, the Gosses Bluff impact crater (Section 18) is a small circular area to the north (near top center) of the Ranges.

Several hundred kilometers from Alice Springs is the country's most famous natural landmark known both by its aboriginal name, Uluru, and by the name given it by English settlers, Ayers Rock. This is the largest monolith in the world, being 3.8 km long and 348 meters high. It consists of arkose (feldspar-rich sandstone), in highly tilted strata. Now a National Park, it boasts several five star hotels within sight to accommodate the hundreds of thousands of visitors now trekking

to it each year. Below we show two high resolution color images of Ayers Rock taken by the IKONOS satellite and an aerial oblique view in natural color.

The dissected plateau country in southeast Australia receives much more rainfall so that the region is heavily forested with eucalyptus, gum, mahogany, and ironwood, as well as tree ferns, palms, and other tropical vegetation near the coast. The Blue Mountains and the Hunter Range make up part of the scene below. The largest city (more than 4 million) in Australia, Sydney, lies between Botany Bay (south) and Port Jackson Bay (north) and extends inland at least 20 km (12 miles). This is how it appears in an Terra MISR image taken just before the Olympics 2000 began in September. The city itself is just right of center.

This Landsat image shows greater Sydney:

Zooming in closer on Sydney, look at this IKONOS image of part of that great city:

The downtown section of Sydney is impressive in this panoramic photograph:

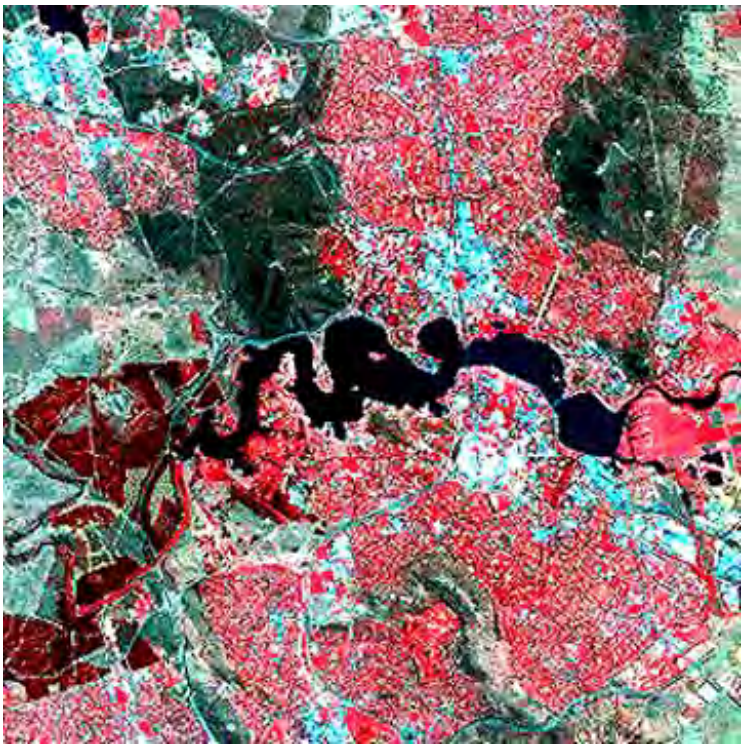
Sydney's most famous landmark is the fabulous Opera House:

This marvelous structure, completed in 1972, shows up as a dominating feature in this Quickbird image of the harbor and downtown buildings of Sydney:

The 2000 Olympic Village in Sydney (outside the edge of the EO-1 image) is sharply singled out by an IKONOS-2 image:

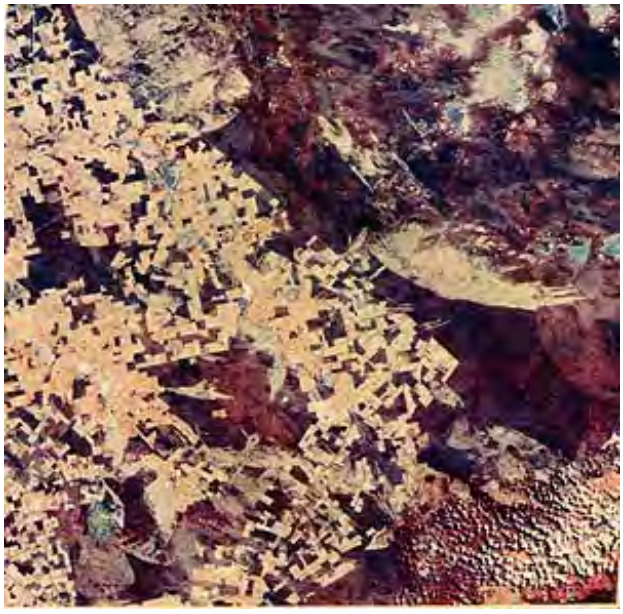
Canberra, Australia's capital, is a newly built complex that has wide open spaces and initially a small population. Everything is built around a central lake:

By the mid '90s this Landsat TM subscene shows major population growth and spread, as the city itself also by then had many planted trees in full growth.



The only big city along the southwest coast of Western Australia is Perth on the Indian Ocean, shown here in a Landsat subscene. Flat plains inland from Perth end abruptly against a spread of low mountains, the Darling Range, heavily forested. The linear nature of the contact of plains with these uplands (part of the Paleozoic crystalline shield), suggest a fault.

We look next at an area about 725 km (450 miles) east of Perth and 160 km (100 miles) inland from the south coast. It is underlain by ancient Archean crystalline and sedimentary rocks that have been worn down to lowlands (plains). Natural vegetation is known locally as mallee scrub, a mixture of saltbush and bluebush. Most of the upper right has been undeveloped but over the years, much of the remaining area has been cleared by burning. Calcium-rich soils there, plus a rainfall of 25-38 cm (10-15 inches) per year, have favored development of numerous interlocking large farms. Wheat is the principal crop. Here in October (Australia's spring), the fields show only the first stages of growth. Between the farmlands and the pristine terrain to the northeast is one of Australia's electrified "rabbit-proof" fences designed to keep that marauding pest away from the croplands. Salt lakes appear in the virgin lands.



When tourists come to Australia, besides Ayer's Rock and Alice Springs, and of course Sydney, the third popular destination is the famed Great Barrier Reef off the coast of Queensland, as seen in this image:

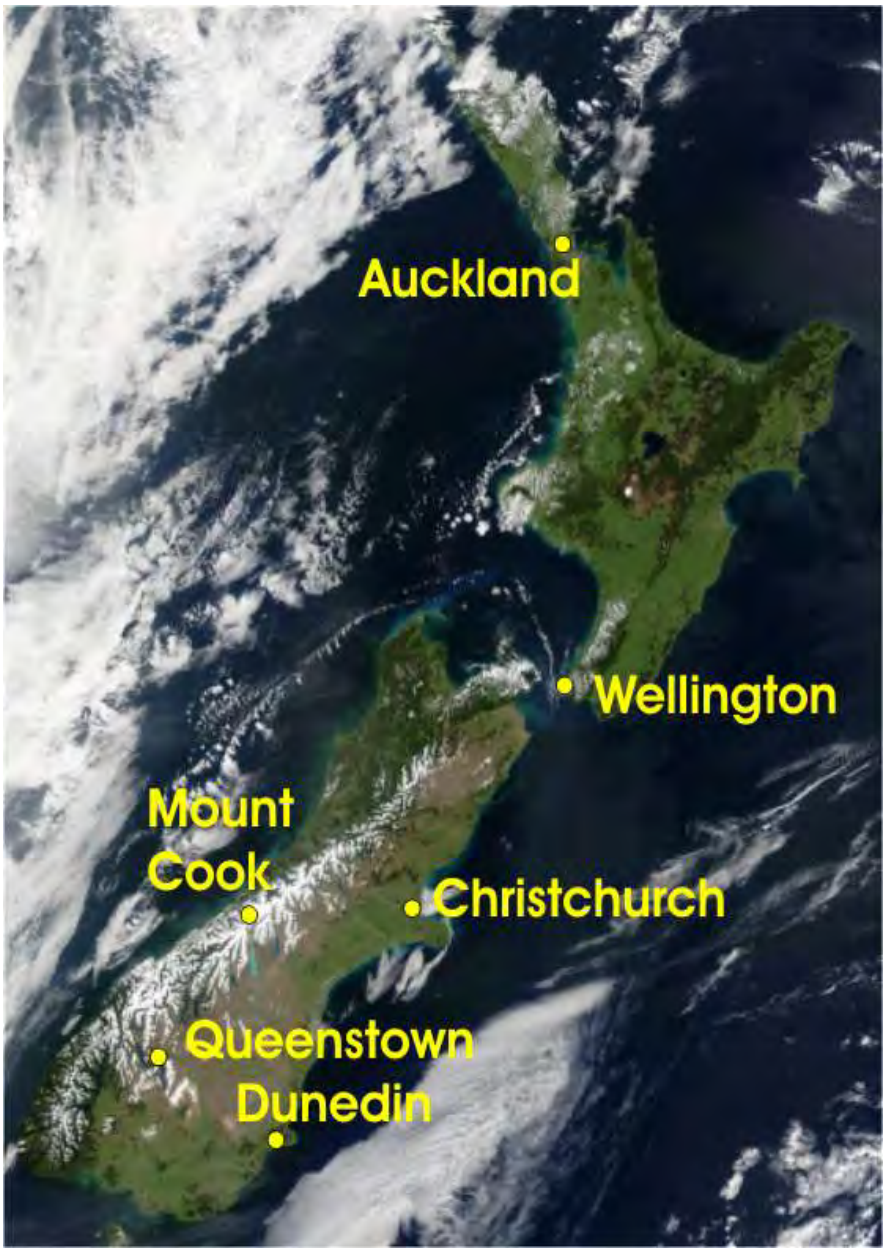
A closer look from space shows parts of the reef complex just north of the city of Cairns - the prime spot for vacationers who wish to explore the reef.

One way is by air; here is an aerial view of the reef:

The best way is to explore parts of the Great Barrier Reef close up, either by sailing over it in a glass-bottomed boat or by snorkeling. This is an indication of what you would see:

The GBR, like much of the world's coral reefs elsewhere, is experiencing damage by bleaching (the white coral stalk in the above photo is an example), brought on both by warmer temperatures and by chemicals added to the seas by man. In this global warming era, there is a real danger that the coral reefs across the globe can by killed.

One of the most beautiful and scenic countries in the world is New Zealand, a pair of islands some 1600 km (1000 miles) southeast of Sydney. These landmasses form near a major plate boundary between the Pacific and Indo-Australian plates. New Zealand consists of two large islands, seen here in this satellite image:



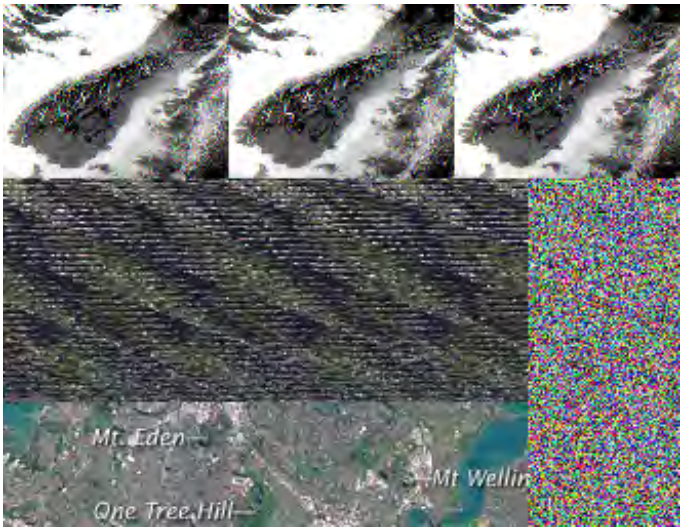
North Island is notable for its several active volcanoes and a major geyser-hot steam field. This Envisat image shows most of North Island in more detail.

The capital of New Zealand is Auckland:

Auckland has grown into a modern city with an impressive skyline:

The second city in population is Wellington. Here it is as seen in an astronaut photo:

But to those who have toured N.Z., South Island is ranked as even more striking because of its spectacular mountains and lush plains. The snow-capped Southern Alps, a chain of complexly folded and metamorphosed rocks elevated by plate collision to current heights above 3000 meters (10000 ft). Their western side is bounded by the great Alpine fault, analogous to the San Andreas of California (and many visitors compare the landscapes of N.Z. to parts of that state). The entire South Island has been emvisaged in a Day-IR HCMM image (see page 9-8), shown next, in which many of the black patches are snow. Note the effect on the regional topography by the Alpine Fault (linear boundary below top center). On its south are high mountains; to the north are low plains.



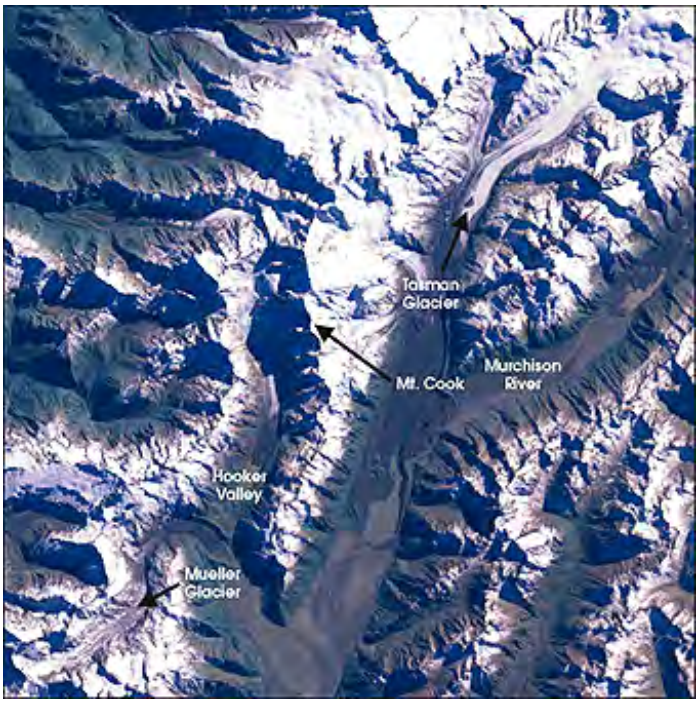
On December 7, 2007 the MODIS instrument on Aqua obtained this cloud-free image of the South Island:

The Alps have been extensively glaciated, producing the rugged peaks and valleys walls that resemble Yosemite in California. As evident in the scene below, the eastern Alps give way to the Canterbury Plains, across which flow (top to bottom) the Waiau, Waimakariri, and Raikaia Rivers. The Banks Peninsula to the east is built up, in part, of several now extinct volcanoes. The city of Christchurch (blackish area) lies at its northwest end. Land use is mainly sheep and cattle husbandry, and wheat, oats, and fodder farming on the Plains.

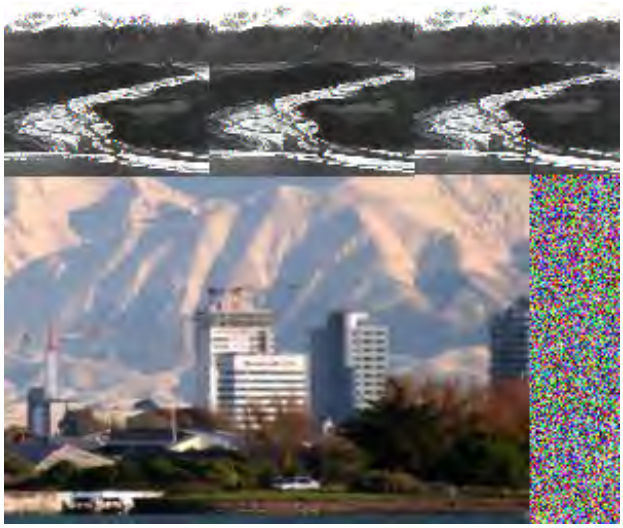
This next photo is a bit of a trick. Christchurch is seen in the foreground but the Alps appear to be close-by. They are actually distant, but a telephoto lens was used to obtain this scene.

Movie-goers, and especially fans of the Lord of the Rings trilogy, should recognize this scenery. The entire Ring was filmed mostly in South Island of New Zealand, but with some scenes in North Island. The actors and crew were based out of South Island's largest town - Christchurch, seen here in an International Space Station photo (the 250000th picture taken from that platform). The circular hills are a dissected old volcanic edifice.

The kinship in appearance of the New Zealand Alps to those in Europe is evident in this Landsat 7 ETM+ image that shows Mt. Cook, at 3754 m (12316 ft) the highest point in this mountain chain, and almost twice as high as the highest mountain in nearby Australia. The ranges on the two islands were once connected before New Zealand split off.

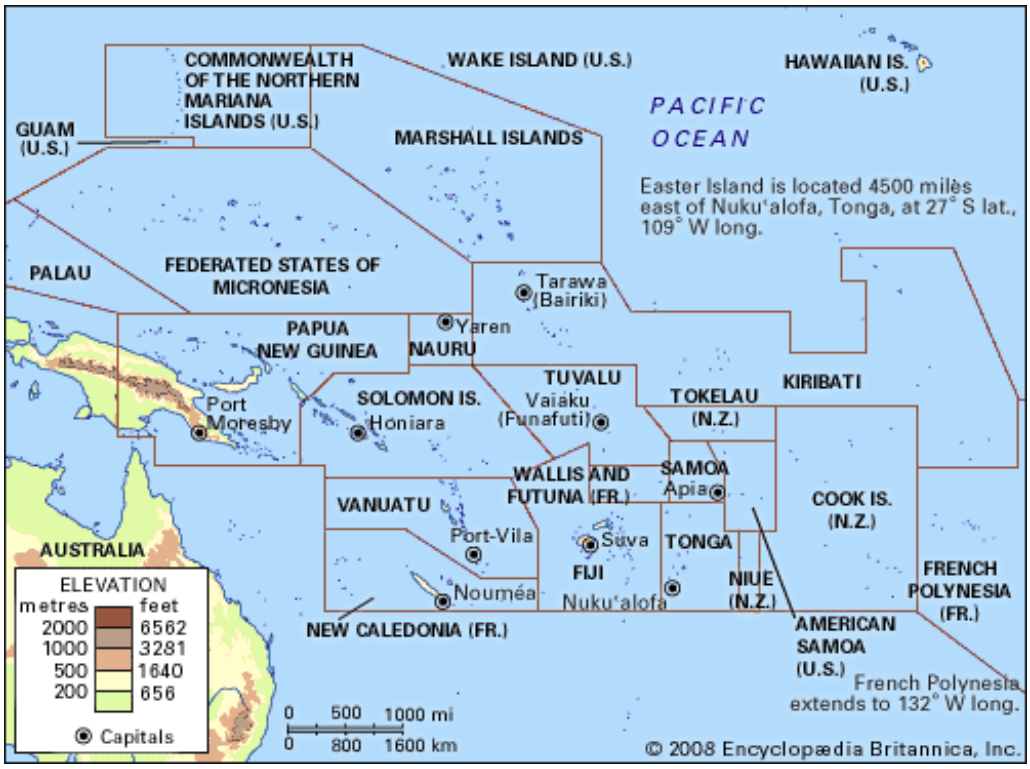


The ground photo below was taken from the coast north of Christchurch looking at the terraces in the Canterbury Plains cut by the Karangama River with the high Alps in the background.



A closer look at the Southern Alps is afforded by these two images, the second being of Mt. Cook, highest in the range, and similar to the Matterhorn in the European Alps:

As one moves out into the great Pacific Basin beyond Australia and New Zealand, many island groups are found scattered about, mainly in low latitudes. Some of the islands shown below on this page are located on this map of Oceania - a general name given to many of the Pacific Ocean islands:



Tahiti is typical of Oceania's islands. It is volcanic in natureIt is located in the Society Islands of French Polynesia. Here it is from space, with the airport at Papeete shown as a thin red strip in the upper left of the island's shore:

Papeete is its largest city. The neighboring island of Moorea rises abruptly from the ocean, as it is the site of a large volcano.

Indeed, many Pacific islands are volcanic (see the Hawaiian Islands as the lead example, on page page 17-3). Some of those have fringes of coral around them. In this latter category, the third largest island is New Caledonia. This photo, taken by an astronaut aboard the ISS, shows a part of the island with its peninsula that contains the largest city, Noumea.

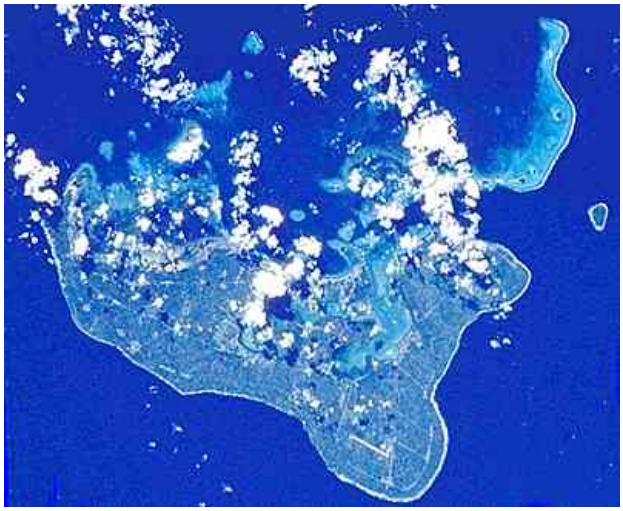
Many Pacific islands are atolls (see page 17-4 for other examples). These may be just fringing

reefs with a lagoon inside, as exemplified by the Arno atoll in the Marshall Islands.

Some atolls are inhabited but the residents need to be supplied with food and necessities by ship since nearly all atolls are not self-sustaining. Atafu atoll (entrusted to New Zealand) is one such example; it has trees (note the coral (lighter blue) in its lagooon:

A typical atoll without a lagoon is Mantangi Island in the Fiji Island group of several hundred. The entire island is a coral reef some 3 km (2 miles) in length. It is fringed with submerged coral (lighter blue) and a white coral sand beach; inland the island is covered with tropical trees and brush.

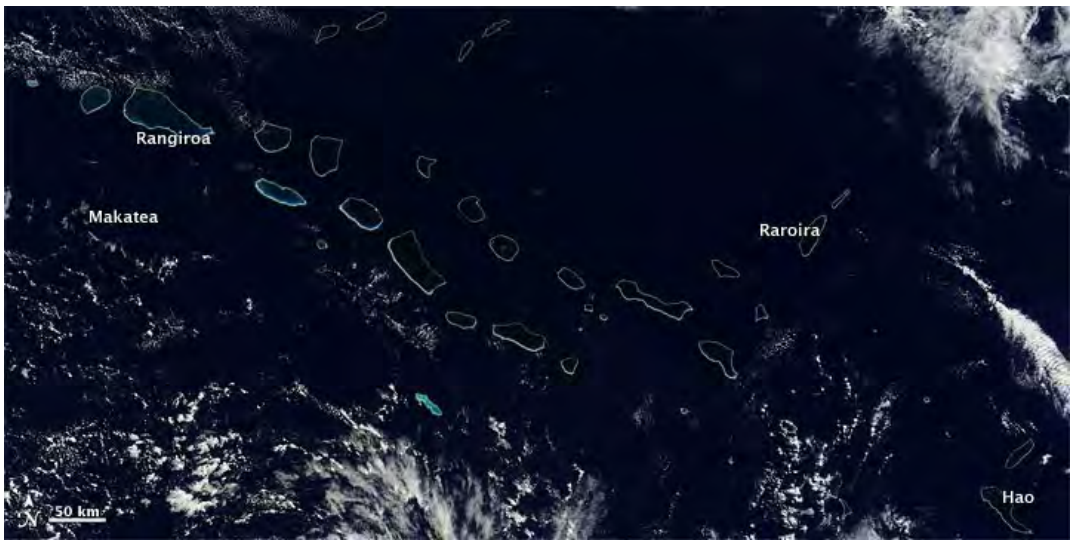
Some atoll islands are well populated, as shown by Tongatapu in the Tonga Islands, a group near New Zealand.



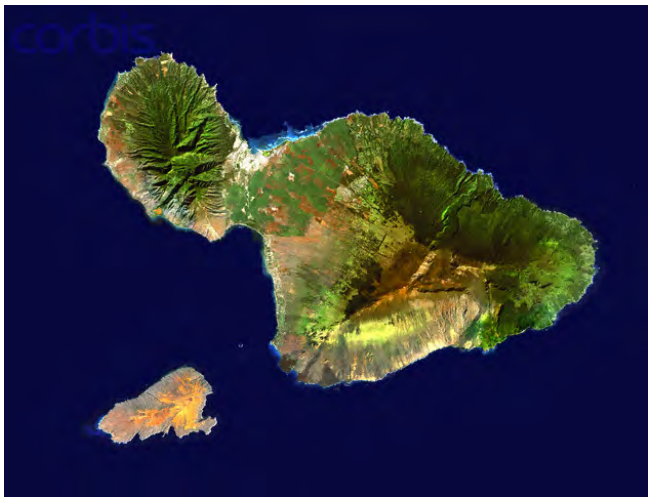
This next image is the atoll of Tetiaroa, in the Society Islands. (These are part of French Polynesia and include Tahiti and Bora Bora. Its claim to fame is that it was purchased outright from the French government by Marlon Brando, the famed Hollywood actor:

Coral atolls can develop in clusters, as shown in this IKONOS image of the Maldives:

Other atolls are aligned in chains, as are the Tuomotu islands in French Polynesia. This archipelago contains 73 individual atolls, the largest of which is Rangiroa, whose top is 80 meters (250 ft) above sealevel. The islands are inhabited, despite the appearance of having little land mass.



The United States possesses a widespread collection of atolls and volcanic islands scattered about the Pacific. Here are two that are well known (Guam sends delegates to the Republican and Democratic conventions); check the captions for identities:



By now you should have formed the impression that the U.S. has many islands in its possession, both in North America and worldwide. Here is a List of American Islands produced by Wikipedia.

Islands are widespread over the Pacific. Some have notably histories in World War II. Midway northwest of the Hawaiian Islands was part of the turning point battle of the Japanese campaign. While it was being attached by Japanese carrier planes, U.S. carriers launched strikes at the Japanese, sinking four of their carriers. Here is Midway, still a U.S. Marine base; it is a very narrow coral reef:

General Douglas MacArthur, along with Admiral Chester Nimitz, made up the duo who devised the "island hopping" strategy that recaptured some islands that the Japanese had occupied earlier in their 1942 spread across the west Pacific. Guadalcanal and Tarawa were the sites of two of the most vicious battles. Tarawa appears here in this satellite image - one wonders why such a tiny island was so vital in the campaign but again the airport was the main motivation.

Several thousand miles east of the Tonga Islands is Easter Island, seen here from space. It is owned by Ecuador. Its claim to fame are the remarkable statues that awed the first explorers after its discovery.

With these last scenes, we conclude our odyssey around the fascinating landscapes and urban centers on our planet Earth.

.

Primary Author: Nicholas M. Short, Sr.

This Section is a change of pace, in which, rather than having to concentrate on heady ideas, we will simply take a pictorial tour - a sort of travelogue - across the U.S. and then around the world. We will look at Landsat and remote sensing images from other satellites, plus where appropriate some ground scenes, in order to "enjoy" a view of our planet's land surfaces as though we were ourselves in the Space Shuttle. The first page introduces several concepts dealing with the physiographic (geographic/geologic) provinces of North America.

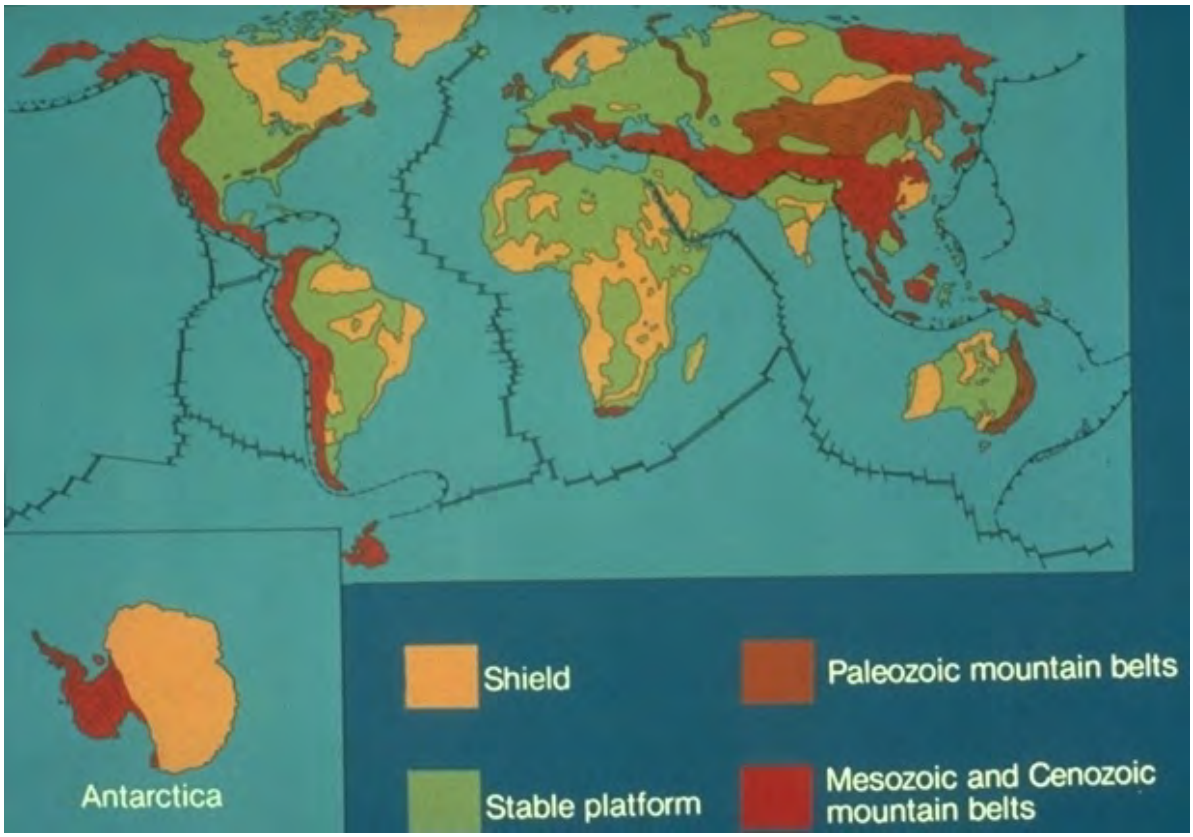
A TOUR OF THE WORLD FROM SPACE

GENERAL BACKGROUND

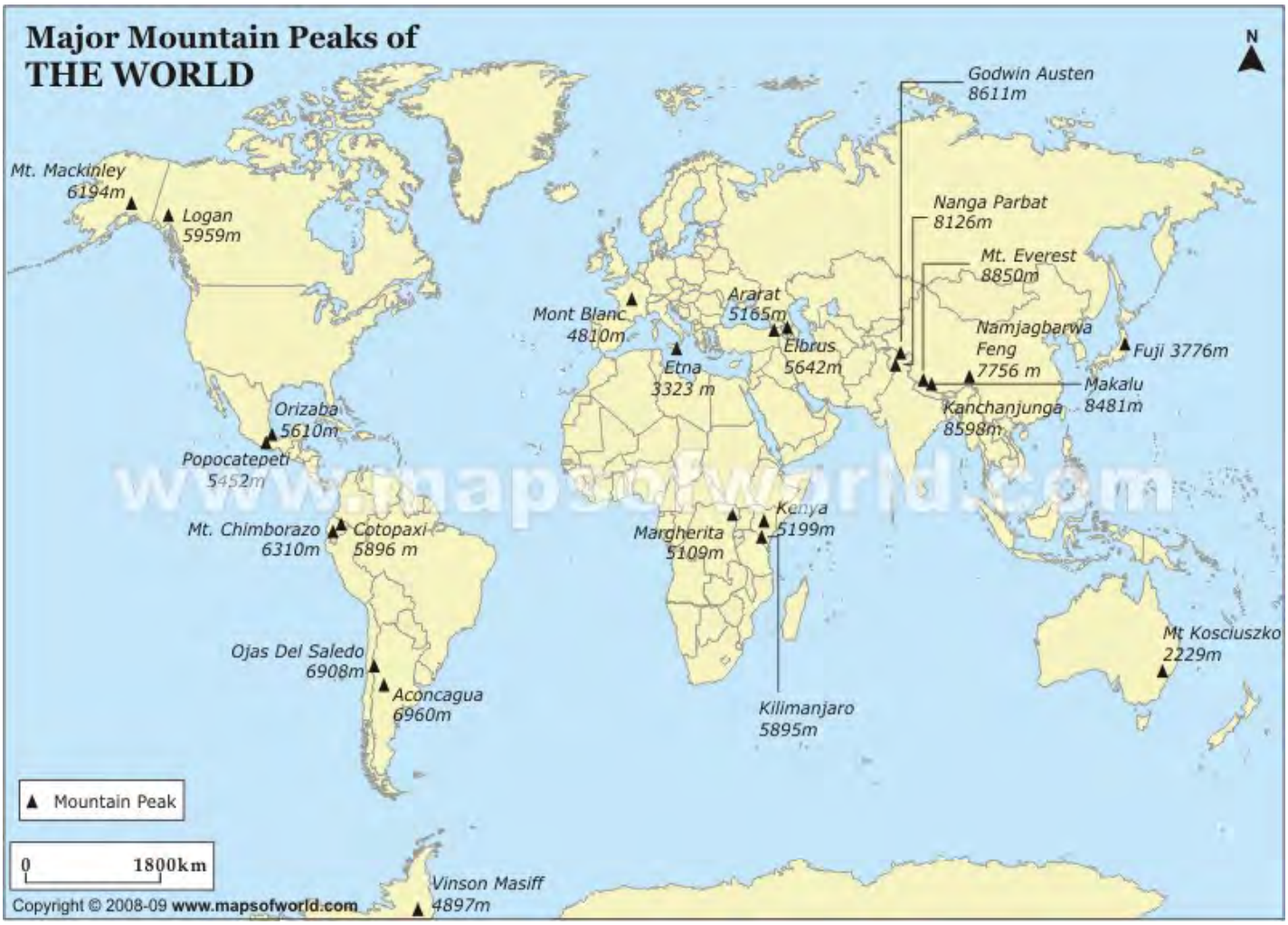
Enough of the heavy stuff for a while! Let's take a breather, have some fun, and just look at some stunning scenes that depict typical landscapes and urban centers of the world. This Section amounts to what can be called a "Travelogue", a pictorial journey across the world looking at places both well known and vistas that give a "flavor" to geographic regions as seen not only from the ground and the air, but now from orbital platforms in space.

The first part of this Section will concentrate on selected regions of the United States from the East to the West Coast. The second part consists of representative views of other countries of the World, organized and displayed by continents beginning with the rest of North America, Central America, South America, Europe, the Middle East and Africa, and Asia.

We will stress two major themes in each part: 1) the natural features of the regions we will visit; and 2) many of the principal and best known cities in various countries. The natural features are grouped into two broad categories, controlled by their underlying geology: 1) mountain belts and shields (igneous and metamorphic rocks; folded/faulted sedimentary rocks), and 2) largely horizontal sedimentary rocks (so-called platform [tectonically unactivated] setting). This map shows the global distribution of category 1 in color and 2 in white (blank).



As an example of some specifics associated with the first theme, we will show satellite imagery and ground photography of some of the highest mountain peaks worldwide:



This map includes many cities we will visit:

So, lets get started on the journey!

PART 1: ACROSS THE UNITED STATES

We set off by going on a simulated air vacation trip from one coast of North America to the other - going east to west, and pretend we can look out over large expanses of the ground, om effect from the Canadian border to the southernmost U.S. In so doing, we will gain a sense of the variety of landscapes that make this one of the most diverse and interesting continents on the globe. We will accomplish this journey by perusing various satellite scenes each typifying some specific region within the 48 conterminous states.

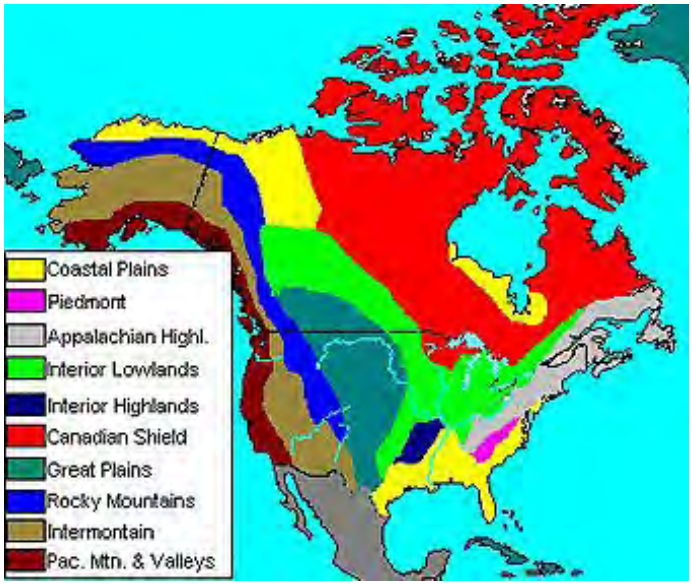
Many people, especially those who live near the East or West Coast of the United States, have flown on commercial aircraft across this country. Those who looked frequently out a window were certainly impressed by the ever-changing landscape. Each region has its own distinct character, expressed by combinations of topography, vegetative cover, and land use/cultural patterns. These characteristics are largely controlled by the underlying geology, climate, vegetative cover, population densities, and particular modes of infrastructure development for supporting human activities.

Some of you have been introduced to these ideas and relationships, if you have taken a course in Physical Geography. One element or theme in that course is the importance of **physiography**, a term which connotes "description of the landscape" at *regional* scales that describes the types and influence of landforms, soils, climate, vegetation, and hydrologic actions on the natural scenery. Each realm is a discrete **physiographic province,** which, once lived in, seems unique, so that we always sense a distinctiveness in the surroundings when visiting a particular province. We show a map of the physiographic provinces in the 48 United States below:

From H.J. de Blij and Peter O. Muller, Physical Geography of the Global Environment, Figure 52-1, 1st Ed. © 1993. Reproduced by permission of John Wiley & Sons, Inc., New York.

6-1: Locate (approximately) your birthplace on the map and determine what physiographic province you live in. ANSWER

This version of the Physiographic Provinces may prove simpler to remember the main subdivisions:



Five of the six mega-provinces in North America (excluding the Arctic Coastal Plains) extend into the U.S. Three national regions are subdivided by two or more provinces, as named. The **Canadian Shield** is part of the ancient craton or core of the continent and is marked by a dominance of Precambrian igneous and metamorphic rocks, which give rise to low hills and subdued mountain ranges. Being in the north, the colder, generally humid climate favors dense boreal forests with evergreen and birch trees. The dominant mega-province in eastern North America is the **Appalachian Highlands**, consisting of several provinces, which contain rocks

ranging in age from about 570 to 245 million years (m.y.) They have been involved in several major episodes of mountain building, culminating in the Appalachian Revolution about 250 m.y. ago. The oldest rocks in the east, many of which are metamorphosed and invaded by igneous intrusions, comprise the **Piedmont** province, consisting of rolling hills and ridges. The structural bending of Paleozoic sedimentary rocks into folds often offset by faults produces a distinctive topography of dissected ridges that generally are now less than 1,220 m (4,000 ft) in elevation. This area encompasses the **Newer Appalachians,**, more commonly called the **Valley and Ridge** province. The surface has been worn down, after uplifts over the past 10 m.y, to form valleys with rolling terrain. To the west of the province is a broad area of uplifted but unfolded sedimentary rocks making up the **Appalachian Plateau** province. Over much of this province, erosion has carved out mountain-like terrain comparable in relief to the adjacent Valley and Ridge. Being in a humid, often warm climate the Appalachians are still heavily forested with deciduous trees except where cleared by humans for development.

The **Interior Plains** are underlain mostly by sedimentary rocks as old as 600 m.y. and young as 10-20 m.y. (in the western parts). These rocks are almost everywhere flat to subhorizontal, so that erosion has not carved out prominent mountain landscapes. The terrain is flattened further in the northern half by deposits from continental glaciers that covered the region during the last three million years. The **Interior Lowlands** seldom reach elevations above 460 m (1500 ft). They now support stretches of forests that remain after farmers cleared much of the land. These deciduous treescapes diminish to the west and south as grasslands take over. The **Great Plains** section consists of deposits of sediment brought eastward from the Rocky Mountains to the west. The deposits are piled up such that the elevations increase westward to 1525+ m (about 5,000 ft). Most of this land is either natural-grass rangeland used for grazing or farmland, with forests now being small and scattered.

In the southern U.S., but extending northward to New Jersey, is the **Gulf-Atlantic Coastal Plain** province, a flat lowlands of nearly horizontal rocks, laid down during the last 60 m.y. as marine waters repeatedly invaded the North American continent during variations in sea level. This region is predominantly farmland, although some parts are still forested.

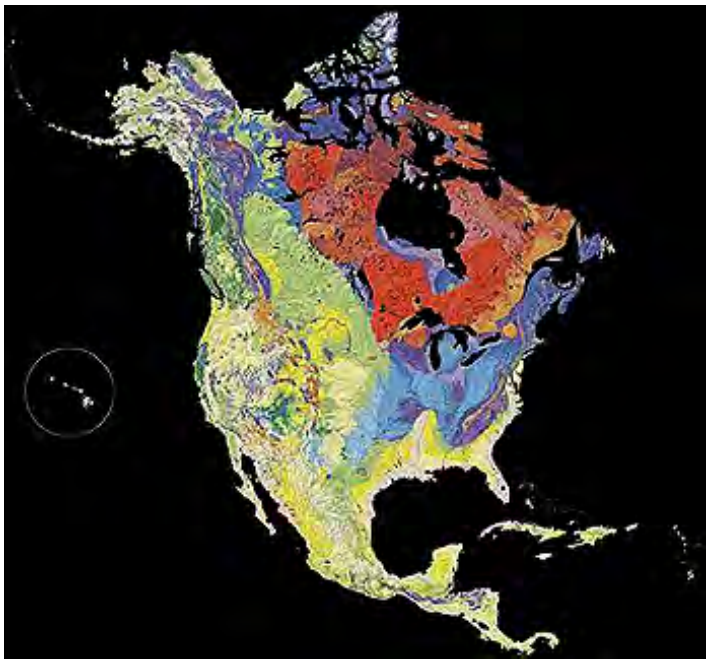
The **Western Mountains** are composed of a number of diverse rock/structural units of differing ages and histories of orogeny (mountain-building). Although the rocks exposed can straddle a wide range of geologic periods, most of the events that produce the units took place in the last 70 million years. Because of differences in structural states, we conveniently break the eastern segment known as the **Rocky Mountains**, which can reach heights up to 4,420 m (14,500 ft), into three sections. The **Colorado Plateau** is an elevated (typically around 1,200 m [4,000 ft]) uplift of largely horizontal rocks, which have been cut into spectacular landscapes mostly covered with sparse desert vegetation. The **Basin and Range** province consists of rugged, often barren rock bodies uplifted along faults into elongated mountains separated by valleys with gentle sloping floors created by erosional fill from the higher ranges. Vegetation is mainly desert types that are subordinate in density of cover to the rocky soils that dominate the surface.

The **Columbia Plateau** is a region capped by volcanic flows (mostly basalts) extruded from fissures over the last 40 m.y. In places, the flat land is cut by deep canyons, but is also broken by older rocks sticking upward as pre-existing mountains that were not engulfed by the volcanic flows (which can be 1,200 m [4,000 ft] in thickness). The landscape in much of California and the western parts of Oregon and Washington is also mostly mountainous. The high **Sierra Nevada** is a great tilted fault block rising as high as 4,420 m (14,500 ft) and supporting fir and pine trees. It connects to the **Cascades** which are marked by numerous stratovolcanoes, more or less active in the last 20 m.y. Along the Pacific coastline are the **Coast Ranges**, still being actively warped by plate tectonic-driven compression. In California, between these mountain belts, is the **Central Valley** (subdivided into the Sacramento [north] and San Joaquin [south] valleys), a structural basin now receiving infill from the highlands. Its warm temperate climate allows extensive farming.

The Physiographic Provinces correlate well with the first-order (general) geology of the U.S. This is evident from the next map that includes not only the U.S. but Canada, Greenland, and the Caribbean. This map shows general structural units related to the provinces. (See also page 2-1b for related information). The term "craton" refers to the stable nucleus of ancient rocks that makes up most of the North American interior; platform denotes sedimentary rocks that cover the craton. Note that there are two main orogenic belts: the Appalachians on the east, with a present day passive tectonic margin, and the Rocky Mountains/Coast Ranges assemblage on the west, with an active margin (subduction; transform faulting).

The map above alludes to something not yet stated. Geographers favor including all of Central America and the Caribbean as part of North America. This is justified by the fact that most of these areas are part of the N. American tectonic plate.

There is a general correspondence between physiographic expression as provinces and the general geology of each region. Here is a look at the U.S. Geological Survey geology map (emphasizes ages of the units) of North America:



As a generalization, red refers to exposed Precambrian rocks, purple to Lower Paleozoic, blue to Upper Paleozoic, greens to Mesozoic rocks, yellows to Cenozoic; and buffs to Pleistocene rocks.

Somewhat more detail is included in the U.S. Geological Survey map of the 48 contiguous United States:

Working together, the U.S. Geological Survey (USGS) and the Environmental Protection Agency (EPA) used Landsat TM data to produce a general land cover map of the 48 contiguous states, published in 1992:

The Physiographic Provinces also correlate rather well with the different major classes of vegetation. Those in turn depend both on soils (which are dependent in various ways on the underlying geology) and climate (which also helps determine which soil types develop). Here is a general Vegetation Map of North America; the names in red are soil types (not treated here; surf the Internet under Soil Classification for further information):

Compare the major vegetation types on the map with this image of all of North and Central America, made from satellite inputs. You should be able to discern subtle to sharp difference in color tones - which correspond mainly to extent of vegetative cover - that match the map units fairly closely:

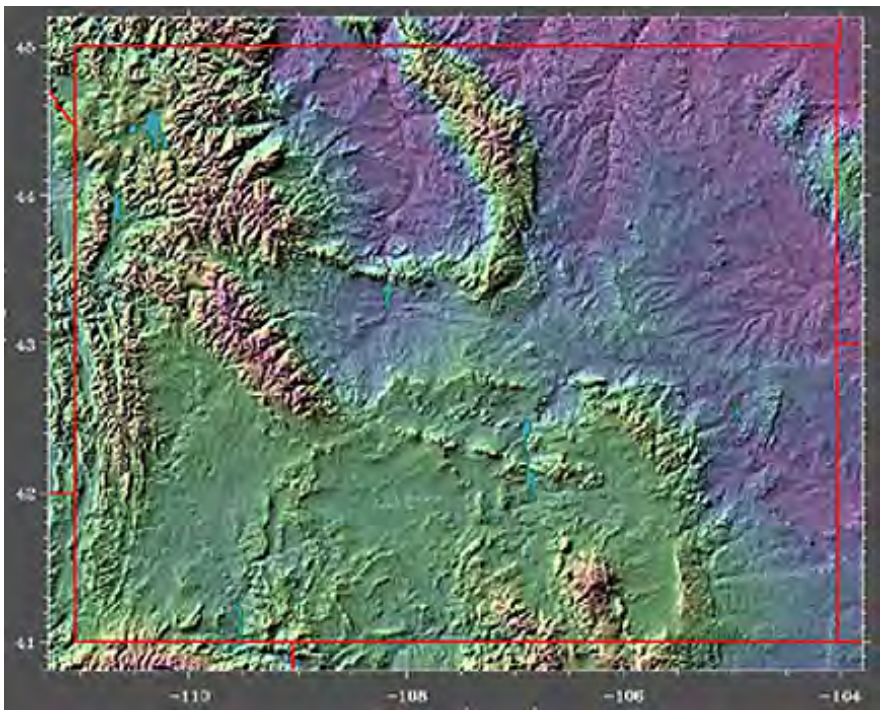
To provide you with a visual overview of the entire country, we display a mosaic of Nimbus Advanced Very-High Resolution Radiometer (AVHRR) false color images (see page 14-2 for a description of this sensor and its satellite platform).

Alternatively, we could have reproduced one of several Landsat mosaics, such as the color versions produced by General Electric Corporation (for the National Geographic Society to celebrate the 1976 Bicentennial) or the first ever (in black and white) put together by the Soil Conservation Service (see page 7-3. Because resolution is not a critical factor at the small scales needed to portray the U.S. from coast to coast, we used the AVHRR mosaic.

Many people have some problems in locating specific cities or major localities, and good maps are not always available. We reproduce a U.S. geographic map here, and will do the same for Canada and Europe later in the Section.

6-2: This will probably prove more difficult, but again locate your birthplace and your current residence, if it is different, on this color mosaic. ANSWER

You may wish to get a better look at the space image appearance of your hometown or town where you now reside that afforded by the U.S. mosaic. An AVHRR image and landform map of each state is accessible on a Web site produced by Applied Physics Laboratory of Johns Hopkins University in Maryland. To show you what to expect, here are the AVHRR and DEM topography (landforms; consult an Atlas to identify these) image you would get for the state of Wyoming.



Although geography is the main theme of this Section, and this page, we will insert here a brief history lesson. We all know that Christopher Columbus is celebrated as the discover of the Americas (in fact, of course, the millions of native tribes people [some being the Indians of whom we are familiar] were here first [for at least 20000 and perhaps 40000+ years], and Norsemen probably visited Newfoundland around 1000 A.D.) In actuality he first landed in October, 1492 on an outermost Bahamas island (called Guanahani by its natives; renamed San Salvador by Columbus). Here are two views from space:

During his first journey, Columbus also observed the islands of Cuba and Hispaniola (today, Haiti and the Dominican Republic).

By his fourth trip, he reached Venezuela and had landed in Mexico, Honduras, and Panama. But he never realized that he had discovered two huge new continents. To him these were islands in the New World which he ultimately did not believe were parts of India but perhaps were off its shores.

It remained for others to determine the vastness of the regions in the New World as it was first settled. Hernando Cortes invaded and conquered Mexico in 1519. Other Spanish explorers discovered Florida and started the first N. American settlement - today's St. Augustine - in 1565; Francisco Coronado and Hernando de Soto led expeditions into the interior (including the Mississippi River valley) of today's United States in the 16th Century. The Englishman John Cabot found the easternmost part of Canada (Nova Scotia) in 1497 - thus was the first to land on North America proper; Jacque Cartier (a Frenchman) sailed along the St. Lawrence River in 1535, thus establishing France's claim to Canada which remained French until the English victory at Quebec City in 1759 followed by the deeding of all of Canada to England in 1763.

The eastern U.S. was first occupied by Europeans with the failed English settlement in today's Virginia at Roanoke Island (1585) and the temporarily successful Jamestown in 1607 (the Virginia Colony thereafter grew fast in population). The English began to settle what they called New England during the early 17th Century (the Mayflower Pilgrims landed at Plymouth Rock in 1620). It is thus fitting to start our trip across the U.S. in Boston, Massachusetts.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

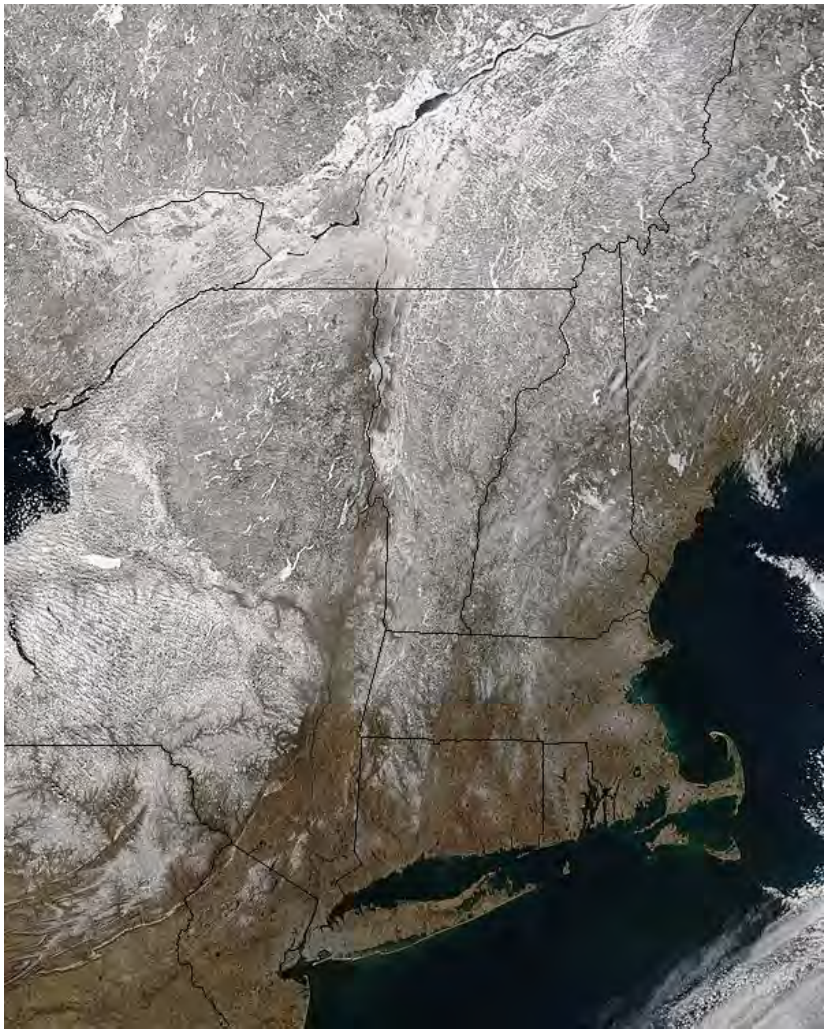
Our sojourn across the U.S., going from east to west, begins at Boston, but will take us up and down the U.S. East Coast as well. We recommend that you keep an Atlas nearby to help in locating the features mentioned in the descriptions that follow. We think this self effort is good practice for learning how to relate space imagery to geographic data sources.

New England and The Atlantic States

With the overview established on the first page for describing and interpreting the various landscapes that make the U.S. so picturesque, we now embark on a simulated flight over the continent, using mainly Landsat images to portray typical surfaces in many of the provinces discussed above. We start on the East Coast in Massachusetts (most of which is also in a lithologically distinct part of the Appalachians) and will proceed westward along a meandering line that brings us into northern California. But note: On this first page we will roam up and down the East Coast, geologically showing regions consisting of young Cenozoic rockss but also areas underlain by ancient Paleozoic rocks that relate to the Appalachians (treated more specifically on the next page).

We will start our transcontinental flight from Logan International Airport in Boston. We see it first as a Digital Globe/Google image and then as an aerial view that might be what an airlines pilot notes on approach:

Let us present a framework for the journey's beginning by looking at this large scene from space, which is a MODIS wintertime image that extends across the New England Maritime physiographic province. The New England states of New Hampshire, Massachusetts and Connecticut were part of the original 13 colonies at the time of the American Revolution. Vermont split from New Hampshire in 1777. Maine was an exclave of Massachusetts until 1820, when as a result of the growing population, it became the 23rd state on March 15 under the Missouri Compromise. This state is almost as large as the 4 original states.



Lets zero in on the starting point of our trans-continental air flight. The summertime Landsat image below covers metropolitan Boston, eastern Massachusetts, Cape Cod, and Providence in Rhode Island. The geology of this region is governed by ancient crystalline rocks that form the lower units in the now much-eroded eastern Appalachian mountain system. The entire region has been glaciated.

6-3: For this image and, if you wish for all subsequent ones examined during this transcontinental overflight, locate the scene in your atlas. This will reveal to you the name of the largest blue area (typical of a metropolitan area) and will aid in correlating text to geography as mapped. What is the main city involved and where is it? What is the blue area city at the left margin? Name the two islands. ANSWER

6-4: Locate the town of Plymouth in Massachusetts; what famous event happened there? Is Gloucester in the image; where? ANSWER

6-5: Out in the ocean north of Sandy Hook is a peculiar curved cloud pattern. It is probably not a natural cloud. Speculate on what might have caused it. ANSWER

Geologically, this region is part of the Appalachian Mountain Belt that passes northward through Nova Scotia and Newfoundland. Most of the rocks underlying the province are igneous and metamorphic. We usually associate these rock types with the deeper parts of an orogenic belt, where temperatures and pressures were higher during their formation (this belt is within the Appalachians; see next page). Their presence at the surface today implies extensive erosion that stripped away the overlying sedimentary rock units after the main episode(s) of mountain-building and uplift. The landscapes typical of New England resemble parts of the Canadian Shield. Flatter areas are interspersed with low mountains, although peaks, such as Mt. Washington in New Hampshire and Mt. Katahdin in Maine, reach elevations exceeding 1,890 m (6,200 ft) and 1,590 m (5,200 ft), respectively. The rocky soils in the region do not favor farming, so that much of the land remains in forests. The predominance of deciduous trees accounts for the widespread reds in this false color composite (above). Most of New England experienced one or more advances of the Pleistocene continental glaciers that removed soils, laid down deposits, and carved out lakes.

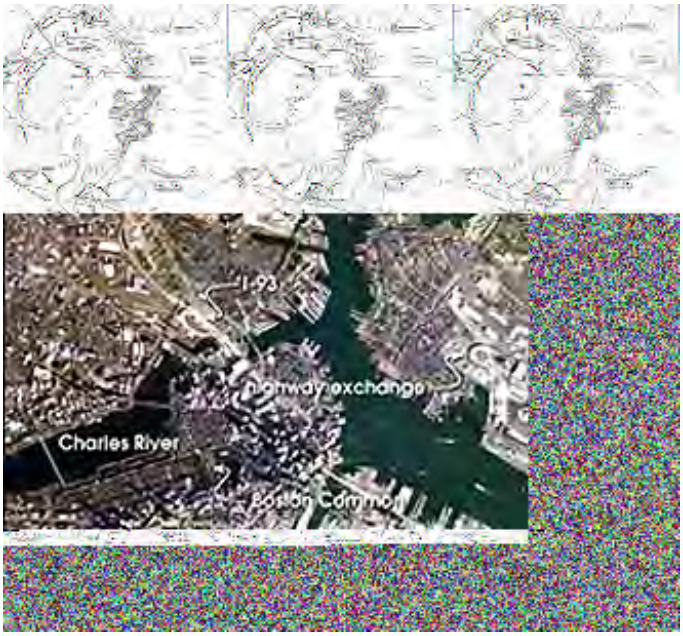
The bluish area along the Massachusetts coast is the central part of greater Boston whose

harbor lies within Boston Bay. As seen by the Landsat-4 TM, that area looks like this:

This map of Greater Boston should help you to get your bearings. Fit the scene above and the scene below in it.

An astronaut photo taken from the International Space Station shows part of inner Boston and sections to its north in some detail. The center of the city is the Boston Common; the large Logan Airport makes this a frequent destination for travelers, including those going to Europe.

Now for a history lesson. Look at this map of Boston in 1775. Compare that with the map above and the ISS image. Before reading on, try to single out the main differences:



There appear to be many changes since the Revolutionary War. At that time, Boston itself was almost an island, with only a narrow stretch of land - the Boston Neck - connecting it to the mainland. Both the Boston Harbor and the Back Bay (Charles River) were much more extensive then. What has happened should be obvious (certainly not sealevel fall, exposing land, since modern sealevel is rising slowly): huge amount of dirt fill have been dumped into the harbors, creating new land and expanding the Neck to its present width. The Charles River has been narrowed (but still wide enough upstream for sailboating). Noodle Island has been added to, enough to allow construction of today's Logan International Airport.

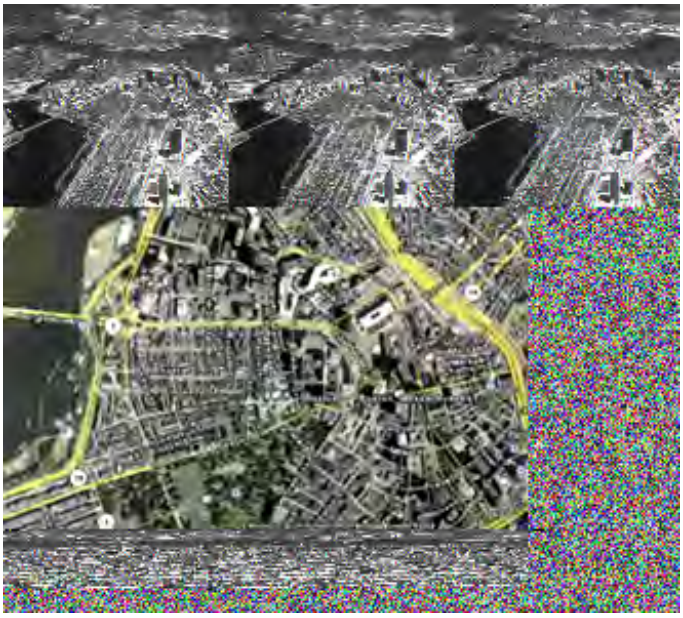
The Revolutionary War began on April 19, 1775 at Lexington and Concord, suburb towns about 35 km west. The first important battle took place on June 17, 1775 at Breeds Hill/Bunker Hill just north of Boston proper (then, a town of about 16000 people). Bunker Hill rose 110 ft from the river. This fight, commanded by British Gen. Wm. Howe, involved several charges by the "Redcoats" under General Thomas Gage that finally drove off the Colonial Minutemen and other militia under General Israel Putnam (although Col. Wm. Prescott was the actual field commander; many historians credit Putnam with the immortal command "Don't fire until you see the whites of their eyes."). Although technically this was an English victory, their heavy losses - 226 dead

(Americans 140 killed) and 828 wounded - shook their confidence in an easy quelling of the Rebellion

Gen. Howe thereafter holed up in Boston, surrounded by Americans to the north and west. His failure to capture the Dorchester Heights proved fatal. In the dead of winter, the Americans led by Henry Knox hauled captured cannon from Fort Ticonderoga in New York (which had been stormed earlier by Ethan Allen and Benedict Arnold) through snowstorms back to Boston - a feat requiring two months. General George Washington, the newly appointed Commander-in-Chief, set the cannon up on Dorchester Heights. The fear of uncontested bombardment eventually forced the British to evacuate Boston (by sea, going first to Nova Scotia, until attacking Long Island and then New York in August) on March 17, 1776 ("Patriots Day"), leaving New England untouched for the rest of the war.

This GoogleEarth image portrays most of present day downtown Boston (area at word "highway.." in the ISS image). Note how some streets curve. One can actually start down a street going west and while staying on this street end up at Boston Harbor on the east. .

An aerial oblique photo shows this downtown area, the Boston Commons, Back Bay, and the John Hancock building (foreground).



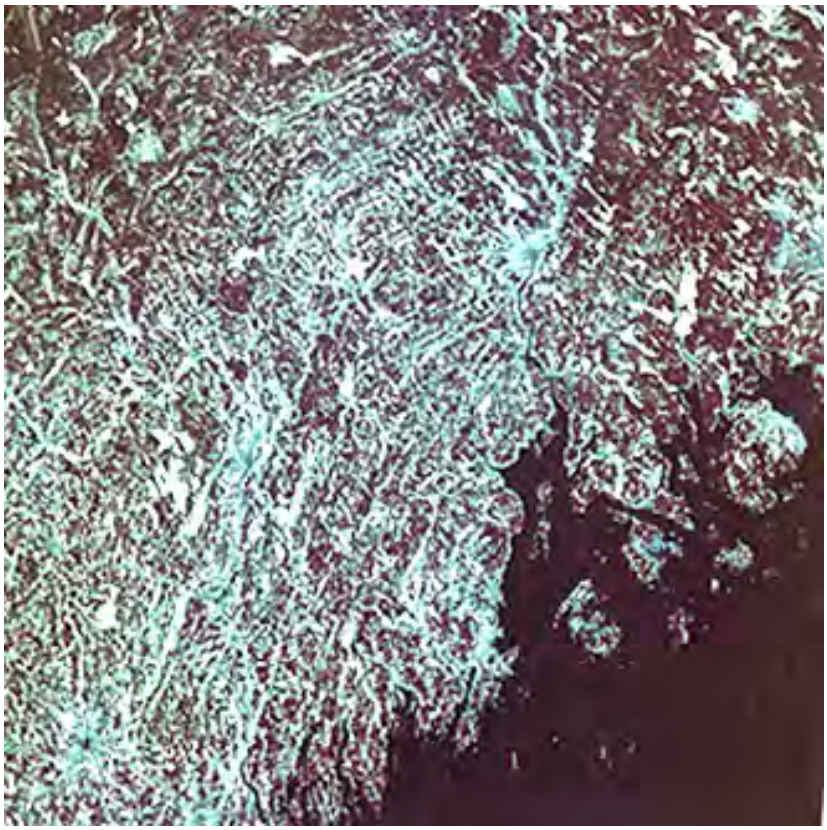
Here are an aerial view showing the campus of the Massachusetts Institute of Technology (MIT) on the Cambridge side of the Charles River and again in a 2 meter color space image made by the Quickbird satellite.

In recent years to offset the traffic snarls and confusion, tunnels under the Bay and broad limited access highways right through town have improved the traffic flow. This Quickbird image shows a part of this transit system, along with striking new modern tall rises:

A conspicuous feature in the Landsat scene is Cape Cod, an east-west landmass built from extensive end morainal deposits that mark the southern boundaries showing how far glacial ice sheets advanced. A closer look by radar shows the many small lakes in the Cape (note the canal shortcut):

The sharp bend that ends at Sandy Hook represents, in part, modifications brought on by ocean-wave action. South of the Cape are the islands of Nantucket and Martha's Vineyard, also originating from fluvioglacial action. All three are noted vacation spots. Lowlands associated with drowned rivers make up Buzzards Bay off the west end of the Cape and Narragansett Bay below Providence, Rhode Island.

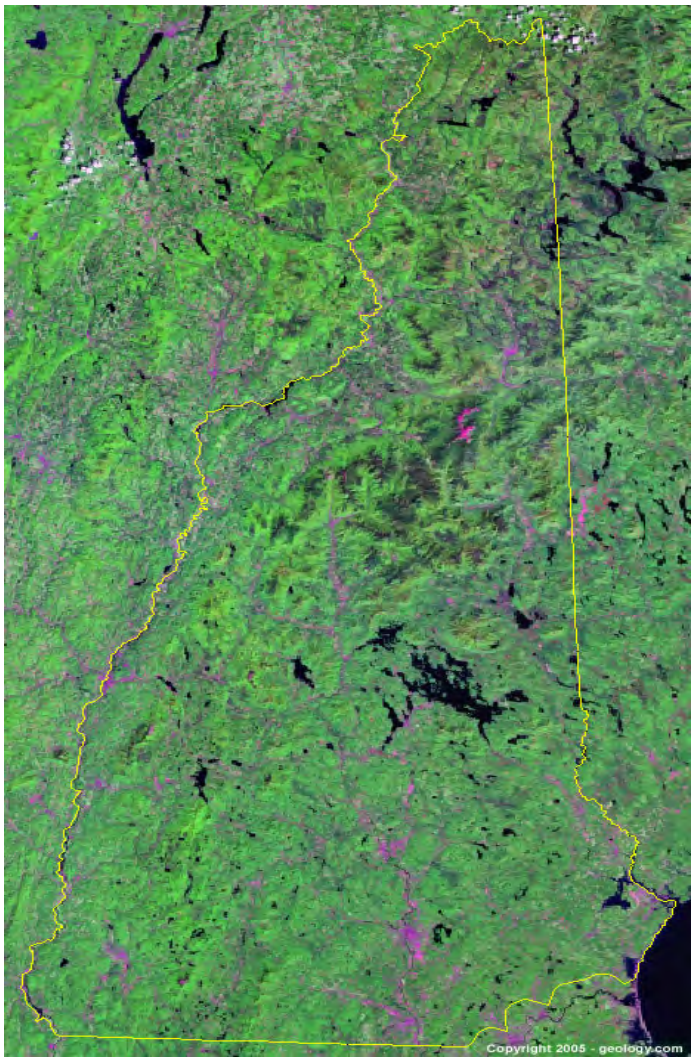
North and east of Boston is the home of the "Down Unders" - the citizens of Maine. Here is coastal and inland Maine in summer and again in winter:



The igneous-metamorphic rocks underlying much of Maine were extensively glaciated in the Pleistocene. Brown areas in the winter scene - set off by snow - are low hills which are covered by fir trees and some birch and maple trees. Lewiston, Brunswick, Augusta, Waterford, and Bangor are in this scene (since your atlas should be open, find them). The indentations caused by the glaciation are partly responsible for the rugged irregular coastline that is being inundated by post-glacial sealevel rise. The last landmass on the right of the winter image that projects into the sea is Mount Desert, on which Bar Harbor is situated.

Mount Desert is the most popular tourist destination in Maine. Here is a space image of the island and a ground scene showing its rugged Cadillac Mountains:

New England is a favorite destination for tourists in part because it has so many lovely, quaint small towns. These abound in Vermont and New Hampshire. All of the latter state and part of Vermont are shown below (note that both states are mainly low mountains and many lakes; the cold winters are ideal for skiers). Beneath it is the town of Portsmouth, NH, near the coast and a splendid place to sample the New England flavor.



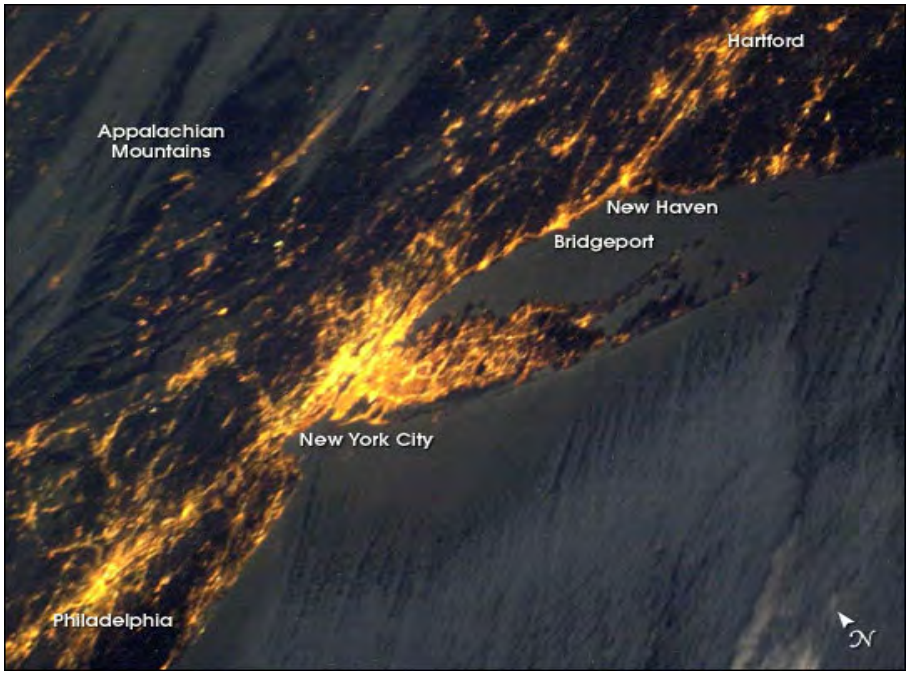
Near the bottom left of the first Landsat scene above is another large blue area - the characteristic signature of a metropolitan region. Its location on Naragansett Bay gives its identity away - it is Providence, capital of Rhode Island, the U.S.'s smallest state. Here is a picture of its downtown buildings:

This is Providence, as seen from space (much of the green is suburban, with a moderately high density of homes):

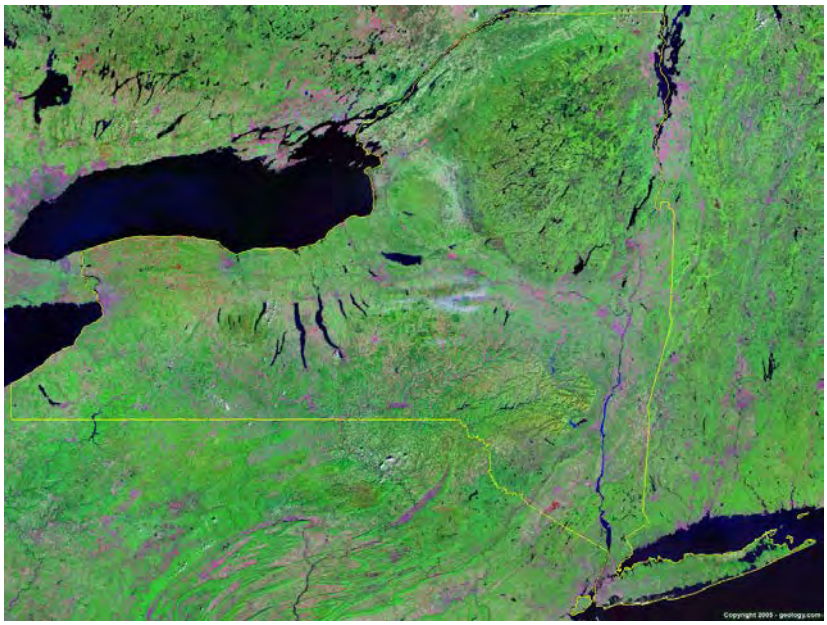
Southern New England is also made up of crystalline rocks that are an extension of the Piedmont to the southwest. The Landsat scene below shows early leafing of trees in Connecticut. The Hudson River is on the west; the smaller Connecticut River to the east. New Haven (home of Yale University) is the blue patch along the upper reach of this latter river. The terrain in Connecticut is mostly low rolling hills, cut into crystalline bedrock (its fracturing is faintly evident).

The capital and largest city in Connecticut is Hartford, on the Connecticut River. This is often cited as the Insurance Capital of the U.S.:

On page 4-3, we commented on the concentration of population in the so-called East Coast Metropolitan Corridor that extends from Boston through New York City, then Philadelphia and Baltimore-Washington to Richmond. Much of that corridor is dramatically highlighted in this nighttime scene from space, because of the bright lights of the cities and their suburbs:



At this early stage in our trip, we will set a precedent that will recur on most of the other pages covering the areas visited in the cross-country journey above the U.S. We will depart significantly from the main geographic region featured on each page to look at other regions often considerable distances away. Let us start with a look with the state of New York, shown here as a mosaic using Landsat and other satellite data. We will look at 4 principal features: 1) New York and Long Island; 2) the Adirondacks; 3) the Finger Lakes; and 4) Buffalo and Niagara Falls.



This is a general physical features map of New York state:

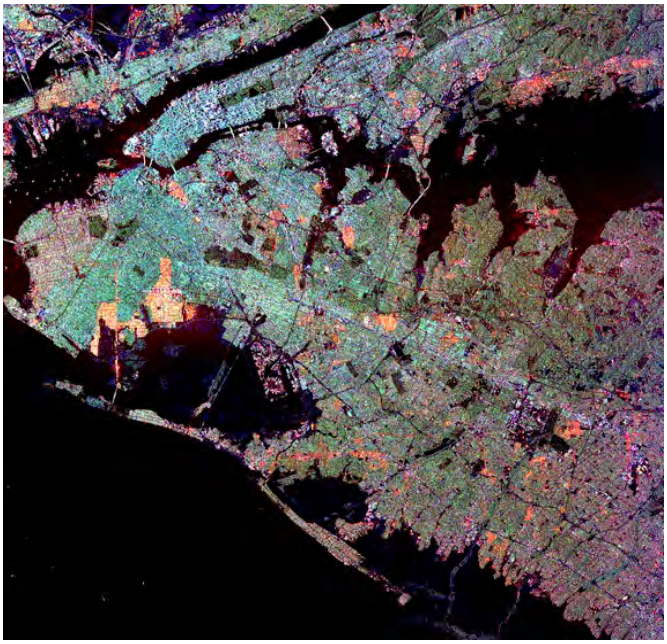
We start first with the greater New York City metropolitan area, which you earlier saw imagery from on page page 4-2. (Check the map on that page for locations of the 5 New York City boroughs.) This next image includes western Long Island, the 5 boroughs of New York city, and many northern New Jersey towns, has a population of more than 21 million. Across from New York City itself are several New Jersey cities with high population densities. This photo, taken from the International Space Station, shows Manhattan, Jersey City, Hoboken, North Bergen, Kearney, and a bit of Newark:

Manhattan, of course, is one of the great cities of the world - the heartstone of cosmopolitan living. This ground photo shows the Empire State Building in midtown and Lower Manhattan in the background.

Many who work in New York City live on Long Island to the east. Long Island, like Cape Cod, owes its origin to Pleistocene glaciation. A series of moraines make up the land mass that lies between Long Island Sound and the Atlantic Ocean. Here it is from space:

The image below is almost an oddity. It is a radar view that shows most of New York City and the western part of Long Island. What was intriguing is the "apparent" jagged coastline of Long

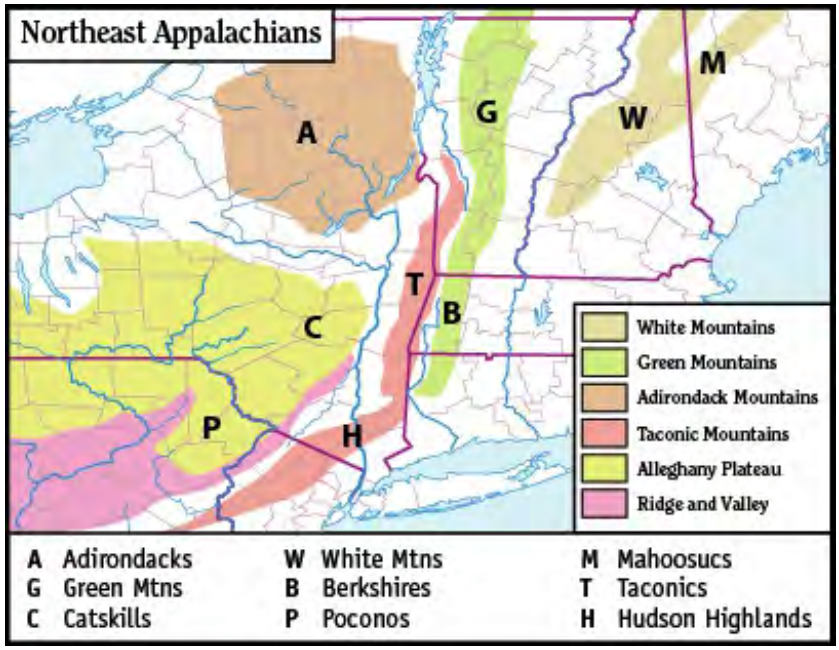
Island Sound, and to a lesser extent the Atlantic coast beaches; this is somewhat illusory as an artifact of the radar imagery.



The Adirondacks and the Catskills are the two principal mountain groups in New York State. They are both displayed in this perspective image which was made from Landsat and SRTM data combined:

The Catskills (right side of the above image) are dissected Paleozoic sedimentary rocks. You saw a Landsat mosaic of the Adirondacks (left side of image) near the bottom of page 2-8.. The Adirondacks are an extension of the Canadian Shield and consist of Proterozoic (Precambrian) metamorphic rocks and some igneous rocks. They are distinctly mountaineous (their highest peak is Mt. Marcy at 1030 m [5345 ft].

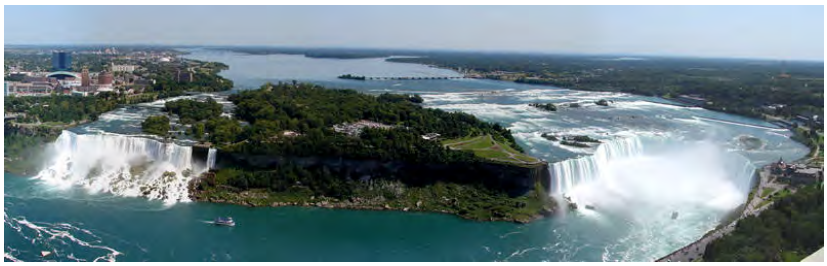
While we are on the subject of mountains in the Northeast, most of the principal groups given specific names are shown in this map:



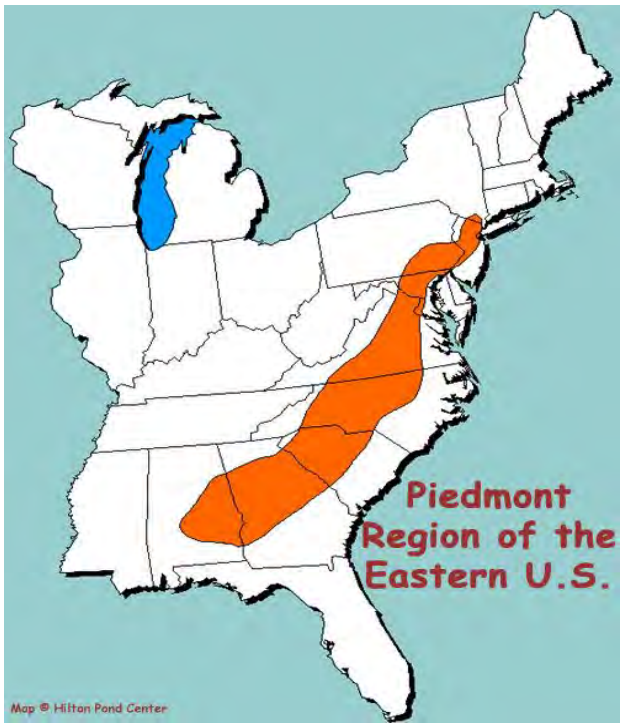
West of these mountains are the Finger Lakes, another popular recreational area. The lakes were scoured out by continental glaciation and then filled by drainage waters after the ice disappeared. We see them in this astronaut photo and identify them from the map below it (note where Syracuse and Rochester are):

At the west end of New York State, between Lakes Erie and Ontario, lies Buffalo and to its north Niagara Falls. We see these first in a Landsat subscene, then the Falls from SPOT and in an IKONOS image:

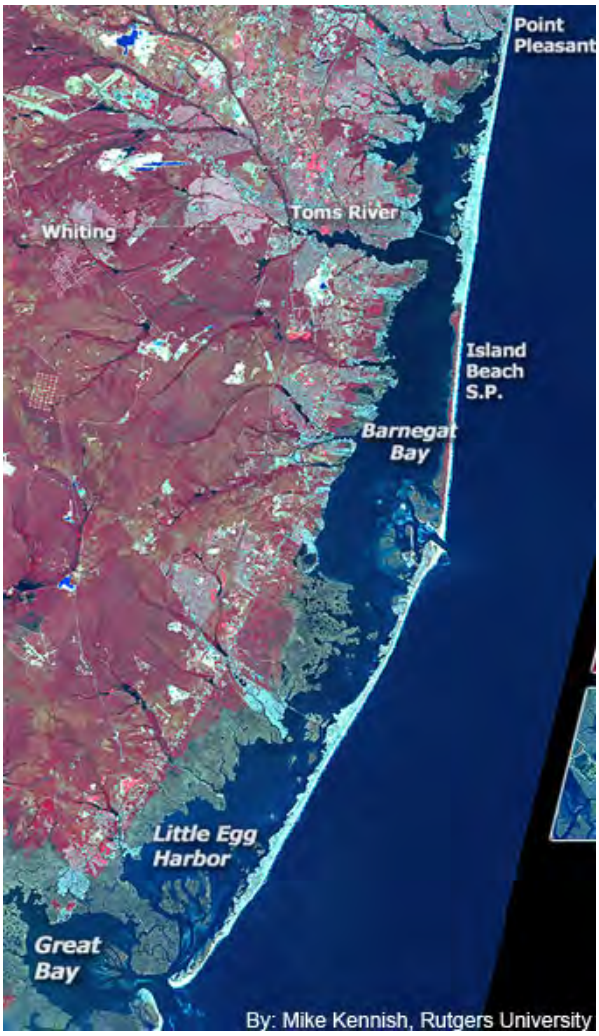
Here is a panoramic view looking south at the American Falls (left) and Canadian Falls (right). The Niagara River (less than 12000 years old) flows from Lake Erie to Lake Ontario (north) and passes over the Niagara Escarpment (Silurian) which is the reason for the Falls (which are slowly receding southward by undercutting erosion):



In the remainder of this page we will examine two segments of the very extensive Atlantic-Gulf Coastal Plains and Piedmont Provinces that begins in New Jersey (to its north, this unit is now almost entirely under the sea as part of the Atlantic Shelf) and continues along the Atlantic Seaboard (Florida is usually considered a special landmass included in Coastal Plains but with a different geological history) and then wraps around the coastline of the Gulf of Mexico, with the Coastal Plains continuing south to the Tex-Mex border (and beyond). The map below shows the Piedmont of the U.S.; every area to its east and south is the Coastal Plains:



Much of New Jersey is rural and supports farming in the Coastal Plains which make up its southern 2/3rds (Piedmont and a little Valley and Ridge in the northern segments). Its coastal region is popular as a seashore vacation spot. Much of its coastline consists of barrier islands; inland are woods still not cleared (the Pine Barrens):



Barrier islands are elongate, wide mounds of sand built up when these particles are carried by waves that strike the coastline, loose their velocity (carrying power), and deposit the load along

the shore. Barrier islands occur along much of the Atlantic coast from New Jersey south to southern Florida.

We look next at the largest city in Delaware ("The First State"), Wilmington. The Christina River flows into the Delaware River (off the image to the right). Parts of the town lie on the Piedmont; other parts to the south are on the Coastal Plain:

As seen from the I-95 Interstate by-pass, this is Wilmington's skyline:

Going southward on the Piedmont, we next note the present day capital of Virginia, and onetime

capital of the Confederacy, Richmond, on the James River:

We travel next to the coastal cities in southeastern Virginia. It may surprise one to see how extensively built up the region is. Look first at the map (Norfolk is not named but is at the Interstate 64 and 564 signs; News is part of Newport News) and then the Landsat image:

The northern edge of the Great Dismal Swamp is in the SW corner of the image. The region contains the home base for the U.S. Navy's Atlantic Fleet (including aircraft carriers). NASA's Langley Research Center (mostly aeronautics) is located in Hampton:

While still in the Coastal Plains, let's look at a popular vacation spot in the East: the Outer Banks of North Carolina. The Banks are a series of barrier islands built up over many millenia by waves carrying sand ashore as deposits that remain stable except during nor'easters (a fierce storm) and hurricanes that erode the beaches.

Then we will move into interior North Carolina, in the Piedmont just west of the Coastal Plains. We will look at three well known cities - Raleigh (home of North Carolina State University), Durham (Duke University), and Chapel Hill (University of North Carolina). In between them is the famed Research Triangle, the largest of its kind in the U.S. First locate them in this Landsat image (use the water reservoirs as a guide), aided by the map:

The largest of the three cities is Raleigh (about 350,000):

This late Fall image taken by Landsat covers the Atlantic Coastal Plain of part of South Carolina, much of which is now farmlands. Coastal pines dominate the dark areas near the coast. The Santee River has been dammed at Lake Marion, which drains into Lake Moultrie.

At the end of the drowned river valley coming off Lake Moultrie is the famed city of Charleston,

rich in ante-bellum homes. This city, along with Savannah, GA and New Orleans, LA have well-preserved sections that are a hallmark of the Deep South. Charleston was much involved in both the Revolutionary and Civil Wars. The main city is a "Peninsula" formed at the confluence of the Ashley and Cooper Rivers. Here is a close-up made from Landsat-7 imagery.

A general aerial view of Charleston appears next. Along the eastern waterfront are some of the old homes of Charleston. These homes postdate those burned in the Revolutionary War, when Sir Henry Clinton captured the city at the start of the final (southern) phase of battle that ended with Yankee victory at Yorktown.

But Charleston is even better known as the town in which the Civil War began. The first shots were a cannonade of artillery fired across the harbor into the Union's Fort Sumper which was on an island loosely attached to the coastal wetlands. The fort was forced to surrender.

Another city of Civil War fame is Savannah, GA, less than 100 miles to the south.

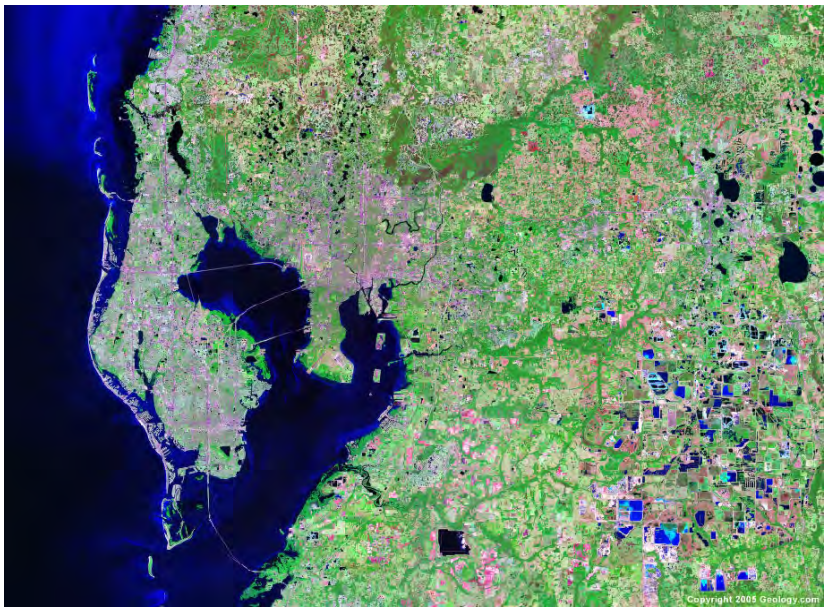
Savannah was the destination of Union General Wm. Tecumseh Sherman at the end of his (in)famous march through Georgia after the fall of Atlanta. The city still has many antebellum homes (he declined to destroy them, unlike his actions at Atlanta), which occupy the

conspicuous city blocks that are true "squares".

We will see several other space images of the Coastal Plains (for example, New Orleans, LA) elsewhere in the Tutorial.

South of Georgia is, of course, Florida. Most of that state does not fit the classic Coastal Plains category. Florida, geologically, is an arch added to the North American continent relatively recently. Most of its bedrock is limestone. Here is an unusual false color image of the entire state:

We saw Miami and Orlando in Section 4. Here we elect to represent Florida by its two big Gulf of Mexico cities - Tampa and St. Petersburg:



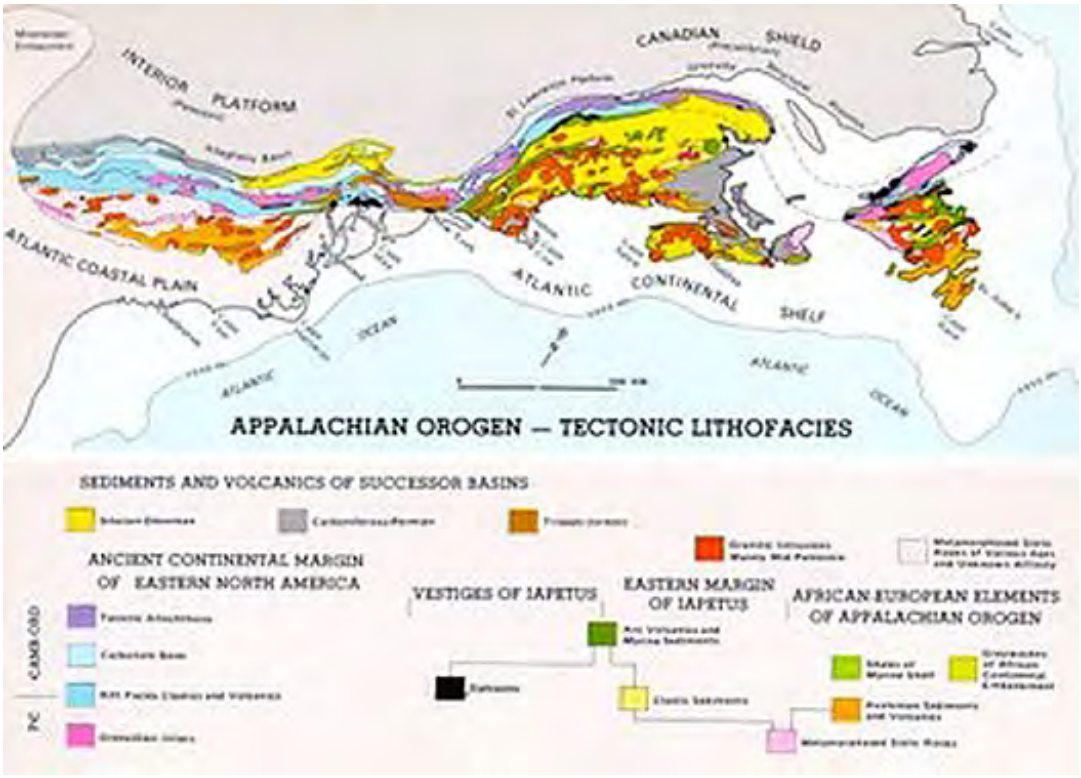
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

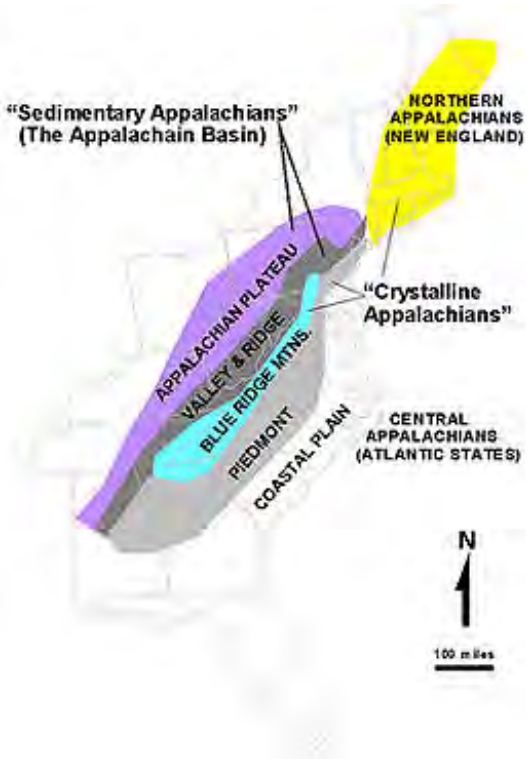
Our imaginary airplane flight now passes over central Pennsylvania. The dominant landscape feature is the zig-zag folds that characterize the Valley and Ridge subprovince of the Appalachians. The structural expression of this mountain system varies along its length as demonstrated by several other space images.

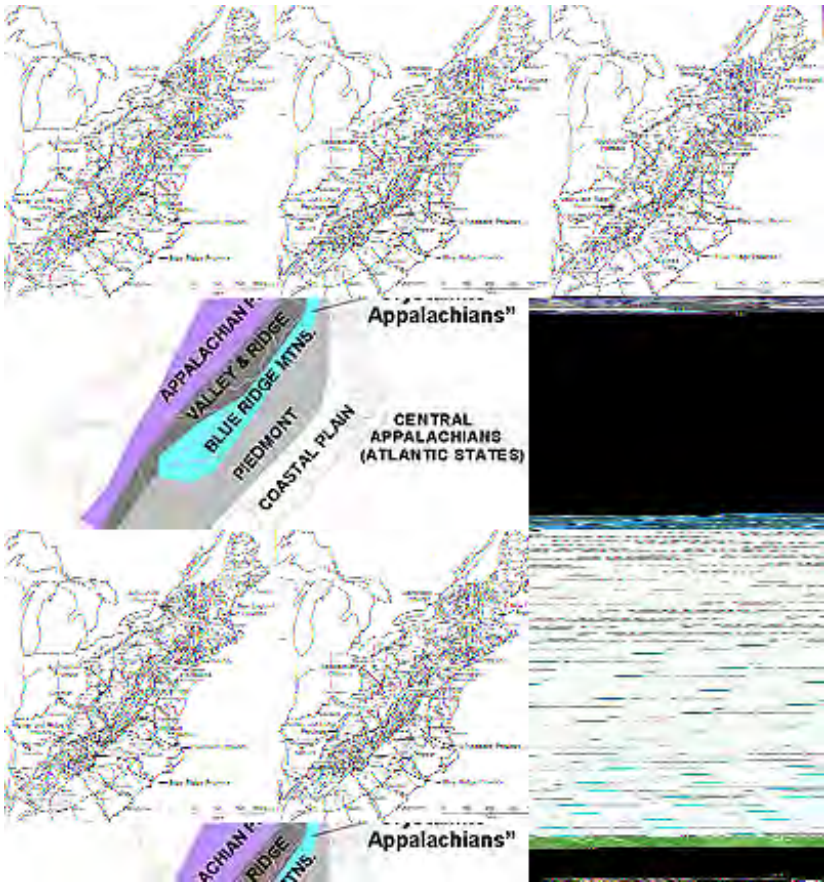
The Pennsylvania Fold Belt; The Appalachians

Our odyssey proceeds west into the heart of the Appalachian fold belt that dominates the landforms of the Appalachians. The U.S. Appalachians extend into northern Georgia and Alabama, south of which they go underground, being buried by sediments comprising the Coastal Plains. Actually, the general orogenic belt that includes the Appalachian Mountains of eastern North America extends westward across the Mississippi River through the Ouachita Mountains of Arkansas/Oklahoma before dying out in Texas. Its northern end passes through Maine into Nova Scotia and Newfoundland and has been traced into the Caledonide mountain systems in the British Isles and Norway. The Atlas Mountains of northwestern Africa also link to the Appalachians. This once continuous belt split at the end of the Paleozoic Age, as the present continents broke apart and their tectonic plates then drifted in several directions. Before then, the Appalachians built up over about 300 million years as subduction zones, back arc basins, and accreted terranes were added to ancient continental nuclei that eventually led to the supercontinent of Pangaea. Three major orogenies mark the history of the Appalachians: Taconic (Ordovician in age), Acadian (late Devonian), and Alleghenian (Permian). Pangaea split apart about 250 million years ago, with one segment (the Appalachians) embedded in the North American Plate and other remnants remaining in the plates that include Europe and Africa (Gondwanaland). The complexity of the geology is evident in this tectonostratigraphic map modified from one made by H. Williams of Memorial University in Newfoundland.

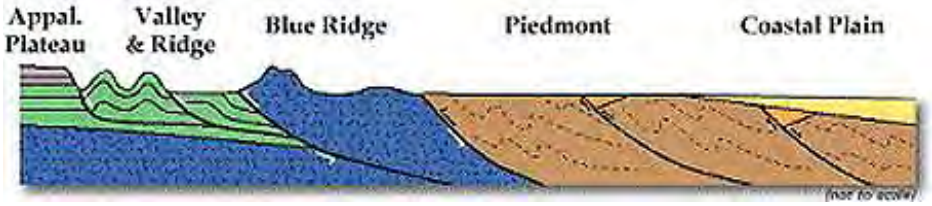


The physiographic subdivisions of the North American Appalachians are indicated in these two maps:

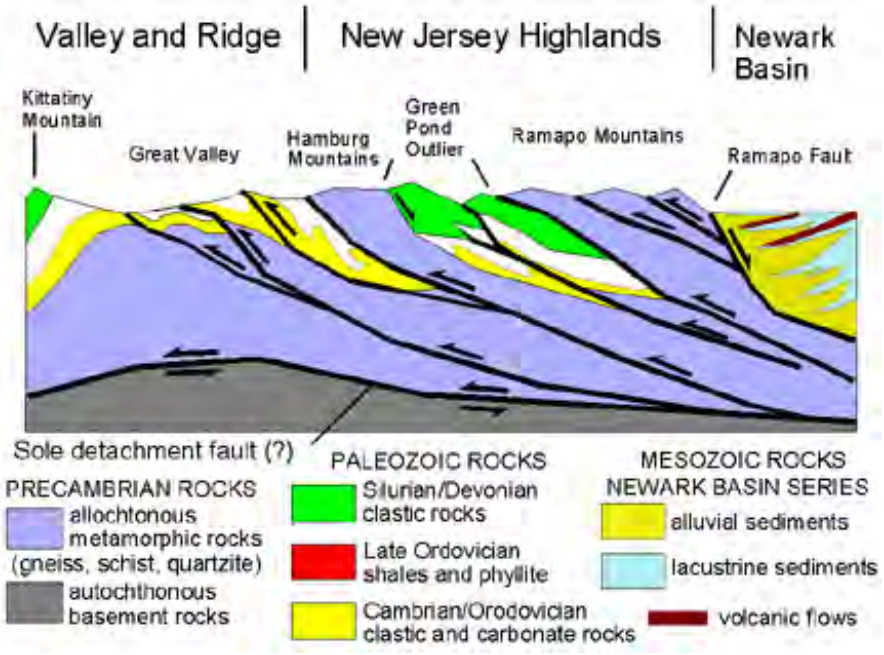




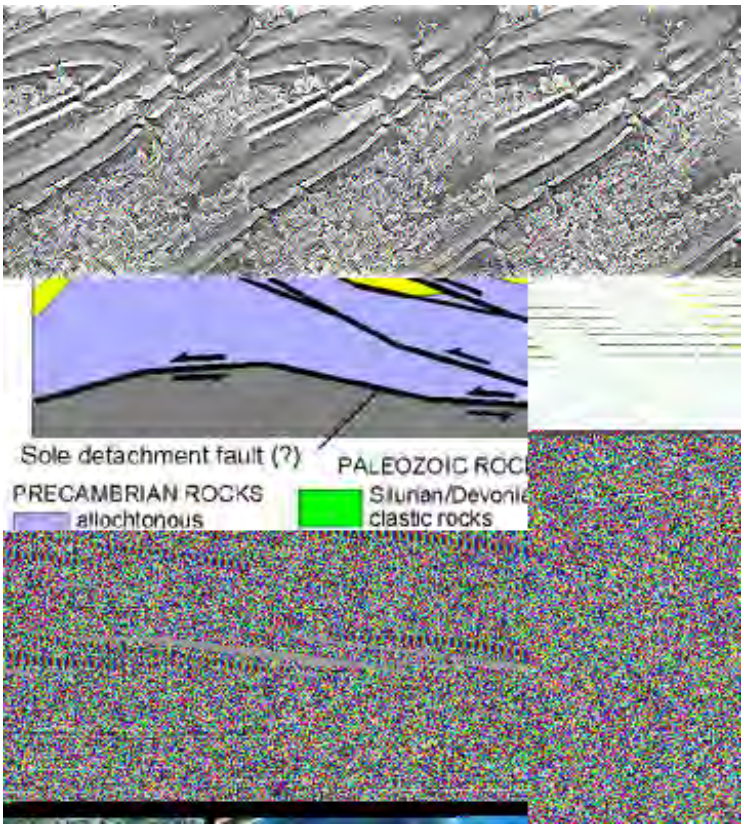
Another illustration that helps to understand the Appalachians as a structural complex is this cross-section of the main geologic components that are now underlying the surface physiography:

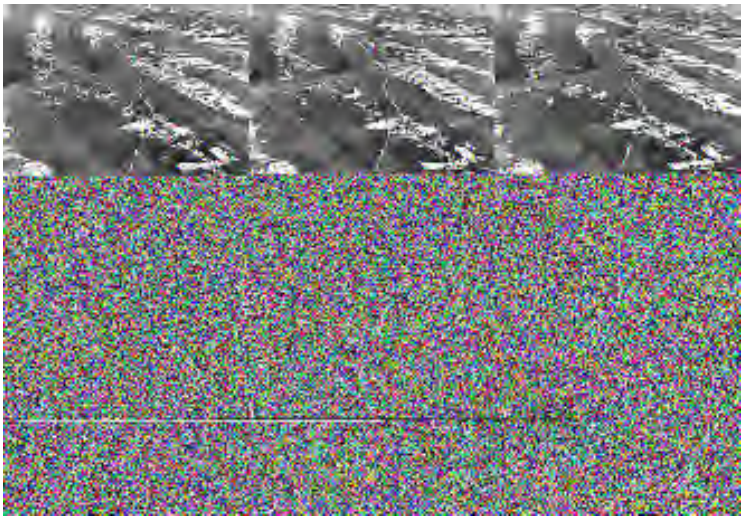


This cross-section indicates that the Appalachians are made up of a series of segments each bounded by a major thrust sheet or belt. That is brought out in more detail in the next figure which shows the sheets present in the Appalachian portion of New Jersey:

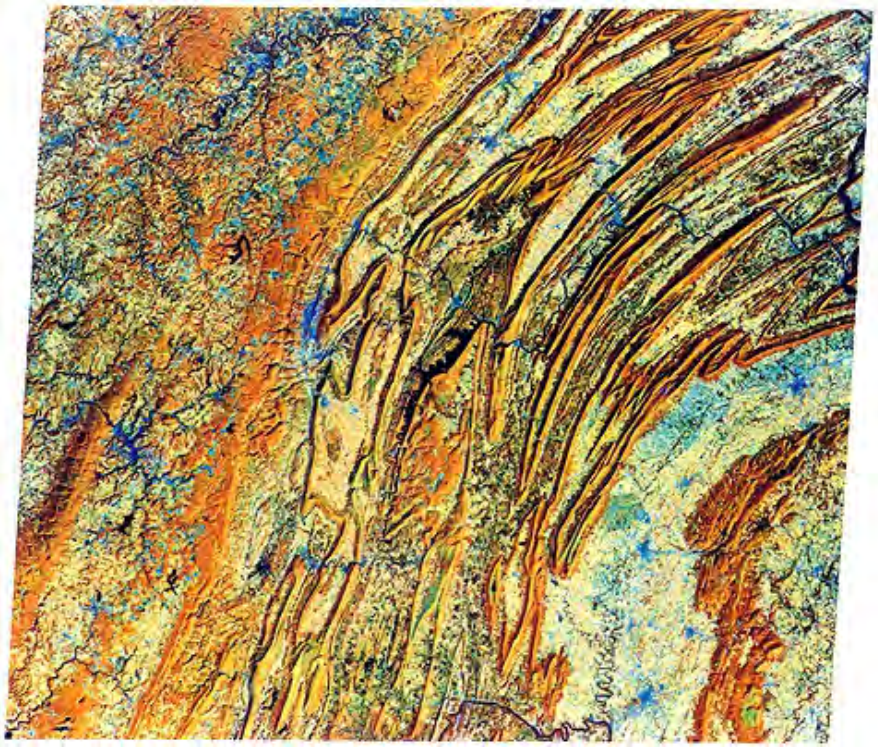


In central Pennsylvania, the segment known as the Valley and Ridge has very distinctive topography: the ridges are curved, some shaped almost like an arrowhead, and have generally uniform elevations (not pointed or irregular like many western mountains). They are even today forest-covered but the valleys in between are flat to rolling and commonly developed for agriculture. The next two aerial photos, the first vertical and the second oblique shows typical expressions of this land use - forested ridges, valleys almost completely given over to farming:

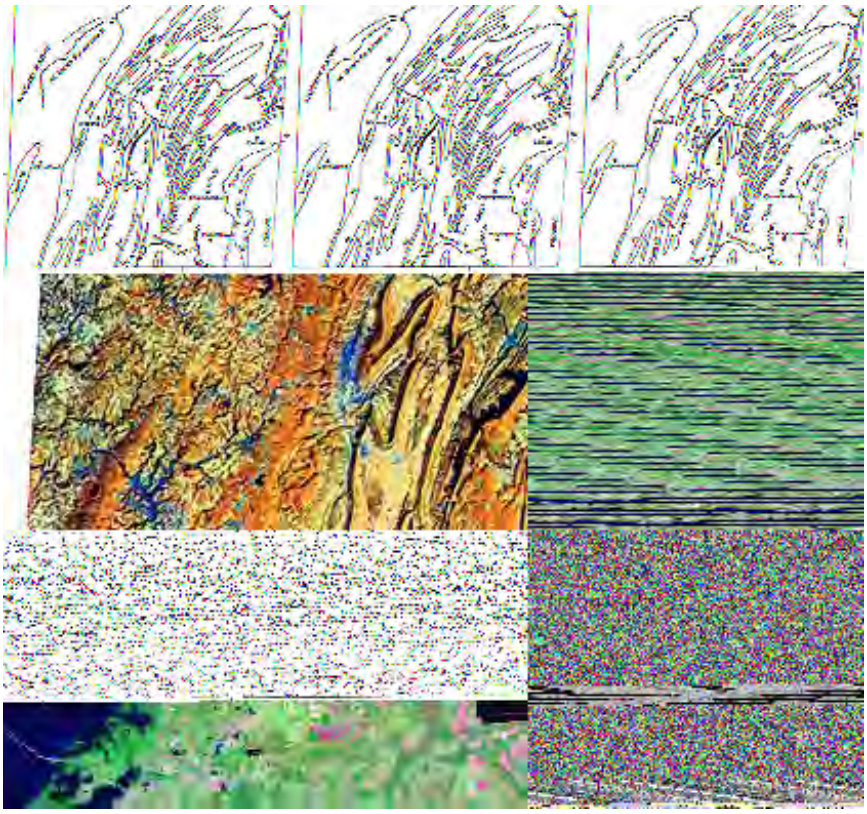




That second photo is included in this unusually colored Fall Landsat MSS composite. Most of the interior of this image depicts the Valley and Ridge province which runs from southernmost New York state to Alabama. The area shown includes the center of Pennsylvania.



The following sketch map names the principal mountain ridges and other features found in this scene:

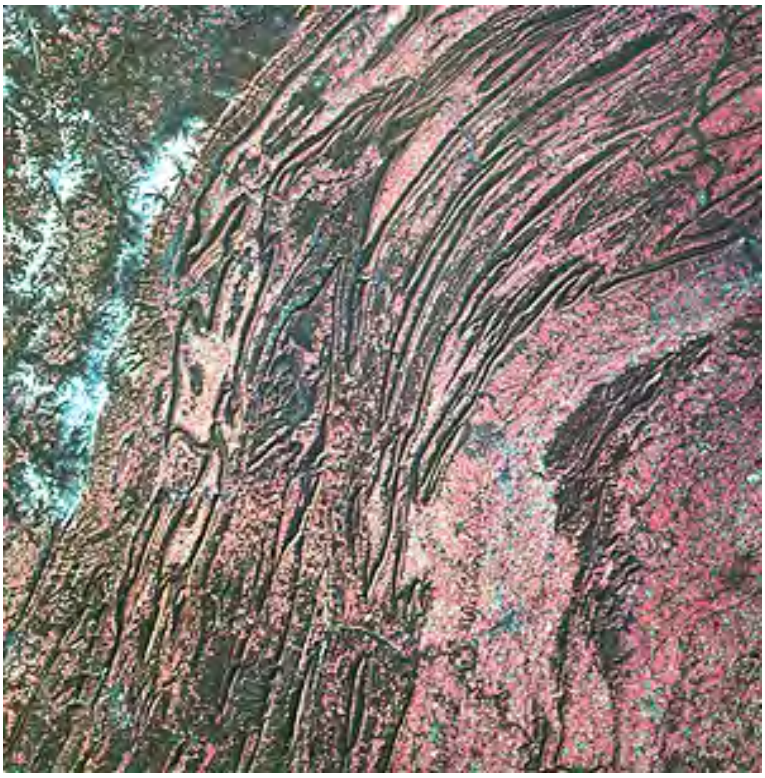


Here, the sedimentary rocks that compiled thicknesses up to 15,250 m (50,000 ft) in parts of the geosynclinal deposition basin during the Paleozoic era, were strongly folded into anticlines (uparches) and synclines (downbends). These rocks were also shoved westward along thrust fault sheets. Erosion since the Appalachian Revolution has repeatedly sculpted ridges out of harder rocks (most commonly, sandstones) and valleys out of weaker rocks (such as shales). This erosion results in a zig-zag pattern developed from folds that also curve sharply as they pinch out, as seen in plan or horizontal perspective. The ridges, few of which are higher than 460 m (1,500 ft), remain heavily forested but the valleys are mostly cleared for agriculture. Toward the lower right corner, shales and limestones of early Paleozoic age underlay a broader lowlands, the Great Valley of Pennsylvania (the northern extension of the Shenandoah Valley of Virginia).

The wide isolated ridge to the southeast is South Mountain/Catoctin Mountain, a complex of igneous-sedimentary-metamorphic rocks, which connects with similar rocks in the Blue Ridge in Virginia. The Appalachian Plateau occupies the western (left) part of the image and continues for several hundred kilometers into eastern Ohio. Its rocks are also Paleozoic sedimentary units that were not strongly folded (the compression effects die down in this region) and are still subhorizontal. However, the rock units were lifted en masse during the orogeny and, more recently, to elevations high enough to foster deep local stream cutting. The stream courses here are typical of the dendritic drainage pattern that develops on elevated terrain underlain by rocks of similar resistance to erosion. The topography of the Appalachian Plateau is distinctly mountainous as indicated in the aerial photo below. Some individual mountains have flat tops in accord with the control by underlying horizontal rocks.

6-6: The town of State College is the home of the Nittany Lions, the football team at Penn State University. It is situated in "Happy Valley" and is overlooked by Nittany Mountain; find it in the Landsat image. Deep in the woods of South Mountain is a famed location where U.S. Presidents like to go; name it. ANSWER

The Valley and Ridge topography is visually brought out more dramatically in winter scenes, when the sun angle is lower, so that shadowing is emphasized. This Landsat-1 image illustrates this effect:



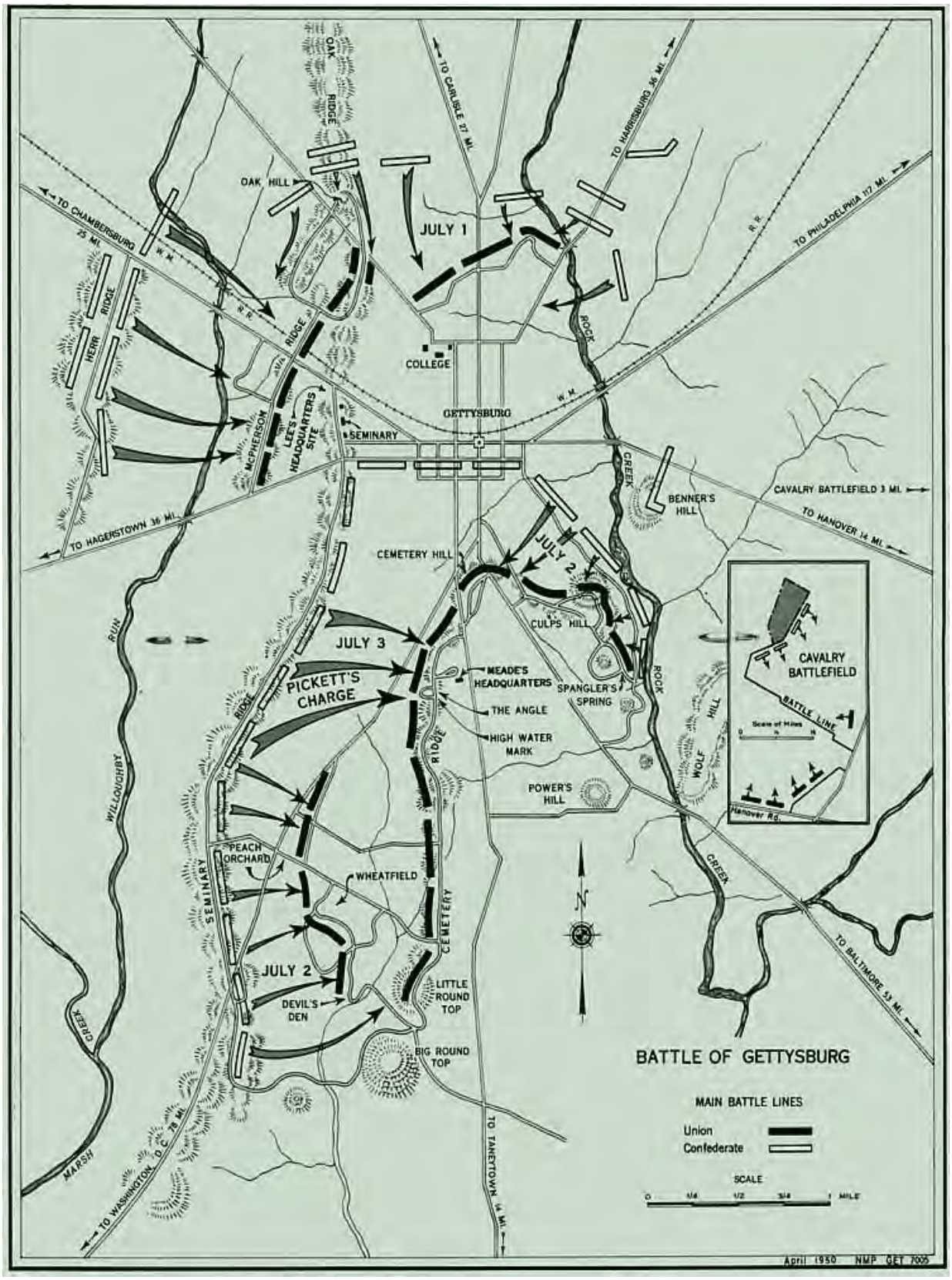
Radar has produced many splendid images of the Folded Appalachians, since the ridges stand out sharply because of the radar shadow effect. This X-band image, acquired from an aircraft by Intera Technologies of Ottawa, Canada, shows a swath 50 km (31 miles) wide that is contained in the Landsat image (you can locate it in that image by the distinctive fold pattern near the center of the radar scene). Look direction is from the west (left). The city of Altoona, PA is the bright, stretched out speckled pattern just above left center. The upper left corner shows the topography of the Appalachian Plateau, another subprovince of this mountain system.

So far, we have concentrated on the geology and physiography of the Appalachians in Pennsylvania. We will digress for a bit to look at an historical event that took place in Pennsylvania on July 1st to the 3rd, 1863. We focus on the town of Gettysburg, at that time a small hamlet where shoe making was, along with farming, the main means of livelihood. Today

the town is larger but is still not visible in this Landsat image (it is near the center bottom):

Even today Gettysburg does not reveal its history when seen in a higher resolution space image:

Yet on those fateful three days in 1863, the Union and Confederate Armies, under Generals Meade and Lee respectively, fought the biggest battle ever waged in the Western Hemisphere. This map summarizes the battle:

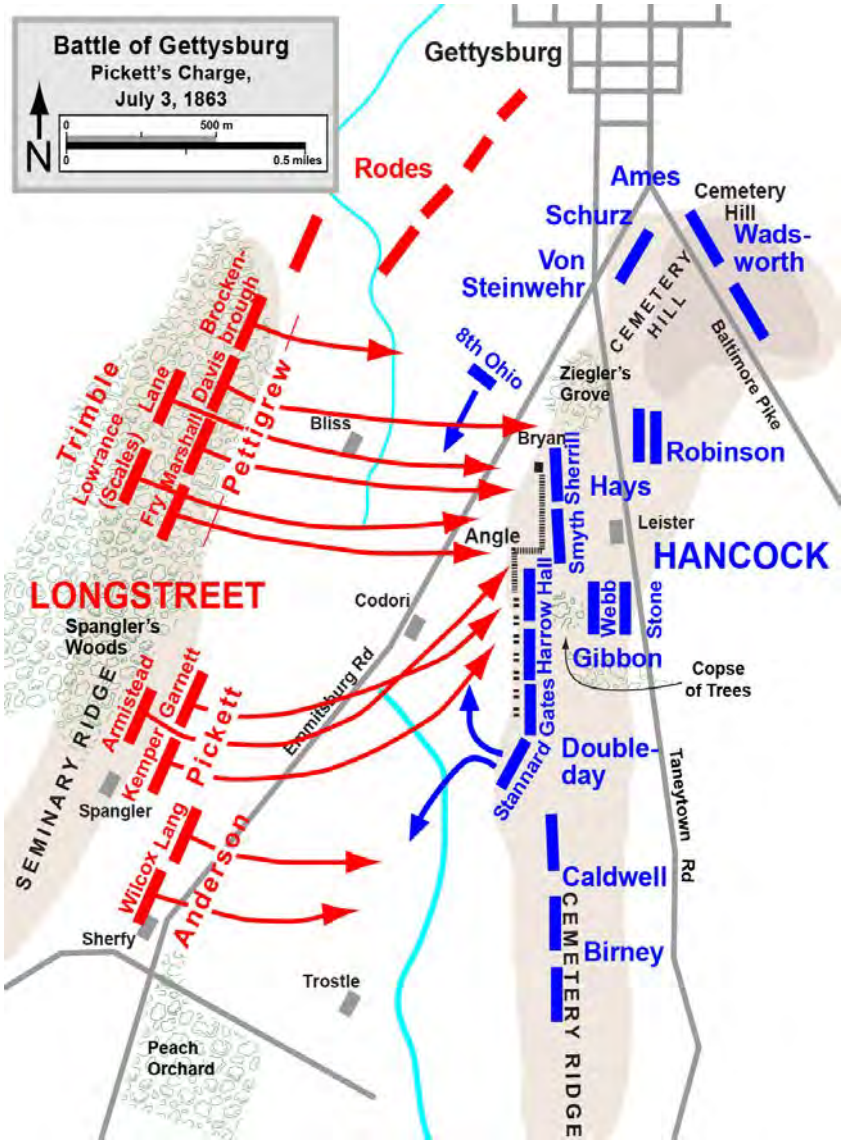


The first day's battle took place west of the town near the Lutheran Seminary. It was a standoff. The second day's battle was crucial. Lee sent Longstreets's army to the south in an effort to get around the Union left flank on Cemetery Ridge. Bloody battles took place in the Devils Den, the Wheat Field and the Peach Orchard.

But the seminal moment of the battle, and perhaps the whole Civil War occurred on the 2nd day. This is superbly re-enacted in the extraordinary Ted Turner movie "Gettysburg", which we strongly recommend that you vie (it's on a DVD). It is based on Michael Shaara's Pulitzer Prize winning novel "The Killer Angels". At Little Round Top, Col. Joshua Lawrence Chamberlain positioned his 20th Maine regiment as the southern anchor of the Union lines. If Col. Ames and his Alabamans of the Confederate Army could overrun the 20th Maine, the whole Union line

would have been "rolled up", probably leading to a huge Confederate victory, and perhaps a decisive situation that would have ended the Civil War in the South's favor. Here is a clip from the movie showing Chamberlain (played by Jeff Daniels) as he led his soldiers on a charge (after issuing the simple command "Fix Bayonets") that captured Big Round Top:

On the 3rd day, General Lee decided on a direct frontal assault from Seminary Ridge against the center of the Union lines on Cemetery Ridge. This is a map synopsis of this most famous of all American charges, which depicts the convergence at the place they named the Angle, with its (then) lone tree.:

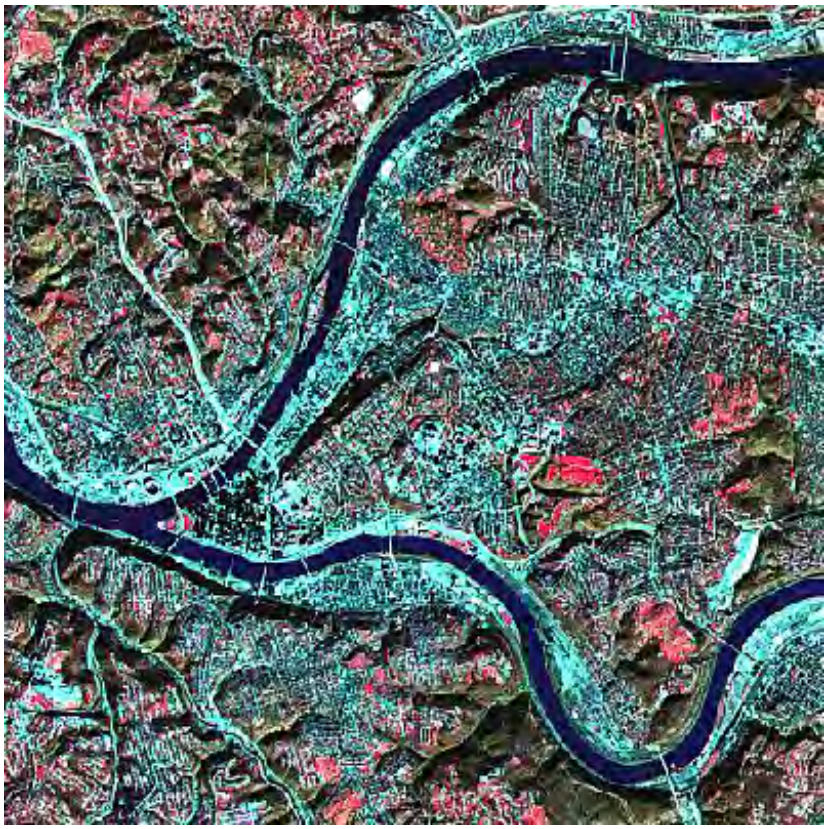


Some 17000 Confederate troops left the woods onto open fields in what is known as *Pickett's Charge*. Some 'Rebs' broke into Union lines but were overwhelmed. This led a a mass retreat after nearly half of the participating Confederates were wounded or killed. Lee's pathetic comment "It's all my fault" summarized the debacle. Gettysburg was the high water mark of the South's fight against the North, and in the view of historians stands as the beginning of the end of the Confederacy. Here is a photo of a re-enactment of Pickett's Charge (this is done every year by volunteers who make such military events one of their hobbies):

All told, some 51000 casualties (killed, wounded, captured) on both sided attest to this biggest of all battles in the Civil War (and the largest ever in the Western Hemisphere). Today Gettysburg is the most visited site in America's military history. Throughout the battlefield are hundreds of monuments, along with cannon. Typical is this scene by the lone tree at the Angle, the point of farthest penetration during Pickett's Charge.

Resuming our journey, let's zero in on Pittsburgh. This wide-angle photo is a panoramic view of the Golden Triangle and downtown Pittsburgh.

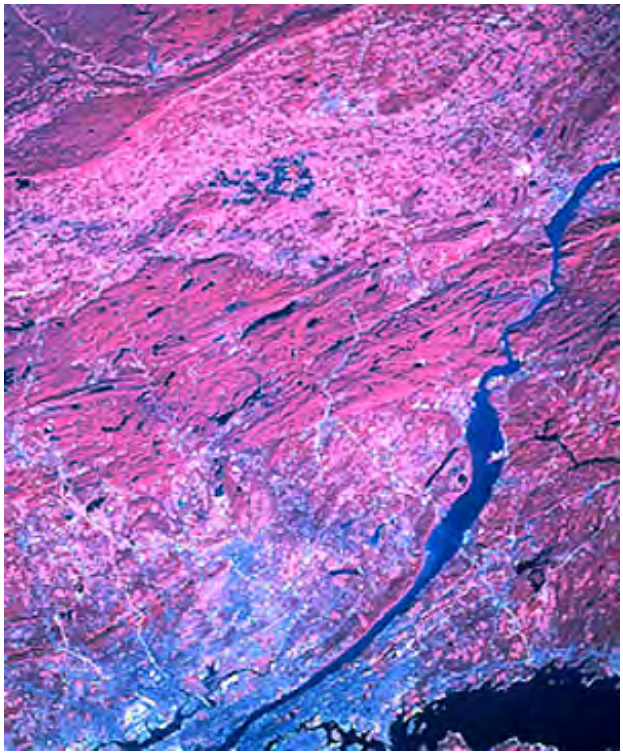
Now, fit that view into this Terra ASTER image of greater Pittsburgh:



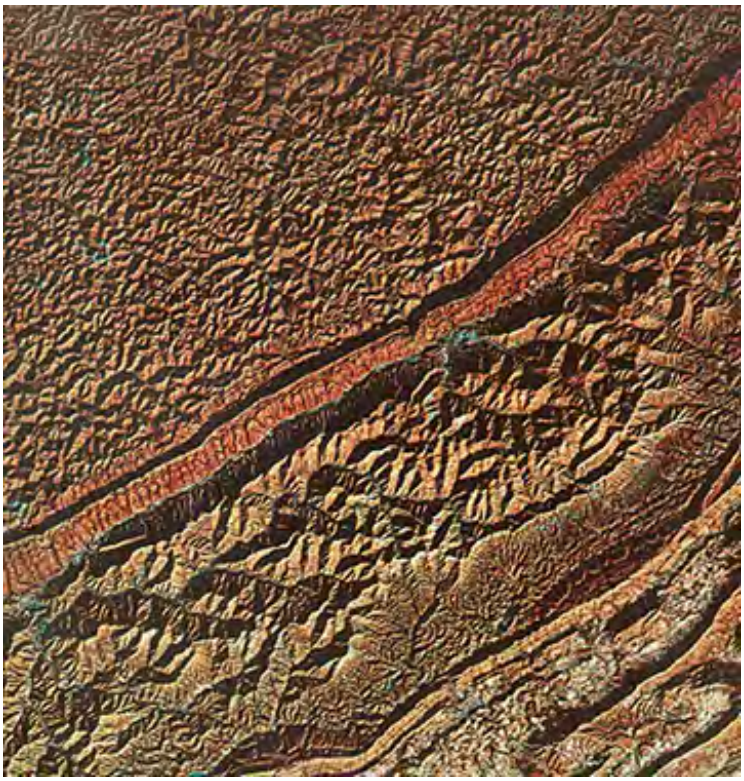
Much more detail is evident in this Quickbird 2 meter resolution image; the stadiums used by the (2009 World Champion) Pittsburgh Steelers and the Pittsburgh Pirates are conspicuous near the left edge:

Two rivers, the Monongahela on the south and the Allegheny on the north, meet at the triangle to form the mighty Ohio River as these waters all move towards their meeting with the Mississippi.

Lets leave Pennsylvania. We now take a detour in our flight to look at parts of the northern, central and southern Appalachians. The first scene - an astronaut infrared photo - shows the Piedmont as seen towards the northwest, with the Greater New York area in the bottom center. Below that is a roadcut that displays strongly folded metamorphic rocks - the most prevalent rock types over most of the Piedmont.

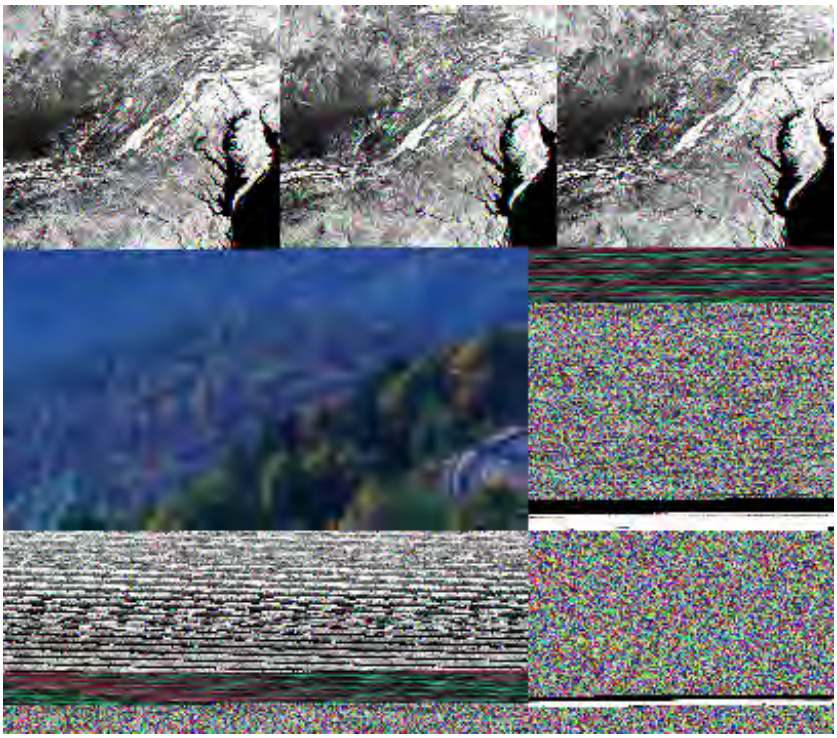


This Landsat scene includes part of western Virginia and eastern Kentucky:



This scene shows two distinctly different topographic expressions: 1) ridges and a broad valley (extension of the Shenandoah Valley), and elongate mountains lying mostly in western Virginia, which here includes the Fold Belt of Pennsylvania, now much narrowed, and 2) dissected plateau rocks forming small hills constructed from horizontal rocks which are incised by a network (dendritic) of small streams. The ground photo below was taken in western Virginia and shows the topography typical of the Blue Ridge Mountains.

The next scene from space shows much of the Appalachian Belt from Pennsylvania through North Carolina, as depicted in this Day-Vis (0.5 - 1.1 µm) image made by the Heat Capacity Mapping Radiometer on HCMM (see Section 9-8 for a description of this 1978 mission).

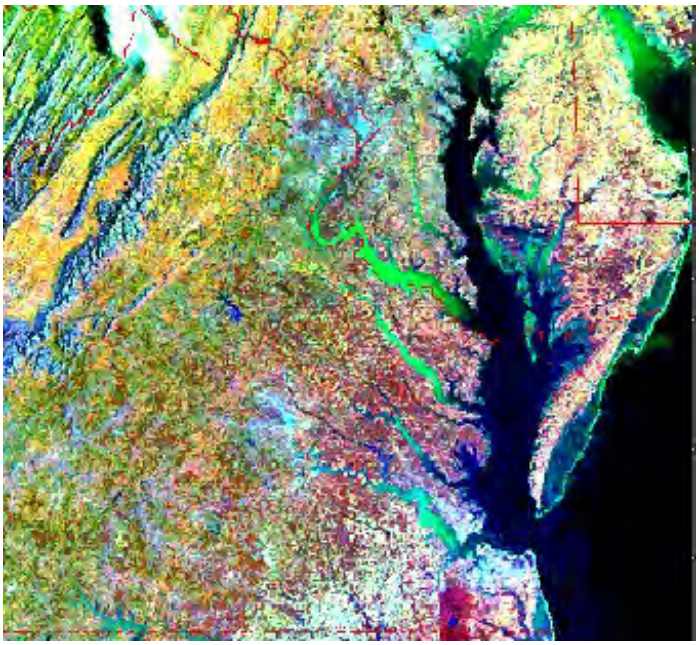


The swath width covered by the image is 715 km (444 miles); spatial resolution is 500 m (1,640 ft). To get your bearings at this scale, locate Washington, D.C. and Pittsburgh, PA in an Atlas. This overview scene is shows the entire central part of the Appalachian Mountain system in a single view. The fold belt of the Valley and Ridge Province is distinctive. At the northern end, in Pennsylvania, the zig-zag pattern described above is determined by the closed or pitching anticlines and synclines. This type of folding largely disappears to the south, being replaced by subparallel ridges separated by valleys. Note the sharp bend or kink in the belt near Roanoke, VA. This feature is, in part, fault-controlled. We can see the Blue Ridge because it is a thin dark line (plus South Mountain) that lies to the east of the Shenandoah Valley in Virginia which continues as the Great Valley through Pennsylvania (both in light tones).

The Piedmont to the southeast is not particularly obvious in this image. Its terrain does not express well because of generally low relief. Its boundary with the Coastal Plains is the Fall Line, so-called because waterfalls form along some streams as they descend from harder Piedmont rocks to softer rocks toward the coast. Its boundary is also indistinct (diffuse), but generally the Piedmont terrain has a darker tone is darker than that of the Coastal Plains (more farming and less trees). The Cumberland Plateau, west of the belt, is much darker because of heavy forestation. As that province grades into the Allegheny Plateau, the dissected morphology remains, but forest cover diminishes.

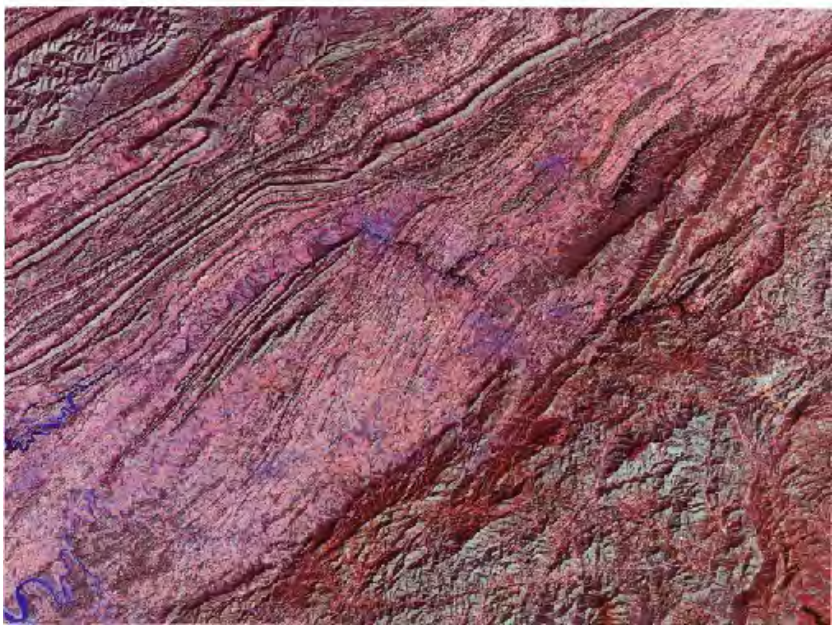
6-7: Where is the Delmarva peninsula; locate the Shenandoah Valley of Civil War battle fame. Where is Pittsburgh. What state occupies the upper left 1/6th of the image? Name all the states that are entirely or partly present in the image. ANSWER

The Piedmont and Coastal Plains extend from New Jersey on the North to Alabama on the South. This next image (made from MODIS data?) of eastern Virginia and Maryland shows how differences in both soil and land use can be discerned by color differences. Inspect this:



The dark blue in the upper left establishes the Blue Ridge. To its east, in orange tones (such tones also mark the Shenandoah Valley), are the Catoctin Mountains and other hills that are the higher part of the Piedmont. The Piedmont to the east is roughly the area that contains considerable darker olive colors. The Coastal Plains further east and south have wine-purples and reds as prevailing colors. Note the area around Washington, D.C. (itself in blue), where the Fall Line runs right through the city.

Lets now venture south along the Appalachians. A section of the Great Smoky Mountains, north of Ashville, occupies the lower right corner of the image. This is an assemblage of relatively homogeneous crystalline rocks that includes the Blue Ridge on their eastern side. Running diagonally through the image center is part of the Tennessee Valley, which is made up of rolling hills underlain by more easily eroded limestones that tie into the Shenandoah Valley to the north. The Valley and Ridge is represented here by continuous folds of steeply dipping rocks that are also involved in the Pine Mountain thrust. The upper left is the beginning of the Cumberland Plateau, a southern extension of the Allegheny Plateau of Pennsylvania. Note the similarities and differences of this subscene and the earlier one. Here the South Mountain counterparts are much wider.



Chattanooga, Tennessee lies within the Appalachian Fold Belt. It was the site of a major battle (Chickamauga) in the Civil War. Here is a Landsat subscene:

What is normally thought of as the southern tip of the Appalachians occurs where the mountains seem to end abruptly in northern Alabama. The city of Birmingham lies right at the tip, located in part in inland Coastal Plains. Look carefully first at this Landsat mosaic of all of Alabama. Birmingham is the red blotch in the upper center.

If you didn't find it, use this enlargement from a different Landsat scene to set the context.

Here is a more detailed look at the city.

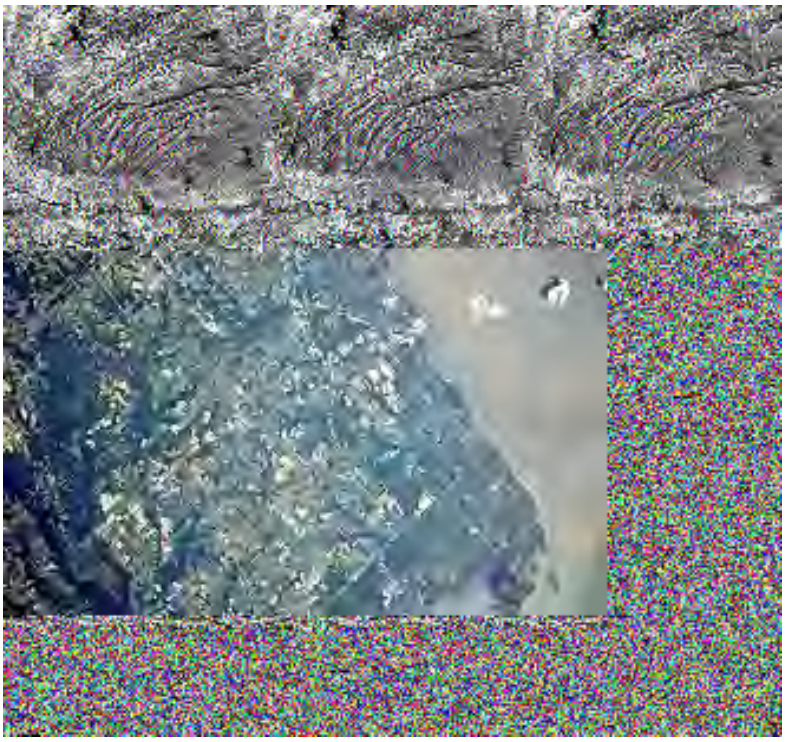
Birmingham has an impressive skyline:

North of Birmingham is the city of Huntsville, AL (near the center). There, the Marshall Space Flight Center, where the Saturn-V rocket was designed, is located. Here is a Landsat-1 view - again, parts of the Appalachian Ridge system and the Appalachian Plateau are evident; the Tennesee River and Bull Shoals also appear.:

Although it lies on the coastal plains of Alabama, this seems a proper place to show an image of Mobile. This city of a half million is situated around Mobile Bay where, in August of 1864, Admiral David Farragut and a Union flotilla defeated the Confederate ships guarding the Bay, thus sealing off the last major seaport through which supplies to the southern states could be landed from overseas sources:

For the writer (NMS), this next image remains an especially fond memory. On the first day that imagery from ERTS-1 was expected to become available to NASA personnel, about 100 program management people and Goddard scientists were gathered at the ERTS Data Receiving Center (in Building 22 at Goddard) to see the results. The initial RBV image was a disappointing polaroid of the Dallas-Fort Worth area (see page 4-5) that showed almost no details. Then, an hour later the first cannister containing a continuous strip of 70mm black and white transparencies representing the first MSS images was brought into the room. Because none of the guests knew how to use the image projection unit, I was delegated to load it for viewing. As I scrolled downward from imagery taken over Minnesota down to southern Missouri,

all views were completely cloud-covered. Disappointing! Then, hallelujah, the next image - the Ouachita Mountains in Arkansas - was completely free of clouds. It was stunning! Much beyond my expectations. Thus did I convert to ERTS/Landsat and to remote sensing in general.



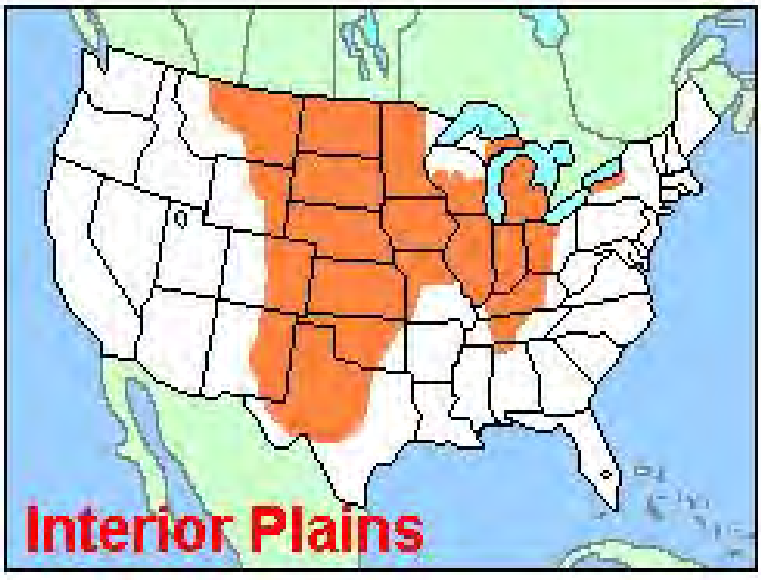
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

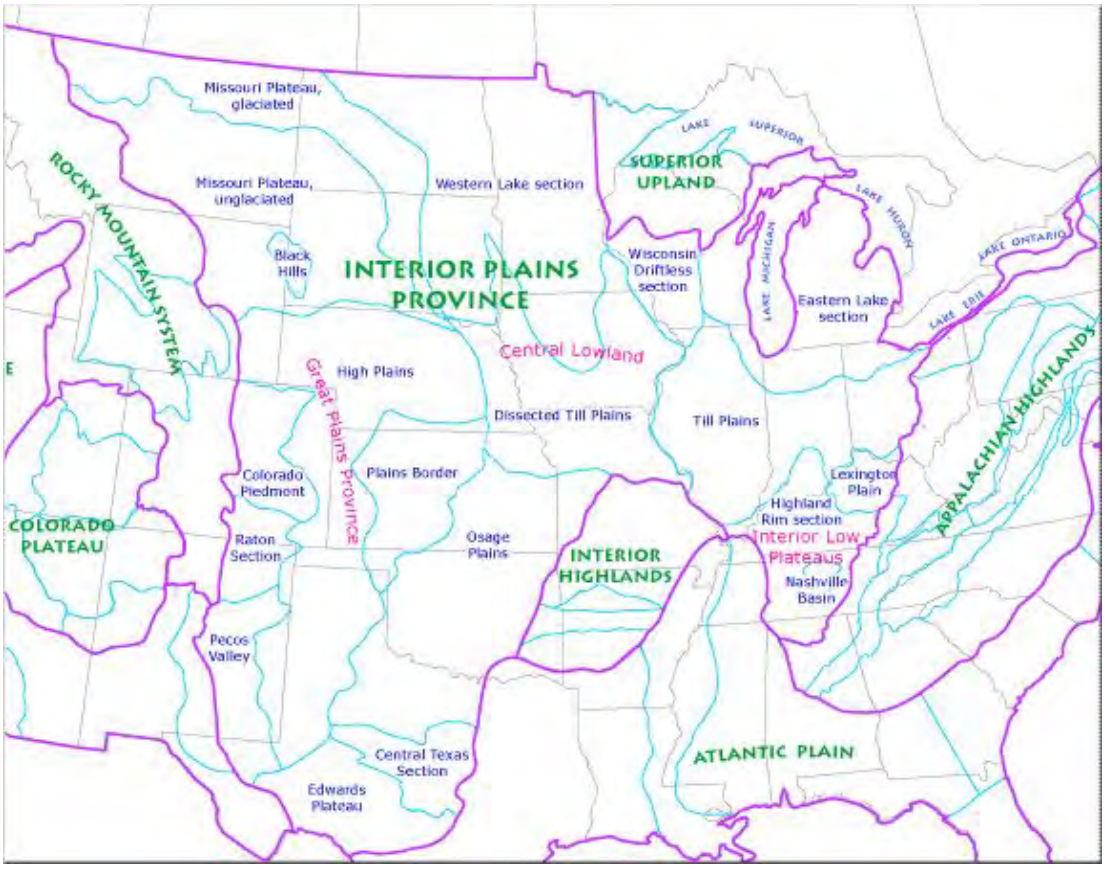
The vast central area of the U.S., into Canada, is a landscape of low, flat to rolling terrain in the Interior Plains. Most ot its eastern 2/3rds forms the Interior Lowlands, discussed on this page. The Lowlands gradually rises westward, from a line passing through eastern Kansas, up to 5000+ feet in the unit known as the Great Plains (next page). Much of the Plains are now converted land use-wise to farming.

The Midwest - The Interior Lowlands:

Proceeding westward, the Appalachian Plateau topography gradually gives way to gentle rolling hills and then (in central Ohio) to flat lands converted principally to farms and urban areas. This is the beginning of the vast Interior Plains of North America, shown first as a single-colored geographic pattern and then in a detailed map which names its subdivisions:



Geographers subdivide the Interior Plains into the Interior Lowlands and the Great Plains on the basis of elevation. The Lowlands are mostly below 1500 feet above sealevel whereas the Great Plains to the west are higher, rising in Colorado to around 5000 feet. The Lowlands, then, are confined to parts of Minnesota, Iowa, Michigan, Ohio, Indiana, Tennessee and Kentucky. Missouri and Arkansas have regions of Lowlands elevations but in the Ozarks (within the Interior Highlands) are higher. Those familiar with the topography of eastern Ohio may be confused by this; that region is hilly but its rocks are horizontal and are an extension of the Appalachian Plateau.



The Interior Plains are largely coincident with the vast Mississippi River Drainage System (other major components are the Missouri and Ohio Rivers), as seen in this map. These rivers have for tens of millions of years been eroding downward into the mostly horizontal sedimentary rocks of Paleozoic, Mesozoic, and Cenozoic ages. The modern Mississippi River system has developed during the Pleistocen Epoch of the Cenozoic Era.

The largest geographic feature in the central U.S. is the group of five Great Lakes, several being in the Lowlands, seen here in this meteorological satellite image:

From left to right (west to east), the lakes are Lake Superior, Lake Michigan, Lake Huron (Georgian Bay is its eastern section), to its south, Lake Erie, then Lake Ontario. Only Lake Michigan lacks a border with Canada. The Great Lakes are the largest complex of freshwater lakes that were scoured out by glacial gouge during the Pleistocene continental ice advance over North America.

The first image depicting Interior Lowlands landforms is just west of the transition between Appalachian-controlled topography and the Interior or Central Lowlands. The scene is that of Cleveland (blackish area on the shore), Lake Erie with partially melted ice in March of 1973, the widespread farmlands (note their distinctive criss-cross pattern) of northern Ohio and southern Ontario, and low hills and lakes (mostly from the last glacial retreat).

Cleveland, Ohio lies astride the southeastern shores of Lake Erie. Here is the central city imaged at 4 meters by IKONOS on April 26, 2000:

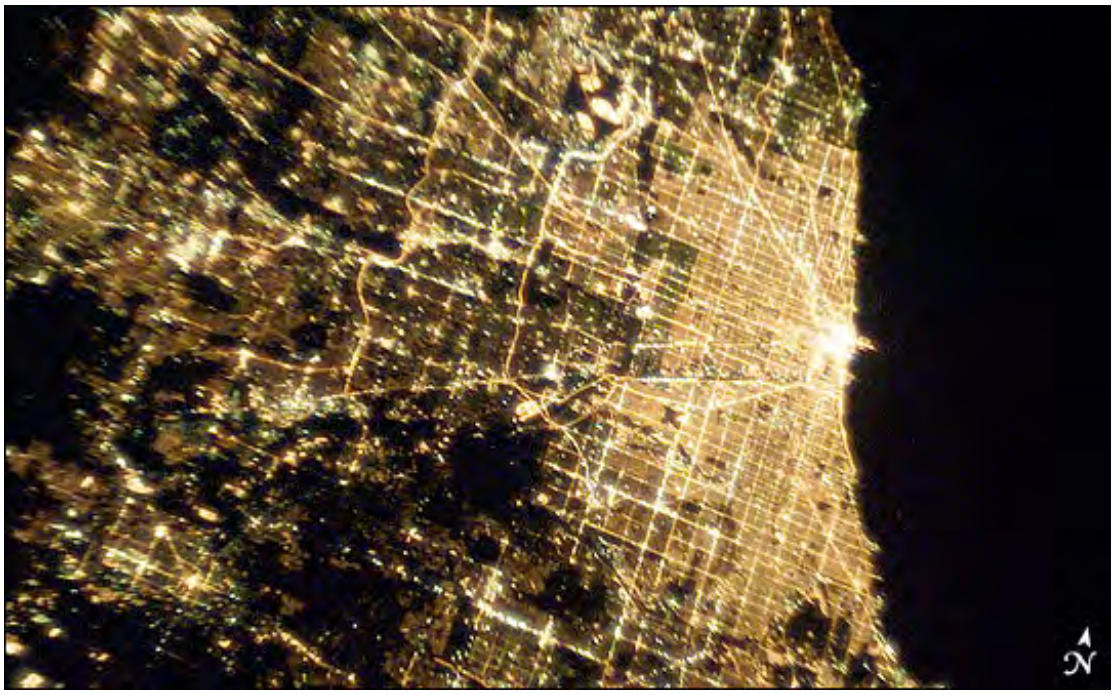
Typical of the Interior Lowlands is the region around Chicago, Illinois, shown in this Landsat MSS scene.

The underlying rock type in the Illinois part of the Interior Province is limestone, but nearly all of the surface here is controlled by glacial deposits. The rich soils from these materials promote farming, of which corn, soybeans, and oats make up the major crops. Most of fields in this scene are rectangular, and many now are fallow (tan-colored) after harvesting. Trees cluster along the banks of rivers such as the Illinois, Des Plaines, and Kankakee that stand out among the farms.

Chicago sits along the southwest tip of Lake Michigan, one of the five Great Lakes. All of them originated as basins cut by glacial scouring and, in places, from damming by moraines. Gary, Indiana is the blue area at the southern tip of the Lake.

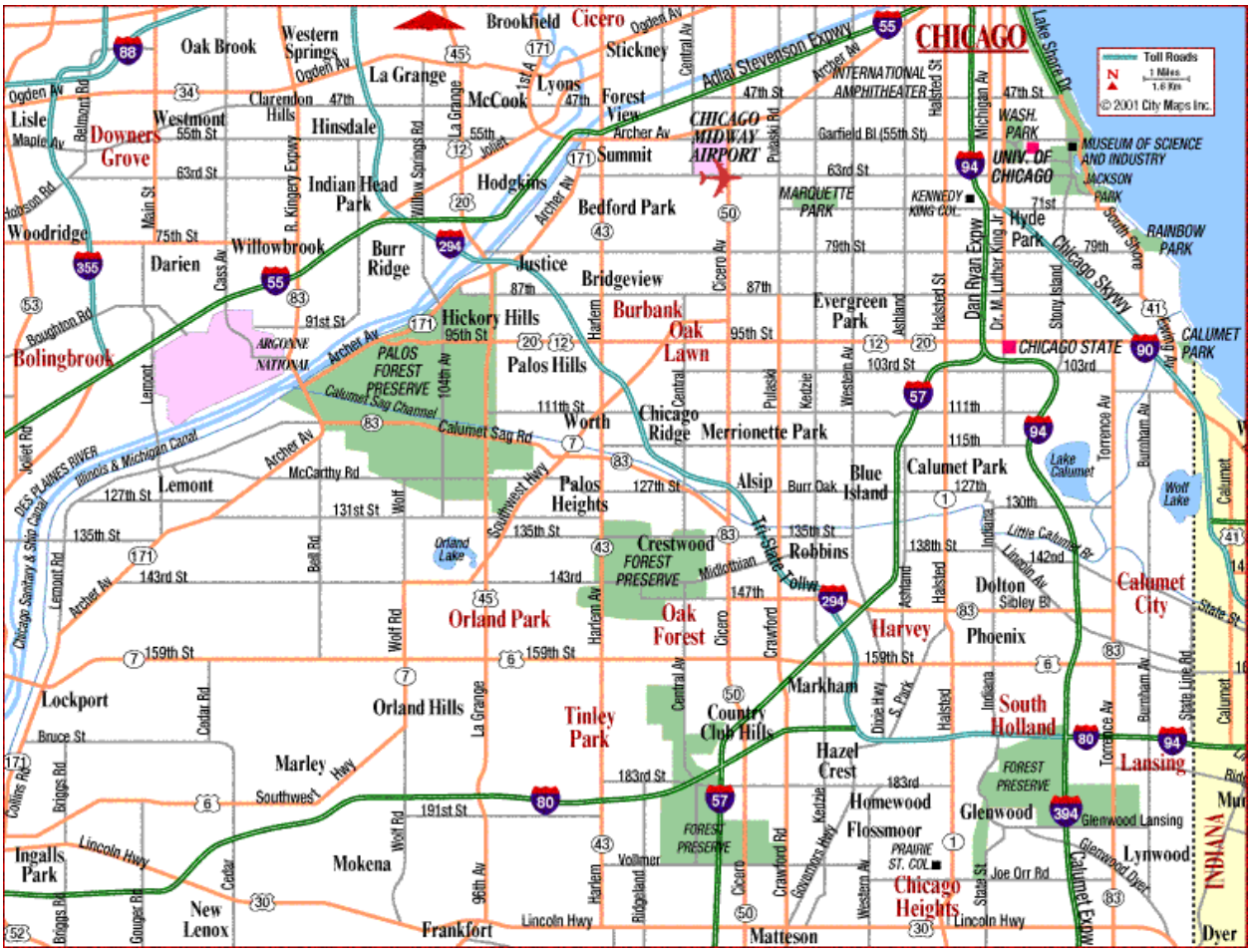
6-8: What time of the year was the Landsat image probably made? Near Gary, IN, note the wispy clouds extending north-northeast into the Lake; what might they be?
ANSWER

Chicago itself appears in these two astronaut photos, one taken in daytime, the other at night:



Chicago is the second largest city in the United States. It has numerous suburban towns, many being well known and in the news. Two maps are needed to show the full metropolitan area:

Metropolitan Chicago, north
Toll Roads
1 Miles
1.6 Km
© 2006 City Maps Inc.
Palatine
Wheeling
PAL-WAUKEE AIRPORT
Northbrook
Glencoe
Winnetka
Kenilworth
Northfield
GLENVIEW NAVAL
Glenview
Wilmette
Rolling Meadows
Arlington Heights
Mount Prospect
Des Plaines
Park Ridge
Niles
Skokie
Evanston
NORTHWESTERN UNIVERSITY
LAKE MICHIGAN
BUSSE FOREST PRESERVE
Schaumburg
Elk Grove Village
CHICAGO O'HARE INTERNATIONAL AIRPORT
Roselle
Itasca
Wood Dale
Rosemont
Schiller Park
Harwood Heights
Lincolnwood
LOYOLA UNIVERSITY
Bensenville
Addison
Stone Park
Elmwood Park
Northlake
River Forest
FRANK LLOYD WRIGHT HOME
Oak Park
DE PAUL UNIV.
CHICAGO
Villa Park
Elmhurst
Glen Ellyn
Lombard
Maywood
Forest Park
Oakbrook Terrace
UNIV. OF ILLINOIS-CHICAGO
SOLDIER FIELD
Chicago Harbor
NAVY PIER
DU PAGE COLLEGE
MORTON ARBORETUM
Oak Brook
Westchester
CHICAGO ZOOLOGICAL PARK
La Grange Park
Riverside
Brookfield
Cicero
COMISKEY PARK
Western Springs
Lyons
Stickney
Forest View
ILLINOIS INST.OF TECH
WRIGLEY FIELD
HORNER PARK
WRIGHT COLLEGE
LINCOLN PARK
LINCOLN ZOO
OAKTON PARK
W. R. HARPER COL.
Tri-State Tollway
Northwest Tollway
West Tollway
Eisenhower Expw
Dan Ryan Expy
Chicago Sanitary & Ship Canal
Adlai Stevenson Expressway
Edens Expw
Lake Shore Dr
Kennedy Expy
Elgin-O'hare Exp
Lake St
Army Trail Rd
Irving Park Rd
Golf Rd
Dundee Rd
Central
Roosevelt
22nd St
31st St
47th St
Cermak Rd
Belmont Av
Grand Av
North Av
Division
Chicago
Washington
Foster
Peterson
Devon
Touhy Av
Oakton St
Euclid Av
Willow Rd
Central St
Emerson
Dempster
Montrose
Irving Park
Fullerton
Lawrence
Montrose Dr
Wise Rd
Thorndale
Higgins
Algonquin
Kirchoff Rd
Northwest Hwy
Milwaukee
Caldwell A
Cumberland
Greenwood
Sunset Ridge
Waukegan
Sheridan Rd
Green Bay
Tower
E Lake Av
W Lake Av
Central Av
Simpson
Morton Grove
Gross Point
Crawford
McCormick
N. SHORE CHANNEL
Ashbury
Sheridan
Halsted
Michigan Av
Western
Ogden
Ashland
Kedzie
Pulaski
Cicero
Central
Narragansett
Harlem
Ogden Av
Cermak
Pershing
Archer Av
AMPHITHEATER
INT'L
Halsted
Ryan Ex
Addison Rd
Glen Ellyn Rd
Main St
Rohlwing Rd
Roselle Rd
Meacham
Grove
Hicks
Rand
Arlington Heights
Elmhurst
Busse
Wolf Rd
Des Plaines River Rd
Rand Rd
Milwaukee Av
East River
Mannheim
25th
1st Av
York Rd
Robert Kingery Expw
St. Charles Rd
Butterfield Rd
Geneva Rd
Lake
Higgins Rd
Palatine Rd
Meacham Rd
Highland
Midwest
Ogden Av
47th St
La Grange
Harlem
Cermak Rd
Ogden Av
Pulaski
Kedzie
Central Av
Archer Av
47th St
Stickney
Forest View
Lyons
Brookfield
Riverside
Cicero
Chicago Sanitary & Ship Canal
Adlai Stevenson Expressway
Dan Ryan Expy
Kennedy Expy
Lake Shore Dr
Michigan Av
Navy Pier
Chicago Harbor
Soldier Field
Univ. of Illinois-Chicago
Comiskey Park
Illinois Inst. of Tech
Int'l Amphitheater
Lincoln Park
Lincoln Zoo
Wrigley Field
Horner Park
Wright College
Lake Shore Dr
Sheridan
Halsted
Lincoln Park
Lincoln Zoo
DE PAUL UNIV.
Wrigley Field
Horner Park
Wright College
Lincolnwood
Devon
Peterson
Foster
Montrose
Irving Park
Fullerton
Grand
Division
Chicago
Washington
Eisenhower Expw
Roosevelt
Cermak
Ogden Av
Pulaski
Kedzie
Central
Narragansett
Harlem
Cermak Rd
Ogden Av
Pulaski
Kedzie
Central Av
Archer Av
47th St
Stickney
Forest View
Lyons
Brookfield
Riverside
Cicero
Chicago Sanitary & Ship Canal
Adlai Stevenson Expressway
Dan Ryan Expy
Kennedy Expy
Lake Shore Dr
Michigan Av
Navy Pier
Chicago Harbor
Soldier Field
Univ. of Illinois-Chicago
Comiskey Park
Illinois Inst. of Tech
Int'l Amphitheater
Lincoln Park
Lincoln Zoo
Wrigley Field
Horner Park
Wright College
Lake Shore Dr
Sheridan
Halsted
Lincoln Park
Lincoln Zoo
DE PAUL UNIV.
Wrigley Field
Horner Park
Wright College
Lincolnwood
Devon
Peterson
Foster
Montrose
Irving Park
Fullerton
Grand
Division
Chicago
Washington
Eisenhower Expw
Roosevelt
Cermak
Ogden Av
Pulaski
Kedzie
Central
Narragansett
Harlem
Cermak Rd
Ogden Av
Pulaski
Kedzie
Central Av
Archer Av
47th St
Stickney
Forest View
Lyons
Brookfield
Riverside
Cicero
Chicago Sanitary & Ship Canal
Adlai Stevenson Expressway
Dan Ryan Expy
Kennedy Expy
Lake Shore Dr
Michigan Av
Navy Pier
Chicago Harbor
Soldier Field
Univ. of Illinois-Chicago
Comiskey Park
Illinois Inst. of Tech
Int'l Amphitheater
Lincoln Park
Lincoln Zoo
Wrigley Field
Horner Park
Wright College



The Chicago Loop (named for the rectangular track layout of elevated trains) reaches to Lake Michigan in the dense central downtown which is visible as blue tones, while much of the surrounding metropolitan areas are expressed by the reds denoting trees in suburban settings. Look in the image for features near the waterfront and compare with those in this aerial oblique photo of that area:

Another major Great Lakes city lies almost due east of Chicago. Detroit, Michigan is built along

the Detroit River which connects Lake Ontario at its northwest end with Lake St. Clair above a peninsula in Canada on which Windsor is situated across river from Detroit. Both cities are shown in this aerial view:

This Landsat-7 ETM+ subscene shows part of this metropolitan complex of nearly 8 million people in natural color.

At the other extreme in population (~16000) is Sault Ste. Marie, at the east end of Michigan's Upper Peninsula. It is one of the oldest towns in the Upper Midwest, tracing its founding to the late 1600's. The term "Sault" is French for cataracts (waterfalls along the St. Mary River):

Wisconsin has several larger cities that one thinks of offhand: Madison, Green Bay, and Milwaukee. Madison is the state capital and the site of the main campus of the University of Wisconsin:

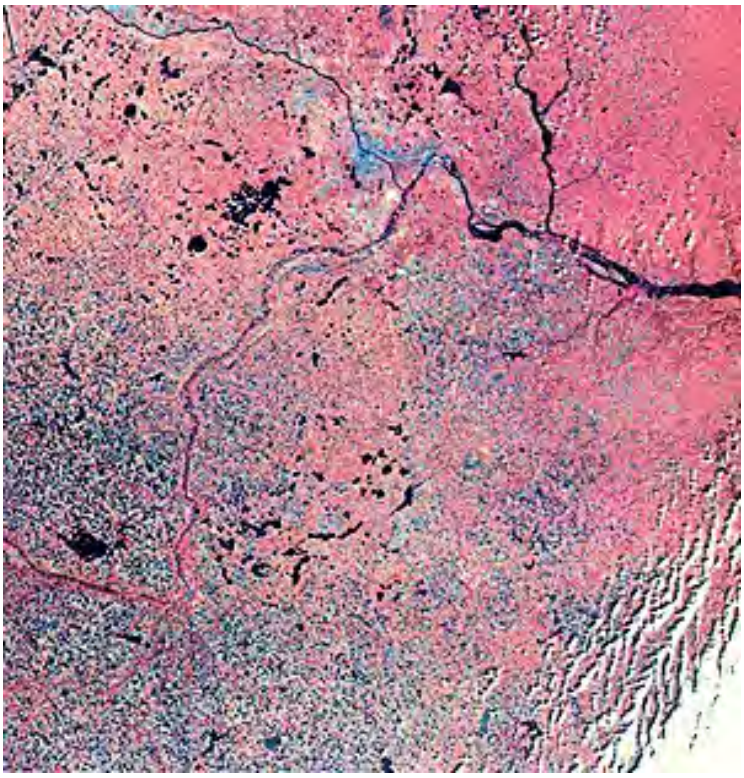
As an emblem of Wisconsin, we have selected a very remarkable small city of about 100000: Green Bay, which is the smallest city to support an NFL team, the Packers - perennial powerhouses that play their football in an often frigid open stadium (Lambeau Field) filled with dedicated and loud fans.

Here is Milwaukee, astride Lake Michigan, from space:

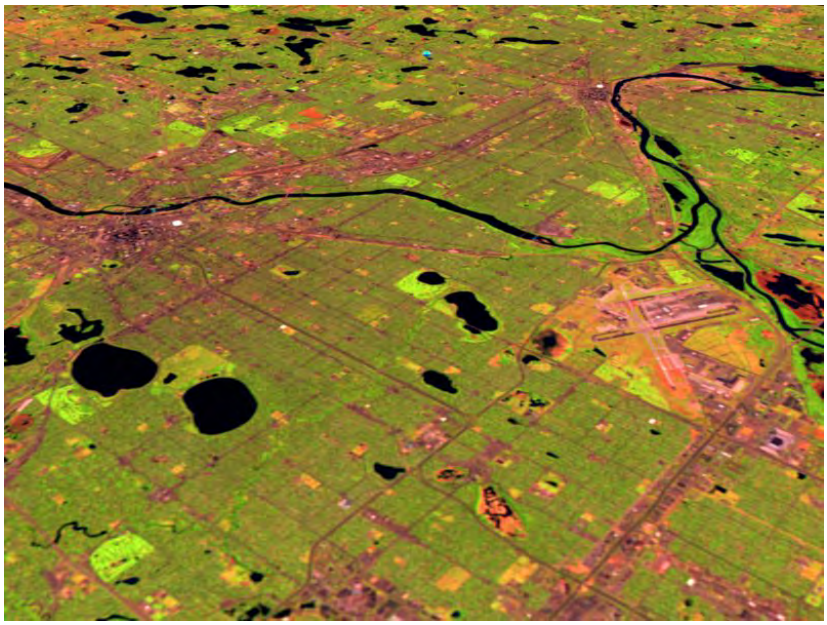
Both Wisconsin and Minnesota have a very large number of lakes developed both by glacial scouring and by water in depressions within the till cover. In Wisconsin, one large lake of reknown is Winnebago:

Part of southeastern Wisconsin never experienced thick ice cover during the Pleistocene, so that after melting no tills or drift were deposited. This is known as the "Driftless Area" near Lacrosse. As photographed from the Space Shuttle, the more irregular dissected topography of this terrain is set apart from its surroundings.

Another area within the glaciated part of the Interior Lowlands is southern Minnesota. In the next scene, farmland with many glacial lakes is predominant. But, Minneapolis and St. Paul just to its east appear (blue patches near top center) along a part of the upper reaches of the Mississippi River (whose headwaters begin to the northwest) that make that mighty stream much narrower in this part of its course. The Minnesota River joins the Mississippi from the southwest; the St. Croix River to the east forms the border there with Wisconsin which then continues southward along the Mississippi. Two Landsat images:



Here are two more Landsat images. The first shows Minneapolis-St.Paul in late winter when almost no vegetation is blooming. The second is an odd false color rendition in which the image is tilted to give a perspective view:

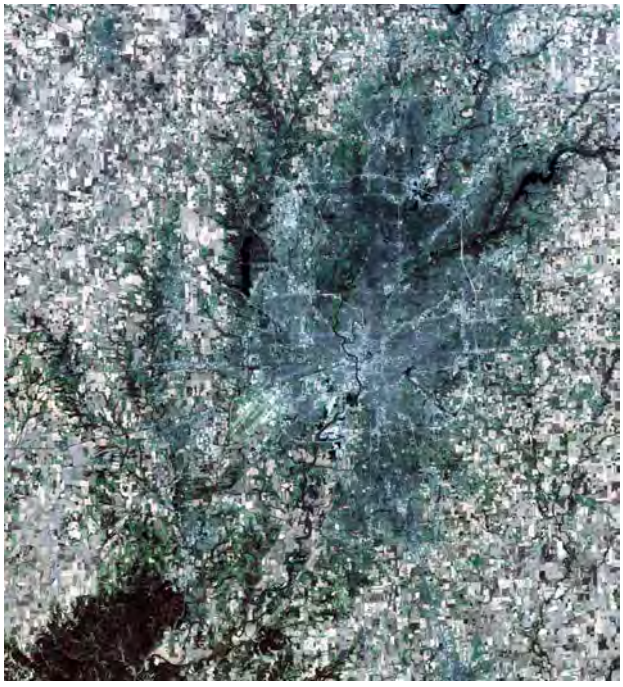


A map of the Twin Cities:

This is Minneapolis and then St.Paul as seen from the air:

St. Paul

Now, let's backtrack a bit and go south and east. In the center of Indiana is its capital, Indianapolis seen here from space and from a ground perspective:

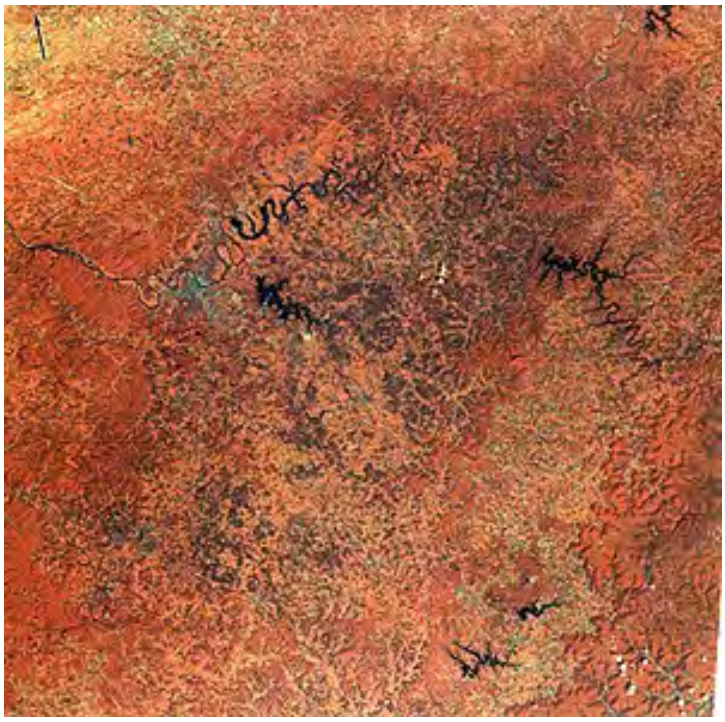


While many sports fans may think first of the Indianapolis Colts, once a year, on Memorial Day in May, Indianapolis is the sports capital of the USA. The famed Indianapolis 500 - for racing cars - is run on that day. Here is the track seen from space.

Now, let's move still further south, first along the Ohio River, which in some places marks the southern boundary of Pleistocene glaciation. First, look at these scenes of Cincinnati, OH., with the conspicuous stadium for the Cincinnati Bengals and the Cincinnati Reds:

Further down the Ohio River is the city of Louisville, Kentucky. Once a year it is the center of attention during the running of the Kentucky Derby (a horse race). The track is near the lower right of this SPOT Image; with an enlarged aerial view beneath:

In Tennessee, the Lowlands continue, as evident in this full Landsat scene. At its southeast (lower right) is the edge of the Cumberland Plateau (assigned to the Appalachian province and equivalent to the Allegheny Plateau); its prominent west-facing scarp continues both southward and into Ohio east of Columbus. Nashville is near the edge of the Nashville Dome, a gentle upwarp of sedimentary rocks (mostly limestones); its broad outline is elliptical, and is indicated by forests (evident in the image) that selectively concentrate on soils from certain sedimentary units.

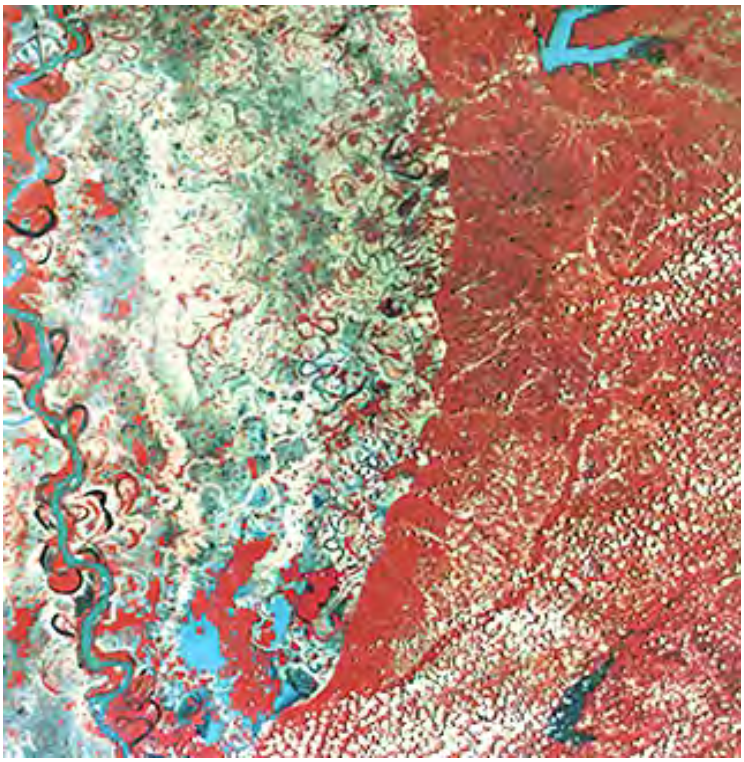


Nashville, Country Music's capital, lies on the Cumberland River:

The Mississippi River services the heartland of the U.S. One of the major urban areas along this river is St. Louis, MO (seen first on page 4-2). Here it is again, extending from St. Charles on the west to the cities of Illinois including East St. Louis and Bellevue and the conspicuous Horseshoe Lake (an oxbow meander).

This Quickbird image shows the downtown area of St. Louis, with its main landmarks being the Arch, and football and baseball stadiums.

Most of the Mississippi River drainage passes through various parts of its vast drainage system within the Interior Plains. But the Lower Mississippi from southernmost Missouri southward is emplaced on a part of the Gulf Coastal Plains known as the Mississippi Embayment (an inflow of marine waters starting about 75 million years ago which deposited thick sediments in a structural trough). In this Landsat-1 view we see the state of Mississippi on the right and Louisiana on the left (along the west bank of the active river). Here, in the Greenville, MS area) the meandering Mississippi River has cut a floodplain about 80 km (50 miles) wide. The forested uplands bedrock to the right consists in part of Eocene sedimentary rocks covered by windblown glacial dust (loess).



Jackson, Mississippi is the capital of the state, seen in this satellite image and from the ground at night:

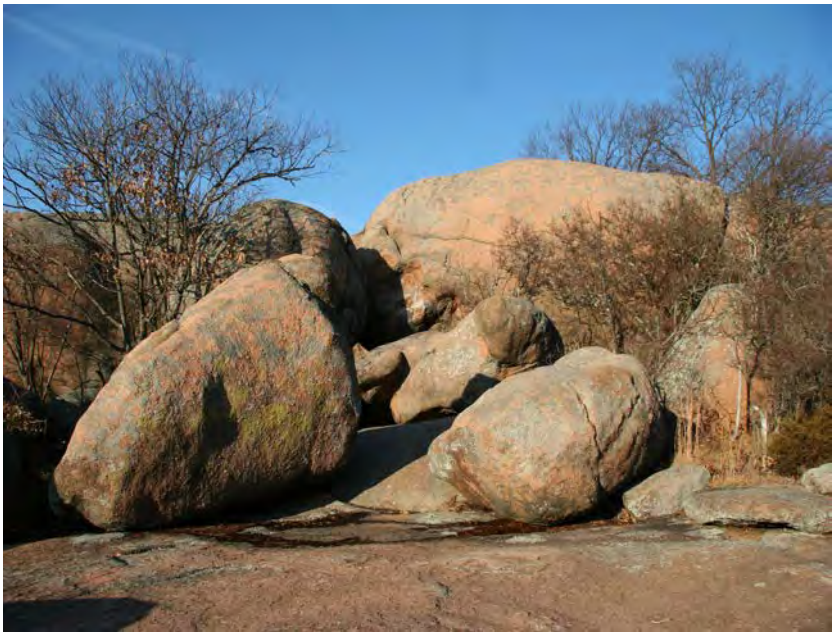
(One wonders what might happen to Jackson if the prominent dam on the Yockinookany River were to experience at catastrophic failure?)

Baton Rouge, on the east bank of the Mississippi River, is the capital of Louisiana and the home of the Louisiana State University, whose Tigers are a perennial power in college football:

The states immediately west of the Mississippi River are considered part of the Interior Lowlands. But, not all of the Interior Plains is flat and low. Some high hills/low mountains occur. The broadest of these lies within southern Missouri and northern Arkansas and is known both as the Interior Highlands and Ozark Plateau. The Landsat scene below is mostly in Arkansas. The darker red area near the middle of the image is the Boston "Mountains", which are actually a dissected plateau of horizontal (mostly Pennsylvanian in age) strata, with relief up to 300 m (1000 ft).

North of the Boston Mountains is the Ozark Plateau. This plateau is actually a broad, gentle dome with its Paleozoic sedimentary rocks only slightly inclined. This map shows the subdivisions of the Ozarks:

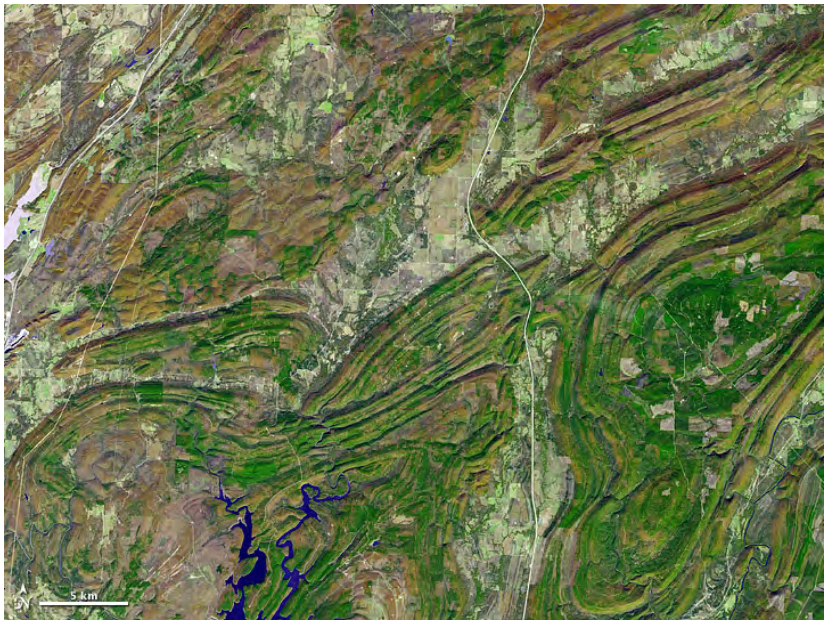
Erosion has produced the Ozark Mountains, seen next. These are technically not true mountains since their rocks are unfolded. At the heart of the Ozarks are the St. Francois Mountains; streams have cut through the sedimentary rocks exposing the granite basement at an outcrop called the Elephant Rocks:

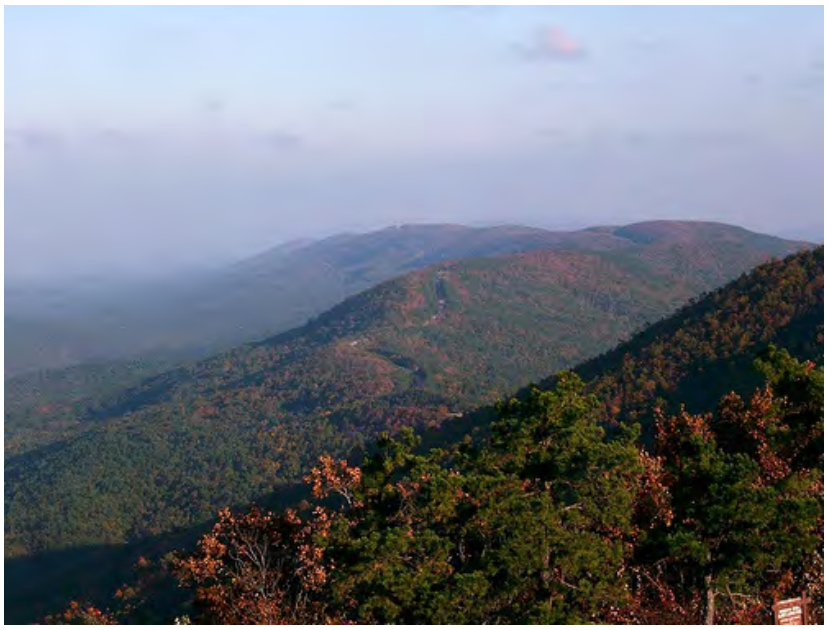


Elephant Rock State Park is about 15 miles from Arcadia, Missouri, outside of which the writer's (NMS) grandparents had a second home. From its north side, I could view Pilot Knob. This conical mountain (elevation 1445 ft) was visible from the Mississippi River 40 miles to the east. The mountain was unique in that it consists of a plug of igneous rocks punched through the surrounding sedimentary rocks. This igneous intrusion contains enriched iron deposits that were mined starting in the 1800s.

Pilot Knob gives its name to a large Civil War battle on September 27, 1864, in which more than 1200 were casualties. The Confederates under Gen. Price won the battle after Union Gen. Ewing ordered Fort Davidson to be blown up.

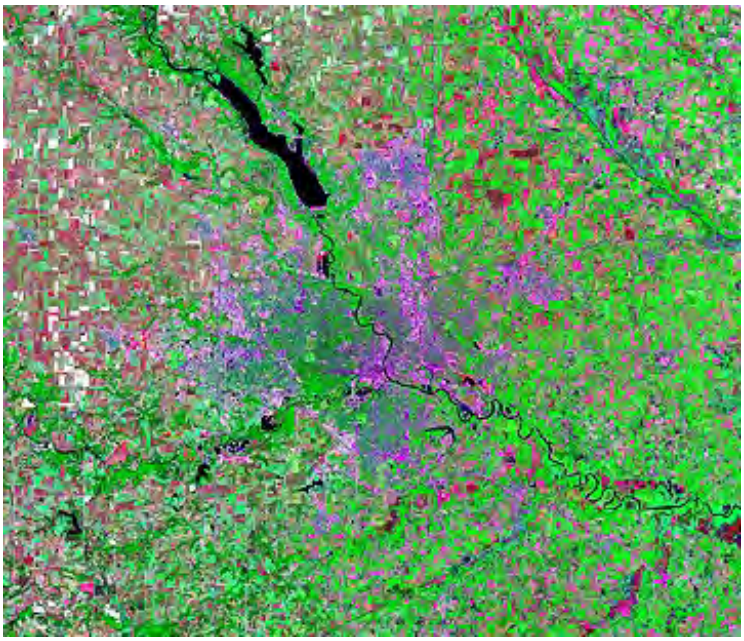
South of the Ozarks, and part of the Interior Highlands, are the Ouachita Mountains in central Arkansas. These mountains are highly deformed into closed folds. They are part of the Appalachian system of late Paleozoic age. Here is an ASTER image of part of the Ouachitas; below that is a ground scene:





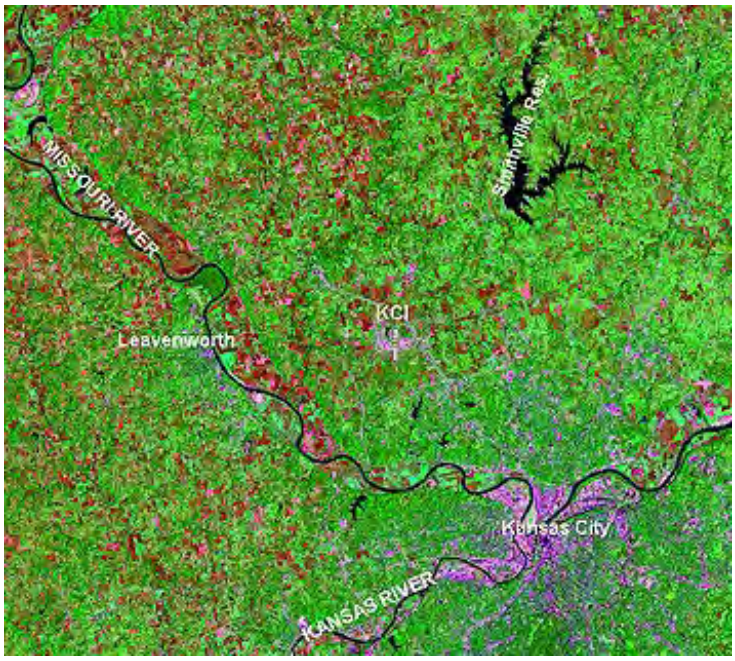
The southern half of Arkansas is part of the Mississippi Embayment, an extension of the Coastal Plains. The state capital, Little Rock, is just south of the Ozarks:

Another state in the westernmost Interior Lowlands is Iowa, sometimes placed specifically in the Central Lowlands subprovince. Most of the state is underlain by Paleozoic rocks (limestones common) covered by Pleistocene till deposits. The capital city is Des Moines, in the center of Iowa:



In the Southeast part of the state is Iowa City, home of the University of Iowa, as shown in this IKONOS image:

During the 19th Century in that period when tens of thousands of Americans followed Horace Greeley's dictum "Go West, young man!", the jumping off places for Prairie Schooners to start their trek across the Great Plains were at the end of eastern rail lines at Kansas City and St. Joseph on the Missouri River at the west end of that state. Here is a space image of the Kansas City area:



Here is a closer look at part of the city with apartment high rises:

This oblique aerial view shows the downtown part of Kansas City:

Primary Author: Nicholas M. Short, Sr.

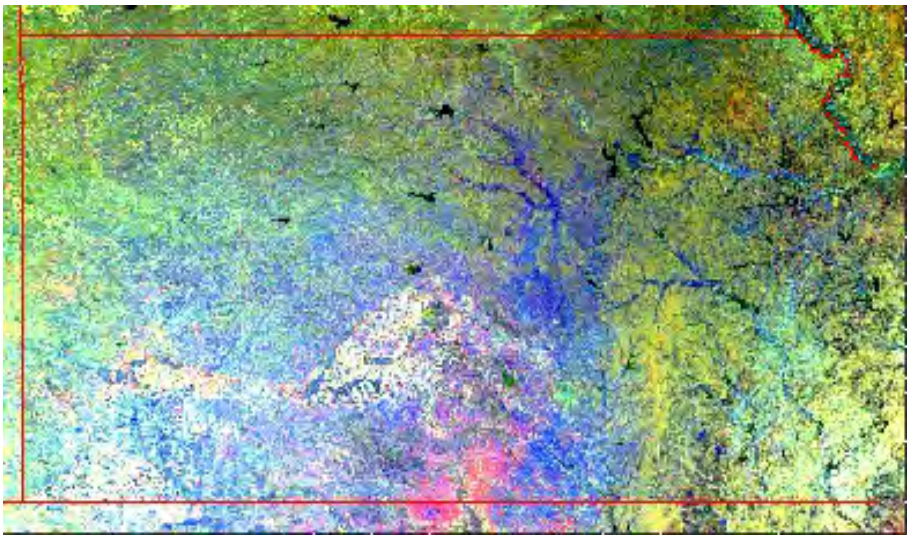
The vast central area of the U.S., into Canada, is a landscape of low, flat to rolling terrain in the Interior Plains. As it rises towards the Rocky Mountains it transitions into the Great Plains. A typical Great Plains view of such country is shown by a space image of Kansas. But, in various areas, especially in western North and South Dakota and eastern Montana, erosion has produce gullies and small canyons. A scene including part of the Black Hills is illustrative of this.

The Interior Lowlands: The Great Plains

Most people believe they enter the Great Plains as they cross into Kansas, Nebraska, and other states along that general longitude. From the geomorphologist's and geographer's viewpoint, the Great Plains is arbitrarily set further west in these states, where lowlands elevations begin to rise to their maximum ascension at the Rocky Mountain Front. This map helps to place the Great Plains physiographic province in context with the rest of the U.S.

Nice topographic map of Great Plains. Source: http://www.lib.ndsu.nodak.edu/govdocs/text/greatplains/cover.jpg; accessed March 3, 2006.

Say "Great Plains" and typically the first thought is agriculture - this region is often called the "Bread Basket of the U.S." because wheat is the dominant crop. While that is valid, the Great Plains can also be thought of as transitional from the vegetation-dominated eastern half of the U.S. to the rock- and soil-dominated western half. Consider this AVHRR scene that shows almost all of Kansas.

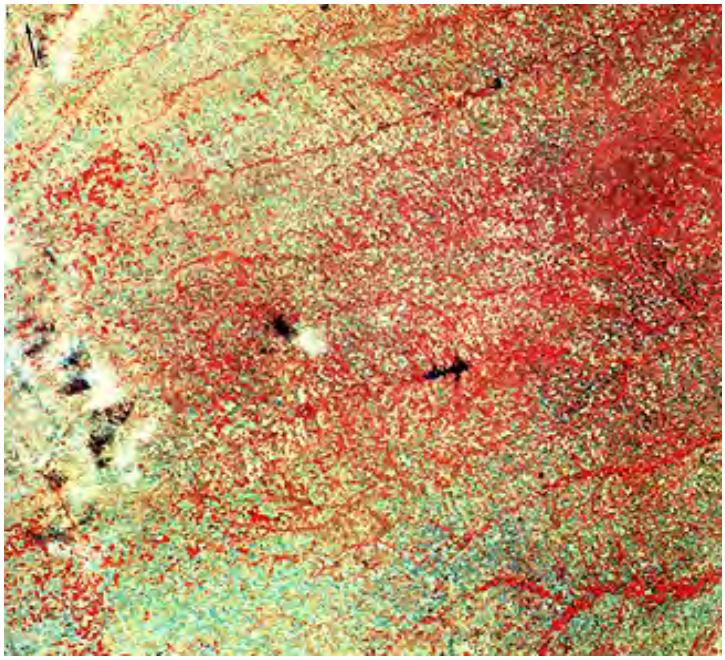


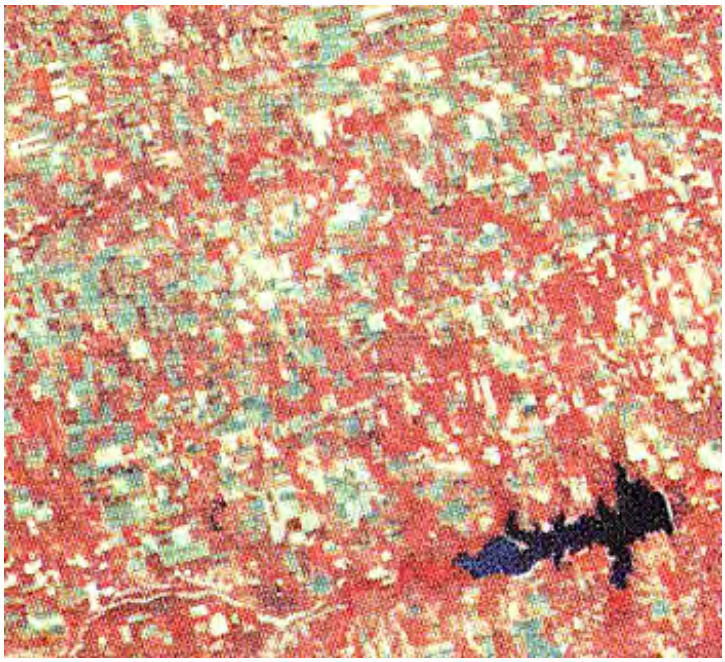
The state's land cover falls roughly into three parts: 1) Eastern (olive): a mix of forests and fields; 2) Central (blue): mostly farmlands; and 3) Western (green): a mix of farms and grasslands, vegetation becoming semi-arid, as elevations rise.

Here is an ASTER image that shows an area of the Great Plains almost totally occupied by farmlands:

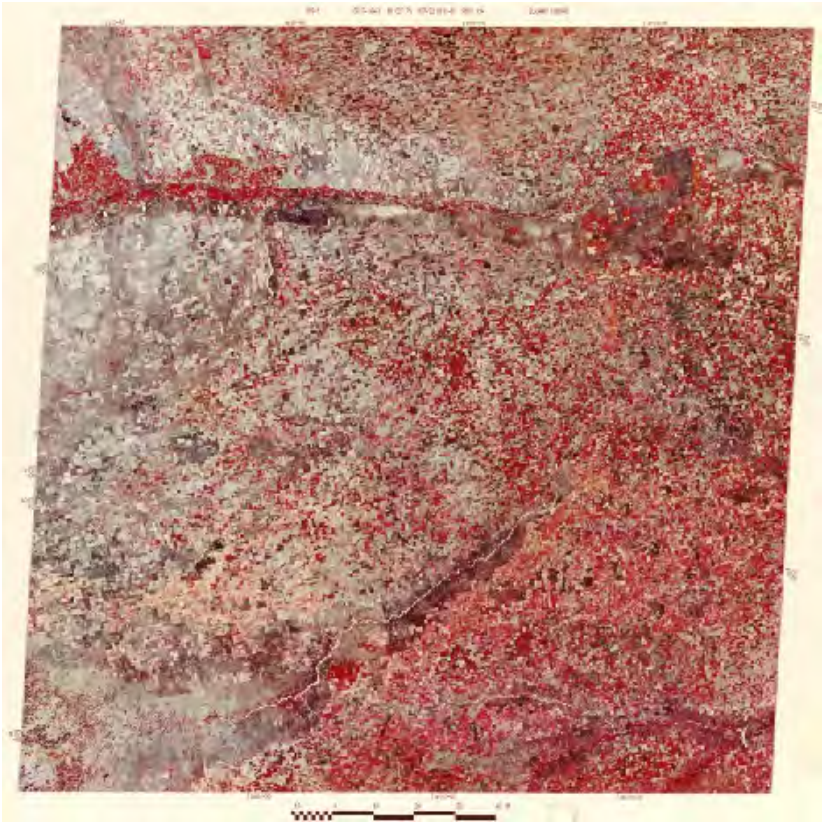
Another mental picture of the Great Plains is that of natural (wild) grasslands which are almost synonimous with the term "prairie", such as depicted here:

So, as our journey continues to the west, the terrain slowly rises in elevation and local hilly surfaces, often with low scarps, begin to appear. This Landsat-1 image is typical of the farmlands in central Kansas. This August. 1973 scene shows fields that contained now harvested winter wheat in blue and spring wheat and other crops such as alfalfa in red. The area contains mostly Tertiary sedimentary rocks. Much of the water is obtained from the vast underground aquifer known as the Ogallala formation (it is being rapidly depleted by overwithdrawal).



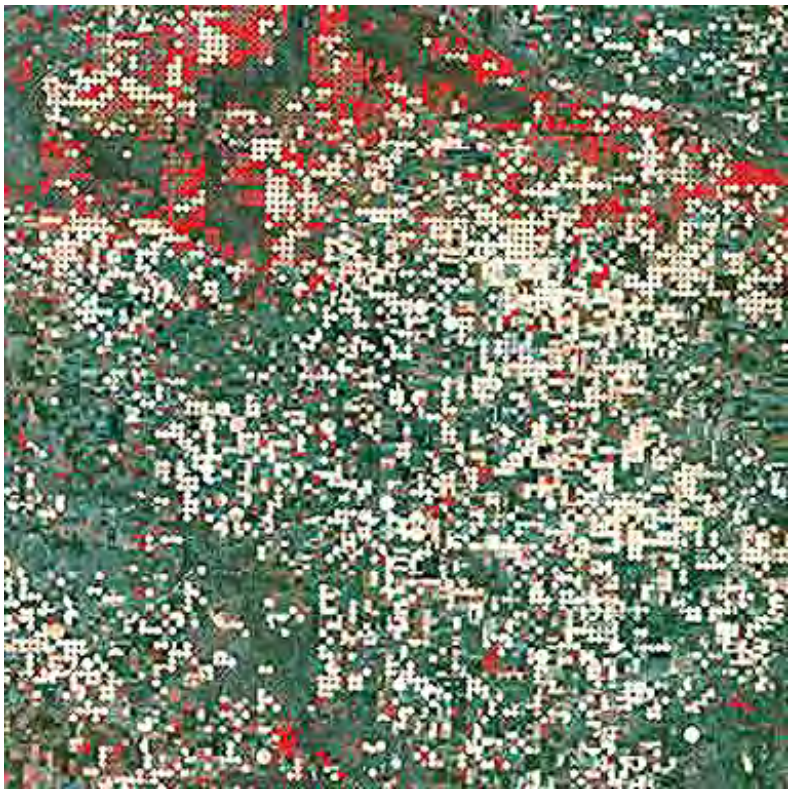


In this next scene, acquired in October, 1972, we are now over the Great Plains of southwestern Kansas. During fall and winter, the ground has a grayish-brown look that intensifies even further to the west in Colorado. This is also evident in parts of this scene, wherever farm crops have not blotted out their underlying soils (i.e., fields currently fallow), or the natural surfaces are not converted to agriculture. An example occurs in the lower left corner where the Cimarron River has developed gullied badlands (dendritic drainage) in soft sediments. Note that many farms are square and are often just one mile on a side. These squares correspond to the section divisions in the Township-Range system of land mapping that was adopted in the 19th century in the United States. Near the upper right corner are clusters of circular pivot-irrigation fields, similar to those we showed on page 3-3. These fields lie along the Arkansas River just to the west of the largest town in this part of the state, Garden City.



A Landsat TM image within which is Garden City, Kansas is another typical view of Great Plains farmlands. The time is early Spring when some crops (in red; probably spring wheat) are ready for harvest, other farms remain fallow before planting. The green areas (not a natural color in this

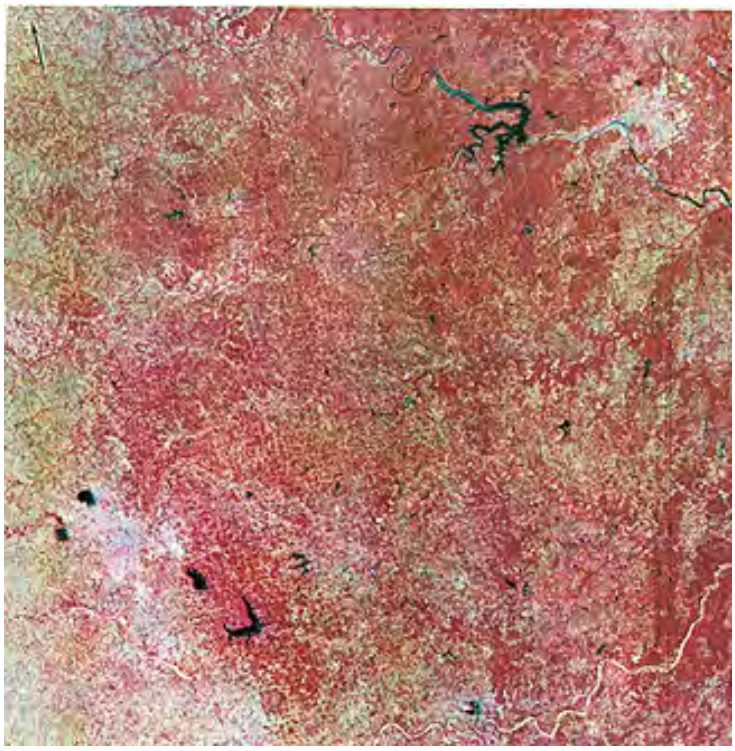
false color rendition) are mainly grasslands.



As a generalization, the Great Plains scenes appear similar to some western Interior Lowlands scenes, both of which, are dominated by farmland. On the ground, the Kansas landscape has a more western look because of the semi-desert vegetation, including grasses and sage-like shrubs. There is more red in the eastern (right) half of the image than to the west, because of differences in crop type and stage. That is, wheat is more common in the western part of the image, and according to the harvesting schedule, the higher and somewhat cooler western lands had been culled earlier in the Fall.

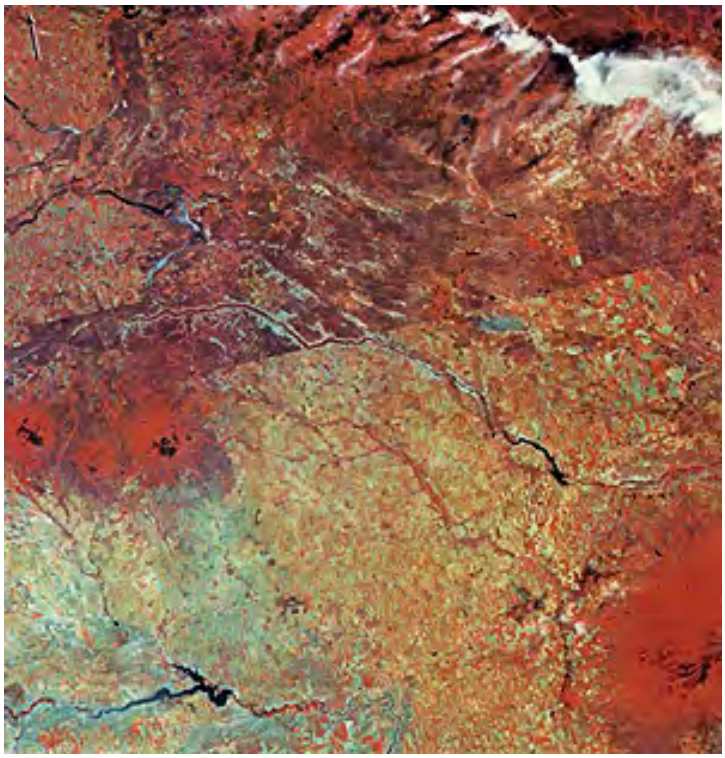
There are few large cities in the Great Plains. Exceptions are Dallas-Fort Worth, which we examined in Section 4, and Kansas City, MO and Omaha, Nebraska which lie near the boundary between the eastern Great Plains and the Interior Lowlands. Here is a color photograph of Omaha, with Council Bluffs, Iowa across the Missouri River, taken by astronauts onboard the International Space Station:

The two largest cities in Oklahoma appear in this next Landsat-1 image. Tulsa is the blue patch in the upper right; Oklahoma City is much harder to see in the lower left. Both oil and agriculture are important in this region, where trees are still fairly common, although giving way on the west side to more western Great Plains vegetation (brush; grasses).



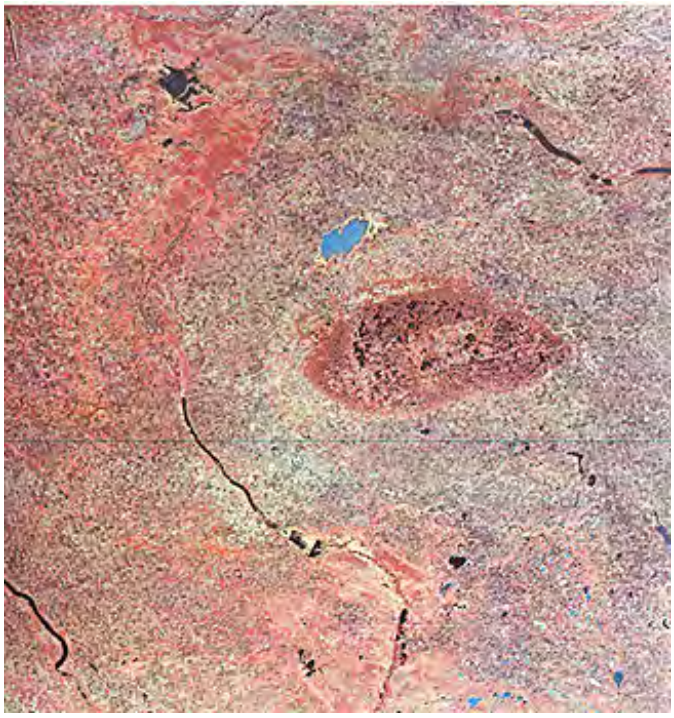
Here is a satellite image that shows Tulsa, Oklahoma:

Where the western Great Plains extends into Canada, a visually striking difference marks the exact border between that country and the U.S., clearly evident in the scene below. The lower half lies within the farmlands (mostly wheat) of eastern Montana (note the elongate shapes of many farms). But across the border in Alberta, Canadian settlers chose to retain the natural vegetation (grasslands) and devote this land use to grazing of cattle (some farms are seen in the upper right). On the U.S. side, the plains surround two older outliers of igneous rocks, the Bearpaw Mountains (lower right) and the Sweet Grass Hills (center left), both forested.



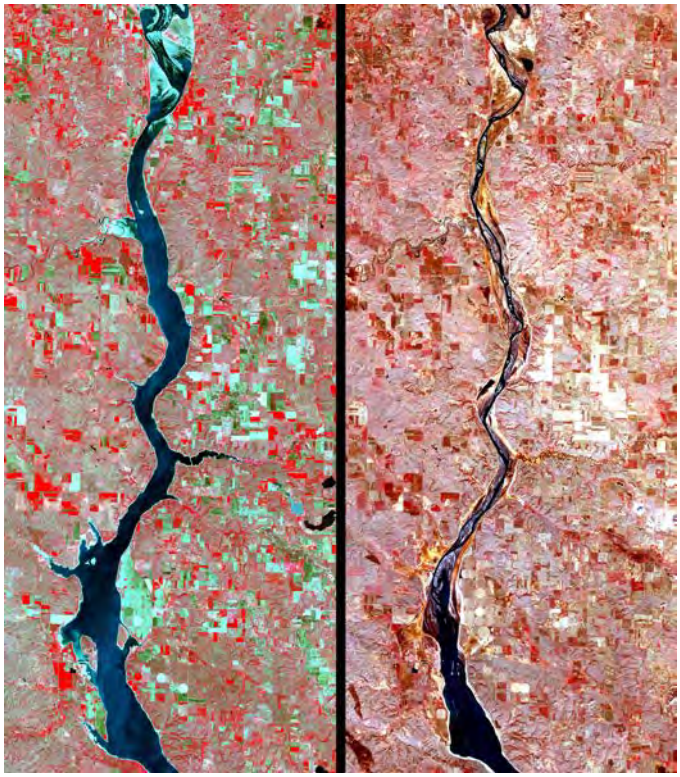
Plains terrain can be modified into badlands and gullies, as seen in this image of western North Dakota near the Montana border:

Most of the northern Interior Lowlands has its topography, vegetation, farm crops, and landscape developed on drifts and tills deposited during the Pleistocene glaciation. The Great Plains in this Landsat-1 scene (in mid-May when vegetation is just emerging) is relatively flat as the overlying till plains is not dissected. Parts of Saskatchewan, Manitoba, and North Dakota are within the scene. However, the Turtle Mountains (elliptical feature) are actually a dissected mesa (of Eocene sandstones) rising about 210 m (700 ft) above the surrounding plains; these wooded mountains were glaciated and have many lakes, making this a popular vacation areas:

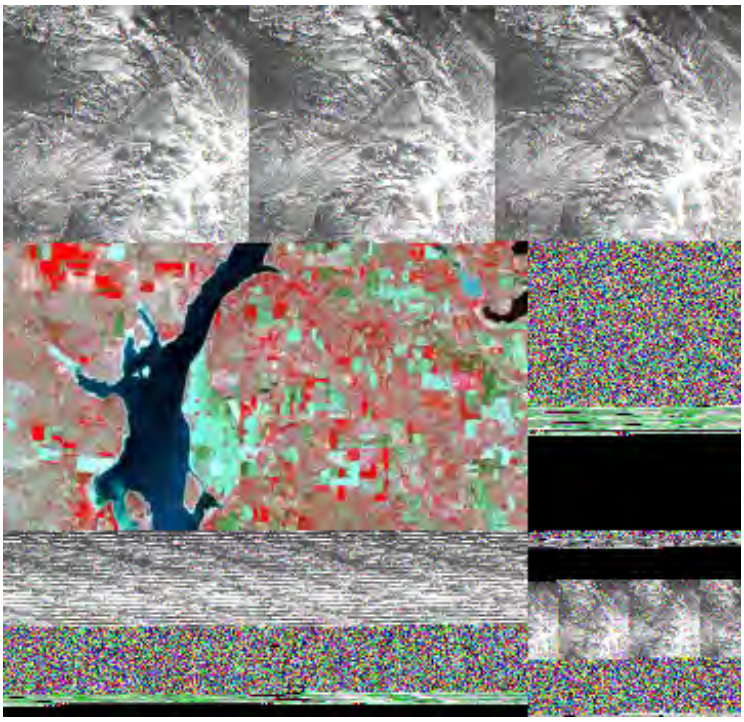


When one thinks of the northern Great Plains, the two Dakotas (North and South) come to mind. Both states actually have their eastern land in the Interior Lowlands but our mental picture sees them as more like the Great Plains. The largest city in North Dakota, Fargo (about 100000), is just west of the Minnesota line;

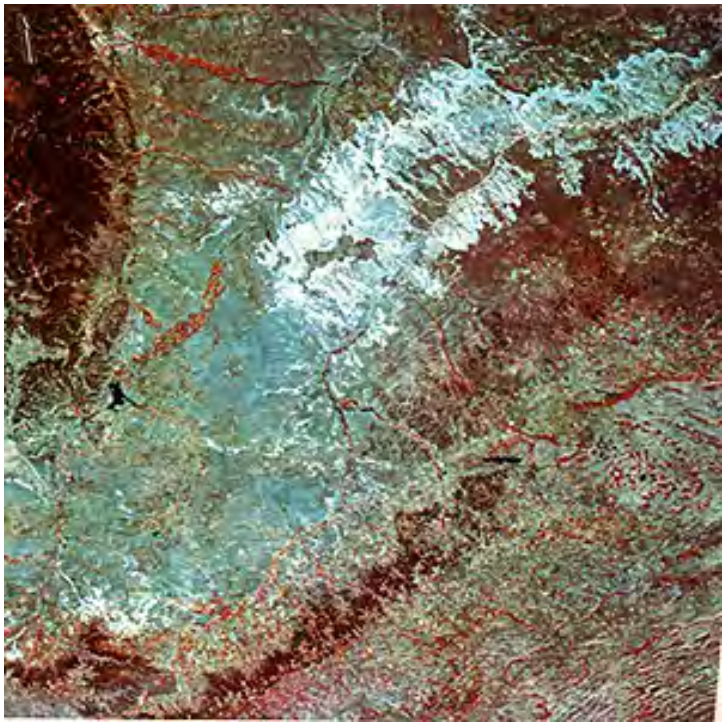
The northern Great Plains has been experiencing a drought of late. The Oahe Reservoir (some 325 km long), normally the fourth largest in the U.S., has lost much of its water as evident in the ASTER image on the right:



It is surprising to some, especially after flying over the Interior and Great Plains, that the region from a broader perspective is not flat but often displays rolling terrain and distinct valleys. This is dramatically revealed in this next portrayal - a Heat Capacity Mapping Mission (HCMM; see page 9-8) Nightime Thermal Infrared view of nearly all of northern Kansas, most of Nebraska, and the southwest corner of Iowa. The Missouri (top) and Platte (center) Rivers control the major drainage. The sense of relief is brought about by the tendency of cooler air (darker) to sink into lower terrain at night, producing an effect somewhat like a shaded relief map.

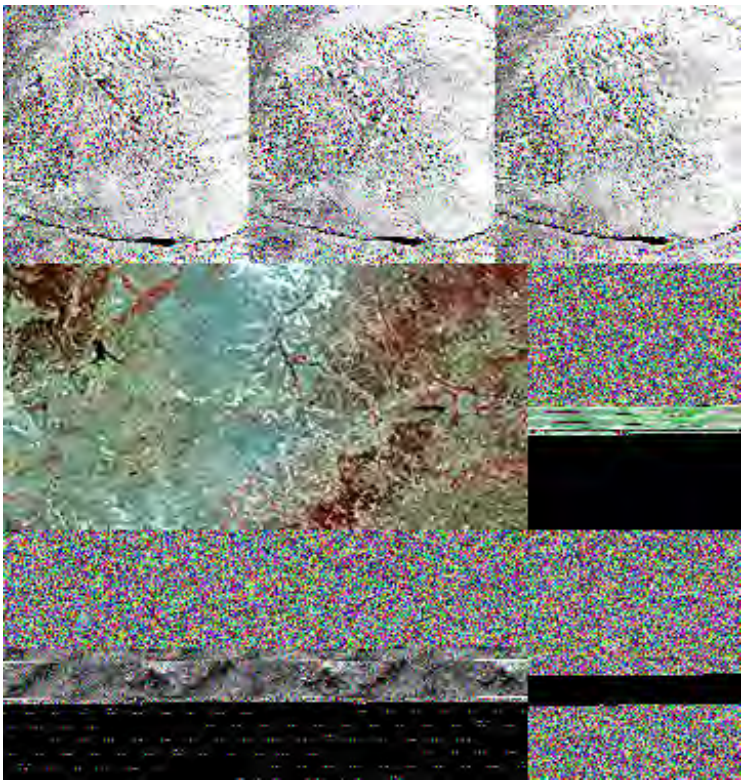


The Great Plains, which grades eastward into the Interior Plains, has a variety of landscapes, including areas that have few farms. A typical example of a more diverse scene appears below. Most of the July 4, 1973 Landsat image is in western South Dakota. On the upper left is a portion of the Black Hills, which rise several thousand feet above the rolling terrain whose elevation exceeds 4000 ft (1100 meters). The Black Hills, home of Mount Rushmore and the Homestake Mines in Deadwood, S.D., are a broad domal uplift exposing igneous and metamorphic rocks over most of the interior. Geologists consider the Black Hills to be an offshoot of the Rocky Mountains, which we will encounter on the next page. The Sand Hills of Nebraska are in the lower right.

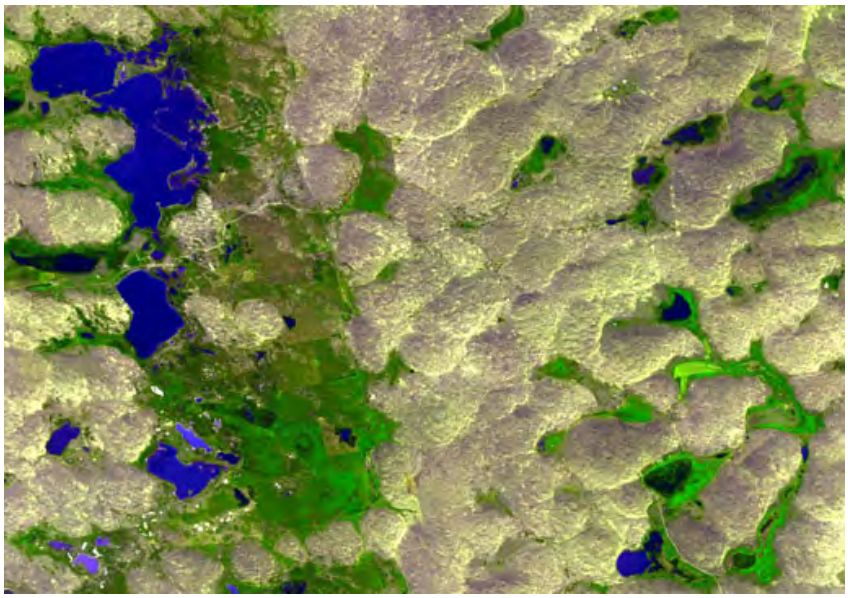


At the lower right corner of the Landsat MSS image (above) is the northwest end of the Sand Hills of Nebraska whose dune fields are discernible. The Sand Hills are one of the largest dune fields in the Western Hemisphere. They formed some 10000 years ago by strong winds in the early post-glacial phase of the last major glaciation. Sand is the interior material in the longitudinal dunes that are capped by loess (fine-grained silt) which allows some vegetation to grow on the dunes themselves. Water is trapped between dunes to form lakes (some evaporate

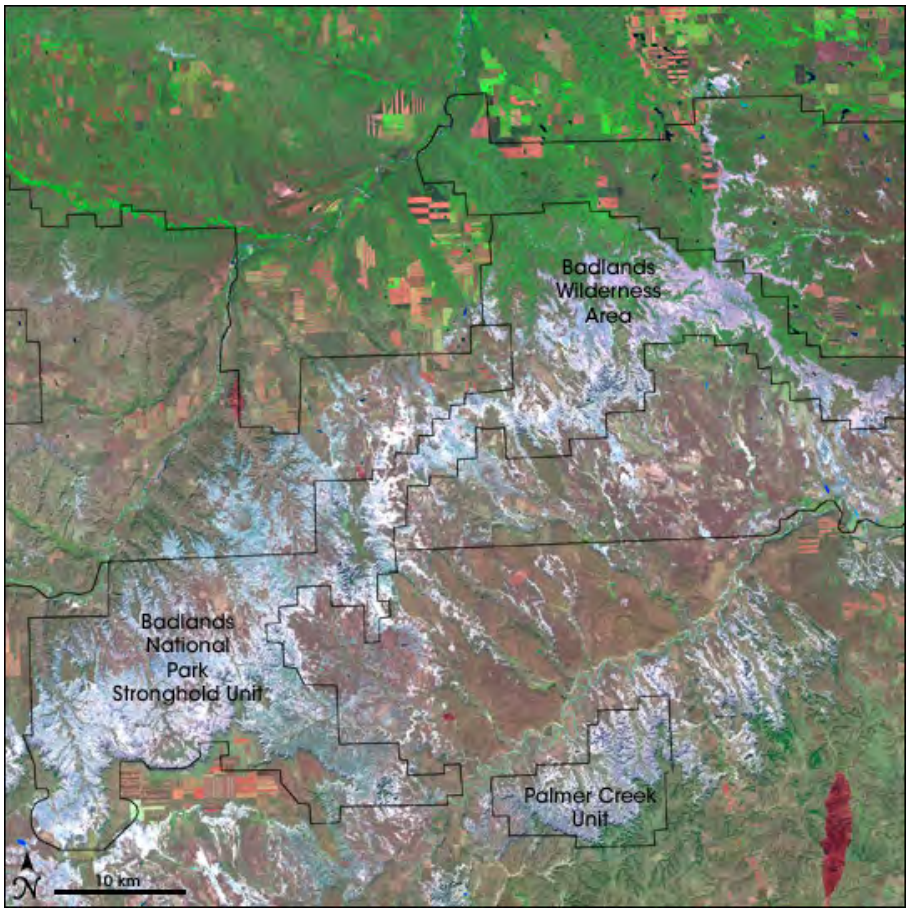
periodically). We show the western 2/3rds of the dune field in this May 15, 1973 Landsat MSS image (Band 5); the main river is the North Platte, the large water body built by damming this river is Lake McConaughy. A ground photo shows typical landscape within the Hills.



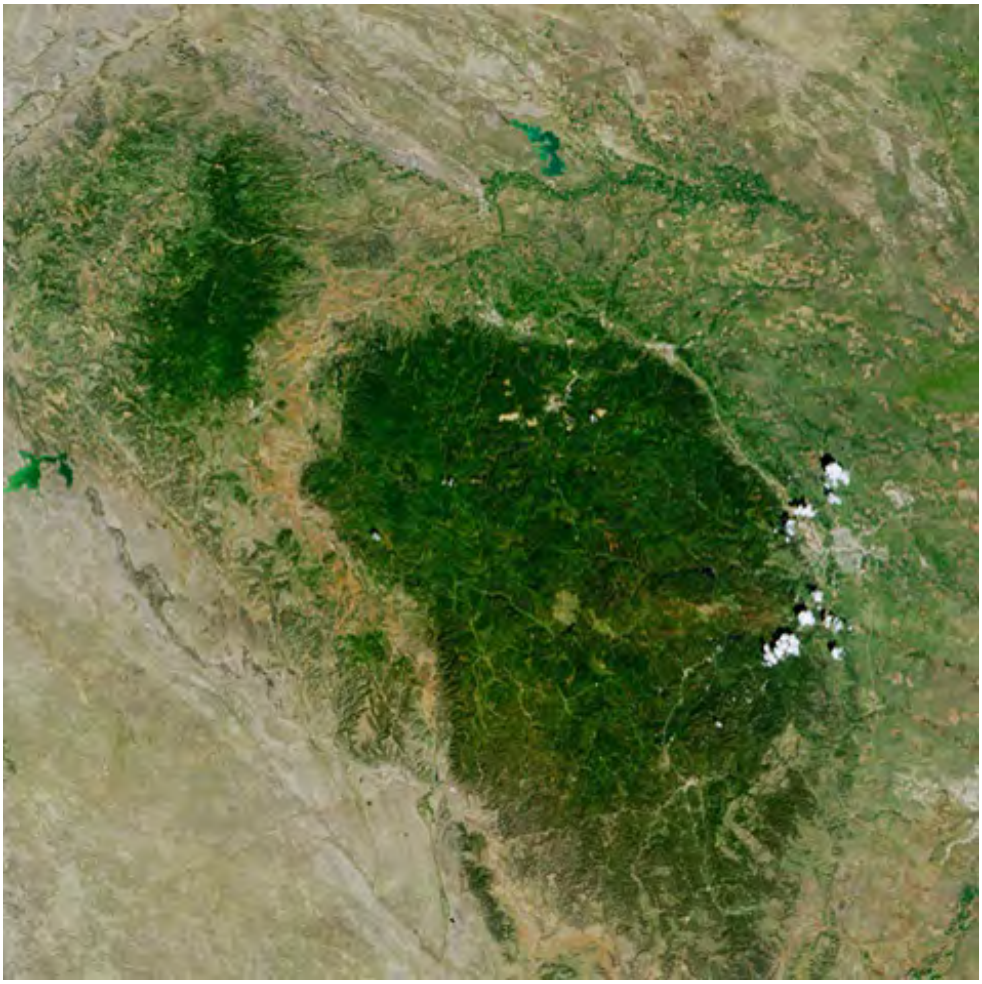
Different impressions of this largest sand dune field east of the Rocky Mountains are given by the next three illustrations: the first, a Terra satellite close-up, the second an aerial view in late winter, and the third a ground scene.



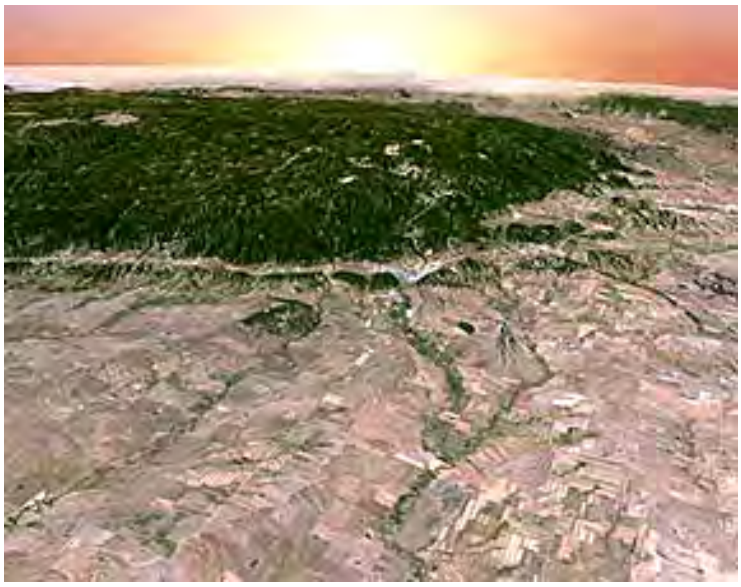
Almost due north is the famous Badlands (bluish-white), an extensively gullied area cut by erosion into soft, easily eroded, often colorful, Tertiary sediments (see figure below). The medium-blue-gray areas are weakly dissected plains with sparse vegetation. This High Plains Tablelands is bounded on its southeast by the forested Pine Ridge escarpment. This Terra satellite view shows more details within the Badlands.


Badlands
Wilderness
Area
Badlands
National
Park
Stronghold Unit
Palmer Creek
Unit
N
10 km

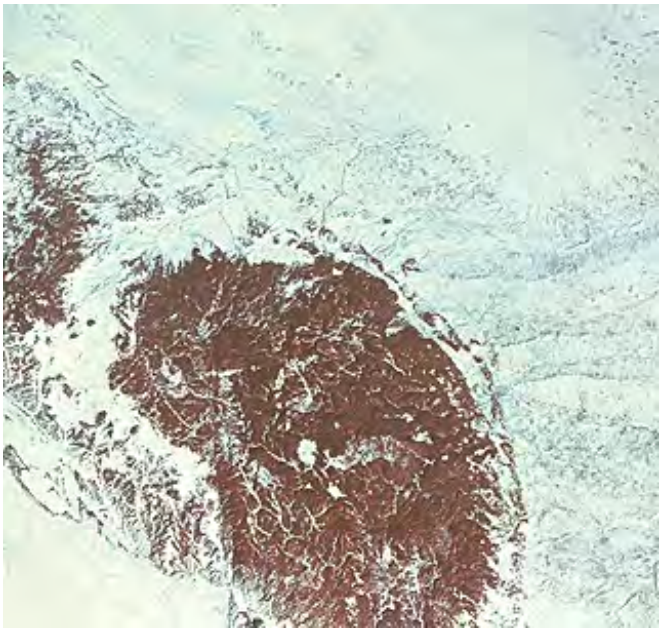
A significant anomaly in the western Great Plains are the Black Hills of South Dakota. This rises to elevations in excess of 2700 meters (8000 ft). They are actually an extension of the Rocky Mountains. They appear green in this Landsat image.



Here is a perspective view looking west at the eastern side of the Black Hills, made from Landsat and DEM data.



The reason that the Black Hills got its name is its general "darkness" owing to a preponderance of dark pine species. This is quite apparent in this Landsat-1 false color image of the Black Hills during wintertime when snow blankets the surrounding plains with snow, which is not visible within the Black Hills because the evergreen canopy prevents its exposure:



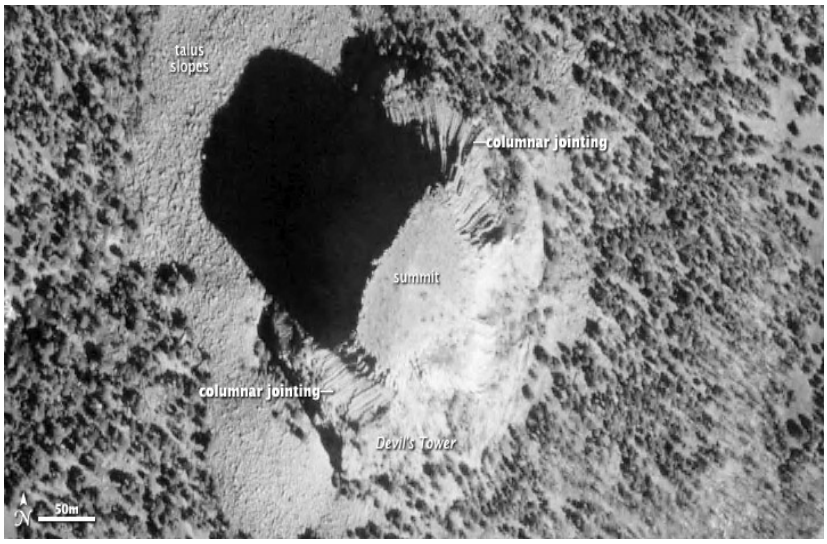
This aerial view shows the main reason for the Black Hills being dark, namely, because most of its vegetation consists of evergreen fir trees:

However, nowadays, in the mood of patriotism brought on by 9/11 and Iraq, the Black Hills have taken on a center stage position as we look back to our roots. On the side of a granite mountain south of Rapid City is the famed rock sculptures of the faces of George Washington, Thomas Jefferson, Teddy Roosevelt, and Abraham Lincoln - the Mount Rushmore Memorial - a National Monument.

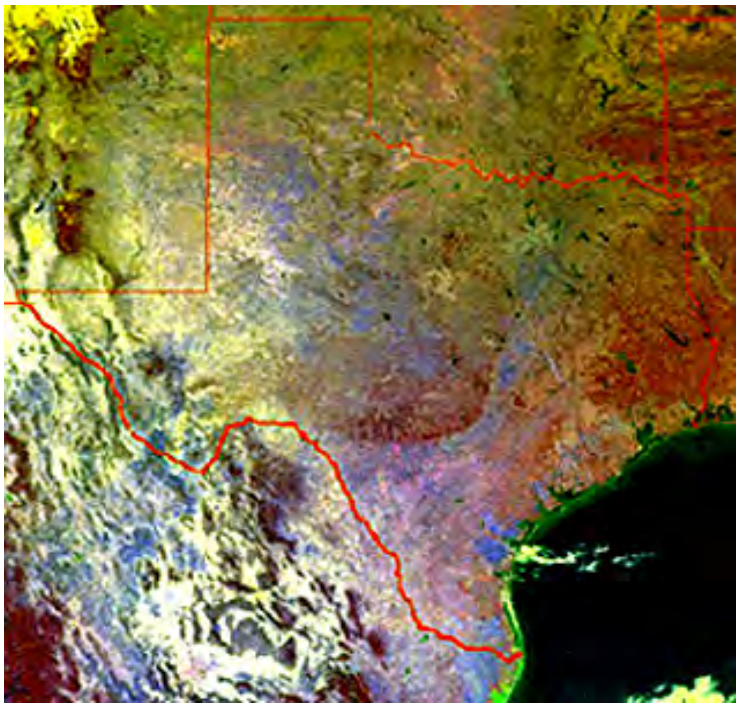
Compare this last image with a ground photo taken through a telescopic lens:

Incidentally, some have questioned the inclusion of Teddy Roosevelt in this quartet. Read his life: perhaps he's not among the greatest (but high up), but few would dispute he is probably the most interesting President owing to his "free spirit" personality.

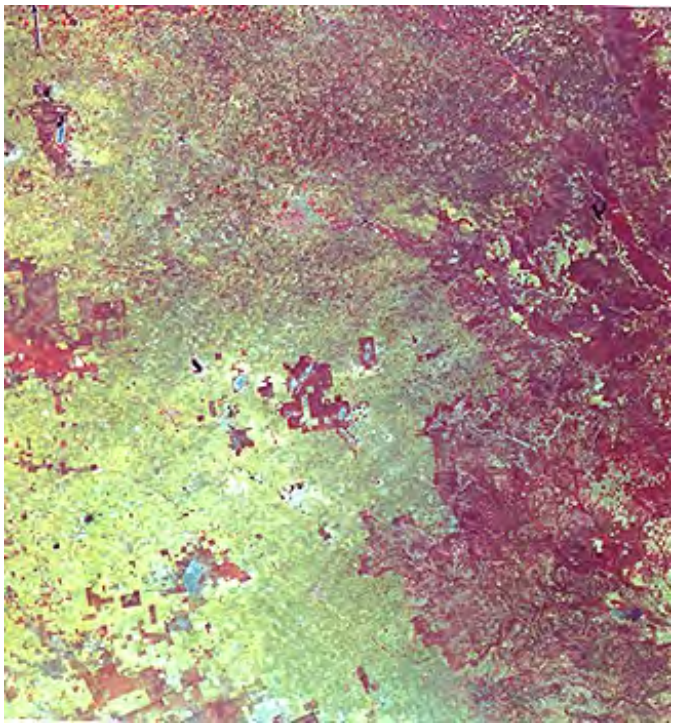
Off the northwest end of the Black Hills is the famed Devils Tower (it was a centerpiece theme in the movie "Close Encounters of the Third Kind"). Devils Tower is actually the exposed central vent of a volcano, with the enclosing volcanic beds having been stripped away by erosion. The volcanic rock is conspicuously marked by "fluting", that is, shrinkage joints produced during cooling of the solidified lava. In the aerial photo below, it is hard to pick out the top of the Tower, but its presence is indicated by its shadow. The Tower (pictured beneath) is nearly flat (it was a popular target for skydiver parachutists until the National Park Service began to arrest the divers [it became a real nuisance to have to have Ranger climbers rescue the divers from the top]).



Lets move off flightline to look at some images from various areas in Texas. This is the largest of the 48 states. Its environmental niches range from pinewoods and deciduous forests in the east through grasslands in the center to desert landscapes in its west. This composite of images made by the NOAA-14 satellite shows the entire state and surroundings (including northeast Mexico):

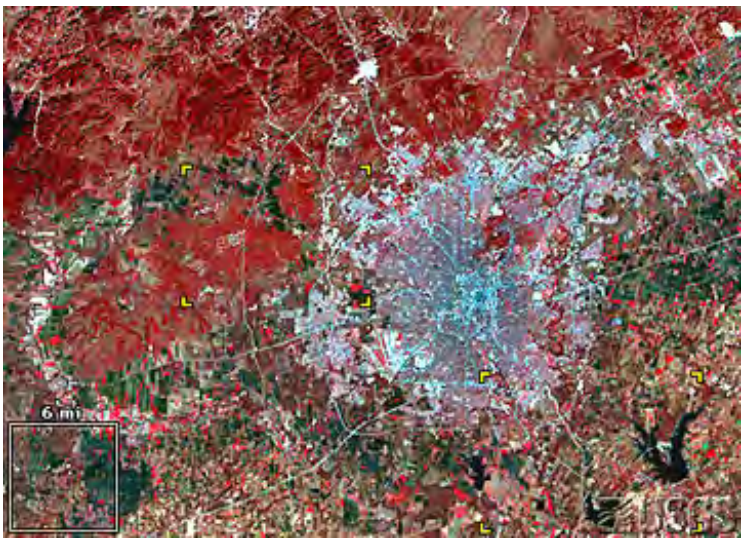


Now look at parts of the Great Plains in Texas. The first image is of the Llano Estacado (Staked Plains) in the Panhandle of Texas. The bluish area locates Lubbock, TX. The Ogallala formation (Miocene in age) is capped by up to 10 m of caliche (calcium carbonate formed by precipitation from upwelling groundwater) and is responsible for the low plateau (in brown) which supports mesquite brushland. The yellowish areas are mainly undeveloped grasslands. Some oil fields - an offshoot of the West Texas petroleum deposits - are present in the scene.



The southernmost section of the Great Plains is the Edwards Plateau in central Texas. Cretaceous Limestones are the dominant rock type. These have been downcut by stream and wind erosion to form canyonlands and gullys in a region of low relief; elevations in this scene range from 600 to 900 m (2000-3000 ft). This is cattle and sheep country rather than farmland. Vegetation is mostly scrub savannah with Junipers and Live Oaks. At the bottom is the Amistead Reservoir formed by damming the Rio Grande. The Pecos River joins it as that river moves south-southeast. Note the canyonland near the left edge of the image.

In the flatlands just south of the Edwards Plateau lies the city of San Antonio.



In the upper left is the forested edge of the Edwards Plateau, bounded by the Balcones Escarpment. Below is a major attraction in San Antonio - the popular River Walk - restaurants and shops along the small Salado River that flows through the town.

The next scene is not in the Great Plains as such but is topographically and ecologically similar to parts of the southern Great Plains in Texas. The region here lies in the Texas Coastal Plains,

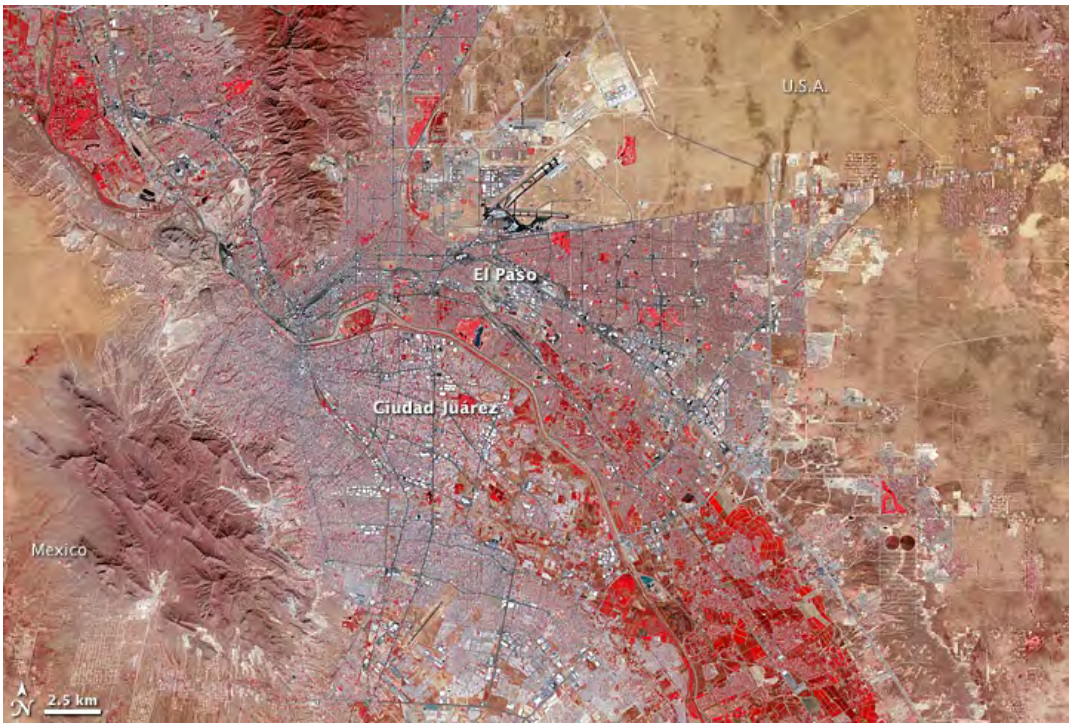
built up, as is nearly all the Coastal Plains province running from the southern U.S. to New Jersey, of Miocene to Pleistocene sedimenary rocks deposited when sea level was higher during marine invasion onto the continent. Corpus Christi (near top center) is found along the mainland side of its Bay. Just north of Corpus Christi is Aransas Pass - winter home of the endangered Whooping Crane. Almost the entire Texas coastline is one continuous string of barrier islands, with lagoons and the Intracoastal Waterway to on their landward side. Best known is Padre Island (bottom to near top), a National Seashore preserve which runs nearly the entire length of this scene. A vast amount of the land south of Corpus Christi was once a huge cattle grazing endeavor, the famed King Ranch, which still exists but has shrunk by sale of some of its holdings.

The NOAA image of Texas, shown above, offers evidence of the more western "flavor" of the state as one moves from San Antonio to the west. This next Landsat-1 scene shows the land around the Big Bend National Park in West Texas; small block fault mountains and volcanic terrain are conspicuous; the Rio Grande is a narrow stream running through the image:

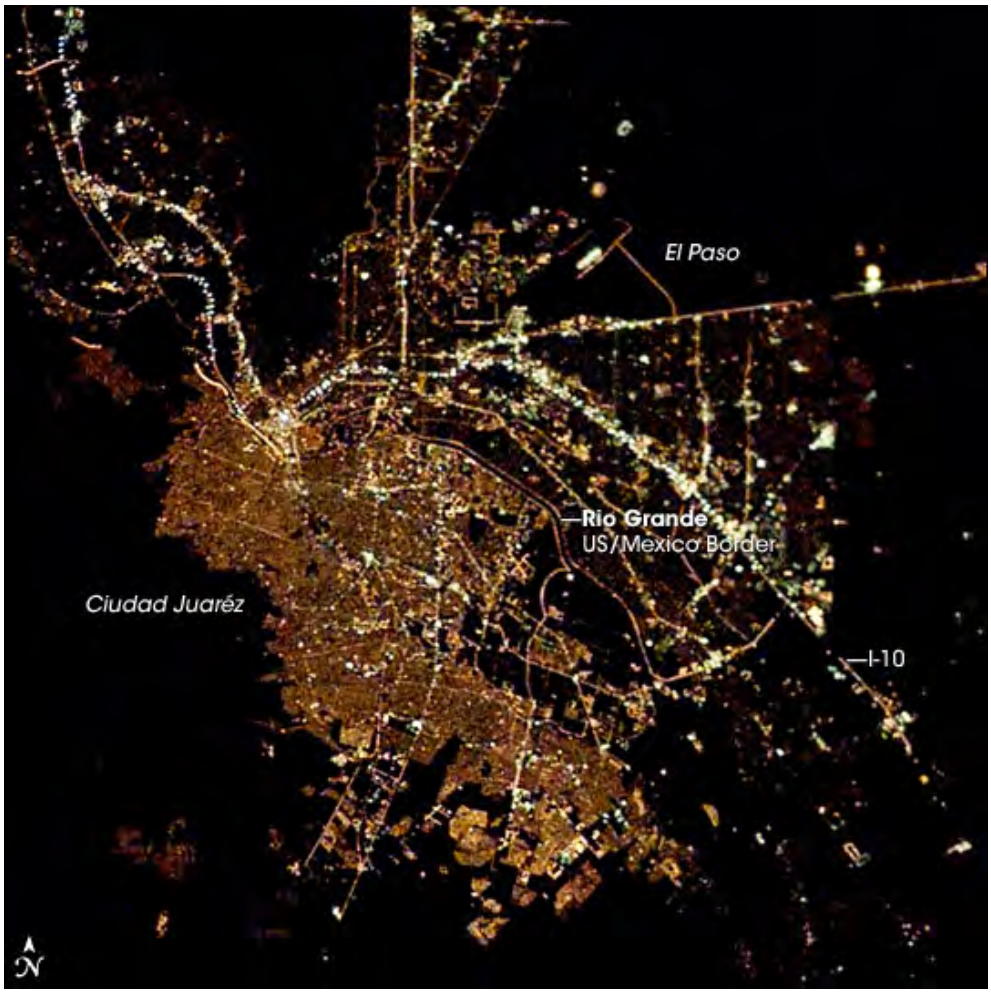
Parts of the Big Bend country are volcanic. The region has perhaps the most rugged scenery in Texas:

This is typical west Texas landscape:

True desert marks the westernmost end of Texas at El Paso, seen in this ASTER image (its Mexican counterpart, Juarez, lies just south of the Rio Grande). El Paso has a population of 775000 while Juarez is up to 1,500,000 people.



Rather mysteriously, in this astronaut photo taken at night, the areal extent of Juarez seems much greater than that of El Paso. A speculation, Juarez has a more active nightlife and thus is lit up in this late night image:



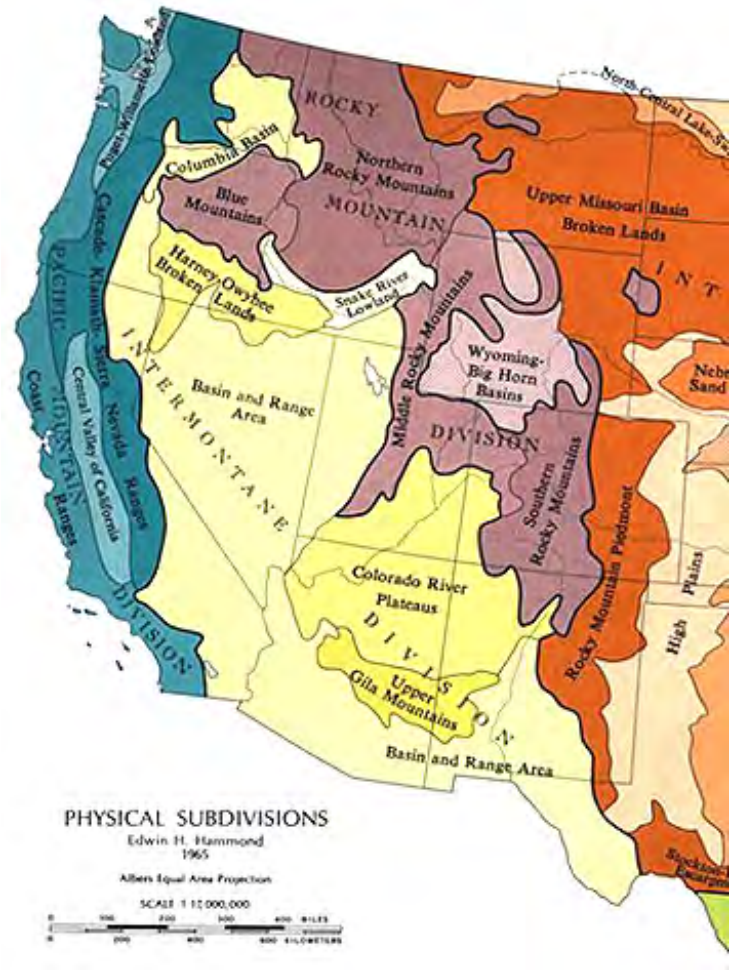
As the western Great Plains approaches the Rocky Mountains, the topography develops gullies (draws), canyons, and a few mesas or plateaus because of increased downward erosion resulting from the higher elevations. Farms diminish rapidly westward and a semi-desert vegetation cover gives hints of true Western U.S. ecosystems, as seen in this image of Southeast Colorado:

Primary Author: Nicholas M. Short, Sr.

In the next flyover, we leave the Great Plains abruptly as the Front Range of the Rocky Mountains rises dramatically from the gently sloping high plains. In the Denver area, much of this zone is taken up by steep-dipping sedimentary rocks that make a feature known as "hogbacks". Both a summer and a winter Landsat MSS full scene show the plains, the Front, and the high mountains of Colorado in the Denver area. Other areas in the Southern Rockies are also examined.

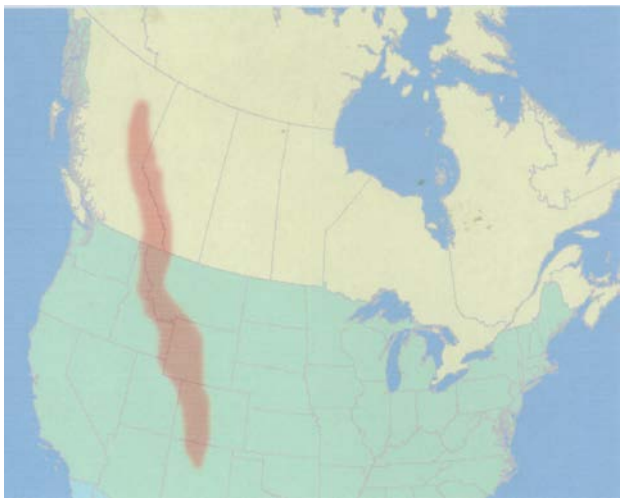
Denver, Colorado and the Southern Rocky Mountains

In the eastern U.S. we passed over an old mountain system - the Appalachian Belt, now rejuvenated by uplift starting some few million years ago. We flew across the vast lowland of the continental interior. From Central Colorado and along the trend from Mexico to Canada, we encounter an array of geological and physiographic expressions of orogenic (mountain-building) events that began in the late Paleozoic and continue today. This map presents an overview of the major landform classes of the western 1/3rd of the U.S.



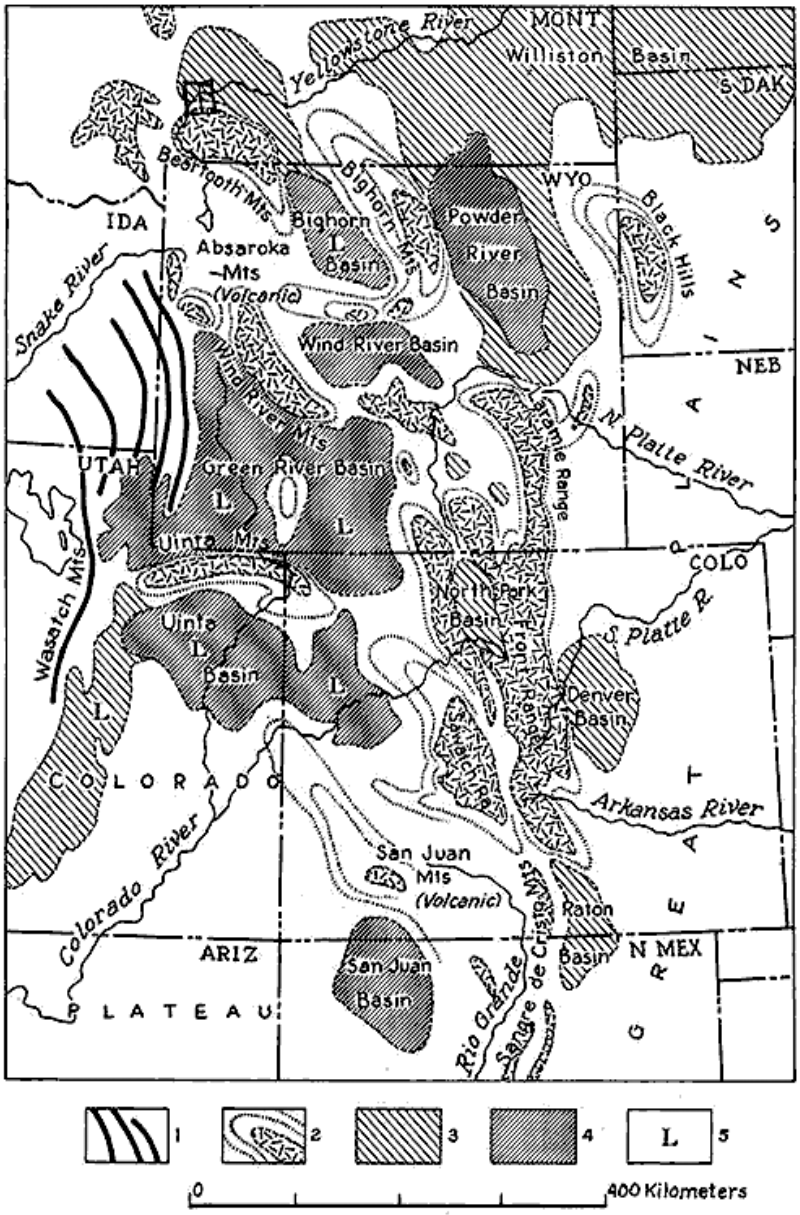
The Rocky Mountains are a vast, generally north-south collection of joined or separated ranges that extend from northern New Mexico well into Canada. The Rockies (as they are often called) are shown here, first in a location map (generalized distribution in rose tone), then within the U.S.

as a shaded relief map and finally as a physiographic subdivisions map:



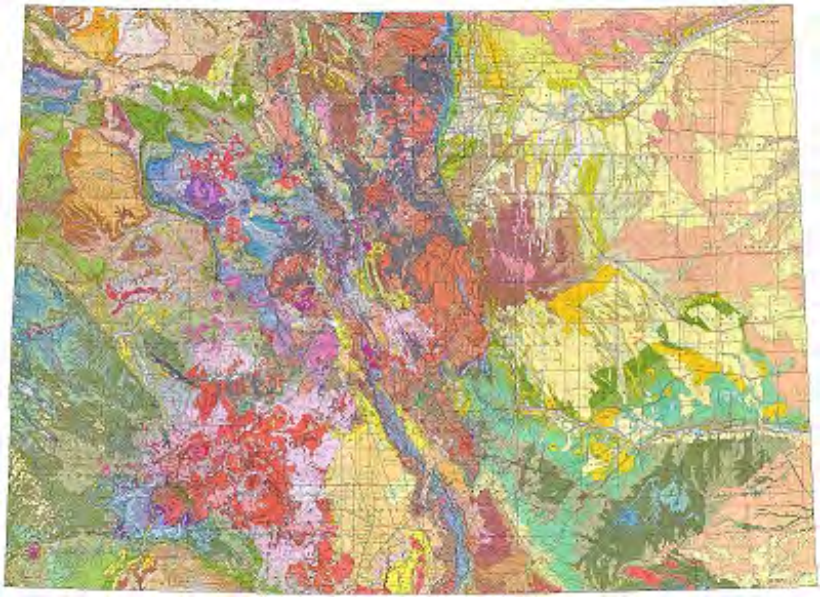
The Colorado-New Mexico-Wyoming segment of the Rockies stands out in imagery taken from space, in part because they are usually heavily wooded (evergreens in particular) and so appear darker:

The major ranges and basins are identified in this map:

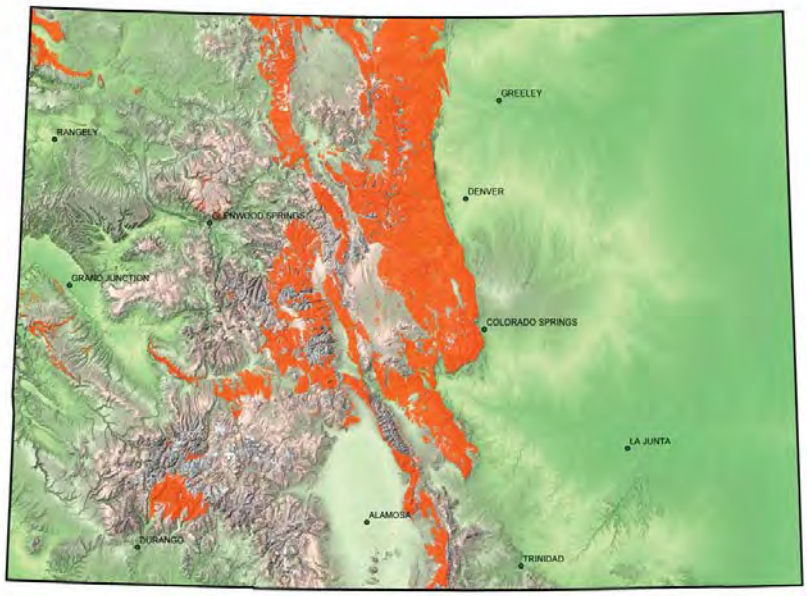


Parts of the Rocky Mountains were being built geologically since the Paleozoic. The main orogeny occurred through subduction to the east at the end of the Mesozoic, with periodic rejuvenation since then. This sketch illustrates the general conditions during the Cretaceous Period as uplift was maximized:

The geologic map (below) of Colorado shows that the core of the Rockies is mainly Precambrian metamorphic and some igneous rocks. The San Juans consist largely of volcanic rocks.

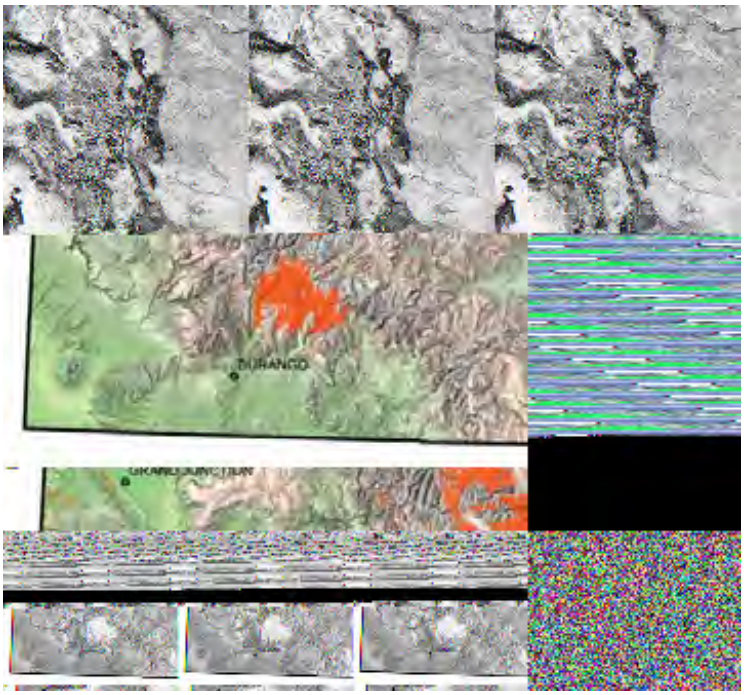


The map below shows the distribution of Precambrian rocks in red. These consist mainly of deeper basement rocks in the North American craton that were uplifted and thrust surfaceward from about 70 million years ago (Cretaceous) to 42 million years ago (Eocene) - a time of mountain building known as the Laramide orogeny. Erosion of overlying rocks has exposed them widely throughout the mountains.



The Rockies in the Southern section, which includes nearly all of the western half of Colorado, are made up mostly of old metamorphic basement rock and some younger granitic intrusions, with remnants of the sedimentary cratonic cover at the time of uplift (climaxing about 68 million years ago). One current hypothesis considers the subducting plate to have been shallower than normal for such an orogenic process.

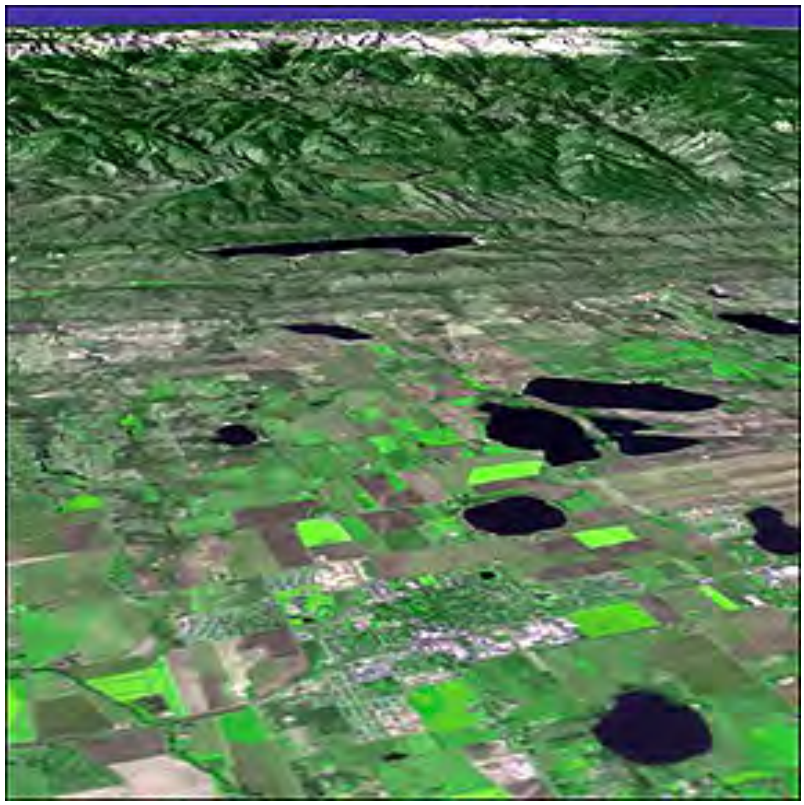
The boundary zone between the Great Plains and the eastern Front of the Rocky Mountains is abrupt, with the transition zone generally less than 1-2 km (0.6-1.2 miles) wide. The 700 km wide scene taken by the Day-Vis channel on the Heat Capacity Mapping Mission (HCMM) sensor shows much of the southern end of the Rockies from northern New Mexico into the southern half of Wyoming. The map below it indicates that the entire state of Colorado is included in the image.



In Colorado, the principal mountain groups are the Front Range, the Sangre de Cristos, the Sawatch Range, and the San Juan Mountains.

In this part of the United States uplifted and folded rocks making up the the main body of the Rocky Mountains occupy a relatively narrow strip passing through the central part of Colorado and New Mexico, and swinging northwest through most of Wyoming (the Wind River Mtns mark the shift that carries the Rockies into Idaho and Montana). This limited width of the orogenic belt comprising the Rockies may seem a surprise because once in these mountains, their lofty grandeur gives an impression that they extend well to the west. In fact, from western Colorado into Utah, mountains of a different type, found in the Colorado Plateau (next page), predominate. The San Juan Mountains look to the traveler much like the Rockies but they are actually a great pile of highly dissected volcanic flow units intermixed with sedimentary rocks and are considered a separate unit.

The Front is well displayed by this perspective natural color image using Landsat-7 data merged with DEM topographic data. The foreground area shown is the western terminus of the High Plains around Greeley, Colorado; the reservoir in the foothills is near Fort Collins.



The Rocky Mountain National Park is in the background, near the snow-capped high mountains near Estes Park. This popular destination for tourists lies within the image below, made from combining a Landsat TM scene with the Space Shuttle's STRM radar data:

It was in Rocky Mt. NP that the writer (NMS) had one of the great adventures of his life. In the summer of 1953, I climbed Long's Peak (14255 ft; 4345 m) along with two friends. This was a challenge but records indicate that this mountain is perhaps the most frequently climbed of all peaks in the U.S. Still it was difficult because of the high altitude. This is a view of Long's Peak

looking eastward from a valley to the west:

A closer look:

Before we began what was to be a full day's expedition to the top and back, we became acclimated to the altitudes by climbing a lower but still challenging Hallett's Peak near our motel

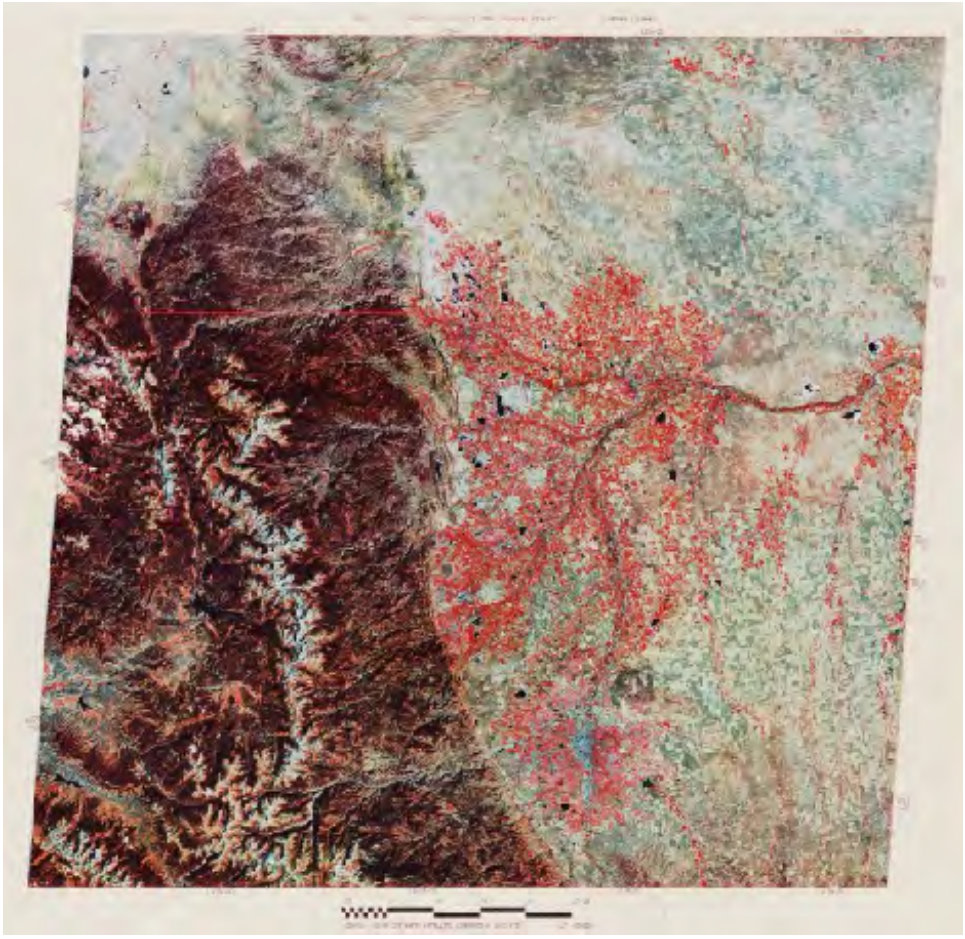
Two days later, the climb began in the afternoon, reaching about 11000 ft. before setting up camp and making dinner. The next morning I awoke at about 5 AM and venture to an overlook at the "Diamond", a name given to the east face of Longs Peak.

As the Sun's first rays lit the east face, I stood in awe of a momentary scene that remains one of the top experiences of my life. About 100 feet from me a Rocky Mountain sheep (a ram) rose, looked at me, nodded (so help me!), and calmly ambled off.

We reached the top by about 1 PM. One could look about in all directions and be looking down (Longs Peak is the highest in the National Park). Here is such a view looking to the northeast:

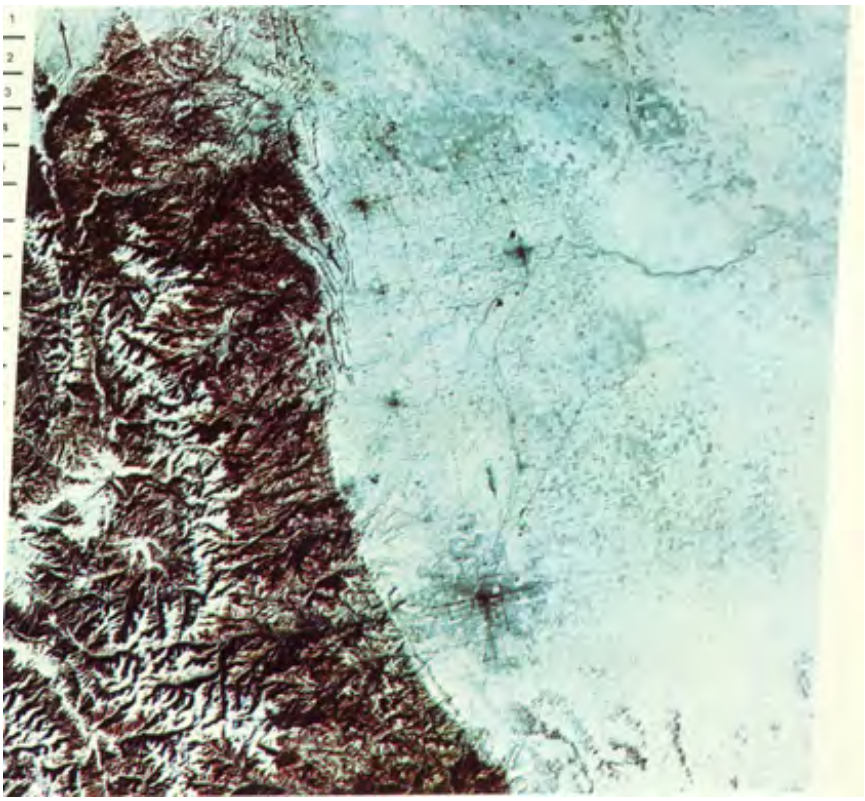
With the usual afternoon storm kicking up by mid-afternoon, we made a quick descent, reached our lodge by dinner and treated ourselves to a victory steak dinner. Thus, the capstone to a marvelous adventure. Nevermind that thousands of others have climbed this Peak - I did it myself and am proud.

The Rocky Mountain National Park is included in this false color composite of central Colorado acquired in early October, 1973 soon after the launch of Landsat 1.



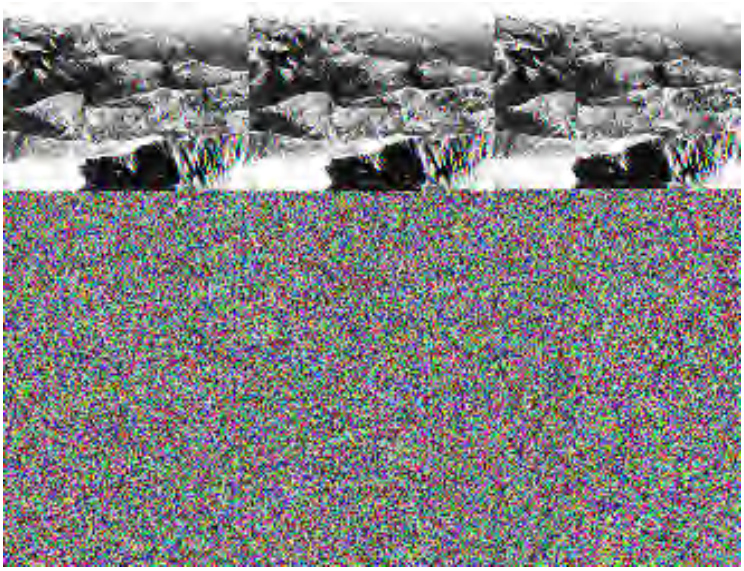
Red colors within Denver and around Boulder and Fort Collins signify trees and crops still bright in the near-IR in this October scene. Their distribution is affected by proximity to the Platte and other rivers. These rivers provide irrigation for the alfalfa, hay, oats, corn, beets, and barley grown in this part of the Plains. Uncultivated areas, such as the Pawnee Grasslands (brownish patch at the right center), contain buffalo and grama grasses.

A January 1973 winter view of Denver places the region in a rather different perspective. One consequence is the "heat island" effect of towns and cities, in which melting and snow-plow removal cause the streets and roofs to appear very dark from the loss of the snow cover.

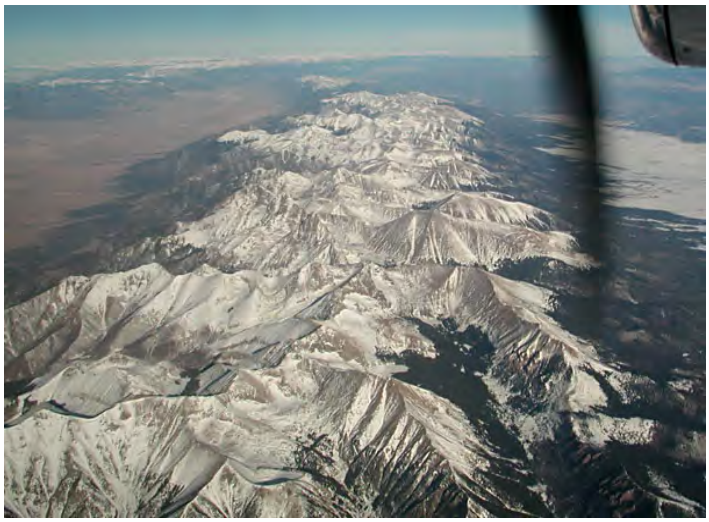


6-9: This winter scene brings out the infrastructure of much of Denver. Four large towns are now visible to the north; name them. Are they visible in the summer scene. Explain the brownish color associated with the Rockies. Is there snow above the treeline? What factor(s) help(s) to indicate drainage patterns. What happened to the lakes so visible in the summer image? ANSWER

The Rocky Mountains form the eastern edge of the North American Cordillera, which is made of groups of diverse, usually complexly-folded and faulted blocks of crust. The blocks are uplifted or thrust against other blocks separated by structural basins. They were deformed and emplaced during periods of major orogenies that were often disconnected in time and place. Rock units in the Rockies were finally compressed and shoved upwards about 65-70 million years ago. The present topography of these mountains is the result of strong erosion thereafter that has lowered their original heights to under 4,420 meters (14,500 ft). Alpine glaciation has steepened and widened already deep valleys, creating the rugged vistas that make this region especially popular for tourists and skiers. The barren (whitish) area along the crest of the Front Range in the October image coincides with the Continental Divide (a high elevation line where rainwater on its eastward side drains into east-flowing rivers that empty into the Gulf of Mexico and on the west side into rivers that ultimately reach the Pacific Ocean) that runs largely above the treeline. The slopes below are forested with Douglas fir, spruce, pine, and aspen trees. Rocky Mountain National Park is in the left center of the image. The southern end of the Medicine Bow Range in Wyoming extends into the image at the top left. The photo below shows a typical landscape within this part of the Rocky Mountains.



Flying over the Colorado Rockies in a commercial flight establishes the distinct character of these mountains.



Denver, CO, appears at the edge of the plains in the lower center in both the Fall and Winter scenes. This burgeoning city is shown first from space and then in an aerial oblique view, with the Rockies in the background.

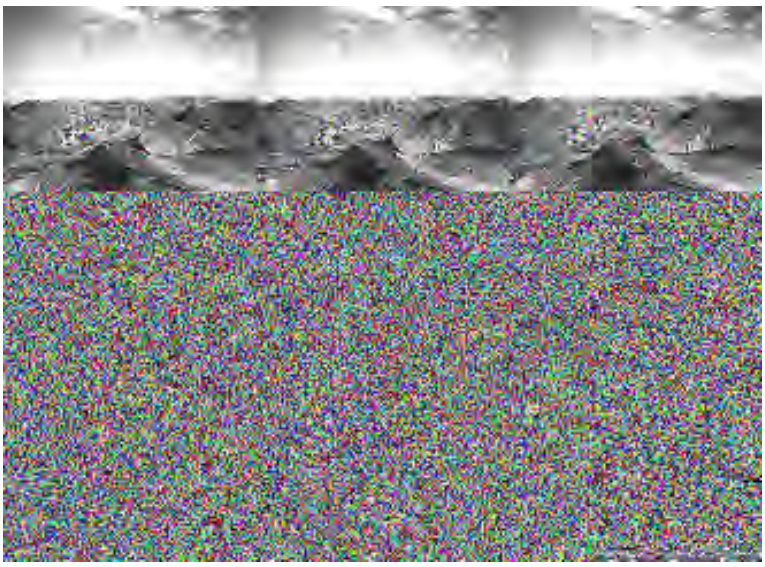
Courtesy: Carolina Map Distributors

The Denver metropolitan area stand out in this nighttime astronaut photo:

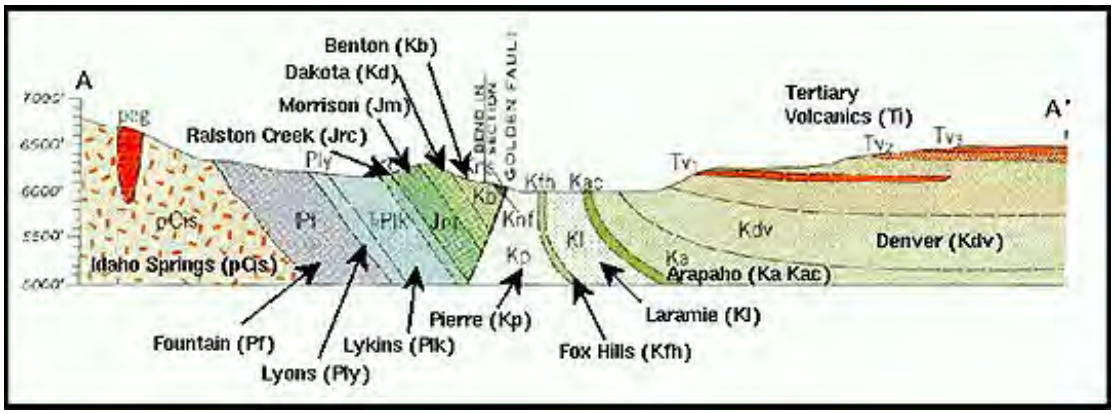
Just west of downtown Denver is Invesco Stadium, home of the Denver Broncos. This IKONOS image shows the details in this part of the city.

To the west of central Denver, the land slowly rises until it abuts against the foothills of the actual Rocky Mountains. This is shown here in a Landsat-7 ETM+ image.

To Denver's west, against the mountains, are narrow outcrops of red sandstone (Lyons Formation) that bend upwards at about 45° to heights up to about 50 meters (160 ft) or more; see labeled "Red Rocks". These landforms are "hogbacks," made up of steeply dipping sedimentary rocks inclined down to the east as the Front Range block of rocks was pushed upwards. The hogbacks are quite evident in this aerial oblique view taken along the Front south of Denver.



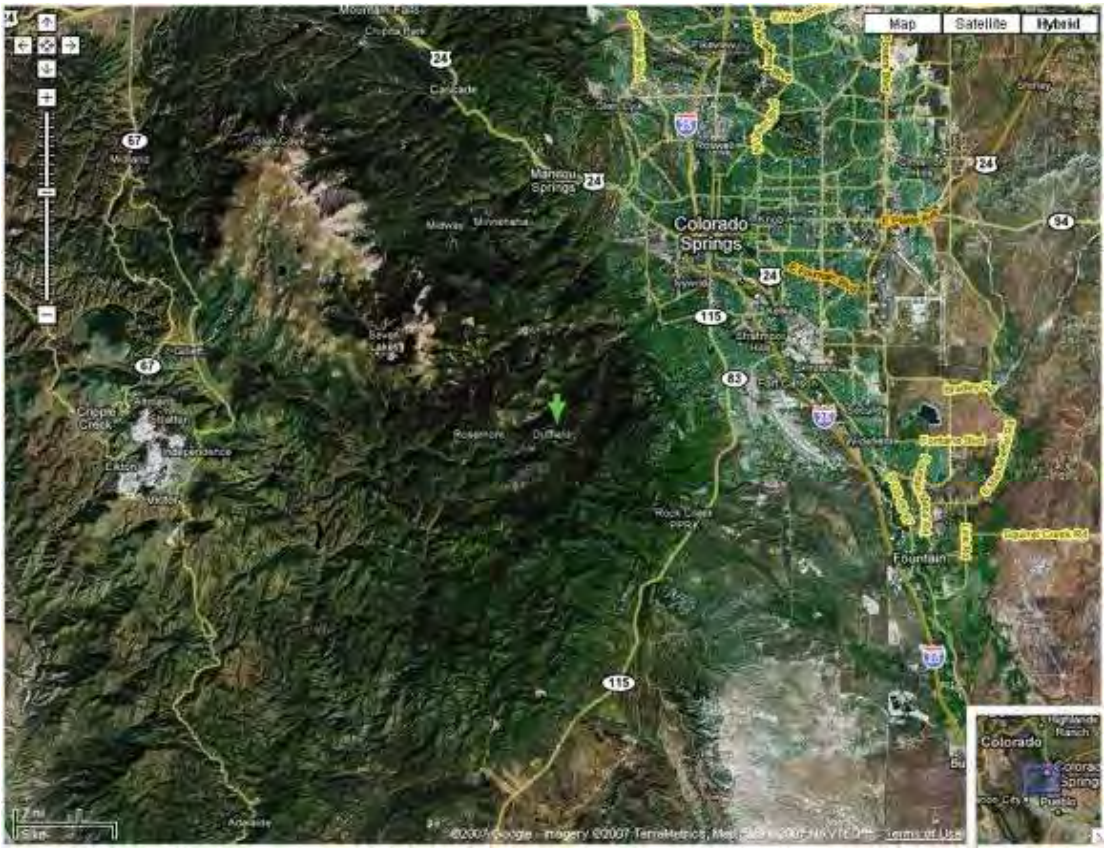
The structural configuration of the rocks at the Front is shown in this geologic cross-section, beneath which is an exposed section of the geologic units at a roadcut along I-70 (there is a sign at the cut which explains the local geology).



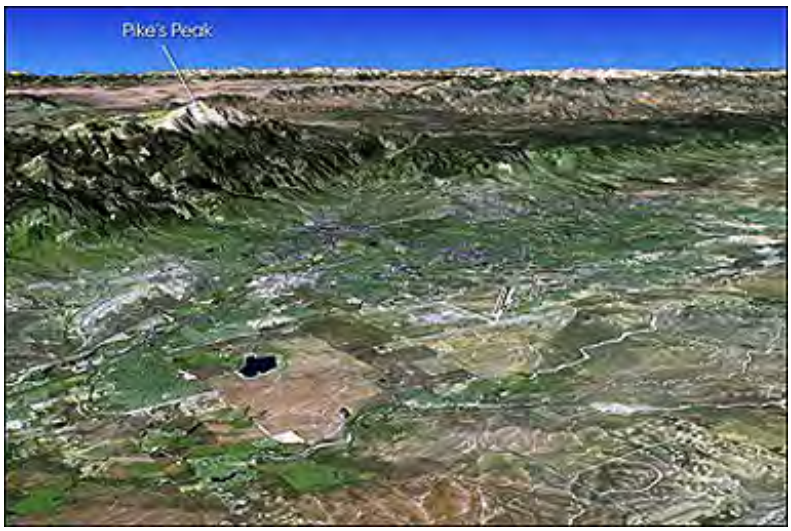
Within the hogbacks just north of Interstate 70 is the famed Red Rocks Amphitheatre, first opened in 1941. There each summer are classical concerts given by the Colorado Symphony and concerts by rock bands and other popular music groups. With hogbacks on either side, the acoustics are surprisingly good in these outdoor conditions. Here is a ground photo of the Red Rocks and a photo of the amphitheatre midst them:

One of the most handsome campuses in North America is that of the University of Colorado, in Boulder, CO. It has the reputation of being a "party school" but there is also quite good science being done there, including a center for Remote Sensing applications, under the direction of Dr. Alexander Goetz. Here is an aerial oblique view of the campus and surrounding homes;

The second largest city in Colorado is Colorado Springs, about 170 km (100 miles) south of Denver, which also is situated at the Rocky Mountain Front. It is the home of the Air Force Academy. Here is a Google satellite image that show where Colorado Springs in context with its Rocky Mountain setting:



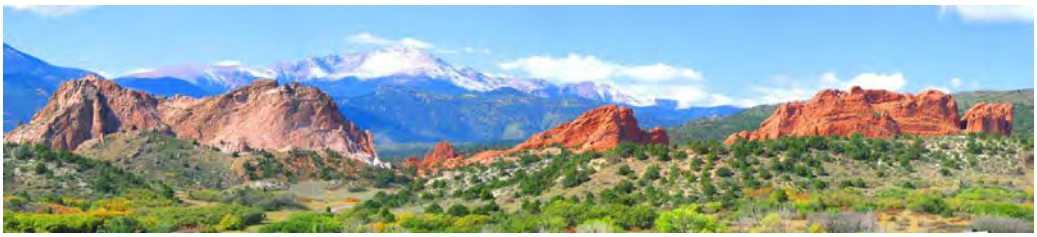
Look next at a Landsat-DEM perspective view that shows Pikes Peak in the high Rockies.



Colorado Springs has grown into a major city, as evidenced in this panoramic photo with Front Range mountains in the background:

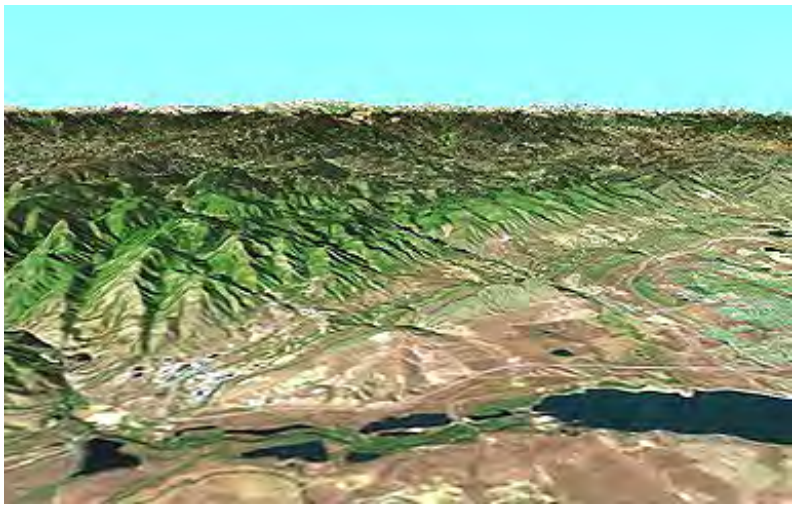
Just north of the city is the beautiful Garden of the Gods, another series of red sandstone

hogbacks (also called "flatirons" from their shape) such as we saw west of Denver.

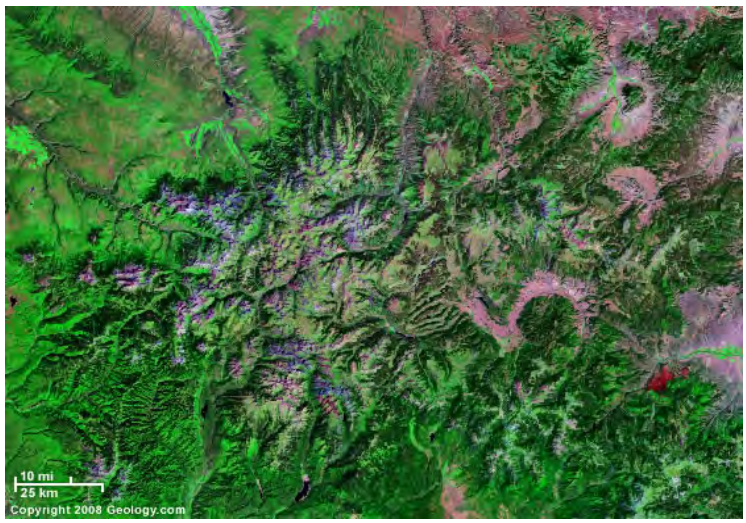


Visitors to Colorado often include a trip to the top of 14000+ ft Pikes Peak. An auto road is open in summer to the top. There is also a cog railway that reaches the summit. The writer (NMS) has taken both journeys.

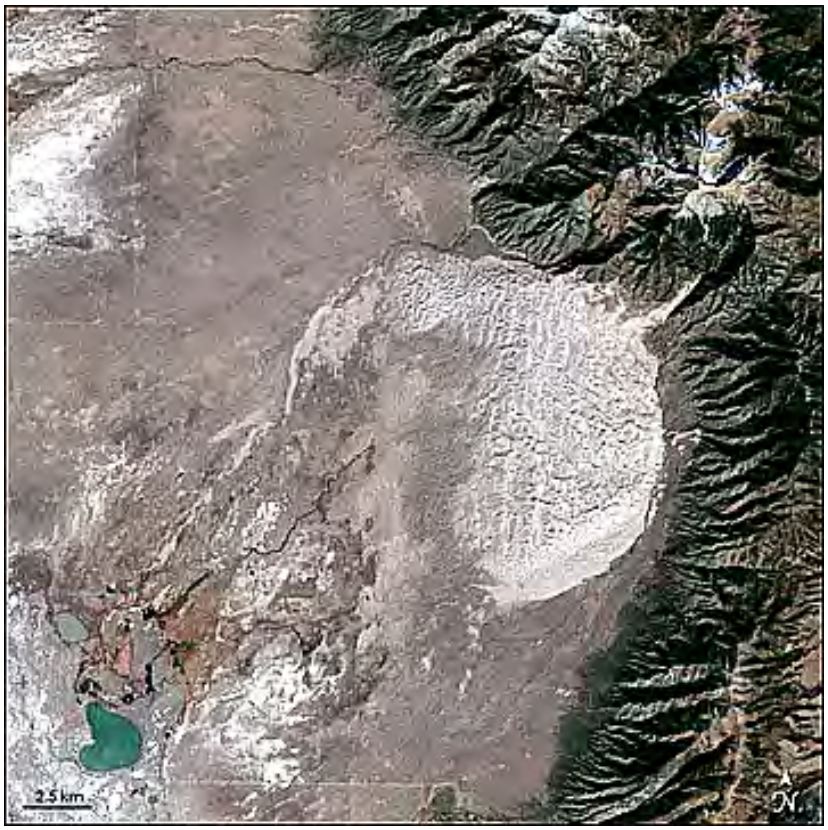
Moving southward from Colorado Springs, note this perspective view (Landsat + DEM) of the Rocky Mountain Front near Walsenburg, CO :



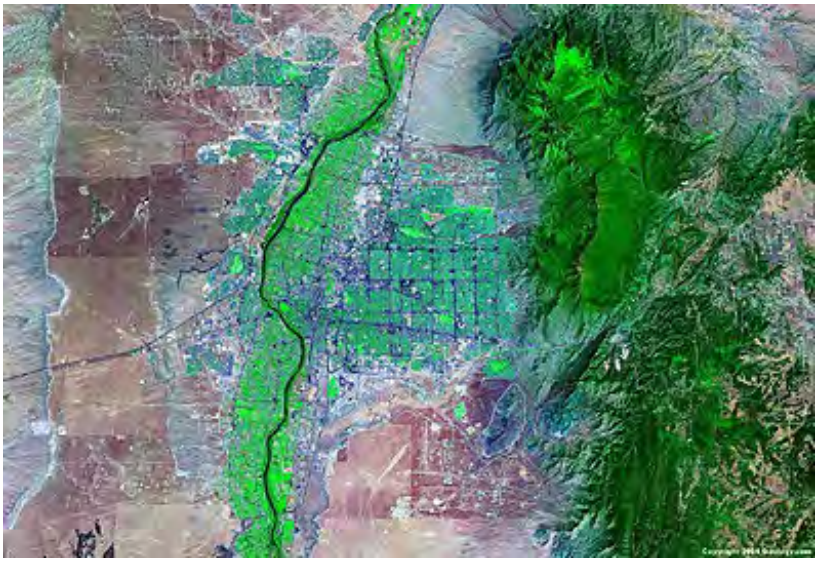
Although a part of the Colorado Rockies, the San Juan mountains in the southwestern part of the state are different geologically. Some igneous/metamorphic rocks, and sedimentary rocks, are present but the main rock types are volcanic. Here is a satellite overview of these mountains, and beneath it is an ASTER image that shows the Creede caldera, a volcanic complex:



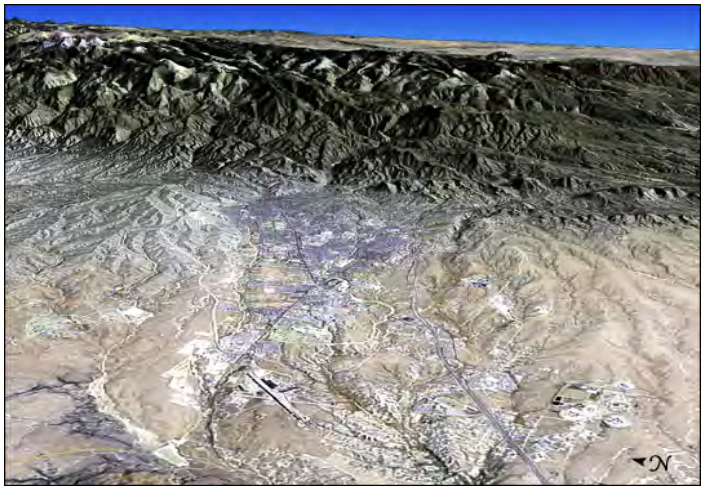
Before leaving Colorado, we point out that on the west side of the Sangre de Cristos above the New Mexico border lies one of America's largest sand dunes fields - The Great Sand Dunes National Monument. Although small in area, windblown sand from western winds piles up against the mountains to produce dunes as high as 220 meters (700 ft), as seen in this Landsat-7 subscene, below which is an IKONOS image that covers a part of the dune field and then a ground scene:


2.5 km
N

Our last look near the south end of the Southern Rocky Mountains is in New Mexico in this Landsat image that includes Albuquerque, the Sandia Mountains to its east, and the Rio Grande to the west:



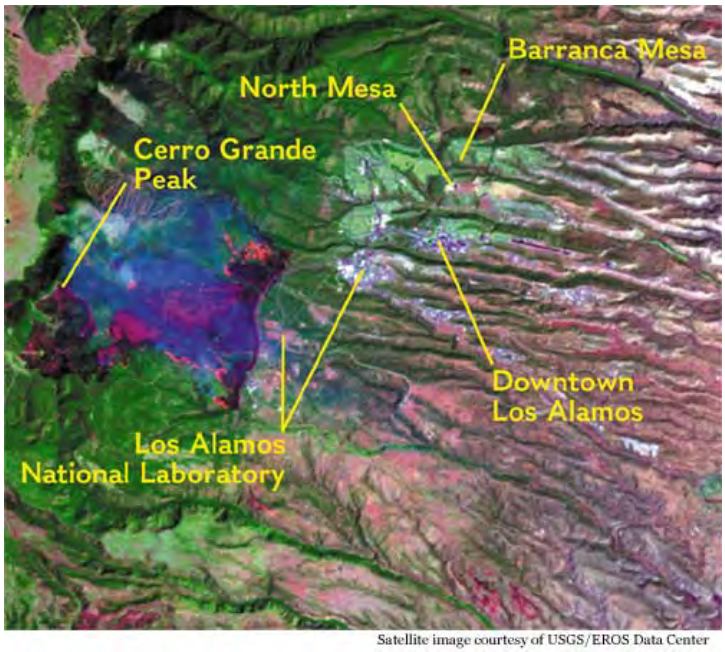
To the northeast is Santa Fe, New Mexico's state capital, seen in this perspective made from a Landsat image. The Glorieta Mountains are in the background to the east:



To the northwest of Santa Fe, are the Jemez Mountains, seen in this satellite view. At their center is the extinct volcano known as the Valles Caldera. On the east side is the town of Los Alamos, built from scratch starting in 1942 to house the thousands of scientists and support

personnel who together developed the first atomic bomb. After World War II, Los Alamos has remained as a research center concerned with atomic energy and its uses, and more recently basic energy research.

Here is a space image of the setting for Los Alamos, and an aerial view of the present day Los Alamos National Laboratory:

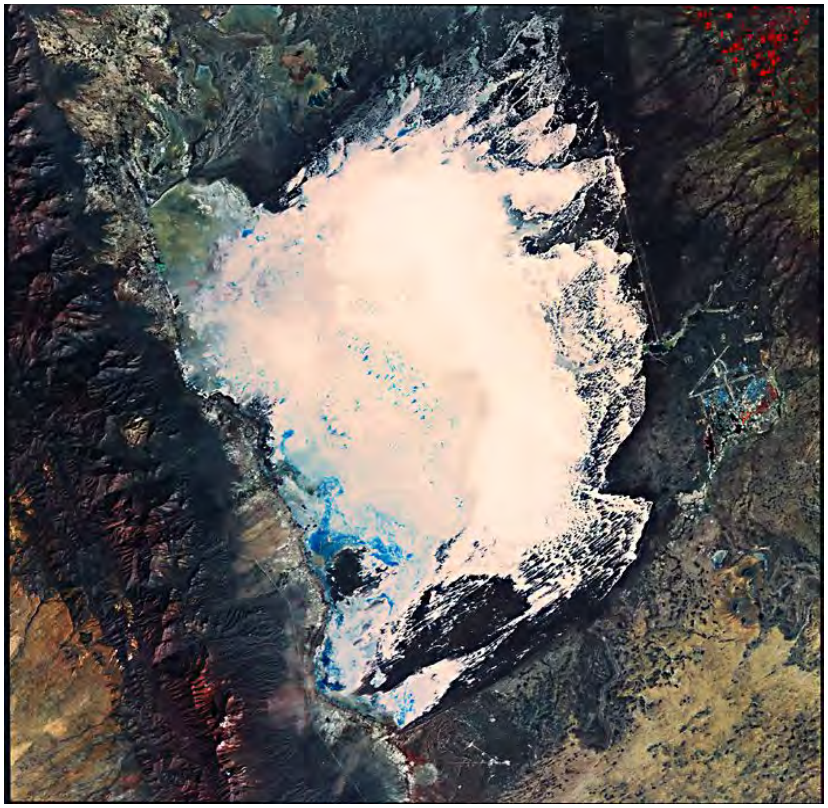


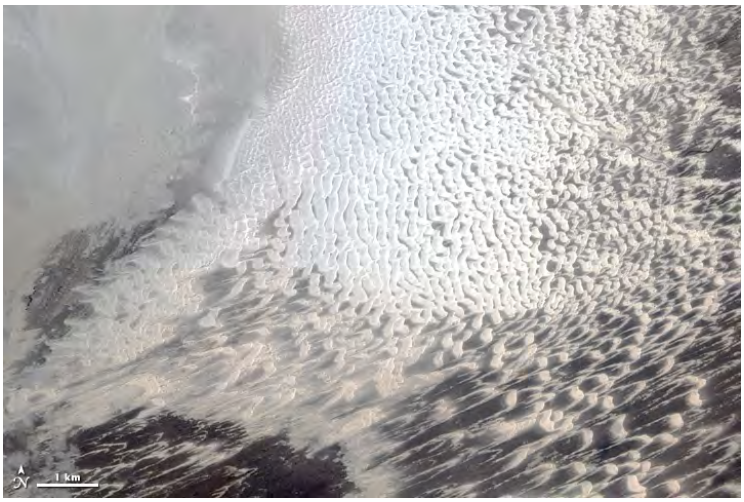
Satellite image courtesy of USGS/EROS Data Center

East of the main surface expression of the Rocky Mountains in New Mexico are several mountain masses that can be considered as uplifted basement rocks and folded sedimentary flank rocks. Below is a Landsat scene that has a true western flavor in which the Rocky Mountain blocks lie between broad lowlands.

The most conspicuous feature in the scene is the White Sands National Monument. The sand is composed not of quartz, as is usual, but of Gypsum ($CaSO_4.2H_2O$). It was there that the U.S. for nearly a decade launched captured German V-2 rockets and some early U.S. rockets. The program was under the direction of Dr. Wernher von Braun, the brilliant scientist-engineer who led the German rocket program during WWII. Also shown are the Sacramento Mountains (dark red), covered mostly with evergreens (Ponderosa Pine); these mountains are the southernmost extention of the Rockies. Mountains of the Basin and Range province's easternmost extension (the San Andres in the center) occur in the center and to the west of the scene (two small segments of the Rio Grande River appear at the left of the image). The towns of Alamagordo and Tucumcari lies at the western base of the Sacramento Mts; not far to the northwest, in the Tularosa Basin is the site of Trinity, where the first atomic bomb was detonated in July 1945. The dark patch is the Malpais basaltic lava field, a young volcanic extrusion of late Pleistocene age.

The gypsiferous sand dunes at White Sands are probably the most interesting in the U.S. and have an historical and contemporary relevance, worthy of some extra inserts on this page. First, we show three views from space:



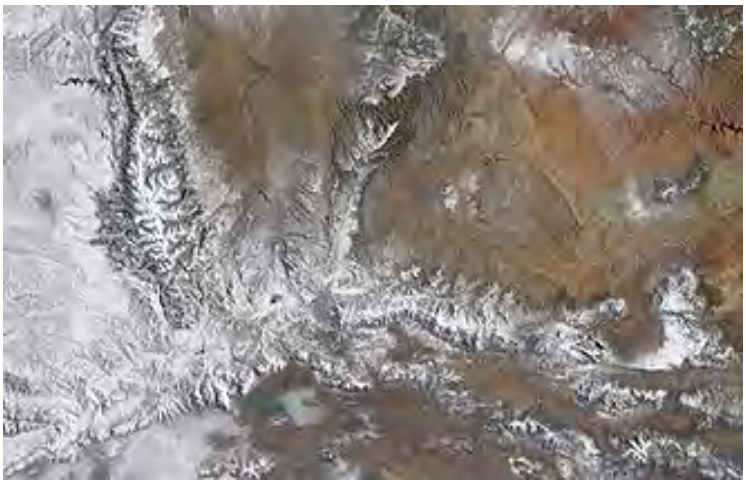


A feel for what it looks like if one were in the midst of these very white sand fields is given by the next two photos.

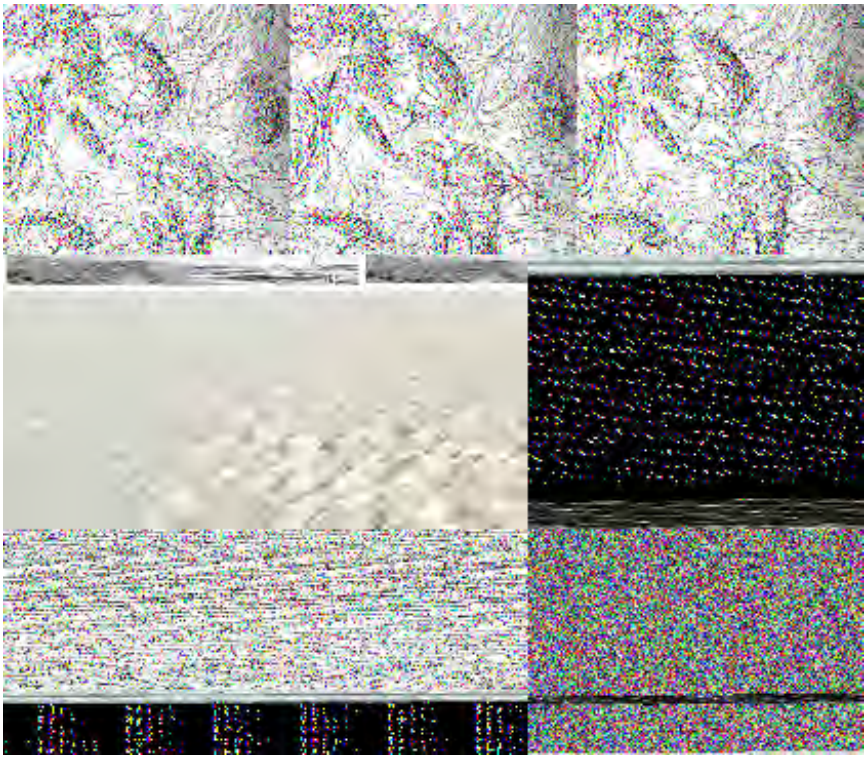
As mentioned above and covered on page 7 of the Introduction, the Whites Sands field was chosen as the site for the first U.S. rocket launches in the postwar era after World War 2. Here is a view of an impending launch from the White Sands Proving Grounds:

The military still uses White Sands as a launch site. NASA, too, has facilities just off White Sands, used to test rocket components, fuels, etc. Here is an aerial view.

Lets now head back north. In eastern Utah, just south of the Wyoming border, is an unusual subdivision of the Rocky Mountains - the Uinta Mountains. This is one of the few East-West-trending ranges in North America. We examine this first in a satellite image, then from an aerial oblique photo:



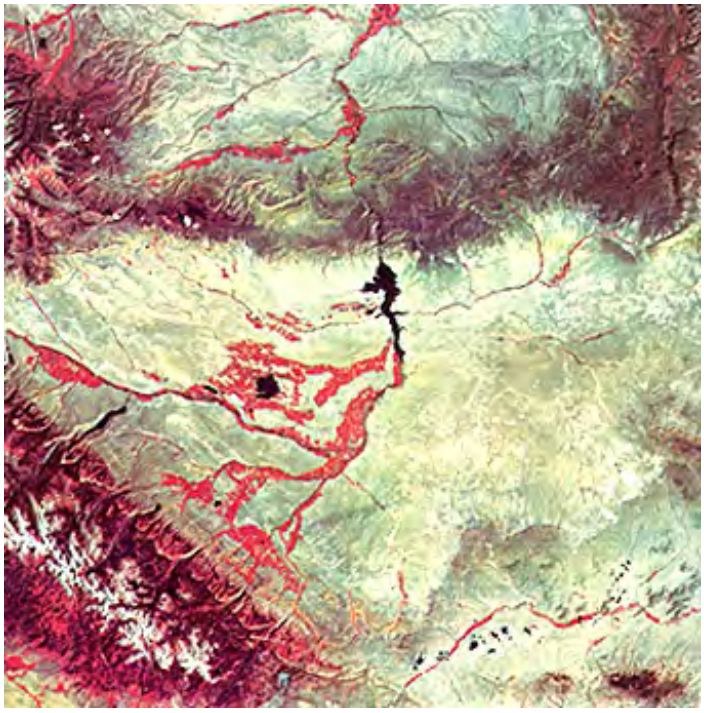
The Middle Rocky Mountains lie mostly in the state of Wyoming. The Rockies there are exposed as a group of separated mountain ranges with deep intermontane basins (filled with erosional sediments from the mountains) in between. This map (from E. Raisz; Landforms of the United States) shows the distribution of these uplifts and basins; read the caption to identify individual ranges:



These self-same ranges appear in the mosaics of Wyoming that we will show on page 7-1.

As an example of the Middle Rocky Mountains, this Landsat-1 scene shows the Wind River

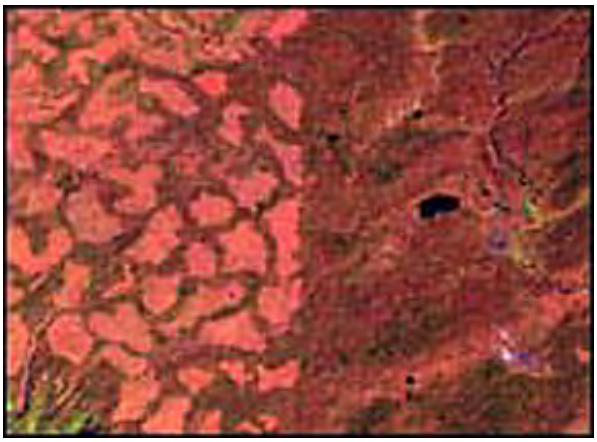
Mountains (left) and the Owl Creek Mountain (near top). The low central area is the Wind River Basin; note Ocean Lake and the Boysen Reservoir:



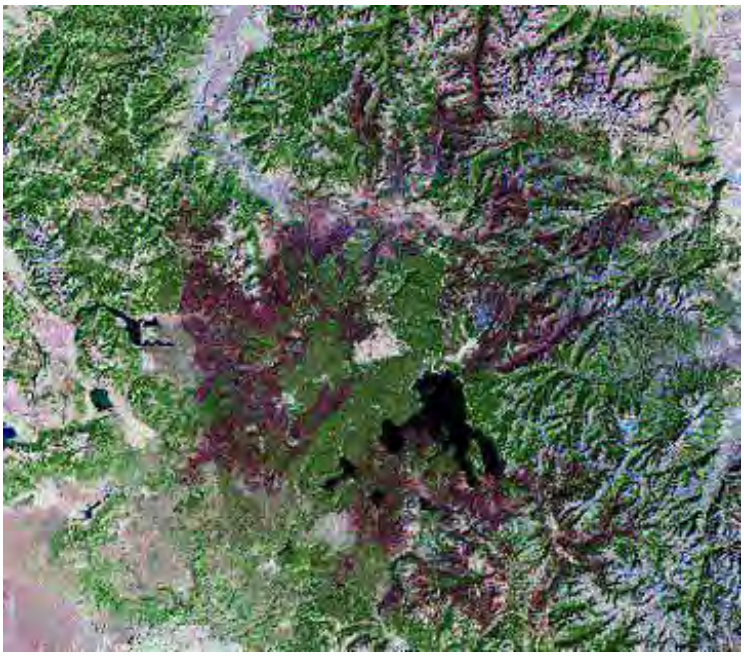
Nestled within the broad stretch of the Middle Rocky Mountains as they pass through northwestern Wyoming is America's first National Park: Yellowstone. This next image shows Yellowstone Park in context with the surrounding mountains:

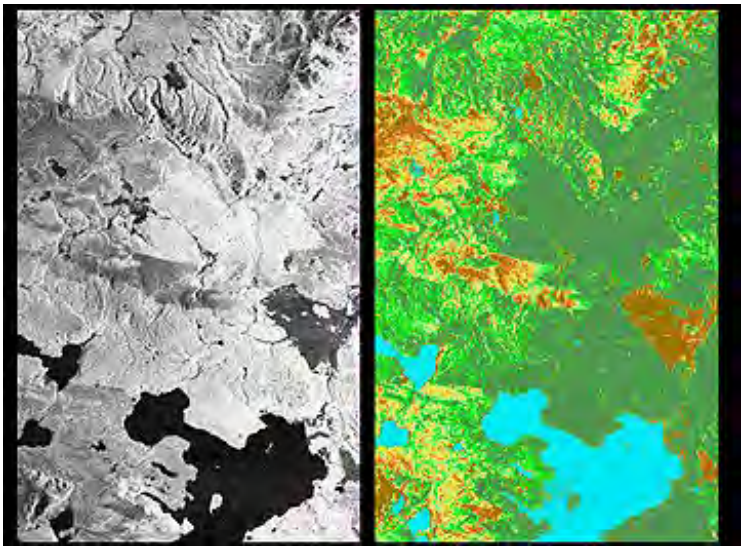
Here is a Landsat view that zeros in on the Park itself; for reference note Yellowstone Lake:

Yellowstone Lake is surrounded by a darker area (evergreens) near the image center. Look in this image about due west (towards the top) of Yellowstone Lake - this is how it appears in a Landsat-7 ETM+ subscene below: grazing and clearcut forests are to the left (west); fully preserved forestlands are on the right; the sharp boundary is related to a fenceline (this effect is similar to the U.S.-Canadian boundary shown on the previous page):

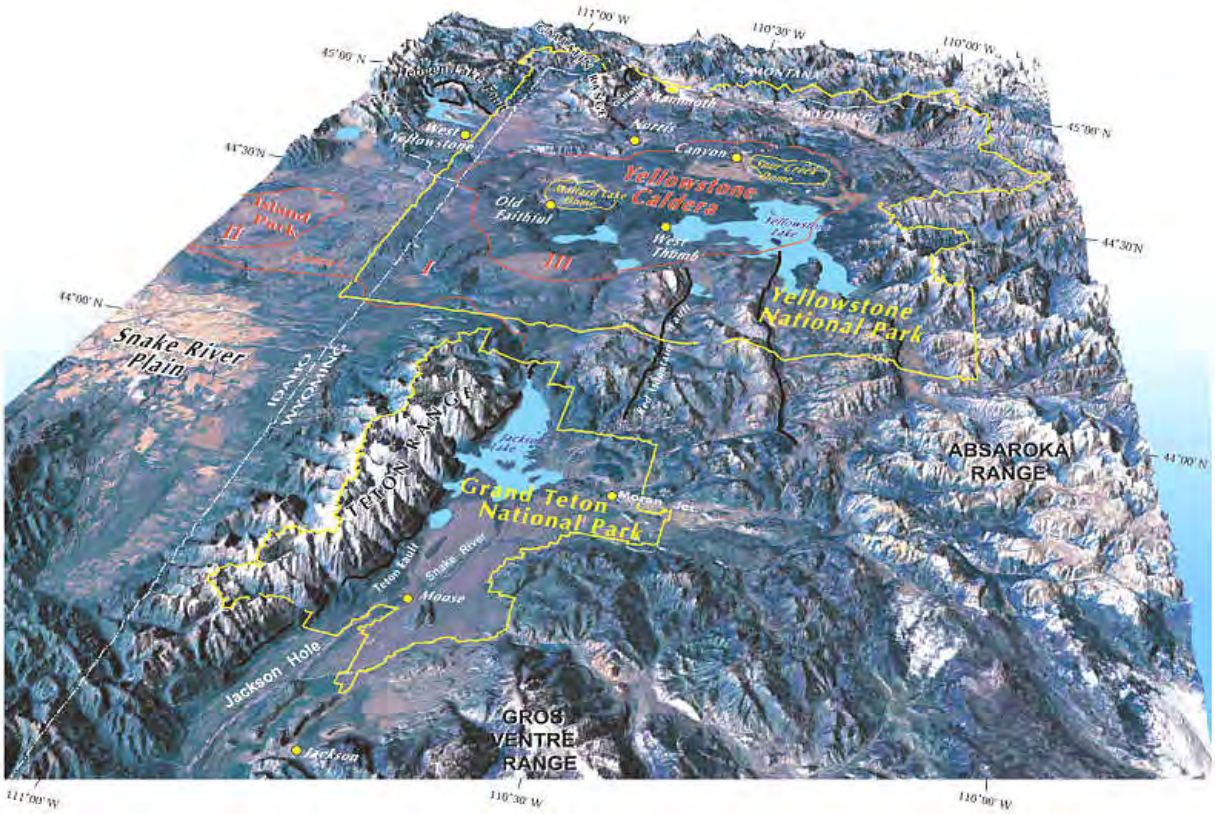


In the mid-1990s, Yellowstone Park and some surrounding areas were plagued with wildfires that extended over huge areas. The next two images - one Landsat, the other SIR-C radar - show some of the burn scars (in dark purple):





As we shall shortly see, Yellowstone owes its special character mainly to recent volcanic activity. This interesting perspective of Yellowstone Park and surrounding mountains was made from satellite imagery.

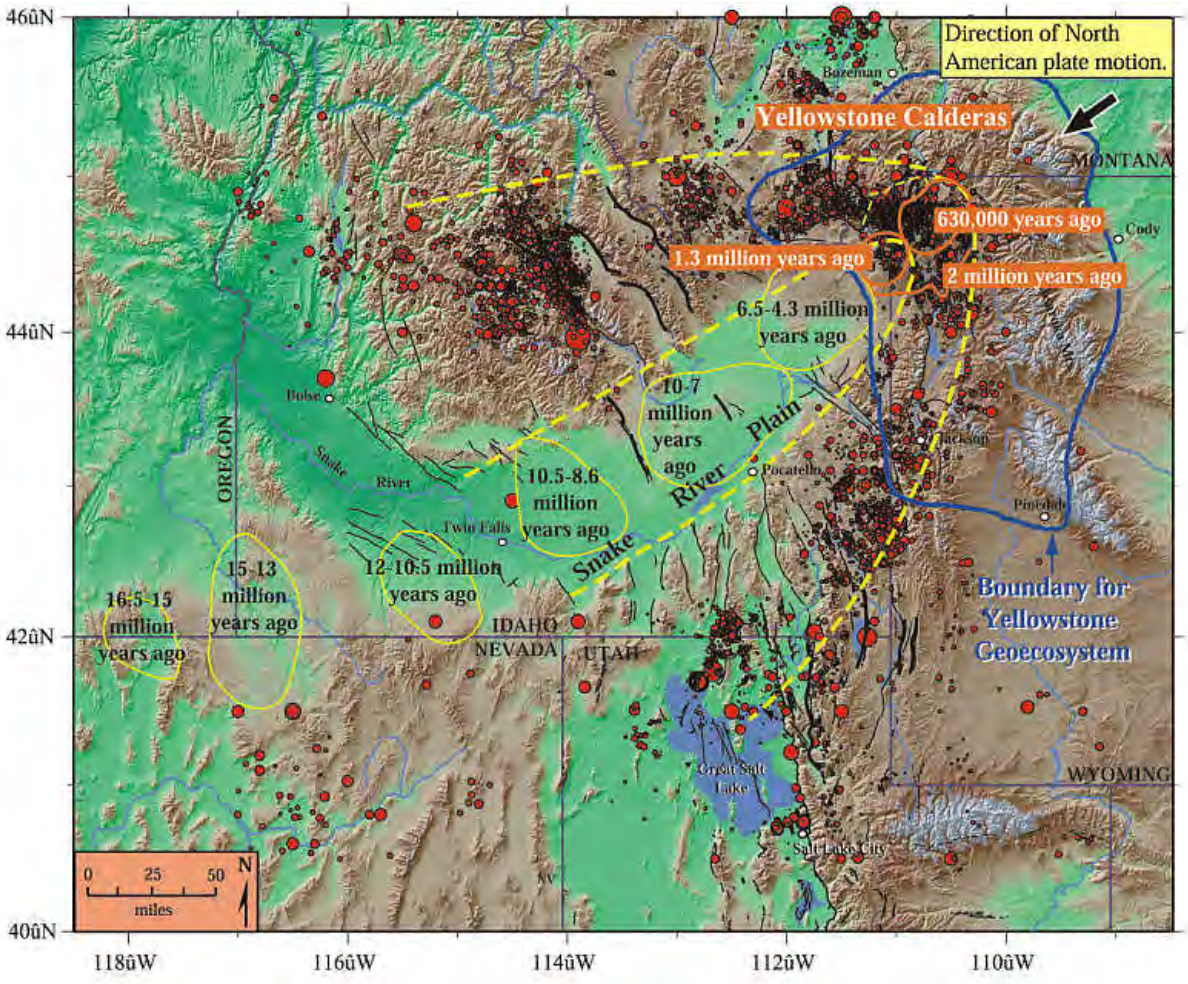


The next two pictures show typical terrain in the Park and Yellowstone Falls (with the yellowish pyroclastic ash deposits that give the park its name):

Of course, Yellowstone Park is most famed because of its geysers (fountains of water spewed out after groundwater has reached its near boiling point after being heated by residual heat from subterranean volcanic activity). Here are the Lion Geyser in eruption and a view of the Norris geyser basin:

So, how do geysers work? Groundwater collects in fissures and pockets below the surface. The water is heated (from deeper magma; see below) to the boiling point. The water expands slightly, relieving the pressure enough to cause some of it to flash into steam. The steam drives the water upward through openings to spill over or erupt vigorously at the surface as hot pools and geysers.

The obvious question you have in mind: What causes the varied phenomena that make Yellowstone almost unique? The answer lies in its special geologic setting. For millions of years this part of the northwest U.S. has been the site of widespread volcanism. This diagram summarizes the relevant activity.

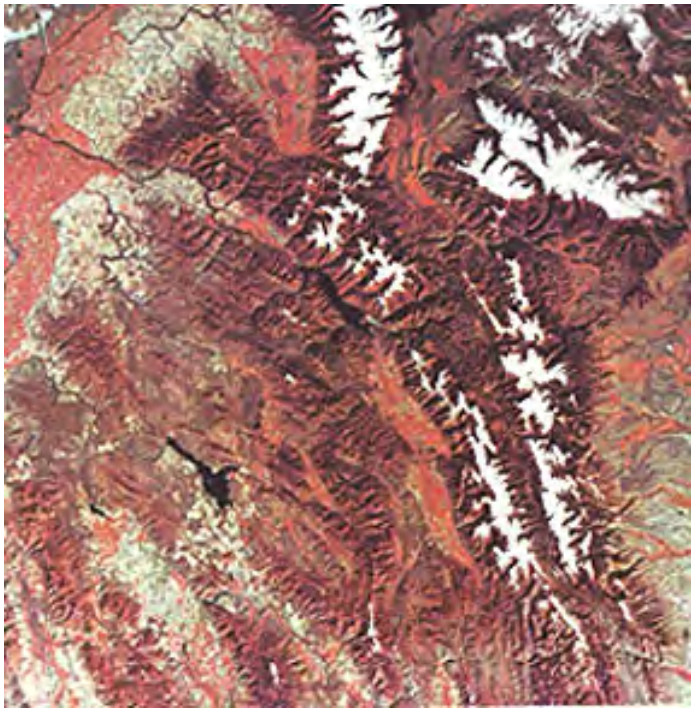


The ultimate cause of the volcanism is a stationary hot spot coming from the Earth's mantle which heats crustal rock to produce large pockets of magma. The North American tectonic plate has been moving to the southwest for millions of years. The path of motion passes over the hot spot which causes this subterranean heating to reach a stage where violent eruption must ensue. The diagram shows a series of previous eruptions progressing back in time along what is today the eastern Snake River Plains of Idaho. The most recent eruption took place within the terrain making up Yellowstone Park.

That eruption took place about 640,000 years ago. Like its predecessors, it was huge (one estimate: 80 times larger than Krakatoa in Indonesia, which was the most violent in the last 500 years). This supervolcano's explosion blew away a large upbent part of the crust (mainly, the volcanic edifice that had been building up) creating a caldera some 48 x 72 km (30 by 45 miles) that then was filled by subsequent lava and ash deposits (so that the present day topography offers little direct evidence of the scar that was left behind). The photo below shows some of the deposits (ash and flows that are dominated by rhyolite volcanics) laid down by the eruption itself and later activity.

The question one naturally poses: If there have been previous major eruptions several times in the last few million years, what can be predicted about the next one? Since there is strong evidence (from earthquake studies) that there is still a large pocket of magma beneath Yellowstone, the likelihood of another super-eruption is fairly high. Just when cannot yet be forecast but it seems probable that it will happen sometime in the next half million years.

The next scene to the south of Yellowstone contains some of the most spectacular scenery in America. See caption for identification of natural features:



The Tetons are the up-down range at top center. Southeast of the Tetons are the E-W Gros Ventre Mountains (those who know French should try to translate these names, given by French trappers in the 18th century). The two curving ranges below are the Wyoming and Hoback ranges. The beginning of the Snake River Plains (page 6-8) appears in the upper left.

The Tetons are dramatically photogenic. Here they are in a panoramic view from the alluvial plains through which the Snake River flows:

This next photo is the most frequent one online when "Grand Tetons" is googled. It is similar to the backdrop that appears in the classic western movie "Shane".

Fit these scenes, which show the beauty of the Tetons close-up, with this perspective view of the entire range, the Snake River Plains, and the Hoback Range to the south as constructed from a Landsat-7 natural color composite and STRM elevation data:

This perspective view discloses the Tetons to be a block fault mountain range in which the east side is thrust upwards as a swinging motion. Thus, the west slope is gently inclined, so much so that bicyclists have reached the crest whereas those climbing from the east side must use ropes and other such advanced scaling equipment.

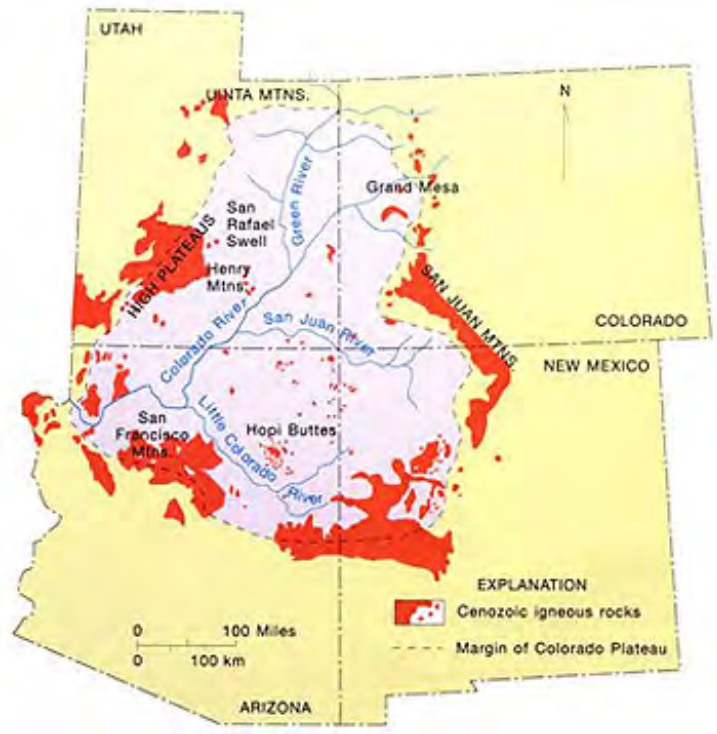
PREVIOUS NEXT

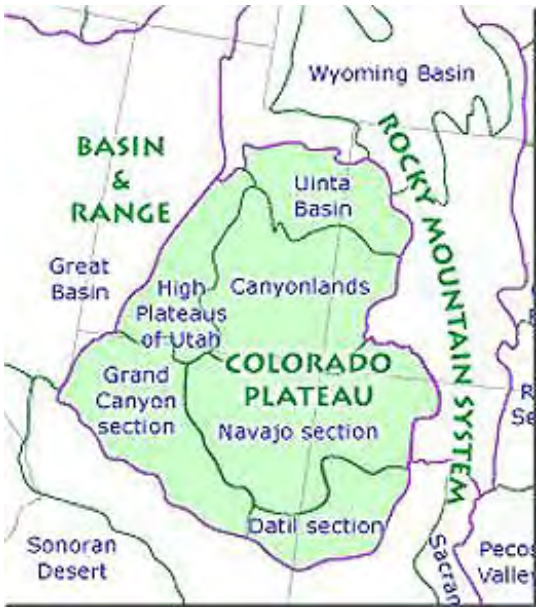
Primary Author: Nicholas M. Short, Sr.

West of the main Rocky Mountain Belt, from northern Utah to northern Arizona, the colorful and varied landscape is controlled by the so-called Colorado Plateau physiographic province. It takes its name from the Colorado River rather than the state. Most rock units are subhorizontal sedimentary rocks but here and there are folds and uplifts that have punctured the plateau. The Plateau has participated in the general deformation of western regions of the U.S. mainly by vertical uplift without folding even as the Basin and Range (see next page) was also uplifted but failed by general faulting of complexly deformed Paleozoic rocks. Satellite images, accompanied by ground photos, give the "flavor" of the flamboyantly scenic regions that comprise the Colorado Plateau.

The Colorado Plateau: Colorado, New Mexico, Arizona and Utah

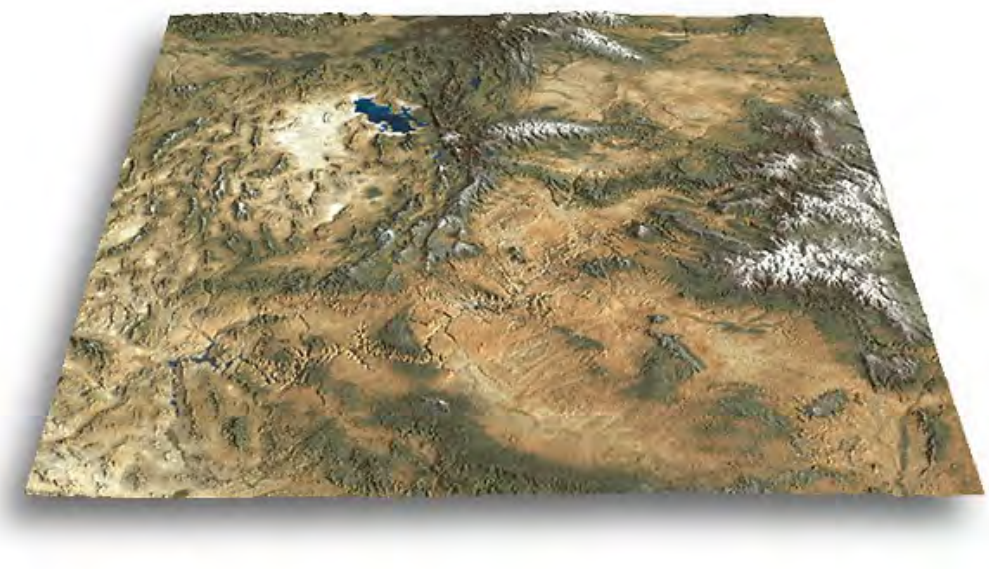
After about 500 km (311 mi) travel westward, our trip swings to the southwest as it moves over the Colorado Plateau. The Plateau includes part of western Colorado and northwestern New Mexico, much of northern Arizona and a substantial part of Utah. Its principal geographic locations are shown in the first map below and its physiographic subdivisions in the second map map:





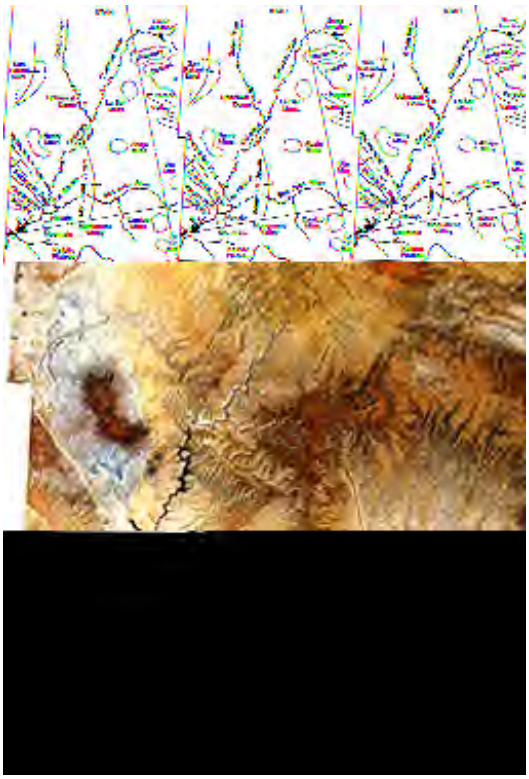
The Colorado Plateau has participated in the general uplift of the interior western U.S. since the Cretaceous. Unlike its eastern neighbor, the southern Rocky Mountains, this segment of the continental crust was thermally heated and subjected to vertical uplift without extensive folding (there are a few small folded warps and some faulting). Thus the "trademark" of the Plateau is the dominance of flat-lying (near horizontal) sedimentary rocks, largely exposed and often without much vegetation. Many of these rock units, ranging from Paleozoic to Cenozoic in age, are colored in shades of red, orange, yellow, and brown. The vistas of bright-colored layers means that the Colorado Plateau probably contains more photogenic scenery than any other part of North America.

This perspective drawing, aided by satellite imagery, gives a good overview of the Colorado Plateau, as well as the Basin and Range to the west:



The Plateau is conspicuous to astronauts in orbit because much of its terrain is brightly colored in reds and browns (north is to the right):

Much of the Plateau is seen in this mosaic of several Landsat images:

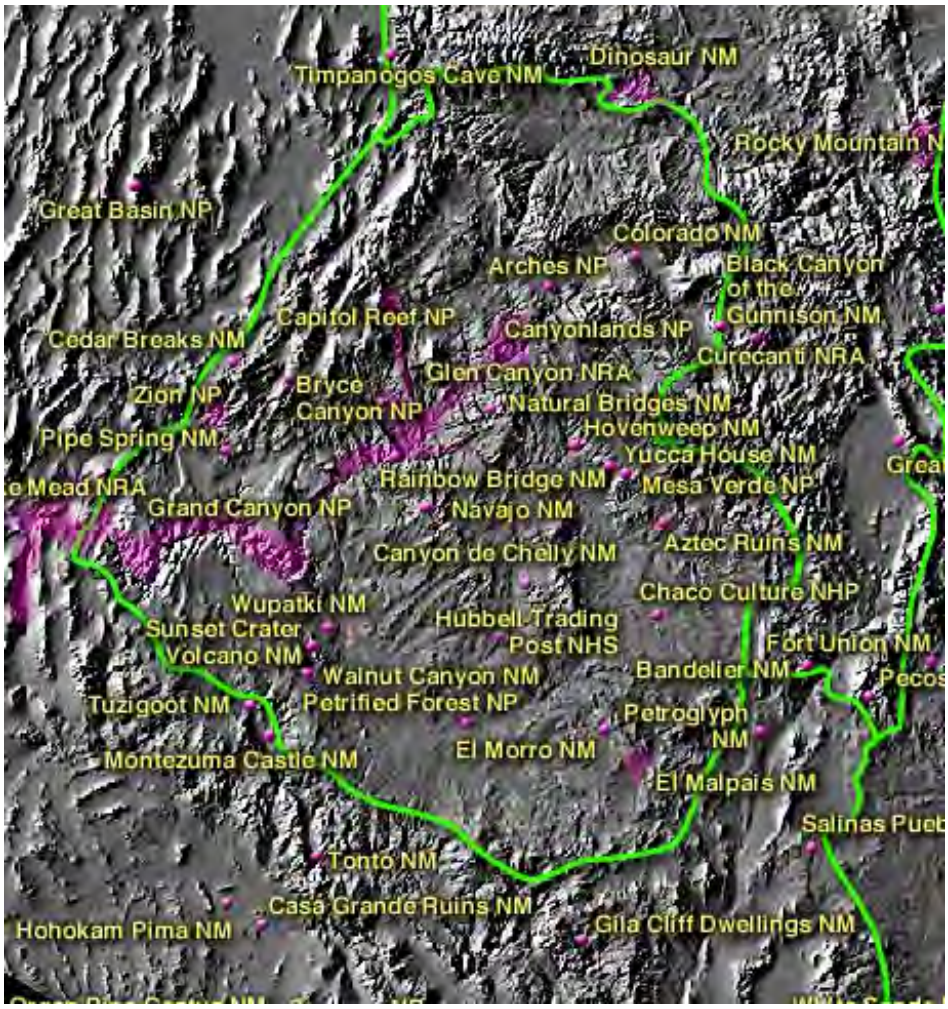


But it helps to examine the Plateau in the context of surrounding provinces. We saw much of the western United States in a HCMM image on the previous page. Now look at this even larger coverage found in a MODIS image:

The central area with reds and browns make up the Colorado Plateau. Most of the Southern

Rocky Mountains appears to the east. Part of the next province we will visit - the Basin and Range is displayed in much of Utah and Southern Arizona. The top of the image shows much of southern Wyoming which is commonly assigned to the Central Rocky Mountains. In the northwest corner of Utah. The large white patch in Utah's northwest (see state outline in black) is the Great Salt Lake Desert, whose saline deposits form a smooth flat surface on which supercharged racing cars have broken international speed records. One surprise: the Grand Canyon (see below) is almost "invisible".

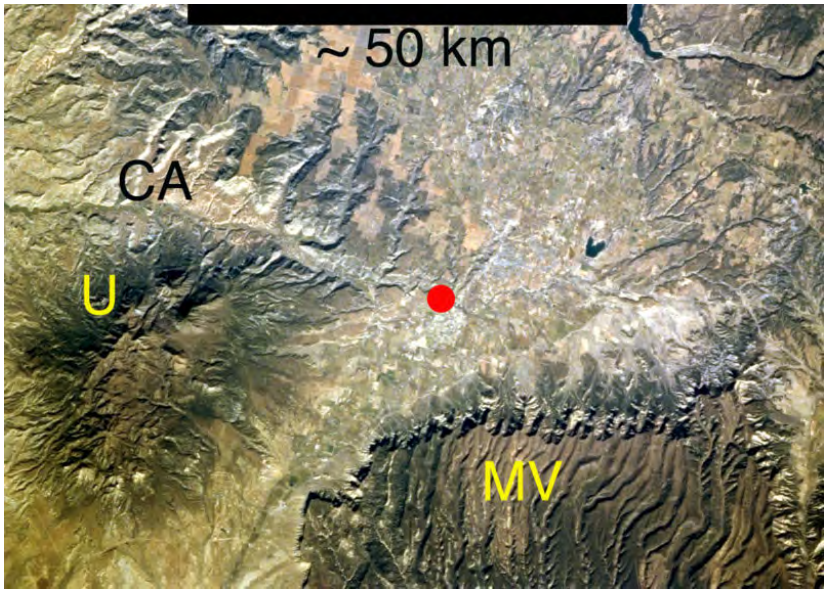
Many people consider the Plateau the most scenic of all provinces in the U.S. because of its marvelous landforms and its colorful rocks. Some of its mainly Upper Paleozoic and Mesozoic rock units are bright reds, oranges, and yellows, whereas others are light to dark gray to brown. Because of their scenic value, these rocks are a major reason for the establishment of numerous national and state parks in the Colorado Plateau:



If one reaches the Plateau by traveling west along Interstate 70, the first large town encountered is Grand Junction, Colorado, near the Colorado National Monument. This town now has about 50,000 residents and dominates the Grand Valley:

Grand Junction lies in a valley north of the Uncompaghre Plateau that is an uplift of Mesozoic flat rocks overlying a segment of the Ancestral Rocky Mountains. Here is a space image; its top part is the east end of the Tavaputs Plateau that exposes the Book Cliffs at its south end:

In southwestern Colorado, within the Plateau, is the Mesa Verde National Monument. Here is the region from space. Beneath it is a photo of Indian cliff dwellings for which the Monument is famed:

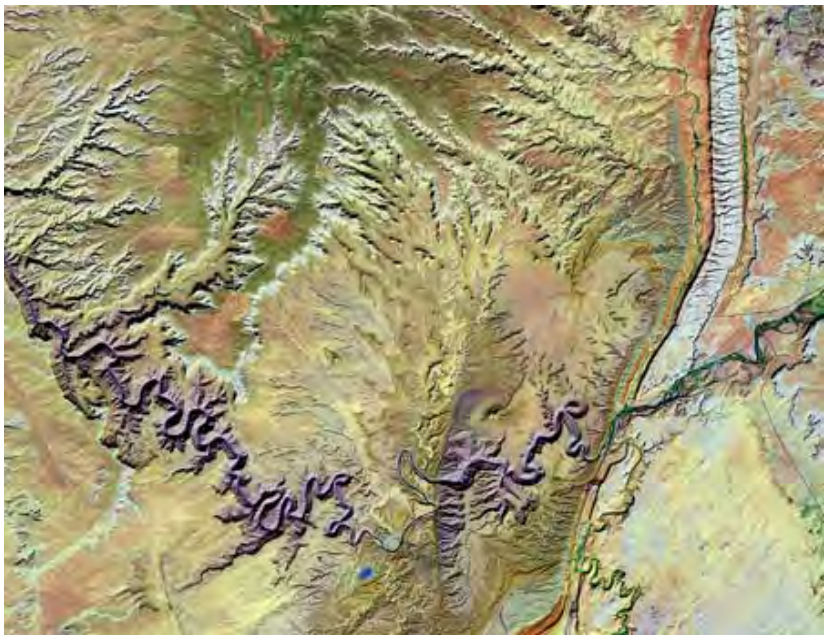


We look next at what is called the Four Corners area of the Plateau. There is a plain benchmark at the exact spot where the four states - Colorado, Utah, Arizona, New Mexico - touch. If you straddle this marker in a supine pose, you will truly have your body in four states simultaneously.

The Landsat scene shown below includes the Four Corners - the only place in the U.S. in which that many states touch each other at one point. Unfortunately, we can't discretely identify the point, but it lies about 30% up and 15% in from the lower right corner of the image.

This false color image, taken in January, approximates some of the colors seen on the ground. At this time of year the sparse vegetative cover of sage, mesquite, and grasses does not produce a typical red signature, so that the surface tones are entirely those of rock and soil. Although not discernible in the image, the Plateau rocks are almost everywhere still in the subhorizontal positions, in which they were deposited as sediments. Because they erode along steep faces or scarps, where cap rock is hard, the layers stack like steps to form plateaus, mesas, and buttes. The Gothic Mesas, just to the right of image center, are typical. Monument Valley begins near the lower left corner. The Plateau has participated in the general deformation of the West chiefly by uplifting without folding. However, Combs Ridge, a prominent monocline (like the Waterpocket Fold) is evident about 15-20% in from the left edge of the image. Near the bottom right corner are the snow-capped Carrizo Mountains, partially volcanic in origin, which rises at Pastora Peak to 2,870 meters (9,414 ft). This barren region has a very low population. Part of the above scene includes the Navajo Indian Reservation. The small towns of Mexican Hat and Bluff in Utah lie along the San Juan River. There is a "streakiness" in much of the lower part of the image. Prevailing winds, re-enforced by joint (fractures) control of landscape erosion, produce this effect.

Another Landsat image covers the region to the west of the previous image. Locate Comb Ridge. To its west are the entrenched meanders of the San Juan River. These have been named the Goosenecks.



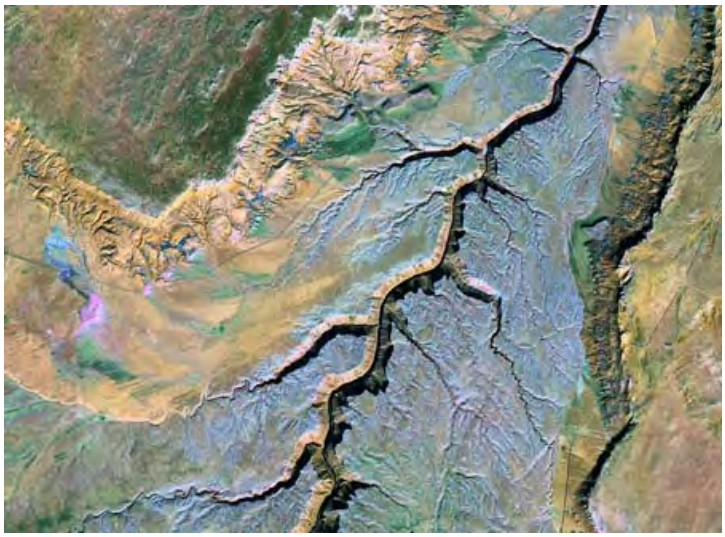
This IKONOS image concentrates on the Goosenecks, also seen in a low altitude oblique aerial photo:

6-10: What causes "goosenecks"? ANSWER

In Section 2, page 2-3, you have already examined in some detail one small area in the Plateau, the Waterpocket Fold in the Capitol Reef National Monument. That fold is a monocline (only one limb inclined); such structures are scattered over the Colorado Plateau, often making up a landform called a hogback, such as pictured here:

Just to the west of the image, the San Juan River joins the Colorado River, upstream from the artificial Lake Powell. This, the second largest of its kind in the U.S., results from flooding of the River behind the Glen Canyon Dam, next to Page, Arizona. The area has become a popular tourist destination, with tours by boat around the lake (one can rent a movable houseboat also) This is Lake Powell, seen from Landsat and from on the lake itself.

Downstream from Page, the Colorado River widens its valley and becomes entrenched. This valley, known as Marble Canyon, is a popular tourist attraction. Here it is shown from space and in an aerial photo:

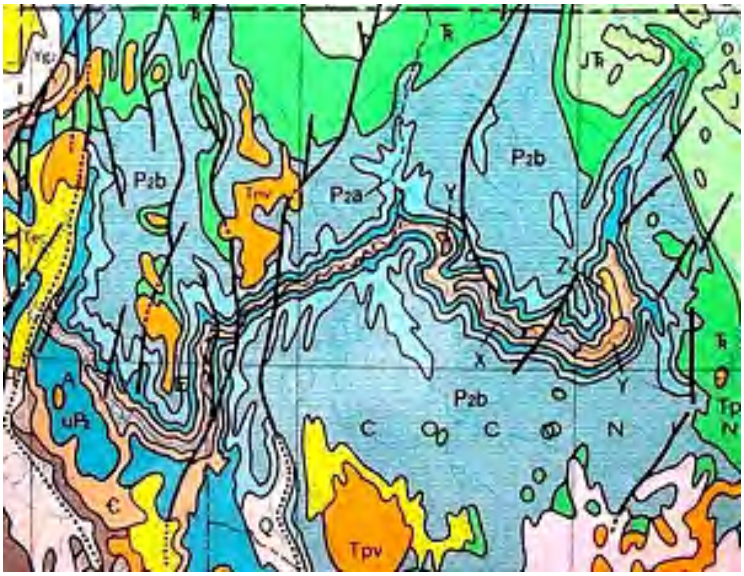


To the southwest of this scene, the Grand Canyon, the most famous feature in the Plateau, exposes typical, multi-colored units. Being several hundred kilometers long, up to 30 km (18 miles) wide, and as deep as 1.6 km (1 mile) deep, this is the largest canyon in terms of volume of excavation (which was accelerated by runoff water from Pleistocene rains and snowfalls) in the world. Yet it is a recent phenomenon, having been cut into the Plateau as the land rose only in the last 6 million years. Its colors, from Paleozoic rocks, makes it a spectacular view that brings millions of tourists per year.

A false color Landsat subscene places this mile-deep gouge into context with the Colorado plateau. The eastern part of the Grand Canyon is well displayed in this Landsat-1 image:

Considering the variety of layered rocks representing 1 Mesozoic Formation and 14 Paleozoic formations, it is surprising that these do not appear as thin bands in the Landsat images and especially the JERS image. Here is an IKONOS image at even higher resolution and again layering - the hallmark of the Grand Canyon as the best exposed stratigraphic sequence in the world - is totally absent (but the pinkish color may be contributed by some units):

A geologic map of the Grand Canyon brings out the effects of layers of different ages, since the map uses different colors to differentiate major stratigraphic intervals:



So, why does space imagery fail to distinguish these layers (the Formations involved are discussed below)? Perhaps this photo provides an explanation:

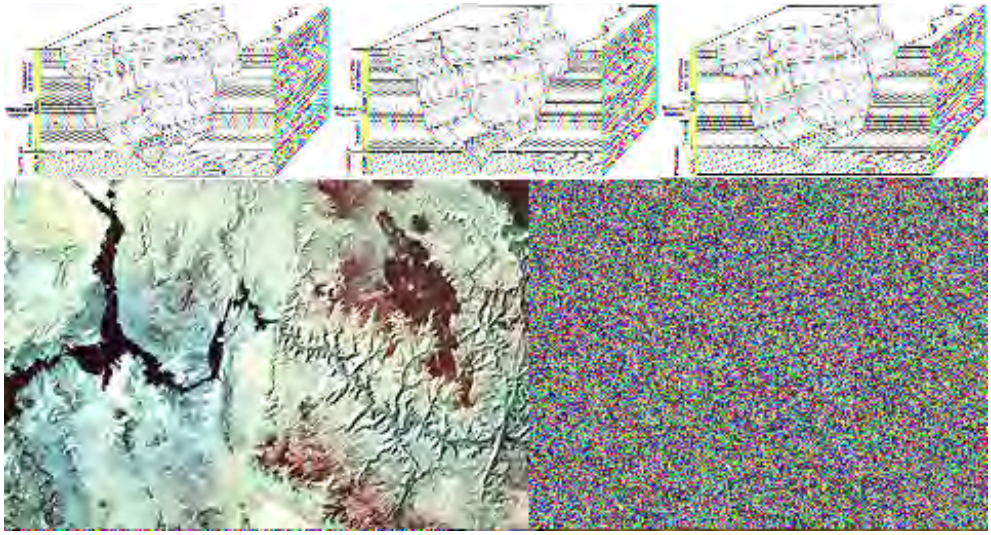
One is struck immediately by the almost uniform reds that seem to color the rocks from the Kaibab Plateau to the gorge through which the Colorado River flows. Yet, in hand specimens, the various units have a range of colors. What seems to have happened is that over the eons, the exposed cliffs and benched slopes have acquired iron staining that imposes uniformity of surface coloration. Often though, standing on the rim, at least some of the various formations can be distinguished by eyeballing the colors. But, the surfaces of these formations do have less color variation than, say, the Waterpocket Fold we examined in Section 2.

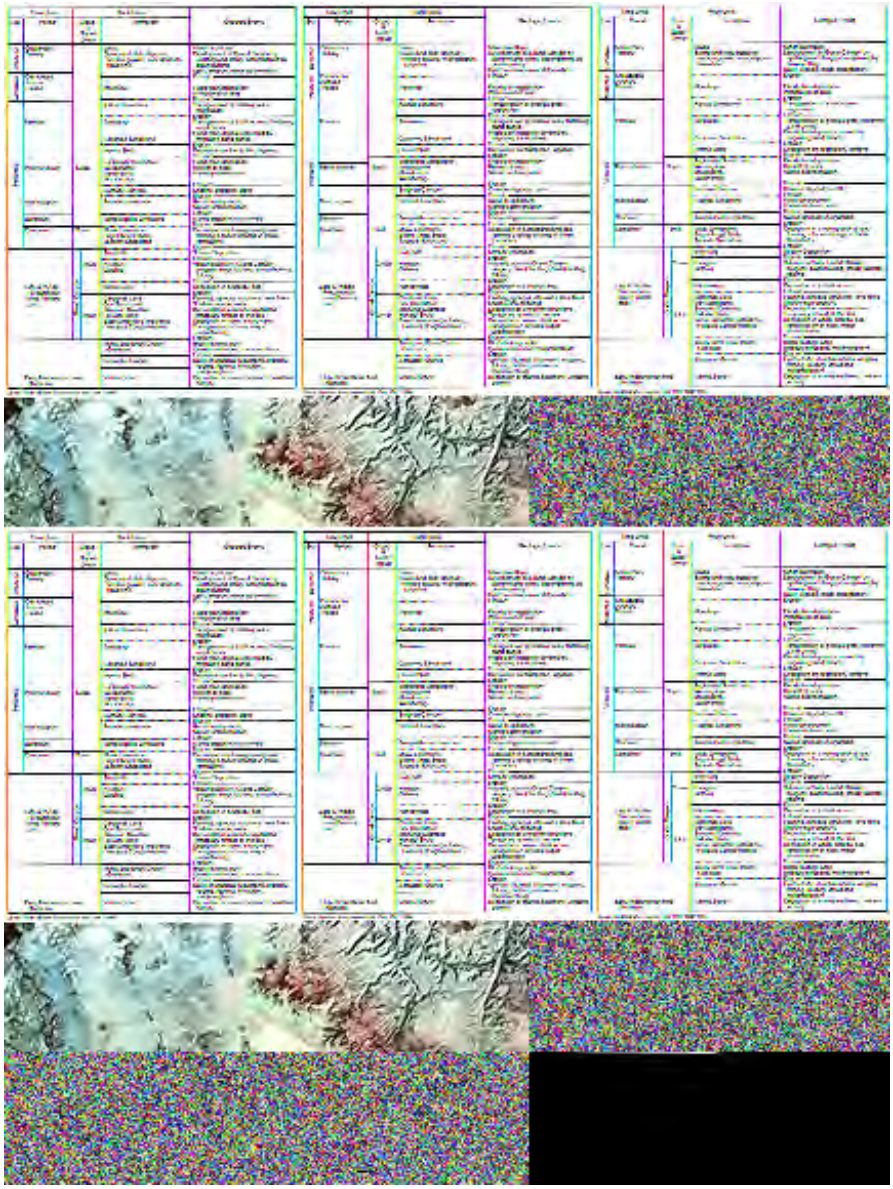
The bottom of the Grand Canyon is a common destination for the intrepid, who either walk down and up some 5000 ft vertically along a trail carved into the canyonwalls or make the trip on the back of a sure-footed burro. The floor of the canyon is generally narrow but, while very hot in summer, has been inhabited by Indian tribes for centuries before its discovery by explorers and its first passage by boat by Wesley Powell, the geologist-explorer in the 1880s. This next image is a high resolution image made by the IKONOS satellite of the darker Precambrian schists and gneisses that rise above the floor of the Grand Canyon.

The horizontal Paleozoic sedimentary rocks rest unconformably on tilted Proterozoic metasedimentary rocks that in turn lie on Archaean rocks dominated by the Vishnu schists. These are exposed in the lower part of the canyon walls seen in this ground photo:

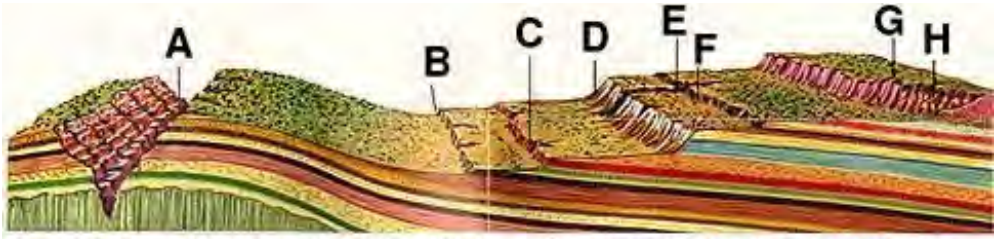
The Grand Canyon continues westward towards southern California but ends just east of Nevada. As it crosses that state, it has been dammed south of Las Vegas (see next page) at Hoover Dam (also called Boulder Dam), behind which Lake Mead has developed. All this including the western Grand Canyon is displayed in this Landsat-1 image

The geologic formations at the Grand Canyon are considered the best exposed type section (a section denotes the sequence of successive age units going upwards from oldest to youngest) in North America. First note this sketch with the main units on the right side of the block diagram. The second figure gives details about the formations so exposed.



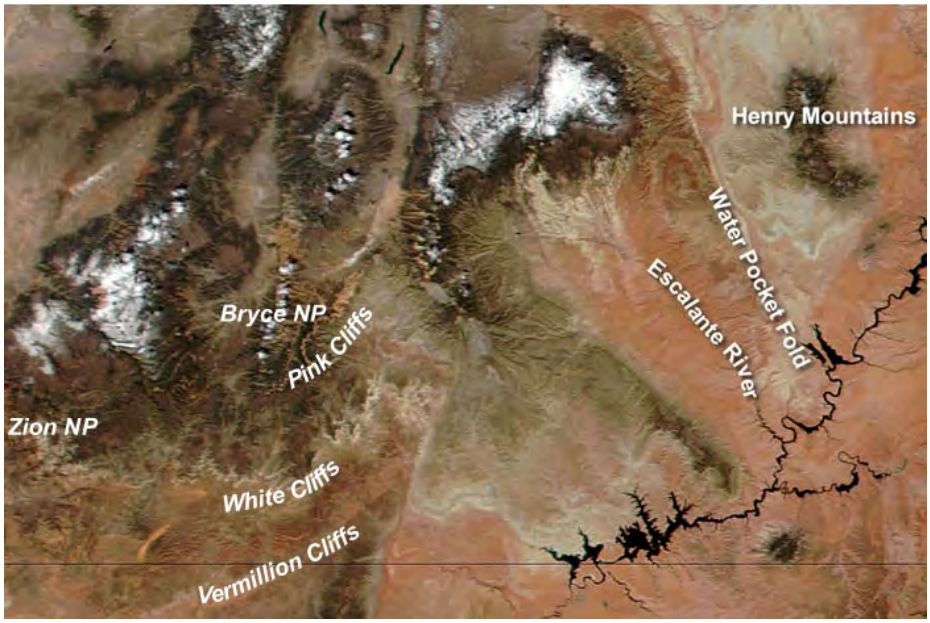


Most of these units can be found elsewhere on the Plateau. They are often distinctly colored, making them into "must see" places that attract tourists - some to National Parks. The rocks of the Plateau tend to tilt northward from the Grand Canyon. Most of these are Mesozoic sandstones and shales. This diagram indicates the general structure of this part of the Plateau:



Grand Canyon (A), Chocolate Cliffs (B), Vermilion Cliffs (C), White Cliffs (D), Zion Canyon (E), Gray Cliffs (F), Pink Cliffs (G), Bryce Canyon (H)

As seen from space:

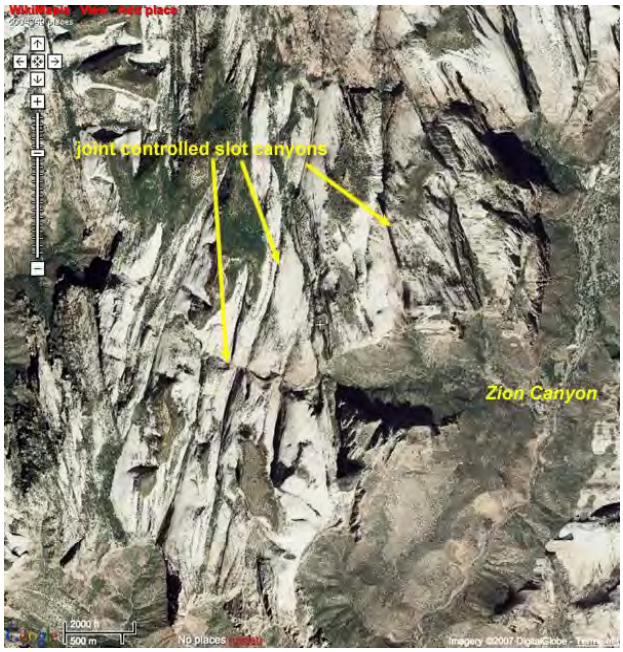


The National Parks at Bryce, Zion (in southwest Utah, respectively east and southeast of Cedar City), and Canyonlands (in southeast Utah west of Moab) also display spectacular colored rocks. Here are examples from space and on the ground; check the caption for identity of Park.

The soft claystones and sandstones making up the Bryce outcrops, carved into marvelous figurines by water erosion, have distinctive red and yellow colors. These are difficult to capture photographically but this is a good example.

Zion National Park is well known for the sandstone units that show distinct cross-bedding.

The rocks in the above satellite image show prominent joints - in fact, better termed megajoints. This higher resolution satellite image shows these in detail.



Here is the valley that exposes these units:

The cross-bedding within the Cretaceous Navajo Formation is evident in these photos:

And the Canyonlands Park:

Similar to Canyonlands is Canyon de Chelly, in Arizona, which is noted for its Indian cliff dwellings.

In Utah is another member of the National Park system, the Arches National Monument:

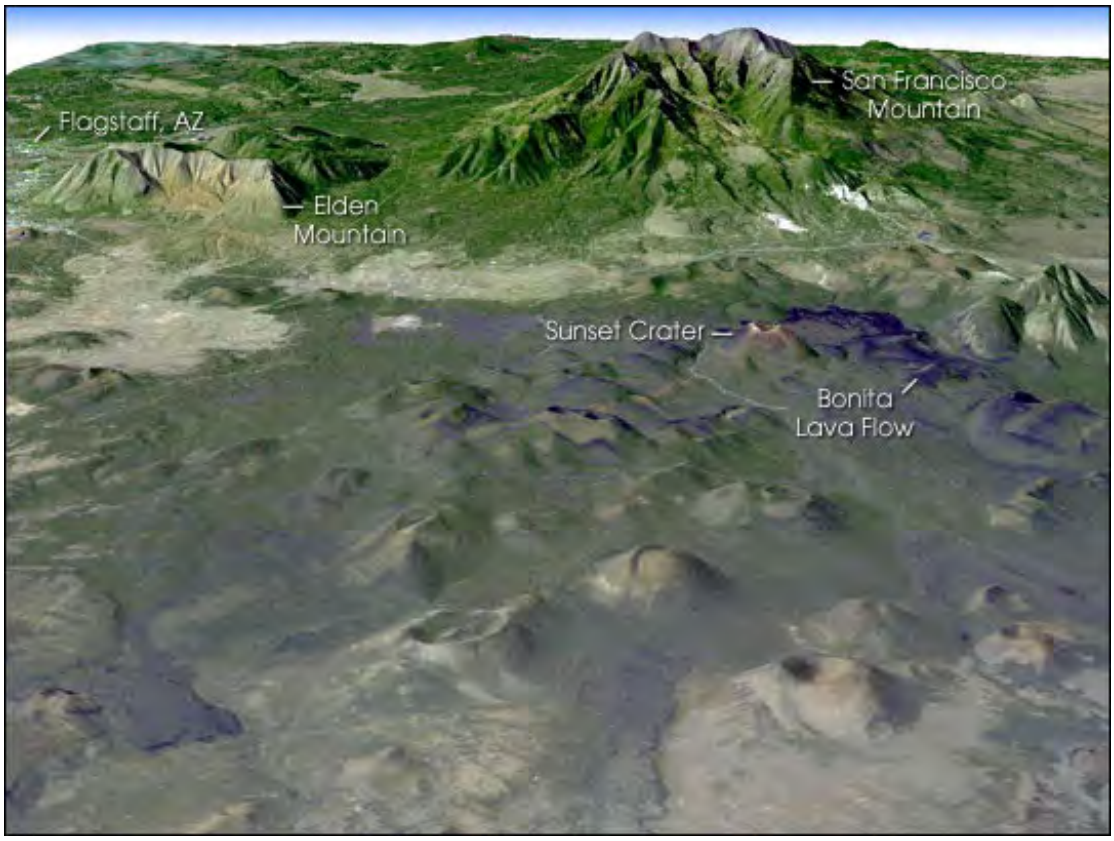
Monument Valley in southern Utah, almost a trademark for that part of the country and site of many western ("cowboy") movies (it was John Wayne's favorite locale), is a landscape dominated by mesas and buttes (both defined as prominences composed of flat rock stacks that are the topographic remnants of stripping away of most of the higher layers from an earlier

plateau cover). Here is a small part of Monument Valley as seen from space, a ground photo of a butte and its neighbors, and a panorami photo that shows a cluster of buttes.

Moving now to the south, here is a satellite image that shows recent volcanism in northern Arizona. This scene contains cinder cones within the San Francisco Volcanic Field.

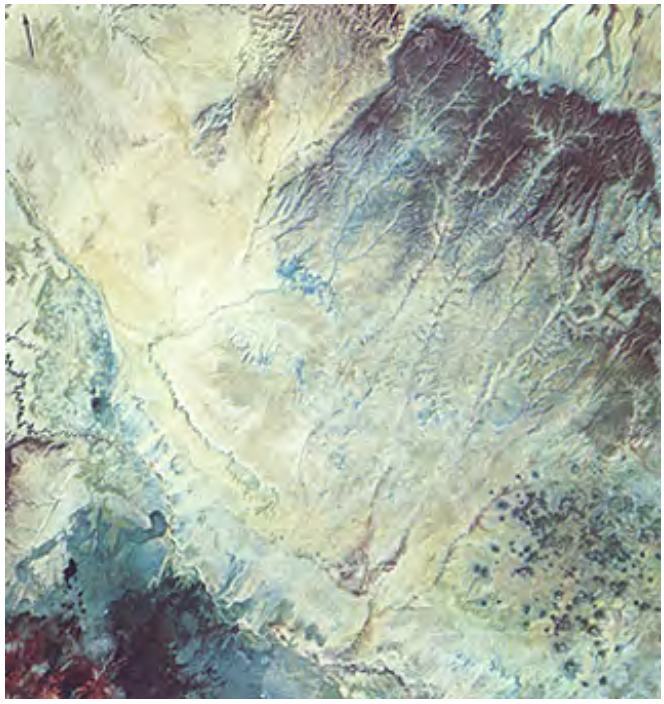
Some of these cinder cones are young, with the last eruption being witnessed by local Indians in the 11th century AD. Two of the best known cones are Sunset Crater and SP crater, shown here;

The field can be pictured as a perspective view, made from an ASTER image

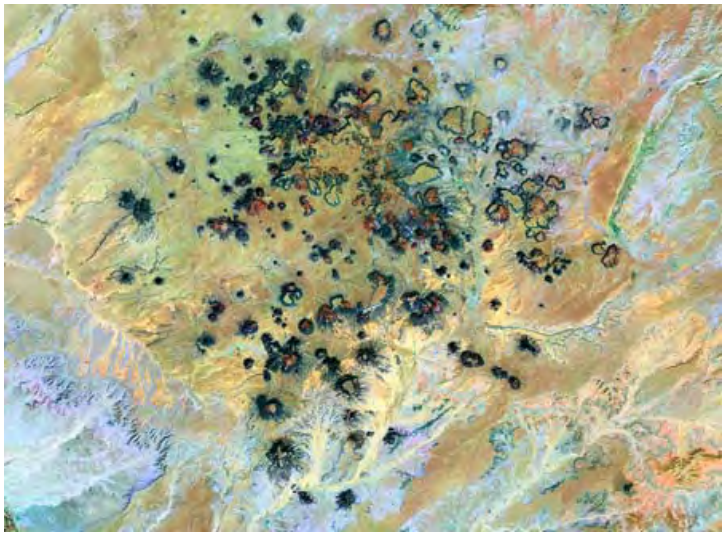


The San Francisco Peaks are actually remnants of an older stratocone complex, now breached by erosion. The highest of these is Mount Humphreys, rising over 12000 feet.

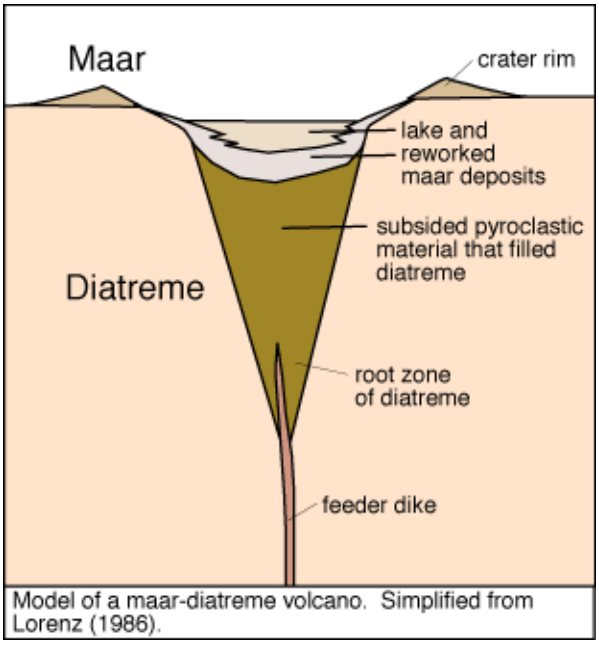
In the vast region of northeast Arizona and a bit of Utah that makes up the Navajo (Indian) Reservation are the Hopi Buttes shown in the lower right corner of this Landsat image:



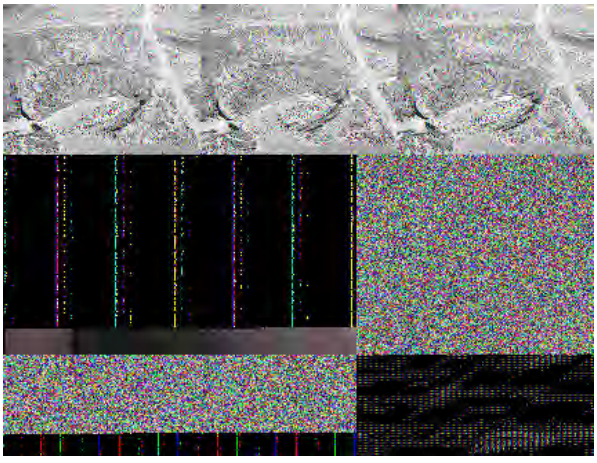
This space image settles on the Hopi Buttes, one of which is seen in the ground photo below it.



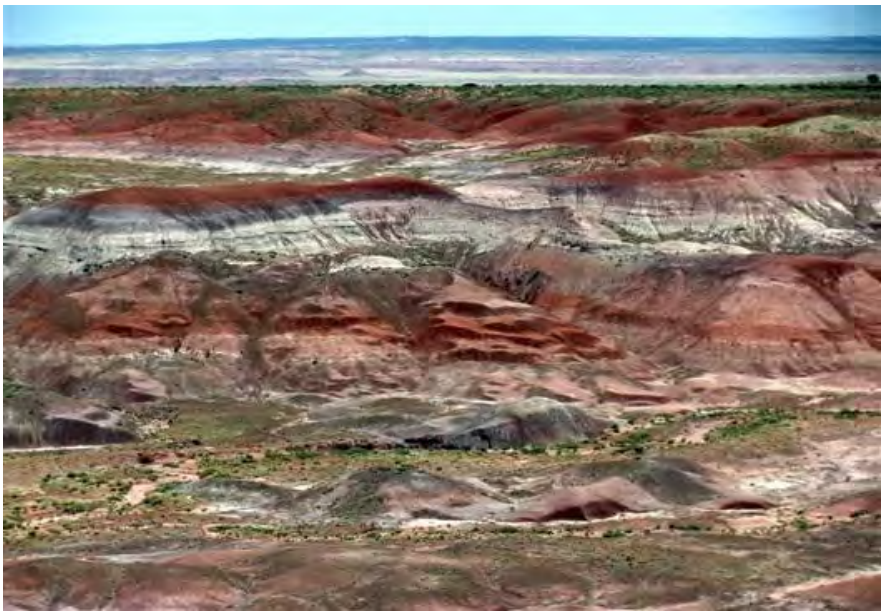
The Hopi Buttes are actually volcanic necks and diatremes that have been exposed as pinnacles after erosion has removed the surrounding rocks. A diatreme usually has a maar at its surface. The maar is an explosion crater formed when volcanic material interacts with shallow ground water which produces steam that blows out the rock. Here is a schematic of a diatreme/maar complex and one of the maars at Hopi Buttes:



Model of a maar-diatreme volcano. Simplified from Lorenz (1986).



The Landsat scene also contains the Painted Desert, noted for its colorful sedimentary rock layers:



The writer's (NMS) favorite spot in all of the western U.S. is along the Mogollon Rim (a popular locale for Zane Grey western novels) which makes up the southern edge of the Plateau. The Rim's face is a nearly continuous cliff where erosion is cutting into the Plateau's underlying sedimentary layers (mostly Mesozoic Formations). Seen below is Sedona, Arizona (about 40 miles south of Flagstaff), a huge tourist attraction and often used as a backdrop in TV commercials. Junipers are the dominant tree. This is a view looking west over the town:

This high altitude aerial photo shows the landscape around Sedona.

And here is Sedona as seen by Quickbird (Google Earth):

This is one 'magical' place on Earth that remains best seen on the ground. No space image can do it full justice.

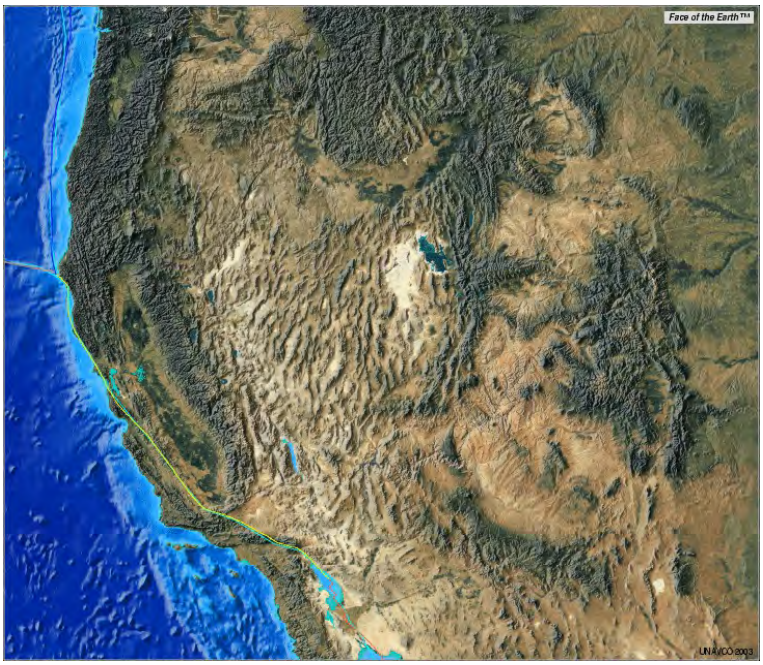
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

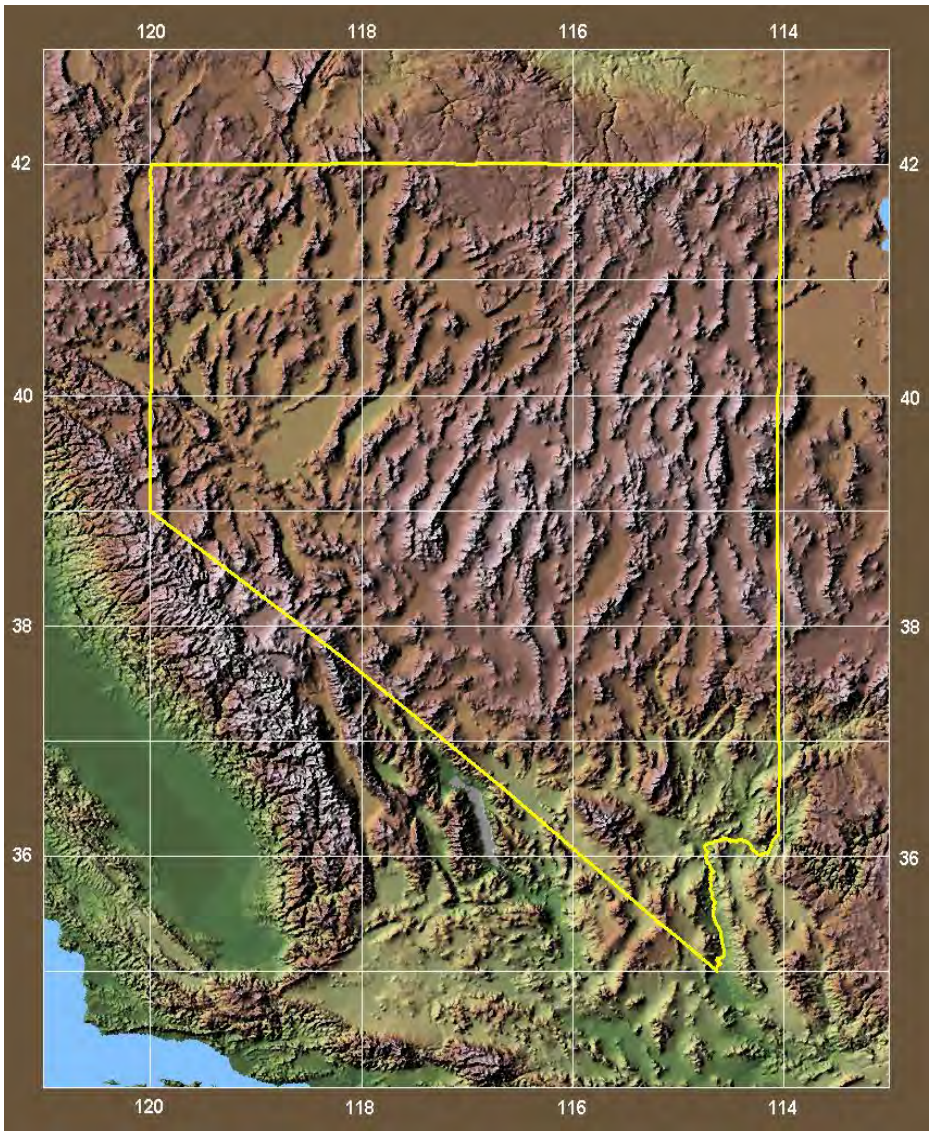
A large region of the west from southeastern Oregon and part of Idaho, through nearly all of Nevada, western Utah into most of southern Arizona and eastern California is known as the Basin and Range province. This province, dominated by uplifted blocks of sedimentary and igneous-metamorphic rocks set between broad, flat valleys, swings eastward through New Mexico into West Texas, and continues southward into Mexico. The Columbia Plateau and the northern Rocky Mountains also are diverse. Typical examples of the varied landscapes are shown on this page.

The Nevada Basin and Range; The Desert Southwest; Northern Rocky Mountains; The Snake River Plains and Columbia Plateau

The geology, physiography, and ecology of the western United States is quite varied. Look at this shaded relief map that shows most of the region. One feature should catch your eye almost at once: the many stand-alone elongated mountain systems trending north-south that cover a wide area centered on Nevada. This is the Basin and Range Province, which owes its character to differential up-down movements of the Earth's crust along normal faults as the region has been uplifting in response to subduction brought about by the westward movement of North America:



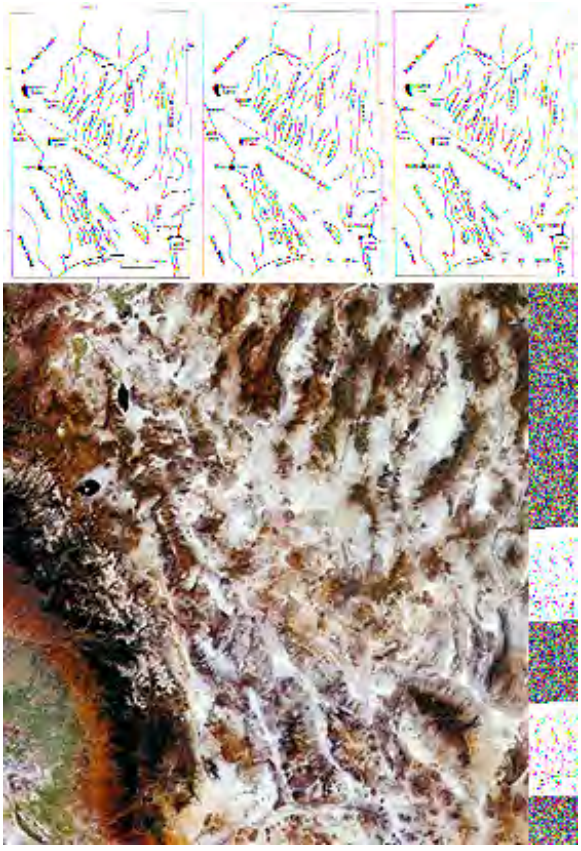
Compare the above map with the color-shaded relief map of most of Nevada.



The bulk of the Basin and Range Province is within Nevada and small parts of California and Colorado. But as this subdivisions map shows, the characteristic block faulting extends into Arizona and New Mexico, and a bit into Mexico itself:

This Landsat image mosaic shows much of the Basin and Range province in most of Nevada but

includes Death Valley, the southern Sierra Nevada block, and a bit of the Great Valley (San Joaquin) near Bakersfield, are in California; see map below the mosaic.



Major C.E. Dutton, an early explorer of the American West, described these mountains as they appeared on a map as resembling "an army of caterpillars crawling northward out of Mexico". That is re-enforced by the next image, a shaded topographic depiction of the region made from DEM data (method described on page 11-5; another example is on page 7-2). This rendition emphasizes the notable flatness of the valley floors in the basins.

This aerial oblique photo brings out the strong parallelism among neighboring mountain ranges in Nevada.

Such characteristic topography in the Basin and Range results from a complex structural history dominated by block faulting (somewhat like the faulting we described in Kenya on page 3-2). The region, as it underwent tensional stresses during uplift of the crust while the Sierra Nevada mountain block to the west and the Southern Rocky Mountains to the east were forming, responded by fracturing. The fractures trend mostly north-south, along which segments dropped down (faulted), leaving adjacent range blocks higher. The present elevations can be greater than 2,700 m (8,900 ft), producing relief of 900 to 1,520 m (3000 to 5000 ft) relative to the basins. Among the major ranges in this scene are the Shoshone, Toiyabe, Toquima, Monitor, Ruby, and Hot Creek Mountains; higher slopes are forested, as suggested by the reds in this September scene.

Although most of these mountains are not high (moderate relief), they can appear imposing as they are approached. Here is the Toiyabe Range as seen from a basin to its east:

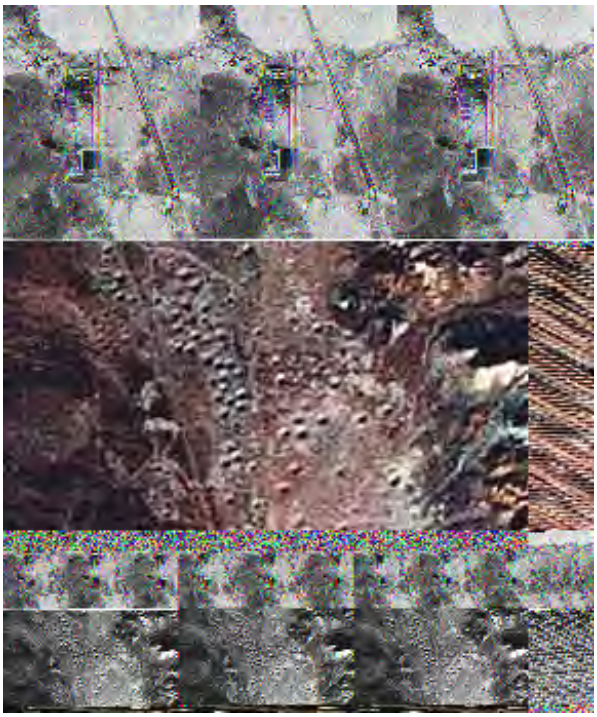
The intermontane basins between ranges are back-filled with great amounts of rock debris descending downslope, so that the valley floors move upwards as the ranges wear down. The central (lowest) surfaces of some of the basins contain playas (deposits of fine sediment left after intermittent lakes evaporate following the rainy season) that are light-toned in the image. Between playas and ranges, along the transitional zone known as a piedmont (literally, "foot of the mountain"), are deposits of coarser sediment (up to cobbles and boulders in size), mixed with clays that make up alluvial fans (dark gray in the image).

The ranges can be quite distinctive, standing out between valleys as bare rock, since many are almost devoid of arboreal vegetation. This scene is in Nevada:

Nevada has experienced relatively recent volcanism. Parts are covered by thick beds of volcanic ash. Basaltic lavas also extrude from vents. This Landsat image shows the Timber Mountain caldera (lower right), another caldera atop a mountain (upper right), and several lava flows:

Most of Nevada is "owned" by the Federal government. That was one reason why a large tract covering hundreds of thousands of acres was selected to develop the Nevada Test Site where first surface and then underground nuclear testing was carried out for several decades (usually in volcanic ash deposits). This "secret" base was entered from Camp Mercury about 75 miles north of Las Vegas (in a recent relaxation of the off-limits policy, since underground tests are no longer carried out, the public can now tour part of this facility). The writer (NMS), when he was part of the Atomic Energy Commission's Plowshare program, visited the NTS many times over five years. The principal areas within NTS, where most of the nuclear explosions took place, are Yucca and Frenchman Flats. This Landsat-4 TM image shows that part of the basin complex, the Rainier Mesa on the left, and limestone mountains to its right. The large white patch in the upper right is Groom Lake, a playa near which is the "infamous" Site 51, supposedly where super secret military tests were carried out (I never visited this area but could hear a rocket engine being tested there). Below the image is an enlargement of the Yucca Flats section of the scene in which are a number of craters. Several, such as Sedan (see Section 18) were explosion craters but the majority were collapse craters caused when an underground nuclear explosion produced a cavity which then failed allowing the alluvium above to sink into it, leaving a circular depression at the surface.

Just to the northeast of Yucca Flats around Groom Lake is the famous (more properly, infamous) Area 51 which is probably the most supersecret stretch of land in the United States. Here are a Russian KVR-1000 photograph and an IKONOS image of facilities within this area, which lies about 110 km (75 miles) northeast of Las Vegas:

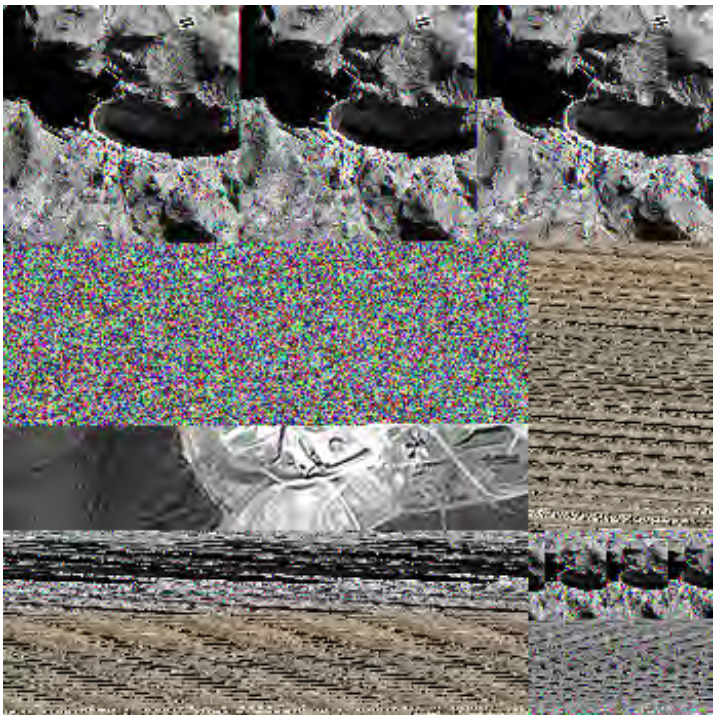


The mystery of the nature of government activities at Area 51 remains. Many think it is where either the U.S. military builds radical new airplanes or keeps captured UFOs. The U.S. Air Force does have a presence there. One confirmed activity is the testing of rocket engines.

Here is a space image of part of southern Nevada that contains Las Vegas (green patch), which we examined earlier on page 4-1. East of L.V. is the Hoover (Boulder) Dam which has created Lake Mead on the Colorado River.

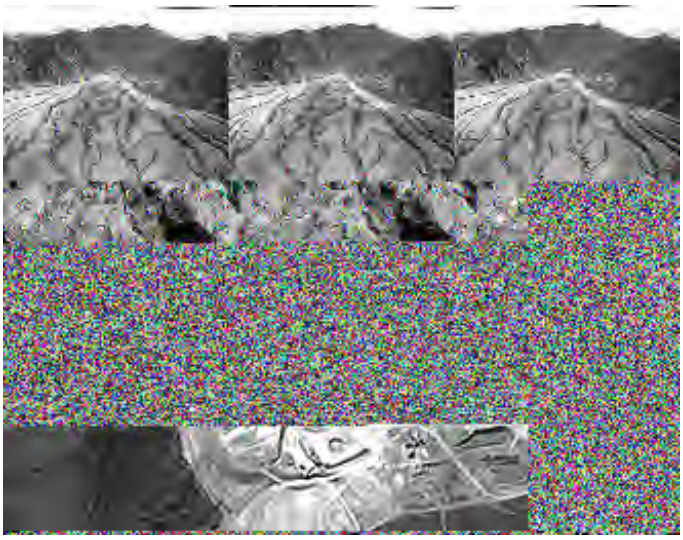
Nearby Lake Mead is the largest manmade freshwater lake in the U.S. This is a close-up satellite view:

This IKONOS high resolution image shows the Hoover dam and the gorge in which it was built in the 1930s:



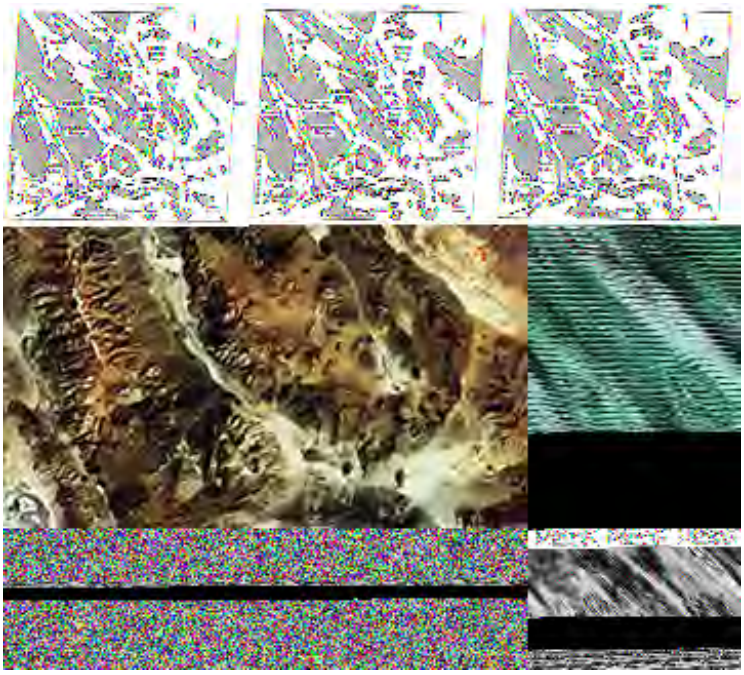
Now, back to the confirmable. Another typical block fault range occurs north of Tucson, AZ and is capped by a pine forest. At the foot of the Santa Catalina Mountains, shown below, is a thick, deeply gullied alluvial fan. You can check the setting of this area in the images of the Tucson, AZ

area on page 4-1.

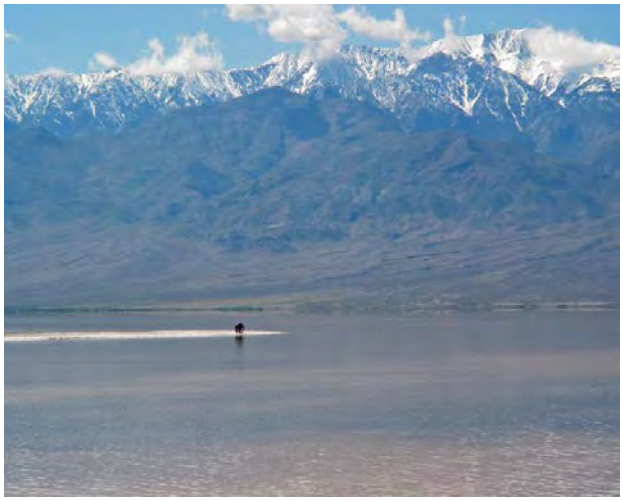


The dry desert air in the Tucson region means that materials are less susceptible to material erosion or degradation. Near the Davis-Monthan Air Force Base is a stretch of open land where hundreds of surplus (retired) military aircraft are kept parked - mainly as a reservoir of spare parts for other planes. Here is a Digital Globe space view of some of these:

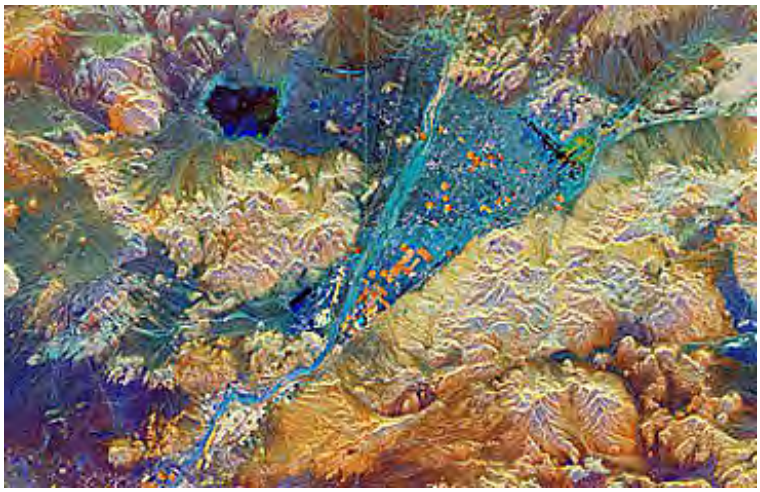
One of the classic areas within the Basin and Range, near its southwest margin, contains Death Valley and other block fault ranges. We show it below along with a map indicating the names of the principal topographic features.



Use this map to locate mountain ranges in this MODIS image of much the same area between the southern tip of the Sierra Nevada and the Spring Mountains near Las Vegas. The Owens Mountains are the long slender range east of the Sierras. Look particularly for the Panamint Mountains, a range whose tallest peak is over 11000 ft, as shown in the photo below the image:



Below the Death Valley area, the Basin and Range ends against the Mojave Desert. That region is located within the structural salient north of the southern California Transverse Ranges (see page 7-2).. Small ranges within the Mojave closely resemble the barren block fault mountains of the Basin and Range types. In west central Mojave Desert one of the larger towns in this desert terrain (where summer temperatures can reach 120° F) is Barstow. It is near several small mountain ranges and is just north of Apple Valley and Victorville at the edge of the San Bernadino Mountains of the Tranverse ranges. Note the alluvial fan (blue) made from an ephermal stream coming from those mountains. Here is a color radar image of an area near Barstow, California, made from SIR-C C- and X-band imagery (see page 8-7).



The Basin and Range type of geology is widespread over western North America (see map above). From Oregon-Idaho through Nevada and Utah, it spreads east into southern Arizona and New Mexico into northern Mexico itself. One area of classic ranges and valleys is around Phoenix, along with Las Vegas the two fastest growing large cities (each over 1,000,000 people) in the southwest U.S. The next image was an experiment done in the early days of Landsat-1 to use MSS Band 7 (IR) to render vegetation green. The fertile (irrigated from reservoir lakes) land in the Salt River Valley within which Phoenix was built shows up as actively growing in this February, 1973 image:

Part of the sprawl of fast-growing Phoenix appears in this aerial oblique photograph:

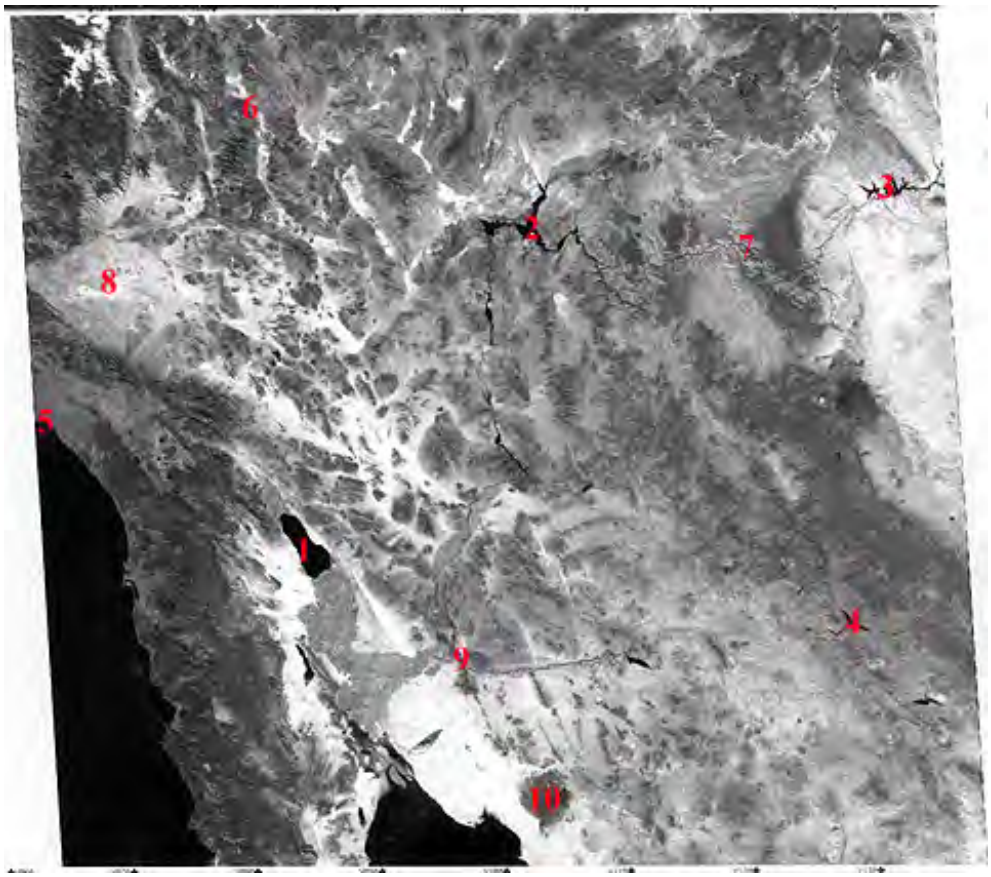
Courtesy: Carolina Map Distributors

Around and south of Phoenix the mountain ranges are small and isolated. North are the high, forested Mazatzal Mountains - a much larger block that includes also the Sierra Ancha and Pinal Mts. In the latter, note a pinkish red patch, which marks mine waste from the Globe-Miami-Superior copper mining district. In these mountains is the dammed Roosevelt Lake: it is a wonder on weekends to find 1000s of power boats being hauled from Phoenix to cruise those waters.

The desert lands described above have a surprising paucity of sand dunes. One dune field that lies near the borders between California and Nevada, and the Mexican state of Sonora was photographed from the International Space Station:

Water from the Colorado River is diverted into the All-American Canal (visible in the image above). This 129 km (80 mile) long canal is the longest of its kind in the world. Part is shown in this ASTER image, which encompasses the uninhabited desert. The photo below shows the canal as it enters the Imperial Valley of California:

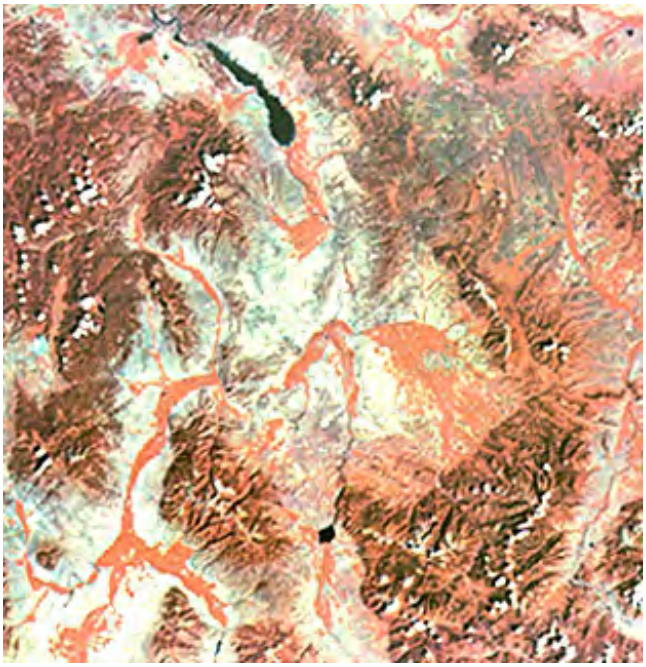
By now, you have gained experience with picking out features in the landscape at the medium scales presented in Landsat imagery. Before finishing our trip, we challenge you to apply your experience by locating or identifying geographic and geologic landmarks in a space image somewhat to the south of our main flight line. This image is much smaller in scale, covering about 1,100 km (684 miles) on a side. Make this a game of finding the places listed below. Look over the black and white Day-Vis HCMM image that shows most of the southwestern U.S. (part of southern California and Nevada, most of Arizona, a bit of Utah, and small segments of the Mexican states of Baja California and Sonora), including some of the Basin and Range.



6-11: Use a U.S or World Atlas to aid in correlating landmarks in the scene with mapped features. To help you get oriented, the red numbers 1, 2, and 3 are the Salton Sea, Lake Mead, and Lake Powell, respectively. Pin these down on an atlas to give you a feel for the scale. Then, relying on the maps, identify what is at or around the numbers 4 through 10. (The answers are below). Finally, without the aid of number guides, try to find the Gulf of California, San Diego, the Imperial Valley, the Sierra Nevada range, Las Vegas, Phoenix, and Tucson. Also, what is the name of the great geologic fault that marks (as a straight tonal boundary) the southern edge of the Mojave Valley? ANSWER

We now swing northward to the interior U.S. West north of the Basin and Range. Several geomorphic provinces come together in northern Utah, northwestern Wyoming, Idaho and Montana. This MODIS wide-angle image shows these states and the landforms within:

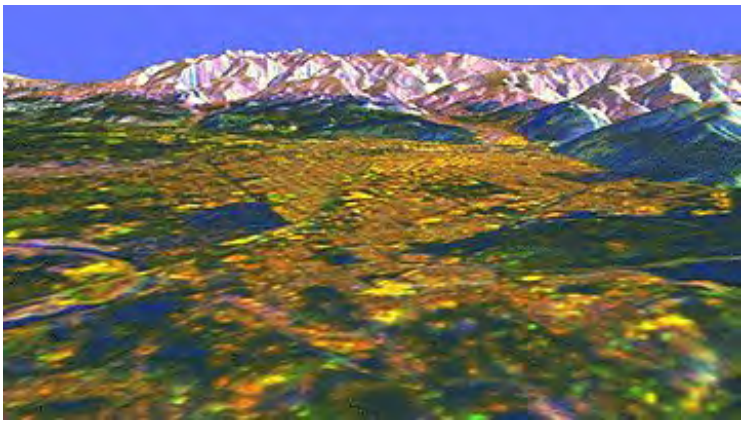
The MODIS image points to a westward deflection of the Rocky Mountain belt of folded rock ranges as it enters Montana. This next Landsat scene shows the style of expression of the mountaneous topography - one of individual blocks separated by broad valleys:



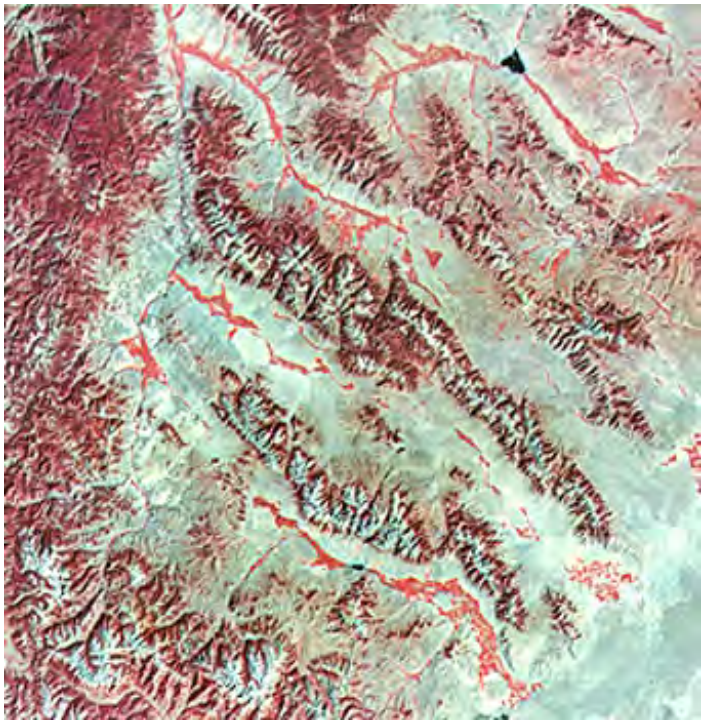
In the image are the towns of Helena, Bozeman, and Livingston. The ranges include Madison, Gallatin, Tobacco Root, and Crazy. The light blue patch just in from the center left margin demarks the tailings pile from the Butte and Anaconda copper mine district.

In northern Montana is the popular Glacier National Park, seen here as a space image and a ground view. The geology is dominated by sedimentary limestones, exposed in rugged peaks now steepened by glaciation. The nearby Lake Louise, in Banff National Park, Alberta, Canada shows the layering of limestones even better:

Within the Northern Rocky Mountains, individual ranges are given their own names. Here are a Landsat image and then a perspective view made from SIR-C data of the valley containing Missoula, MT (the state's largest town) as it is bounded to the east by the Mission Range.



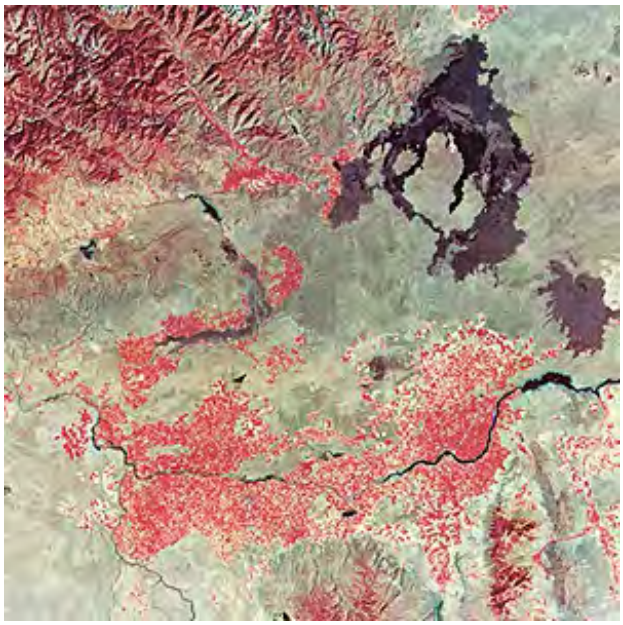
The Idaho segment of the Northern Rocky Mountains begins with several NW-trending individual ranges, separated by valleys and a more continuous mountain mass to the west, as seen on the left side of this Landsat-1 image:



Idaho' state capital, Boise, is typical of a small western mountain city (but big enough for Boise State University's football team to have been in the top 25). We see it first in a ground panorama and then from space:

On previous pages in this Section note was given to sand dunes - a hallmark of deserts and desert like physiographies - in various states including Nebraska and Colorado. Idaho, although more north and in a climate that has a semi-temporate feel, also has sand dunes. The field shown below lies between Idaho Springs and Yellowstone Park. It was produced during the last ice age less than 10000 years ago.

The northern end of the Basin and Range abuts against the Snake River Plains in Idaho, seen here in this Landsat-1 full scene:



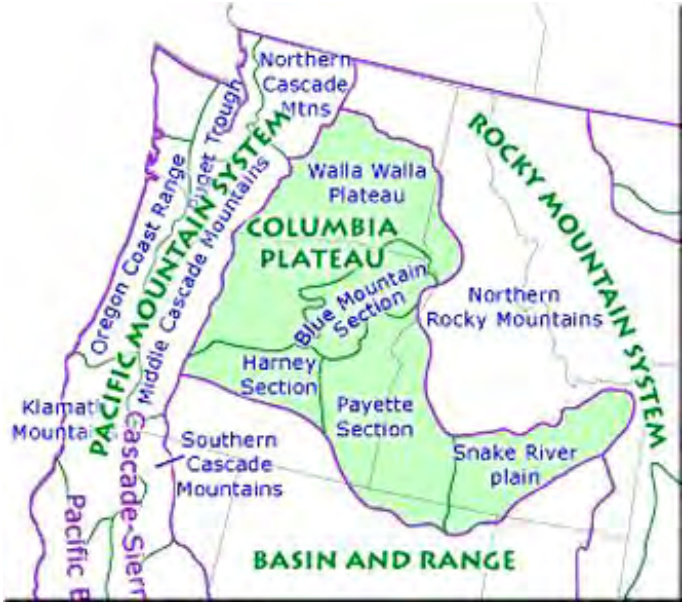
The Idaho Rocky Mountains are seen near the top. The Plains themselves are a series of bluish-gray lava flows (mostly basalt) that began some 10 million years ago. More recent is the large Craters of the Moon National Monument - the conspicuous black blotches - seen in two Landsat views. Water from the Snake River is used to irrigate the farms that produce Idaho potatoes, sugar beets, barley, wheat.

This colorized SIR-C radar image takes a closer look at the Craters of the Moon.

In 1963, when the writer (NMS) was working in the AEC's Plowshare program (at Livermore, CA) for engineering uses of nuclear explosives, I was given the responsibility of finding a site in the plains basalts for Project Schooner - a proposed 100 kiloton nuclear explosion in basalt designed to make a crater (this was a prototype experiment to determine feasibility of using nuclear devices to dig a second Panama Canal in Central America). The most promising area was Bruneau Canyon just west of the left edge of the above false color Landsat image. After more than a month on site, and several expensive drill holes, I recommended abandoning this once promising area (the county where the detonation would have taken place has 25000 cows and less than 5000 people - logistically favorable) when unexpected lake beds were found buried by younger flows near the surface - these would have compromised the diagnostics of the

cratering event by creating multiple echoes. (The beds consisted of loose, fine-powdered dust; the dust clogged local roads and got into auto engines; one day 4 of 5 field vehicles were out of commission requiring us to leave the area clinging to a drilling water truck). Here is a space view of the Bruneau Canyon area and the Canyon itself.

The Snake River Plains are the eastern extension of the Columbia Plateau. That physiographic unit is dominated by a series of flat-lying stacked units of basalt that were extruded over a wide area in Tertiary and Quaternary times. Here is a map of the Province.



The best exposure of basaltic lava flows in the Columbia Plateau is Hells Canyon, a gorge cut by the Snake River.

Typical Plateau terrain, consisting of relatively gentle rolling plains, is shown in this astronaut photo that displays the junction of the Snake River (right) with the Columbia River just south of Richland, Washington. The Hanford nuclear energy research facility lies to its north.

Up the Columbia River some 110 km (75 miles) is Wenatchee, WA, which nestles against the

eastern Casacades as it juxtaposes against the west edge of the Columbia Plateau province.

Both the Basin and Range and Columbia Plateaus conjoin with Pacific Coast physiographic units, examined on page 6-9.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

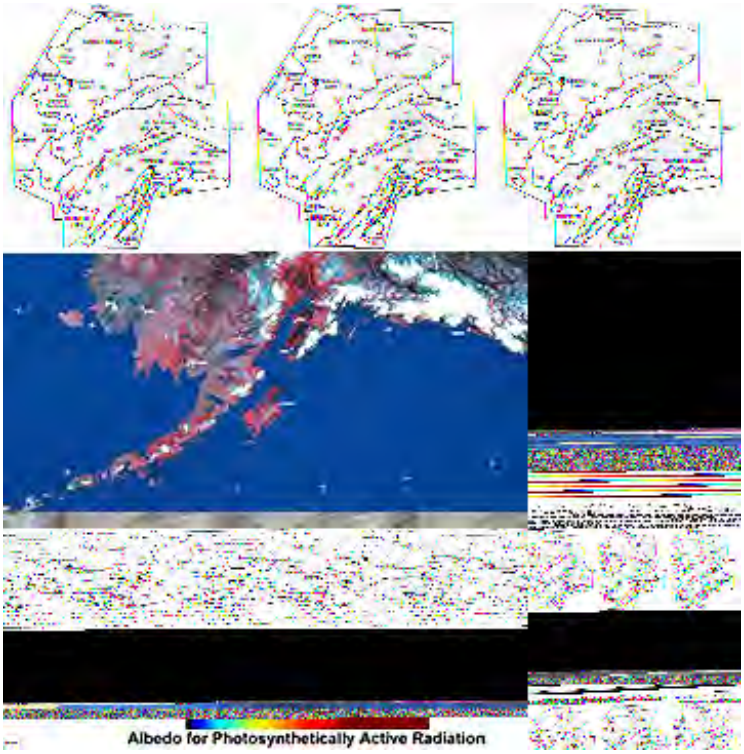
Some of the most spectacular scenery in the 50 states now making up the United States is found in the last two states to be added to the Union - Alaska and Hawaii. Both are mountainous, in part because they are experiencing significant tectonic activity. Both contain active volcanoes. We will conclude our U.S. journey by examining each state.

Alaska and Hawaii; American Pacific Territories

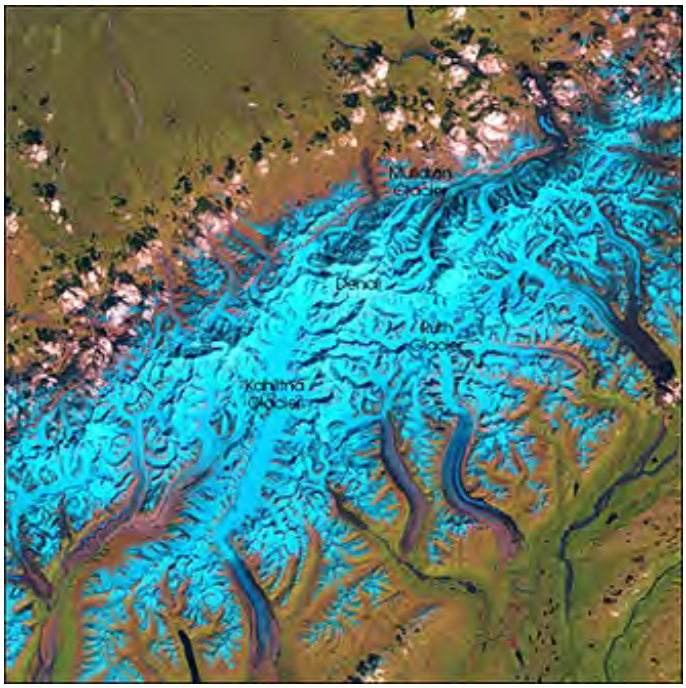
Having crossed the "48", one might expect the trip to be finished. But, the citizens of the two newest United States would probably be miffed by this apparent overlook. So, we will take off again, to visit Alaska and Hawaii.

Alaska, the 49th state, was admitted to the Union in January, 1959. Here is a nearly complete view of the largest (twice the size of Texas) of these United States, in a color Landsat mosaic (see page 7-3 for an earlier black and white version):

Its key geographic and urban features are shown in this map:

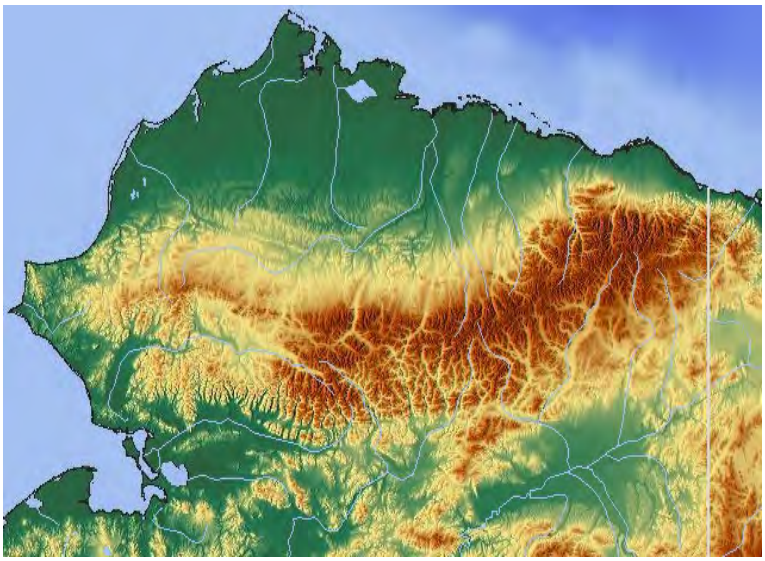


Lets look at the Alaska Range closer up using this Landsat 5 image. Mt. Denali (once named McKinley to honor the assassinated U.S. President, but now called by the Indian name "Denali") at 6160 m (20320 ft) is the highest point in North America. This mountain is treacherous to climb because of sudden, fierce storms; many lives have been lost in the attempt.



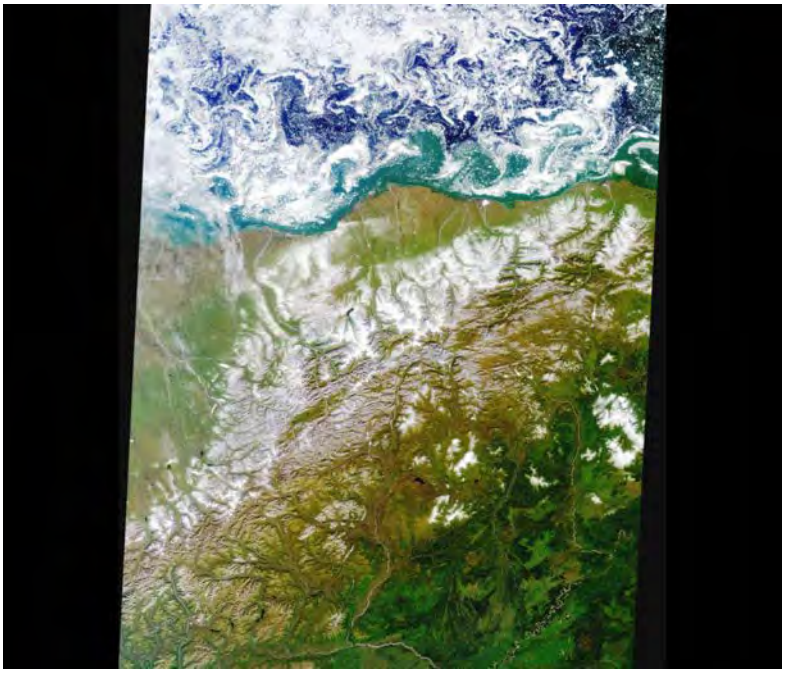
This magnificent peak is impressive from the ground:

The Brooks Range is the northernmost mountain chain in Alaska. It stands out in this topographic map, which suggests its landforms are created by glaciation. The aerial photo beneath it confirms that observation:



The highest peak in the Brooks Range is Mt. Chamberlain (2749m; 9020 ft). Here is a ground view of this mountain (in the far ridge):

On the North Slope of Alaska is the great Arctic Wildlife Refuge, which is the scene of years of 'infighting' about whether to open it up for oil and gas drilling or maintain it as a pristine nature preserve, supporting millions of caribou and other game. Here is a MISR image of the Refuge:



The northernmost town in the United States is Barrow, on the North Slope. It is located in the tundra midst thermokarst lakes. Here it is pictured by an EO-1 Hyperion image:

One of the big three cities in Alaska is Anchorage (the others: Juneau [the capital] and Fairbanks). Locate it in the map above. Here is a SIR-C view:

Seen from Cook Inlet, the central part of Anchorage is set off strikingly against the spectacular background of the Chugach Mountains.

Fairbanks is the home of the University of Alaska. We see it first as a false color satellite subscene, and then from the air:

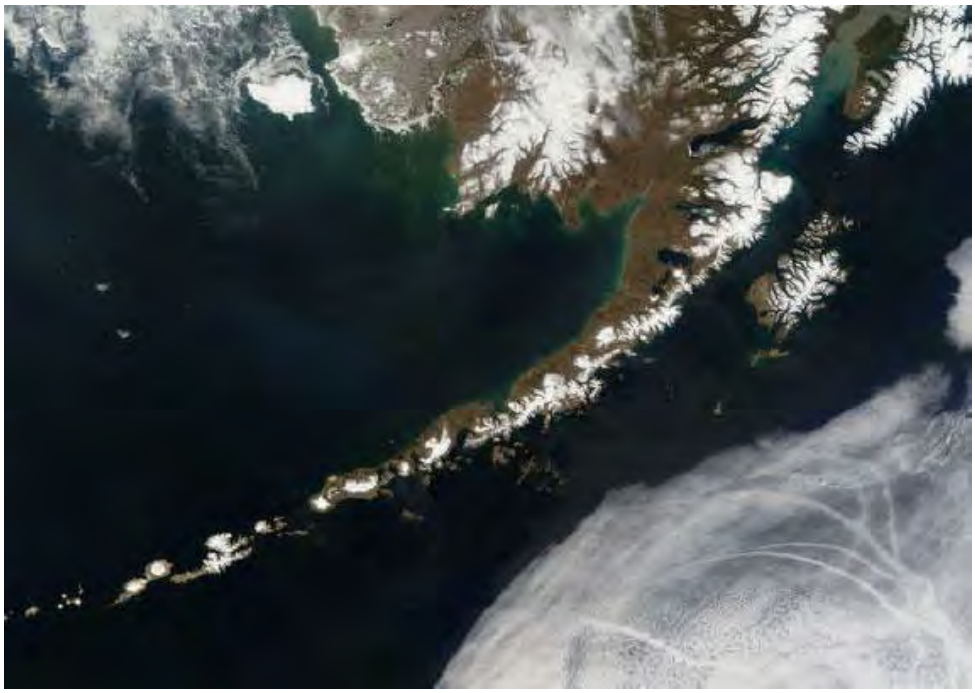
The rugged glaciated Chugach mountains extend to the southern coast of Alaska. Part if the range is shown here in this MODIS image:

Here are the Chugach mountains near Port Valdez (where an oil tanker spilled hundreds of thousands of oil, creating a major disaster to wildlife).

Just off the bottom right of the above Alaska mosaic, to the south, is the long "panhandle" of Alaska that is nestled against high mountains of the Rocky Mountain chain. Through it run the waterways of the Inland Passage, popular as the main theme of Alaskan Cruises. One destination is Alaska's capital, Juneau, seen here in a Landsat image. Part of the ice field to its east is the Skagway Glacier:

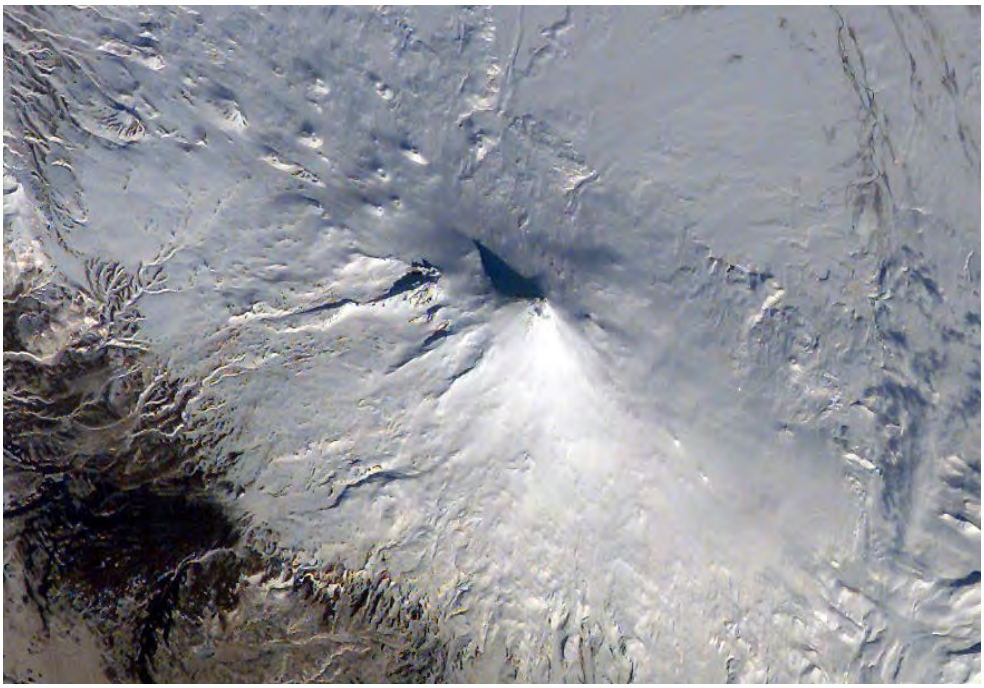
This is Juneau, Alaska's capital, from the ground:

Alaska is huge - double the area of Texas. It also extends, as part of North America, almost to the latitude of Japan. From Alaska's mainland on the east there is a long, narrow chain of islands known as the Aleutians. These are the best example on Earth of what is known as an Island Arc - a sequence of volcanoes along a classic subduction zone (to the south). Here is a MODIS image of much of this chain, followed by an International Space Station photo of the Cleveland Volcano in eruption, and then a ground photo of a several of the volcanic stratocones:

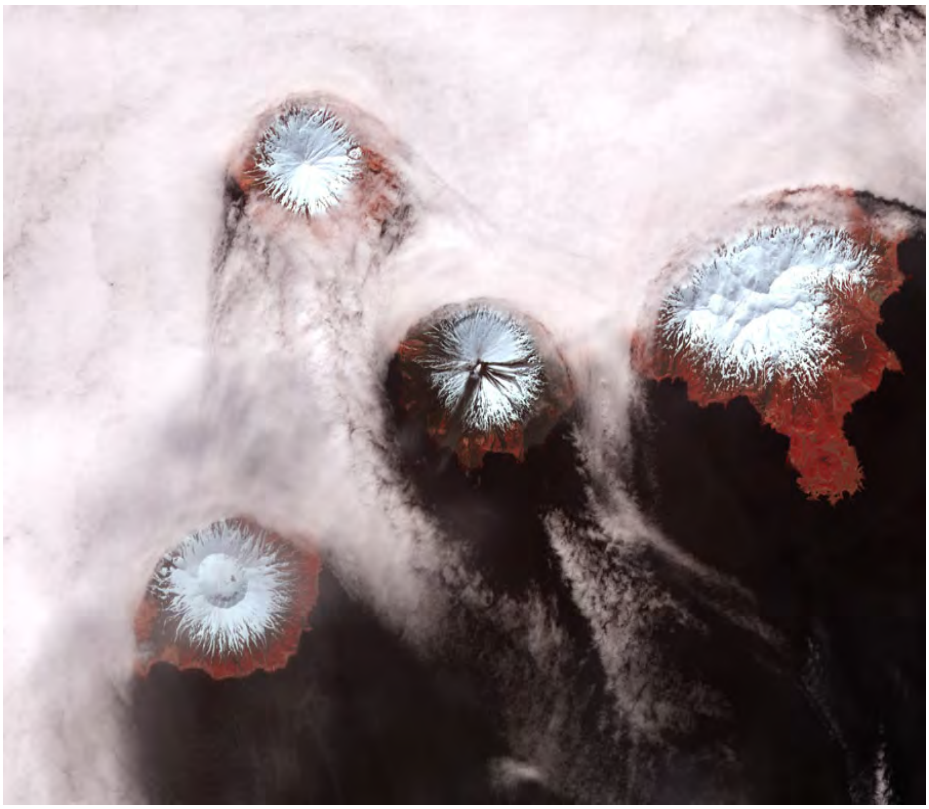


The westernmost village in North America is Unalaska, in the Aleutians:

The largest of the Aleutian islands is Unimak (4120 km2; 113 km long). It is noted for its active volcano - Shishaldin:



Sometimes the Aleutians are shrouded in fog, but the taller volcanic peaks can rise above this cover, as seen here:



The westernmost island in the Aleutians is Attu. This is one of two such island occupied for a time by the Japanese Army in World War 2. Today it is home to a radar station. It is also the destination for bird watchers because it is visited by many Asian species but being part of North America a birder can count these on their U.S. life list (the writer, a birder, once planned to go there [by air] in the early 2000s). The island is desolate but one can move around on foot or by bicycle. Here it is from space and from the ground:

Hawaii was the last state added to the Union, in December of 1959. It is a cluster of many small islands and six large ones. Here is a Landsat mosaic, with a map below it:

Best known of its many islands is Oahu, seen in this Landsat image:

Using satellite imagery and Digital Elevation Model (DEM) data, this perspective view of Honolulu shows the inlet to Pearl Harbor which, on December 7, 1941 (the "date that will live in infamy"), was partially sealed off by the Japanese sneak attack at dawn that brought the U.S. into World War II:

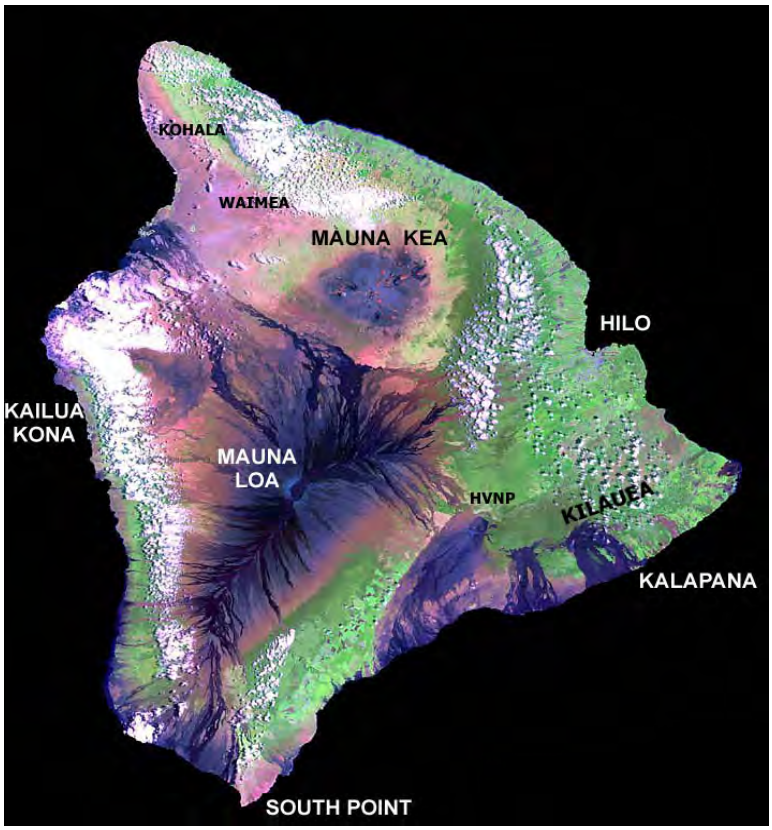
Honolulu is the capital of the state. We first saw it on page 4-3. Here is its downtown:

Honolulu is the home to the largest U.S. naval base in the Pacific - Pearl Harbor. On December 6, 1941 six battleships were there when the Japanese Navy attacked by air (fortunately, the

American aircraft carriers were on manoveurs elsewhere). More than 4000 were killed on the "Day that will live in Infamy" that brought the United States into World War II. Here is a view of Pearl Harbor from space and a ground photo of the U.S.S. Arizona and other ships aflame shortly after the bombing:

The best known natural landmark in Honolulu is Diamond Head, a ring around an extinct caldera, seen here is this aerial view:

The big island of Hawaii is shown in this Landsat image (compare with the one shown on page 17-3):



As mentioned elsewhere in this Tutorial, the big island is just the latest in the Hawaiian chain to have formed by eruption of basalt from the sea floor. In this region the Pacific tectonic plate is moving northwest. Each island and seamount in the chain resulted when that part of the crust was atop an active hot spot made up of a widespread mantle plume (molten magma that emerges as lava. The center of volcanic activity is Kilauea, whose caldera is shown here:

Another of the islands deserves special mention. Molokai is considered the most pristine of the larger islands, having a population around 7000. It was once the home of a famed leper colony, run by Father Damien. It now welcomes tourists.

This image does not clearly show its other claim to fame, indicated by this ground photo.

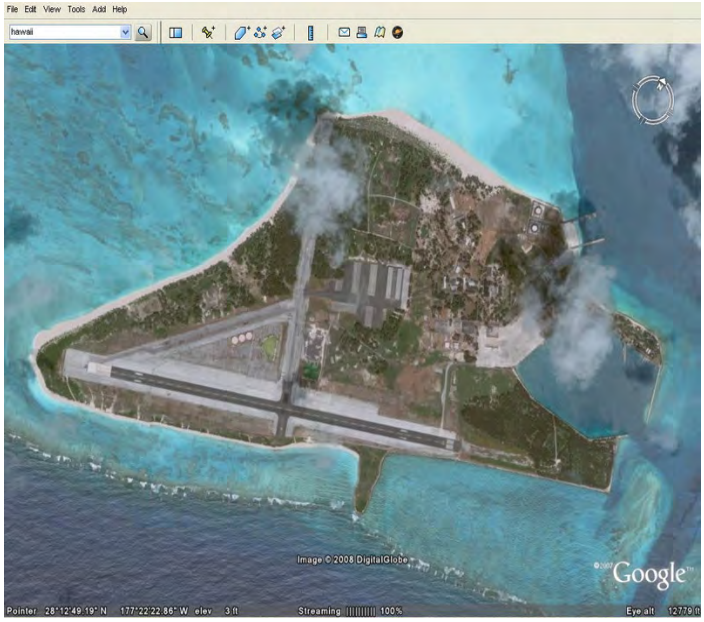
These are the highest steep cliffs in the world. Internet sites are ambiguous about the actual heights: several cited "greater than 3000 feet"; 2000 feet was also found. The cliffs make up most of the northern shore. Regardless, their origin was as spectacular as the cliffs themselve. Some time in prehistory Molokai was perhaps twice its present size. For reasons unknown, a huge section of the island gave way in a massive landslide, producing the cliffs now seen. Large blocks of the separated land have been found in the ocean tens of kilometers from Molokai.

The United States possesses a widespread collection of atolls and volcanic islands scattered about the Pacific. Several are used as military bases. Here are two views of one that is well known: Guam (it sends delegates to the Republican and Democratic conventions):

Guam appears to be uninhabited in the above two views but it actually has many residences as evident in this Quickbird image, which also shows a golf course:

American Samoa is part of the Samoan Islands (the others are an independent state) in the western Pacific. The American island, with about 60000 residents, was in the news in late September of 2009 because of a tsunami that killed more than 100 of its residents. Here is a space image of Tutuila, the main island, and a ground photo of the harbor near Pago Pago, its main city:

Some American Pacific islands have notably histories in World War II. Midway northwest of the Hawaiian Islands was part of the turning point battle of the Japanese campaign. While it was being attached by Japanese carrier planes, U.S. carriers launched strikes at the Japanese, sinking four of their carriers (the U.S. lost one). The Battle of Midway remains the biggest sea battle ever fought. Midway is actually two islands above water on a larger coral reef.



Less well known to most Americans are the islands of Saipan and Tinian in the Marianas in the western Pacific. These belonged to the Japanese at the start of World War II. They were invaded and captured by American marines near the end of that war and have remained in U.S. hands since (they elected to become part of the United States in 1986). It was from Tinian that the B-29 aircraft, the famed Enola Gay, took off to drop the first nuclear bomb ever on Hiroshima in Japan. Here they are as seen from space (Saipan is at the top):

With these scenes of the newest of the United States, we end our sojourn across the 48 contiguous United States and the two states beyond. If you have crossed the U.S. before, you can use this tour as a reminder of what you saw. If such a trip is in your future, this tour is a splendid preview of what you could look for. Bon voyage!

But wait! You should really be aiming to become a world traveler. For the U.S., there already is an Atlas of Landsat scenes covering the 48 contiguous states. Would be nice to have a World Atlas composed of Landsat images (except, since up to 11000 are needed to cover all land areas, such an atlas would have more pages the a multi-book Encyclopedia; however, National Geographic has a global Atlas that uses space imagery). To entice you to think "worldwide", the remainder of this Section will take you on a space-based tour of the countries beyond the United States.

To affirm that you have now developed the skills needed to "geographically locate" a Landsat image, we are going to give you another (littler) test. On the next page are 8 full and four partial Landsat scenes, taken from all (except the Antarctic) of the other continents besides North America. They are labeled from A to H. Using any means at your disposal (but a World Atlas is best) try to locate each scene. To help you in this, we will put a few key word hints beneath each image.

So, to proceed with the game, press "Game". This will get you in the proper frame of mind to then pass onto the second part of this Section - the travelogue through the continents that

helps you, through looking at various characteristic cities and landscapes, to gain a perspective on the wide varieties of scenery that make up the land surfaces of planet Earth.

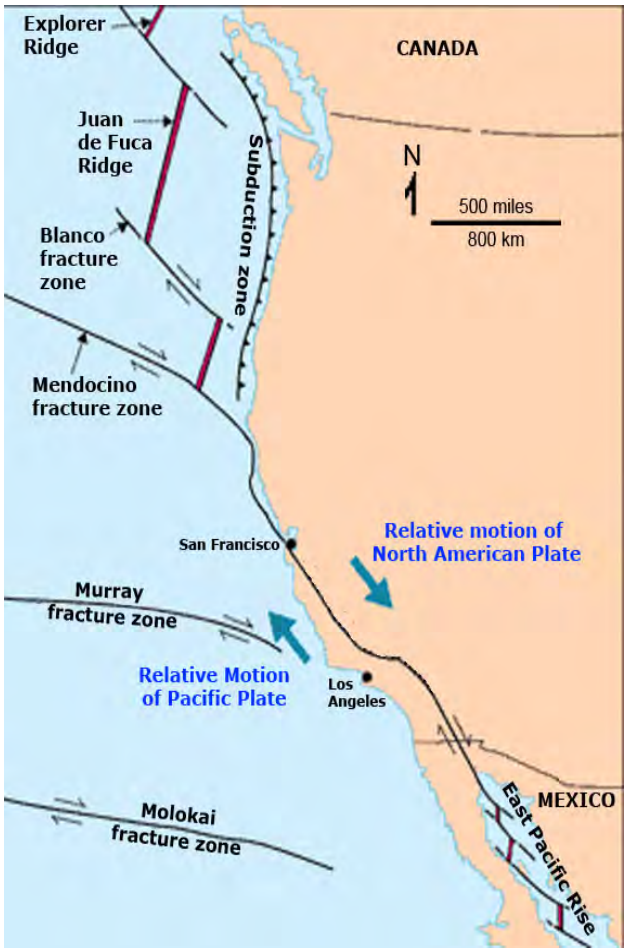
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The Basin and Range in Nevada abruptly terminates against the Sierra Nevada, rugged glaciated mountains pushed up from the crust as a tilted block whose western side continues as a structural basin (Sacramento and San Joaquin Valleys) now filled with erosion debris from the mountains. More inwash comes from the Coastal Ranges, a recent series of folded/faulted mountains pushed up by the east-moving Pacific Plate against this western edge of the North American Plate (now complicated by lateral movements along the San Andreas and other faults). Our journey ends in San Francisco but we also look at Coastal Ranges and the Cascades in Oregon and Washington.

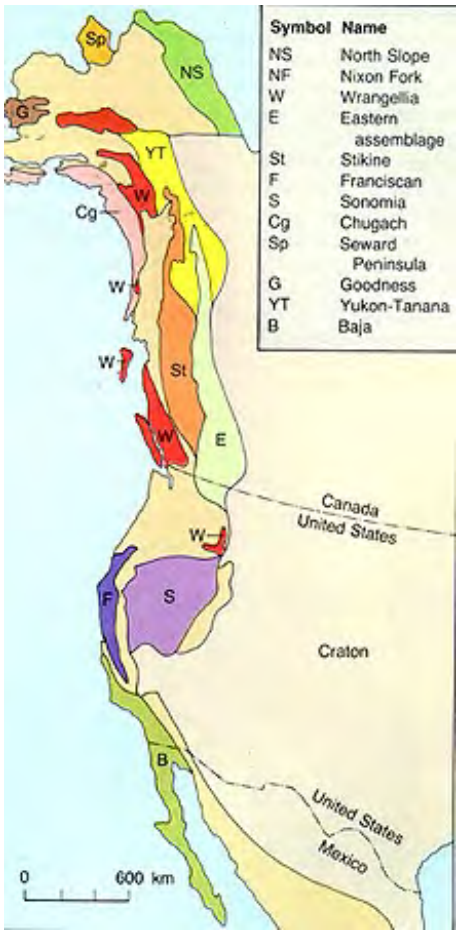
The Far West: The Pacific Coast; The Sierra Nevada Range; The San Francisco Bay Area, CA; The Cascades; Seattle, WA

Our imaginary flight finally nears the West Coast of the United States. It is a region of active geology (both tectonic and volcanic) and magnificent, varied scenery. The entire coast from Baja California (and into Mexico and Central America) is along an active plate margin. For California, Oregon, and Washington, the tectonic framework is shown in this figure, which plots the location of the great transform fault, the San Andreas, and a subduction zone from northern California past Washington:



The mountain units are part of a vast array of folded and tilted blocks, and accreted terranes that extend from Baja California on the western edge of the North American tectonic plate to

well into western Canada northward into southern Alaska. The map below shows the major terranes that were driven onto the western edge of the North American continent. The nature of these terranes is discussed in detail in Section 17, starting with page 17-6. It is worth noting here that all of the western continental edge is an active plate margin. From San Francisco southward the boundary is complex, involving a failure to subduct so that movement is lateral along fault systems (San Andreas et al.). From northern California over most of the remaining coast into Alaska, the boundary is of the subduction type (mostly of the Pacific plate, but from California to southernmost British Columbia a small separate oceanic crustal block, named the Juan de Fuca plate, is subducting and causing the volcanoes of the Cascades to be emplaced as plate scraping produces heat to melt rock.

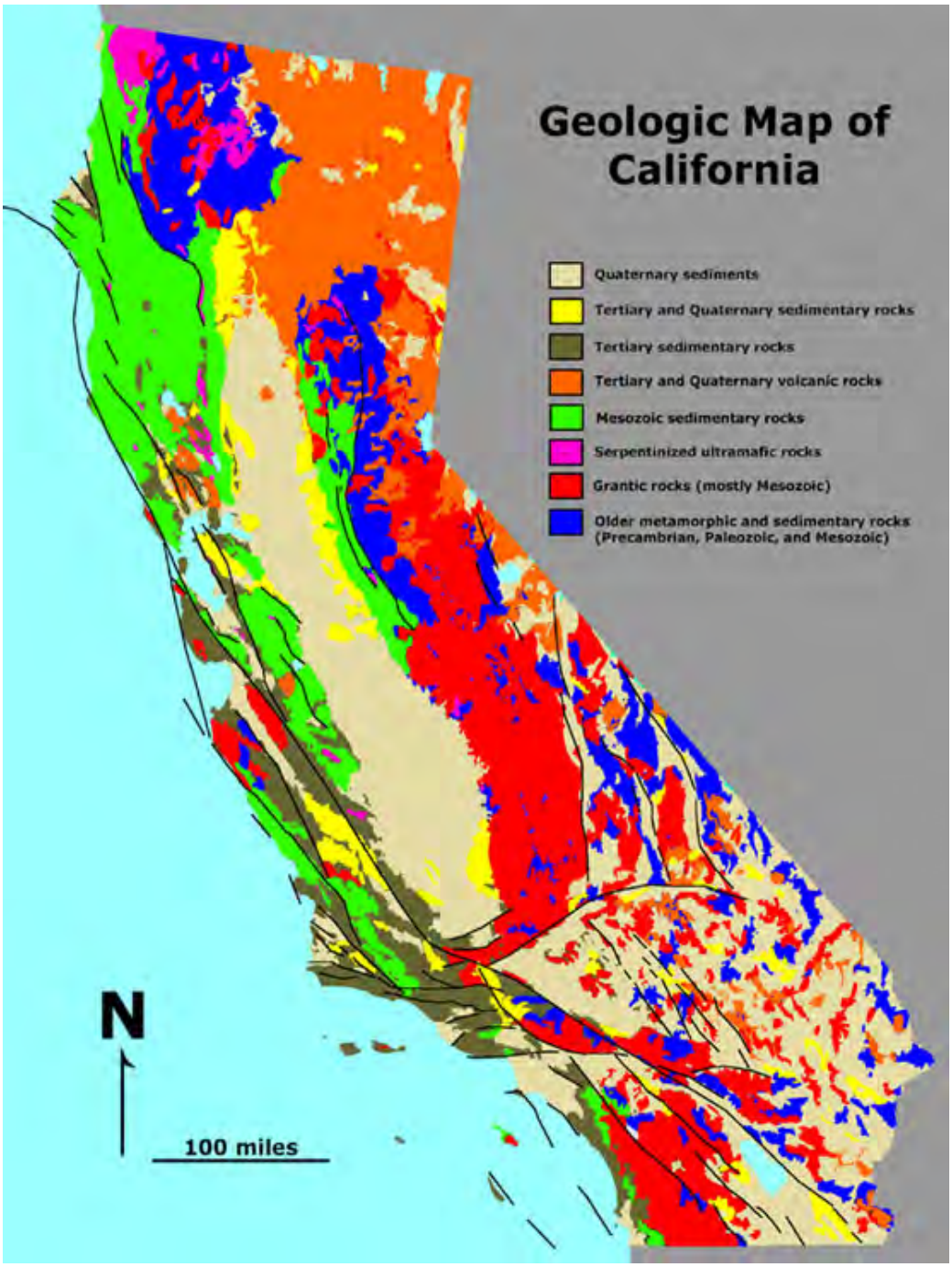


This map shows the major subdivisions of the Pacific Ranges:

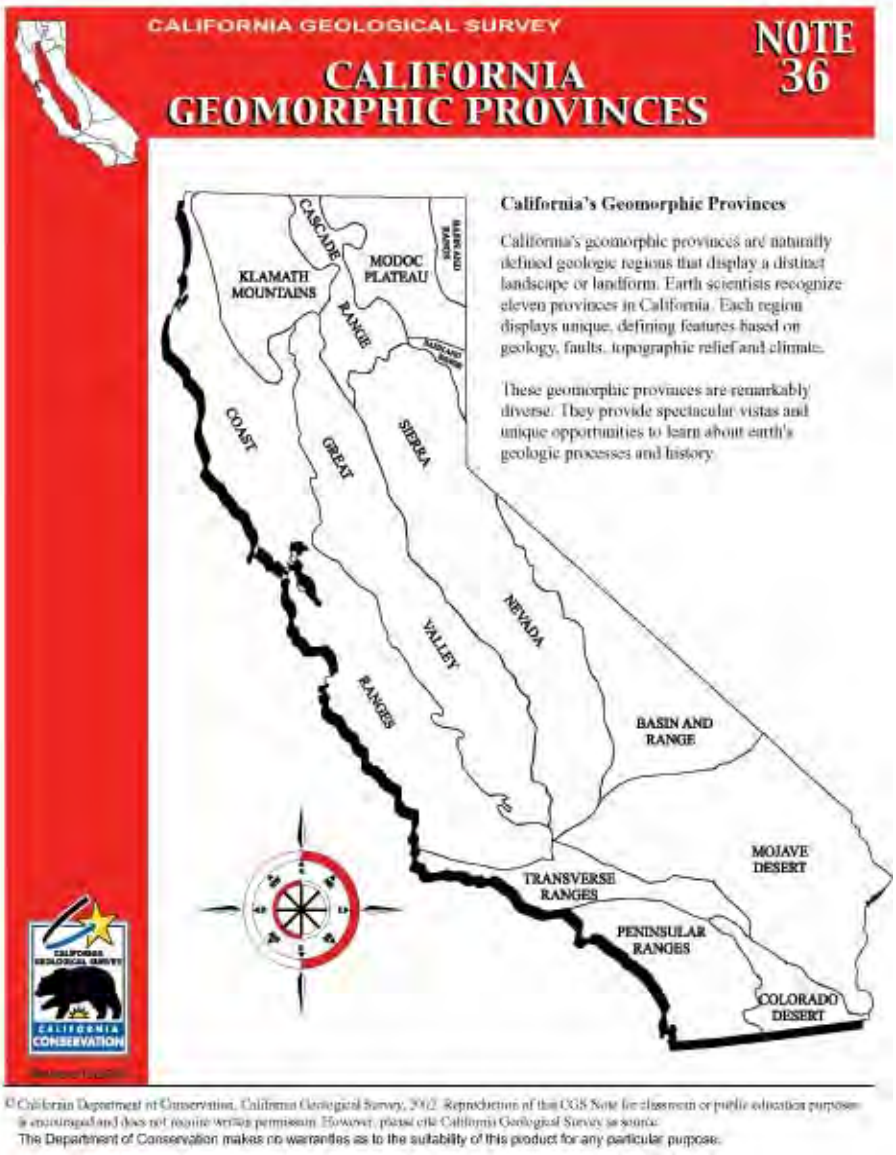
Our aircraft is headed toward San Francisco. The scene below will introduce you to California - a fabled place to both the Americans and to the world at large. With its 40 million people, its development has made the state the 5th largest economy in the world. Lets take a look at the entire state as seen in one of the early Landsat mosaics made by the General Electric Space Division lab near Beltsville, MD.:

The physiographic/geographic features of California are shown in this map:

Since we will concentrate on California we will set up a geologic framework for that state using this map.

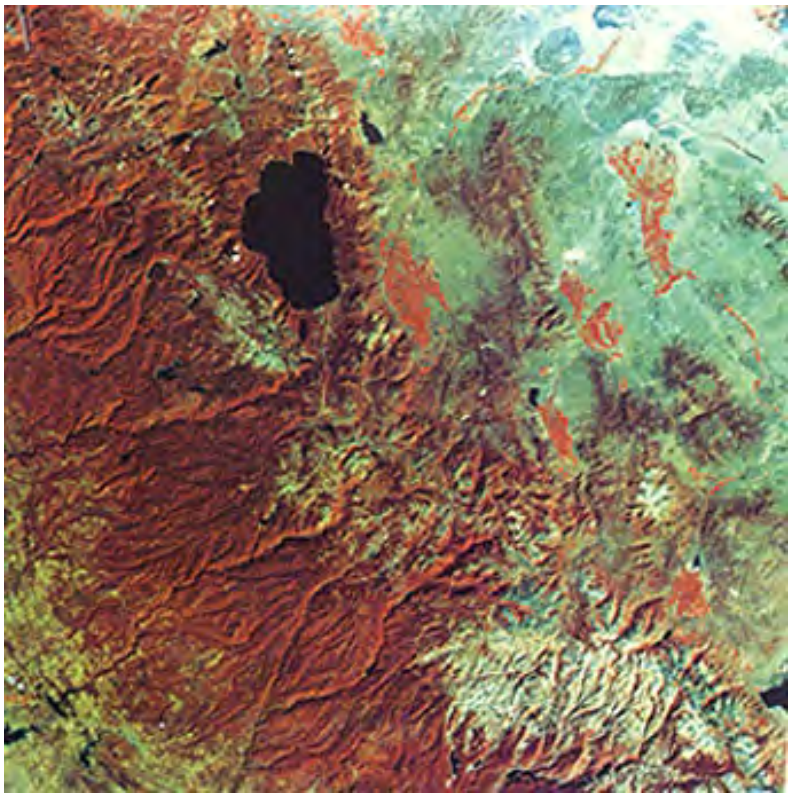


The major physiographic (geomorphic) units for California correlate well with the geologic units:

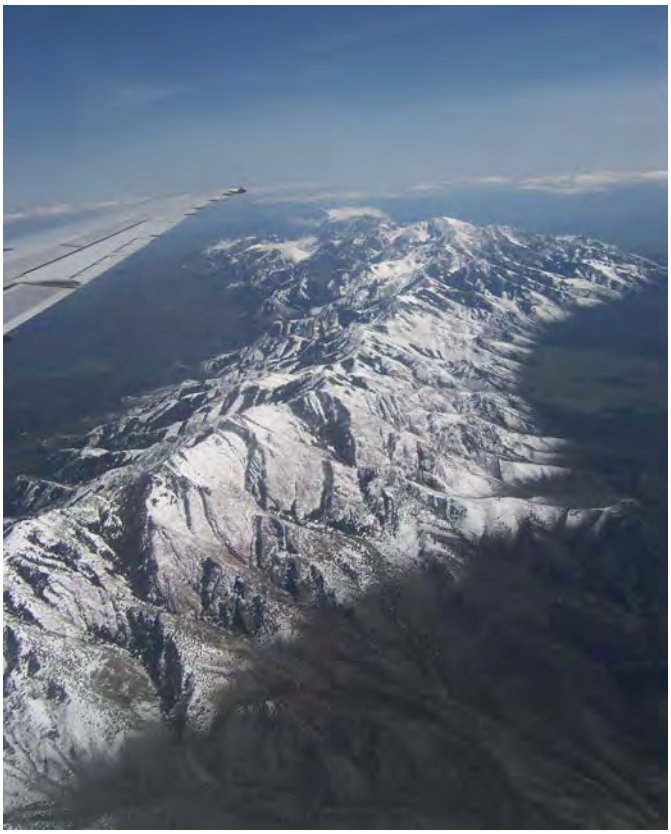


© California Department of Conservation, California Geological Survey, 2002. Reproduction of this CGS Note for classroom or public education purposes is encouraged and does not require written permission. However, please cite California Geological Survey as source.
The Department of Conservation makes no warranties as to the suitability of this product for any particular purpose.

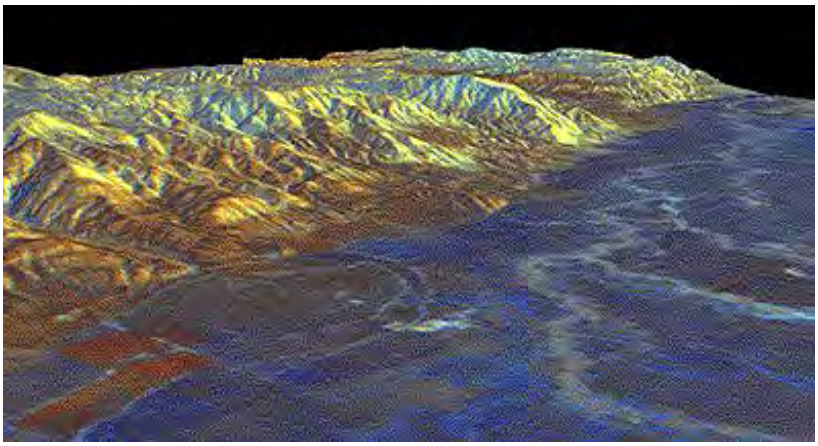
Entering the state from Nevada, we proceed across the Sierra Nevada Range, the large, strongly glaciated tilt block of crust composed mostly of igneous and metamorphic rocks associated with numerous interpenetrating batholiths. This mountain range rises precipitously to 4,421 m (14,494 ft) at Mt Whitney - the highest point in the 48 states - on its east side (south of this scene). The east side tends to be steep-fronted, being fault-bounded, whereas the western slopes are gentler in inclination, although still deeply dissected by streams. The scene below shows off the northern Sierras with Reno at top center, volcanics in the westernmost Basin and Range appearing in bluish tones, Lake Tahoe nestled in the High Sierras, and many deep valleys. Dark firs and other vegetation persist up to about 2750 m (9000 ft), above which trees are absent and rock bare (light tan).



The Sierra Nevada is impressive as viewed by the astronauts from space. It is also compelling to see from the air. These two images confirm this:



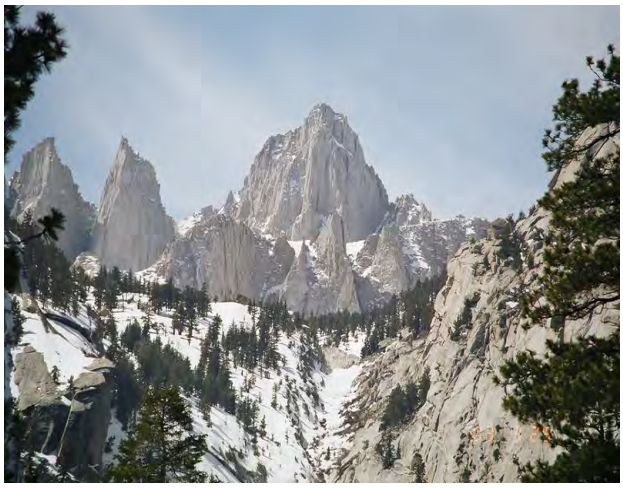
The eastern front of the Sierras rises steeply from the Owens Valley to its east (to the south of the Landsat image. This image was made by combining a SIR-C image with topographic data derived from its altimeter:



Many "westerns" or "cowboy" movies are filmed along the east side of the Sierra Nevadas. The distinctive vegetation (semi-desert) and the gray granite rocks are clues as to the setting being eastern California. This is an example:

The next image shows the section of the east front of the Sierra Nevada that includes Mount Whitney and the town of Lone Pine:

The highest mountain in the continental 48 state is Mount Whitney at 14505 ft. Here it is:



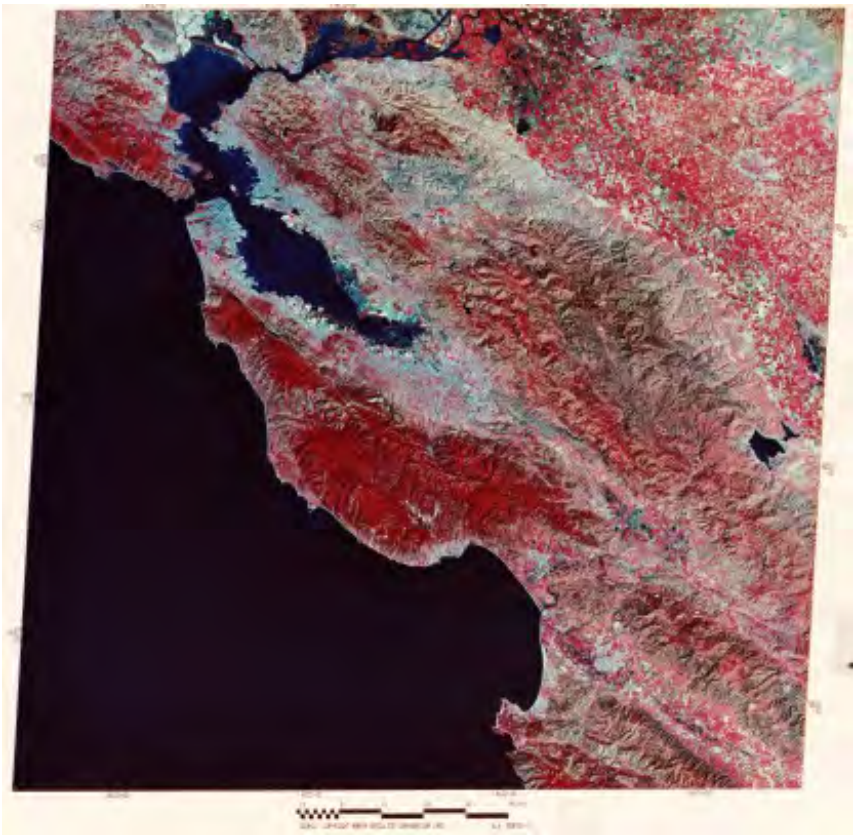
In the high country, glaciation has carved out deep canyons and rugged landscapes. This ground view of the Kings Canyon National Park area to the south shows why it is popular with rock climbers:

One of the best known National Parks is Yosemite. Its most featured visual landmark is the steep-faced (from glacial erosion) El Capitan, seen here from space by the IKONOS satellite:

The foothills of the Sierra Nevada have a notably different topography than the High Sierra:

Both the High Sierra and its foothills were the focal point of the "mad" gold rush of 1849. Gold has been continuously mined ever since. Much gold is found as placer deposits, i.e., found in river wash and sediment as nuggets and grains after being weathered out from the Mother Lode (mostly quartz veins in the granite complex). Active placer mining, using streams of water to wash through the river beds and separate the heavy gold from the sediment, is still going on today. This ASTER image shows the placer piles along the Yuba River in the foothills.

We then pass west over rolling ground in the Sierra Nevada Foothills across the Central Valley into the Coast Ranges around the San Francisco Bay area, all shown in this October, 1972 Landsat image.



The Coast Ranges are generally hills to mountains that are not very high. We saw in Section 1 (Morro Bay) that a hallmark of these ranges is the brown grasses that cover the landscape during all but the rainy season. But woodlands are present in some of the higher areas of these mountains (which have been forming during much of the Cenozoic):

Most of the landmarks we describe next are seen in this high-altitude aerial oblique IR image taken by a camera onboard a NASA U-2 aircraft. The view from this perspective looks east-northeast from a point over the Pacific Ocean. In the distance are the snow-capped Sierra Nevadas and the Sacramento Valley before it. Try to match specific features in this IR photo (which, incidentally, is blue-dominated because the sky blue effect was not compensated by a haze filter) with their counterparts in the Landsat image.

The Central Valley, known in its northern half as the Sacramento Valley (the southern half is the San Joaquin Valley), is one of the great agricultural regions of the world. The numerous farms visible in the above imagery attest to this dominant land use. Cash crops include beans, cotton, rice, barley, and sugar beets. Stockton, Fresno and Modesto are three large valley towns.

The Pacific Coast Ranges are part of the Western Cordillera (sometimes known in Canada as

the Pacific Cordillera and also as the Canadian Cordillera), which includes the Rocky Mountains, Columbia Mountains, Interior Mountains, the Interior Plateau, Sierra Nevada Mountains, the , and other ranges and various plateaus and basins. The Pacific Coast Ranges designation, however, only applies to the Western System of the Western Cordillera, which comprises the Saint Elias Mountains, Coast Mountains, Insular Mountains, Olympic Mountains, Cascade Range, Oregon Coast Range, California Coast Ranges, Transverse Ranges, Peninsular Ranges, and the Sierra Nevada Range. The term "Coast Range" is used by the United States Geological Survey to refer only to the ranges south from the Strait of Juan de Fuca in Washington to the California-Mexico border; and only the ranges west of Puget Sound, the Williamette valley, the Sacramento and San Joaquin valleys or 'California Central Valley' (thereby excluding the Sierra Nevada and Cascade Ranges).

The Coast Ranges are the most recent products of still active orogeny that results from the collision of the Pacific and North American lithospheric plates. The mountains consist of granular (clastic) sedimentary rocks in places metamorphosed by high temperatures and pressures or intruded by igneous magmas when previously they were buried. Their rocks have been severely deformed, but the ranges generally rise only a few thousand feet above the Pacific sea level. Passing near the coast but inland is the great wrench (horizontal movement) fracture known as the San Andreas Fault (in the Landsat image look for the abrupt termination of red [from California oaks, redwoods, and other vegetation] in the mountains south of San Franciso), which is the plane dividing the north-moving Pacific side of the fault from the more static North American side.

Two coastal bays stand out in the Landsat image: Monterey Bay near the bottom (noted for its rich aquatic fauna and pelagic bird life) and San Francisco Bay, which to its north becomes San Pablo Bay, along which San Francisco (west), Oakland (east), San Jose (south) and many other Bay Area cities are situated. The fault-bounded hills to the east are part of the Diablo Range, which becomes progressively less vegetated towards the Great Valley (at this time of year it is covered with the same brown, dried grasses we examined in the Morro Bay subscene).

Most of San Francisco appears in grayish-blue tones that indicate a sparsity of trees in the city (rather surprising to any resident or visitor who is struck by its great beauty). The exceptions are the lush vegetation (many eucalyptus trees) in Golden Gate Park (long rectangle) and the Presidio (a now deactivated military facility). Your monitor resolution may allow you also to pick out thin (blue) lines of ships (mothballed Liberty Ships and others from World War II) right of the narrow Carquinez Straits off the east end of San Pablo Bay.

In the southern part of the Bay are large patches with lighter blue or green tints. These are salt evaporating basins. The greenish ones are actually basins with red sediments (bright tones in Thematic Mapper Band 3 that thus pass green light in an RGB:432 transparency). A closer look at these basins is afforded by the MISR sensor onboard the Terra spacecraft.

Many consider San Francisco to be the most beautiful city in the U.S. Here is a photo that shows the central city from the tourist vista called Twin Peaks.

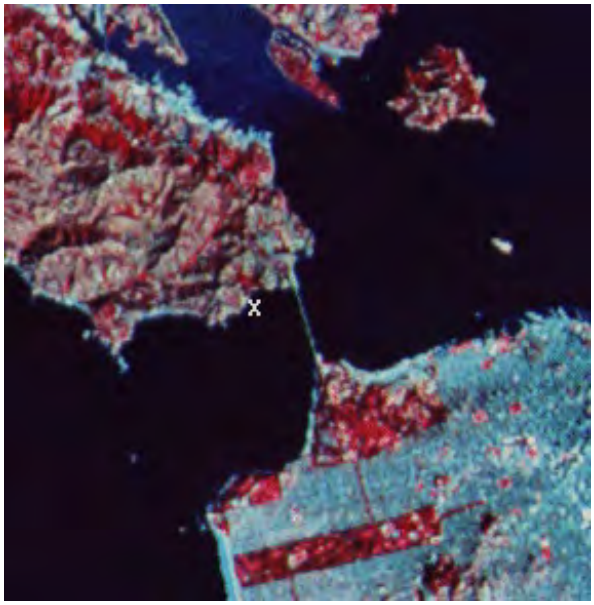
Lets take a closer look from space by enlarging the Bay Area portion of the full Landsat image.

6-12: In the above subscene, find Berkeley, Oakland, Hayward; San Jose; the San Francisco Airport, Palo Alto, and Mount Diablo. ANSWER

It is worth noting that Oakland and San Jose are major cities, with populations greater than a half million each, and both support major league sports.

We carry this enlargement even further to highlight the landmark bridge leading into the

Peninsula from the north, the Golden Gate Bridge. The view (lower image) of "the City" through the bridge is from the vantage of a small park on the Marin County side (at X in upper image). The Presidio and Golden Gate Park (rectangle) stand out because of their many trees (eucalyptus is common).



Such a well known landmark as the Golden Gate Bridge is an obvious prime target for high resolution imagery obtained from a satellite. The IKONOS system has produced this 2 meter image of the Golden Gate Bridge:

San Francisco is equally impressive when looking at its east side, from a southeast vantage point across the San Francisco Bay; the span is called the Bay Bridge and extends to the east (right) into Oakland:

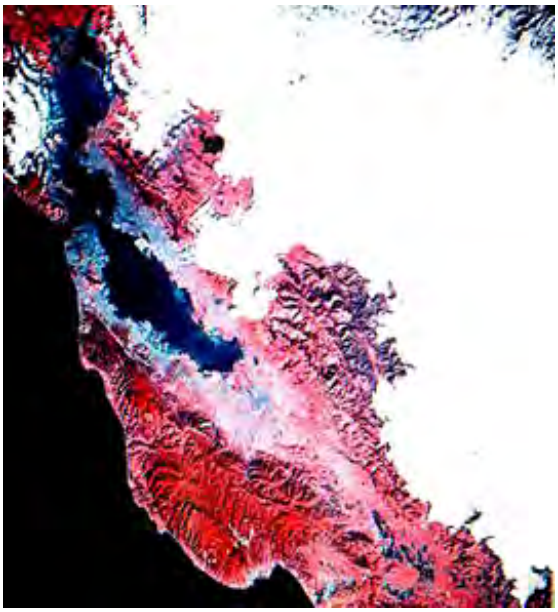
Courtesy: Carolina Map Distributors

A remarkable image of downtown San Francisco with its many tall building (some visible in the above photo) has been taken by the IKONOS satellite (SpaceImaging, Inc.). This color version has a spatial resolution of 2 meters:

Across the Bay near Oakland is Berkeley, the home of the University of California's flagship

campus in the northern part of the state.

Before taking leave of the City by the Bay and central California, we present one more interesting Landsat image. It shows the Bay area clear of fog (normally in summer fog rolls in every day) but the Great Valley to the east is entirely fog shrouded. This is the so-called tule fog that often develops inland under the right meteorological conditions. Note that this fog has spilled over the low pass where Highway 50 crosses the eastern Coast Range into the Livermore Valley.



One of the best known places in California - and for many a visual introduction to its beauties - is the Pebble Beach Golf course between Monterey and Carmel, about 100 miles south of San Francisco. Each year a major golf tournament takes place there in winter. Here is a view looking south:

Those living, or visiting, in California - if they have the time - often travel to or from San Franciso to Los Angeles via the famed Highway 1 that follows the coast much of the way. The scenery is spectacular (frequently used in TV commercials). The most scenic stretch is the Big Sur, south of Monterey. Here is a ground view:

Venturing to the southern part of California, we have shown images and ground scenes from the Los Angeles area and around San Diego in Section 4 of the Tutorial. The areas in and around Los Angeles were emphasized. We can see the locations of major cities in the area in this nighttime astronaut photo:

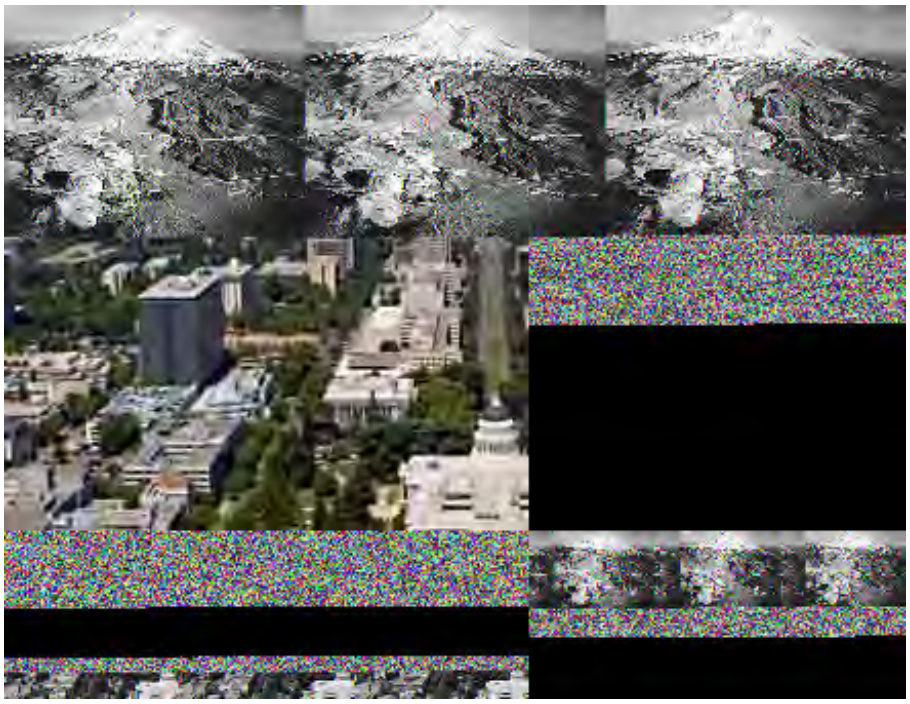
Parts of southern California are quite mountaineous. Here is a scene showing the San Bernadino mountains east of Los Angeles:

Off the shores of Southern California are several large islands, sometimes referred to as the Channel Islands. These are actually the above sealevel tops of largely submarine mountaneous protrusions on the shelf. Santa Cruz Island lies just south of Santa Barbara.

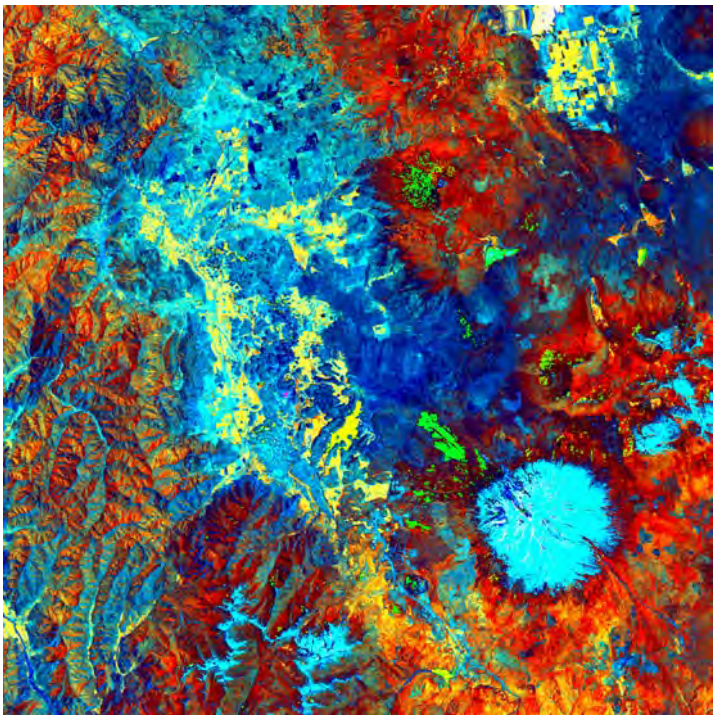
It is appropriate to show one more scene in the southern half of the state - Santa Barbara. This town is often referred to as the "American Riviera" because of its setting with mountains (Santa Ynez Range) as a backdrop. Here are two views of this lovely town:

Let's start to leave the Great State of California to go northward by first traveling up the Sacramento Valley to the state's capital at Sacramento, seen here in two space and one aerial view:

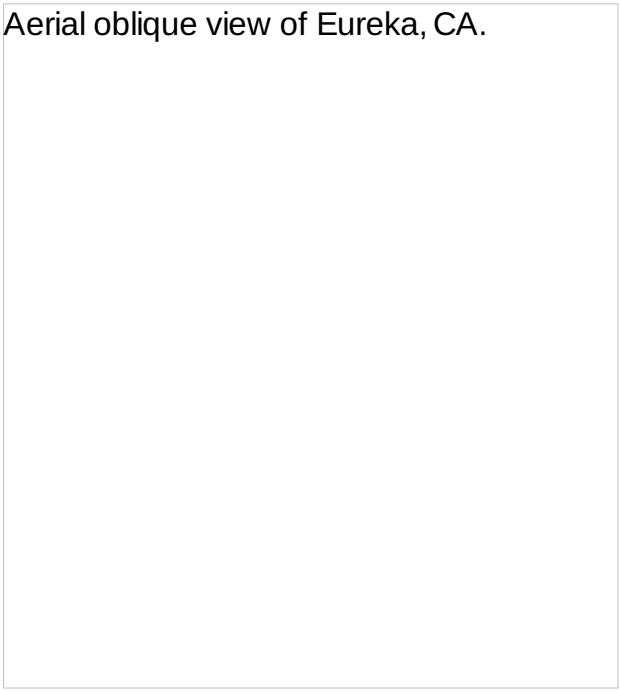
California contains at least one active volcano. Mt Lassen in the Cascades at the north end of the Sierra Nevadas last erupted in 1915. One of the most beautiful of all Cascade Volcanoes is Mt. Shasta (the larger white patch near the top) seen here in a photo taken from the ground looking north and then in an astronaut photo taken from space:



The photo is rather bland. This next image is a processed digital image of a satellite data set in which some of the ground features now stand out owing to their color assignments.



To the west is the scenic coastline of California. The largest town on the coast anywhere north of San Francisco up to the Canadian border is Eureka (population 42000+), seen in this IKONOS image and in an aerial view, which shows how the town is situated on the nearly enclosed Humboldt Bay.


Aerial oblique view of Eureka, CA.

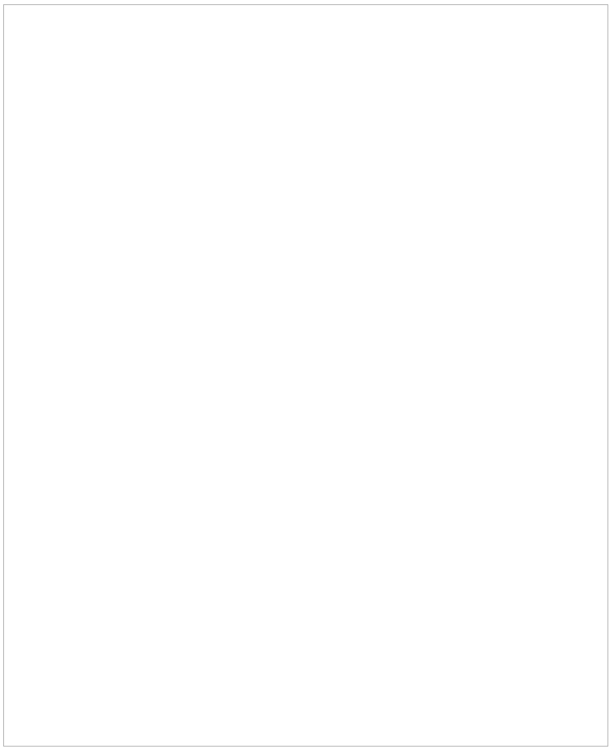
Before leaving the Golden State, look at this: a first-order Landuse/landcover map of the state was produced by the NASA Ames Research Center in cooperation with the NASA Jet Propulsion Laboratory. Here it is (admittedly hard to read at your screen scale):

Land Cover Map of California.

JPL has a short survey of the various ways that California has been monitored from space. Access through the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes "California from Space", September 6, 2003. To start it, once found, click on the

blue RealVideo link.

Whoa! Some viewers of this Tutorial who live in the Pacific Northwest may feel slighted in that we haven't shown any imagery so far that covers their fast-growing region. We can ameliorate this seeming oversight by allowing our airplane to change course in Nevada and head northwest into Oregon. The view below is a mosaic (production of this type of composite scene is treated in the next section) of images covering most of the western halves of Oregon and Washington:



The large river that divides the mosaic is the Columbia which is also the state border between Oregon and Washington. The near vertical set of white patches marks snow found on the higher elevations of the Cascades. The large white circular patch in the Oregon Cascades is Mount Hood, a few 10s of miles south of the Columbia River. North of the river are Mount Rainier and Mount St. Helens, among the numerous active and dormant, variably eroded, stratovolcanoes that make up the High Cascades. In northwest Washington is an E-W body of water off the Pacific Ocean that ends to the east with various islands and several inlets including the N-S Puget Sound along which Seattle is located. Near the upper left is the Canadian island of Vancouver. From the Cascades westward the land is heavily wooded (including abundant firs and other evergreens), as indicated by the widespread reds. East of the Cascades, the land is sparsely covered by trees - a consequence of the rapid reduction in rainfall as clouds pass over the Cascades, dumping their moisture, and producing a so-called "rainshadow" that creates a semi-desert ecosystem.

One Landsat scene in southern Oregon (just below the bottom of the mosaic) contains several features of special interest both within and east of the Cascades. Look:

Landsat-1 image containing Crater Lake.

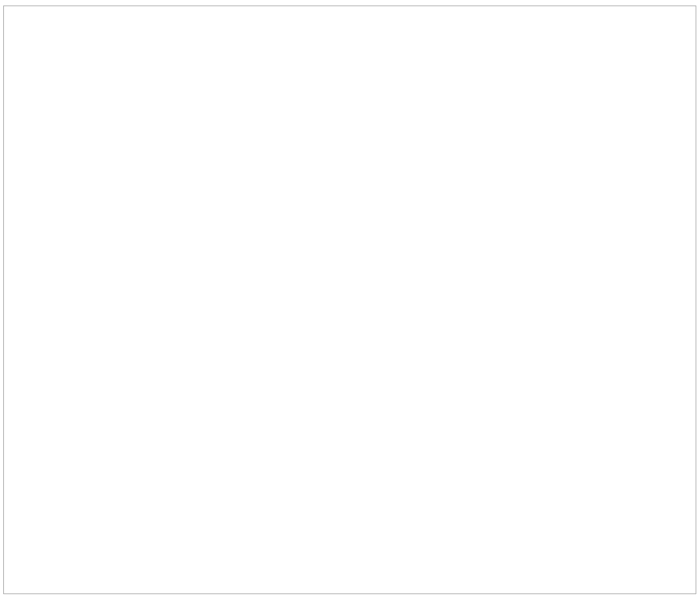
The focal point in this scene is the circular lake known as Crater Lake (near left center edge). This 12 km (8 miles) wide lake lies within a now extinct volcano, known as Mt. Mazama, that collapsed and erupted violently about 10000 years ago, perhaps before humans lived in the region. Thick ash beds resulted - much more than was expelled from Mt. St. Helens. Here is an astronaut photo of Crater Lake taken from the International Space Station:

Crater Lake from space.

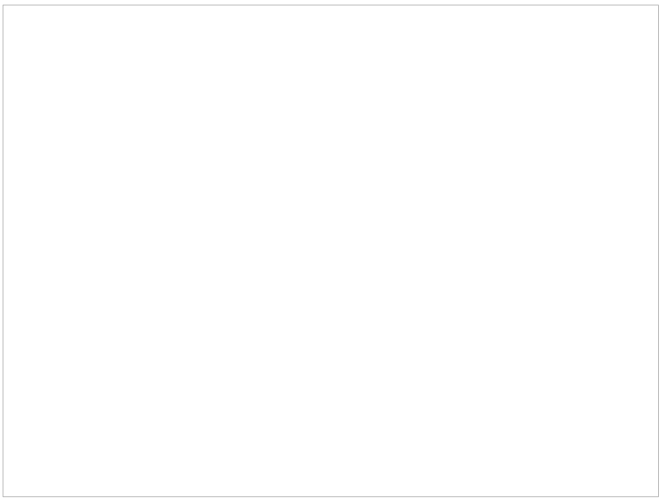
In the Landsat image, Klamath Lake lies near the bottom. The blue lake to the east is Summer Lake. The Great Sandy Desert (dark blue) is a series of volcanic flows that are roughly coeval with the Snake River Plains to the east.

When the writer (NMS) first saw this Landsat image, he was struck by the conspicuous large elliptical feature northeast of Crater Lake. It resembled a huge volcanic caldera. Visiting it in 1966 during a field conference on volcanism, I learned that the western part known as the Walker Rim was a cliff composed of volcanic flows. Solid evidence *for* a caldera was not obvious. This may be a coincidental artifact of several topographic features that because of vegetation distribution gives the impression of ellipticity.

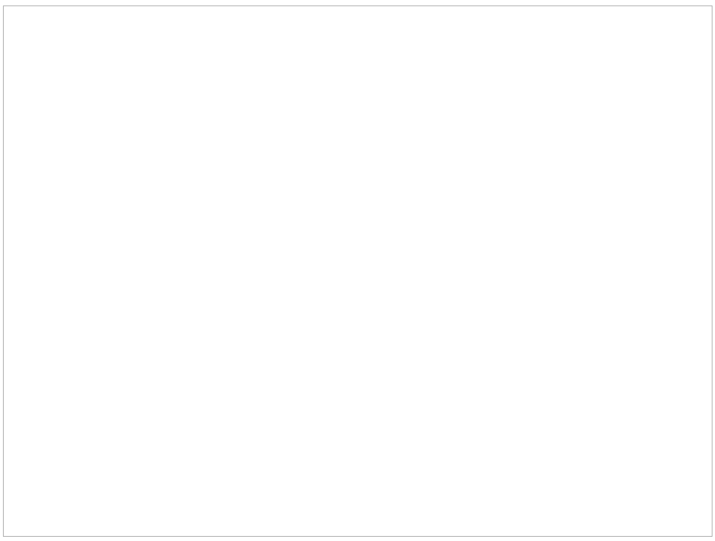
Portland, Oregon lies within the land between the confluence of the Willamette River and the Columbia River. Here this is seen in these two Landsat-7 scenes and then in a perspective view using a Landsat-5 (RGB = 542) subscene looking almost south.



Another view of Portland, OR from space.



The city is shown in this aerial oblique photo:



Looking east, Portland's skyline is beautifully enhanced by Oregon's tallest mountain, Mt. Hood, a stratovolcano:

Now let us look for a moment at Seattle, Washington, and the embayed land to its west, most hosting residential areas, as seen first in this ASTER image and then in this aerial photo of the waterfront looking northeast (note the Space Needle and Mount Rainier):

Panorama of Seattle.

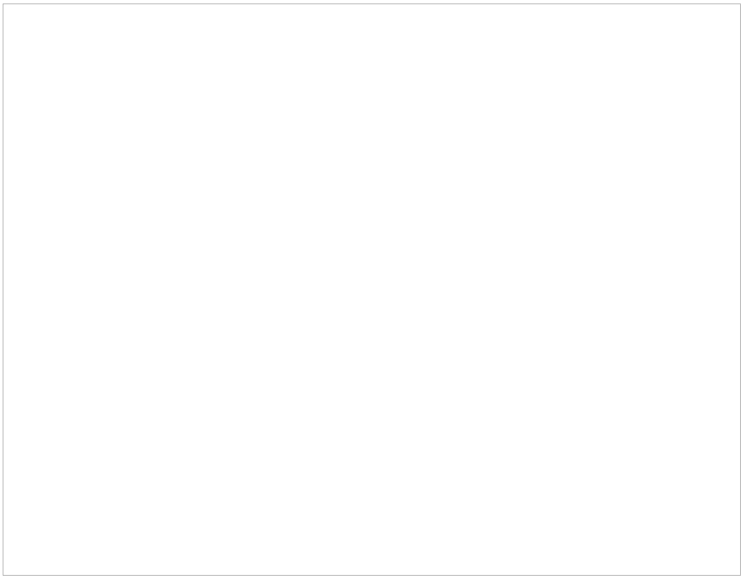
Courtesy: Carolina Map Distributors

West of Seattle is the Olympic Range, one of the most beautiful mountain groups in the 48 states. Here are a space image and a ground scene:

The Olympic Range from Space.

Ground view of the Olympic Range

East of Washington's Cascade Mountains lies a broad lowlands which includes the so-called Channeled Scablands (see page 17-4 for an explanation). These show up in the Landsat image as the bands of redder surface that have streamlike patterns. On the ground the scablands are dissected basalt plateau rocks and windblown loess:



Typical topography in the Scablands.

On the next page we will continue our odyssey of inspection of states adjacent to the Pacific coast, along with a look at several islands that are U.S. territories.

PREVIOUS NEXT

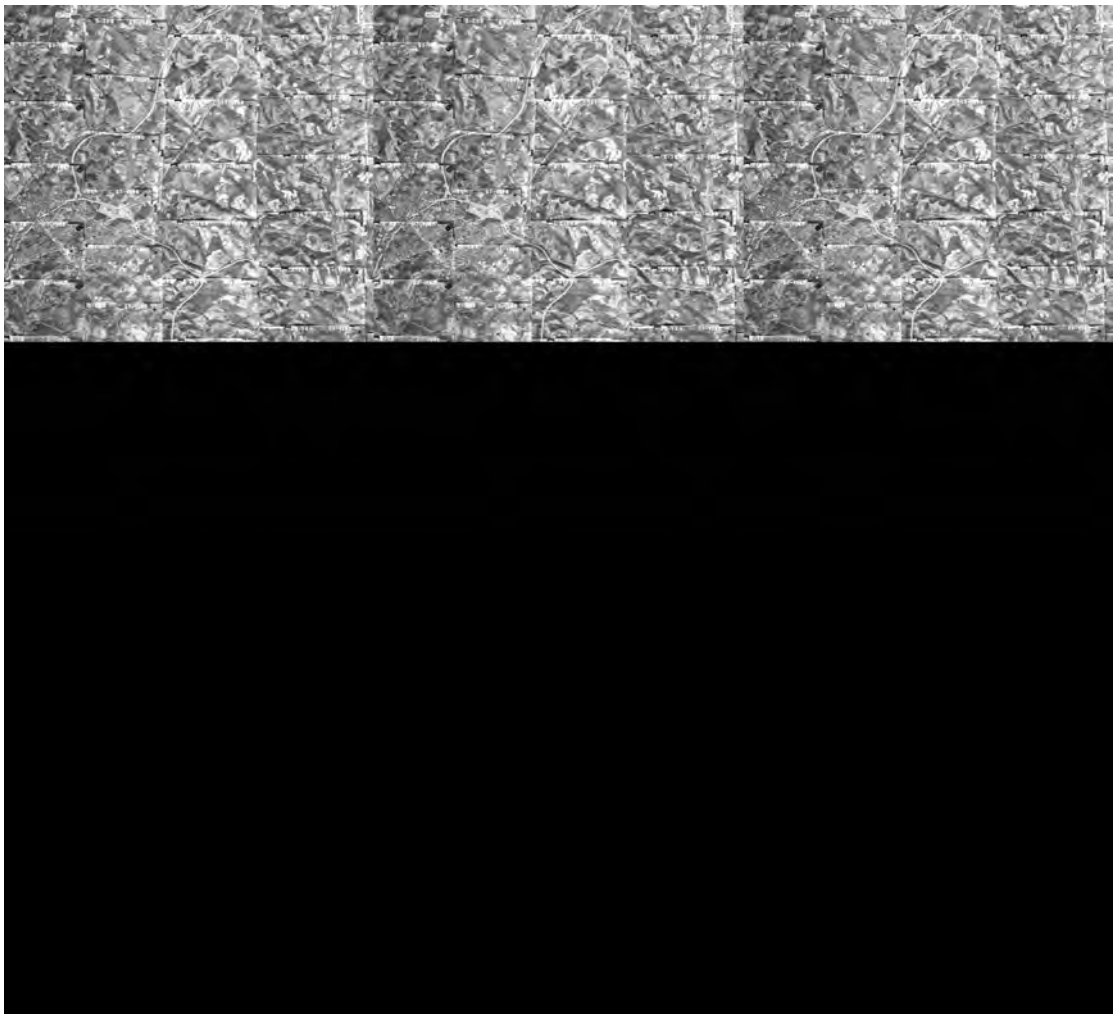
Primary Author: Nicholas M. Short, Sr.

Aerial photos have four principal uses: 1) to show what is in each photo; 2) to function as a visible base for mapping; 3) to serve as a mate for stereo viewing; and 4) to be joined together in mosaics that retain each photo's resolution (unless reduced in scale by being rephotographed) but cover a much larger area. This last use recasts a collection of contiguous aerial photos into a synoptic product that offers regional coverage. However, unless very careful photo processing is applied, the mosaics will likely show a patchy quality because of vignette darkening towards the edges (see Section 10). Space images, such as Landsat, SPOT, IRS, JERS, and others can also be combined into mosaics that can embrace very large areas, even to continental scale. Since the images were recorded in digital format, the data can be reprocessed to minimize and even eliminate the differences in tone, etc. from center outwards.

REGIONAL STUDIES: USE OF MOSAICS

How Mosaics are Made

The notion of getting a much larger picture of a region by pasting images of individual scenes into a single composite goes back to the early days of aerial The individual photos can be laid out side by side, commonly with some overlap, to construct mosaics. The process is a bit like jigsaw puzzles, except that we usually know the position of each air photo in advance. During an aerial mission (see page 10-1, which discusses aerial photography) the planners predetermine the flight lines for the airplane to follow, usually a back and forth pattern, much as a farmer plows a field. Here is a typical mosaic made without any attempt to compensate for the distortion in tonal balance and geometry that increase from each photo's center outward (the term "uncontrolled" applies to this lack of correction):



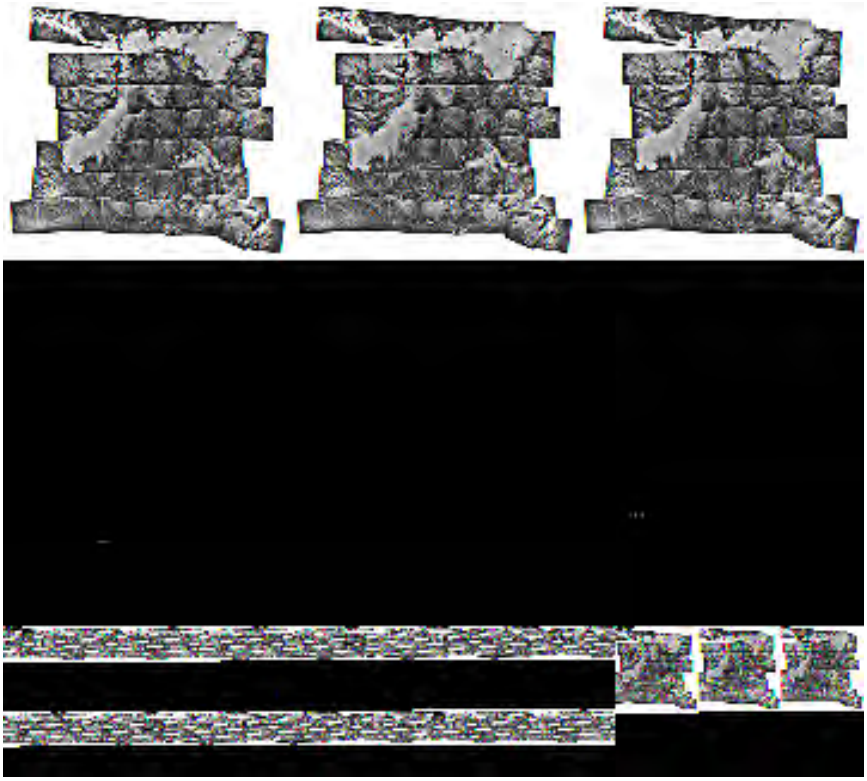
The aircraft is flown such that the aerial camera takes the photographs in sequence, so that there is always about 50% overlap (common area) between each successive picture, normally accomplished by an automated camera shutter at a timing interval controlled by altitude, air speed, and camera properties. On the next parallel line the pilot attempts to traverse the ground at a lateral distance that produces up to 40% sidelap. Both end-to-end and side overlaps allow for obtaining a three-dimensional or stereo effect (see page 11-3 for stereoscopic procedures and examples of 3-D views). However, the resulting pictures almost always display distortions, which comes from inexact navigation and aircraft wobble (in pitch, roll, and yaw) caused by turbulence.. There is also a notable distortion in an image outward from its center both because the camera is supposed to look vertically downward, but at times it's canted off center, and the geometry of displacement imparted to objects towards the outside of a photo related to the increasing slant distances from center to edge. This effect decreases with higher altitudes (thus is lessened in space images) and we can also adjust the focal length and other camera geometry to diminish it. Tonal brightness also can be discernibly lower from the interior of the photo towards its boundaries.

In practice, mosaics can be constructed from the central parts of the component photographs, from which the photospecialists trim enough to remove the more distorted, overlapping parts. Thus, they approximate rectified orthophotos. Large photo-mosaics have an added problem: the crews frequently fly the assigned flight lines on different days (or, less commonly, at different times of the same day), so that lighting and weather conditions often are not uniform. This problem is especially severe if the overflights happen weeks apart, in which case vegetation changes, along with the sun orientation due to seasonal shifts.. Photographic processing can compensate for some of these deviations, but uncontrolled (minimal adjustments) mosaics are typically rather patchy.

To retain maximum resolution, one must put together a photo-mosaic from individual scenes that we have not reduced in size. Consider making a mosaic from aerial photos having a scale of

1:62,500 (scale is discussed on page 10-1). Each photo, approximately 30 cm (12 in) on a side, covers almost 2.6 sq km (1 sq mi) of ground surface. At this scale, to create a mosaic representing ground dimensions of 6,400 sq km (2,500 sq mi), equivalent to 80 km (50 mi) on a side, we would need a "billboard" 15 m (50 ft) wide and high, requiring 2,500 photos, neglecting trimming requirements. This is impractical, so we almost invariably reduce mosaics in size and hence in scale and therefore, they have a notably lower resolution.

The scene below is a solid example of a typical *uncontrolled* mosaic. This is a series of high-altitude aerial photos (each about 18 km [12 mi] on a side) taken by NASA's U-2 aircraft along seven flight lines during late spring of 1972 in support of the writer's study of the geology of central Wyoming.



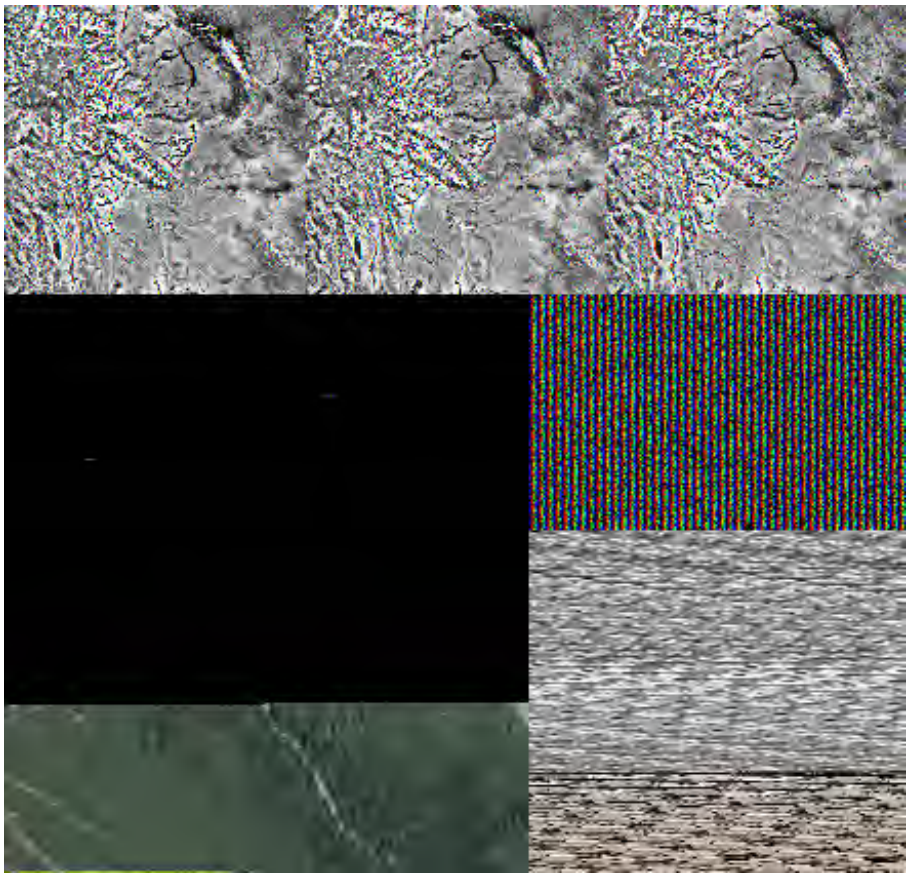
7-1: Evaluate the black and white aerial photo mosaic in terms of its properties and usability. ANSWER

Compare this scene, centered on the Wind River Basin, with this Landsat 1 Band 6 MSS scene that includes, and extends beyond, the mosaic. This was the prime image taken by the writer (NMS) into the field in central Wyoming just over a month after launch of ERTS-2 (to my knowledge I was the first to take any ERTS product to the area it covered so as to check on its usability):

7-2: What photo property does this ERTS image NOT have that is evident in the individual photos making up the U-2 mosaic ANSWER

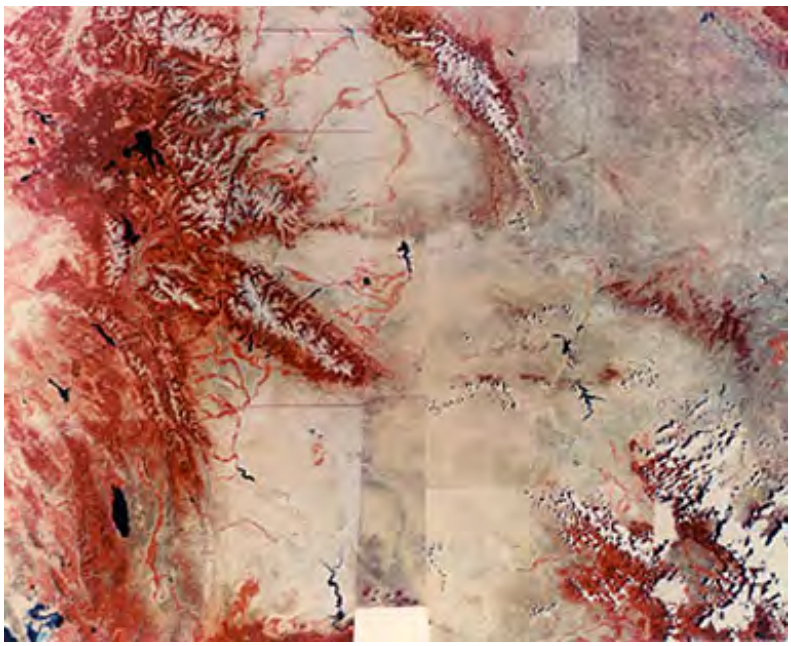
To match the image with the U-2 mosaic, look for the Boysen Reservoir and Ocean Lake (round) that stand out in the MSS view but are hard to see in the mosaic. In that mosaic, the Owl Creek Mountains, with partial snow cover, lie along the top; the Sweetwater River is at the bottom; a cloud bank appears in left center. When examined full size on a light table, this black and white scene shows more ground detail, even in that tonal mode, than does the Landsat full image but the even-toned nature of the latter compensates somewhat for the information quality.

Shortly after returning from the field in 1973, the writer (NMS) arranged with a support contractor, GE Space Sciences Lab in Beltsville, MD, to reprocess black and white images of western and central Wyoming and part of adjacent Utah to optimize contrast and then to match tonal levels at joins so as to produce an ERTS mosaic - possibly the first ever made with this quality of space imagery. The result was both pleasing to the eye and informative. Here is that product:



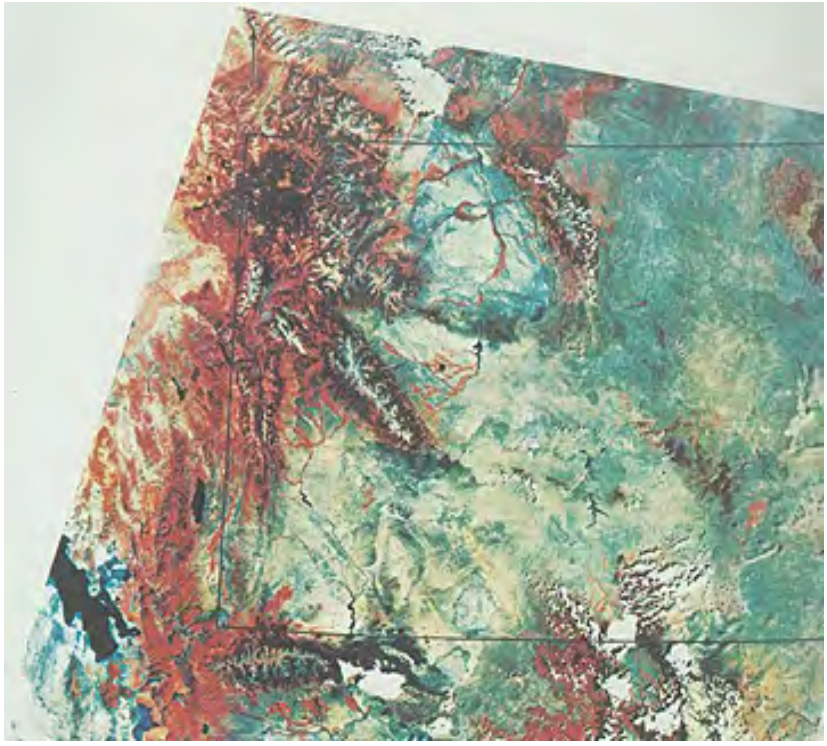
It's hard to see any join lines in this mosaic. Main reason: the lack of vignetting, i.e., tonal gradation to the edges, because of the high altitude of Landsat relative to its image base size means that the outer parts have nearly the same tonal balance as the inner.

To the writer's knowledge, the next image is the first color mosaic of a State in the U.S. - Wyoming - ever made, again at the GE facility. This was used as a base for data plotting in the Wyoming project he conducted with members of the Geology Department of the University of Wyoming:



7-3: Why is this mosaic still somewhat patchy? ANSWER

As more Landsat images were acquired, and GE honed its skills in processing them, a new mosaic of Wyoming and parts of Utah and Montana was produced:



We close this page with a bit of history: The first mosaics made from satellite imagery were of images taken by meteorological satellites. Here is a mosaic showing cloud cover over much of the Earth, laid out in a rectangular projection, using images acquired by the TIROS weather satellite:

Primary Author: Nicholas M. Short, Sr.

On this page we are introduced to three mosaics, each of different origin. The first is a controlled (carefully reprocessed to match edges) aerial mosaic of southern California. The second is part of the first mosaic of the U.S. made from Landsat Band 6 images. The third is made from digitized elevation data (DEM) that allows construction of an image product in which light and dark tones are assigned by artificial shadowing so that a quasi-three dimensional picture results. To illustrate the other extreme, an enlargement of a small part of a single Landsat image of Los Angeles is presented to indicate the versatility of space imagery for a wide range of scales.

Mosaics of Los Angeles, the Mojave Desert, and the Southwest U.S.

We next view one of the best aerial photomosaics this writer (NMS) has ever seen. It was made by Aero-Service, Incorporated, through a NASA contract, during the early days of Landsat when comparing space images and aerial counterparts was a point of interest.

The scene shows the Los Angeles basin, the Transverse Ranges (especially the San Andreas fault), and the adjacent segments of California, which is broadly the same "real estate" as the first image in Section 4, page 4-1. On that page, look at the Atlas map for part of the Los Angeles image as an aid to locating the geographic features, and note the roughly equivalent area as seen by radar, next to the map.

The black and white photomosaic, which includes the interior parts of about 8,000 photos, when reduced to a 23 x 28 cm (9 x 11 in) print, has moderately better resolution than the Thematic Mapper color composite, but the differences are not great. Because of excellent edge matching

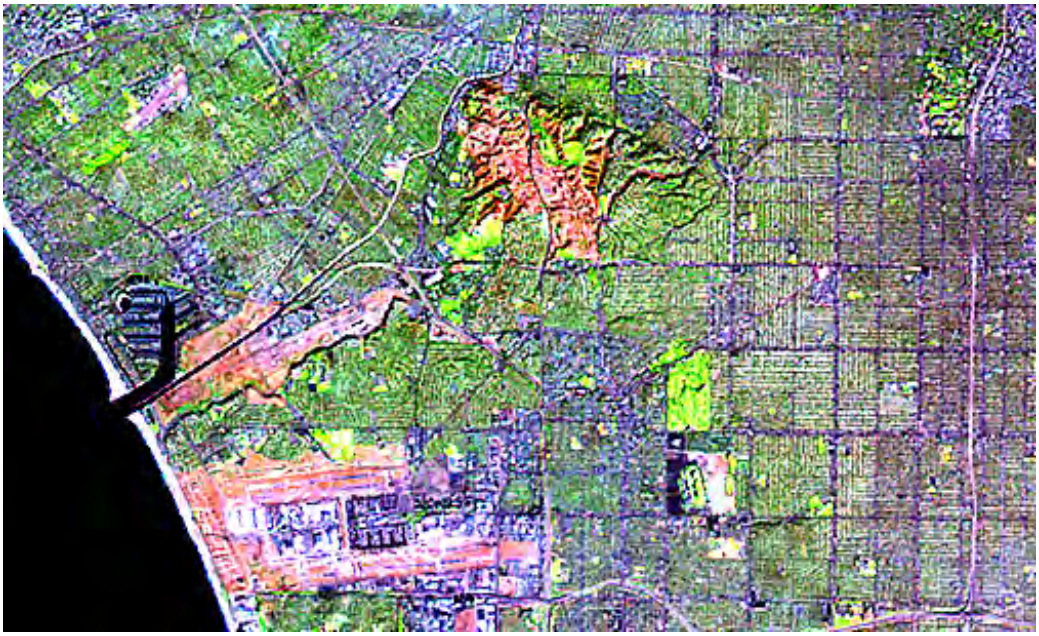
efforts, this mosaic shows little blockiness or variations in tone within each image component nor any discontinuities at the margin. In this sense, its internal purity approaches the even lighting that exists in a Landsat image, which takes about 28 seconds to scan. By varying contrast and other factors during image production, we can make a single Landsat image look similar to a photomosaic of this quality, but the latter is panchromatic (film sensitive to all visible colors). Thus it has a different mix of brightness' (tonal patterns) than that of the more narrow spectral region in a black and white image from a particular Landsat spectral band. We can approximate a panchromatic film image by calculating the first principal component from the multiple bands in a Landsat scene and then writing this on a film recorder.

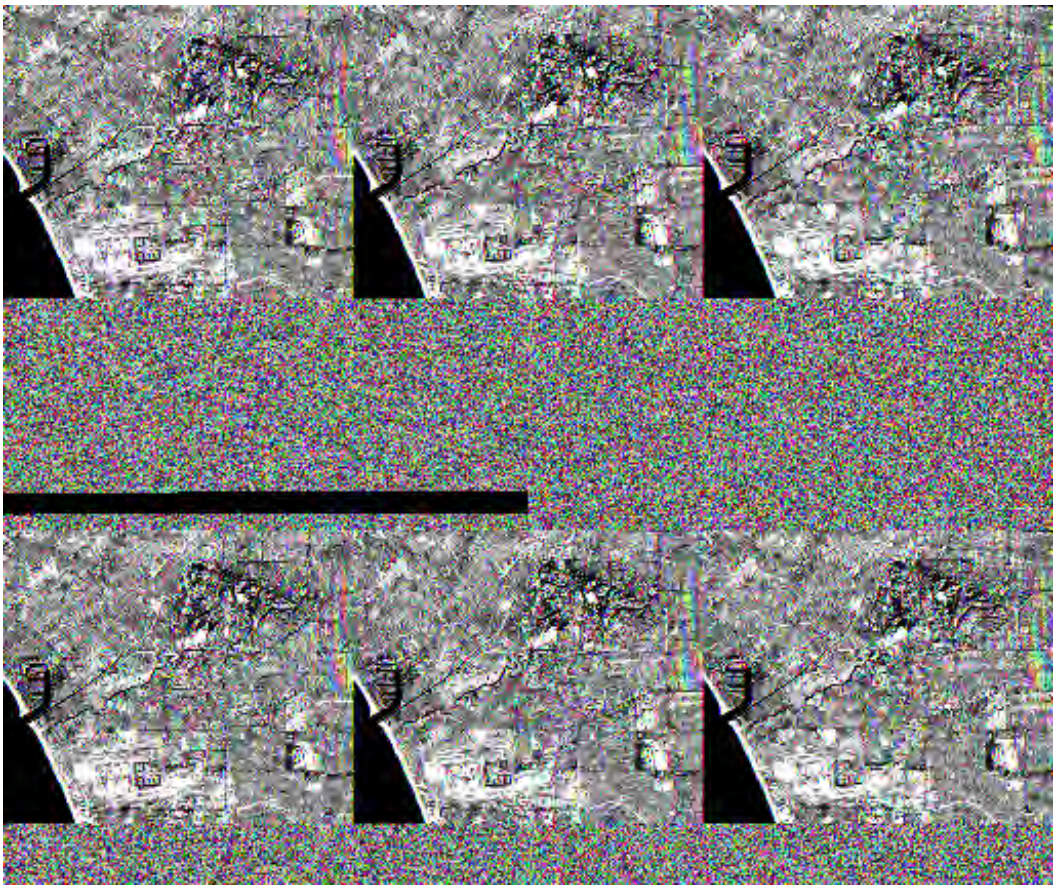
7-4: Comment on any flaws or evidence of the individual photos used to make this mosaic. ANSWER

We can easily make space mosaics from Landsat (or SPOT or other imaging satellite systems) inputs. A sequence of space images is continuous and has no overlap. Generally, the next orbital line, obtained at roughly the same time the next day, gives rise to about 10% sidelap at the equator, increasing to about 40% near the poles because of longitudinal convergence. The obvious problem in constructing a mosaic that is several orbital paths wide is that the space paths are days apart, so that the likelihood of constant weather conditions (especially minimal cloud cover) goes down as the size of the mosaic increases. We can lessen this problem by using images with few clouds taken at different times of year, but then sun angle, vegetation state, and other factors affect the picture. So, good scene matches over disparate dates are unlikely. But, since Landsat has operated for many years, the probability goes up of getting cloud-free images at generally the same time of year over a large area. However, unlike film-based primary imagery (aerial photos), the digital format for the Landsat (SPOT, etc.) images permits a number of cosmetic enhancements by computer processing of the raw data, so that we can virtually eliminate variations in natural lighting conditions inherent to Landsat imagery acquired on different dates.

Enlarged Subscene of Los Angeles

Let's digress for a moment and consider the other extreme, the versatility of Landsat subscene images when *enlarged* to show localized details. This is evidenced by this next pair of images made from (top) a 30 m Landsat 7 ETM of west-central Los Angeles and (bottom) a 15 m Landsat 7 panchromatic image covering part of the upper image.





7-5: What do you think (this is subjective) is the most obvious spatial difference between the upper and lower ETM images? ANSWER

The degree to which we can extract useful information from images at 30 m resolution became apparent in a situation experienced by this writer. In an enlargement of a Thematic Mapper scene containing Baltimore, Maryland, and Washington, District of Columbia, from which he produced a subset around Baltimore, he spotted the backyard of his home near Columbia, MD, because of its well-watered lawn (strong IR reflectance) relative to those of neighbors on either side whose uncared-for grasses had turned brown in the summer heat. He noticed a small black spot to the south in the image which he surmised was the apron of a swimming pool at a neighbor's home three doors down, not visible to him because of a fence. When he asked the neighbor if the pool had an asphalt apron, the "yes" answer convinced him of the power of TM imagery to invade the "hidden" for arcane information. Still, the Landsat TM resolution is a far cry from that of some military spy systems, which, it is rumored, can read license plates and locate golf balls on a course. And, as we have seen already, now commercial satellites are supplying imagery with better than a meter spatial resolution.

MSS Mosaic from San Francisco to San Diego

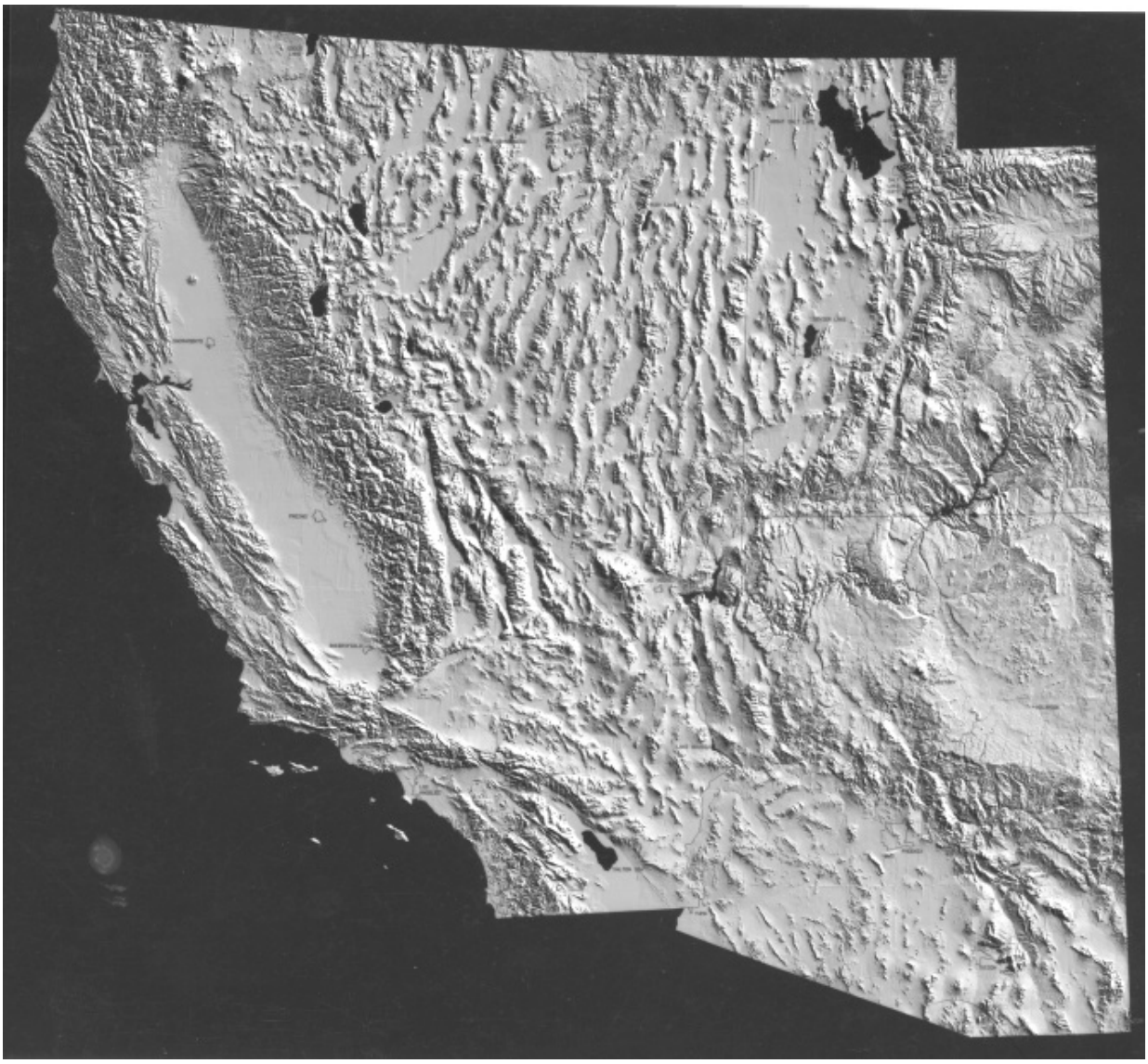
As a demonstration project during the first year of Landsat operations, the feasibility of making composite images covering large areas was established in this SCS Landat mosaic made from 22 cloud-free images selected from five orbital paths, that includes all of California to north of San Francisco, a part of southern Nevada, and a bit of Baja California in Mexico.

7-6: Can you find at least two parts of this mosaic in which the individual Landsat images don't quite fit tonally. How would this space image mosaic differ from the aerial mosaic above? ANSWER

Digital Elevation Models: Shaded Relief Image of CA, NV, AZ, UT

Digital Elevation Models (DEM) are digitized values of elevations (Z) at specified points (X and Y) in an array of X and Y coordinates (tied to some projection system, such as Universal Transverse Mercator). The values are codified by some DN system and are arranged in raster format. They can be used to construct topographic maps (contoured; see page 11-5 for a broader discussion of DEM uses) or shaded relief maps or other modes of showing surfaces. A good Internet review of DEMs is offered by the Softwright Company.

Compare the above mosaic with a shaded relief, topographic image produced from Digital Elevation Model (DEM) data using a northeast direction of illumination at 30° inclination from the horizontal.



This DEM image has a scale of approximately 1:7,600,000 and a resolution of 305 m (1,000 ft), and depicts the entire states of California, Nevada, Utah and Arizona. The DEM image, with its controlled artificial shadowing, graphically shows the structural expressions of mountain-scale landforms in the physiographic provinces described on page 6-1. These provinces include elements of the western Rocky Mountains, Basin and Range, Colorado Plateau, Sierra Nevada and California Coast Ranges. Note also the flatness of valley floors corresponding to the Basins and to the Central Valley. This digital rendition also highlights the linear character of the San Andreas fault. Topographic relief is less obvious in the Landsat mosaic, but the structural features emphasized in the DEM image generally have visual counterparts in the mosaic, because elevated terrain tends to be vegetated (hence, dark in this red-band version). Thus, we see that tonal contrast substitutes for the shadowing effect in the DEM version. Areas used for agriculture (e.g., the Central Valley and Imperial Valley) show a distinctive light-dark mottled pattern. All such land use patterns are absent in the DEM version which purports only to show raw topography. At the scales and resolutions of both image mosaics, it is difficult to see any tonal or geometric patterns that would disclose the large concentrations of population. However, an alien observer from outer space presented with the Landsat image might guess at the existence of intelligent beings from the field patterns of growing crops.

7-7: Why does this DEM image appear tonally different compared, say, to the Landsat black and white mosaic? ANSWER

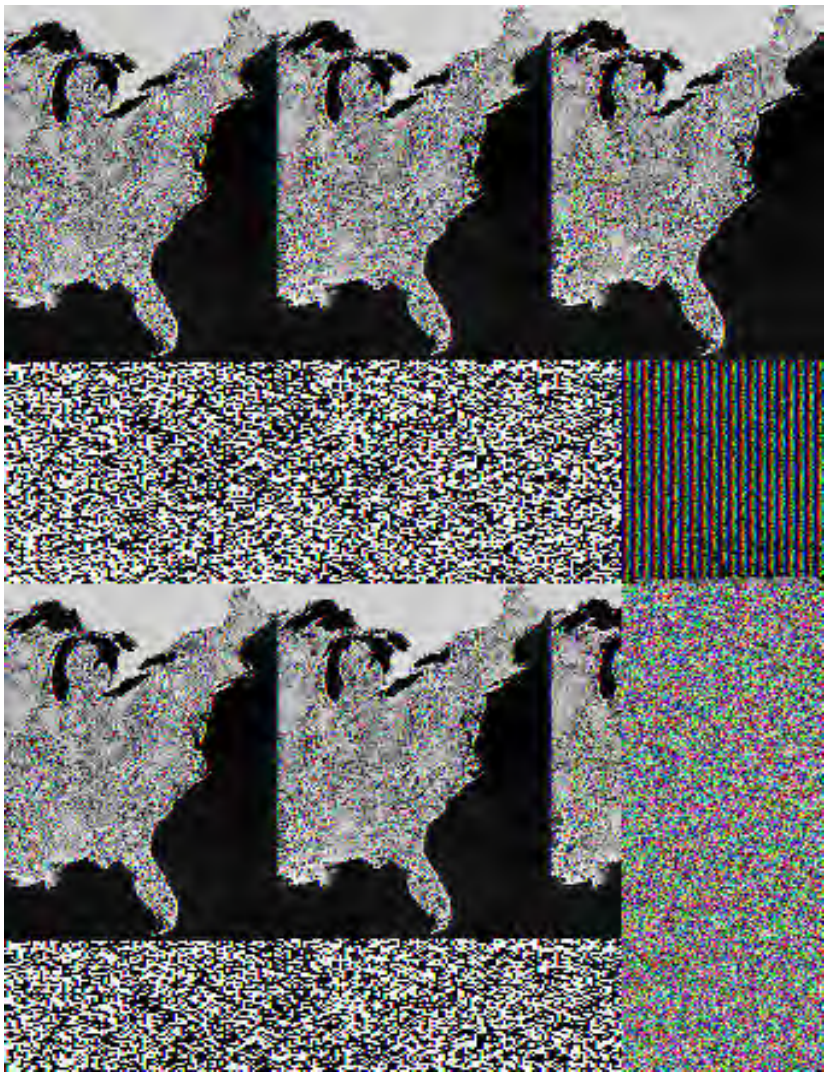
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

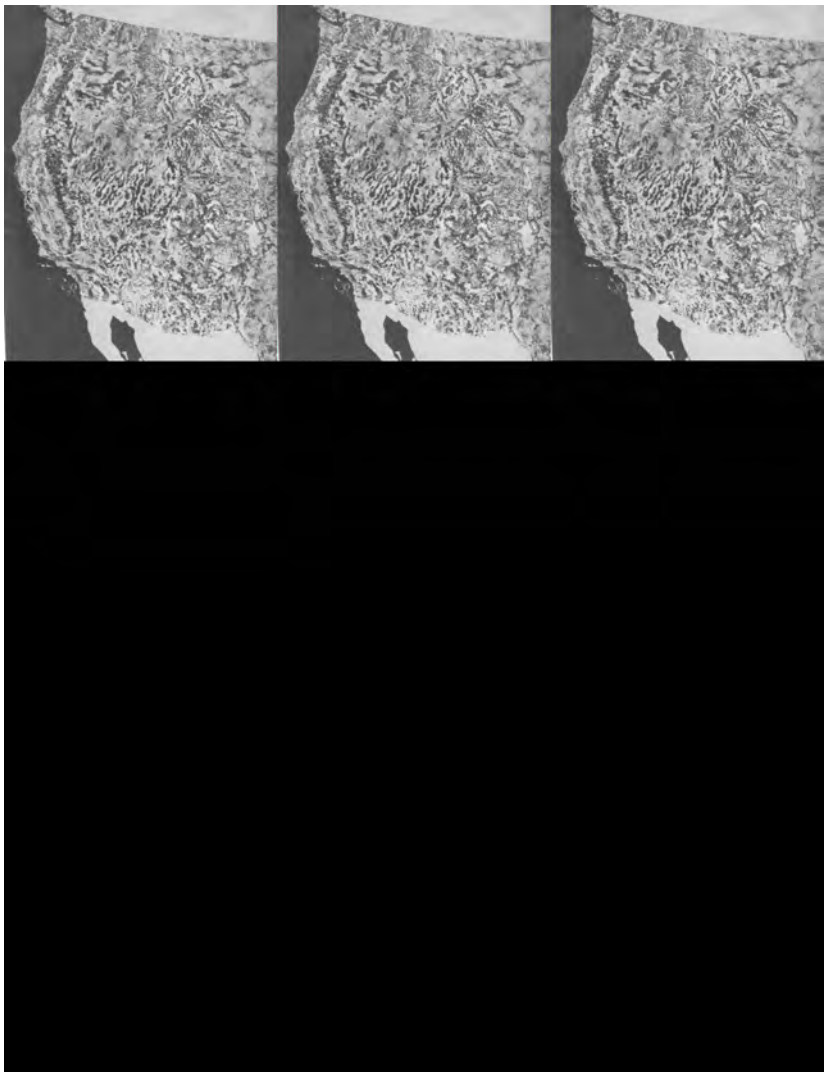
Additional mosaics of the U.S., Alaska, and regions in Europe and Asia demonstrate the value of combining space images to present views of large segments of the Earth's continental surfaces. A series of images shows Afghanistan and its neighbors - the focus of attention in the first year of the War on Terrorism.

MSS Mosaics of the United States; Alaska and Mexico; International Mosaics and large scenes including Afghanistan

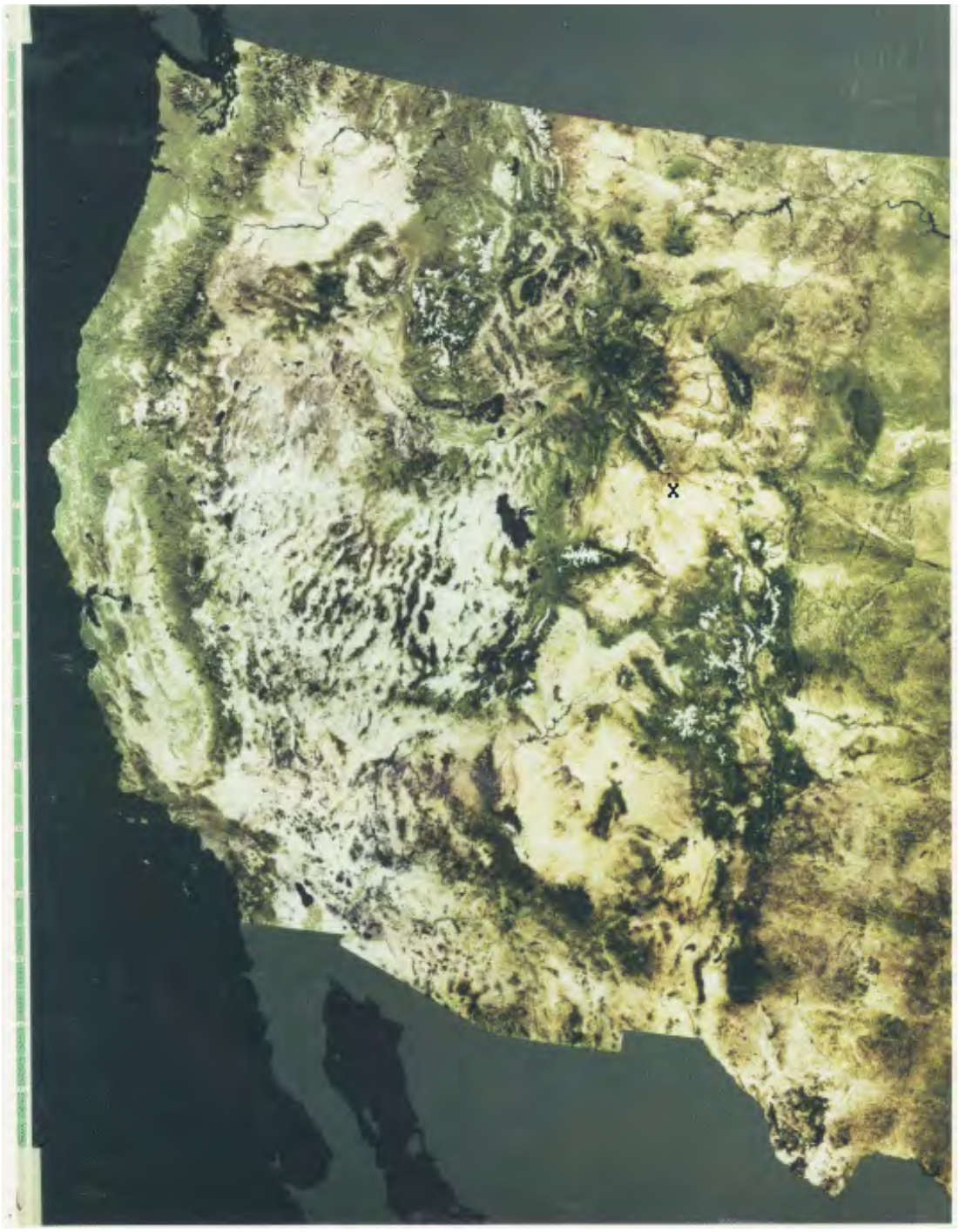
Mosaics of entire continents or subcontinents became plausible when enough cloud-free Landsat images were acquired and reprocessed (to minimize tonal differences) The U.S. Department of Agriculture's Soil Conservation Service (SCS) put together the first Landsat "image" of the entire United States in 1974, as part of a NASA contract to prepare the mosaic in time for the U.S. Bicentennial year (1976). The idea for this was conceived and initiated by the writer (NMS). The contract work was monitored by his colleague, Arthur Anderson. Using 595 Band 5 black and white images, the result was the first higher resolution space portrait of the 48 States (Alaska and Hawaii were also done). The original photo version was many feet in dimension; what appears here has been greatly reduced in size, so that much detail present in the original is lost. Appearing here first is the eastern half of the United States; the version shown is somewhat degraded because the only copy available to the writer is the inside fly leaf of the book, Mission to Earth.



The western half (from the Rockies to the West Coast) appears better in part because of the sharper contrast in black and white (resulting from the dark appearance of vegetated surfaces [Band 5 was used] and the light tones in the deserts and lowlands):



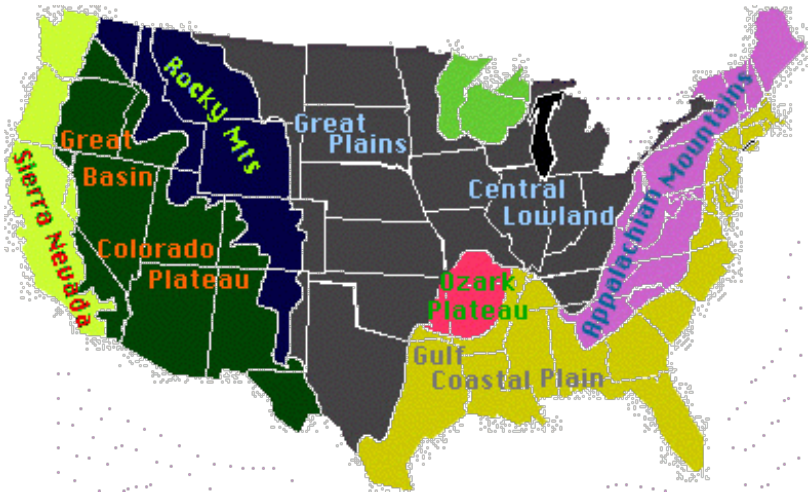
The General Electric Company produced the first false color rendition for NASA and the National Geographic Society (NGS). We reproduce here a portion of the NGS Magazine's Landsat "map" of the U.S., which was part of its Bicentennial (1976) anniversary issue, in the natural color version.



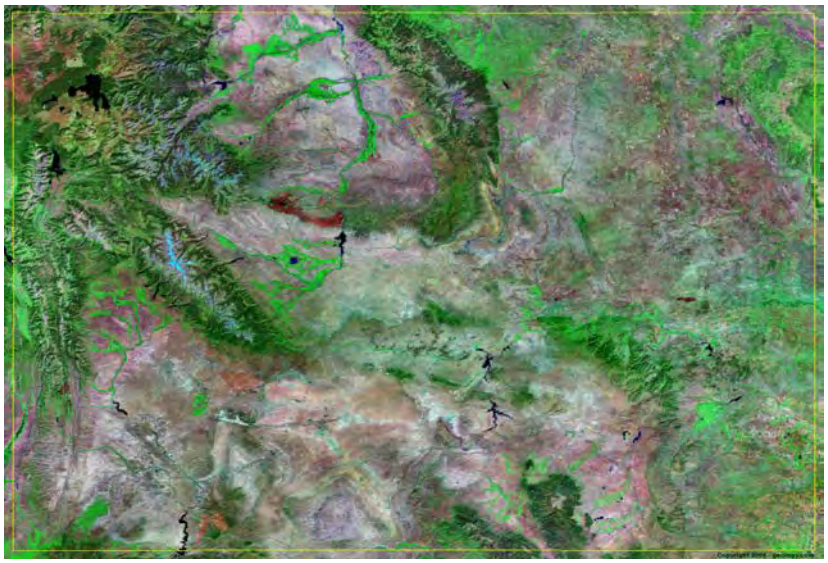
In making this mosaic for NGC (there is also a false color red version), GE projected the MSS Band 7 through a green filter and Bands 4 and 5 through blue and red filters). It shows all of the western U.S. into the Great Plains beyond Denver and the Black Hills. Blackish areas on land are either water bodies (such as the Great Salt Lake) or vegetated mountain slopes. Obviously, relatively little detail in the landscape is visible at this scale (~1:11,000,000) but the impression of the overall geomorphology and geologic structures that set the western U.S. apart is striking (more so when the full size - 14 ft wide - is viewed). Compare this rendition with the equivalent area in the NOAA Advanced Very High Resolution Radiometer image presented on page 6-1.

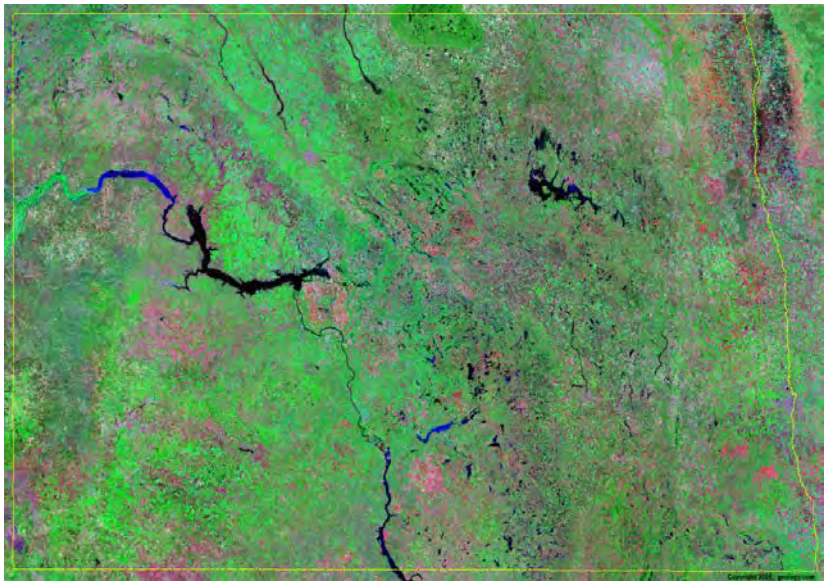
7-8: While this is a good natural color approximation, something does not ring true. Why; what's missing? ANSWER

We showed a colorful false color mosaic of the entire U.S. on page 6-1. Here is another one made from Landsat imagery that shows the U.S., Mexico, and part of Lower Canada in approximate natural color. The projection appears to a conical one, probably a modified Mercator. Look at this image and then try to pick out the major geographic subdivisions of the U.S. using the figure below as a guide.



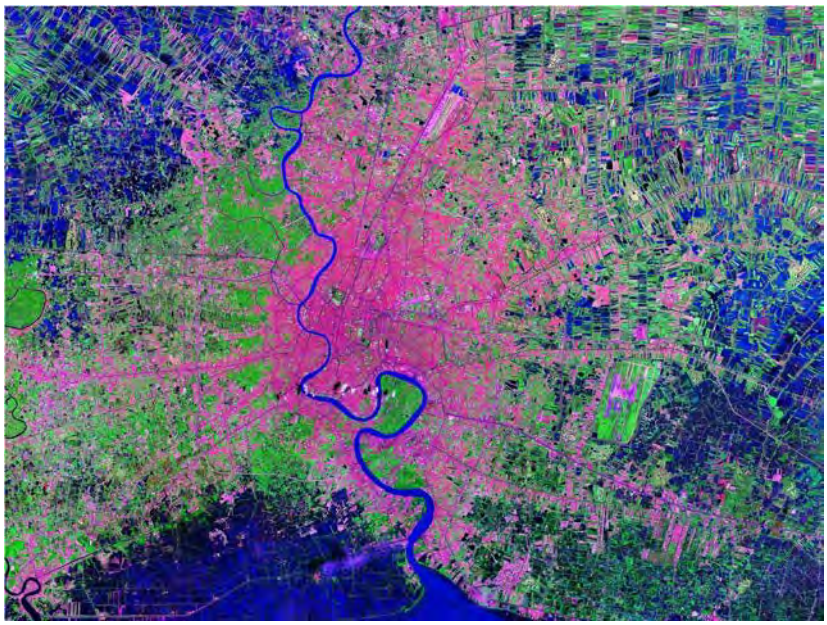
Mosaics of each of the 50 states of the United States and of countries across the world can be found on the Internet. The best known, and most reliable, source is geology.com. Mosaics from that source of two U.S. states are shown here, followed by a comment:





The writer's comment is this: The makers of these mosaics, which use Landsat images as inputs, have arbitrarily selected a color scheme which overemphasizes greens and reds at the expense of other colors. In other words, they are not anywhere close to true color renditions. Green represents vegetation, red represents non-vegetation (blue is used for water). While "form" detail is fairly good, class detail is not. After awhile, this color scheme gets boring. Compare these to other mosaics in this Section, and draw your own conclusions.

To make that point even further, look at geology.com's rendition of cities, here Bangkok, Thailand. The urban areas are uniformly red, active vegetation is green, and the blue represents, in this case, rice fields that are kept submerged in water. /p>



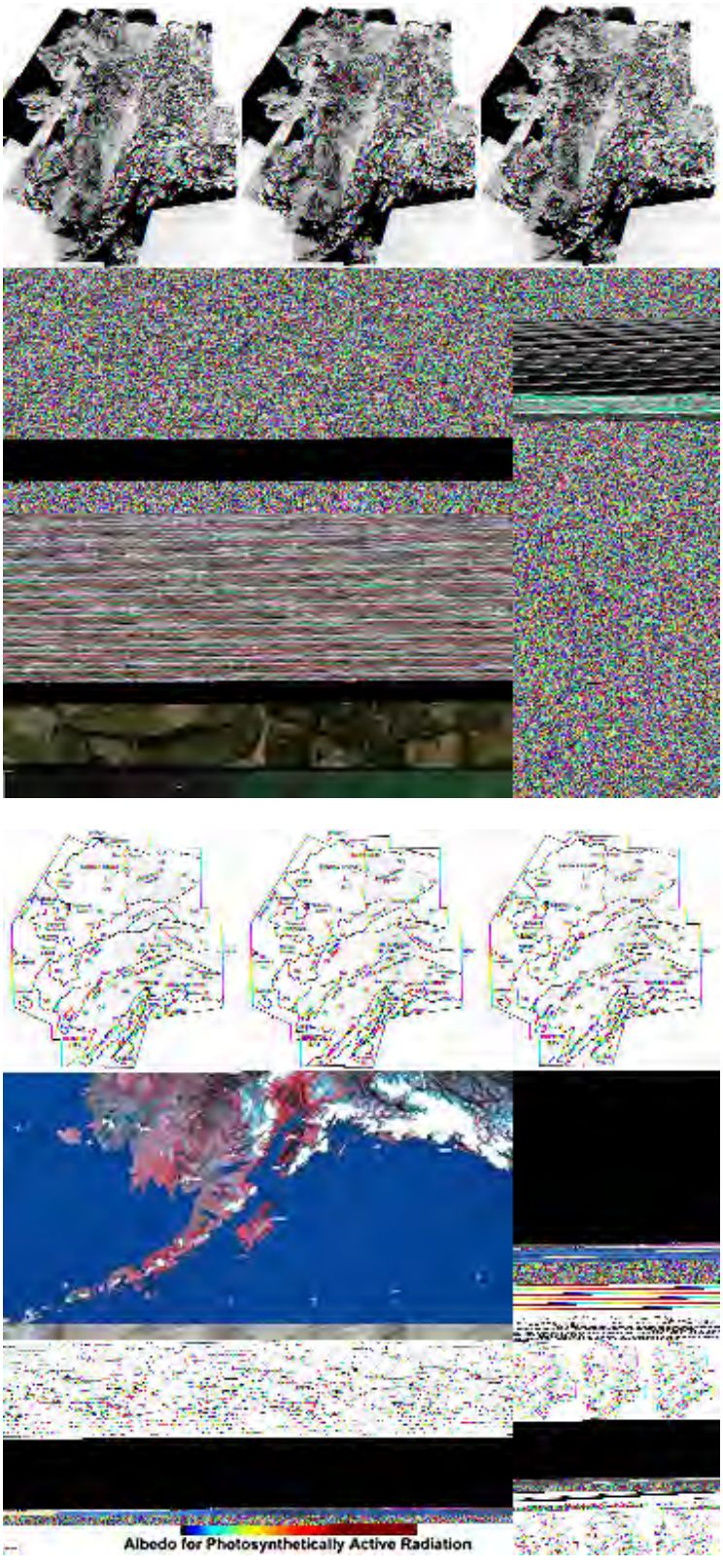
These two Landsat TM images of the same general area show how the color assignments can provide either a more pleasing and/or a more informative rendition:

This critique is somewhat arbitrary. The geology.com version does disclose two important pieces

of information: It better separates metropolitan from rural and it does reveal the condition of the rice paddies (which may be seasonal). Interpretations can be subjective.

Alaska and Mexico Mosaics

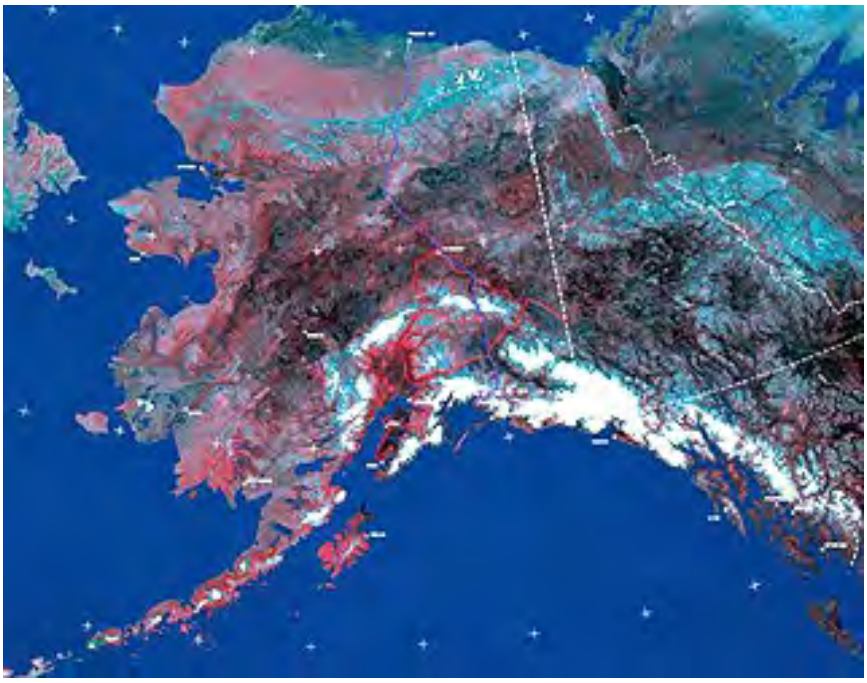
After the U.S. black and white mosaic was made, someone realized that the states of Hawaii and Alaska had been overlooked. An effort to correct this soon produced the missing states. Hawaii was easy. The Alaska mosaic was more problematic as at the time there was a missing strip of coverage and the western Aleutians were also not covered. Still, the mosaic shown here is imposing; a map of its main features is placed below it:



Notice on the map a number of symbols, e.g., OS. These refer to the generalized locations of

different structural terranes - great blocks of crust that at different times in the distant past crashed into that part of the North American plate (both before and after its split from Pangaea) and thus built up the region by accretion. The concept of terranes is discussed in much more detail in Section 17 (second half). For now, we will just identify these terranes by name: Agm = Angayucham; Ch = Chugach; End = Endicott; Kyk = Koyukuk; NxF = Nixon Fort; OS = Overlap Sequence; Pn = Peninsular; Por = Porcupine; PW = Prince William; Sew = Seward; Tog = Togiak; Toz = Tozina; Wr = Warngellia; YT = Yukon/Tinana; Yuk = Yukon. Major faults are numbered.

On page 6-9 we saw a false color Landsat mosaic of all of Alaska except for the outer Aleutians. This is reproduced here:



7-9: What does this mosaic tell you that is absent in the Soil Conservation Service mosaic ANSWER

A standard false color mosaic of Mexico, extending into parts of the southern United States, appears thusly:

South America

Mosaics have now been produced for entire continents. North America has been constructed from Landsat images, and several images of the United States appear in this Tutorial. Using AVHRR data (see page 14-2), all of South America has been mosaicked; the Andes and the Amazon in particular stand out:

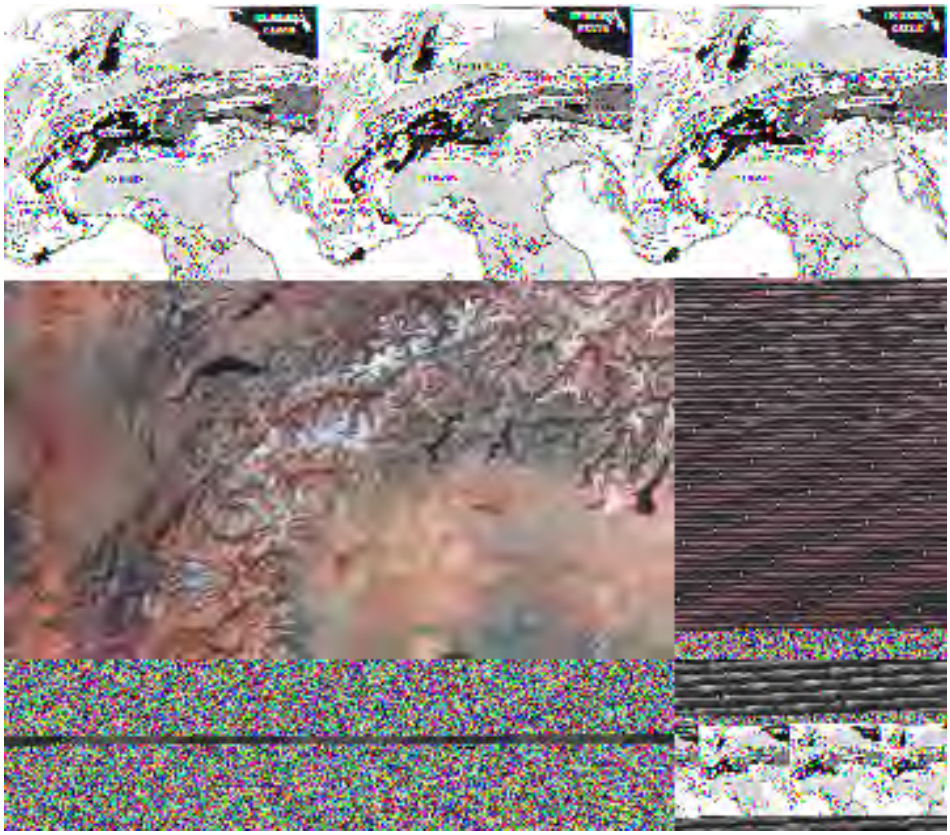
7-10: Can you pick out the general regions in this mosaic that correspond to the Amazon rain forest; how did you do this? ANSWER

Color Mosaics of Europe: The Alps and Italy.

Since the advent of high resolution space imagery, beginning with Landsat, mosaics of entire continents have appeared, using various satellites as input. A search through the Internet shows that Europe has a number of high quality mosaics. Here are two (note that much of European Russia is included):

Another magnificent mosaic, made as a variant of true color, shows the entire Alps (Alpine Mountain Chain) across southern Europe, including all of Switzerland and Austria, and parts of eastern France, southern Germany, Czechoslovakia, Slovenia, Croatia, and northern Italy. Westermann Satellite Maps prepared this mosaic, which, shortly after its production, became a standard aid for teaching geography to students in many European countries.

Below is a schematic map that shows most of the same area as in the Alps mosaic; this map defines the geologic/geomorphic segments of the Alps as well as for adjacent terrains. Try to match the main units to their topographic expression in the mosaic.



7-11: Locate the Apennines on the mosaic, using the map. Develop a single item argument that supports this statement: The Apennines are not nearly as high as the Alps? ANSWER

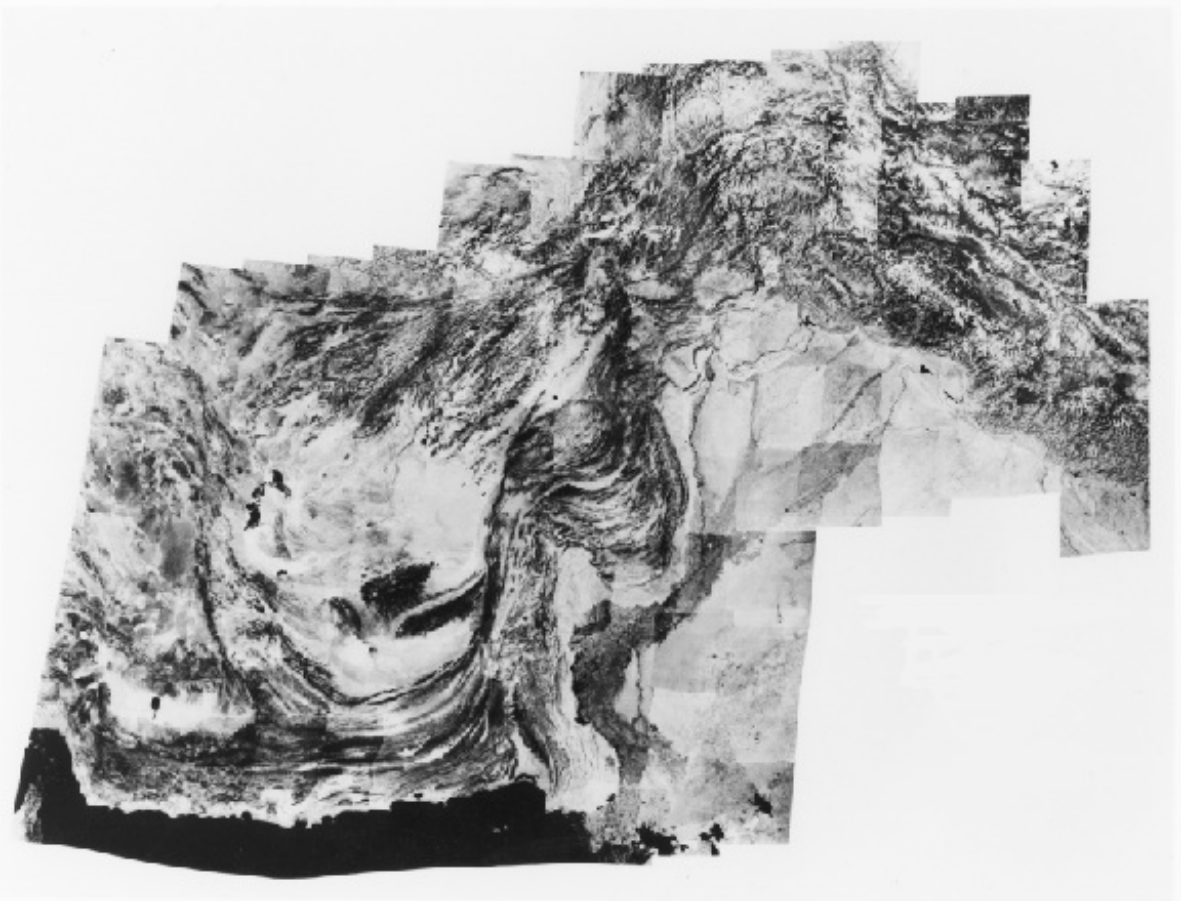
While we are focused on the Alps, compare this next scene, made from Landsat-7 images, of a part of the Alps by fitting it in to the above mosaic.

This next mosaic shows all of the Italian "boot", plus the islands of Sicily, Sardinia, and Corsica, and extends into the Alps:

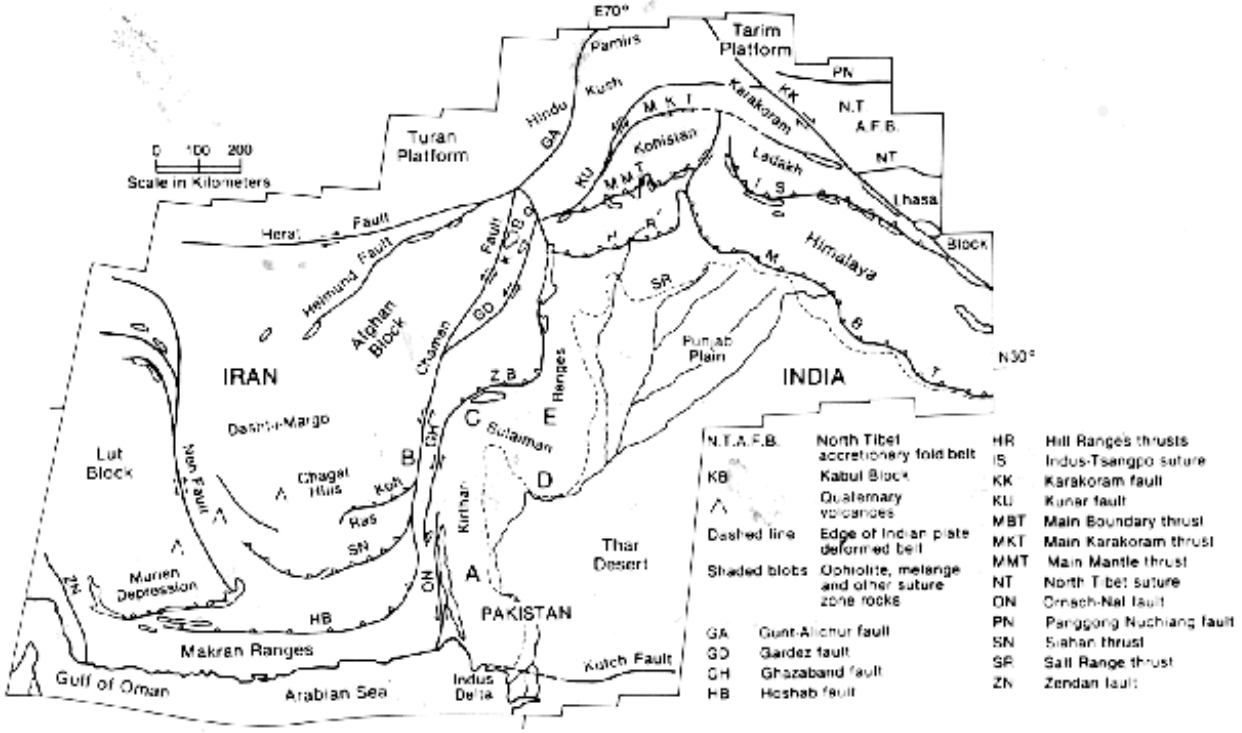
Elsewhere in the Tutorial (page6-13), we showed a color mosaic of all of Africa. This is repeated here as another example of the now sophisticated practice of continental-scale mosaicking:

Mosaics of Iran, Pakistan, India, and Afghanistan

We look next at one of the most sensational mosaics yet made from Landsat images, assembled by R.D. Lawrence and R.O. Rogers of the Geology Department at Oregon State University.



The region displayed in this black and white, red band (Band 5) mosaic includes nearly all of Pakistan, part of eastern Iran and a bit of southern Afghanistan, and a smaller segment of India, extending over about 2,100 km (1,305 mi) from west to east. The writer (NMS) has sketched in the principal structural features on the locator map below.



The continuous mountain belt starting with the east-west Makran Ranges, then the north-trending Kirthar and Sulaiman Ranges in Pakistan, and the great bend of the Hindu Kush joining the north-west-trending western Himalayas, is the consequence of the monstrous, still continuing, collision of the Indian subcontinent during the last 130 million years as it drifted north, striking the underbelly of Asia along a vast subduction zone.

7-12: This question is a departure into the realm of ancient history in which the Landsat mosaic helps to show what the problems were in the events that led to the demise of the Macedonian Alexander the Great. Maybe you know the story; if not, consult an encyclopedia for a synopsis of his life. What was the situation he faced, and how did the imagery help you to appreciate its gravity? ANSWER

The region just to the north of the top center of this mosaic has became famous - and infamous - to Americans since September 11, 2001. The country of Afghanistan is in many ways similar to Pakistan, with desert, a group of central mountains, and mountains on its southeast border with Pakistan. Afghanistan lies within this 4 km resolution metsat AVHRR image of the broad region that includes Saudi Arabia, Iran, Pakistan, Afghanistan, and the several "Stans" such as Turkestan - countries once part of the Soviet Union

This next view zeroes in on the region that includes all of Afghanistan.

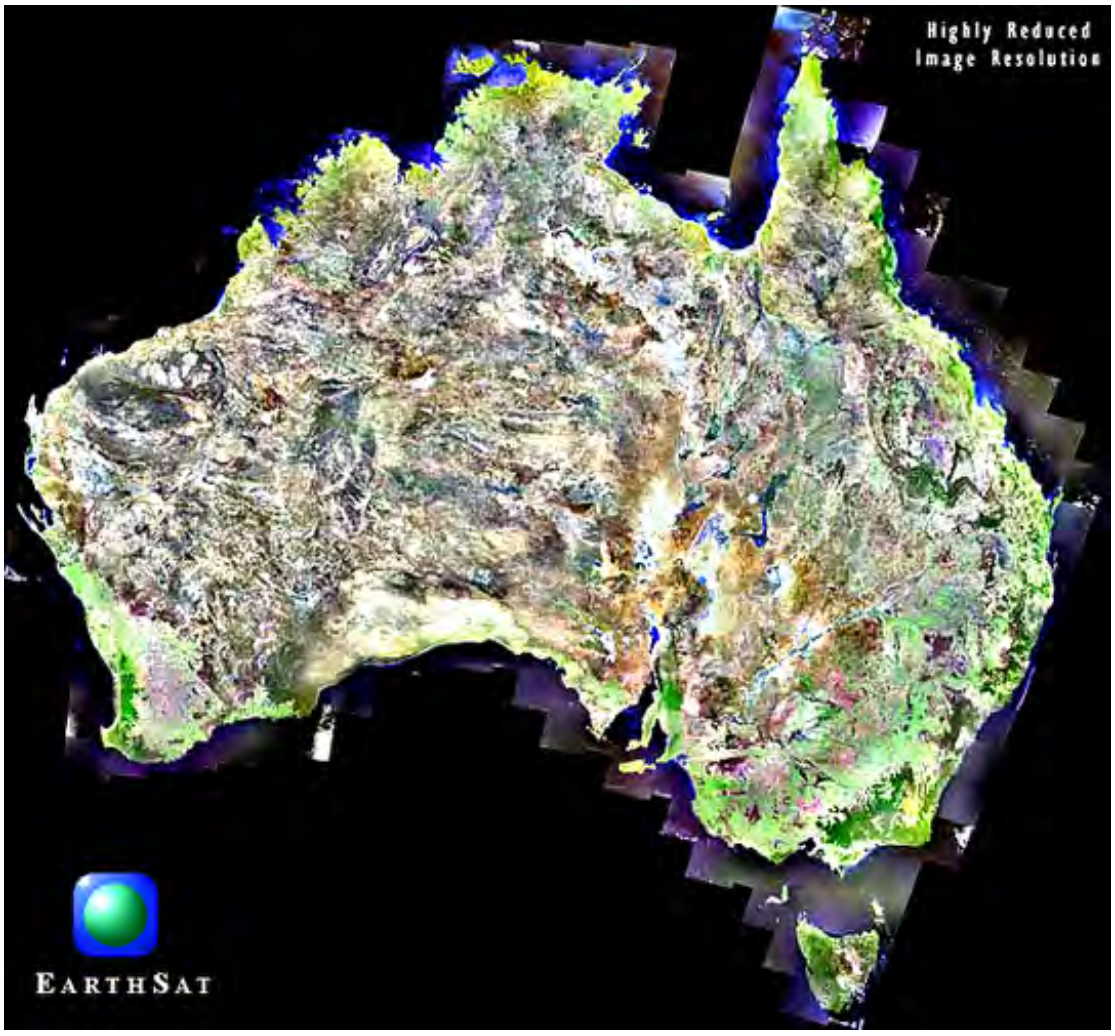
Compare the Metsat image with this mosaic made from Landsat images:

7-13: In the above images, try to find the area encompassing the Afghanistan-Pakistan border where Osama bin Laden and his Al-Queda terrorists are believed to be hiding since the 2002 U.S.-led campaign to overthrow the Taliban. ANSWER

The mountain terrain in central Afghanistan is strongly depicted in this reprocessed single Landsat image:

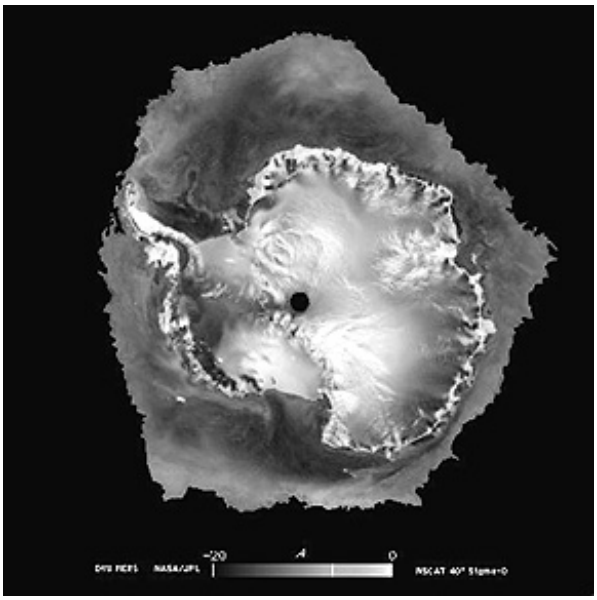
Just to its northeast is this Landsat subimage that includes Kabul and several of the Afghan towns much in the news since the War on Terrorism:

We turn next to one of the most stunning ever produced: the entire continent of Australia. Landsat images were carefully reprocessed and joined. The writer has seen a large copy of the original, on the wall of the lobby of Building 33 at Goddard Space Flight Center. He asked John Everett of EarthSat Corp. in Rockville, MD (which produced this product) for permission to use in this Tutorial. Here it is, displayed somewhat larger than normal (but please note that the original gives a much better impression of details):



The Antarctic and the Arctic

The entire Antarctic continent and surrounding ice shelves has been imaged from sets of radar images acquired over six days in September of 1996 by the NASA Scatterometer (NSCAT), as shown below, with the South Pole appearing as a drawn-in black dot:

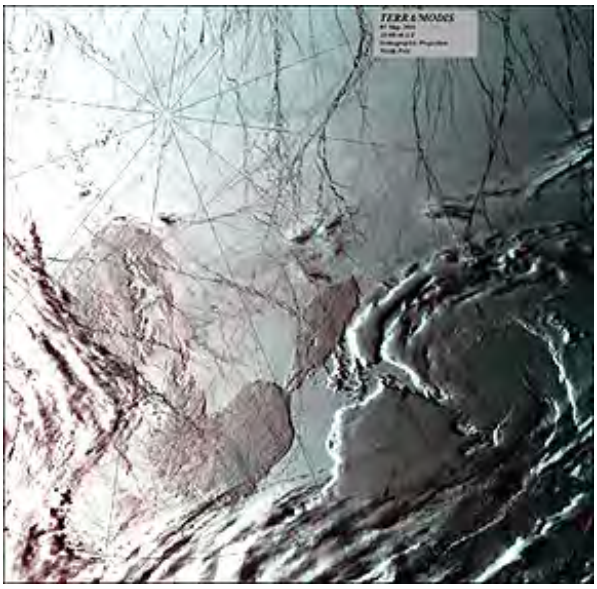


This whole continent, about the size of the U.S. and Canada combined, has been imaged in color by specially processing AVHRR scenes acquired by meteorological satellites. This mosaicked version, produced by Pat Chavez's image processiong group at the U.S. Geological Survey, is shown here in its annotated form:

It is a bit of a surprise to those (most of us) who know little by the Antarctic continent to learn that it has high mountains. The high point of the continent, the Vinson massif, at 4897 m (16067 ft) in the Ellesmere Range, has been climbed. Here it is in an ASTER image.

Turning to the other extreme, here is a very high resolution (2.4 m) of the U.S. Antarctic Research Base at McMurdo Sound (bottom center of the above mosaic) as imaged by the QuickBird satellite (see Overview):

The MODIS instrument on Terra (page 16-9) has obtained a view of the North Pole (in the image below the longitudinal lines converging at it are added artificially), even though it does not pass directly over it (in the image below it has looked slightly sidewards). The polar Arctic is not a land mass but a sea covered with ice. While the scene roughly resembles the ice sheet covering central Antarctica, the give-away that the Arctic has sea ice is the sets of leads (fractures) that result from ice cover breakup and movement.



The ability to produce mosaics from high resolution imagery such as Landsat opens a new dimension in the study of Earth's land masses at continental and regional scales.

Displays on your computer screen lose much of the detail evident at the resolution in the Building 33 wall version of Australia, but when we prepare a large space photomosaic, either at the scale (1:1,000,000), at which it is normally printed or at a moderately reduced scale (say, 1:2,500,000) to keep it's size manageable, we retain most of the original information. The chief users of mosaics, such as those shown above, are geographers, structural geologists, and geomorphologists, all of whom benefit from the panoramic vistas afforded by seeing local features in a regional context.

These mosaics also dovetail nicely with the lower resolution views of continents at the hemispherical scale provided by meteorological satellites, such as Nimbus (see page 14-4) and GOES (see page 14-7).

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Electromagnetic radiation at long wavelengths (0.1 to 30 centimeters) falls into a segment of the spectrum commonly called the microwave region. At still longer wavelengths (centimeters to meters) the radiation is known as radio waves (these can be generated by manmade transmitters or occur naturally [e.g., beamed from energetic stars]). Remote sensing has utilized passive microwaves, emanating from thermally activated bodies. But, in much more common use (since World War II) is another manmade device, radar, an active (transmitter-produced) microwave system that sends out radiation, some of which is reflected back to a receiver. The varying signal, which changes with the positions and shapes of target bodies, and is influenced by their properties, can be used to form kinds of images that superficially resemble those recorded by Landsat-like sensors. This first page introduces certain basic principles, describes the common radar bands in use, and shows a typical radar image.

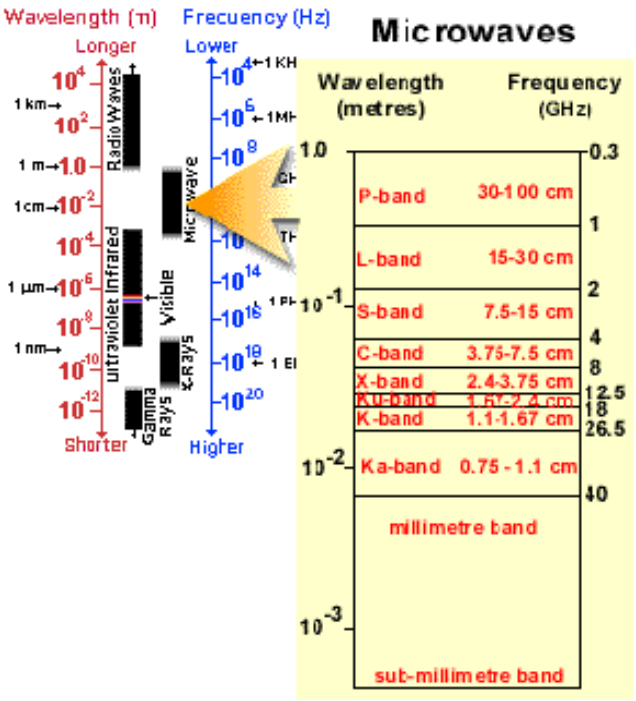
RADAR AND MICROWAVE REMOTE SENSING*

Radar Defined

Radar is an acronym for Radio Detection and Ranging. Radar is an active sensor systems. It generates its own illumination as an outgoing signal that interacts with the target such that some of the signal is returned as backscatter that is picked up by the same antenna that emitted the radar beam. Radar operates in part of the microwave region of the electromagnetic spectrum, specifically in the frequency interval from 40,000 to 300 megahertz (MHz). The latter frequency extends into the higher frequencies of the broadcast-radio region. Commonly used frequencies and their corresponding wavelengths are specified by a band nomenclature, as follows:

- **Ka Band: Frequncy 40,000-26,000 MHz; Wavelength (0.8-1.1 cm)**
- **K Band: 26,500-18,500 MHz; (1.1-1.7 cm)**
- **X Band: 12,500-8,000 MHz; (2.4-3.8 cm)**
- **C Band: 8,000-4,000 MHz; (3.8-7.5 cm)**
- **L Band: 2,000-1,000 MHz; (15.0-30.0 cm)**
- **P Band: 1,000- 300 MHz; (30.0-100.0 cm)**

This chart summarizes the above information on Bands in the Microwave segment of the EM Spectrum:



Unlike other sensors that passively sense radiation from targets illuminated by the Sun or thermal sources, radar generates its own illumination (hence, it is active; another example is the flash camera) by sending bursts or pulses of EM energy that reflect off of a target. A fraction of the reflected energy then returns to the radar's receiving antenna, which collects it and passes it to an electronic processing system. A radar system is a ranging device that measures distances as a function of round trip travel times (at light speed) of a directed beam of pulses (the signal, whose strength is measured in decibels, dB) spread out over specific distances. In this way radar determines the direction and distance from the instrument (fixed or moving) to an energy-scattering object. We can also derive information about target shapes and certain diagnostic physical properties of materials at and just below the surface, or from within the atmosphere, by analyzing signal modifications.

8-1: **In the Introduction, you learned that wavelength times frequency = the speed of light. By choosing the proper units, this equation can be derived: Wavelength (in cm) = 30/frequency (in GHz or 10^9cycles per second). Given a frequency of 20 GHz, what Wave Band would this radar pulse fall into? ANSWER**

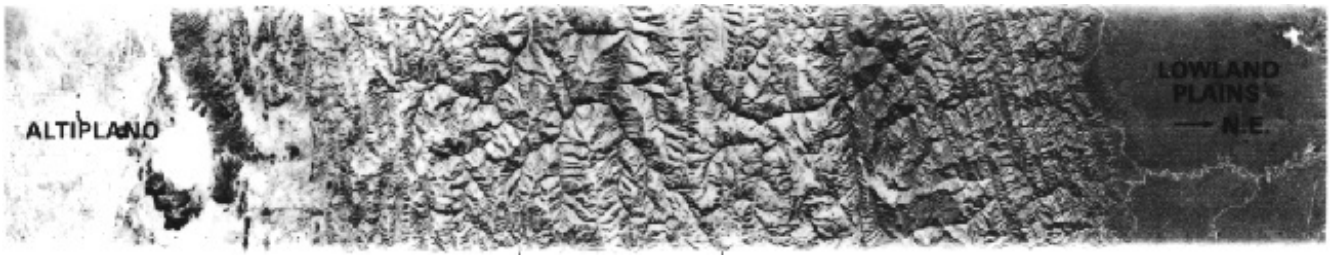
By supplying its own illumination, radar can function day and night and, for some wavelengths, without significant interference from blocking atmospheric conditions (e.g., clouds). These characteristics prompted radar development in World War II for tracking (attacking) aircraft and ships. Ground (fixed) and airborne (mobile) radar systems are used extensively today for marine navigation and air traffic control. Imaging radar, mounted on air or space platforms, has proven especially useful in mapping cloud-shrouded land surfaces (a landforms map of Panama used this approach). This use also permits expressing surface shapes in regions heavily covered by vegetation (penetrated by some bands).

8-2:Why did radar play such a vital role in the Allies' victory over Germany in World War II? ANSWER

As we shall see in this Section, radar has become increasingly important in various applications of remote sensing. The military continues to be a prime user. Several satellites operating now have radar has their principal sensor. Most of us are most familiar with radar's use in Meteorology, mainly as a tracker of storms, rainfall, and advancing fronts. We will skip further reference to this last use in this Section; it is defered until Section 14 that deals with Meteorological Remote Sensing.

This ability to mirror ground surfaces for displaying topography is a prime use of radar for a

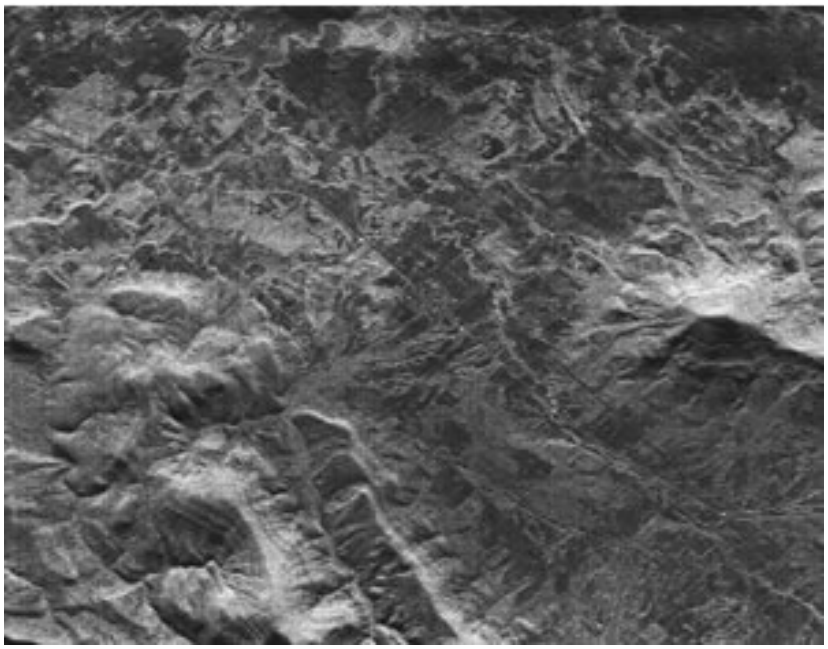
variety of applications. Some radars operate on moving platforms; others are fixed on the ground. The usefulness in determining land surface configurations is strikingly confirmed in the radar image strip (SIR-A system on the Shuttle) shown here, extending 300 km (200 mi) to the northeast (right side) across the folded and dissected South American Andes in Bolivia. The image covers from the high plains (Altiplano) on the west to the lowlands (Amazon Basin) on the east. (Scroll to see this right end.)



8-3: **Care to guess why the Low plains are dark and the High (Alti) plains are light.**

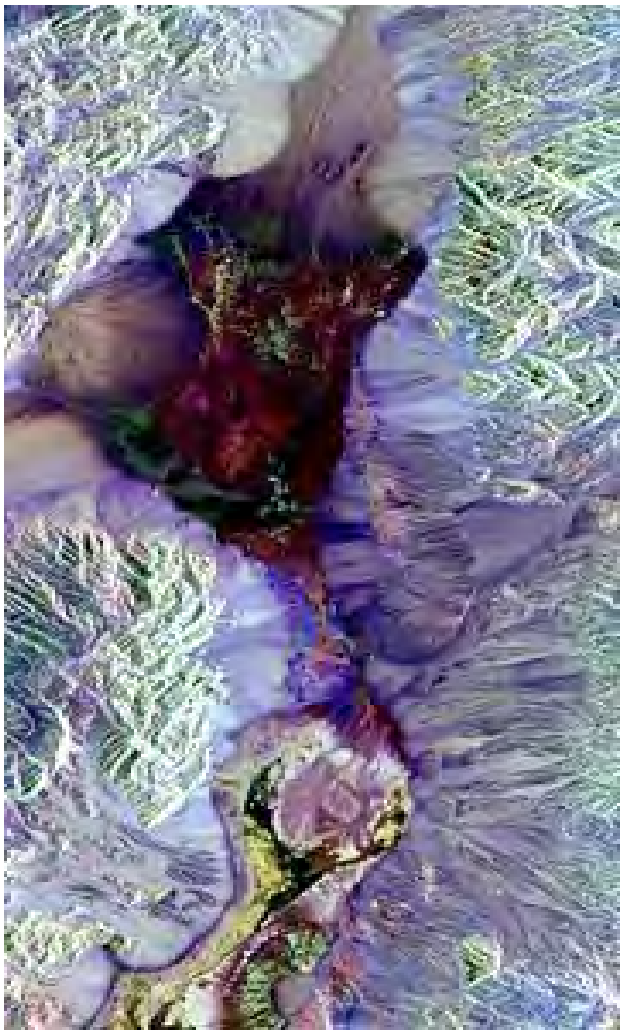
Radar systems, as well as passive sensors operating at microwave wavelengths (slightly shorter than radar), are also effective in detecting soil moisture, storms-clouds-rain, and sea states.

Radar has operated on the ground and from aircraft for more than a half century. These are typical radar images taken from moving aircraft:

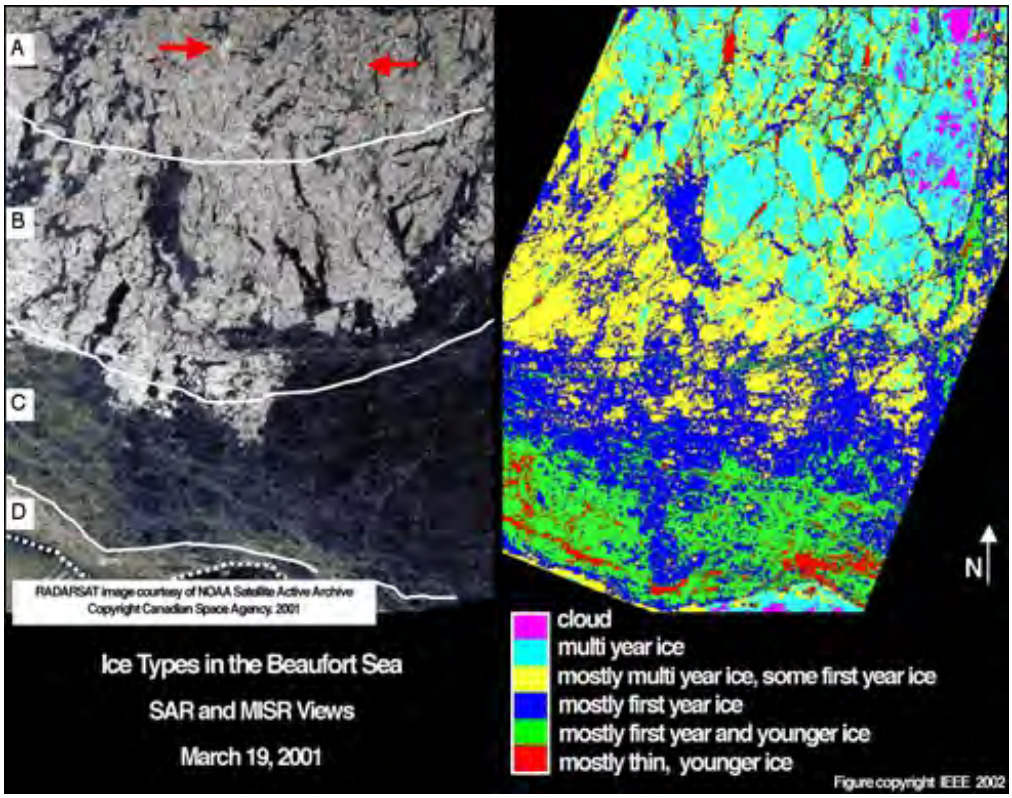


Radar systems are now being routinely sent into space to image the entire Earth. Some of these are multiband and multipolarization radars; this allows color composites to be constructed. On the left is a typical multiband satellite radar (SAR or Synthetic Aperture Radar) image, showing part of Death Valley in California, and surrounding mountains; a Landsat counterpart is on the right:

Here is a closer look at part of Death Valley in another SAR image;



Multiband radar data sets, sometimes combined with other modes of satellite imagery, can be used to classify features in a scene, using the classifier approach described in Section 1. As an example, here is a classification of sea ice (by age) in the Beaufort Sea:



---

* This unit is adapted and expanded from the section on Radar Systems (pp. 367- 374) in the Landsat Tutorial Workbook, NASA RP 1078, 1982. For an excellent, up-to-date review of Imaging Radar Technology, covering both air and space systems, click on the NASA JPL Radar Page

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

The main components of a radar instrument are described. Using a general diagram, the mode of operation of a radar system is explained. The difference between a SLAR and a SAR type radar is clarified. Terms in common use that indicate directional look properties of radar are defined. Two types of resolution are discussed and the importance of pulse intensities in interpreting returned radar signals is stated.

How Radar Works

A typical radar system consists of the following components:

(1) A pulse generator that discharges timed pulses of microwave/radio energy

(2) A transmitter

(3) A duplexer that alternates the signals involved between transmitted and received

(4) A directional antenna that shapes and focuses each pulse into a stream

(5) The same antenna which picks up returned pulses and sends them to a receiver that determines the time delays between their transmission and return, and converts (and amplifies) them into video signals

(6) A recording device which stores them for later processing

(7) An electronic signal processing system that analyzes the signal and prepares it for display

(8) A realtime analog display on a cathode ray tube (CRT), or a TVlike monitor, or on film exposed by a moving light beam to record the image.

Each pulse lasts only microseconds (typically there are about 1,500 pulses per second). Pulse length–an important factor along with bandwidth in setting the system resolution–is the distance traveled during the pulse generation. The duplexer separates the outgoing and returned pulses (i.e., eliminates their mutual interferences) by blocking reception during transmission and vice versa. The antenna on a ground system is generally a parabolic dish; the dish can rotate (sweep), commonly over a 360° range, or is fixed and looking outward.

Radar antennas on aircraft are usually mounted on the underside of the platform so as to direct their beam to the side of the airplane in a direction normal to the flight path. For aircraft, this mode of operation is implied in the acronym SLAR, for Side Looking Airborne Radar. A real aperture radar system (RAR), also know as brute force radar, operates with an antenna that has a discrete physical length. (Aperture is analogous to Field of View.) For a SLAR radar this is a long (about 5-6 m) antenna, *usually* shaped as a section of a cylinder wall. This type produces a beam of noncoherent pulses and uses its length to obtain the desired resolution (related to angular beamwidth) in the azimuthal (flight line) direction.

The antennas can be large, although weight especially is a limit to the size put on an airplane. The antenna used in the SIR-C radar system was designed to fit into the cargo bay of its host Space Shuttle. This is the antenna before it was mated to the Shuttle:

So, for many operational setups there is an upper limit to the size of the antenna. But the system spatial resolution is proportional to length of the antenna, thus getting an antenna as long as practical is an objective in designing the radar. A "trick" can increase the effective length, and hence resolution, by simulating a real aperture through electronic means for integrating the pulse echos into a composite signal. This is the approach used by Synthetic Aperture Radar (SAR), which is always associated with moving platforms carrying the system. SAR utilizes both recording and processing techniques to generate a signal that acts as though it has an "apparent" length greater than the antenna itself. At any instant the transmitted beam propagates outward within a fan-shaped plane, perpendicular to the flight line; this amounts to a "broad beam" through which targets components advance. The movement of the antenna, be it on aircraft or spacecraft, is involved in increasing the effective length simulating a real aperture by integrating the pulse echos into a composite signal. Mathematical manipulation, including Fourier Analysis, is needed to handle the pulse stream associated with the forward motion. More information on SAR is found at the Wikipedia website on SAR.

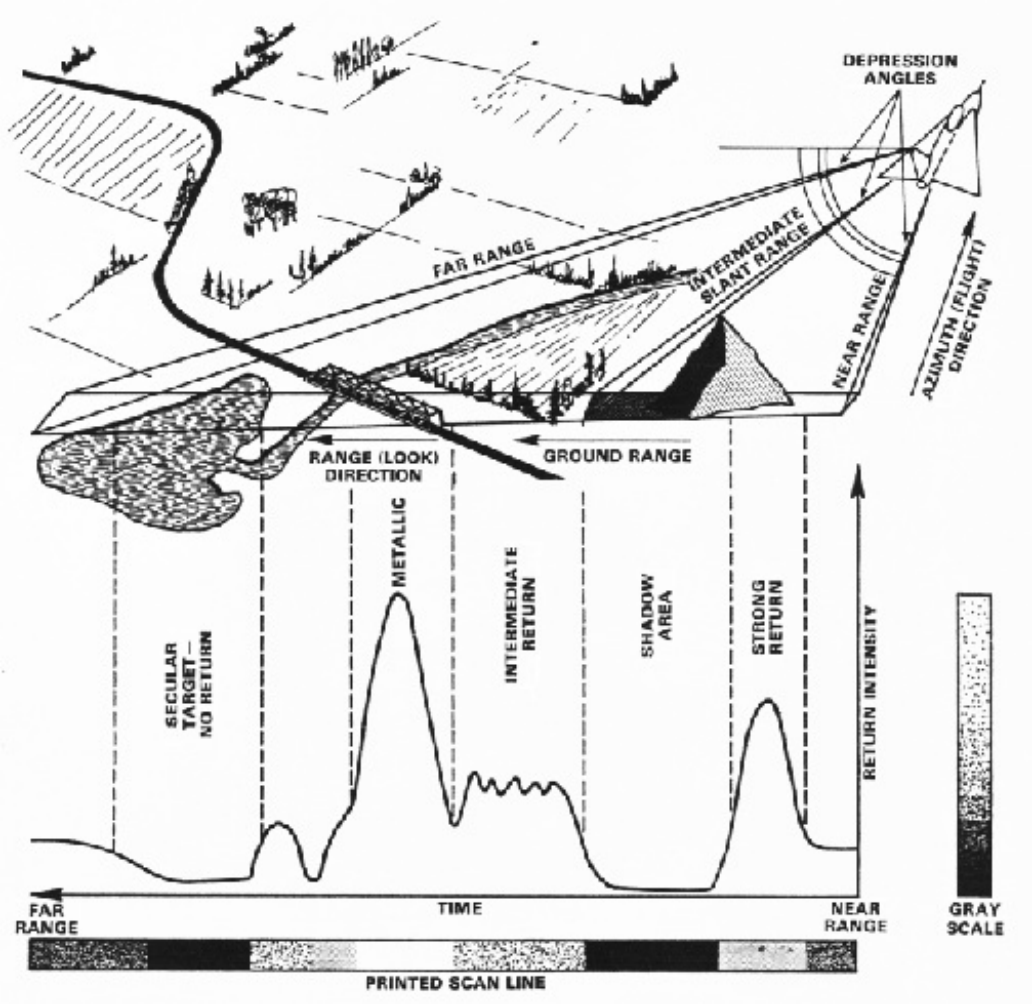
SIR-A, B, C, Radarsat, ERS-1,2, JERS-1, and Envisat are the main examples of space satellites that use SAR.

Another type of radar system is exclusive to conditions in which there is relative motion between platform and target. This system depends on the *Doppler* effect (apparent frequency shift due to the target's or the radar-vehicle's velocity) to determine azimuthal resolution. As most know, when a target is moving towards the observer, and emits a signal such as a sound, the frequency continues to rise as the target moves closer - the sound increases in pitch; movement away lowers the frequency. Doppler radar is used in the so-called Radar Guns used by the police to gauge auto speeds based on the rate of frequency shift; in this case the radar beam is continuous rather than pulsed.

For a SLAR system, there is a Doppler component in the returned beam that results in changing frequencies, which give rise to variations in phase and amplitude in the returned pulses. As coherent pulses transmitted from the radar source reflect from the ground to the advancing platform (aircraft or spacecraft), the target acts as if it were in apparent (relative) motion. The system analyzes the moderated pulses and recombines them to synthesize signals. The radar records these data for later processing by optical (using coherent laser light) or digital correlation methods; Fast Fourier Transform analysis is also applied. More about Doppler radar is provided by Wikipedia.

8-4: **What is the main practical advantage over SAR over SLAR? ANSWER**

Let us now consider the beam characteristics of a typical radar system, as well as the nature and interpretation of the signal returns, as displayed on film or a monitor. The following illustration describes this process (from Sabins, 1987):



The aircraft moves forward at some altitude above the terrain in an *azimuthal direction,* while the pulses spread outward in the *range (look) direction*. Any line-of-sight from the radar to some ground point within the terrain strip defines the *slant range* to that point. The distance between the aircraft nadir (directly below) line and any ground target point is its *ground range*. The ground point closest to the aircraft flight trace, at which sensing begins, is the *near range* limit; at this distance the pulsed signal has the shortest roundtrip transit time. The pulsed ground point at the greatest distance normal to the flight path fixes the *far range*; the delay between the instant of pulse release and its return is longest at the far range.

At the radar antenna, the angle between a horizontal plane (essentially, parallel to a level surface) and a given slant range direction is called the *depression angle* β (Beta) for any point along that directional line (a mnemonic for that is "lowering your head down from staring forward when depressed"). We refer to the complementary angle (measured from a vertical plane) as the *look angle* (a good mnemonic is to think of looking up from staring at your feet [vertically downward]). The *incidence angle* at any point within the range is the angle between the radar beam direction (of look) and a line perpendicular (normal) to the surface, which can be inclined at any angle (which varies with slope orientation in non-flat topography). The depression angle decreases outward from near to far range. Pulse travel times increase outward between these limits. The duration of a single pulse determines the resolution at a given slant range. This range resolution is effectively the minimum distance between two reflecting points along the azimuthal direction that the radar can identify as separate, at that range. Range resolution gets poorer outward for a specific pulse duration. Thus the resolution increases (gets better) with increasing depression angles (it's optimum, close-in).

8-5: **Both the Sun and radar are examples of an active illuminating source. Aside from the differences in peak wavelengths, what is the most obvious difference in the way each illuminates its full target? ANSWER**

The duration of a single pulse determines the two types of resolution at a given slant range. This range resolution (effectively, the minimum distance between two reflecting points along the look direction at that range at which these may be sensed as separate and distinct) gets poorer outward for a specific pulse duration. The formula for *range resolution* is:

$$\mathbf{R_r(in\ cm) = \tau c/2\cos\beta}\ ,$$

where τ (Tau, in Greek) is the pulse length (in microseconds), c is the speed of light (3×10^8 m/sec) and β (Beta) is the depression angle. There is a second measure, the *azimuth resolution*, which at any specific slant range point expresses the minimal size of an object along the direction of the flight path that can be resolved; it too varies with depression angle (i.e., slant distance outward). It is given by:

$$\mathbf{R_a(in\ cm) = 0.7S\lambda/D,}$$

where S is the slant range (in km), λ (small lambda) is the system wavelength, and D is the effective length of the antenna in centimeters.

8-6: **Given a C-band radar with wavelength of 5 cm that generates pulses 0.1 microseconds duration, what is its range resolution for a depression angle of 50 ° its antenna is 5000 cm in length, and it looks outward at the altitude it is flown for slant distances from 5 km to 15 km - for these conditions calculate the resolutions at the corresponding near and far ranges.** **ANSWER**

Pulse intensities of returned signals within the beam-swept strip are plotted in the lower half of the above figure. The pulses undergo varying degrees of backscattering when they reach an object. A smooth, or specular, surface at low angles to the look direction (subparallel) scatters most of the pulses away from the receiver, so that its image expression is dark. A rough surface, in contrast, scatters the pulse beam over many directions, a fraction of which returns to the radar. The amount of returned signal determines the relative lightness (related to signal amplitude) of the image tone. The quantity of returned energy (as backscatter echos) also depends on the size of the target relative to the signal's wavelength. Objects with dimensions similar to the wavelength appear bright (as though rough), while smaller ones are dark (smooth).

In the above diagram, note first the intensity peak in the tracing associated with the steep slope of the mountain side facing the passing aircraft. At this low incidence angle, a significant part of the transmitted pulses is reflected directly back to the receiver. However, the beam fails to reach (illuminate) the opposing mountain slope (back side) leading to no return (black) from this shadowed area or if the slope is so inclined as to receive some illumination at high incidence the returned signal is weak (dark gray). (For a mountain with some average slope and a given height, the shadow length increases with decreasing depression angle.) The next feature encountered is vegetation, which typically consists of irregular-oriented surfaces, with some leaves facing the radar and others in different positions. These objects in the plant together behave as somewhat rough and diffuse surfaces, scattering the beam but also returning variable signals of intermediate intensities, depending on leaf shape and size, tree shape, continuity of canopy, etc. The metal bridge, with its smooth surfaces, is a strong reflector (buildings, with their edges and corners, also tend to behave that way but the nature of their exterior materials somewhat reduces the returns). The lake, with its smooth surface, functions as a specular reflector to divert most of the signal away from the receiver in this far range position. Smooth surfaces at near range locations will return more of the signal. The behavior of a surface, whether smooth, intermediate, or rough, in terms of the height h of small surface objects, can be determined by a "less than" formula. Peake and Oliver give these criteria: For a surface to be smooth, this applies: h < wavelength/25 x sin β; to be rough: h > wavelength/4.4 x sin β (remember, β is the depression angle).

8-7: **Calculate the smoothness and roughness values for a wavelength of 15 cm and a depression angle of 60°.** **ANSWER**

8-8: **How will the circular fields, watered by the pivot irrigation described in Section 3, appear (gray level wise) in a radar image? ANSWER**

The signal trace shown in the figure represents a single scan line, which is composed of pixels, each corresponding to a resolution-determined area on the ground. The succession of scan lines produces an image by varying either the light intensities on a display (itself made up of screen-resolution pixels), or the density levels in a film, in proportion to the signal intensities. On either film or conventional black and white monitors, strong intensity peaks show as light tones and weak returned signals are dark.

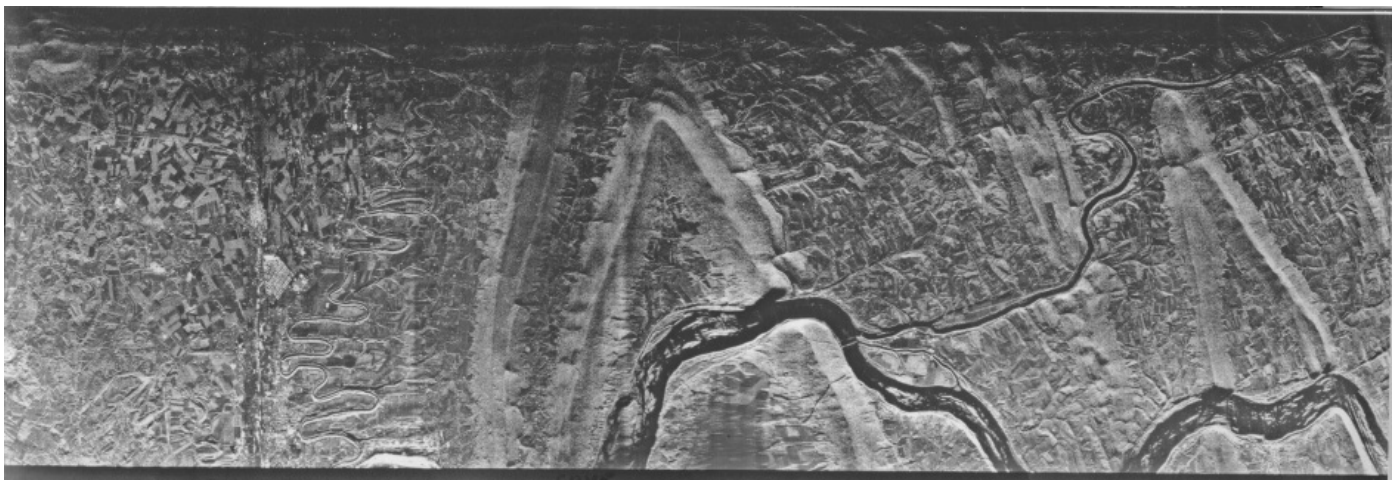
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Several examples of radar images taken from aircraft missions are shown. Two of these are mosaics and one image pair demonstrates the stereo capability possible with radar imagery.

Harrisburg, PA and Nigeria/Cameroon Radar Images

Armed with this background in basic principles and nomenclature, you are now ready to learn more about radar image interpretation. Examine the aircraft SLAR image (K-band) of Harrisburg, PA (near the bottom left) shown below:



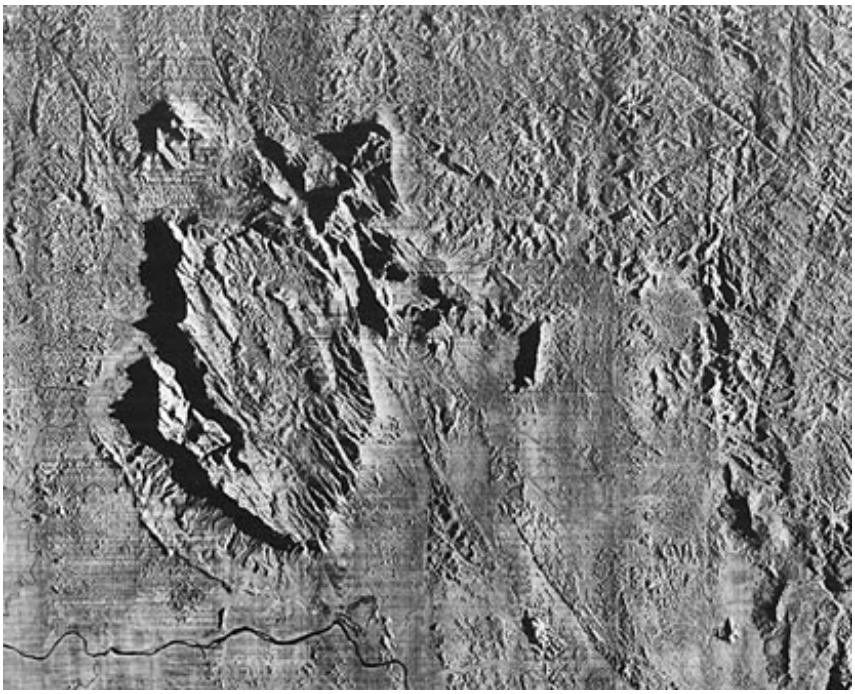
The image was acquired by a SLAR system; look angle increases from the bottom to the top - thus the base is at image bottom. In the image, moist and bare fields show up in various darker shades but those with crops have lighter tones. The more rugged terrain (ridges with notable topographic relief) resemble those in an aerial photo, in which hill shadows are prominent, such as at lower Sun angles. The brighter slopes are oriented towards the aircraft, which flew along the bottom of image, and are at near range. Note that as the look angle increases (towards top), a general darkening (because of an overall drop in signal return), and more distorting of terrain features, occur.

By flying the radar along parallel flight lines, so that it always illuminates from the same direction (thus, the aircraft acquires signals only while flying in one direction) and maintains appropriate incidence angles, it acquires strips of images with identical shading characteristics, which allows us to merge them into mosaics. The image below is a composite of X-band radar strips taken during a flight series over the same Harrisburg area shown above, now combined in a mosaic. Note the different orientation.

8-9: **Fit the first larger-scale image of Harrisburg (above) into this mosaic. What don't you see in both these images that is often present in Landsat images on a transient basis? In the K-band image, something in the valley near the Juniata River is evident that you probably can't see in the X-band mosaic; what?** **ANSWER**

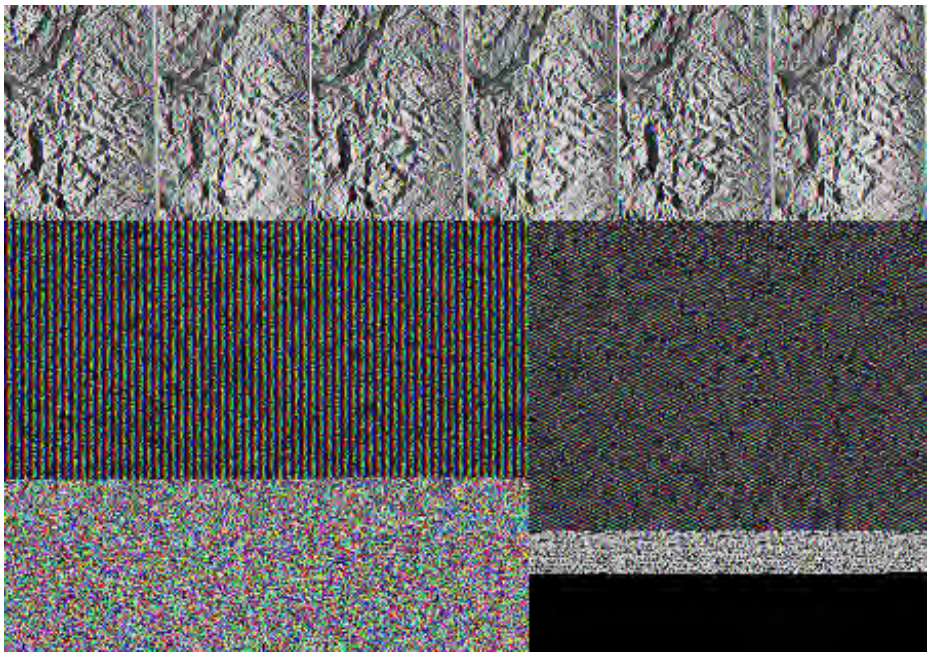
The image below is a composite of radar strips taken during a flight series over an uneven hilly terrain making up part of an igneous/metamorphic shield in the Nigeria/Cameroon border region of western Africa. This mosaic is about 97 km (60 miles) on a side.

Even though the surface is heavily vegetated (tropical forests), the penetrating radar shows mainly its topographic character and adeptly displays fracture systems that have been weathered out into narrow valleys. Another example in which lineaments are prominent is in this radar image that shows a plateau in Venezuela:



Flight crews can set successive flight lines to obtain sidelap imagery, in which part of the same area in each line strip appears in the adjacent strip but at different angles. In principle, two side-looking radars operating simultaneously at different look angles from the same platform can

achieve a comparable effect. This allows viewing in stereo which, despite distortions in terrain shapes, further aids in characterizing and analyzing topography. Examine the two images (from the western U.S.) below, stereoscopically to check out this effect. To do this, print this page and cut the two images apart, then place them at a proper separation. If you are not familiar with how to view two scenes with a stereoscope, check the first part of Section 11, especially page 11-3.



From T.M. Lillesand and R.W. Kieffer, Remote Sensing and Image Interpretation, 2nd Ed., © 1987. Reproduced by permission of J. Wiley & Sons, New York.

8-10: **Where is the highest point in the stereo pair? How could you have deduced this without viewing the two images through a stereoscope? ANSWER**

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

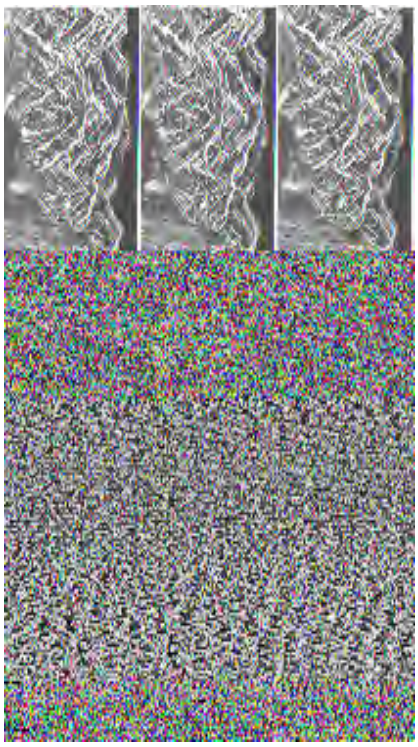
The conditions known as layover and foreshortening, common in radar images, are displayed and explained. The effects of radar beams of varying look angle and look direction, when they illuminate target surfaces with different angles (such as hillslopes), can significantly modify the appearance of and information in a radar image.

Foreshortening and Layover

The look direction, along which the radar beam is traveling, can always be determined by the position of bright slopes in hilly terrain - like those that are sunlit indicating where the Sun is azimuthally, the radar platform will be to the right (looking in the opposite direction) at approximately 90° to the pattern of elongated bright slopes.

Unlike solar illumination which, coming from a distant source, sends its signal as continuous parallel rays of light (photons) onto a sensed surface, radar sends a discontinuous (intermittent) series of photon pulses from a point source that then spreads out as an angular beam. Moreover, the return signal is closely tied to the transmitted signal in that the travel time to the target and back relates to the history of the pulse. The position of any part of the target relative to the near and far ranges affects its dimensions. As a generality, the width of a surface, be it horizontal or inclined, *appears* to increase as the beam spreads from near to far.

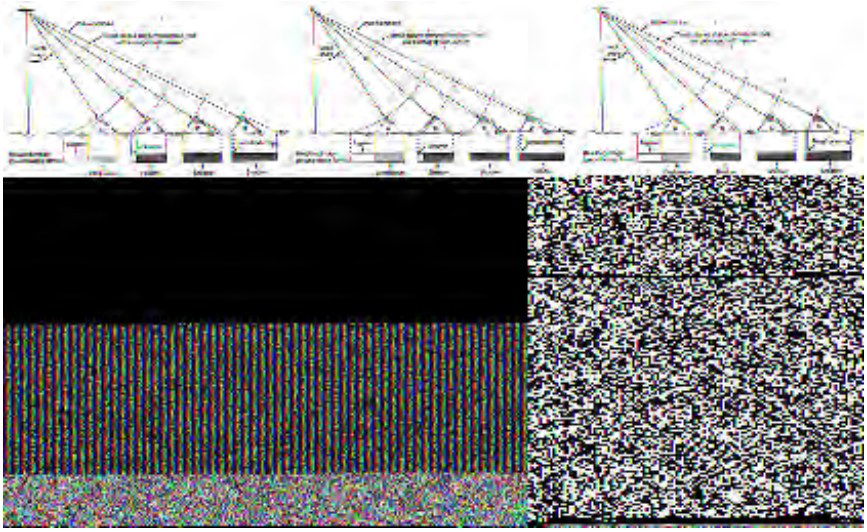
For topographic surfaces other than flat, the irregularies can have a significant effect on the appearance of the resulting image. In radar images where the terrain is quite mountainous with high relief, unusual geometric characteristics may appear. Hill or ridge slopes facing the radar are subject to a distorted appearance. The terms *layover* and *foreshortening* apply to this appearance. Both are expressed as a compression or "thinning" of slopes on the facing (bright-toned) side and an elongation on the side that is shadowed. In most instances, layover and foreshortening produce the same end result visually. This slope displacement is more pronounced in the near range part of a scene than in the far. This radar image is characteristic of this distortion effect:



The image above is typical of the layover seen in aircraft radar images of mountainous terrain. The beam is directed from right to left. The visual pattern gives the impression that the mountains, whose opposing surfaces have similar angles, are analogous to the special topographic feature known as "flatiron" form (hogbacks along the Rocky Mountain Front resulting where rock units are dipping in one direction are this geomorphic type).

8-11: **From what direction is the incoming radar illumination coming? Describe the physical appearance of the mountains, i.e., characterize the distortion.** **ANSWER**

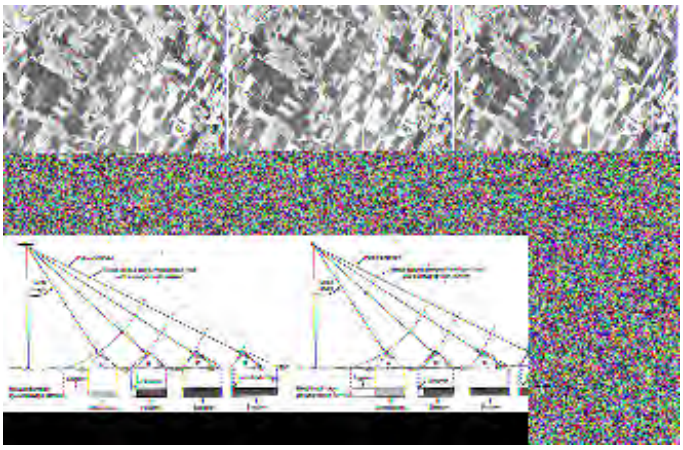
All foreslopes (those facing the incoming beam) are shortened to some extent in radar images. Visually, these slopes appear distorted, with the facing slopes seeming to lean toward the radar platform, as though they are steeper. As depression angles increase, the geometry makes the slope lengths (top to bottom) appear to progressively decrease, thus increasing the degree of foreshortening. Slopes on the opposite side of mountains with ridges will generally not be illuminated and are thus rendered as shadows (no signal returns and therefore dark). Radar-blocked shadows become wider and darken as look angles increase, so they too are distorted from their projected dimensions. These ideas are summarized in the following diagram, adapted from a figure by A.J. Lewis, 1976, as reproduced in the 4th Edition of *Remote Sensing and Image Interpretation*, by T.M. Lillesand and R.W. Kiefer:



In topographic feature A, the beam reaches the top of the slope facing the wave front before it

reaches the slope base. Thus the upper beam is reflected to the receiving antenna before the lower beam, producing the layover effect. This is most pronounced in near range encounters and decreases as the beam encounter tend towards far range; thus layover has dimininished in feature B. Layover is the phenomenon that results when the radar wavefront reaches the top of a slope before the bottom. Foreshortening begins when the wavefront reaches the bottom before the top, such as where the facing slope is less steep than the impinging wave front. The angle of the opposing face (back side) relative to the look angle produces a third phenomenon known as radar shadowing. Shadow length and darkness increases from feature A through D as the look angle becomes progressively less than the fixed back slope angle.

Another kind of distortion occurs in flat terrain. Geometric figures experience dimensional changes as the Look Direction proceeds from near to far slant range. This is expressed as compression of regularly shaped features (square crop fields may distort into a rhombus). This is evident by the gradual shift from square agricultural fields to rectangular ones in this image, from left to right:

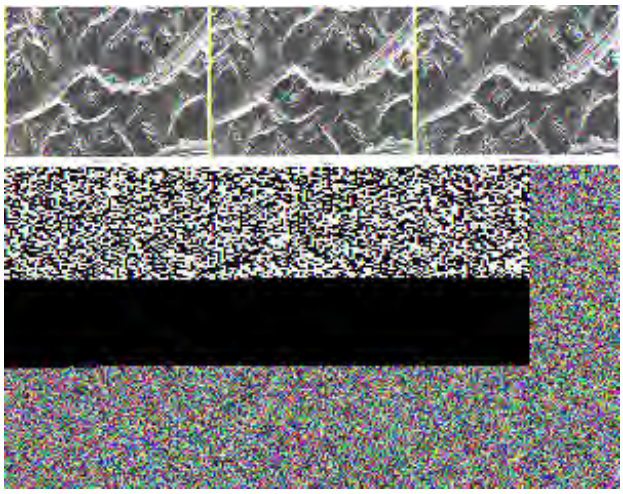


We can convert these images to ground range images if we know independent information on topography. Still other distortions come from erratic aircraft motions during flight (which are nil from stable space platforms). These can also compensated for these by further processing.

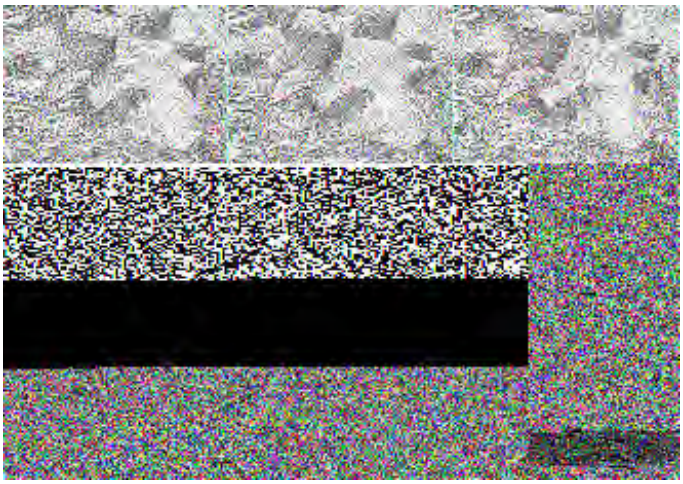
Radar images made from aircraft are more likely to show layover and foreshortening than do images acquired from space radar systems. At higher altitudes, satellites have higher depression angles so that radiation reaches fore and back slopes at about the same time. This also minimizes differences related to near and far range timing.

The foreshortening effects tend to diminish and appear more uniform throughout the image when the radar unit is well above the surface. This is the case for radar on orbiting satellites. The next image displays this diminuted effect well. It's a Seasat image of part of the Pine Mountain thrust in North Carolina:

But, distortion does occur where the geometry of the scene relative to the radar beam brings about some differences in arrival times of the wavefront. Here is layover in part of the Alaska Range as imaged by Seasat:

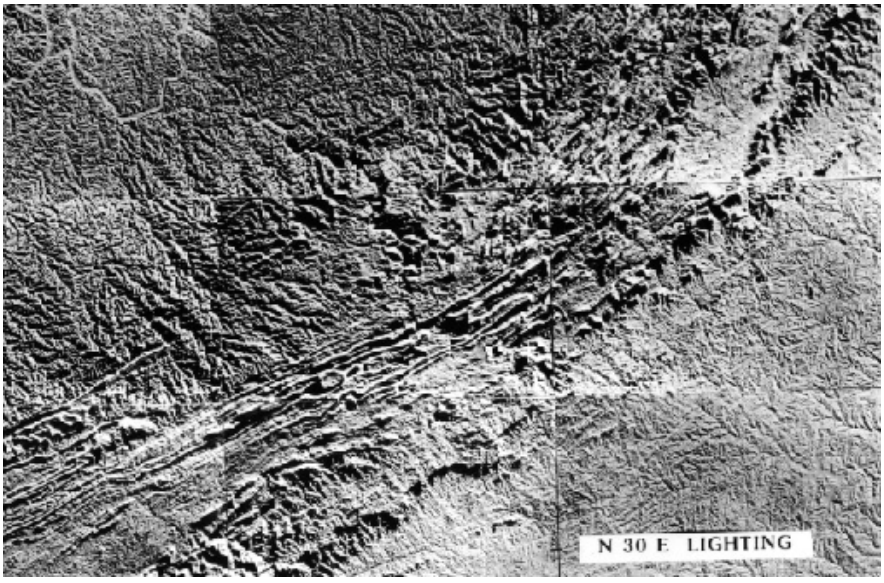


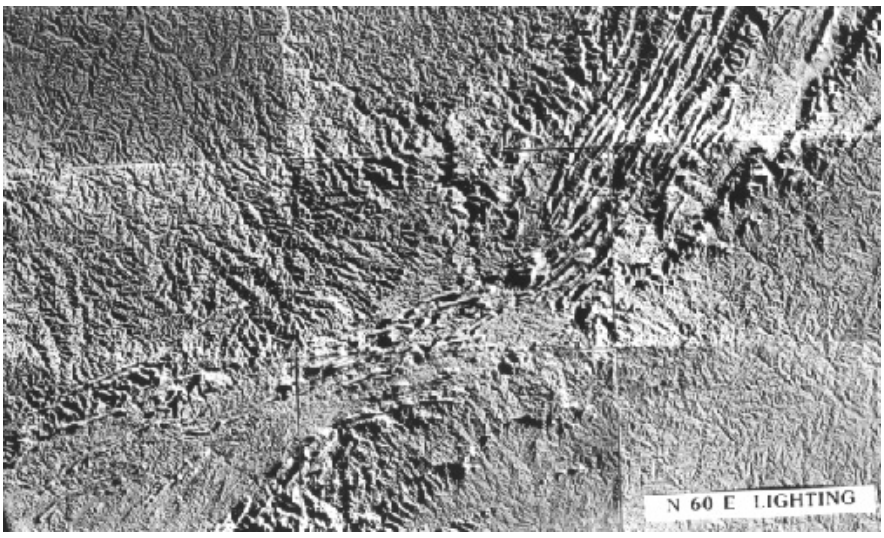
Another mode of image distortion, not strictly geometric, is speckling. Speckling produces a granularity to a radar image that can be distracting. This "noise" results from sporadic, almost random, variations caused by small, but recordable, phase shifts in the radar signal which may be related to ground irregularities, such as plowed soil in a field that results in point sources which backscatter in various directions. The speckling can be minimized by resampling methods in image processing. Here is an example of speckling:



Effect of Illumination Direction

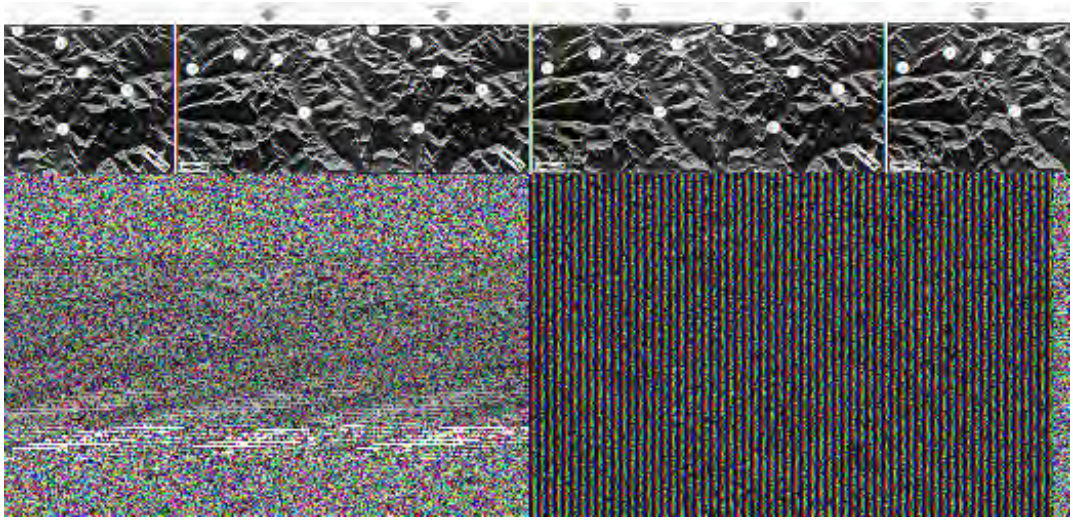
Linear features in rolling or mountainous terrains, such as long, straight valleys or ridge crests, normally stand out with a combination of bright slope-shadow effects. But their patterns change depending on their orientations relative to the flight line or look direction. We illustrate this effect with an interesting analog experiment conducted by Professor Donald Wise (University of Massachusetts), who used a three-dimensional topographic map covered with dark powder that was illuminated at low artificial light angles from several directions:





A simple glance at the two images shows an immediately obvious difference: In the top image, ridges and valleys that trend N 60° E are strongly enhanced by lighting coming from the N 30° E direction , whereas these features when illuminated from the N 60° E direction are subdued in expression but those that trend N 30° E (upper part of ridge trend) now stand out. Maximum visual emphasis occurs when illumination is perpendicular to the trend of a linear feature. Other features at varying orientations are visible in each scene but with subdued expression. As we saw on page 2-8, this phenomenon - that linear trends are greatly influenced by illumination azimuth and angle - is quite pronounced in Landsat images, producing a directional bias. That bias is evident from the plots of fractures on azimuthal rose diagrams, where those oriented northeast-southwest (roughly perpendicular to mid-morning sun angles) tend to dominate the distribution of orientations. This trend can be an advantage in airborne radar imagery, because we can chose flight line directions to underscore and accentuate certain directions of interest to optimize detection of fractures in all orientations.

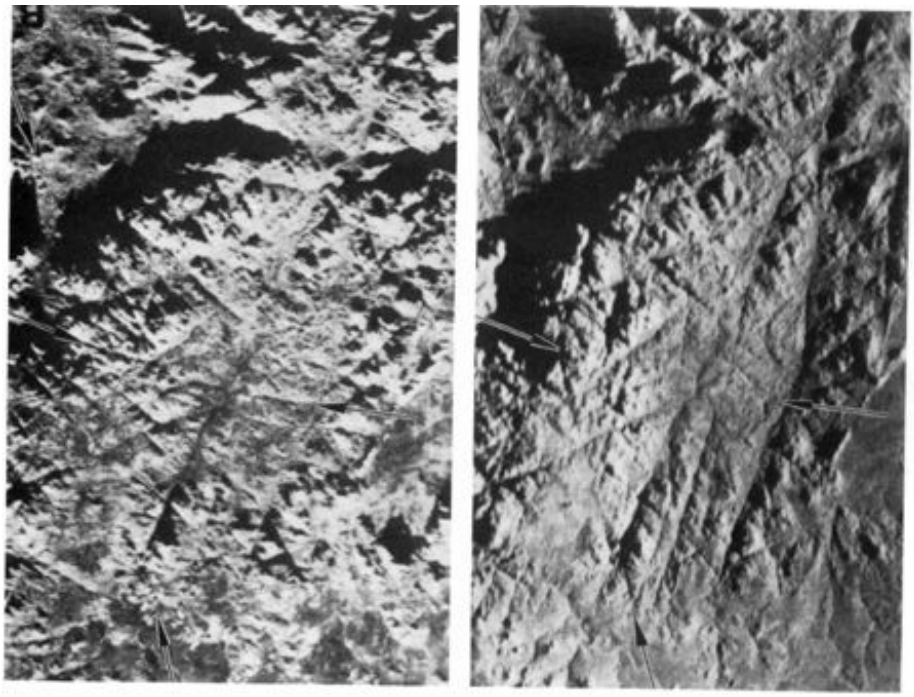
This dependency on direction of illumination is used to good advantage in conducting radar flights over terrains that are sensitive to differences in orientation of mountain ridges, fault valleys, and other conditions where linear features are being sought. The next (aerial) radar image clearly indicates that a different viewpoint (and information extraction) results when the illumination direction shifts relative to the orientation of the scene (north as reference).



From Drury, S.A., *Image Interpretation in Geology*, 1987, Allen & Unwin

The effects can be seen by keying in on several features, such as the stream valley at (e), as these "rotate" with the scene shift indicated by the north arrows. This is the same topography but its appearance, as controlled by changing slope illumination, seems quite different in the two view. Note that foreshortening lessens from top to bottom; front illuminated slopes close to the top (near range) display layover.

To bring this point home, consider this image of Precambrian rocks in the African nation of Nigeria.



In the left image, the look direction is from the bottom; in the right image illumination is from the right. Same scene - but with notable differences in the expression of topography and structural features.

8-12: **Comment on several of the more obvious differences?** **ANSWER**

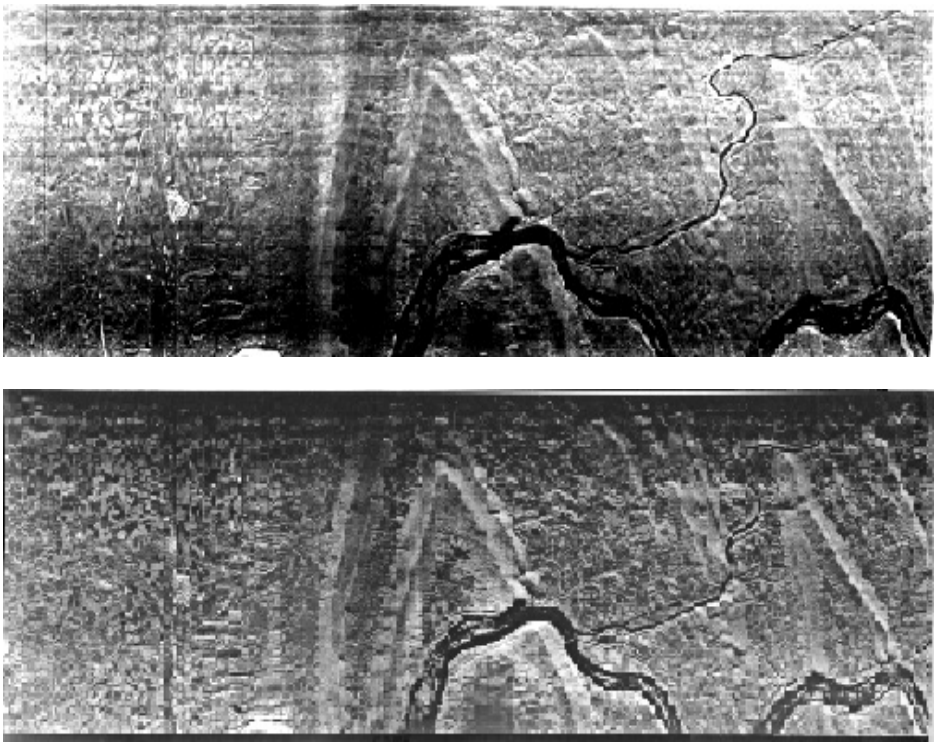
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The mode of polarization in the outgoing signal, and then its interaction with the target, will determine the compound polarization of the returned signal. A parameter called the dielectric constant has an effect on the intensity of signal return.Surface roughness also affects the signal character. Depending on radar wavelength and on the size/shape aspects of target features, different radar bands have differing degrees of penetrability through clouds, forest canopy, and surface materials.

Radar Polarization (Harrisburg, PA); Penetration of Foliage

Radar image tones may also vary in another systematic and controllable way. When a pulse of photon energy leaves the transmitter, its electrical field vector can be made to vibrate in either a horizontal (H) or a vertical (V) direction depending on antenna design. Most reflected pulses are parallel-polarized, i.e., return with the same direction of electric field vibration as the transmitted pulse. Thus, we get either a HH or VV polarization pairing of the transmitted and returned signals. However, upon striking the target, the pulses can undergo depolarization to some extent, so that reflections with different vibration directions return. A second antenna picks up cross-polarization that is orthogonal to the transmitted direction, leading to either a VH or HV mode (first letter refers to the transmitted signal). Some ground features appear about the same in either parallel or cross-polarized images. But, vegetation, in particular, tends to show different degrees of image brightness in HV or VH modes, because of depolarization by multiple reflecting branches and leaves. Compare the HV image of Harrisburg shown below with the previously shown (p. 8-3) K-band image of that area, obtained simultaneously in the HH mode.



8-13: **Describe the principal differences you note between the HV and HH modes.**
ANSWER

Other factors contribute to the brightness or intensity of the returned signal. Two material properties provide clues about composition and surface state by the manner in which these attributes interact with the incoming pulses. One property is the dielectric constant (its symbol is the small Greek letter, *κ*, kappa), which is the ratio of the capacitance of a material to that of a vacuum (yielding dimensionless numbers; the value of κ is arbitrarily set to 1 for a vacuum). It is a measure of both the conductivity and reflectivity in terms of the electrical response of materials. This electrical property describes a material's capability (capacity) to hold a charge, which also measures the material's ability to polarize when subjected to an electric field. Radar waves penetrate deeper into materials with low dielectric constants and reflect more efficiently from those with high constants. Values for *κ* range from 3 to 16 for most dry rocks and soils, and up to 80 for water with impurities. Moist soils have values typically between 30 and 60. Thus, variation in reflected-pulse intensities may indicate differences in soil moisture, other factors being constant. Variations among rocks is generally too small to distinguish most types by this property alone.

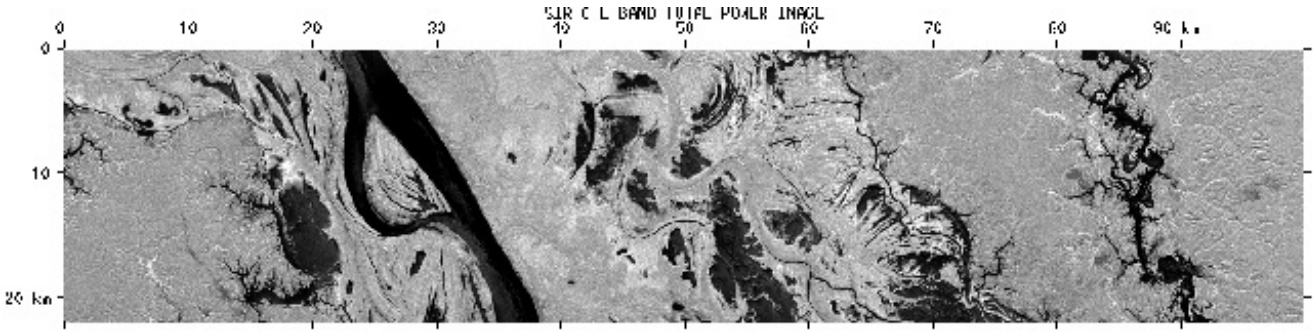
The second material property is roughness. Materials differ from one another in their natural or cultivated state of surface roughness. Roughness, in this sense, refers to minute irregularities that relate either to textures of the surfaces or of objects on them (such as, closely-spaced vegetation that may have a variety of shapes). Examples include the surficial character of pitted materials, granular soils, gravel, grass blades, and other covering objects whose surfaces have dimensional variability on the order of millimeters to centimeters. The height of an irregularity, together with radar wavelength and grazing angle at the point of contact, determines the behavior of a surface as smooth (specular reflector), intermediate, or rough (diffuse reflector). To quantify the effect of different wave bands, a surface with a given small irregularity height (in cm) will reflect Ka band (λ = 0.85 cm), X band (λ = 3 cm), and L band (λ = 25 cm) radar waves as if it were a smooth, intermediate, and rough surface, respectively. Other height variations produce different responses, from combinations of "all smooth" to "all rough" for the several bands used. This situation means a radar, broadcasting three bands simultaneously in a quasi-multispectral mode, can produce color composites, if we assign a color to each band (see below). Patterns of relative intensities (gray levels) for images made from different bands may serve as diagnostic tonal signatures for diverse materials whose surfaces show contrasted roughness.

As an example of multiband color composites, here is one made from SIR-C radar data that shows the Roter Kamm impact structure in western Africa:

Radar Penetration

Radar wavelengths also influence penetrability below target tops to ground surfaces. Depth of penetration increases with wavelength, λ. L and P band radar penetrate deeper than K or X bands. In forests, shorter wavelengths, such as C band, reflect mainly from the first leaves encountered, thus, from the canopy tops. At longer wavelengths, most tree leaves are too small to have much influence on backscatter, although branches will interact, so that canopies are penetrated to varying degrees. The image below, acquired by the SIR-C SAR during its Shuttle mission, shows how L-band radar has penetrated through the dense, continuous tropical vegetation cover of the Amazon Basin in Brazil to image the gently rolling terrain beneath.

>



Signals for all common radar bands pass through the fine droplets of moisture making up clouds, so that this condensation is effectively transparent or invisible to the beam. However, large ice crystals or raindrops do backscatter K-band radiation. Weather radar relies on this band to "picture" clouds and determine their movement using the Doppler effect (the apparent frequency shift that depends on motion toward or away from the receiver).

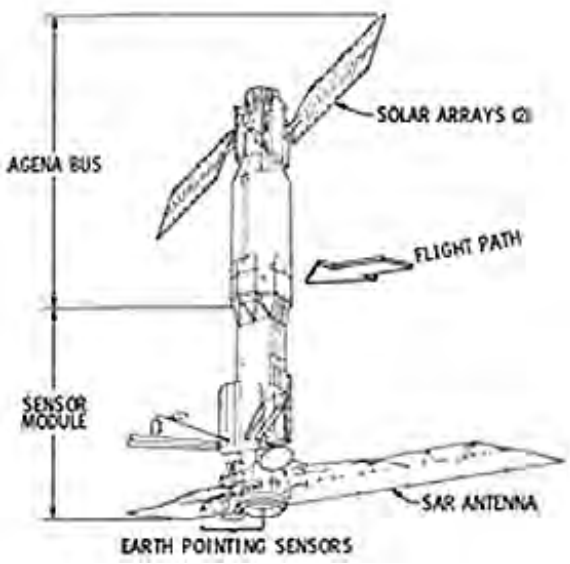
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

The first non-military operation of radar from a spacecraft was the Seasat mission developed by JPL and launched (on a satellite) in 1976. This L-band instrument had a high depression angle which helped to reduce shadowing but heightened foreshortening. Selected areas worldwide were imaged until Seasat's radar failed on the 99th day. Although the intended use was primarily to monitor "sea state", Seasat obtained many fine examples of land imagery. However, the ocean observations, coupled with a radar altimeter onboard, produced a view of the marine surface which had surprising implications about the ocean floor.

Seasat Images

Military aircraft were the first to fly radar units during World War II. These units used very long (1 to 10 m) wavelengths that extended into the radio region of the EM spectrum. Airborne SLAR systems were developed soon thereafter. The first radar systems in space had military applications, so the resulting imagery was classified. The first civilian use of spaceborne radar was the Seasat system operated for NASA by the Jet Propulsion Laboratory (JPL). A synopsis of its history is found at this JPL site. The spacecraft and its systems are shown here:

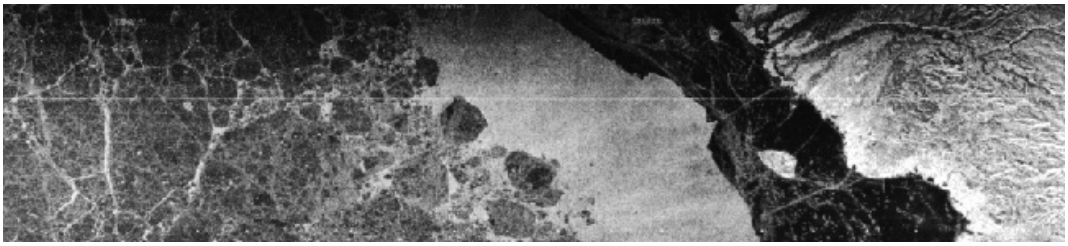


Launched on June 26, 1978, the onboard SAR radar lasted only 99 days before a circuit failed, causing it and other sensors to quit. During normal operations, its L-band (23.5 cm) transmitter produced a focused 1 x 6 degree HH-polarized beam pointed starboard (right) at 20 degree off nadir (vertical). From a slightly elliptical, nearly polar orbit at a nominal altitude of 790 km (491 mi), Seasat's radar had an outward swath width of 100 km (62 mi), giving an image that was printed as four individual strips of 25 km each and then combined to make a full-width image. Along any track (retraced every 24 days), this swath had near and far boundaries that lay between 24 and 240 km (15-150 mi) off nadir. The high depression angles (between 67 and 73 degrees), reduced shadow effects in rugged terrains but induced notable layover. The resolutions achieved by the radar depended on the method by which the synthetic aperture data was processed. With an optical correlator, image resolutions were as low as 70-80 m (230-262 ft), but a digital correlator improved the resolution to about 25 m (82 ft).

Seasat's principal mission was to study various properties at and near the ocean surface, including sea surface temperature, wind speeds, and wave heights. The SAR, with its low look

angle, was designed to measure directions and wavelengths of ocean waves exceeding 50 m (164 ft) in fetch (distance between crests), and to look at sea ice. Below is an image made with the digital correlator of waves off Alaska's southern coastline near Yakutat (note the glaciers on land).

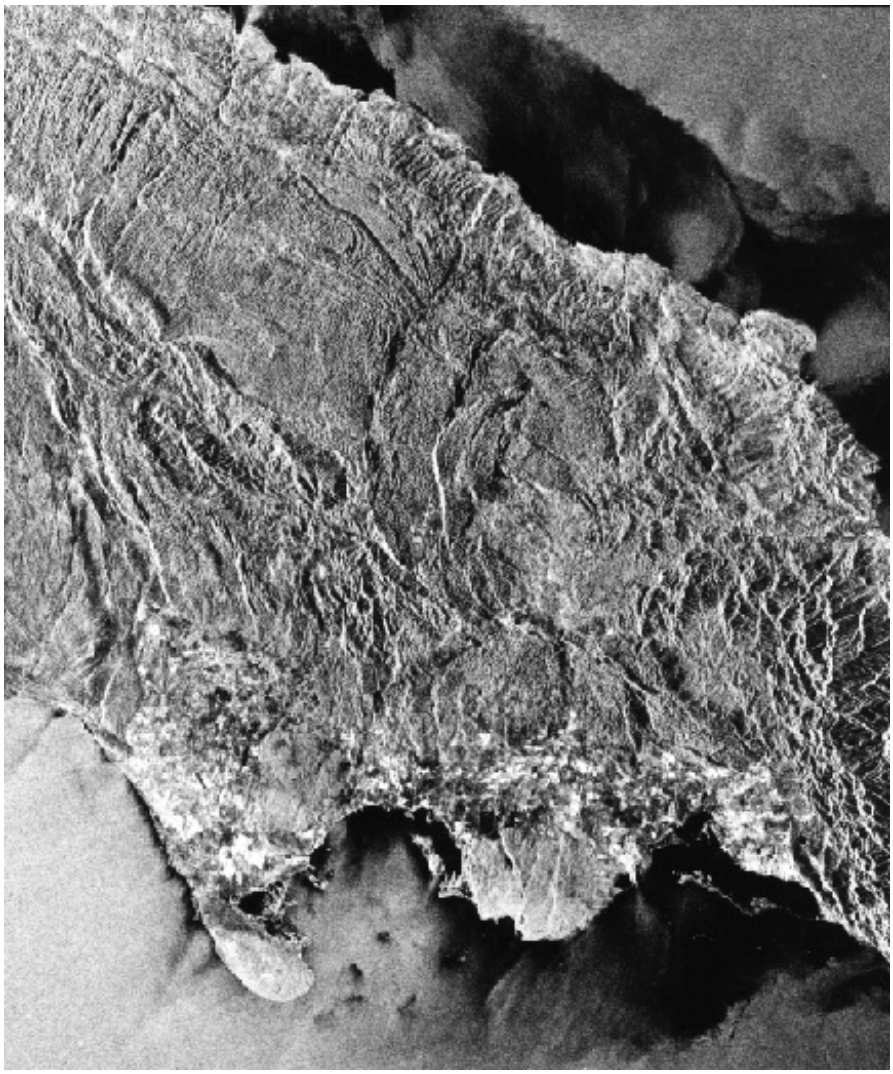
This Seasat image strip below shows sea ice off Banks Island, Canada.



8-14: **Look carefully at the upper scene, focusing on the mountains. Do you see any examples of layover (extreme foreshortening) that causes one topographic feature to appear as though overlapping another? ANSWER**

8-15 In the Banks Island strip, there is both first year and multiyear ice. How do they differ tonally? Hint: think what can happen to ice year after year. ANSWER

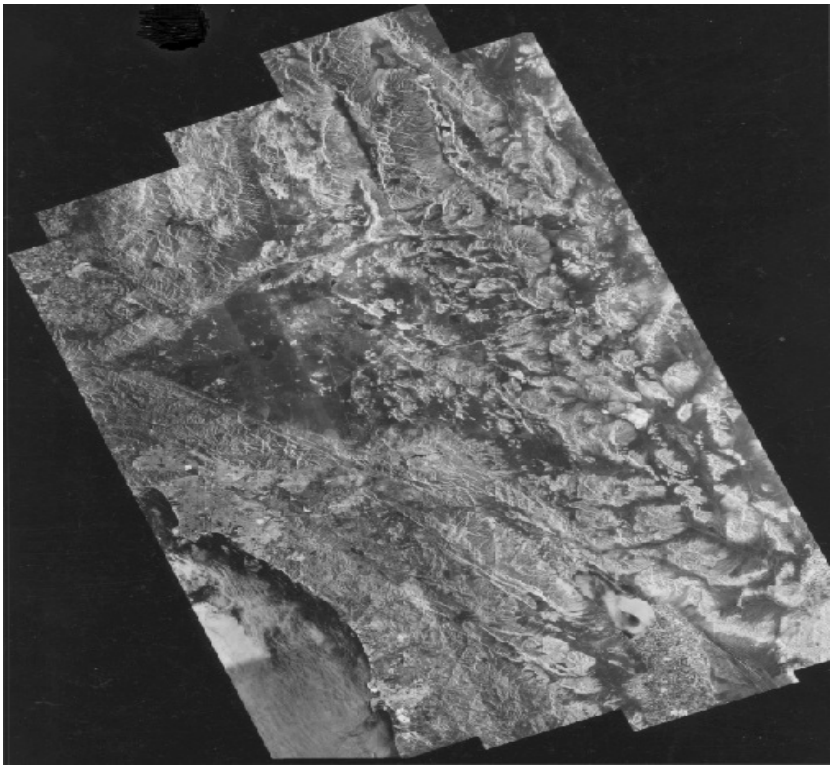
We already showed two Seasat images of land surfaces - the Pine Mountain and Harrisburg scenes. As another example, examine this Seasat image of central Jamaica in the Caribbean obtained through cloud cover on August 8, 1978; there is also some vegetation penetration.



The image shows a variety of landforms: some of mountains that are foreshortened, areas of land use, and roughened offshore waters. Limestone beds govern much of the island's underlying geology. In this area of heavy rainfall, these beds readily dissolve to form typical karst topography.

Seasat was adept at showing both onland and submerged landforms. This next image covers much of southern Florida which, like Jamaica, is a limestone platform. Most of the Everglades shows here as dark (low signal returns) where water flow has spread out but the hardwood tree hammocks (streamlined elongate forms) are bright (white). The Florida Keys form a line of connected reefs at the bottom leading to Key West and in Florida Bay to the north the reefs (some submerged) are revealed by radar to be an interlocking network.

Similar to aerial radar imagery, space radar imagery is well suited to mosaics, as strikingly depicted below in this JPL mosaic of southern California. Use the Digital Elevation Map mosaic of nearly the same region that we presented in Section 7 on Mosaics, as a locator guide (i.e., find Los Angeles).

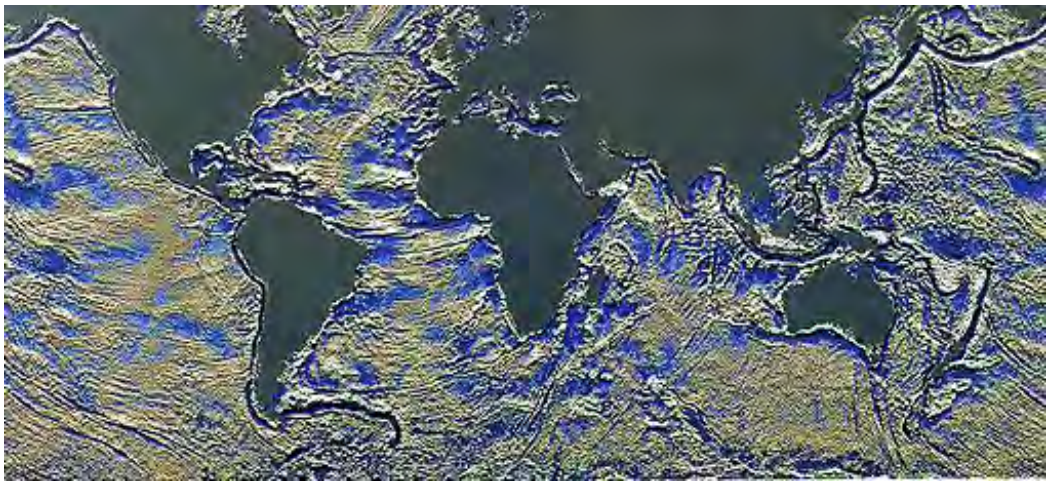


8-16: **Why is the Mojave Desert (the great arrowhead-like salient) shown as dark in this Seasat image? ANSWER**

We can also co-register Seasat radar imagery with Landsat imagery. Below, we superimposed part of a Landsat image of dissected Allegheny Plateau in West Virginia (alone in the upper right), on a Seasat image giving a new impression of apparent relief by virtue of the light tones in the foreshortened foreslopes.

Seasat included a second instrument, a radar altimeter, designed to measure height variations of the ocean surface and large waves imposed on the water. The generated pulses are sent directly downward (as close to vertical, or nadir, as possible [slight variations as the spacecraft wobbled]) and a large fraction of the energy is then reflected straight up. This is timed, so knowing pulse speed (that of light), the total duration or transit time is a measure of distance from Seasat to surface and back). Slight variations in transit time along track (orbital path) represent differences in elevation over the distance traversed.

This led to a remarkable result which proved a sensation among geoscientists and oceanographers. Look at this image, produced at the Lamont-Doherty Geological Observatory of Columbia University, New York:



The map shows the morphological features of the ocean floor at a gross scale. These features are largely the mid-ocean ridges, the trenches, and the many transverse faults (those that segment the ridges at high angles) that emanate from either side of the ridges as the sea floor spreads in both directions away from these ridges. Amazingly, both the faults, which have cliff-like expressions. and the ridges have differential gravitational effects (being topographic highs and lows) that cause the surface waters to mirror these differences. This is what is picked up by the radar altimeter. The map has been color-manipulated to highlight (using blues and yellows)

these differences. This altimeter map is very similar to the (integrated) maps produced over the years by depth soundings and other geophysical measurements of the ocean floors.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

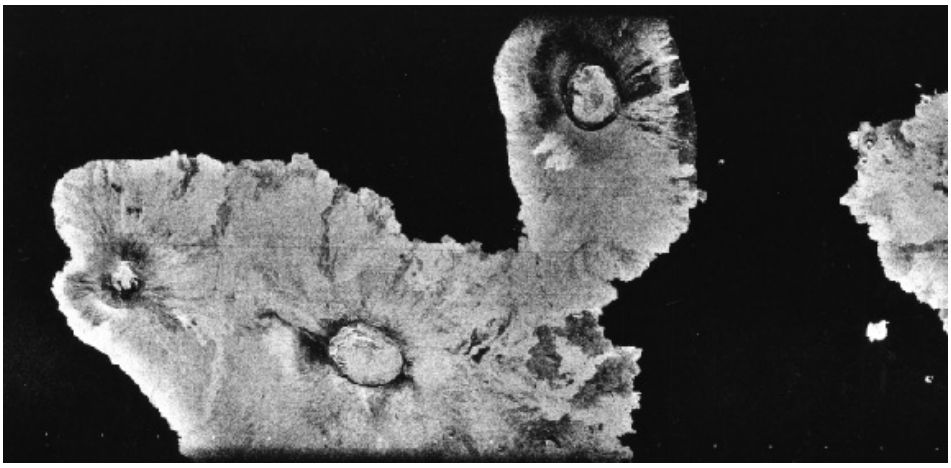
The follow-on to Seasat by NASA-JPL was the SIR (Shuttle Imaging Radar) series: SIR-A; SIR-B; and SIR-C (flown twice on the Shuttle). They differed in bands used, in depression angles, and polarization modes. These different operating conditions produced images that could be made into color composites and, in some instances, into stereo pairs from which perspective (oblique) views could be constructed using altimeter data. The next system developed by JPL was TOPEX-POSEIDON (flown on a satellite), with a strong emphasis on gathering data sets from which topographic information is the output. Other nations have also launched satellites with radar systems onboard. Examples from all of these are presented: Radarsat, ERS-2; Envisat ASAR; ALMAZ; JERS.

SIR-A, -B, and -C on the Space Shuttle; Other Radar Systems

A quick look at Shuttle Imaging Radar is presented in a JPL video. Access it through the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes "Shuttle Imaging Radar", May 8, 2002. To start it, once found, click on the blue RealVideo link.

The next step in NASA/JPL's radar entry into space came from the SIR (Shuttle Imaging Radar) series flown on four Space Shuttle missions. SIR-A used an L-band SAR HH-polarized system, which was capable of 40 m resolution in images whose swath widths were 50 km (31 mi). The near to far range depression angles for this fixed-look radar were 43° and 37°, respectively. These relatively small angles diminished foreshortening and layover effects. This setting has led to some spectacular imagery, such as this color version of folds in the eastern edge of the Atlas Mountains in Morocco.

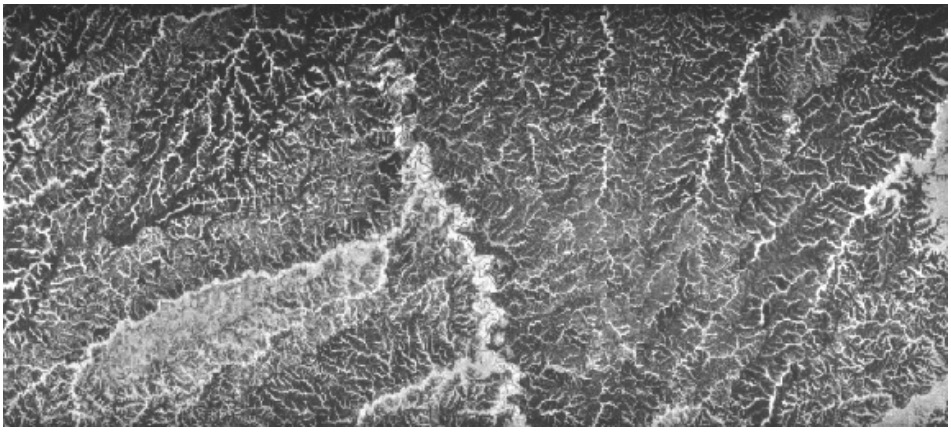
Next, we display a somewhat simpler, but still intriguing, geology, displaying the basaltic volcanic calderas, Volcan Alcedo (top) and Sierra Negra (left center) on Isabela Island, the largest of the Galapagos.



Try to locate the radar image in the following view (with north on the right) of much of the Galapagos group of islands made using the MISR sensor on Terra:

8-17: **What is unusual about this image (clue: think of the volcanoes themselves)?**
ANSWER

Below, the remarkable display of dendritic drainage in the SIR-A image of east-central Colombia results from uplands covered with grass that (because the blades are small) strongly scatters the radar beam (thus dark), whereas the streams stand out as bright because their tree-lined channels produce double bounce reflections between the smooth water and tree trunks.



One property of radar pulses gave rise to an extraordinary image acquired from SIR-A in

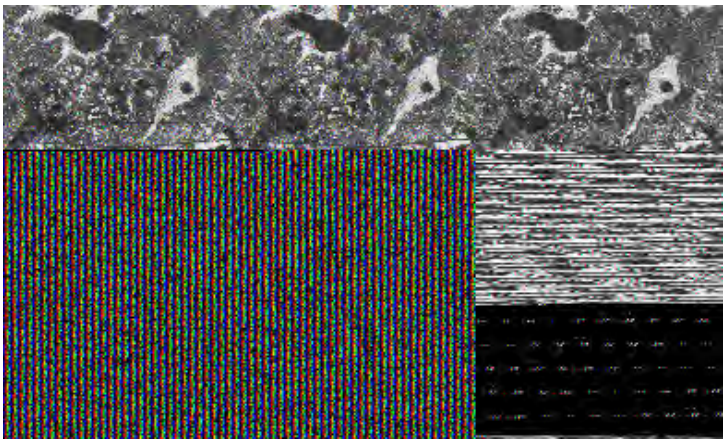
November, 1981. The color scene below is a Landsat subimage of the Selma Sand Sheet in the Sahara Desert within northwestern Sudan. Because dry sand has a low dielectric constant, radar waves penetrate through these small particles to depths of several meters (about 10 ft). The inset radar strip trending northeast actually images bedrock at several meters depth below the loose alluvial sand and gravel which acts as though almost invisible. It reveals a channeled subsurface topography, with valleys that correlate to specularly reflecting surfaces and uplands shown as brighter.

SIR-C radar continued this unmasking to the bedrock in Egypt because of this ability to penetrate the thin sand cover. This area of the Nile River shows to its left the underlying drainage pattern and to its right the rock units involved in a structurally complex geologic terrain:

Both Seasat and SIR-A were L-band radars. They differed mainly in altitude of operation and depression angle: Seasat at 790 km, angle = 67-73°; SIR-A at 250 km, angle = 37-43°; their spatial resolutions were similar. It is interesting to compare the same scene as witnessed by each system. Seasat had a near polar orbit, whereas SIR-A was confined (by Shuttle orbital configuration) to latitudes less than 38°. Look at this pair of views of the California coast and mountain ranges near Santa Barbara, with Seasat on top and SIR-A on bottom.

8-18: **Compare the two radar system images, commenting especially on differences (and why)? ANSWER**

SIR-B operated in 1984 over eight days aboard another Shuttle mission. It differed from SIR-A in having a variable look angle that ranged between 15° and 55°. Here is a SIR-B image L-band image, taken at a 28° incidence angle, of a forested area in northern Florida.

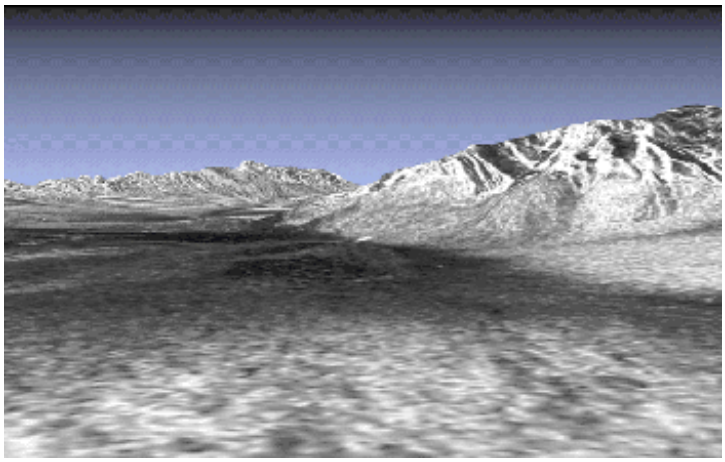


In April and October of 1994, a more versatile system flew twice on Shuttle missions. This system was JPL's SIR-C, which had L- and C-band radars, each capable of HH, VV, HV, and VH polarizations, and an X-band (X-SAR) instrument, supplied by German and Italian organizations, that was in the VV mode. All of these radars had variable look angles that imaged sidewards between 20° and 65°, producing resolutions between 10 and 25 m. One advantage of this multiband system was the ability to combine different bands and polarizations into color composites. JPL's SIR-C Web Site describes how to create composites. You can reach it by clicking here. It contains a wealth of information and imagery, including our next set of images showing the Kliuchevskoi Volcano in Kamchatka (Russian Siberia) as captured by SIR-C in three polarization modes (L-band HH = blue; L-band HV = green; C-band HV = red); on the left is a photo of the volcano taken at the same time by one of the Shuttle astronauts.

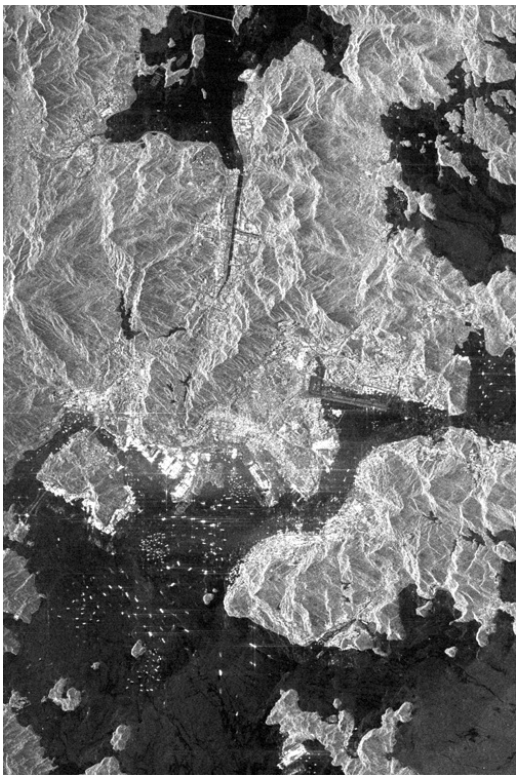
Next, we show a multiband (multifrequency) image of San Francisco, CA, made from L-bands HH (red) and HV (green) and C-band HV (blue). It's one of the most pleasing images to the eye and it shows the city layout, surrounded by water which is one reason why so many people want to live in the Bay Area after visiting it.

8-19: **Name the bridges you can find in this image.**

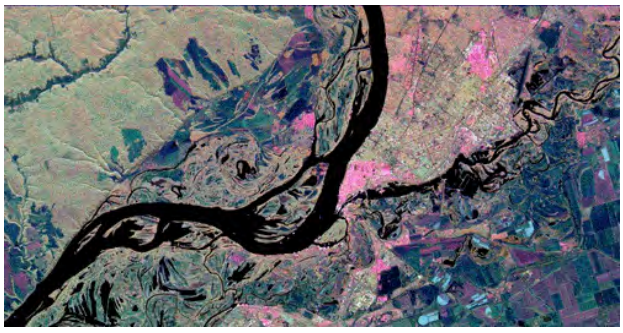
We can process SIR-C radar imagery taken on two dates (or with two antennas) using interferometric techniques that use signal phase differences to determine differences in distance to point targets to yield information on topographic variations. When combined with Digital Elevation Model (DEM) data (see page 11-5), single band or color composite radar images can show perspective views (page 11-8) as analglyphs (requiring stereo color glasses) (page 11-10), or even in simulated flyby videos. A perspective view of Death Valley and adjacent mountains made from SIR-C imagery is a good example.



The X-SAR instrument on SIR-C was supplied by Deutsches Zentrum fur Luft und Raumfahrt (DLR). This X-band (3 cm) radar operates in the VV mode. Here is an image of Hong Kong and adjacent mountains. Note the many ships in the waters near the city (Kowloon).



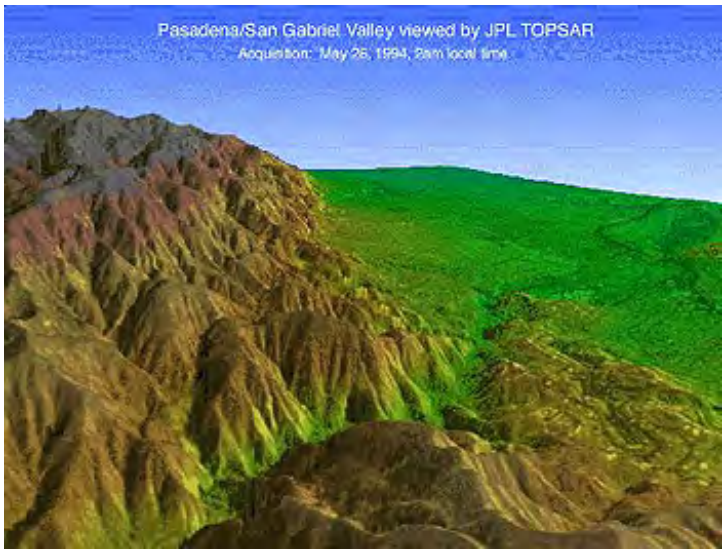
The next image was made from all three SIR-C bands: X-band = blue; C-band = green; L-band = red. The scene shows the city of Samarkind in Russia along the Volga River.



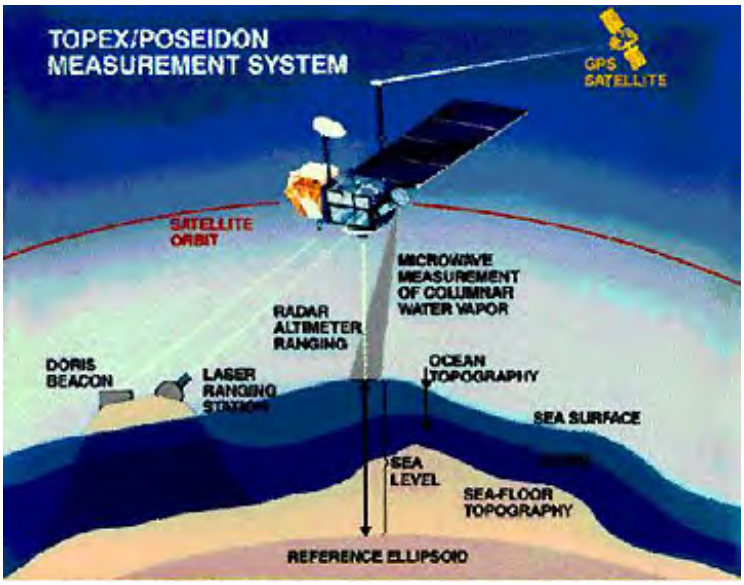
Another JPL radar system, the SRTM (Shuttle Radar Topography Mission) flown in the year 2000, will be discussed in a full page (11-10) in Section 11.

TOPEX/Poseidon

In conjunction with the SIR-C program, JPL flies an airborne system called AIRSAR/TOPSAR that includes an altimeter. One of its principal missions has been to simulate imagery similar to that produced from the TOPEX/Poseidon program. From this system, we present a multiband perspective view of the mountains just north of JPL's home in Pasadena, CA.

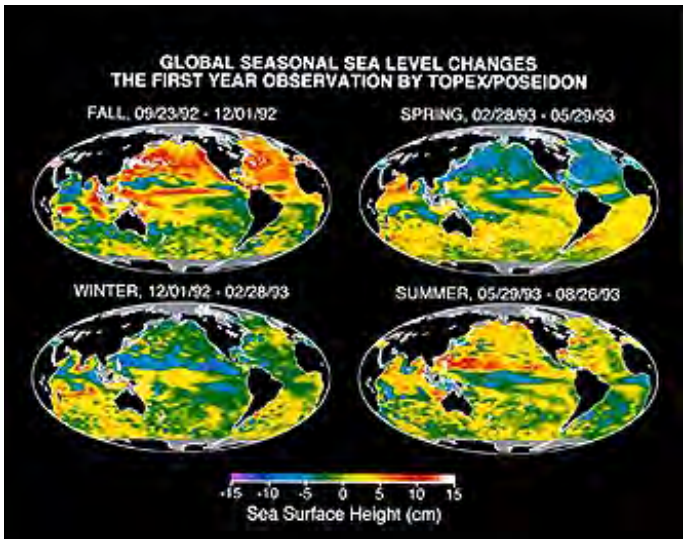


NASA/JPL, in conjunction with the French Center for National Space Studies (CNES), has placed a radar altimeter in space on the Topex/Poseidon mission launched on August 10, 1992. This JPL diagram summarizes the general mission configuration for TOPEX/Poseidon (T/P):

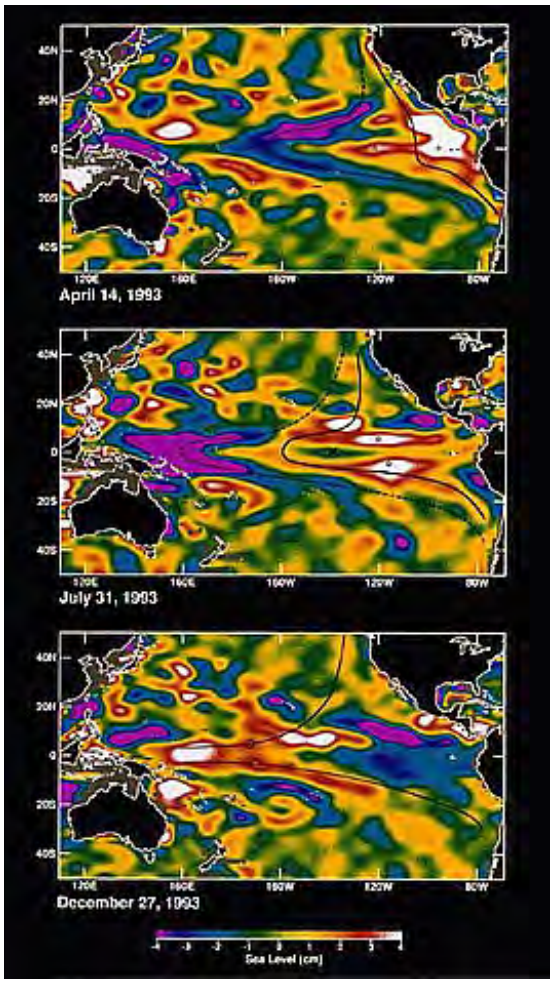


Pointing straight down (at nadir), this dual frequency (13.6 and 5.3 GHz) instrument transmits a narrow beam of pulses whose variations in round-trip transit time represent changes in altitude or (for oceans) wave heights along the 3-4 km swath line (successive lines are spaced about 345 km [214 miles] apart at equatorial crossings). The TOPEX altimeter can discriminate elevation differences of 13 cm. JPL operates TOPEX primarily for oceanographic studies, measuring the effects of wind on waves, and the influence of currents and tides on marine surfaces, and relating these to global climate change mechanisms. A second French altimeter and a microwave radiometer (for atmospheric water measurements) are among the six instruments onboard. The data gathered by TOPEX are not normally displayed as images but are used to produce maps of regions or even global hemispheres, as illustrated in the examples below (see page 14-12 for other examples).

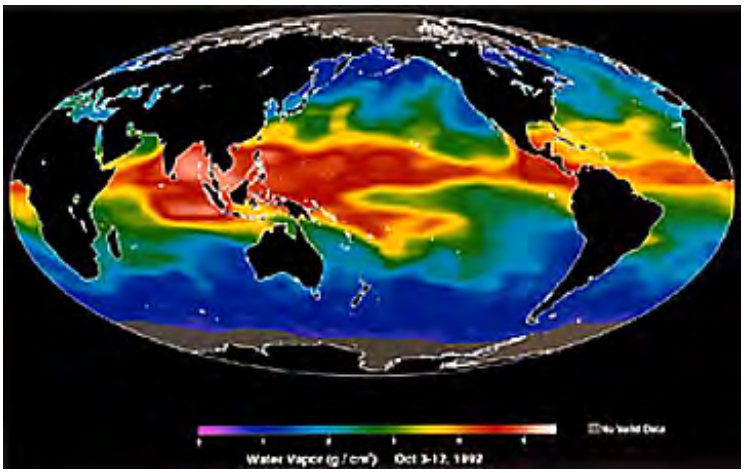
The first image (map) shows T/P sea surface height (SSH) variations (departures from a general mean sea level) for 4 seasons in late 1992-1993.



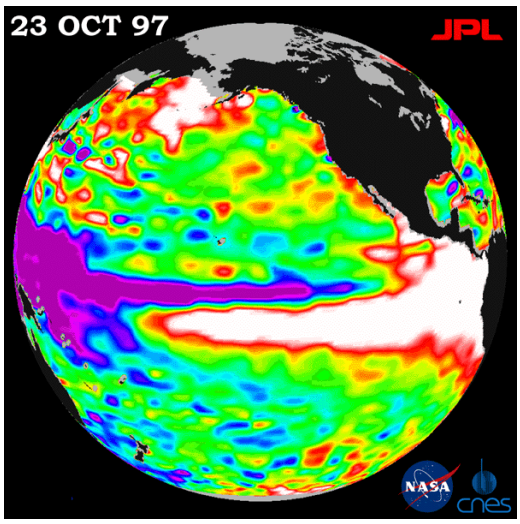
These three T/P maps zero in on height variations in the mid-Pacific Ocean on three days in 1993:



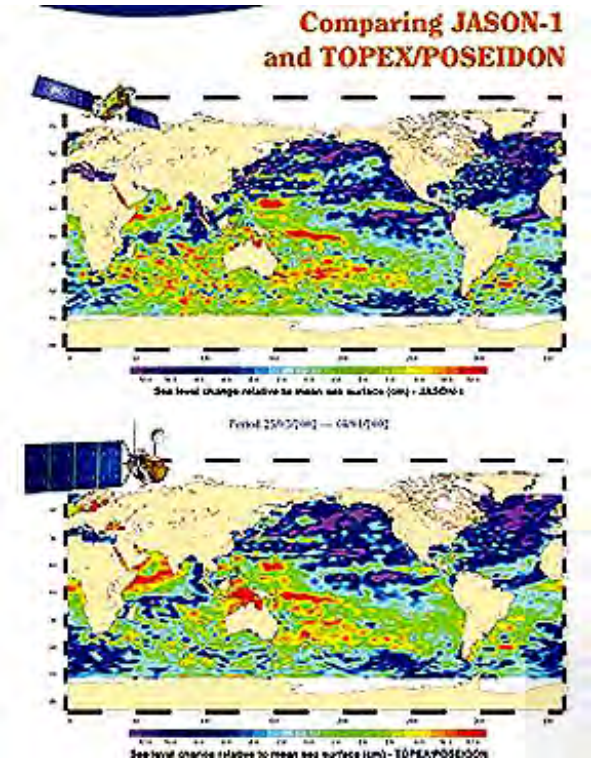
One of the instruments on T/P is used to measure atmospheric water vapor - data used to make a needed correction to improve SSH calculations. Here is a global map of water vapor content.



One of the main tasks assigned to T/P was determination of surface water temperatures using altimetric data and other imputs. This has proved especially vital in monitoring the variations of El Niño. This October 1997 shows a broad equatorial band of very warm water (white) in the eastern Pacific adjacent to a cold band (purple) in the western Pacific that corresponds to a near-maximum time of year for that year's El Niño.



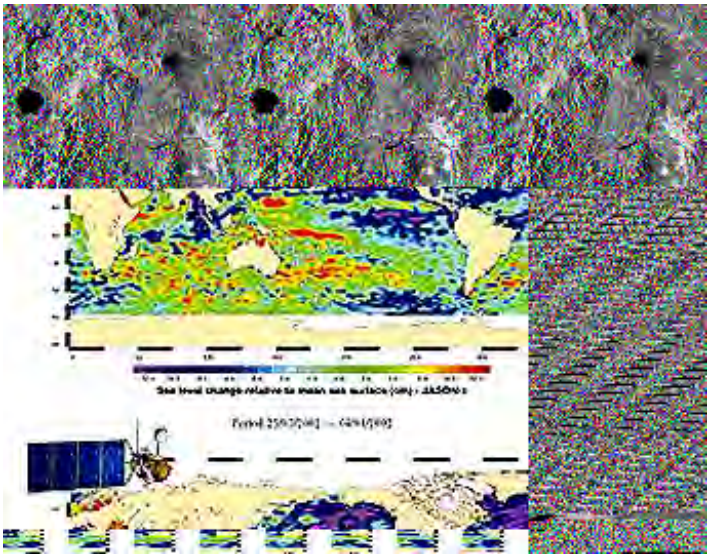
A T/P spacecraft is still operating. In late 2001, a follow-up NASA/CNES satellite, called Jason-1, with greater surface height resolution, was launched (see page 14-12). As a preview, compare these two images (Jason on top; T/P on bottom) taken over a short time period in Fall of 1992.



International Radar Systems: Radarsat; ERS; ALMAZ; JERS

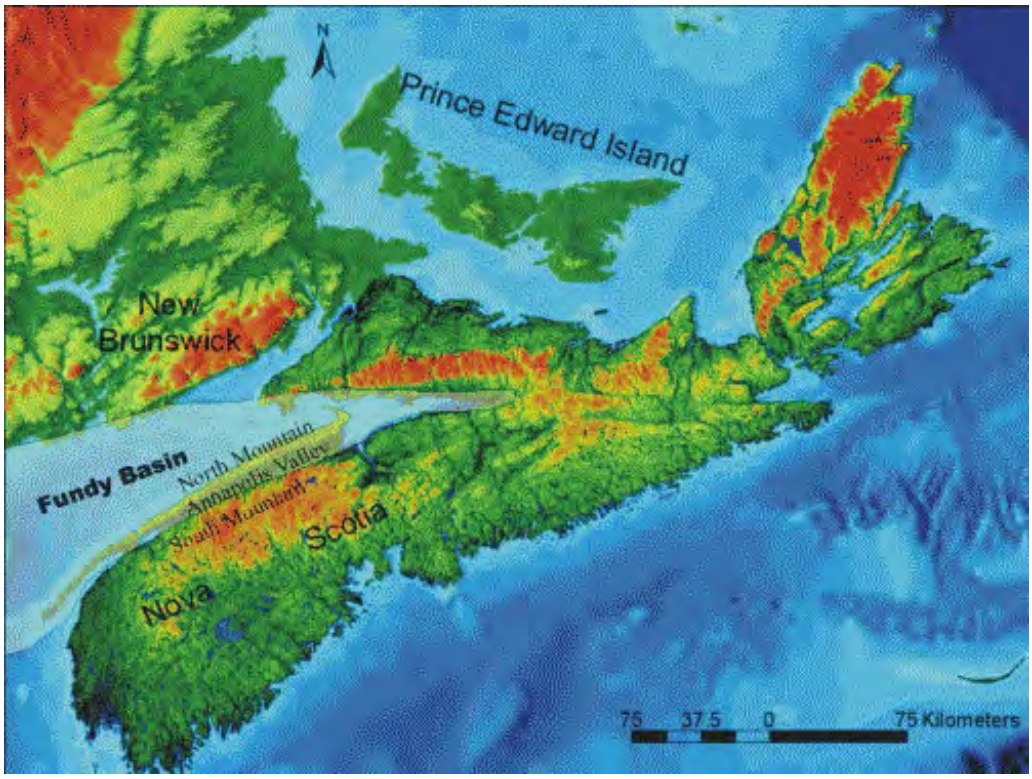
As part of its ongoing remote sensing program, the Canadian Space Agency on November 4, 1995, launched its first Radarsat into a near-polar orbit at a height of 798 km (about 500 mi); this Web site also describes Radarsat2. This is a C-band SAR (5.3 GHz; wavelength 56 mm), whose look angle can range between 10° and 58° to provide swath widths between 35 and 500 km (22 to 311 mi), providing variable resolution centering around 25 m, but ranging from 9 to 100 m as the look angle varies. The first image collected covered Cape Breton in northern Nova Scotia, shown heretofore on Page I-25.

This next Radarsat image covers a 70 km wide area in the northern part of Honshu, the main island of Japan. Note Lake Tazawako, the volcano (Mt. Iwati) and the small city of Morioka (white patch near lower right).

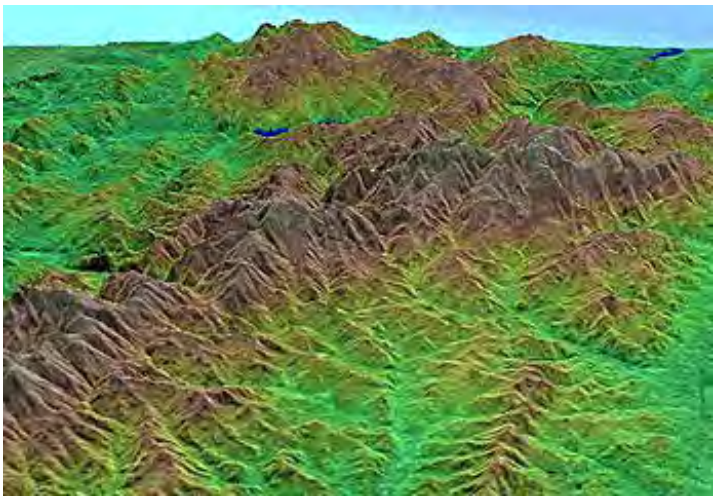


Single band images can be colorized (density-sliced by assigning colors to different gray levels) to help bring out features whose radar returns vary; to some degree in such an image, specific colors may actually relate to separable classes or features. This radar image of the Mekong Delta in South Vietnam shows some color/feature correlation:

Additional Radarsats are operating. This is a recent Radarsat-7 image of Nova Scotia and surrounding parts of eastern Canada:



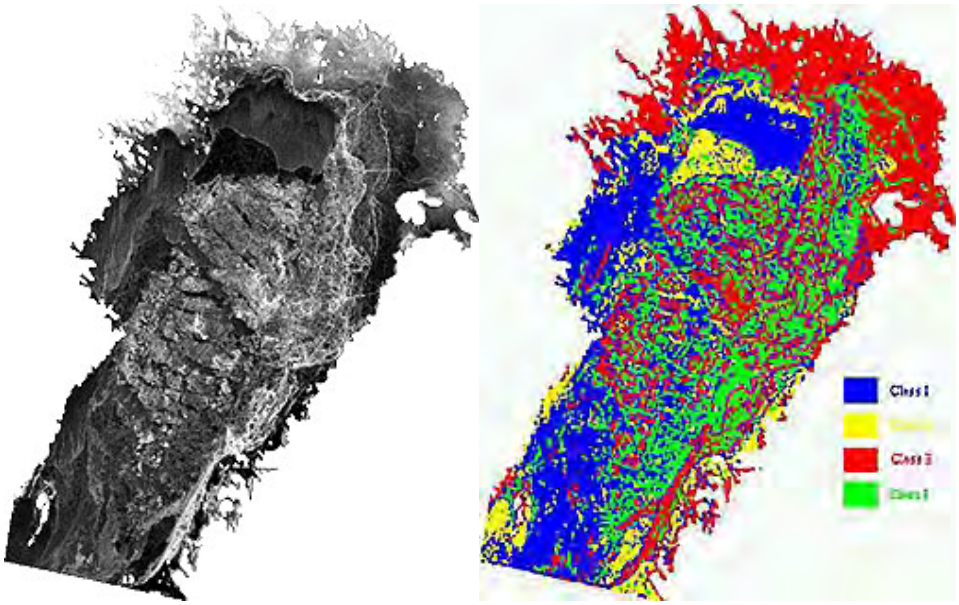
As with most of the other systems previously described, users can convert images from Radarsat using separate topographic data into perspective views. This next image was made from a radar scan over the Tuzla district in Bosnia-Herzogovina, in the Balkans.



Making mosaics of radar images is a fairly easy job since the variability of lighting from different Sun angles as occurs during the changing seasons, which characterizes systems like Landsat and SPOT, is not a factor with radar (although this constancy of radar illumination must be fixed by the same Look Angles and proper overlap conditions). Sixteen hundred Radarsat images have been combined to make a continental-scale mosaic of the African continent:

One of the prime tasks for which Radarsat was flown by the Canadians is to monitor sea ice in their northern (Arctic) waters. The distribution and characteristics of sea ice affect both the (limited) operations of ocean shipping and the extent to which open waters will affect climate (and vice versa). Here is an image of sea ice around Ellesmere Island west of northern Greenland:

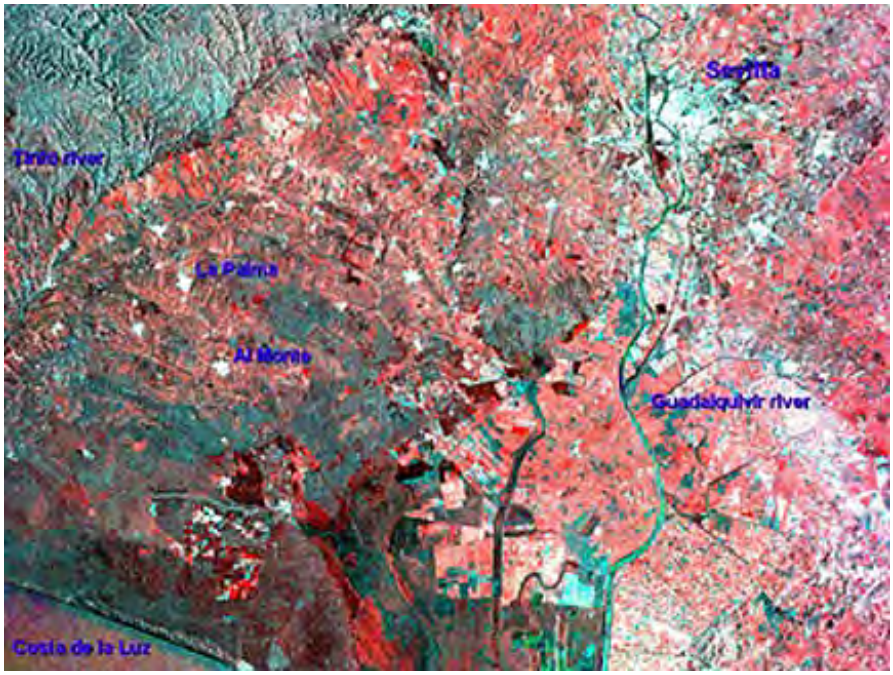
Different types of ice cover can be distinguished when on the ice, or looking from a ship or an airplane. Classes can be set up, based on whether the ice is smooth (fast ice), or deformed in differnt ways and degrees. A classification (right image below) of a small area of Canadian Arctic ocean (left image) met with considerable success.



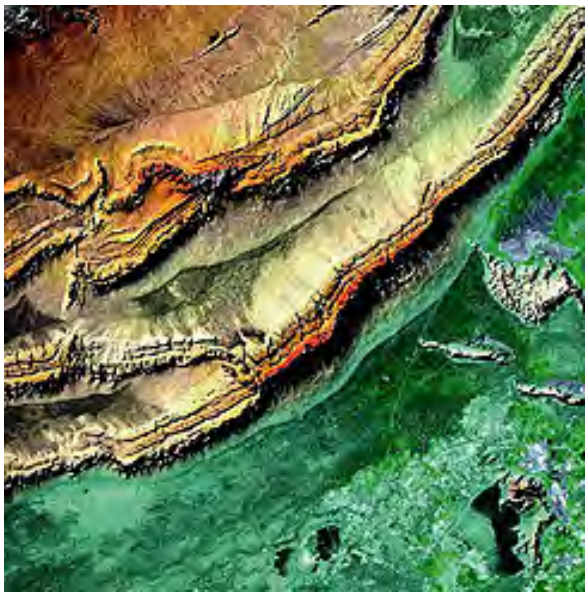
In the legend on the right, blue indicates open water, yellow shows fast ice, green is assigned to somewhat deformed ice, and red denotes notably deformed ice.

The Europeans and Japanese have now flown radars on unmanned space platforms. Some information about the European missions is available in JPL's Radar Home Page under the topic Earth Resources Satellites (ERS). The European Space Agency launched ERS-1, with a complement of sensors, in July, 1991, at a nominal altitude of 800 km (500 mi). Along with a radar scatterometer, it carried a C-band, VV SAR with a fixed look angle extending from 20° to 26°. This next scene shows a color composite, made from multidate images, of the farmlands (including vineyards) along the Rhine near the city of Darmstadt.

A nearly identical SAR was on ERS-2, which launched in April, 1995. Users can construct innovative color composites from the single band, single polarization radar by using images from different dates. We show here an example of this process: a multi-date image of Sevilla, Spain, in which an ERS-1 image on Nov. 3, 1993, is assigned to red; an ERS-1 image on June 9, 1995, is green, and an ERS-2 image on June 10, 1995, is blue. The city of Seville appears in a cyan tone in the upper right, as does the Sierra de Aracena in the upper left, and agricultural fields (bright, but barren) show as red in the rest of the image.

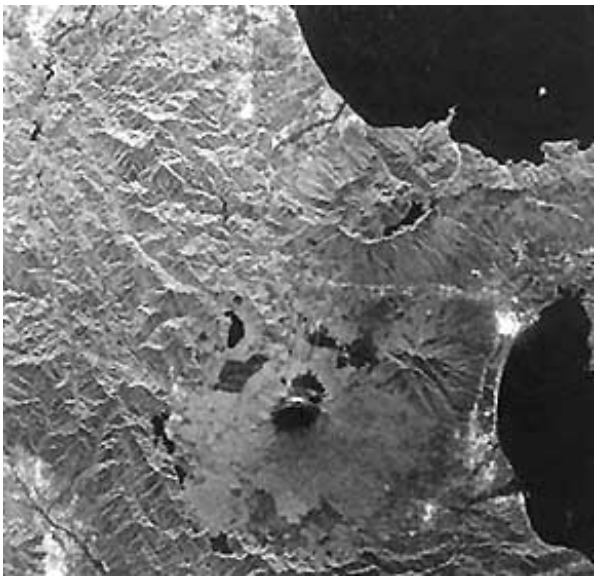


One of the most impressive radar products from ERS-2 shows tilted beds in western China:



The Russian Space program, since it entered the commercial marketplace, has released (through SovInformSputnik) some of the radar images acquired by its ALMAZ radar satellites. An example shows the mouth of the Elbe River along the northern Germany coastline:

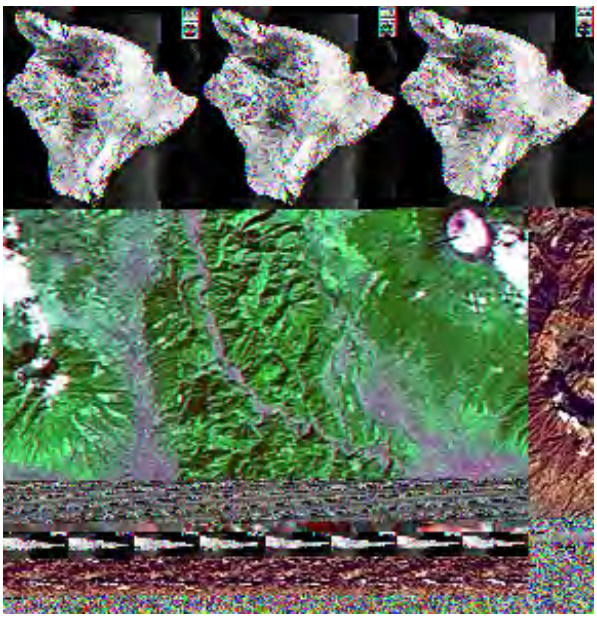
JERS-1, launched to a 570 km (354 mi) orbit on February 11, 1992, by the Japanese Space Agency, contained a seven band optical sensor and a SAR. The latter was L-band with HH polarization. It had a fixed look angle view between 32° and 38°, yielding a swath width of 75 km and a mean resolution of 18 m. One of its first images covered Mt. Fuji, a stratocone volcano west of Tokyo, as shown here.



8-20: **Here is a puzzler. I have determined that the very bright patches are the small city of Fujiyoshida. How did I do that (remember to rely on your World Atlas).ANSWER**

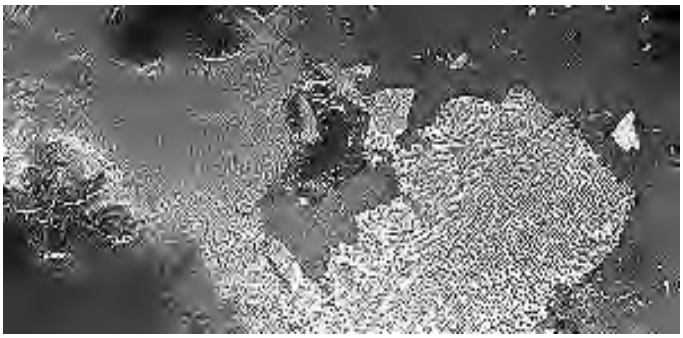
By way of comparison, here is a false color composite of Fujiyama made by the Japanese MOS-1 (Marine Observation Satellite; see page 14-12).

A spectacular JERS radar image is this mosaic scene of the Big Island of Hawaii:



The latest entry into commercially available radar imagery is the ASAR (Advanced SAR) sensor onboard ESA's Envisat (see page 16-10a for details. Two tasks of that instrument are 1) to observe sea state, and 2) to monitor high latitude ice conditions. The first is exemplified by this view of the Atlantic Ocean waters around the Canary Islands:

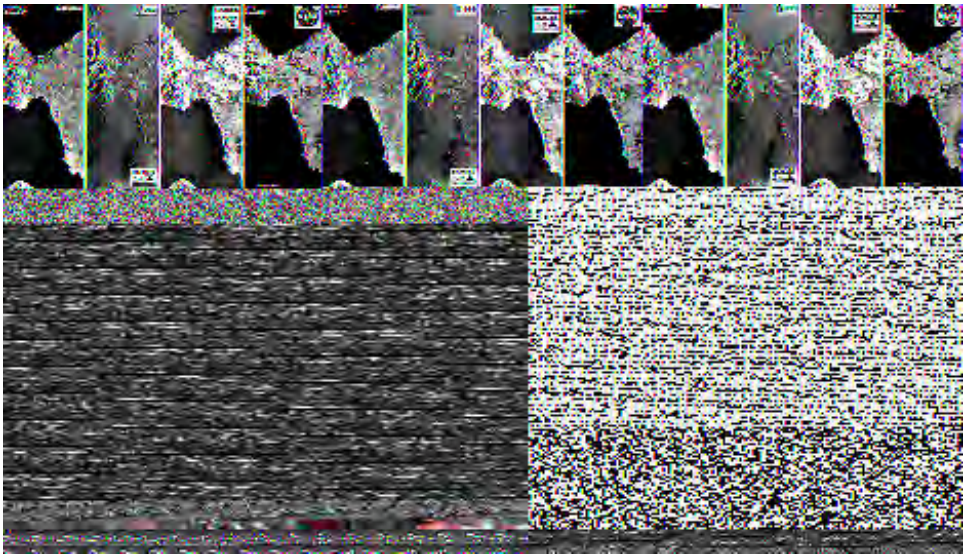
The second is illustrated by this image of ice, open water, and land in the Arctic:



And, as we have seen before in the Overview, the C-Band ASAR is capable of producing colorized images, such as this view of part of the Volga River in Russia, using different viewing modes; the image appears flat (no relief) because this is part of the steppe plains of western Russia and the Ukraine which has little variation in elevations - hence no conspicuous hilly irregularities.

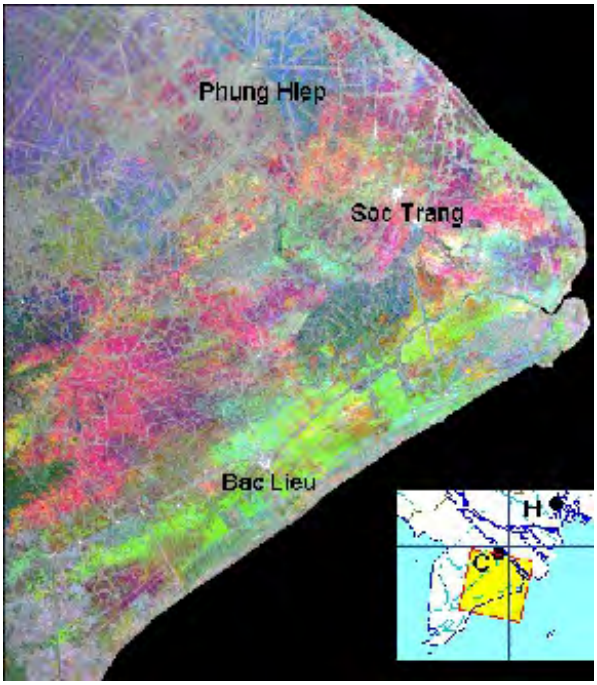
The Germans have put up their own radar satellite, named TerraSAR. Here is an image of Bonn, on the Rhine:

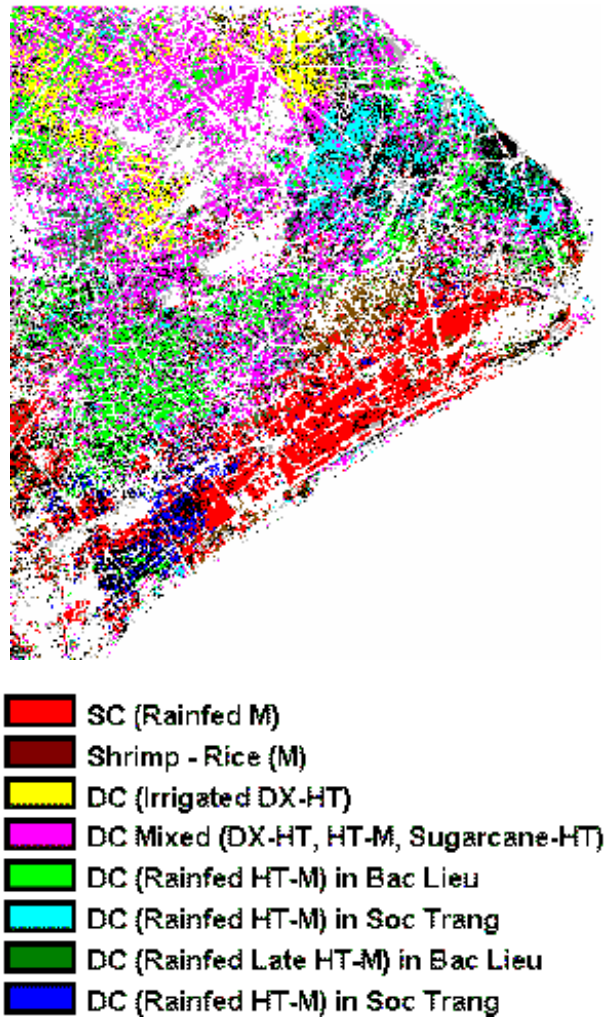
Each radar satellite tends to show a given scene in its own particular way, depending on band(s) used, Look Angle, polarization and other factors. Compare this small part of the island of Maui in the Hawaiian Islands as seen by (left to right) Radarsat, ERS-2, JERS-1, and SIR-C.



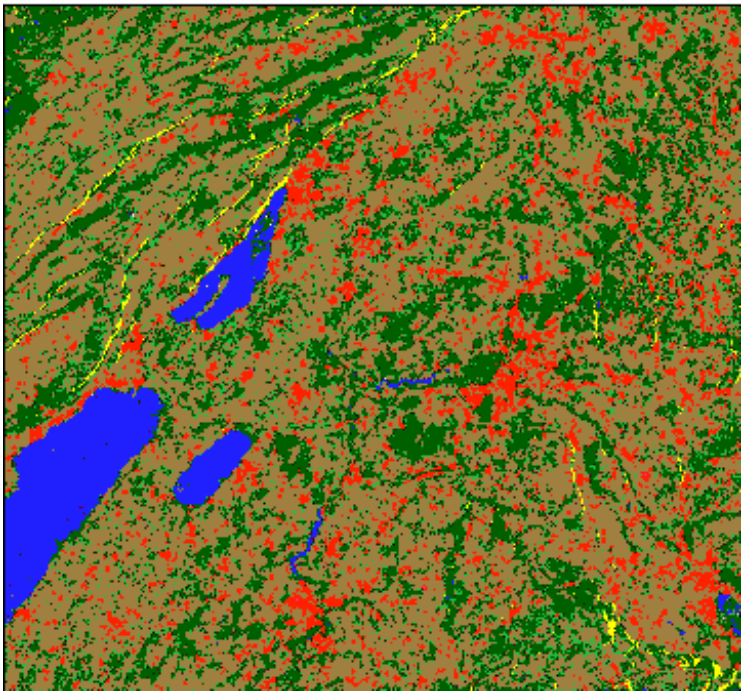
The advent of radar systems into space, following their effective demilitarization worldwide, provides the remote sensing communities with a powerful source of environmental and mapping data that are obtainable over any part of the Earth. With altimeter or interferometric processing, radar presents a new capability to generate topographic maps for parts of the global land surface, viewable from near-polar orbits. Information on several aspects of ocean surface states is also a valuable payoff. The prospects of using multi-frequency, multi-polarization beams to obtain distinctive radar signatures offers another means to identify materials that are separable on the basis of dielectric constants, surficial roughness, and other properties.

Thus, classification of features and classes in a scene is feasible with multi- radar imagery. We saw two examples in Section 3. Here are two more: First, we show part of the rice-growing region around the Mekong Delta in Vietnam. The top image is a color composite of ERS-2 radar data. The bottom image is a classification of the different rice fields, at various stages of maturation (DC refers to the Double Cropping practice), which thus predicts when harvesting is optimal:





The second is a land use map of the region around Bern, Switzerland, classified using ERS SAR data:



Color coding used:

- water
- urban area
- forest 1 (dense/coniferous)
- forest 2 (open/deciduous)
- sparse vegetation
- moisture change / freezing
- mechanical cultivation
- layover area

The final page in this Section considers remote sensing of passive microwave radiation, and then a non-mirowave topic, Lidar, which uses radiation of various shorter wavelengths including visible, IR, and UV.

PREVIOUS RST NEXT

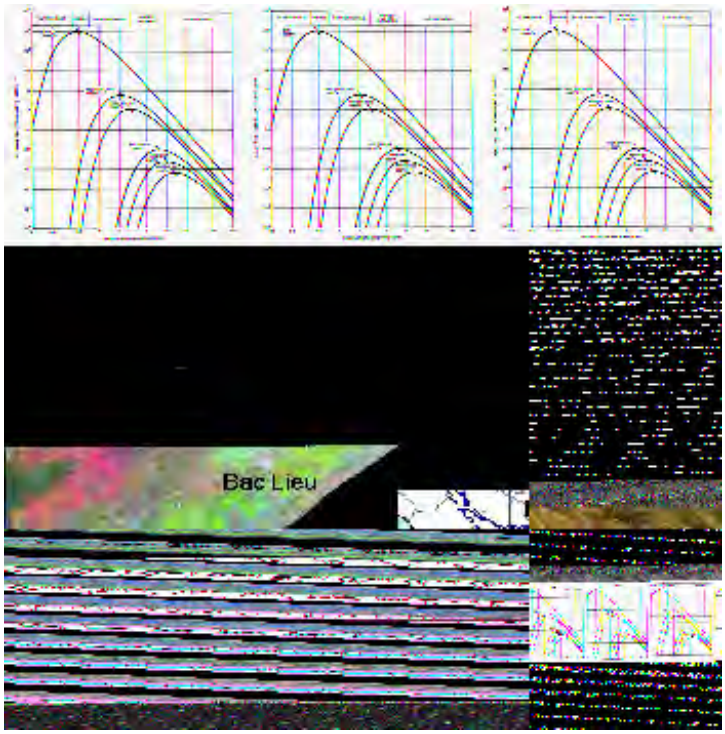
Primary Author: Nicholas M. Short, Sr.

Microwave radiation, induced by thermal heating, is emitted from the Earth's land, seas, and atmosphere. Passive microwave detectors measure brightness temperatures whose values and variations can be correlated with different materials, e.g., moisture content in soils. An application continuing over the last several decades is determination of sea ice conditions in the Arctic and Antarctic. Also reviewed on this page is a non-radar technique called ***lidar****, which depends on measuring laser light pulse round-trip times. From this altitude variations can be calculated. Besides topographic uses, lidar has proved useful in atmospheric studies, tree canopy characteristics, and oil spill detection (through induced fluorescence).*

Passive Microwave; Lidar

Although active microwave systems, i.e., radar, are the more commonly used sensors for this region of the spectrum, passive microwave sensors also have provided information about the Earth's surface, its oceans, and its atmosphere. Air- and space-borne sensors have operated for several decades. They measure directly radiation incited by thermal states in these media and hence are representative of natural phenomena inherent to the materials (hence, passive).

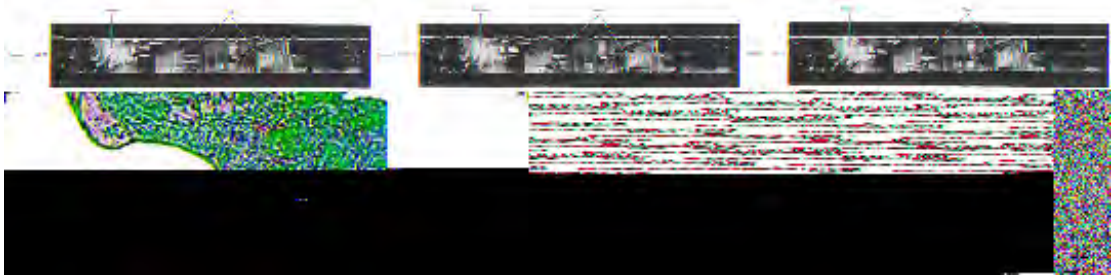
The principle underlying passive microwave radiation is implicit in the following spectral curves that show relative intensities of radiation (radiances) as a function of wavelength for materials with different intrinsic temperatures:



All of these curves have similar shapes, but as expected, the hotter the radiating object, the greater the intensity. Note, too, that the peaks of the curves shift systematically to the left as the objects increase in kinetic temperature. This shift is the consequence of Wien's Displacement Law that we examine in more detail on page 9-2 in the Thermal Remote Sensing Section. Technically, the radiation shown above is that of blackbodies at different temperatures. Natural materials are graybodies whose temperatures depart somewhat from perfect

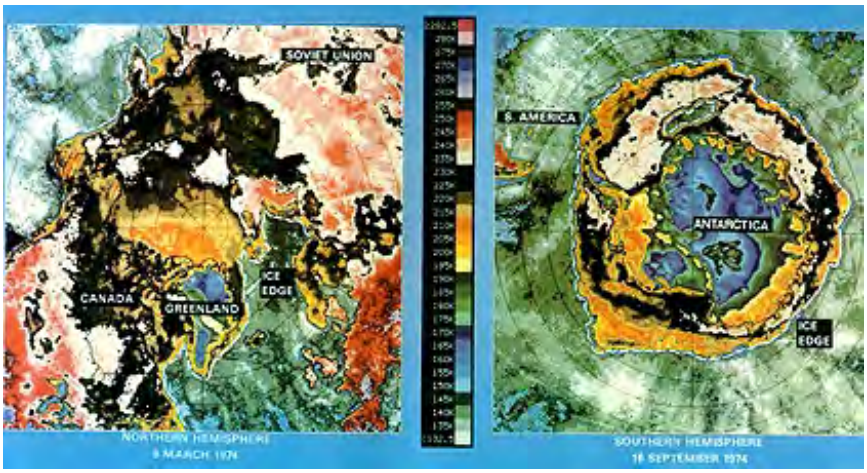
blackbodies. The important point here is that there is radiation from thermal bodies even at longer wavelengths (right part of the curves and beyond) that extend into the microwave region. This radiation, which is emissive, is generally much weaker in intensity compared with shorter wavelength outputs but is still detectable by sensitive instruments and also is not much attenuated by the atmosphere. The temperatures measured by these instruments are *brightness temperatures*. . In this family of curves, land, water, air, and ice have different brightness temperatures and thus we can separate them in many cases, and sometimes uniquely identify them.

The wavelength segment of the blackbody curves employed in passive microwave detectors is generally between 0.15 and 30 cm. In frequency units (the normal way to express a radiation interval), this interval translates to a range between 1 and 200 GHz. Frequencies most commonly used center on 1, 4, 6, 10, 18, 21, 37, 55, 90, 157, and 183 GHz (thus the multispectral mode is feasible), but the signal beamwidth is usually wide, in order to gather sufficient amounts of the weak radiation. The spatial resolution of the instrument also tends to be low (commonly, in kilometers from space and in meters from aircraft-mounted sensors), allowing the sensor to work with large sampling areas that provide enough radiation for ready detection. The sensors are typically radiometers that require large collection antennas (fixed or movable). On moving platforms, the fixed antenna operates along a single linear track so that it generates intensity profiles rather than images. Scanning radiometers differ by having the antenna move sidewards to produce multiple tracking lines. The result can be a swath, in which the variations in intensity, when converted to photographic gray levels, yield images that resemble those formed in visible, near-IR, and thermal-IR regions. Here is an example:

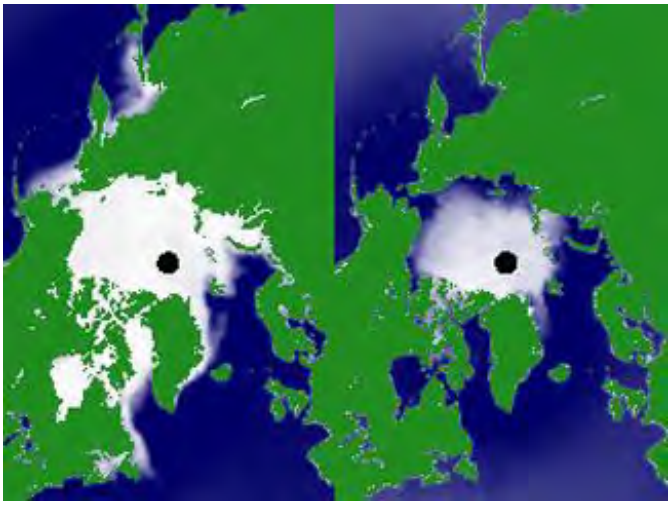


From T.M. Lillesand and R.W. Kieffer, Remote Sensing and Image Interpretation, 2nd Ed., © 1987. Reproduced by permission of J. Wiley & Sons, New York.

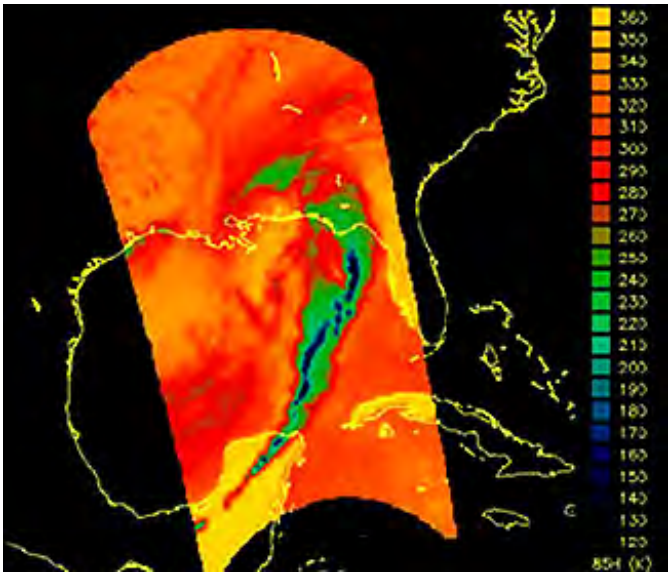
On land, passive microwave surveys are particularly effective for detecting soil moisture and temperature, because of sensitivity to water. Microwave radiation from below thin soil (overburden) cover is contributed by near-surface bedrock geology. Assessing snow melt conditions is another use. Tracking sea ice distribution and conditions is a prime oceanographic application. Another marine use is for assessing sea surface temperature. Passive microwave sensors are important components on some meteorological satellites, being well suited to obtain temperature profiles through the atmosphere, as well as water vapor and ozone distributions and precipitation conditions. The ESMR and SMMR scanning microwave radiometers flown on Nimbus metsats are described on page 14-4. Here is an ESMR (Electronically Scanned Microwave Radiometer), single band (19.3 GHz) image received from the Nimbus 5 metsat that provided sea ice data for both polar regions. Open water appears in gray-green tones.



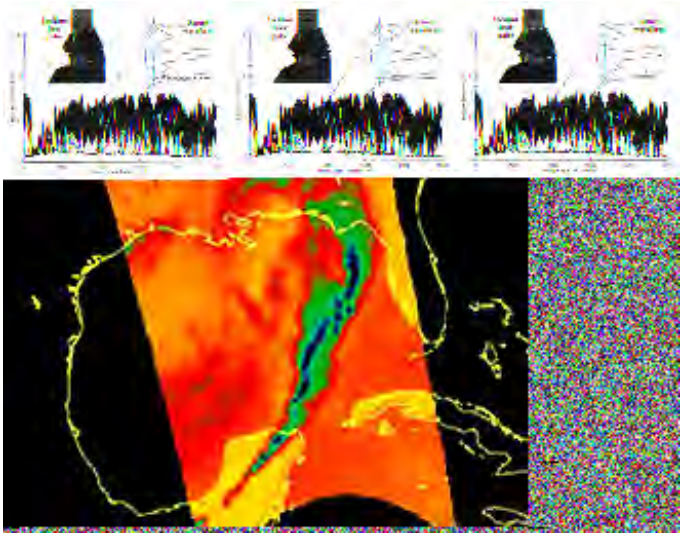
Several other satellites, e.g., TRMM, that use passive microwave to learn more about meteorological conditions, are discussed on page 14-5. One of particular versatility is the SSM/I (Special Sensor Microwave/Imager) on the DMSP series of Dept. of Defense satellites. It can measure such variables as wind velocity, soil moisture, and rainfall. One of its prime functions, useful for both military and civilian needs, is to monitor sea ice, much in the manner of the ESMR discussed above. Here is a SSM/I-derived map of polar ice in the Arctic, with the left panel showing winter distribution and the right depicting summer ice.



SSM/I uses several of its channels to determine brightness temperatures over land and sea. This next image shows the variation of such temperatures (in degrees Kelvin) over land in the southern U.S. and the waters of the Gulf of Mexico.

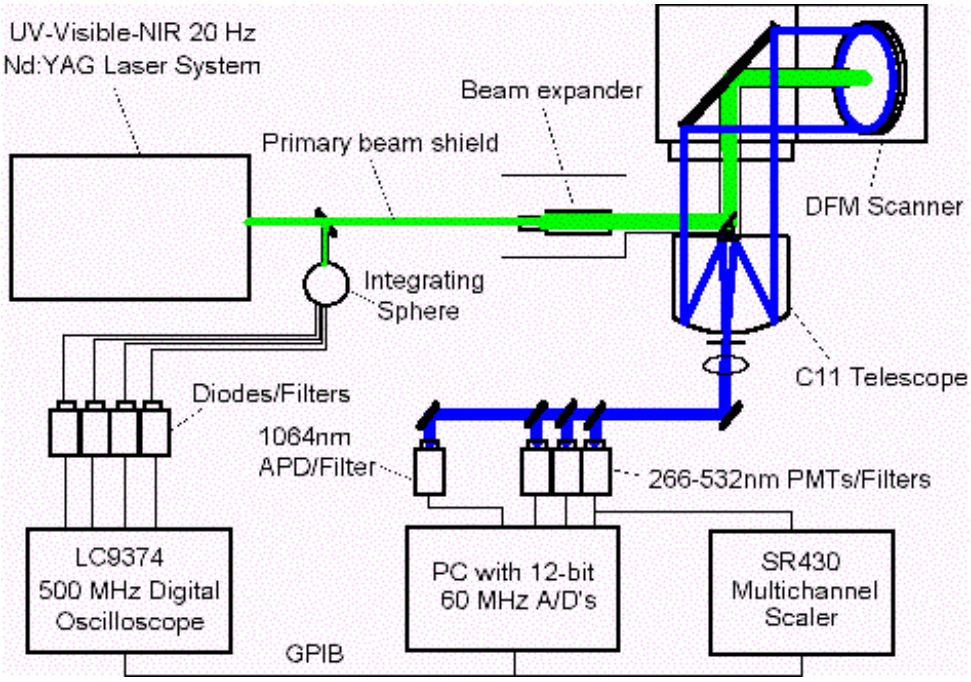


Another active sensor system, similar in some respects to radar, is Lidar (light detection and ranging). A Lidar transmits coherent **laser** light, at various visible or NIR (Near-IR) wavelengths, as a series of pulses (100s per second) to the surface, from which some of the light reflects. In this sense, it is similar to radar. Travel times for the round-trip are the measured parameter. We can operate Lidar instruments as profiliers and as scanners, day and night. Lidar serves either as a ranging device to determine altitudes (topography mapping) and depths (in water) or as a particle analyzer within the atmosphere. Light penetrates certain targets, so that one prime use is to assess tree canopy conditions. Returns are obtained from tree tops, from within the trees, and from the ground, as shown in this diagram:

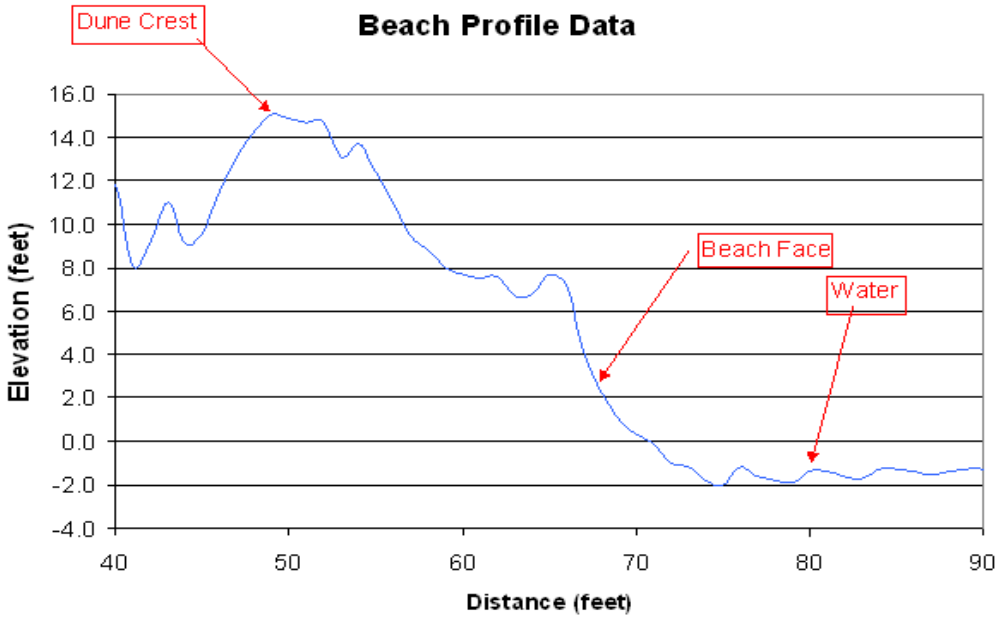


Since Lidar images are not used elsehere in the Tutorial, we confine coverage to the rest of this page. A Google search on the Internet will turn up considerable relevant information on Lidar. Two Wikipedia websites give good overviews about Lidarand Laser.

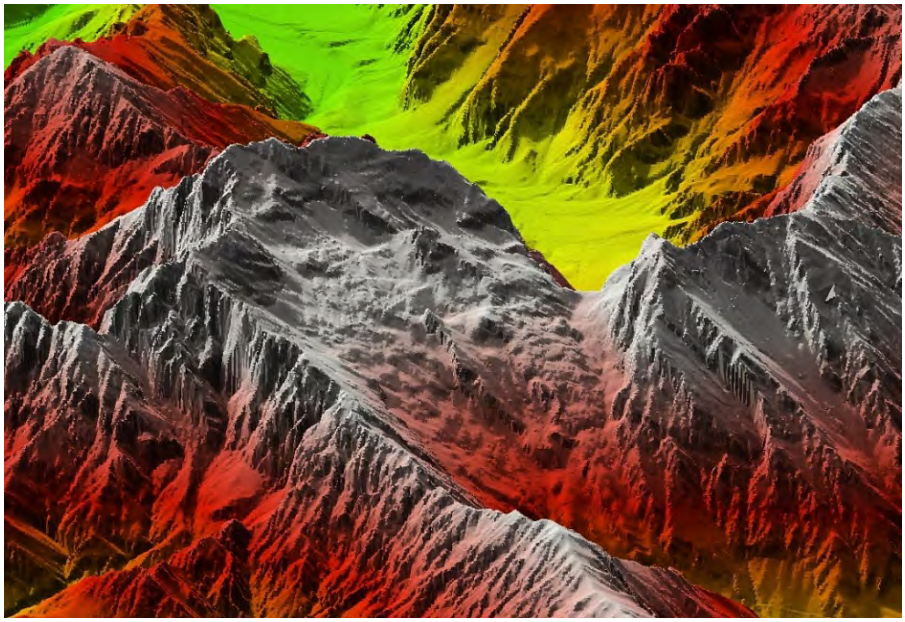
Lidars flown on aircraft function as scanners. The "footprint" of a Lidar can be either circular or linear. This is a diagram of a typical scanning Lidar system:



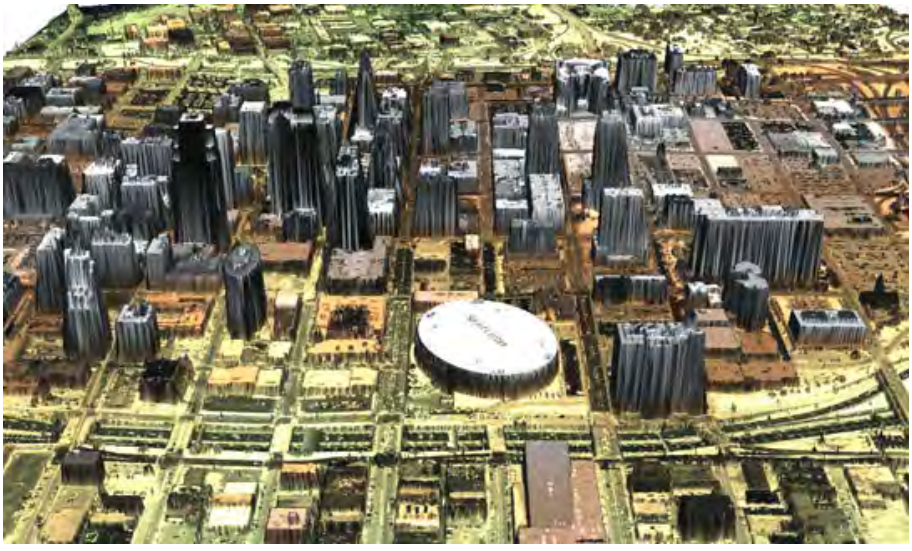
When used in the scanning mode, the platform (commonly aircraft) will follow a series of side by side flight lines. The data can be shown as individual lines, and plotted as profiles

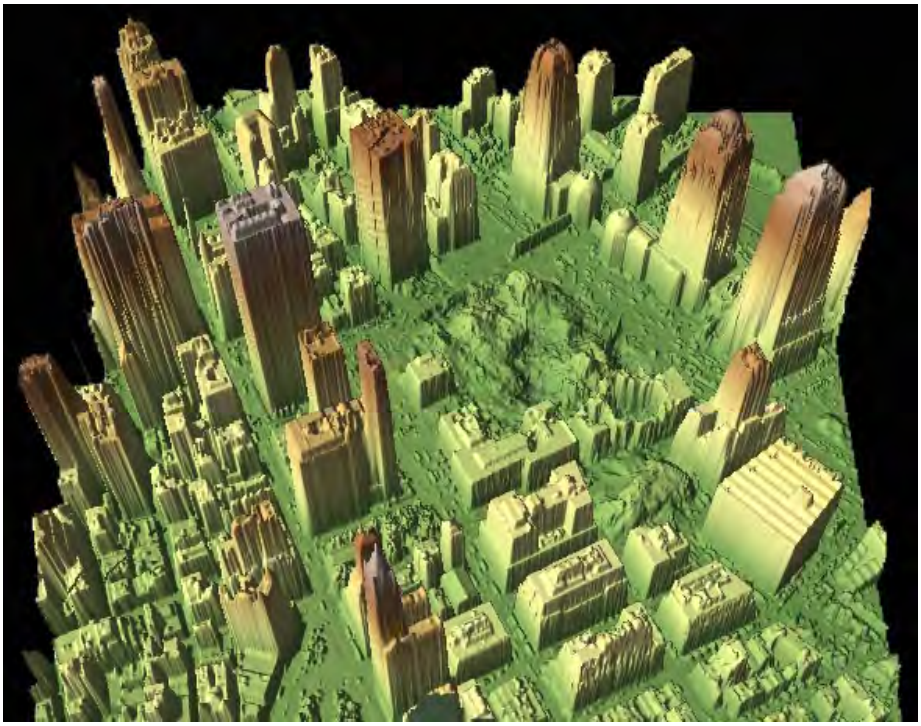


The succession of parallel lines allows for two-dimensional "maps" in which the 3rd dimension appears as a sense of height. Here is a Lidar image that shows topography; the colors are arbitrarily assigned (usually the practice, as Lidar does not operate as a multispectral image).

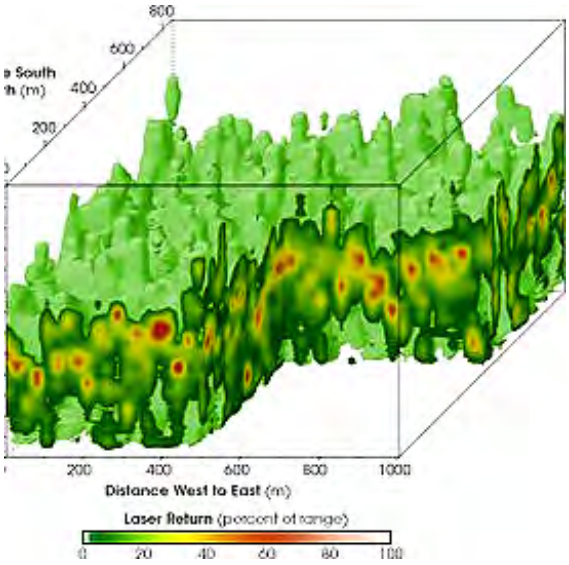


The 3-D effect is visually striking when cities are imaged. Two examples are Kansas City, MO and Lower Manhattan (New York City) after the World Trade Center bombing



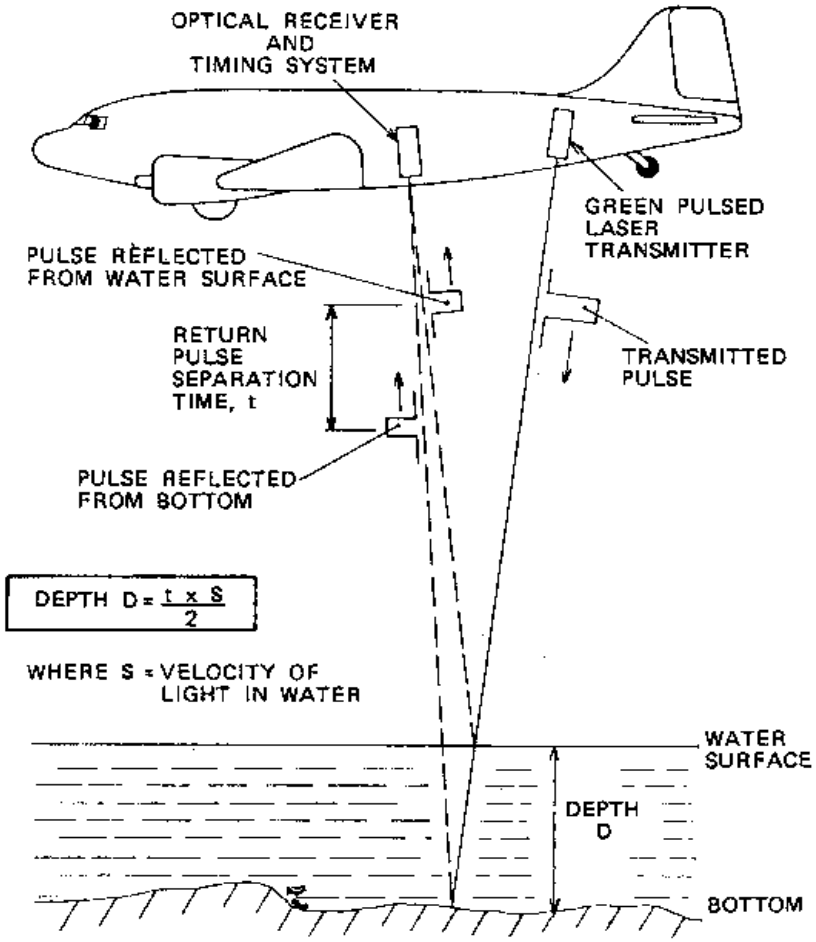


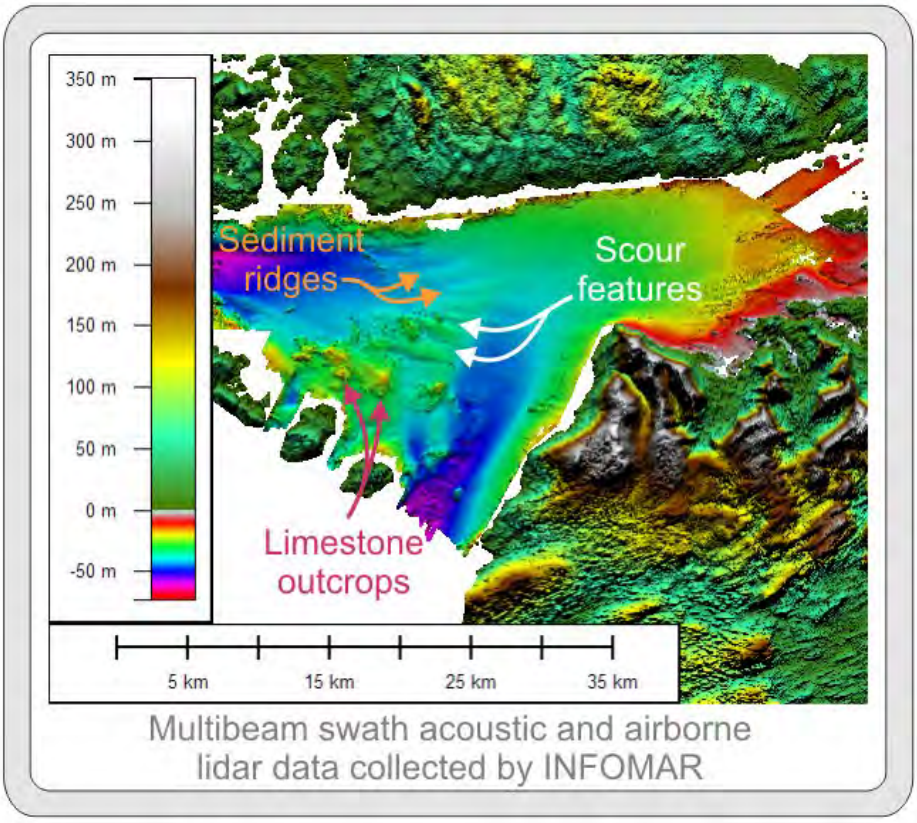
Height data from Lidar measurements can be shown visually in block diagrams such as this, which shows a small part of the forest near La Selva, Costa Rica:



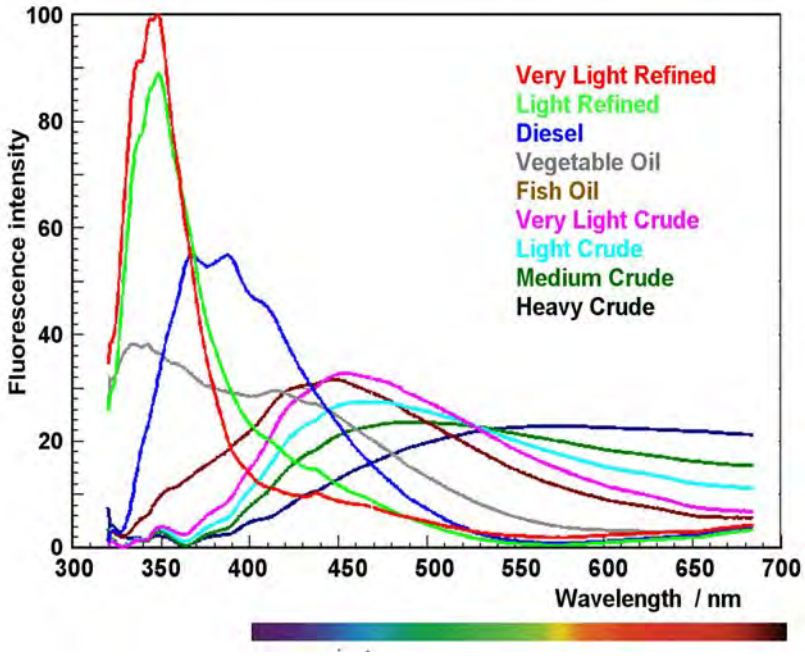
We can interpret these data to indicate the amount of biomass associated mainly with the leaves. This information is important in determining the global condition of vegetation that governs production of CO_2 and O_2, key factors in sustaining life, and now a great concern because deforestation in the tropics and temperate zones depletes these gaseous resources.

Lidar can also penetrate shallow water bodies to give information (usually as profiles) on water depths. The difference in time delay between returns from the surface and from the water bottom (from which returns will be weaker) indicates the thickness (depth) of the water column at any point. This is shown in the next diagram, followed by a maplike rendition of land elevation and water depth near Galway in Ireland:



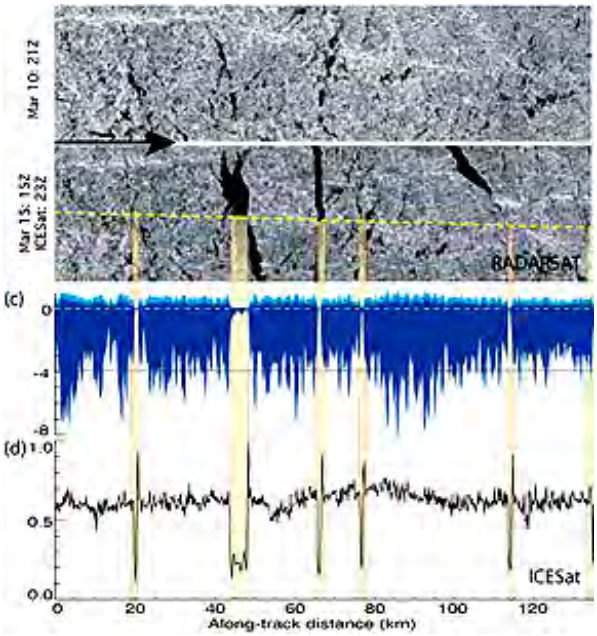


Certain Lidar wavelengths cause materials to fluoresce (emit light radiation at different wavelengths than that of the incoming beam), which tuned detectors can pick out. Oil slicks respond in this way, and chlorophyll in sealife also fluoresces. The diagram below indicates that different oils have different spectral response curves:

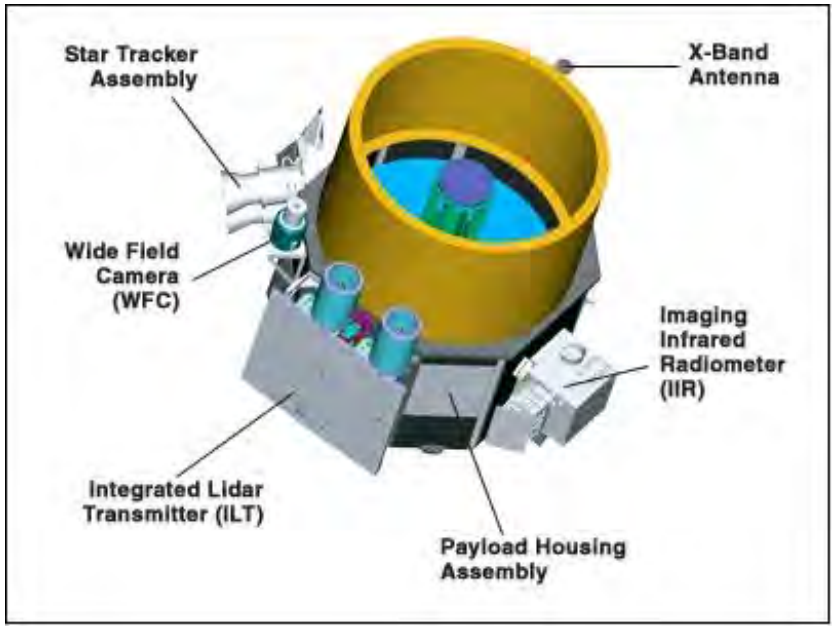


Lidar operates normally from aircraft platforms but a spaceborne sensor (LITE, for Light In-space Technology Experiment) was flown on the Shuttle in 1994; it monitored clouds and measured atmospheric particles. Since then several other Lidar systems have been flown in space Another Lidar used in space was Alissa, onboard the Russian MIR station. If approved, the Vegetation Canopy Lidar, will fly on the first of a series of low-cost satellite missions in NASA's new Earth System Science Pathfinder (ESSP) program.

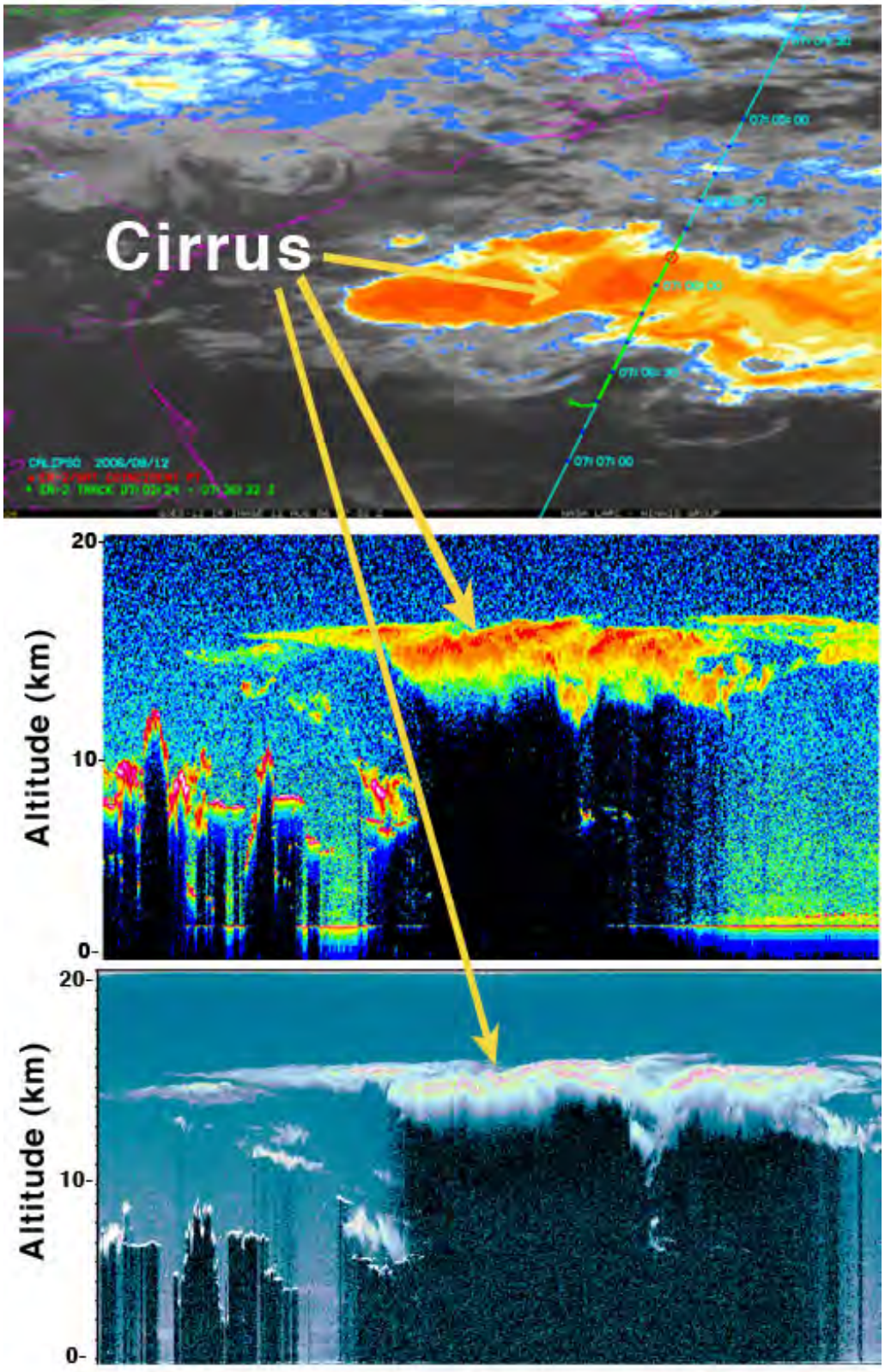
A laser altimeter (GLAS) was one of the sensors launched on January 13, 2003 in the EOS series (see page 16-7). Its role is covered at the University of Texas GLAS website. GLAS is being used mainly to monitor seaice but it also gathers meteorological data. Here is ice profile made by GLAS; the images are areas it traversed, as seen by MODIS:



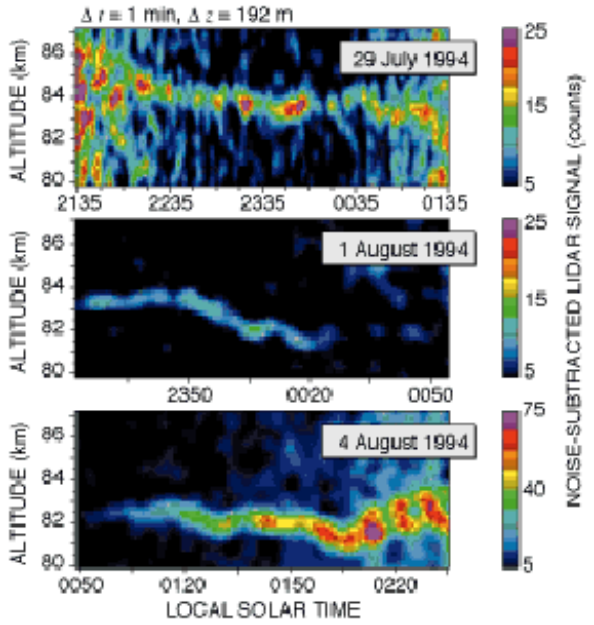
A Lidar instrument was onboard Calipso (Cloud Aerosol Lidar and Infrared Pathfinder Satellite Observation) launched on April 28, 2006. Here is the satellite and a drawing showing its payload.

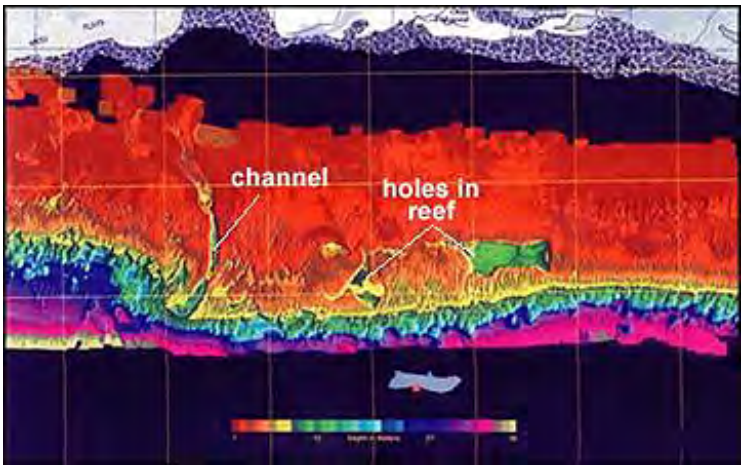


Calipso data are displayed in several formats. The diagram below shows a Calipso cross-section of aerosols in the atmosphere (middle), with a GOES satellite image of the same orbital path, indicating cloud cover (top), and a plot of Lidar data from a simultaneous aircraft (ER-1) underflight (bottom):

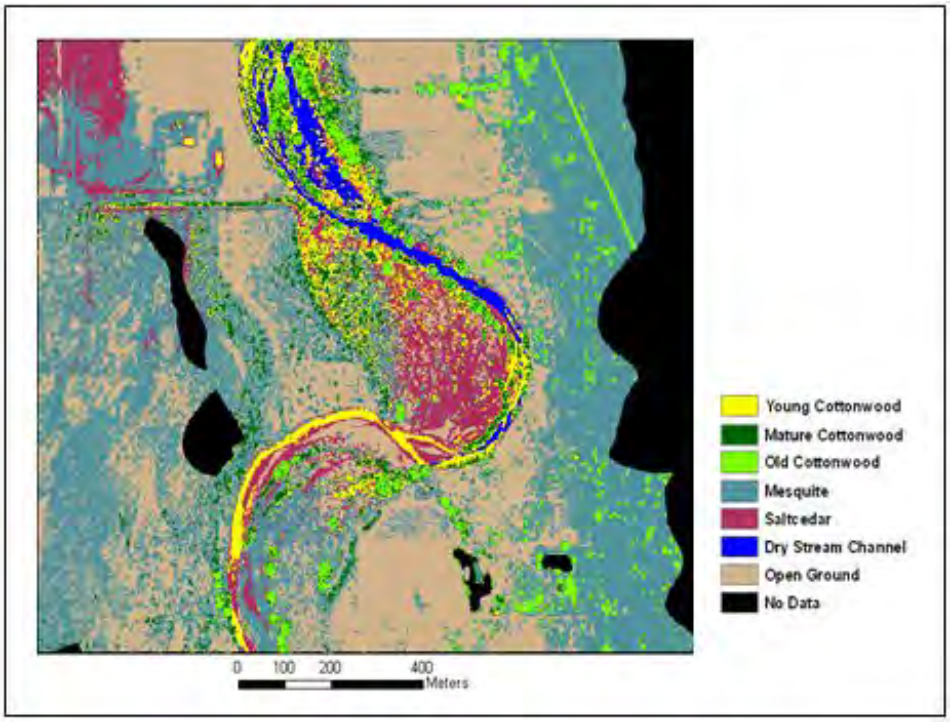


Here are three more diverse examples of Lidar use; read the captions for descriptions.





As experience with Lidar as a versatile remote sensor increases, new modes of use are coming online. By using lasers with different wavelengths, one can now produce multiband color images and can classify the features in an image. Consider this example:



Lidar is finding many applications that "ordinary people" can utilize. You may have encountered a Lidar use if you were stopped by the police using Lidar instead of radar.

We have reached the end of this Section, hoping that you have come to the realization that the microwave region of the spectrum can do many useful things, besides cook your food. Radar, in particular, is now a mainstay for both civilian and military space observations, being invaluable because of its independence from most weather conditions (but a reminder: radar is a powerful tool in weather forecasting). As experience is gained, we are learning how to conduct feature classifications offering much the same information content as extracted from Visible and Near-IR imagery. Next, we need to become acquainted with the power of thermal remote sensing for identifying classes on the ground and conditions in the oceans.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The Earth's surface and atmosphere radiate thermal energy outward owing to heating by solar irradiation and by internal heat flow (generally a very minor contribution). Sensors that measure this emitted radiation in parts of the thermal region of the spectrum can produce very informative data (especially as images) that provide both distinctive signatures and ,indirectly, indications of properties of materials that are diagnostic. The atmospheric windows passing thermal radiation are indicated. The Planck Blackbody Radiation Law, the basis for thermal remote sensing, is presented.

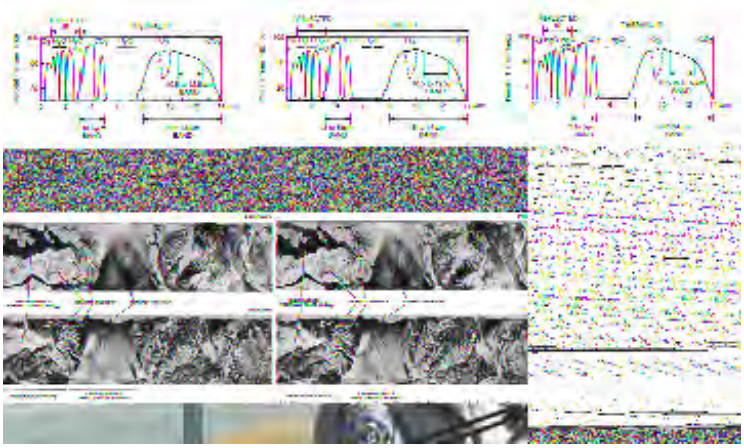
THE WARM EARTH: THERMAL REMOTE SENSING

Before you start this section, you might profit from a look at the brief vignette on thermal infrared imaging as produced by JPL. As always, access through the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes "Infrared: More than your Eyes can See", January 13, 2003. To start it, once found, click on the blue RealVideo link.

We have already shown several previews of imagery that depict the thermal state of the Earth's surface. The Landsat Thematic Mapper Band 6 produces images that show the relative differences in emitted thermal energy from surfaces of varying nature and orientation. On page 1-3, we showed this effect in the obviously warmer slopes in the Morro Bay scene and in Section 2, page 2-4 the surprisingly cool surfaces of the bright, white sandstone at the Waterpocket Fold. In seeming contradiction, the warmer surfaces associated with dark shales showed clearly in the Waterpocket Fold image. As we shall demonstrate, these thermal effects also differ, sometimes drastically, in images taken during the warmer days and cooler nights. This is clearly the case in this airborne thermal scanner image of a valley-mountain terrain in southern California, in which the same features may appear in sharply contrasting tones between the dawn and day scenes:

One usually obtains considerable information from thermal imagery, especially that generated as multispectral images. Some of this information supplements images obtained from reflected radiation and some stands alone as an information source. In this Section review, we delve rather extensively into the theory and practice of thermal remote sensing and examine some striking examples obtained from land and water targets. However, we postpone our discussion of using thermal sensors to obtain temperature profiles or find other thermal properties in the atmosphere until we consider meteorological satellites (Section 14).

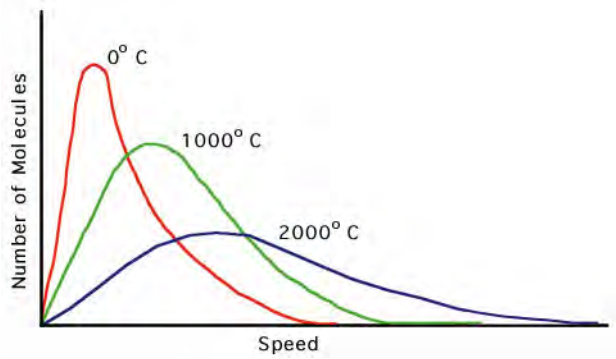
Remote sensing of direct temperature effects is carried out by sensing radiation emitted from matter in the thermal infrared region of the spectrum. Most thermal sensing of solids and liquids occurs in two atmospheric windows, where absorption is a minimum, as shown in this spectral plot taken from Sabins (Remote Sensing: Principles and Interpretation, 1987). Note that the segment of the spectrum labeled Reflected IR is considered to be non-thermal, although objects viewed in that range have discrete temperatures.

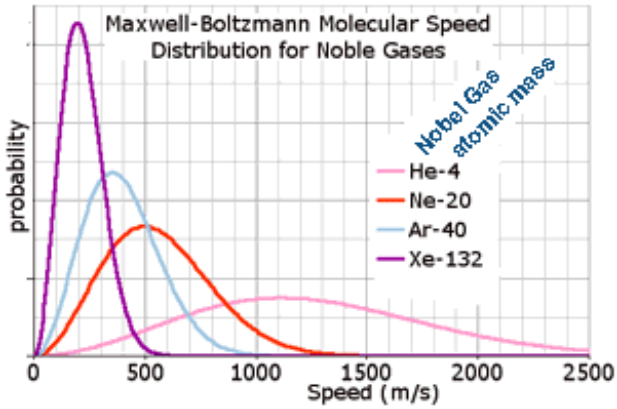


The windows normally used from aircraft platforms are in the 3-5 µm and 8-14 µm wavelength regions. Some spaceborne sensors commonly use transmission windows between 3 and 4 µm and between 10.5 - 12.5 µm. None of the windows transmits 100 percent because water vapor and carbon dioxide absorb some of the energy across the spectrum and ozone absorbs energy in the 10.5-12.5 µm interval. In addition, solar reflectance contaminates the 3-4 µm window to some degree during daylight hours, so we use it for Earth studies only when measurements are made at night.

The Concept of Temperature

Temperature is a measure of the motion of atoms and molecules in a substance. At absolute zero (-273 degrees Kelvin) all atomic/molecular motion has ceased. As a body is heated, its constituent particles begin to vibrate over very short distances (submicroscopic). This motion can be expressed in terms of velocity (speed); for gases, in particular, individual molecules move at various kinetic energies (K.E. = $1/2mv^2$). The Maxwell-Boltzmann Distribution function is a probability plot of the range of speeds. It can be applied to a specific gas for different temperatures (top plot) or to several gases of different composition at the same temperature:





Particle motions occur in gases, liquids, and solids which can be heated to various temperatures. For gases, temperature T plays a role as expressed by the Perfect Gas Law PV = nRT, where P is pressure, V is Volume, n is the number of moles of the gas in a specific sample (n = m/M in which m is the mass involved and M is its molecular weight), and R is the Universal Gas Constant (R = 8.31 Joules/mol/deg). Another relationship that directly involves the Kinetic Energy of a gas has two expressions, the second derived from the first: 1) $T = 2/3(N_m \ln R)(1/2m_0\{v^2\}_{av})$, where N_m is the number of molecules in a given sample and ln refers to the natural log (of R); this equation states that the temperature of the gas (K) is proportional to the mean kinetic energy (summed from the range of velocities) of translational motion per molecule of the gas; 2) replacing $(N_m \ln R)$ with k, the Boltzmann constant, one gets: $T = (2/3k)(1/2m_0\{v^2\}_{av})$; $k = R/N_A = 1.38 \times 10^{-23}$(Joules/Kelvin)/molecule (note: N_A is Avogadro's Number given as 6.02×10^{23} molecules per mole of any gas).

Nothing specific about temperatures in solids and liquids will be discussed here but some information about this can be found at these Wikipedia and UCAR sites.

Planck Radiation (Blackbody) Law

The concept of a *Perfect Blackbody (BB)* relates to an ideal material that completely absorbs all incident radiation, converting it to internal energy that gives rise to a characteristic temperature profile. Therefore a BB does not participate in any transmittance or reflectance but emits (re-radiates) the absorbed energy at the maximum possible rate per unit area. The amount of this radiant energy varies with temperature and wavelength(s).

In 1900, Max Planck published a paper on the thermal properties of blackbodies that became one of the foundation stones from which quantum physics was built on. The Planck Radiation Law gives the rate at which Blackbody objects radiate thermal energy:

$$E(\lambda, T) = \frac{2\pi hc^2}{\lambda^5 (e^{hc/\lambda kT} - 1)}$$

where [E] = $W/m^2/m$

λ ... Wavelength
c ... Speed of light
k ...Boltzmann's constant
h ...Planck's constant
e = 2.718 ...Base of Natural Logarithms

[T] = Kelvin
[λ] = meters
$h = 6.626 \times 10^{-34}$ Js
$c = 2.998 \times 10^{8}$ m/s
$k = 1.381 \times 10^{-23}$ J/K

Depending on what terms (parameters) and unit systems are used, the Planck equation can be written in various ways. As often used in remote sensing calculations, this differential form is given as:

$$P_\lambda d\lambda = \frac{c_1 \lambda^{-5}}{e^{c_2/\lambda T} - 1} d\lambda$$

$$\begin{aligned} c_1 &= 2\pi hc^2 &&= 3.74183 \times 10^{-16} \mathrm{W\,m^2}, \\ c_2 &= hc/k &&= 1.4388 \times 10^{-2} \mathrm{m\,K}. \end{aligned}$$

where P_λ = E_λ = spectral emission in $W/m^2/m$ at a wavelength λ.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

This page reviews several of the fundamentals of thermal behavior of materials. The Wien Displacement Law demonstrates that peak radiative intensity will occur at different wavelengths depending on the temperature of the radiating body. The concepts of emissivity and of blackbody radiation are discussed. A diagram shows that the radiant temperature measured by a sensor will vary considerably if emissivities are sharply different, even if the different bodies have similar kinetic temperatures.

The Wien Displacement Law and Emissivity Effects

The above figure (also shown on page 8-8) plots spectral wavelength versus emitted radiance (as intensity) from thermal radiators at various peak radiant temperatures ranging from that of the Sun to the Earth's surface (average ambient temperature and sea ice). The hotter the radiating body, the greater is its radiance (intensity on the ordinate) over its range of wavelengths, and the shorter is its peak emission wavelength. The relation between peak wavelength and radiant body temperature is the Wien Displacement Law:

$$\lambda_m T = 2898$$

where λ_m is the wavelength at maximum radiant emittance and T is the absolute temperature expressed in degrees Kelvin (°C + 273). The constant, 2898, is in units of μm °K. It is also given

as (rounded off) 0.29 cm °K. For the Sun, with a photospheric radiant temperature of about 6000 °K, this peak is in the visible (centered on 0.58 µm). A forest fire peaks around 5.0 µm. The Earth's surface, as observed from space, peaks within the 8-14 µm interval.

9-1: **What is the peak wavelength for a lamp that glows at 1800° C? ANSWER**

A parenthetical comment about the Sun's color, as human's see it. We perceive it as yellow-orange but from the above diagram its peak is within the green visible range. Why then doesn't it appear green? The answer relates to the eye's response to a mix of colors. For the Sun all visible colors are emitted, with green slightly dominant. To an astronaut in orbit above the atmosphere, the Sun appears almost white - a mix of all visible light waves. Steel heated to 6800 °C would glow in a near white color. But on Earth the astronaut's retina would describe it as orange, since his eye cannot separate the stronger influence of green from adjacent wavelengths in the different colors since all have nearly the same intensity. (For the Sun to actually glow as a green incandescent body, the wavelengths on either side of the green region would have to be much less intense, i.e, depart from a radiant Black Body.) The orange results from subtraction of blue and blue-green owing to selective scattering of those wavelengths by the atmosphere - hence they are largely removed, shifting the color into orange. Even that color varies, as the Sun extends low above the horizon after dawn or around sunset, since the greater density of air as one looks near the horizon causes further scattering, leading to a more reddish Sun (which also seems larger, as does the Moon near the horizon, owing to refraction-controlled bending differences).

A body's temperature can represent one thermal state but be expressed by two temperatures: the first is its internal temperature (from the kinetic motion of its atoms) as measured by an inserted thermometer whereas the second is the external temperature measured by its emitted radiation. The radiant flux F_B (rate of flow of EM energy, commonly measured as Watts [W - a unit of power; 1 Watt = 1 Joule per second] per square centimeter) emanating from a blackbody is related to its internal (kinetic) temperature T_k (temperature in Kelvin units) by the Stefan-Boltzmann Law, which in simplified form is given as $\mathbf{F_B = \sigma T_k^4}$, where σ (the Greek letter, small sigma, σ, or "s") is a constant given as 5.67×10^{-12} W(atts) × cm^{-2} × K(elvin)$^{-4}$. Strictly, this equation holds only for perfect blackbodies; for other bodies (so-called real or "graybodies"), the radiant flux will always be less than the blackbody flux, as calculated by $\mathbf{F_R = \varepsilon \sigma T_k^4}$, where ε is defined in the next paragraph.

9-2: **Calculate the value of F_r for a blackbody having a temperature of 17° C. ANSWER**

The quantity of radiant emission, and thus the effective temperature that is measured externally as radiation, also depends on the emissivity **ε** (the small Greek letter "epsilon") of the object in the spectral region of interest. Emissivity is a dimensionless number that expresses the ratio of the radiant flux of a real material F_R to the radiant flux of a perfect blackbody F_B (one that completely absorbs incoming radiant energy, with none being partitioned into transmitted or reflected components), or $F_R/F_B = \boldsymbol{\varepsilon}$. It is a measure of the efficiency of emitted radiance of any real body to that of a perfect radiator (for which **ε** = 1.0). Values of **ε** vary from 0 to 1 and are spectrally dependent, i.e., can change with λ. Here is an example comparing the spectral radiant emittance of the common mineral Quartz to a perfect B.B. when they are at thermal equilibrium at a given temperature (here, at 600 °K).

From T.M. Lillesand and R.W. Kieffer, Remote Sensing and Image Interpretation, 2nd Ed., © 1987. Reproduced by permission of J. Wiley & Sons, New York.

9-3: **What is the radiant flux F_R for a real body having the same kinetic temperature as in question 9-2, and also having an emissivity of 0.9.** **ANSWER**

The sharp decrease in ε_λ in the 8-10 µm region, noted for quartz and other silicates, is a "reststrahlen" effect (decreased emission) related to thermally induced stretching vibrations within silicon-oxygen bonds. In general, for opaque materials, $\boldsymbol{\varepsilon_\lambda = 1 - \rho_\lambda}$, where ρ (Greek rho) is the material's optical reflectance. Therefore, as ρ_λ ----> 1, with high reflectance of radiation (poor absorptance), the emittance will be lowered (thus, thermal radiation decreases). Water, which has a high emissivity in the thermal infrared in the 8-10 µm interval, is a poor reflector over that range; quartz (and many silicate rocks) is a good emitter at lower thermal wavelengths but poor in this interval. From this, one might predict that rock surfaces would appear darker than water in the 8-10 µm interval but this holds only for certain conditions, as we will shortly see.

The radiant (sensed) temperature T_R differs from a body's kinetic (internal) temperature $T_K$according to the relation $\boldsymbol{T_R = \varepsilon^{1/4} T_K}$;as stated above, for real bodies (graybodies) radiant temperatures are always less than kinetic temperatures. Thus, checking the figure below, the radiant temperature is significantly higher for a blackened surface (high **ε**) than for a shiny surface (lower **ε**), even if the two materials are at the same kinetic temperature.

From F. F. Sabins, Jr., Remote Sensing: Principles and Interpretation. 2nd Ed., © 1987. Reproduced by permission of W.H. Freeman & Co., New York City.

9-4: **For the real body we considered in question 9-3, what is its radiant temperature?**
ANSWER

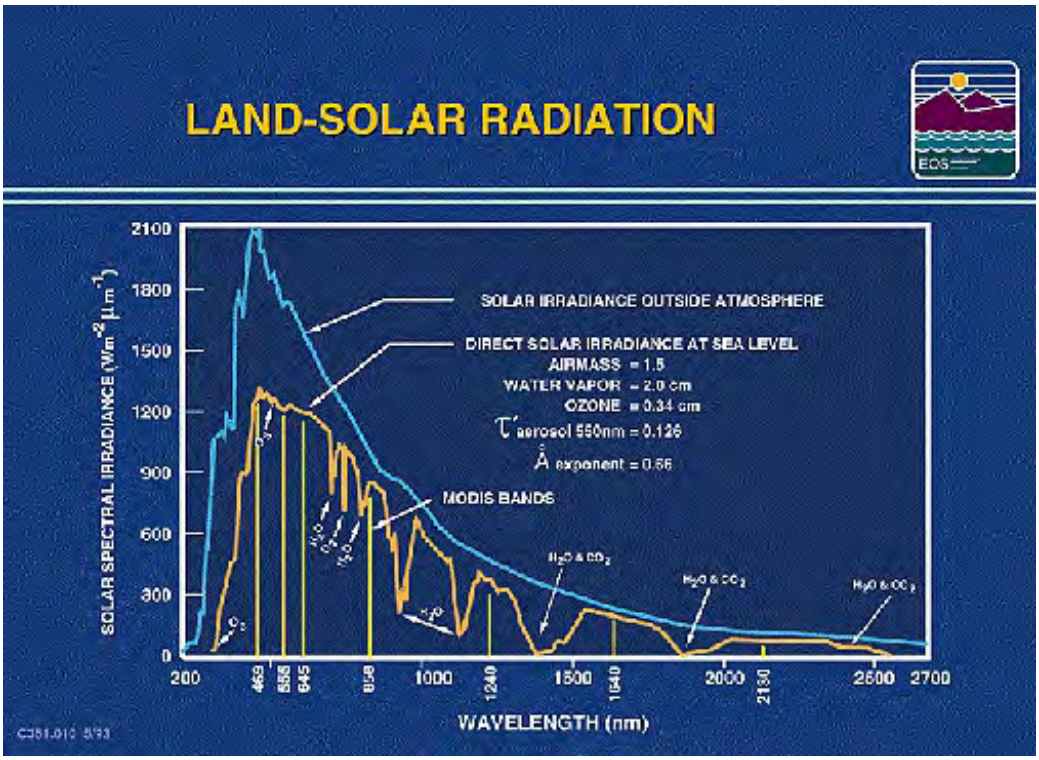
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

A plot indicates that the incoming solar radiance, distributed over the Visible-Near IR, presents one curve for its measurement at the top of the atmosphere and a second, different (reduced intensities) after the irradiation reaches the Earth's surface. The meaning of several thermal parameters - Heat Capacity; Thermal Conductivity and Diffusivity, and Thermal Inertia - is explored. A list of factors that influence variations in these parameters and the measured thermal radiances is outlined.

Solar Irradiation as a Heating Mechanism

The amount of solar radiation reflected from land and sea surfaces, as well as the amount absorbed, depends partly on that portion of energy from the Sun that reaches these surfaces. On page 9-2, we stated that in going from lower to higher wavelengths, this radiance rises rapidly to a peak at 480 nanometers (0.48 µm), then trails off to near zero through wavelengths out to about 2600 nanometers (2.6 µm). The plot below confirms that distribution and also shows many of the principal water, carbon dioxide, and oxygen absorption bands.



A thermal sensor detects radiant energy from a surface target, heated through radiation (solar insolation and sky radiance), convection (atmospheric circulation) and conduction (through the ground). Most sensed heat from surfaces has its origin in solar illumination, that varies with diurnal and seasonal changes, as well as cloud cover, but there is also a small, nearly constant, contribution from internal heat flux from Earth's interior (mostly from radioactive decay). Heat transfers into and out of near surface layers because of external heating by the thermal processes of conduction, convection, and radiation.

Heat Capacity; Thermal Conductivity; Thermal Inertia

A primary objective of temperature measurements and related thermal responses is to infer something about the nature of the composition and other physical attributes of materials at the Earth's surface and in its atmosphere. For any material, certain internal properties play important roles in governing the temperature of a body at equilibrium with its surroundings.

These properties include:

- Heat Capacity (C): The measure of the increase in thermal energy content (Q) per degree of temperature rise. It denotes the capacity of a material to store heat, and we give it cgs units of calories per cubic cm. per degree Centigrade (recall from physics that a calorie [cal] is the quantity of heat needed to raise one gram of water by one degree Centigrade). We calculate heat capacity as the ratio of the amount of heat energy, in calories, required to raise a given volume of a material by one degree Centigrade (at a standard temperature of 15° Centigrade) to the amount needed to raise the same volume of water by one degree Centigrade. A related quantity, specific heat (c), is defined as $C = c/\rho$ (units are calories per gram per degree Centigrade) where ρ (rho) = density. This property associates Heat Capacity with the thermal energy required to raise a mass of one gram of water by one degree Centigrade.
- Thermal Conductivity (K): The rate at which heat passes through a specific thickness of a substance, measured as the calories delivered in one second across a one centimeter square area through a thickness of one cm at a temperature gradient of one degree Centigrade (units: calories per centimeter per second per degree Centigrade)
- Thermal Inertia (P): The resistance of a material to temperature change, indicated by the time dependent variations in temperature during a full heating/cooling cycle (a 24-hour day for Earth); it is defined as $P = (K c \rho)^{1/2} = c\rho\,(k)^{1/2}$. (The term k, related to conductivity K, is known as thermal diffusivity, and has units of centimeters squared per second; this parameter governs the rate of temperature change within a material; it is a measure of a substance's ability to transfer heat in and out of that portion that received solar heating during the day and cools at night). P is a measure of the heat transfer rate across a boundary between two materials. e.g., air/soil. Because materials with high P possess a strong inertial resistance to temperature fluctuations at a surface boundary, they show less temperature variation per heating/cooling cycle than those with lower thermal inertia.

In this chart, we show some characteristic values of these intrinsic thermal properties:

	Water	Sandy Soil	Basalt	Stainless Steel
K	0.0014	0.0014	0.0050	0.030
c	1.0	0.24	0.20	0.12
d	1.0	1.82	2.80	7.83
P	0.038	0.024	0.053	0.168

9-5: **Of the materials in this table, which will show the largest temperature fluctuation during a 24-hr heating/cooling cycle; which the smallest? ANSWER**

9-6: **We give here the values (without their units; see above) for Specific Heat; Thermal Conductivity; and Density (in that order) for these three rock types: Limestone: 0.17, 0.0048, 2.5; Sandstone: 0.47, 0.0125, 2.5; Shale: 0.391, 0.0042, 2.3. Calculate the Thermal Inertia P for each type. Are the three P values different enough that, assuming a gray scale from 0 to 100, they would show up as sufficiently separable gray tones (assigned to each P value) on a black and white image of a target containing these three rocks? There is also a river passing through the scene; its values are 1.00, 0.0013, 1.0; will it show up as distinctly different? ANSWER**

Interpreting thermal data and images of temperature distribution over an area is complex. In many instances, we must look for patterns of relative temperature differences rather than the absolute values, because of the many complex factors that make quantitative determinations difficult, such as:

- Number and distribution of different material classes in an instantaneous field of view
- Variations in the angle of thermal insolation relative to sensor position
- Dependency of thermal response on composition, density and texture of the materials
- Emissivities of the surface materials
- Contributions from geothermal (internal) heat flux; usually small and local
- Topographic irregularities including elevation, slope angle, and aspect (surface direction relative to the Sun's position)
- Rainfall history, soil-moisture content, and evaporative cooling effects near the surface
- Vegetation canopy characteristics, including height, leaf geometry, and plant shape
- Leaf temperatures as a function of evapotranspiration and plant stress
- Near surface (1 to 3 meters) air temperature; relative humidity; and wind effects
- Temperature history of the atmosphere above the surface zone
- Cloud-cover history (during heating/cooling cycle)
- Absorption and re-emission of thermal radiation by aerosols, water vapor, and air gases

9-7: **All of the above factors play a role but some are more influential in determining the radiant temperatures than others. List, in your opinion, the five most important of these.** **ANSWER**

Some factors have fixed or constant effects, while others vary with each sensor overpass. One can sometimes correct for the influence of some of the variable factors, but this is difficult to do routinely. Measurements made at isolated individual points in a scene and extrapolated to the general scene have limited validity.

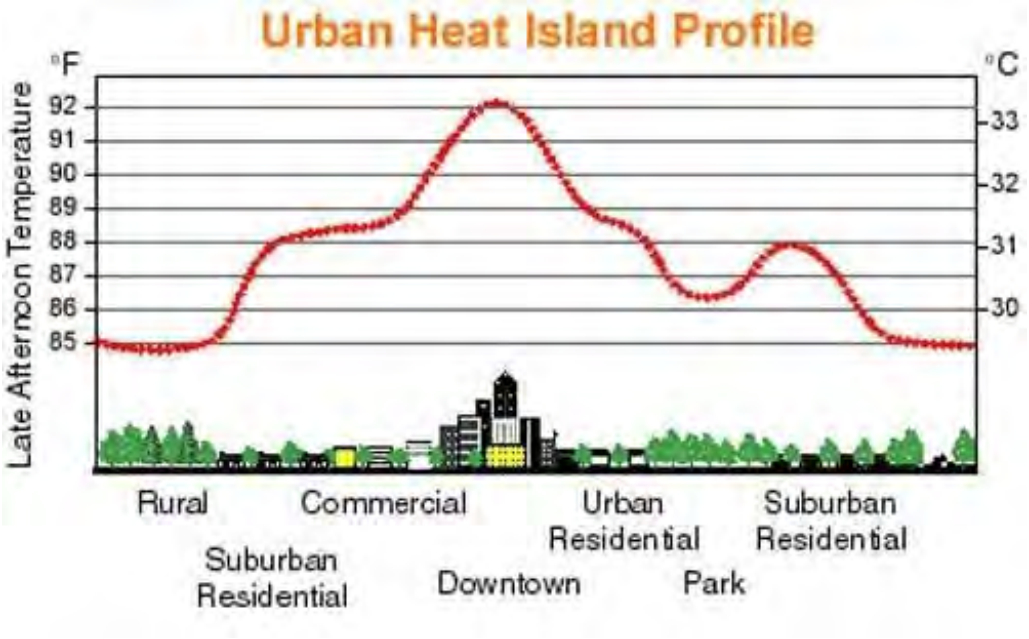
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

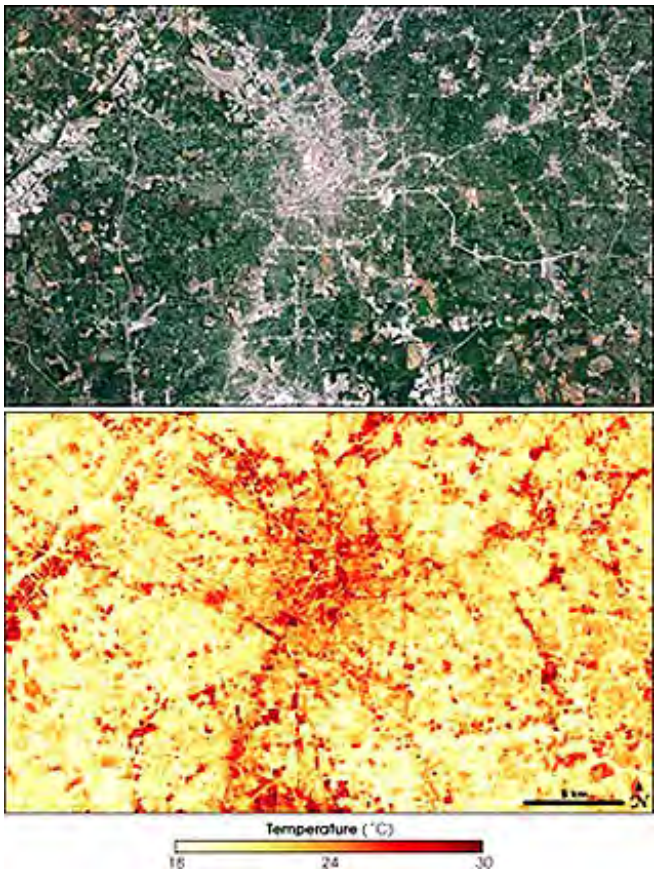
The thermal state of a material is very sensitive to the time of day (or night), since most heating is done by the Sun, aided by heat transfer from the atmosphere. The change in radiant temperature during a diurnal heating cycle varies with the materials involved. Changes in temperature during the cycle will depend on the thermal properties of each material. The heating history of a soil, as a function of depth, is particularly important since most land surfaces are covered by various types of soils.

Land Class and Diurnal Heating Effects

The reader no doubt is fully aware of the fact that the warmth (specific temperature) felt inside an urban area is higher - often notably - than experienced at the same time in the countryside. This is simply a manifestation of the so-called "urban heat island effect". An urban heat island (UHI) is a metropolitan area which is significantly warmer than its surrounding rural areas. The rise in temperature is due to a combination of factors, among which are these: Asphalt and other materials give off heat; human activities - such as heating buildings in winter and operation of motor vehicles - produce increases in temperature; buildings also modify air circulation, often inhibiting cooling; air pollution gases can trap heat, and they prevent heat from effectively radiating into the atmosphere; reduction in vegetation leads to less evapotranspiration (a cooling mechanism). This plot illustrates temperature differences due to the UHI:

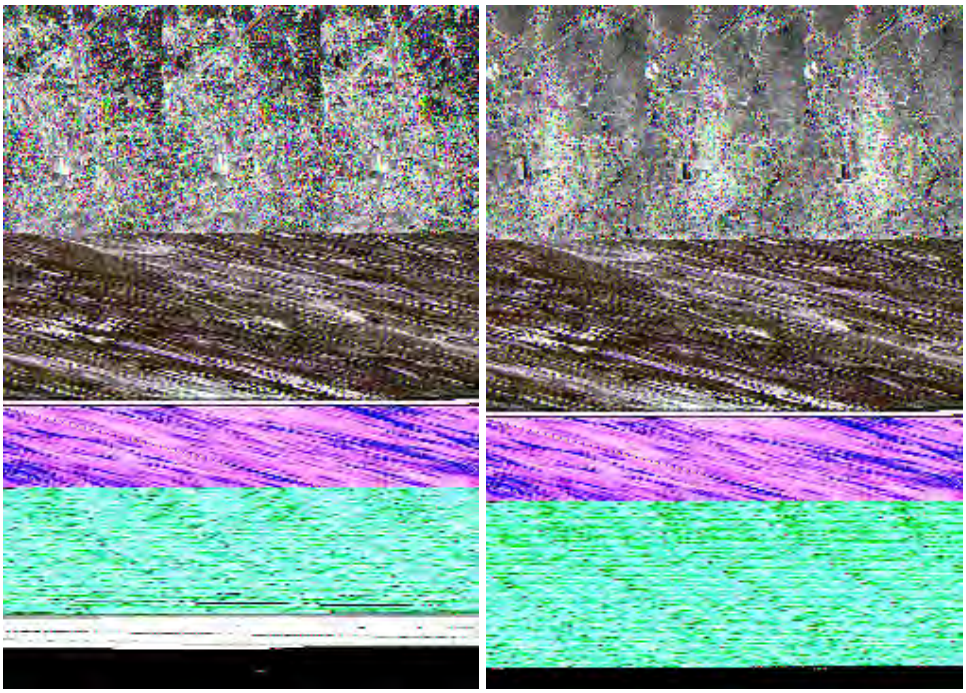


Temperatures can be several degrees (F or C) hotter than rural areas. This holds for both day and night. A pair of Landsat-7 images - the upper true color, the lower a color-coded map of temperature variations as measured by the thermal band on the ETM+ sensor - illustrates this effect for Atlanta, for an observation made on September 28, 2000.



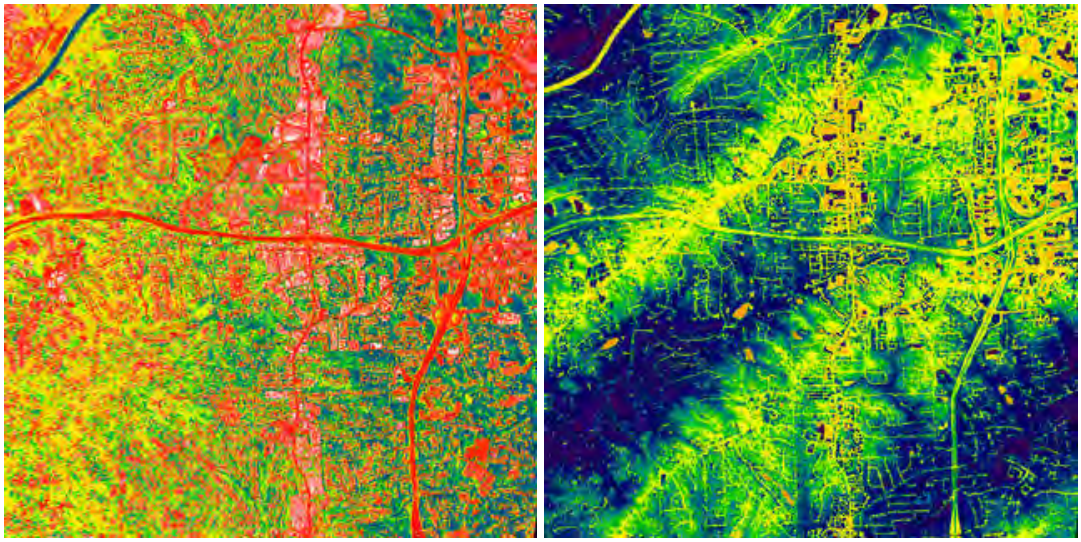
It has been stated earlier in this Section that thermal images will vary considerably in appearance depending on whether they are acquired during the warm part of the day or after a night of absence of the Sun and resultant cooling of the atmosphere as well as heat loss from the surface and shallow depths beneath.

This is evident in these images of central Atlanta, GA. taken during the day (left, or top) and then just before dawn (right, or bottom). The thermal sensor was flown on an aircraft.

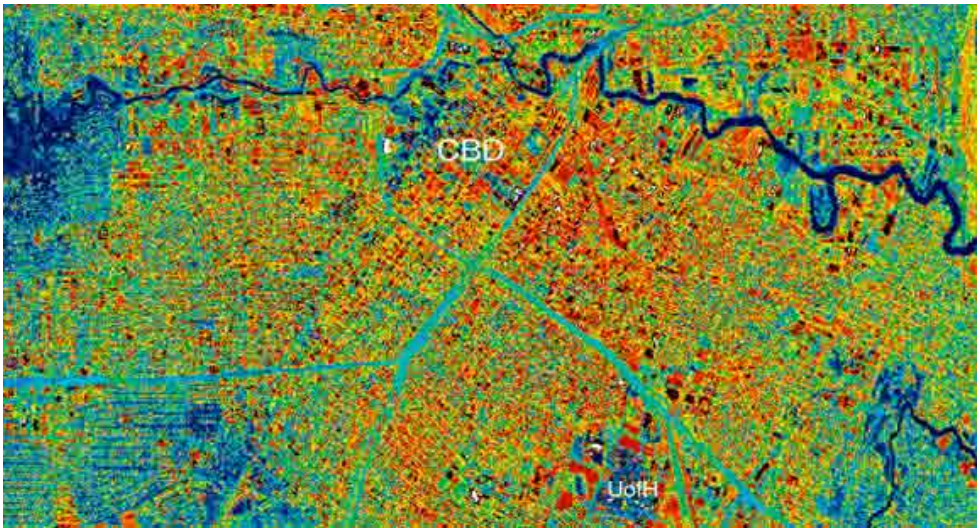


In the day thermal image, contrasts between heated buildings and streets and areas in shadow create a scene that resembles an aerial photo. But in the dawn image, differences in temperature have decreased sharply (no shadows), although a part of that image is brighter, representing a local "heat island" effect. Also, in the dawn image, many streets are evident because, being asphalt paved, they absorb more heat and remain warmer through the night.

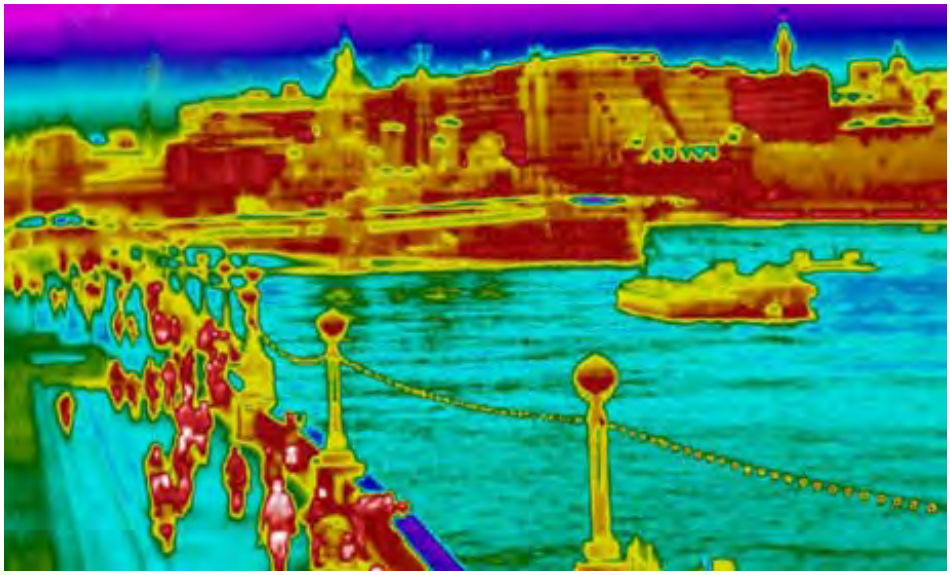
The variations within a different part of Atlanta between midday and late night are brought out better when the images are assigned colors for the temperatures (as is the convention reds are hottest and blues coolest):



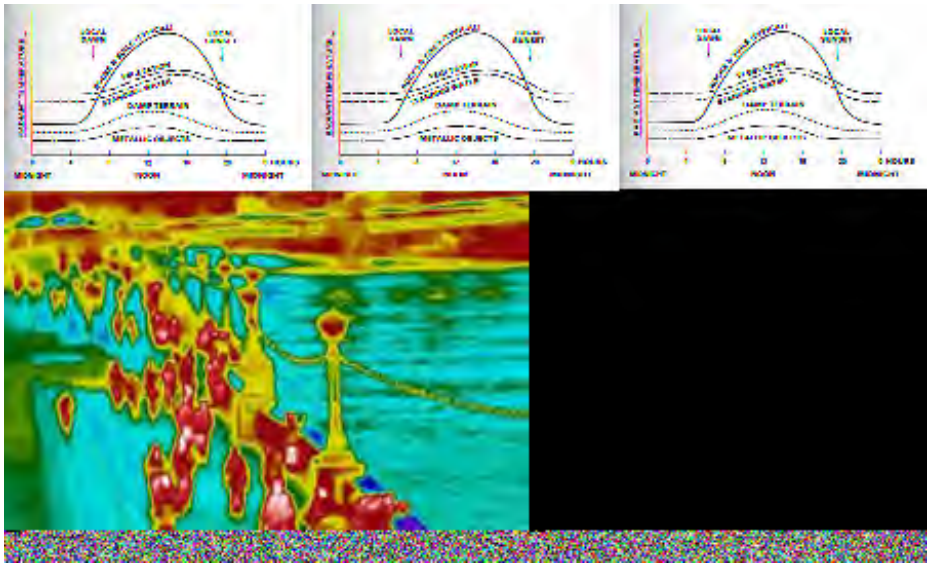
Higher resolution thermal imagery can distinguish temperature differences between buildings and surroundings, as is shown in this image of central Houston, TX:



Temperature differences can be assessed using thermal cameras on the ground, as shown in this image of part of London:



Lets look into the theory behind these temperature changes: Unlike remote sensing of reflected light from surfaces in which only the topmost layers (a few molecular layers thick) are involved, thermal remote sensing includes energy variations extending to varying shallow depths below the ground surface. This takes time and is the normal consequence of heating during the day and cooling at night. The most critical consideration in analyzing and interpreting thermal data and imagery is that of knowing the physical and temporal conditions that heat the near surface layers. Over the seasons, minor shifts in the mean temperature in bedrock can occur to depths of 10 m (33 ft) or more. Solar radiation and heat transfer from the air significantly heat materials at and immediately below the surface during the day. Temperatures usually drop at night primarily by radiative cooling (maximum radiative cooling occurs under cloudless conditions), accompanied by some conduction and convection. During a single daily (diurnal) cycle, the near surface layers (commonly, unconsolidated soils) experience alternate heating and cooling to depths typically between 50 and 100 cm (20-40 in). The daily mean surface temperature is commonly near the mean air temperature. Observed temperature changes are induced mainly by changes during the diurnal heating cycle, but seasonal differences (averages and range) in temperature and local meteorological conditions also affect the cycle response from day to day.

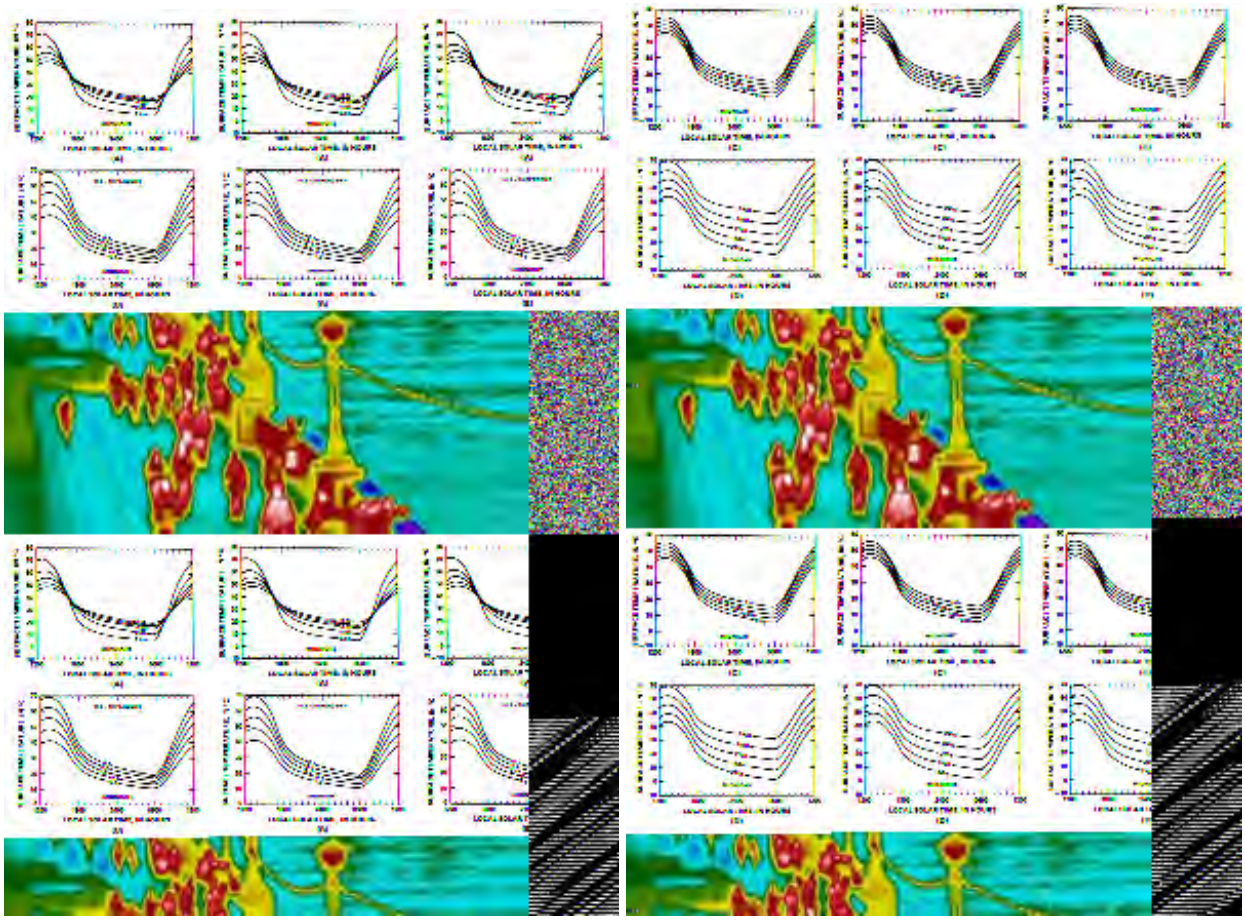


Changes in radiant temperatures of five surface-cover types during a 24-hour thermal cycle. From F.F. Sabins, Jr., Remote Sensing: Principles and Interpretation. 2nd Ed., © 1987. Reproduced by permission of W.H. Freeman & Co., New York City.

The curves shown here summarize the qualitative changes in radiant temperature during a 24-hr cycle, beginning and ending at local midnight, for five general classes of materials found at the surface. From these curves we can estimate the relative gray levels that a thermal sensor could record, as a function of the material and the time of day. Given two thermal images of the same locale, taken 12 hours apart, about noon and about midnight, we might determine the identities of co-registered pixels, based on their temperatures and thermal inertias.

9-8 Letting the gray levels in a thermal image vary with temperature, predict the relative gray tones you would expect for a desert surface (rock and soil) and standing water at 4:00 AM and 2 PM. ANSWER

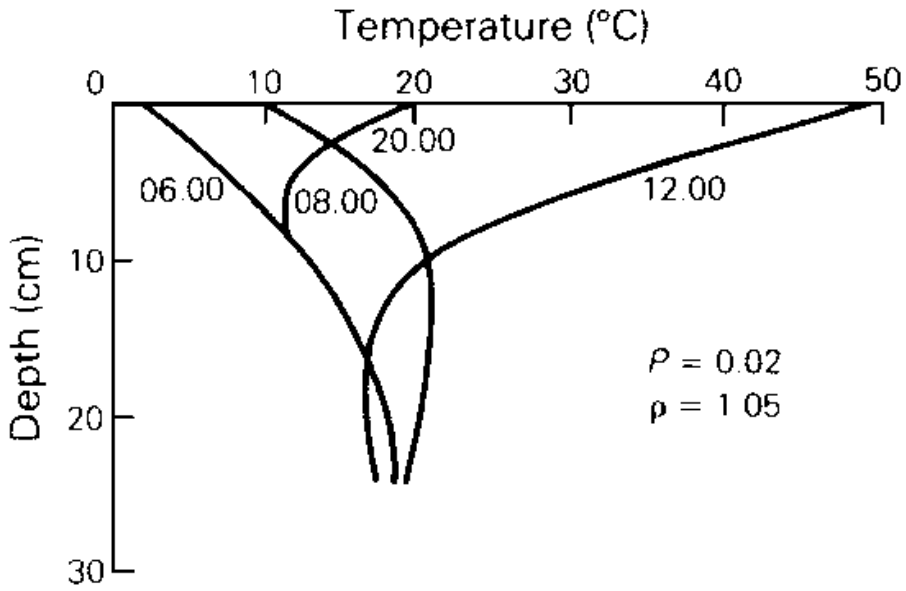
Differences in thermal inertia, reflectances, and emissivity of various materials and variable atmospheric radiance are important factors that modify the measured surface temperatures. Consider the four sets of curves plotted below:



The top curves (A) show temperature variations resulting solely from differences in thermal inertias of materials, with other factors held constant. Note the distinct crossover points. Values of P much higher than 0.05 (commonly, metallic objects) produce diurnal curves that approach a straight line passing through the crossover points. This effect is consistent with the earlier statement that materials with high thermal inertias undergo smaller radiant temperature changes during a full heating/cooling cycle. Curves in B show the effects of different reflectances (in terms of albedo - the ratio of reflected solar radiation to incident radiation, in percent. The maximum and minimum daily temperatures and their differences (δT) increase with decreasing reflectance of solar insolation. In C, the curves represent changes associated with different *emissivities*; where most natural materials have values of **ε** ranging from 0.80 to 0.98. Curves in D indicate the temperature effects introduced by the atmospheric radiances, which are caused by re-emission from water vapor, a common source of error. From these sets of curves, we learn that natural surface materials show considerable modifications in radiant temperatures because of these variables.

9-9: **Why do the above curves reach their temperature minima around 0600 hours (6 AM)? Make up a general statement or rule relating the change in surface temperature with changing values for each of the four property variables in these curves?** **ANSWER**

The interplay among these varying properties combines to affect the rate of heating, total heat exchanged, and temperature profile (gradient = dT/dz, where z is thickness or depth) within the layers in the top 30 cm (1.2 ft) or so beneath the Earth's surface. This profile can change considerably during the diurnal cycle. Different temperature vs depth profiles characterize different times of the day and night, as summarized in this diagram (for a low density soil with very low thermal inertia):



Regardless of the time of day, temperature profiles in the heating zone converge on a nearly steady value at some depth below 30-50 cm (depending on whether it is rock or soil, and its moisture content), but that convergent temperature slowly increases with depth into the Earth as the geothermal gradient takes over. Before dawn (06:00 hours), the surface zone (top 50 cm or so) shows a diminishing rate of increase in ground temperatures (dT/dz is positive overall). In the morning sunlight (08:00 hrs), the uppermost 10 cm (4 in) of a typical soil begins to heat rapidly upwards, forming a negative dT/dz at shallow depth that reverts to a positive gradient in the lower near surface. By shortly after noon (12:00 hrs), the surface layers reach maximum temperature while the lower layers experience successively less warming with increasing depth; this creates a negative gradient. Following sunset into the night (20:00 hrs), as heat leaves, the gradient remains negative so that temperatures decrease at the surface but revert to a steady warmth down to the base of the diurnal change zone.

9-10: **At what time of day does the soil show the least variation in temperature; the most? ANSWER**

The maximum and minimum temperatures at the observed surface (topmost centimeter or so) depend mostly on the extent of buildup of the heat reservoir and its storage capacity through the affected layers. For materials that have the same density and albedo but different conductivities and/or specific heats, the difference in predawn (lowest T values) and midday (highest T values) increases with increasing conductivity and decreasing specific heat. The actual magnitude of the difference decreases with depth, down to the stable (convergent) value at the base. The effect of increasing the material density (holding other variables constant in $P = (\rho Kc)^{1/2}$) is to require more thermal energy to heat the additional mass in a given volume, so that less heat transfers to lower layers. With increasing density, the total added heat (derived from Sun and air) distributes over a decreased thickness of surficial layers. So, in this instance, the maximum T reached at midday is lower, and the minimum at night is greater. Thus, the spread of daily temperature extremes (max δT's) tends to be lower relative to less dense materials. The corresponding rise in thermal inertia for denser materials simply expresses their increased resistance (sometimes referred to as thermal impedance) to temperature changes (smaller δTs as heating or cooling progresses). For materials with higher thermal conductivitys, more heat transfers to greater depths, less remains concentrated at the surface, and, again, temperature extremes diminish, as evidenced by lower daytime surface temperatures and higher nighttime temperatures compared with materials having lower Ks (e.g., soils).

In general, it is difficult to compensate for, correct, or otherwise remove effects of many of the factors mentioned above. Consequently, temperatures and derivative functions, such as

apparent thermal inertia (ATI; "apparent" is a qualifier, indicating that we don't obtain true values unless we take into account the influence of atmospheric processes and other factors) are approximate and subject to (sometimes serious) errors. Field measurements of the more critical variables help to alleviate the uncertainties. We can incorporate these, and other sources of ancillary data, into mathematical *phenomenological models* that attempt to duplicate the roles played by the physical factors.

PREVIOUS NEXT

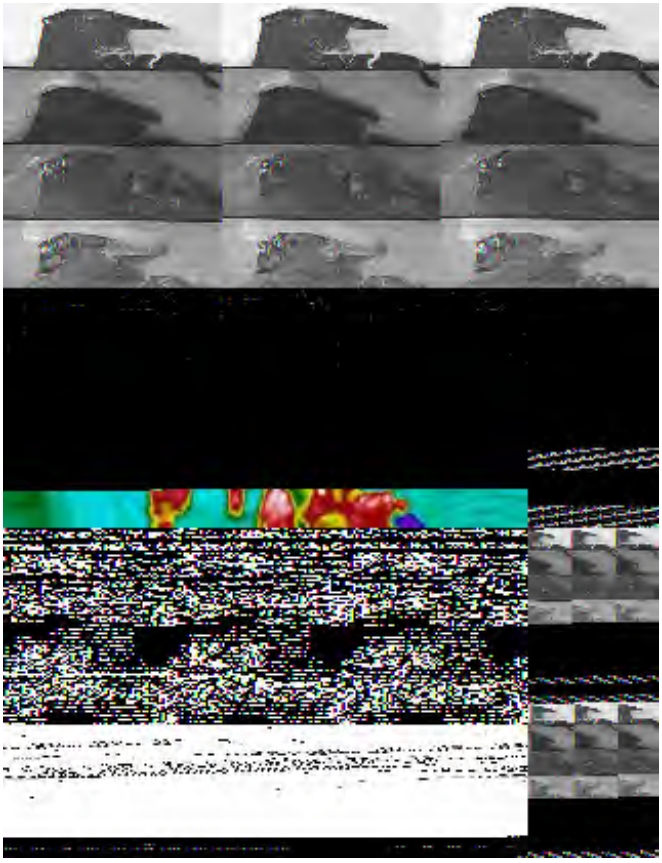
Primary Author: Nicholas M. Short, Sr.

This page treats several subjects. First, the thermal behavior of water is examined. Then, the nature of thermal sensors, in terms of detectors and their environment and operational conditions from air or space platforms, is investigated.

Thermal Properties of Water

Water has very dark to medium gray tones in day thermal-IR images and moderately light tones in night thermal images, compared with the land. This simply means it is cooler in the day and warmer in the night than most other materials in the scene.This response is due in part to a rather high thermal inertia, relative to typical land surfaces, as controlled largely by water's high specific heat. Thus, it heats less during the day and holds that heat more at night (an obvious condition swimmers experience), giving rise under many meteorological conditions to intrinsically cooler daytime temperatures and often warmer nighttime temperatures than most materials on land. Also, being nonsolid, water in natural settings (rivers, lakes, oceans) is likely to experience disruption of its thermal gradient by convection (e.g., upwelling) and turbulence (e.g., wave action) that tend towards mixing and homogenation, so that its near-surface temperatures vary by only a few degrees at most (temperature "smoothing").

As discussed on the previous page, the appearance in thermal images of both water and land depend on the time of day. Below is a sequence of 4 thermal images taken by an airborne Daedalus thermal scanner that show a land area next to the Delaware River, on the New Jersey side.



The images were processed to emphasize water differences, thus subduing the land expression

of thermal variability. The top three images were taken on December 28, 1979. The first on top was acquired during a pass at 06:00 hours when the air temperature was -9° C; the water relative to land was notably warmer (above freezing). The second image was obtained at 08:00 with the water still warmer than the land; the bright streak near the land's upper left point is hot effluent from a power plant. At 14:00 hours (3rd strip), the air temperature had climbed to -2° C and thermal contrast between land and water is near minimal; note the detail evident in the power plant, which is a hot building. The effluent from the plant is now pointing downstream as ebb tide occurs. The power plant detail persists in the bottom image, taken on December 29th at 11:00 when the air temperature was -4°: C. and the tide was near high. The scenes taken at different times and dates show the temperature variations that occur in the river water, and to a lesser extent on the land

As we saw on page 3-1, water in moist soils tends to keep them cooler than drier soils, so that black and white thermal images show darker patterns where the amount of moisture is higher.

Thermal Sensors

Some reminders about thermal sensors: For scanners designed to sense in the 8 to 14 µm interval, the detector is usually an alloy of mercury-cadmium-tellurium (HgCdTe) that acts as a photoconductor in response to incoming photons in this thermal energy range. Mercury-doped germanium is also used for this interval, although it is effective over a broader range, down to about 2 µm. Over the 3-5 µm interval, indium-antimony (InSb) is the alloy used in detectors operating in that range. Efficient operation requires onboard cooling of this detector to temperatures between 30° and 77° K, depending on the detector type. This temperature range is maintained either with cooling agents, such as liquid nitrogen or helium (in a container called a Dewar, that encloses the detector) or, for some spacecraft designs, with radiant cooling systems that take advantage of the cold vacuum of outer space. Detectors need this cooling to improve their signal-to-noise (S/N) ratio to a level at which they have a stable signal response. This signal is, of course, an electrical current related to changes in detector resistance that are proportional to the radiant energy.

To obtain a quantitative expression of radiant temperatures, the detector response must be calibrated. Calibration requires sources (e.g., thermistors) at different temperatures near the extremes expected from the ground to provide a correction function. The scanner normally has a glow tube or other device in which a wire passes a current that causes it to glow (giving off radiant energy) at some temperature. Thermal scanners commonly use two such thermistors: one glowing at a temperature near the low value anticipated from most targets, and the other glowing near the high value. These temperature/radiance relations are usually determined in advance in the laboratory, prior to the scanner becoming operational. Aircraft scanners require periodic recalibration.

In operation, the image signal goes either to a separate recording unit or through a chopper for sampling from the main beam. The radiant temperatures are not normally converted to kinetic temperatures because we don't usually know the emissivities of the diverse surface materials sufficiently well to permit this.

9-11: **What is an obvious disadvantage to the use of electric temperature devices, such as lamps, in spacecraft thermal sensors as a means of calibration? ANSWER**

Now that we've warmed up to this subject, lets look at some notably hot images, unless you decide to cool it and go on to something else!

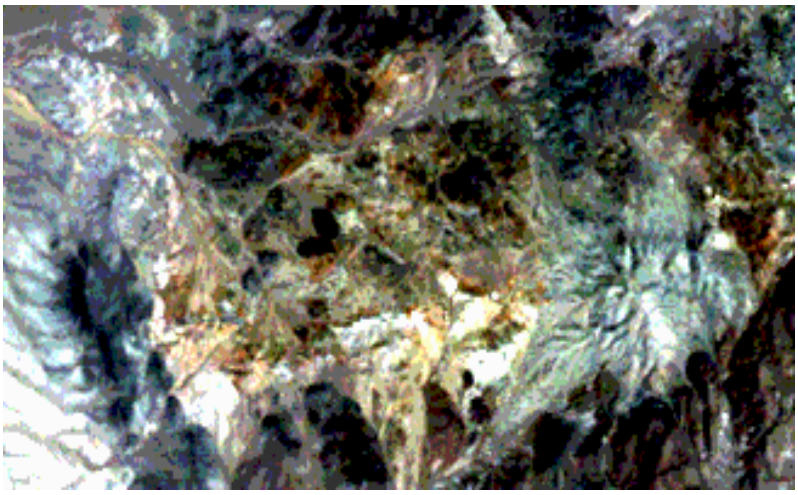
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

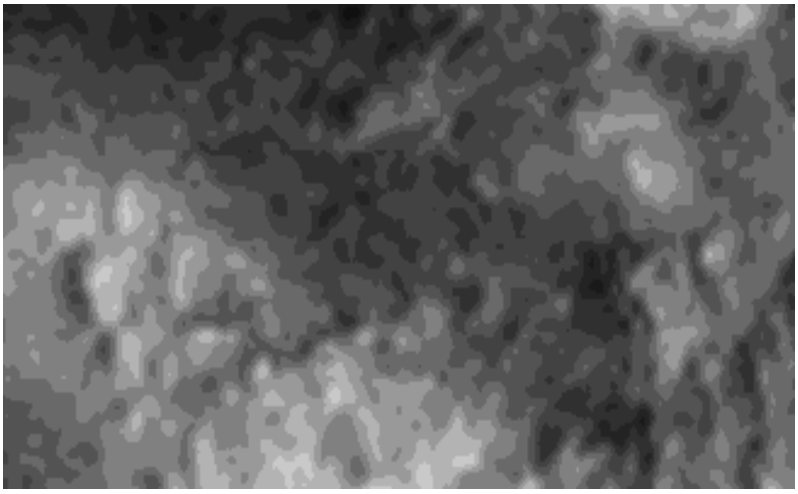
Two examples of the thermal band (TM 6) on Landsat introduce the appearance in satellite imagery of surfaces showing variations in temperature. The first is the familiar White Mountain scene studied in Section 5. Variations in temperature during a daytime pass over the site do not coincide with some of the alteration and rock type features but some correlation with dark surfaces is evident. A nighttime pass over Lake Erie exposes the expected thermal state of water as a heat retainer during land cooling.

White Mountain Thermal Features

Refresh your memory of the White Mountain scene, as recalled here in a color composite.



Three broad geological categories dominate the scene: dark volcanics, alteration products derived from these, and the limestone mountain. The Landsat thermal Band 6 image greatly simplifies the apparent scene content.

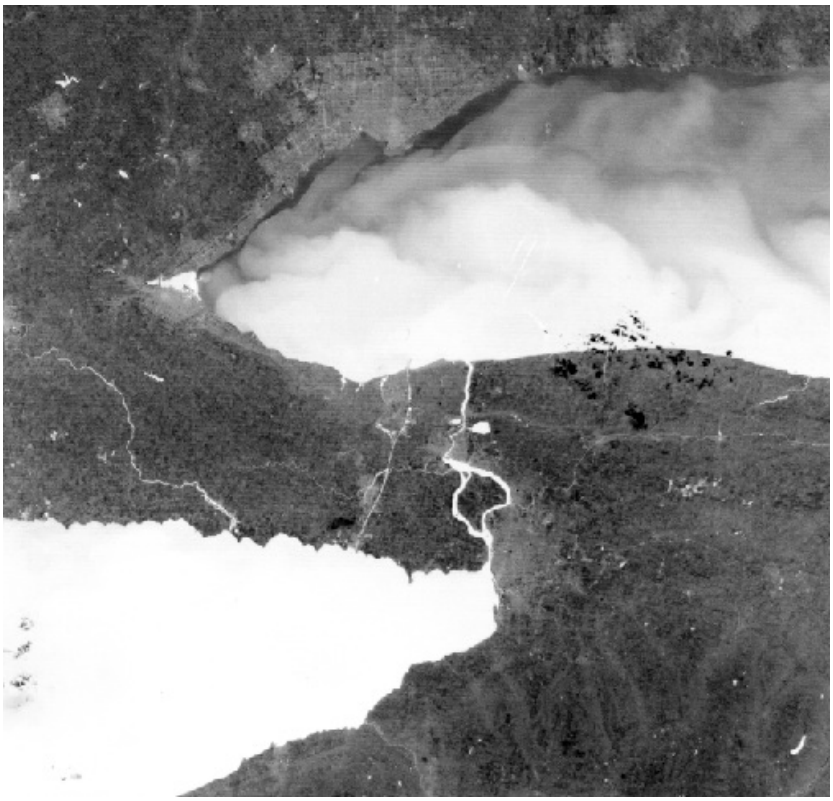


The warmest areas (shown in lighter tones in keeping with the convention that hot areas display in whites and light grays, and cold areas are in dark tones) are from the volcanics, most of which are dark basalts. These rocks are imperfect black bodies that absorb much of the solar radiation and re-emit it as strong radiators. The altered zones appear mostly in dark tones, implying that they absorb less solar energy, so that they have lower radiant temperatures. The limestones, in

the field grayish surfaces with superficial brownish-red alteration, appear as moderately light tones, which are hard to distinguish from the volcanic expressions.

Lakes Erie/Ontario TM 6

For experimental reasons, operators activate Thematic Mapper Band 6 on Landsat occasionally at night to obtain thermal images. One example, a full scene acquired at 9:32 P.M. on August 22, 1982 shows the familiar east half of Lake Erie (left lake), and the western part of Lake Ontario (top lake).



The land appears moderately cool (darker tones), with little detail, although the cities of Buffalo, NY (east tip of Lake Erie), and Toronto (top center) and Hamilton, Ontario (west end of Lake Ontario; locally hot because of steel mill effluents) may be discernible on a computer monitor from street patterns and slightly lighter (warmer) tones. A mottled pattern of variably warmer bands characterizes Lake Ontario. These bands relate to thermal overturning effects (thermoclines), which are possible in this deeper (237 m [777 ft]) lake. Lake Erie is uniformly "hot" because its shallowness (less than 67 m [220 ft]) inhibits this type of circulation. Warm rivers, such as the Niagara connecting the two lakes, stand in contrast to the land.

9-12: **Where is Niagara Falls? Any visual clues to where it should be? Is there anything in the imagery that might give you a clue as to which way the Niagara River is flowing? Is Toronto larger than Buffalo? What tone do the few clouds in the scene have? How might clouds be distinguish from smoke in this night scene? ANSWER**

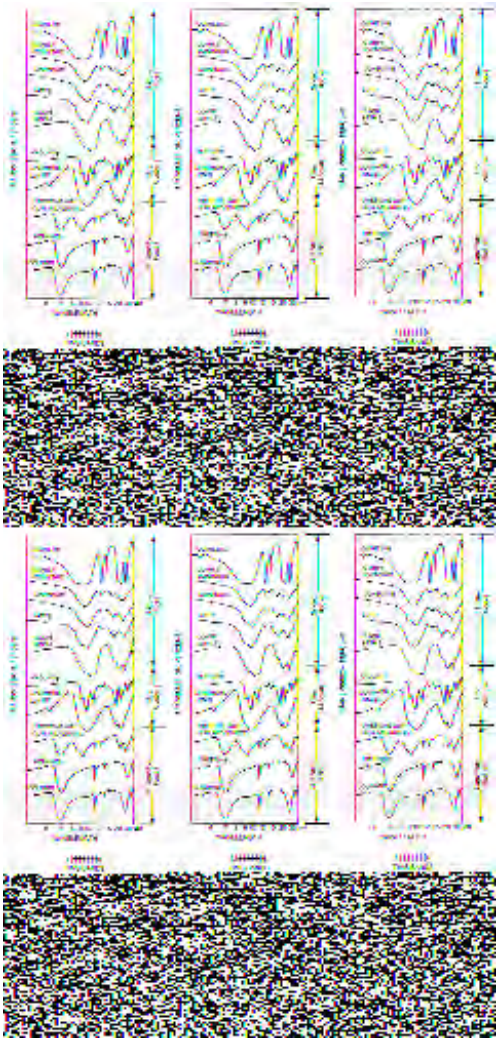
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The Thermal Infrared Multispectral Scanner (TIMS) was developed by NASA-JPL as an airborne prototype for an eventual spaceborne sensor of similar capabilities. It divides the 8-14 µm thermal interval into 6 bands. The airborne version, with its high resolution, proved capable of actually identifying some individual minerals and many rock types. Images from Death Valley, California taken both day and night clearly show a strong correlation with geologic materials at that test site. A day TIMS image of lava flows on Mauna Loa, Hawaii proved able to distinguish different flows owing to varying degrees of weathering between younger and older rock units.

TIMS Thermal Scenes: Death Valley; Mauna Loa

Now we turn to a powerful thermal sensor: an airborne instrument known as TIMS, for Thermal IR Multispectral Scanner, operated for NASA by JPL. Ground equivalent resolution is 18 m (59 ft). As its name indicates, TIMS breaks the 8-12 µm interval into six bands: (1) 8.2-8.6, (2) 8.6-9.0, (3) 9.0- 9.4, (4) 9.4-10.2, (5) 10.2-11.2, and (6) 11.2-12.2 µm wide. From these bands, we can produce three-band color composites by various combinations of bands. Although broad, these gather varying amounts of spectral thermal energy that are related to more discrete (narrower) absorption features associated with minerals that occur in rocks and soils.



From Kahle (1984), Figure 4 in F.F. Sabins, Jr., Remote Sensing: Principles and Interpretation. 2nd Ed., © 1987. Reproduced by permission of W.H. Freeman & Co., New York City.

9-13: **What do the silicates and clay minerals share in common? What distinguishes a quartzite (a cemented sandstone dominated by quartz [SiO_2]) from a monzonite (a granitic rock with quartz and feldspars, and perhaps some mica)? How do the carbonates differ from the other rocks in the above plot? ANSWER**

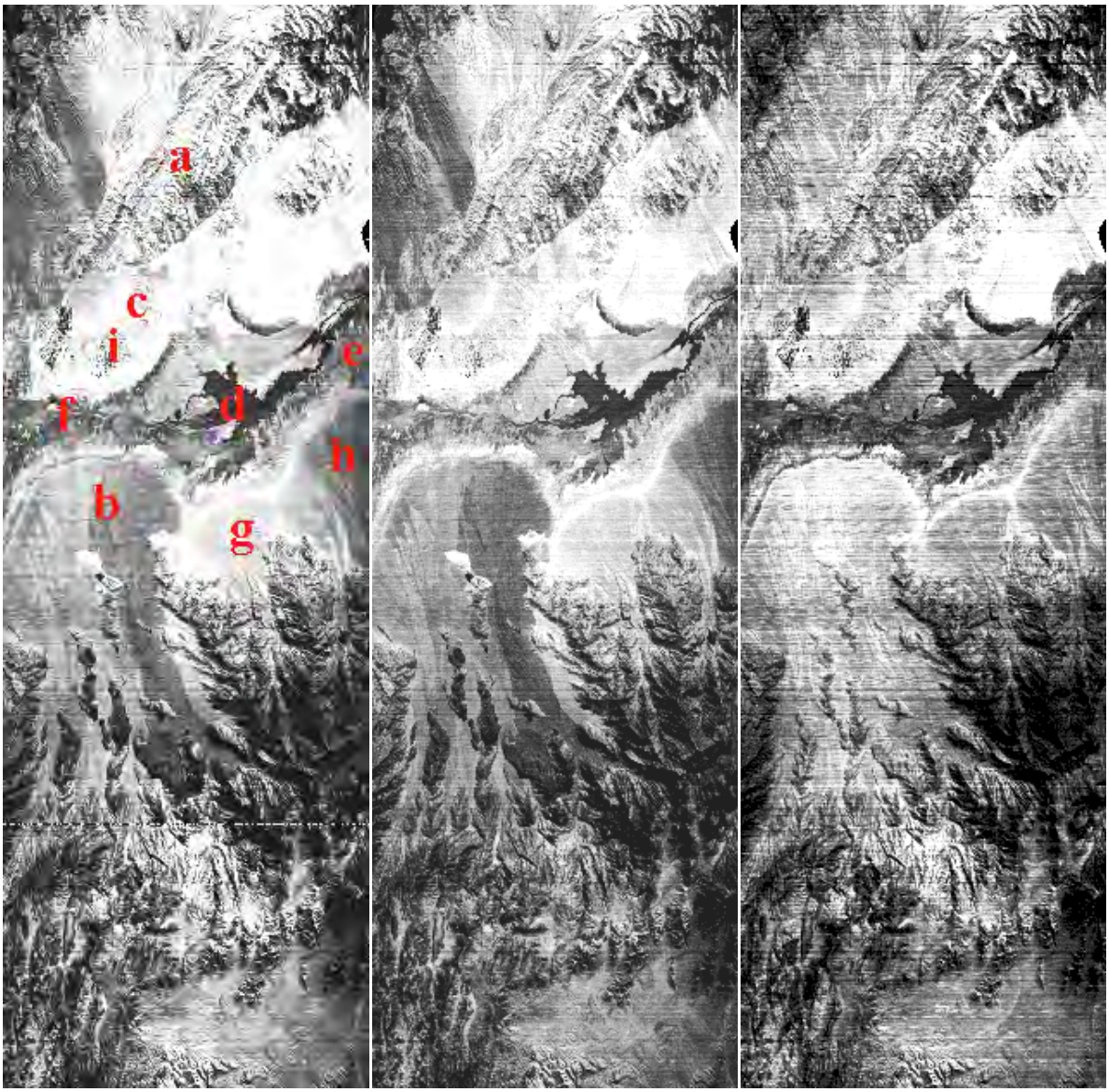
TIMS has flown over many test sites, of which one of the most informative extends across part of Death Valley in southeastern California. Here the Great Basin meets the Mojave Desert to the south. The Earth's crust breaks along tensional faults, allowing segments to downdrop into intervening basins bounded on either side by mountain ranges. As these ranges wear down, debris flows into the basins, partially filling them. Death Valley occupies such a structural basin, within which is the lowest point in the Western Hemisphere (86 m or 282 ft below sealevel). Once largely covered by lakes, the valley is now one of the driest and warmest spots on Earth, with summer temperatures often exceeding 120 °F.

Next, we show part of a Landsat Thematic Mapper false color composite of the northern end of Death Valley.

The high mountain block on the left is the Panamint Range, made up of complexly folded and faulted Paleozoic sedimentary rocks. A small segment of the Funeral Range appears at the upper right. The valley contains playas (deposits from ephemeral [usually dry] lakes) and salt pans. Saline deposits, formed from evaporation of intermittent stream waters, include sodium borates (source of the borax hauled out by the famed 20-mule teams) and sulphates, gypsum, and rock salt, mixed with mudstones, carbonates, and floodplain sediments. Water today is at a premium, but note the red patch near right center (a golf course!) at Furnace Creek, headquarters for the Death Valley National Monument, located at the edge of a large alluvial fan with active, vegetated distributaries (shown here in an aerial photo).

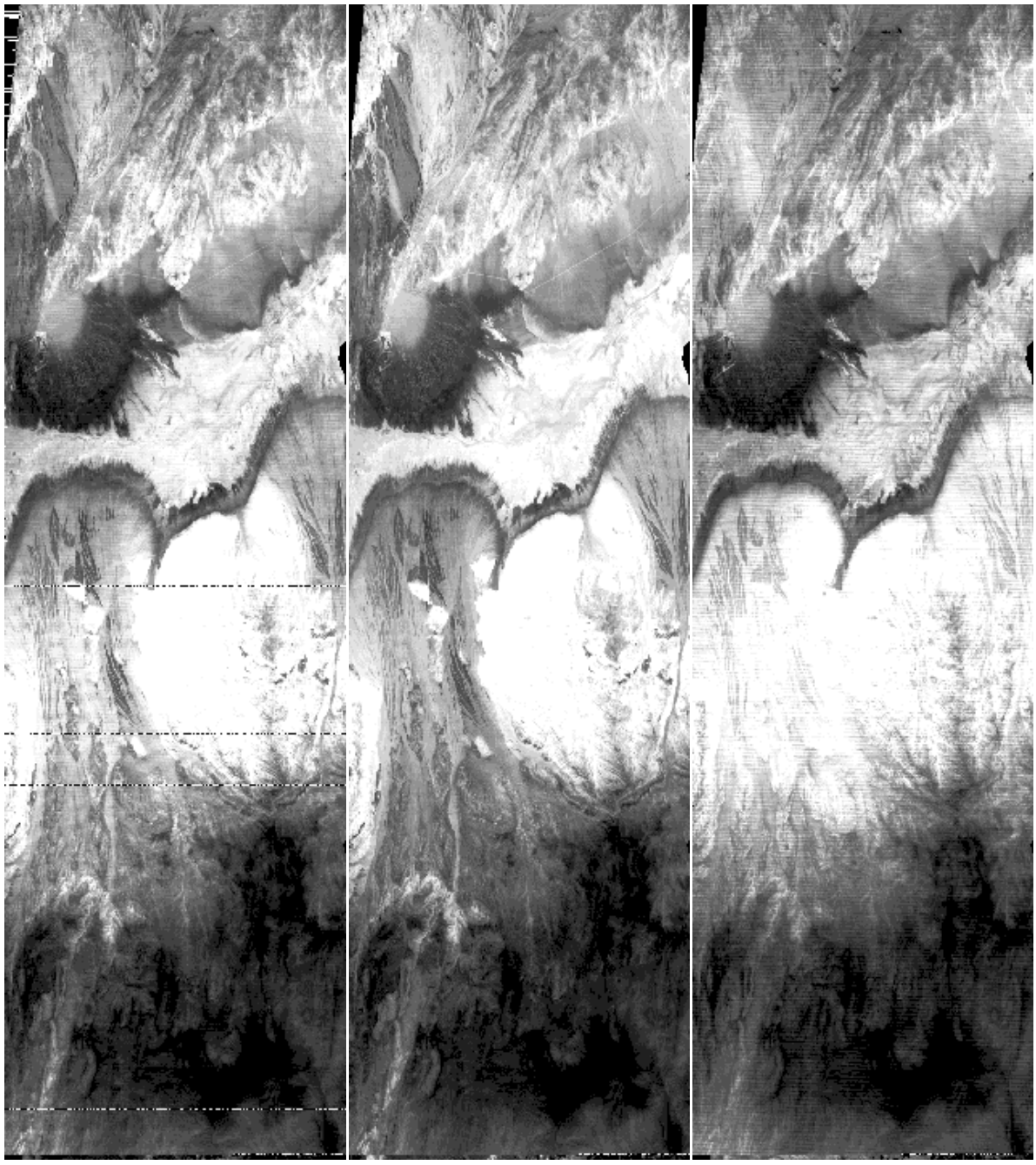
The effect of rainfall runoff over the millenia is especially evident in the numerous other alluvial fans spread over the valley pediment surfaces. These fans formed when mountain streams dropped their loads, because their gradients were abruptly lowered upon entering the valleys, causing deposition from the resulting velocity checks.

A TIMS daytime overflight along an E-W line on August 17, 1982 highlighted several of these arid landforms. We examine separate images made from Bands 1, 3, and 6 and a color composite of these. The TIMS aircraft executed a nighttime flight over the same areas shortly thereafter. We display the three daytime images here side by side: The red letters refer to specific features that are listed in the table below.



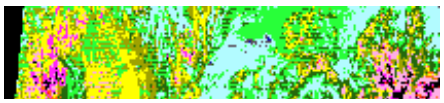
Band 1 Band 3 Band 6

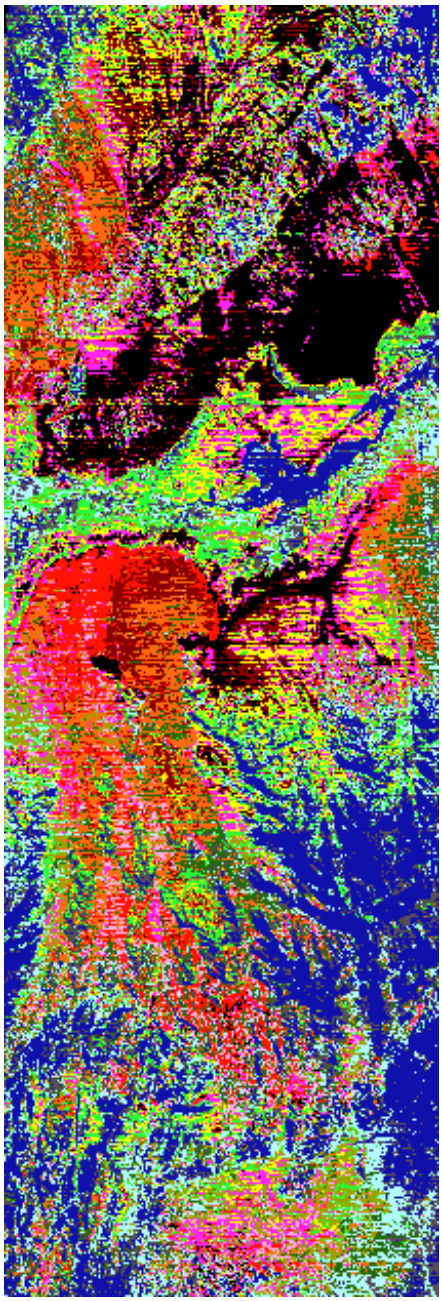
We oriented the above images with north on the left side of each. The area covered is approximately that of the lower right quadrant of the Landsat subscene shown above. Visible in each is an outlier of the Panamint Range (bottom right) and a small segment of the Black Range (top right), two prominent alluvial fans (lower), the conspicuous dark and light fan on which Furnace Creek resides, darker alluvial fill in the basin, and the sharply contrasting light tones of salt deposits. At first glance, the three bands seem similar to one another in their general tonal distribution. But, when comparing equivalent points, subtle to sometimes notable changes in tone are clear. Compare these tonal patterns with those from the night thermal IR overflight, seen here:

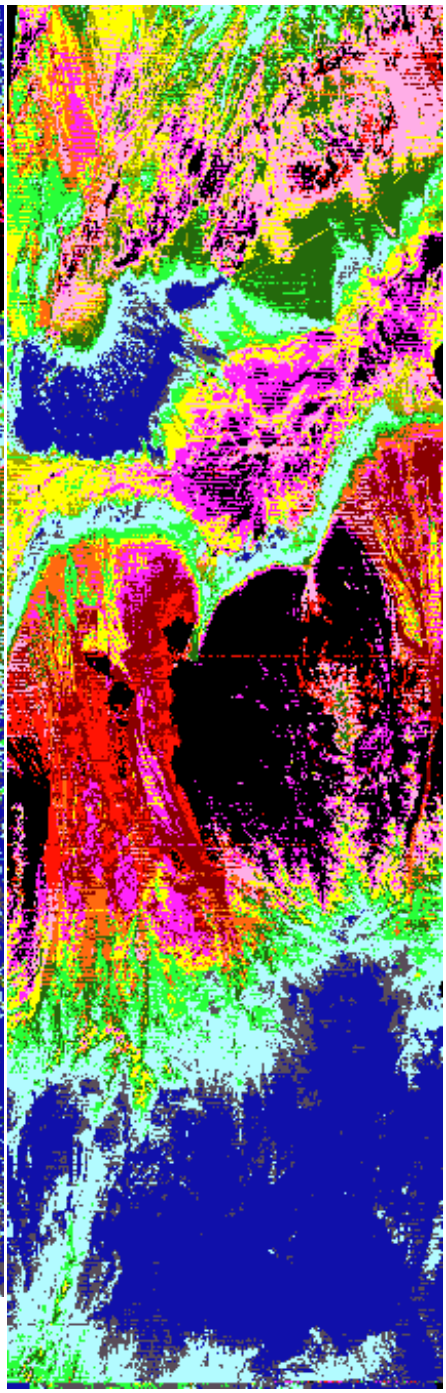


Band 1 Band 3 Band 6

Next, examine the day and night color composites made from these bands:







Day Composite Night Composite

We can analyze these images visually (qualitatively) in terms of tonal variations and colors. The following table reports this (letters locate features in the Day Band 1 image):

Day	1	3	6	CC	Night	1	3	6	CC
a. Mountains	m-d	m- d	m-d	mixed		l-m	l- m	l-m	mh-h
b. West Fans	m	m&d	m-l	h		m	m	l	h
c. Furn Crk Fan	l	l- m	m	vh		d	d	d	c
d. Dark Fill	m	m	m- d	mh		l	l	l- m	mh
e. Salt 1	d	d	d	c		-	vague	-	-
f. Salt 2	l	m	l- m	w		-	-	-	-
g. Interfan	l	l	l- m	h		l	l	m	h
h. Fan Rim	m-d	m- l	d	w		l	m-d	d	mc
Code:	A. Day -	l = light (tone); m = medium; d = dark							

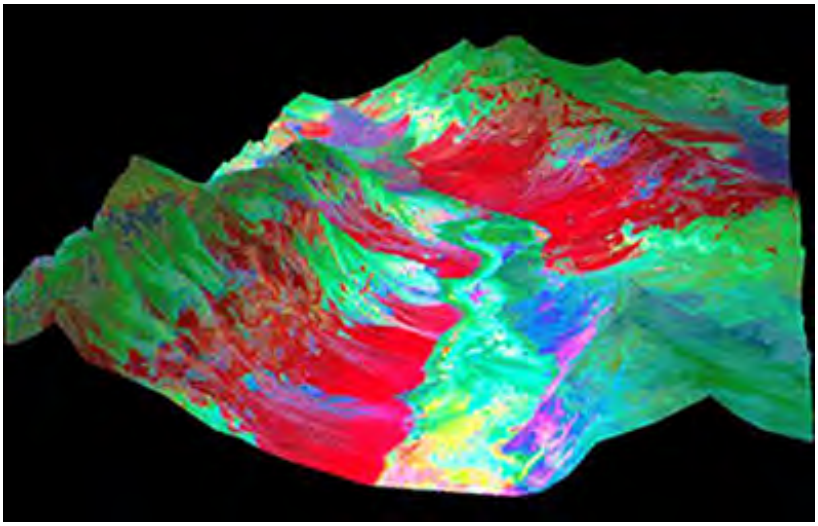
	B. Night	c = cold (blue); m = moderate (green);
		w = warm (yellow); h = hot (pink to red);
		v = very hot (white)

A few comments may help to interpret these thermal states. As one would surmise, the mountains (**a**) show variegated tones and colors because of the variations in slope, aspect, elevation, rock types, and other factors. The large fans (**b**) stand out in moderate to light tones and are hot in both the day and the night scenes. The fan at Furnace Creek (**c**) appears in lighter day tones and dark in all night bands. Thus, it is quite hot in the day composite and very cool in the night composite, which we expect from loose, unconsolidated materials that readily gain, then lose heat. The darker valley fill (**d**) absorbs heat in the day and retains it at night. The peculiar-looking salt deposit at e, easily separated in the visible Landsat image, has a dark daytime pattern and a cool color in the composite. The saline area at the end of the Furnace Creek fan is less well defined and appears warm in the composite. Neither saline area (**e** and **f**) show clearly in the night band and composite images. A darker fan on the west side, referred to as interfan (**g**), is as warm or slightly hotter than the fans on either side. Its darkness (black body effect) infers a higher radiant temperature. The deposits just beyond the western fans (**h**) make up a band in the Landsat image that displays moderately well in the day thermal images, being in darker tones. They are even more noticeable in the night images, as dark tones.

The above TIMS scenes are part of the data set acquired by JPL for their Death Valley Test Site. From those data, they also prepared a larger-area TIMS scene that is much publicized. We reproduce that scene below, noting that they made it by assigning TIMS 1 = blue, 3 = green, and 5 = red. Again, observe that alluvial fans appear in reds, lavender, and blue-greens; saline soils in yellow; and other saline deposits in blues and greens. The golf course also appears in blue.

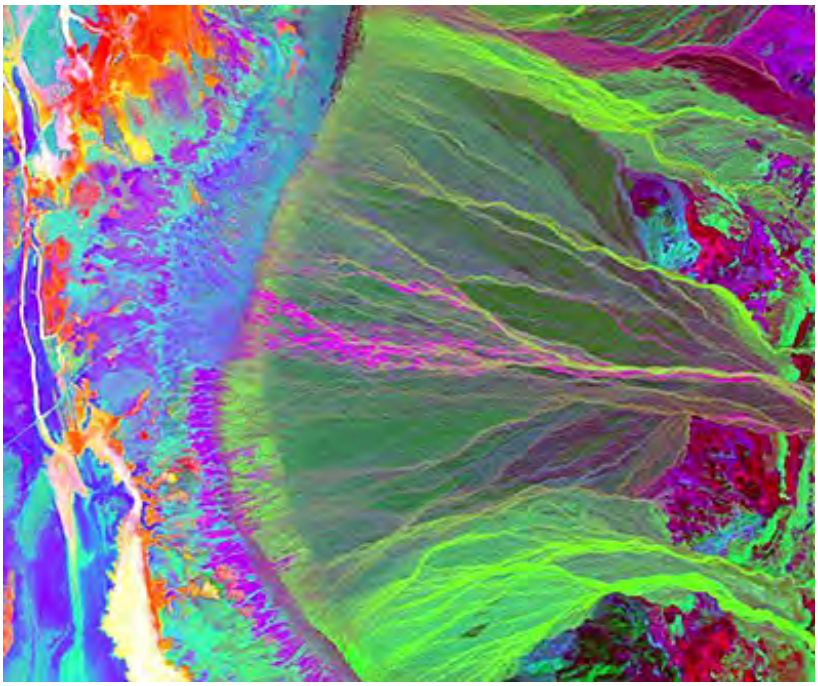
9-14: **Locate the Furnace Creek fan (c) in each image, day and night, and also the fan labelled "b". Describe the differences between the two fans, and try to account for these. ANSWER**

Verification of the ability to discriminate and identify materials by the multiband thermal IR approach demonstrated by TIMS, using a space-based system of similar capabilities, is evident in this next image. This is a display of the Death Valley region imaged at night, made by draping into a DEM-developed perspective view these three ASTER (Terra; page 16-10) thermal bands: Blue = Band 10 (8.3 µm); Green = Band 12 (9.1 µm); Red = Band 13 (10.6 µm).



The salt deposits in the central valley - mostly halides, with some nitrates, sulphates, and borates, are rendered in yellow, purple, pink, and bluish-green. The high mountains, such as the Panamints, composed of carbonates and shales, with some sandstones and metamorphics, are colored as greens. The prominent reds along uplands flanks are the stream deposits ending in alluvial fans; this unconsolidated debris is enriched in silica (such as quartz sand grains) produced through weathering.

While it is a departure from the thermal theme of this Section, we append here a colorful and informative image of a small part of Death Valley made by performing a Principal Components Analysis of AVIRIS hyperspectral data (see Sections 5 and 13) collected at D.V. A large number of individual narrow bands were utilized in making the PCA image constructed from the first three Principal Components;



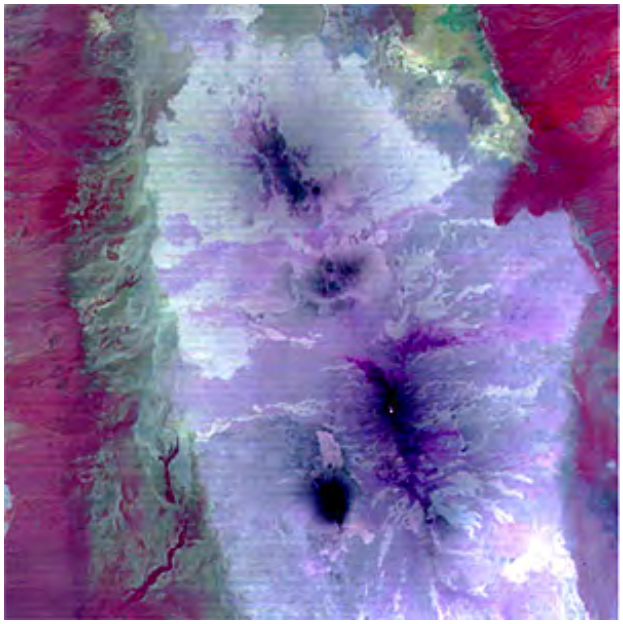
Another TIMS image from an island more than five thousand kilometers (3000 miles) to the west superbly reveals the power of thermal remote sensing. The scene below extends across a series of solidified basaltic lava flows coming from the great shield volcano Mauna Loa on the big island of Hawaii.

Most of the distinguishable flows have emerged from the volcano at various times in the last 200 years. The thermal differences sensed, as displayed by the selected colors, are due to several factors: variations in the glassiness of the original flow; variations in surface textures and porosity; and variations in degrees of weathering (mainly as a function of age), which are expressed in part by subtle differences in the rock color imposed by surface alteration. In visible light, all these flows tend to be dark in reflectance because of their inherent blackness, but their radiant temperatures vary because of the above factors.

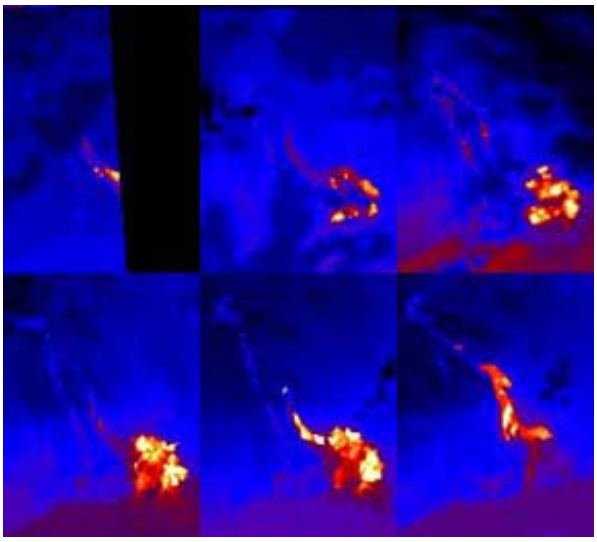
9-15: **In trying to determine the relative age of each flow with respect to the others, what other factor will help in determining the sequence? ANSWER**

As we noted in Section 8 (Radar), surface roughness plays a role also in the reflection of radar pulses, such that degrees of roughness - and thus differences in volcano flow types - give rise to distinctively different signatures. SIR-C imaged much of Mauna Loa in October of 1994, showing the many different flows, some of which are just those in the TIMS image above. In the image below, the radar color composite is made from L-band HH = red; L-band HV = green; and C-band HV = blue. According to the published interpretation by JPL, pahoehoe (ropy) flows appear as red and smoother flows as yellow to white (some other color(s) may correspond to aa [chunky] flows). In any event, radar can distinguish flow types but for different reasons than thermal sensing. Compare this image with that shown on page 17-3.

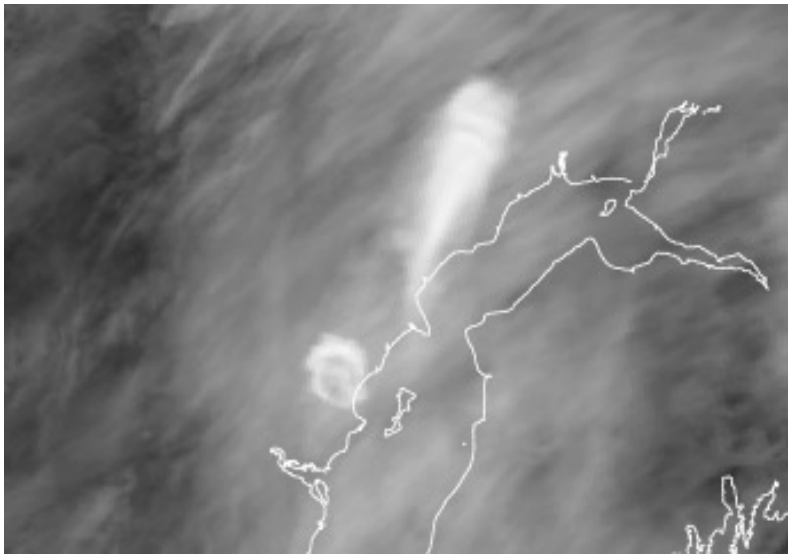
TIMS was developed and proof-of-concept tested with the intention of flying something similar to it in space, as, for example, on the Shuttle (which has yet to happen). However, a multichannel thermal sensor is part of the ASTER sensor system on the ESSE program's Terra (see page 16-10). Here is an image of the Afar area in Ethiopia, put into color using these three ASTER bands: Band 10 = Blue; Band 12 = Green; Band 14 = Red. The scene is 60 by 60 km (~ 40 x 40 miles) wide; resolution is 90 meters. Reds indicate rocks high in silica; blues and purples denote lower silica content (basalts).



Terra's ASTER is effective at pinpointing active, hot lava flows as we saw earlier at Mount Etna. Here is a series of Band 14 ASTER images of a continuing flow from the Kilauea volcanic rift on the Island of Hawaii, obtained between May 13, 2000 and January 1, 2001. The date of each image is listed in the caption.



The thermal band on Terra's MODIS is also capable of monitoring volcano effects. Mt. Redoubt in Alaska erupted in late March of 2009. Its ash plumes posed a hazard to aircraft. MODIS was able to track the ash as it moved across the state, as shown here:



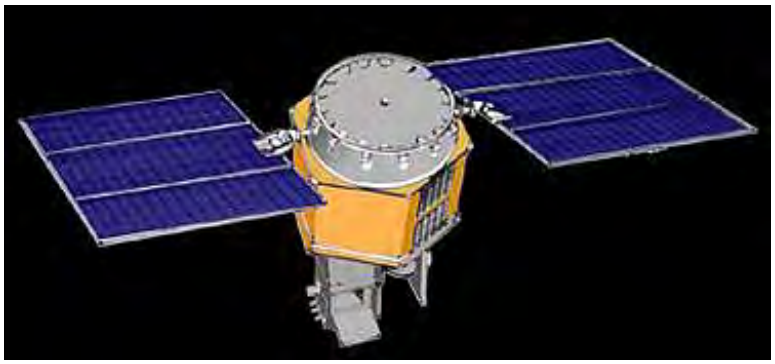
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The Heat Capacity Mapping Mission (HCMM) was an experimental satellite program dedicated to determining how useful temperature measurements could be in identifying materials and thermal conditions on the land and sea. Although day and night thermal images were valuable visual guides, the determination of one of the physical properties of materials, thermal inertia, was the ultimate goal. The scenes acquired covered large areas at higher resolution than meteorological satellites. Many proposed applications were proven valid. The principles of interpretation can also be applied to thermal data gathered routinely by meteorological satellites; examples of three thermal band images of clouds, and the corresponding visible image, taken in June 2000 by the GOES-11 geostationary satellite illustrate the features evident at different wavelengths.

The Heat Capacity Mapping Mission (HCMM); Weather Satellites

On April 26, 1978, NASA launched a sensor system capable of measuring temperatures and albedo, from which the apparent thermal inertia (ATI) of parcels of the Earth's land and sea surfaces can be estimated. This was the Heat Capacity Mapping Mission (HCMM) satellite, shown here:.



HCMM used a two channel radiometer: one channel detected visible to near- IR radiation, 0.5 to 1.1 µm, and the second detected emitted thermal IR, 10.5-12.5 µm interval. Following a near-polar, Sun-synchronous, retrograde orbit, at a nominal altitude of 620 km (385 mi), the satellite passed south to north over acquisition targets in the early afternoon (about 2:00 P.M. at the equator) along pathways inclined at 7.86° to longitudinal lines. Night passes were also inclined at 7.86° , but in an opposing orientation relative to longitudinal lines, so the day and night paths were symmetrical. A night pass, descending north to south, covered its target between 1:30 and 2:30 A.M. local time. It imaged the same area about 12 hours later (one-half diurnal cycle in eight orbits) at latitudes from 0° to 20° and 35° to 78° . It imaged again 36 hours apart (one and a half cycles) between 20° and 35° latitude (of course, cloud cover compromised some images). The repeat cycle, covering approximately the same area in the day-night pattern, was 16 days. The altitude and sensor optics produced a swath width of 715 km (445 mi) and a spatial resolution of 500 m (1640 ft) for the visible channel and 600 m (1968 ft) for the thermal channel, whose sensitivity (ability to discriminate temperature differences) was about 0.4° K at 280°K.

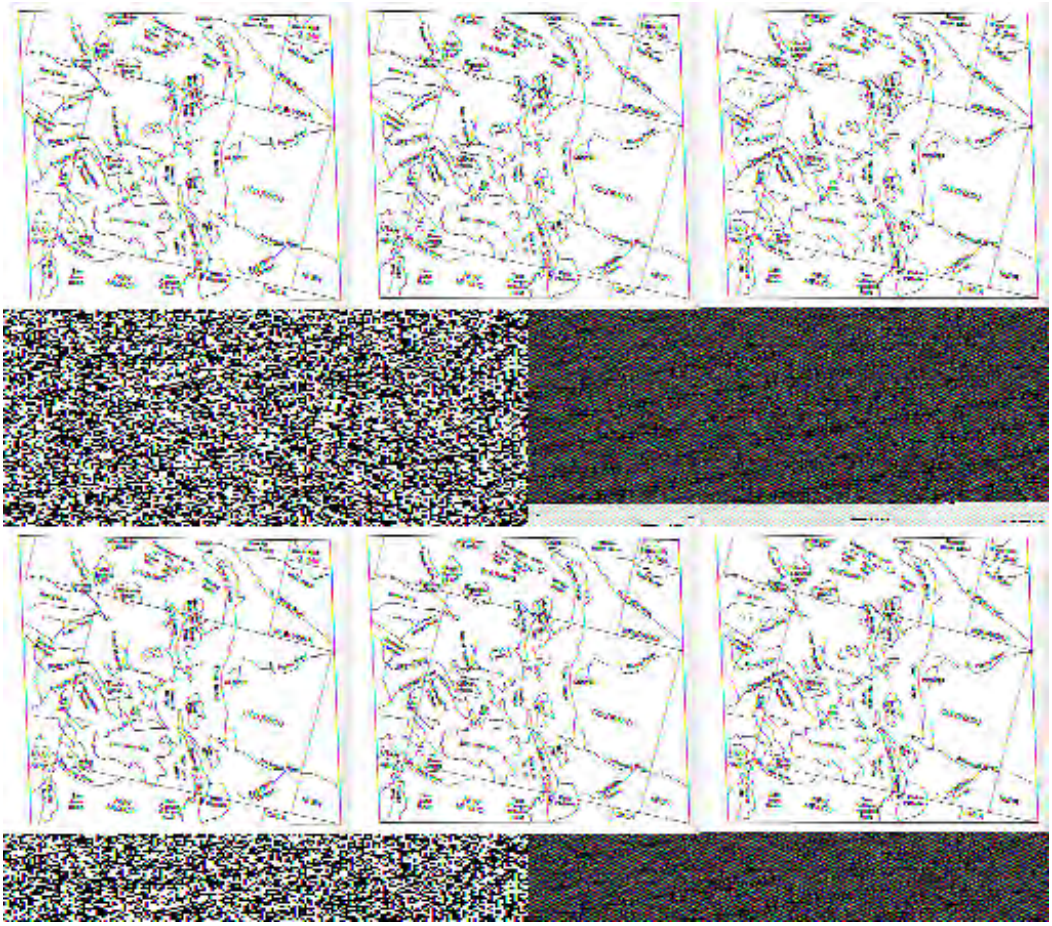
Because the day and night passes over the same areas along paths were acquired at opposing inclinations, and individual pixels scanned by the thermal sensor therefore do not precisely cover the same ground areas (Instanteous Fields Of View), i.e., do not coincide, it is necessary to co-register equivalent ground plots using a computer program that uses specific control points

(features recognizable at the 600 m resolution) to tie the two scenes together. Once the pixel sets are registered, then we can calculate the temperature differences (δT) for a particular cycle from the values obtained at the mid-afternoon and middle of night times (note that δT will not be maximum since the coolest time is several hours later, around dawn). We can derive the apparent albedo 'a' (ranging from 0 to 1 [the maximum is equivalent to 100% reflection]) from the Day Visible channel values. These values are the sensor-obtained values needed to calculate ATI according to the formula: ATI = NC(1-a)/δT, where N is a scaling factor (set at 1000) and C is a constant related to the solar flux (irradiance), generalized for a given latitude. There are no discrete units for ATI; the values are relative. Images made to display variations in ATI for a scene show high values (small δ Ts and/or low albedos) as light tones and low as dark tones. Day and Night temperature images follow the usual convention of light as warm and dark as cool. We must interpret these images cautiously, because many of the variables mentioned earlier in this Section can affect the temperature state at any or all point(s) and usually we can't control them or measure them. ATI, then, is only an approximation to actual thermal inertia, since its calculation does not take into account atmospheric properties and other factors difficult to ascertain from space.

9-16: **Calculate the ATI for a material that attains a temperature of 30° C in the day and 15 °C at night. The apparent albedo is 0.3. For this case, at a latitude of 40° N and a date of June 21st (summer solstice), the constant C = 1.605. (Note: in some instances, the formula is further simplified by setting N = 1; this gives ATI values in the same general range as we observed earlier for P).** **ANSWER**

HCMM was the first in a series of lower cost experimental satellites, the Applications Explorer Missions, flown into the 1980s. There were 54 funded investigations designed to determine the usefulness of thermal data in Geology, Agriculture, and Land Use. Unfortunately, at the conclusion of these investigations, with final reports submitted, funds to hold a conference of investigators were depleted. This compromised the intention to publicize the value of this mode of thermal remote sensing. Although the writer (NMS) had almost no experience with thermal remote sensing, he was asked to prepare a folio summarizing both the HCMM program and the investigator results. (For this, I learned fast!) The outcome was NASA Special Publication SP-465 (264 pp), entitled "The HCMM Anthology", which contained a gallery of HCMM images and an integrated summary of the investigations. The information on the remainder of this page is extracted from that document.

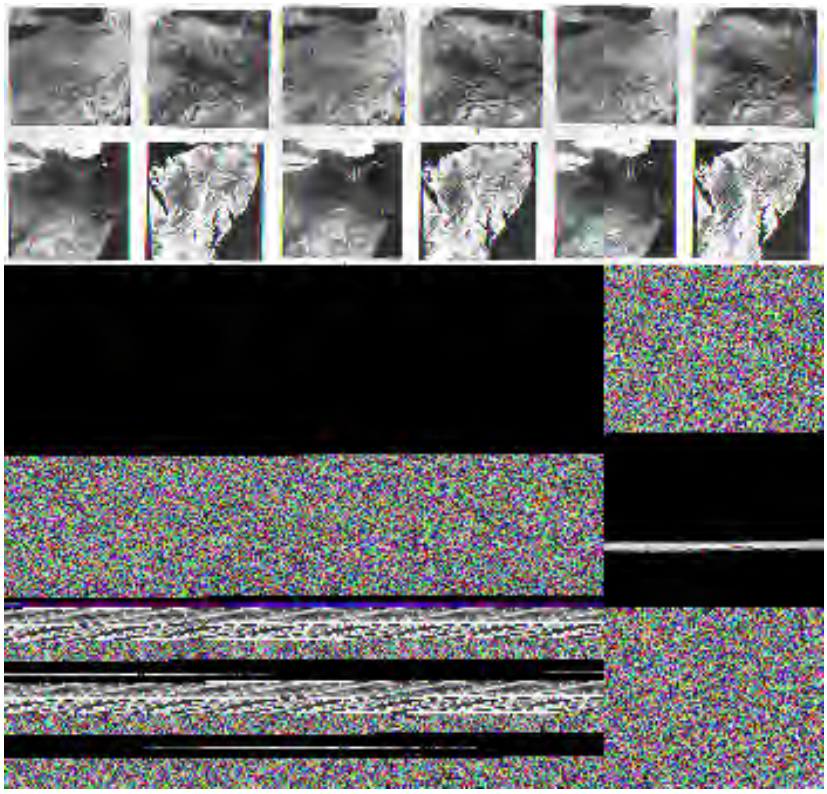
We begin our look at HCMM images with this instructive full Day Visible scene that extends for more than 640 km (398 mi) across and in so doing, completely encompasses the state of Colorado as well as parts of surrounding states, from the Great Plains in Kansas, west into Utah and north into Wyoming, as located in the accompanying map.



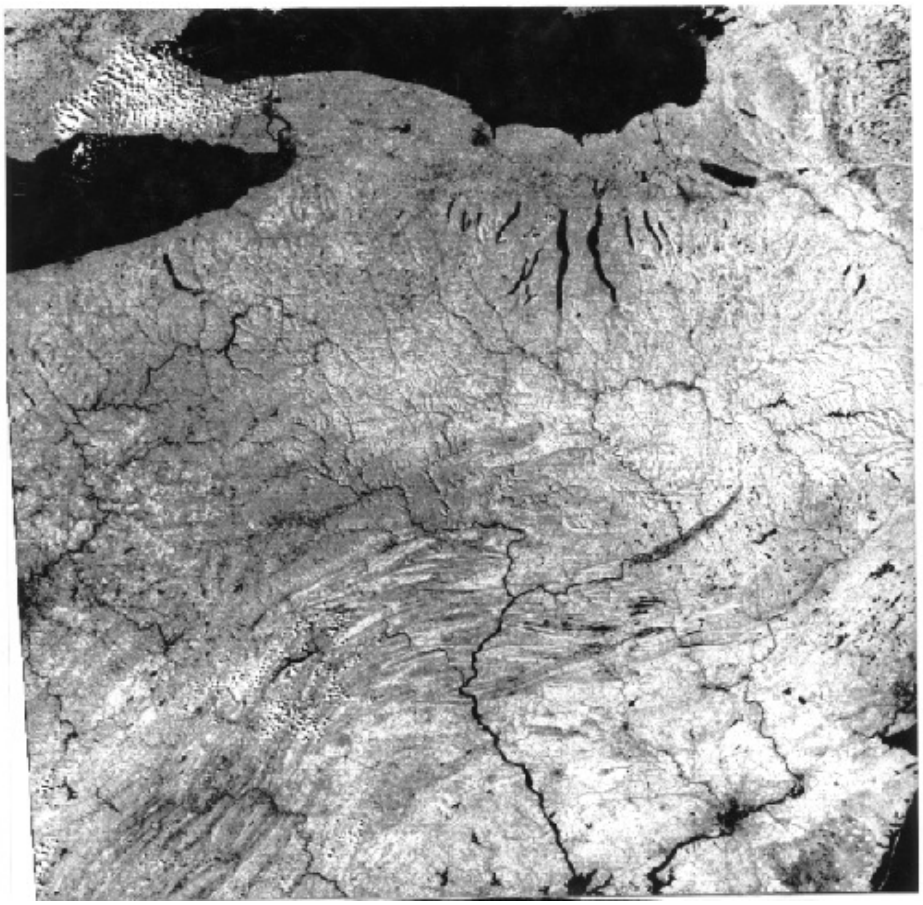
The Rocky Mountains stand out by their dark tones, caused mainly by coniferous vegetation, in sharp contrast to the lighter tones associated with the Plains and Basins. The lofty Rockies are somehow less impressive when flattened in this image.

To gain a feel for the thermal images, let's examine several HCMM scenes that capture part of the northeast U.S. and neighboring Canada. First, an overview: the four images shown cover much of the states of New York and Pennsylvania. The images are, respectively, a Day-Vis

(upper left); Day-IR (upper right); Night-IR (lower left); Apparent Thermal Inertia (ATI) (lower right):



Lets look at each in more detail, using enlargements. Each covers a 467 km (290 mi) wide swath. The first image is extracted from a full Day Visible-Near IR image; all four to be shown were obtained on September 26, 1978.



The largest scale features are Lakes Ontario and Erie. The Finger Lakes of New York are evident. The main rivers in the scene are the Susquehanna and Delaware Rivers, visible mainly in the lower right third of the image. Philadelphia and New York City appear as dark tones.

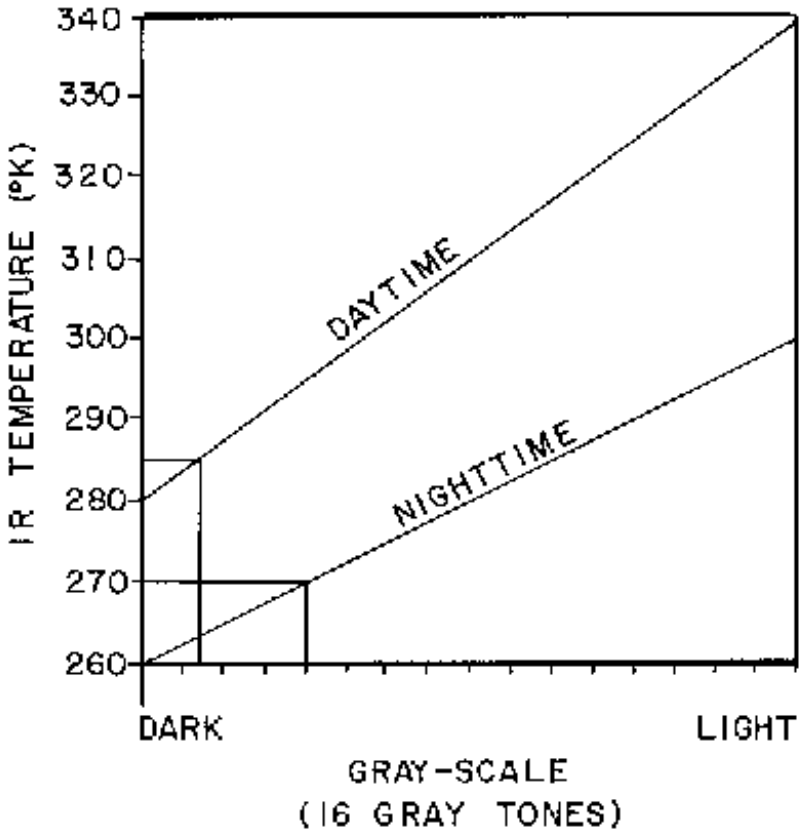
Compare this image with the MSS Band 5 view of the same area shown in the Introduction. The regional geology is dominated by the Coastal Plains and Piedmont (lower right), the fold belt of the Appalachians, the Appalachian Plateau, and glaciated upstate New York into Canada (upper left). In the right center is a narrow, curved dark pattern that we identify as the Wyoming Valley of eastern Pennsylvania (Wilkes-Barre/Scranton areas), a major anthracite coal belt. Note the few cumulus clouds.

The Day Thermal image taken at the same time presents distinct differences.

The thermal structure of the lakes is evident. Similar to the Thematic Mapper Band 6 image, Lake Erie is warmer than Lake Ontario. The white clouds west of Buffalo are depicted as very dark (cold) in the thermal image, consistent with their generally low temperatures as condensation in the atmosphere. Much of the land shows in medium grays. Very dark areas associate with the fold ridges and with sections of the Appalachian Plateau. These areas relate to heavy deciduous tree cover in these mountainous areas. The trees cool their surroundings by evapotranspiration. Five major metropolitan areas - Buffalo, Rochester, Syracuse, New York City/New Jersey, and Philadelphia/Wilmington - stand out as very light tones, indicating warmer temperatures. A few rural areas also are light (warm) for reasons not obvious. But the Wyoming Valley stands out from its surroundings by its very light tone, which is the result of higher radiant temperatures caused by the widespread dark shale, mixed with black coal dust (blackbody effect).

Next, we show a nearly cloud-free full image of nearly the same area, but extending into western New England, across Long Island and down to the mid-Chesapeake Bay. This image in the night of November 2, 1978. Lakes, rivers, and the ocean appear much warmer than the cooler land. This is an apparent paradox: even though the water has cooled somewhat during the night, it normally experiences smaller δTs than the land.

During the day water is notably colder than the land, and in the night it appears warmer than the land, because of its heat retention and the commonly larger drop in land surface temperatures. This plot may clarify this idea.



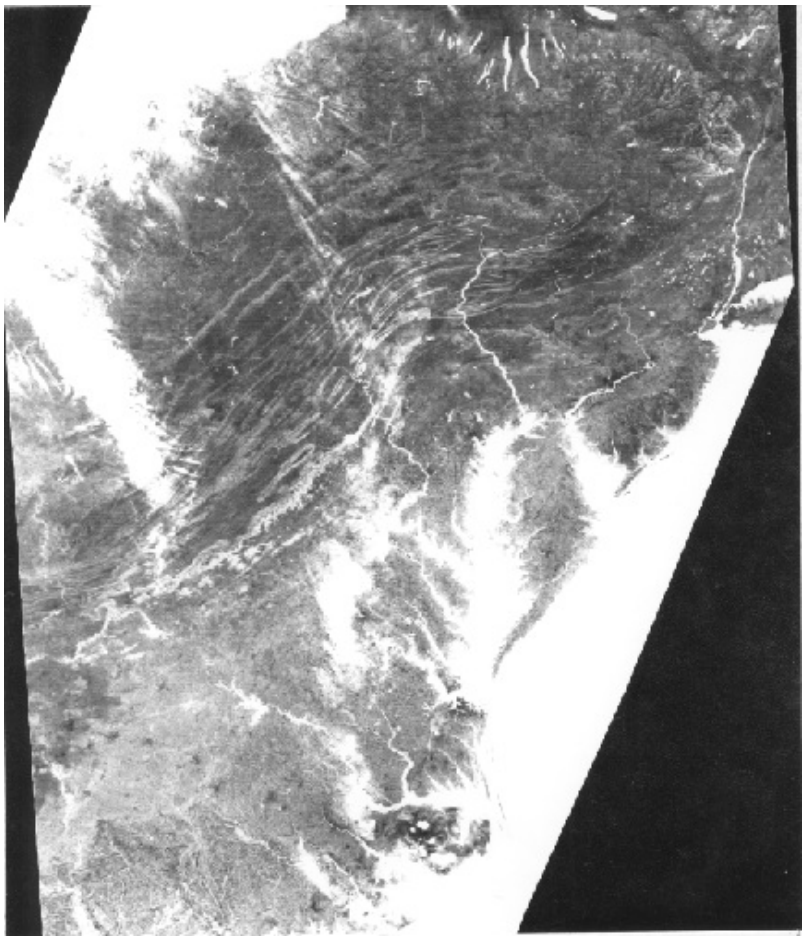
This plot shows the assignments of gray tones as a function of radiant temperatures for the Day and Night images. Here, a temperature of 285°K on the land has a darker gray level than a temperature of 270°K for that same surface in the night. If (case not shown in the plot), water during the day had a temperature of 280°K and dropped at night to 275°K, we can extrapolate these values to the two straight-line plots: the 280°K gray level would be very dark and the

275°K level would be lighter than the corresponding land value. This reasoning accounts for the relative land-water gray tones noted in the Day-IR and Night-IR images shown above.

In the Night-IR image, cities are slightly warmer than their surroundings. The ridges in the folded Appalachians appear warmer, largely because they have dropped their leaves (thus, no longer cooling by evapotranspiration), and underlying rock units contribute to the thermal response. The Wyoming Valley is not emphasized by warm emission from the coal/dark soil surface and is only locatable from the ridge pattern enclosing it. Some valleys appear conspicuously dark, perhaps because of cold-air drainage from uplands. Dark cloud banks (cold) are evident north of Lake Ontario.

9-17: **Bring back to memory your knowledge of the Harrisburg region you built up from taking the test at the end of Section 1. In the visible and thermal images shown on this page, locate features from the test scenes and note their tonal patterns in the thermal images. ANSWER**

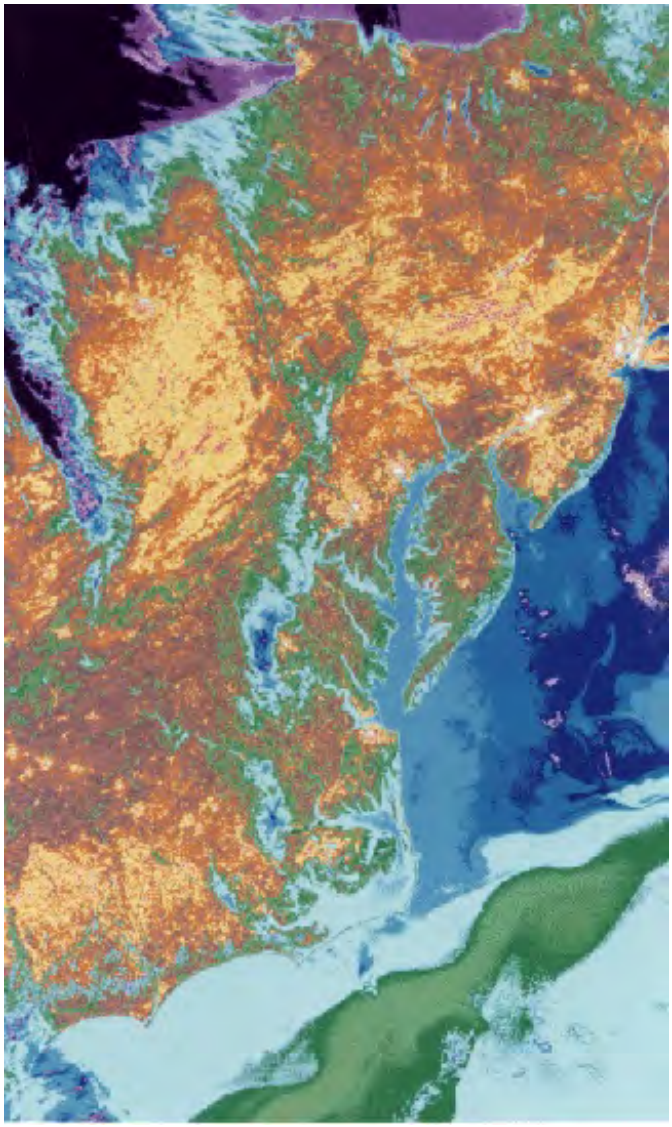
The first ever space-acquired ATI image came from satellite observations of this eastern U.S. area (HCMM Day images on May 11 and a night image on June 11, 1978).



Water has a very bright tone, as do clouds (upper left) and snow (not in this image). The value of (1 - a) is near the maximum for water (with a moderate δT), hence a large ATI. But, for clouds and snow, having high albedos, thus tending to lower ATI, the δT is quite low, countering the (1 - a) effect by raising ATI (because of its position in the denominator). A typical ratio of (1 - a)/δT for water is 0.98/3 = 32.7; for snow is 0.40/2 = 20.0; and for moderately reflective soils is 0.70/20 = 3.5. As a generalization, vegetation has moderate to low ATIs, as do many soils, while basalt has a very low and granite a moderate-to-high ATI. Heavily forested areas in the Appalachians are dark, denoting low ATIs. Those soils in the Piedmont and Coastal Plains, where they are exposed better because of fewer trees, have somewhat higher ATIs. We can crudely separate the Piedmont rom the Plains by its lighter tones.

The same Day-thermal image, which we used as partial input in making the following ATI image,

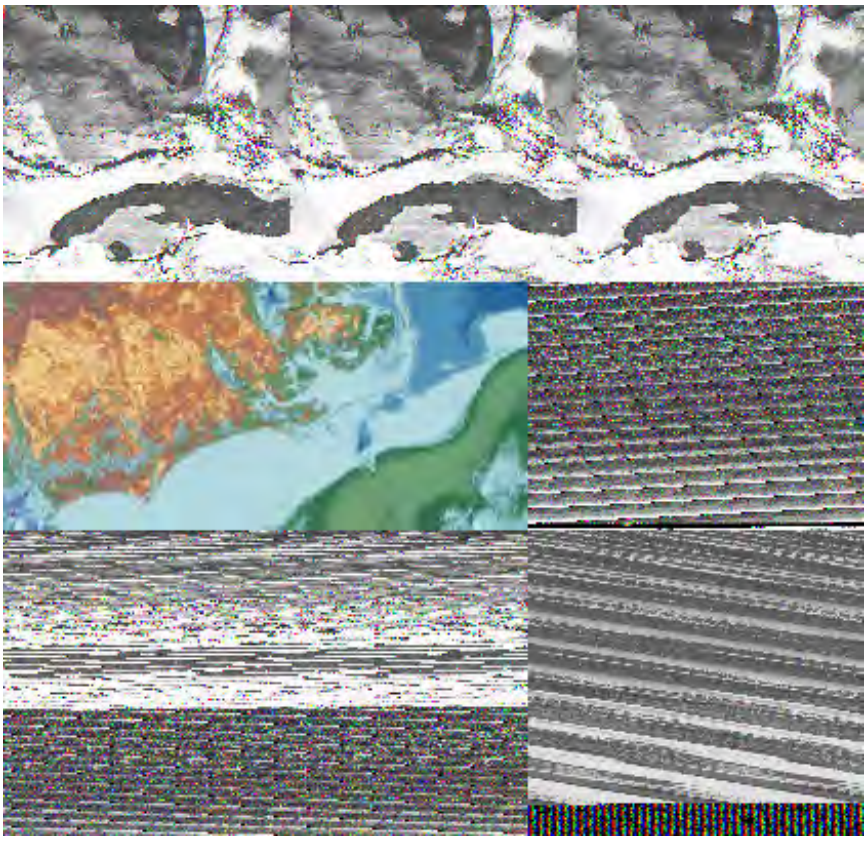
reveals a prominent thermal pattern in the Atlantic ocean, when we assign colors to the different calibrated temperatures.



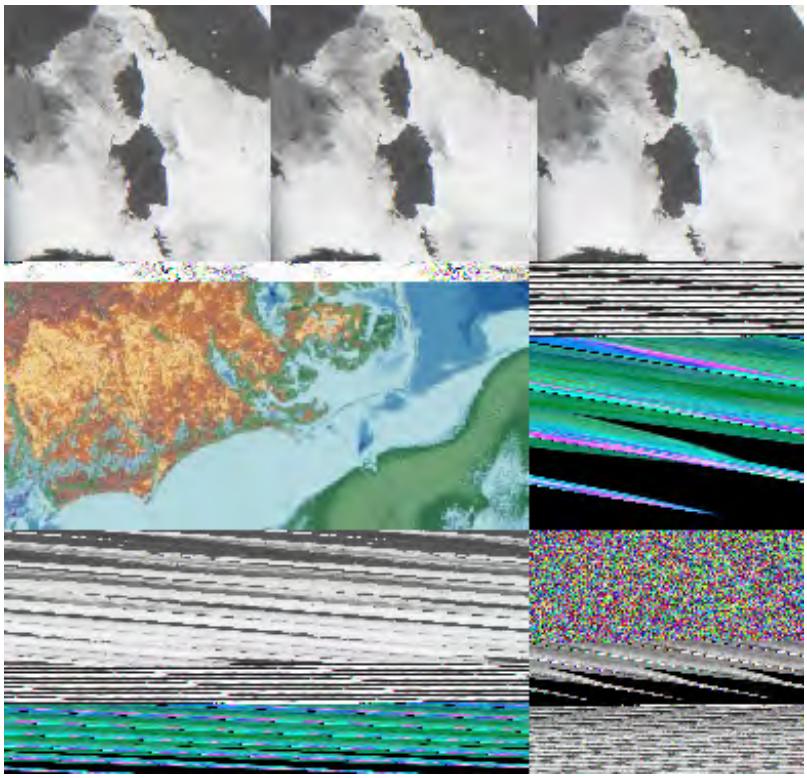
9-18: **Account for the black, purple, and blue colors on the continent. Where is the coldest part of the Atlantic Ocean in this scene? What is responsible for the green band in that ocean?** **ANSWER**

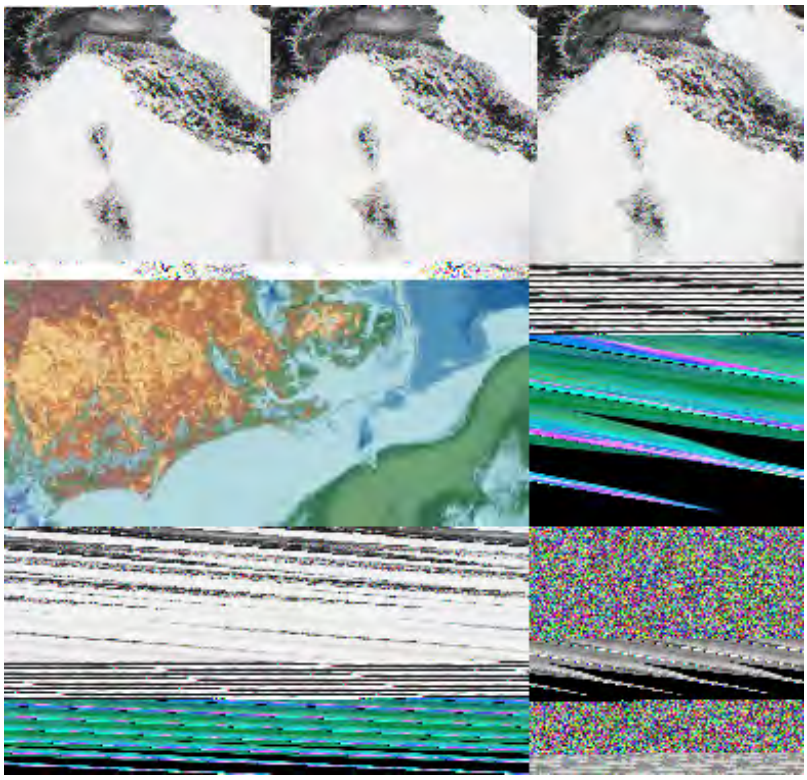
Several shades of darker blue mark zones of colder water. Near the bottom, within a lighter blue body, is a greenish curving pattern that represents the somewhat warmer Gulf Stream moving northward enroute to the Outer Banks off Newfoundland.

A Night Thermal-IR image made on December 20, 1978 of the Gulf Stream east of Florida and of waters around the Bahamas, Cuba, and the Gulf of Mexico indicate rather complex thermal patterns in the seawater. Some of this may relate to thin cloud cover over stretches of the ocean. Note that all land is dark (cooler), so that it is hard even to find the Bahamas.



Nighttime thermal images taken at different times of the year can be quite different in tonal representations and hence in temperatures. Look at this pair of Night-IR HCMM images of the western Mediterranean in which the northern half of Italy, Sardinia, Corsica, the French Riviera and some of the Alps are displayed. Note the main differences.



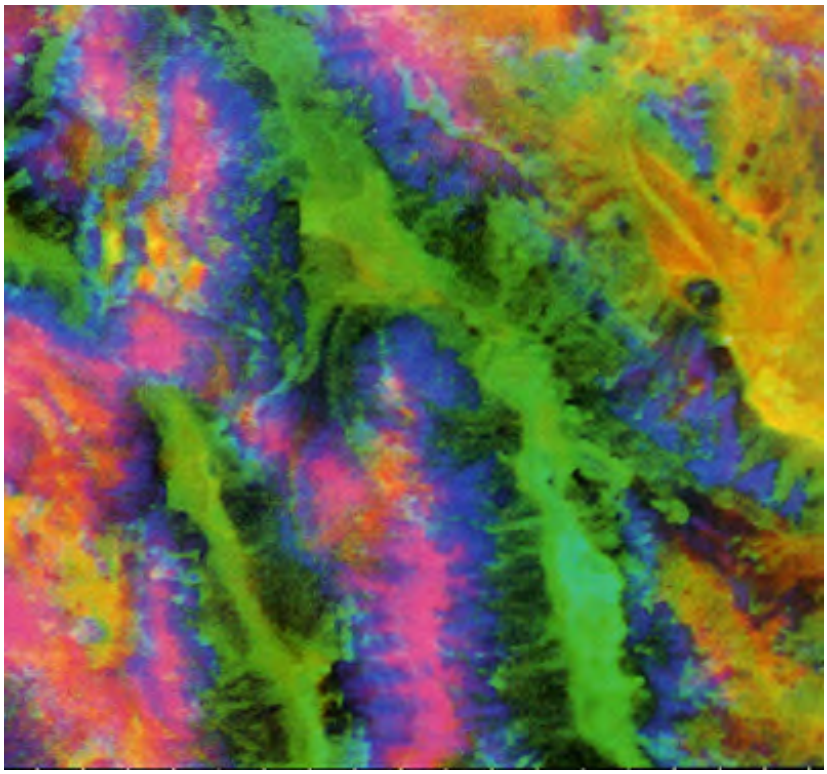


In the July image, the land is almost uniformly cooler than the seawater. In the seawater, subtle thermal patterns, to some extent the result of variable winds in different areas, are visible. The cooler gray pattern on the left results in part from Rhone River water entering the Mediterranean. Blackish areas over the sea are cold clouds. The November scene was processed to bring out temperature variations on the land. A contrast stretch lumped lighter tones in the seawater into a single value, thus shifting the darker pixels into a stretch that emphasizes smaller temperature difference. On the land, higher elevations are darker, valley/lowlands are lighter.

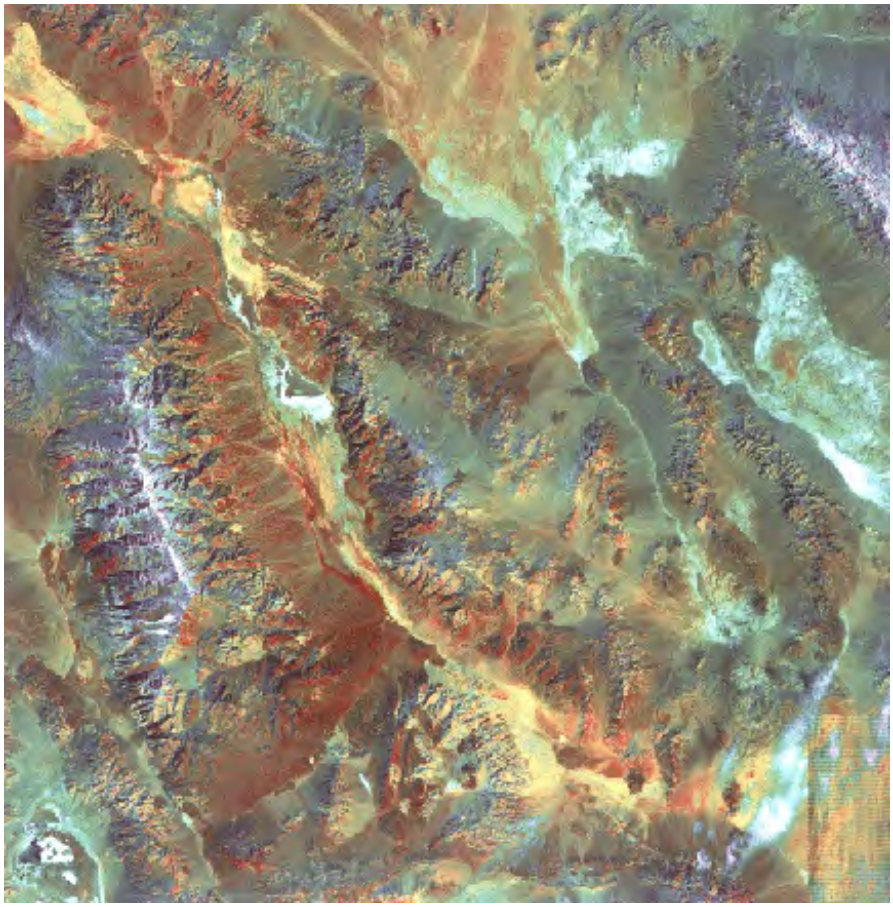
JPL operated HCMM to prove, or get further insights into, applications in most of the same disciplines addressed by the Landsat program. Among the objectives successfully investigated were:

- Production of thermal maps useful in distinguishing rock types and locating resources
- Measurements of plant-canopy temperatures to determine stress and transpiration
- Monitoring soil moisture content and changes during seasonal and diurnal cycles
- Pinpointing natural and man-made effluents and observing thermal gradients in water
- Prediction of runoff during repeated coverage of snow fields
- Correlation of urban heat island effects with local climatological changes

To evaluate two geological examples of what HCMM can accomplish, we turn first again to the Death Valley region, now expanded in this HCMM color composite to cover a wider area. In this rendition, the bands were assigned these colors: Day IR = blue; Day Vis = green; Night IR = red. As interpreted and Anne Kahle of JPL, many rock units can be discriminated as groups by the different colors. But, in fact, a group may contain several rock types or categories that are considered distinctly different in composition and character by geologists.

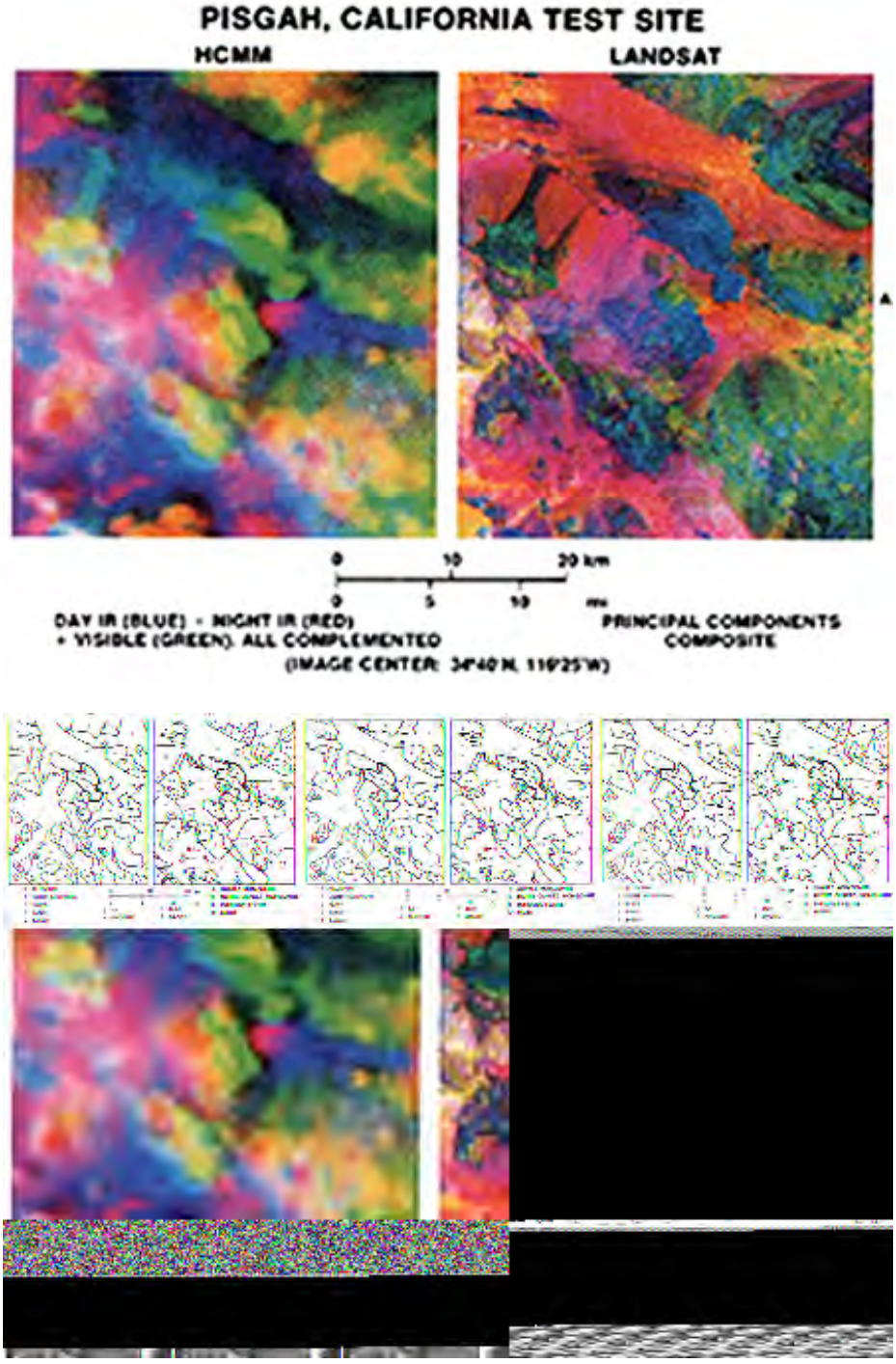


Dr. Rupert Haydn of Bavaria has made a fascinating special product in which the Death Valley Landsat scene (here cropped to show only the northwestern 60%; D.V. is in the upper left quadrant), comes from a TM 4 image (I), convolved with the TM 6 thermal image (H) and a HCMM Day IR image (S). These letters in parenthesis refer to a different method for making color composites known as the IHS system, where the three primary colors are assigned to derived intensity (I), hue (H), and saturation (S) parameters.



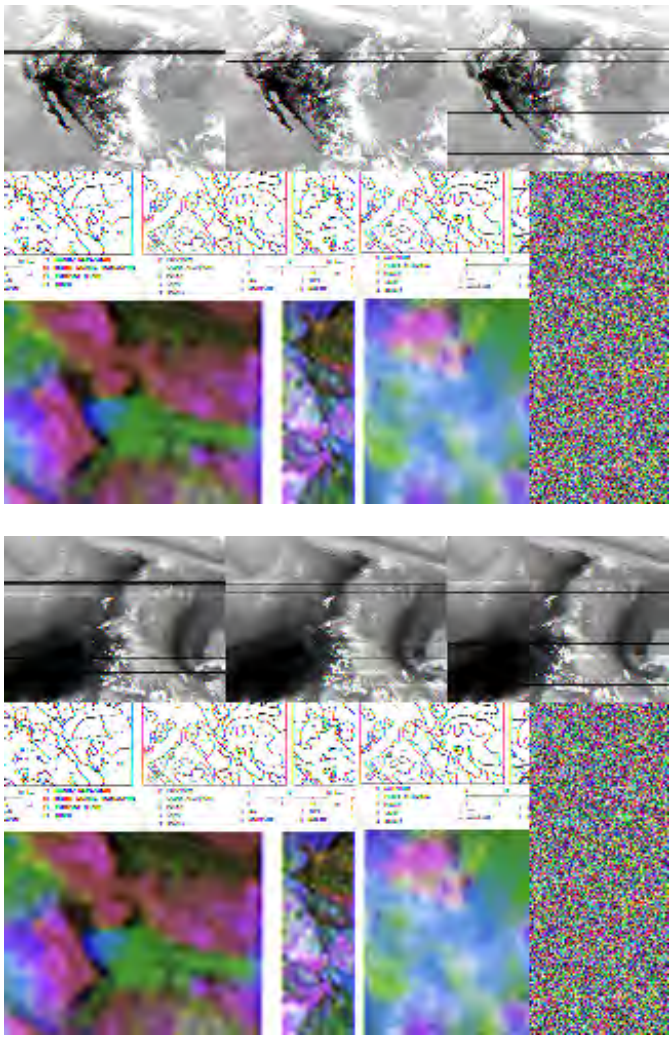
Again, reds denote hot, blues cool, and yellows and greens intermediate temperatures. As with the Thermal IR Multispectral Scanner imagery, the alluvial fans show in reds, some salt deposits in blues and greens, and certain playa beds in yellow.

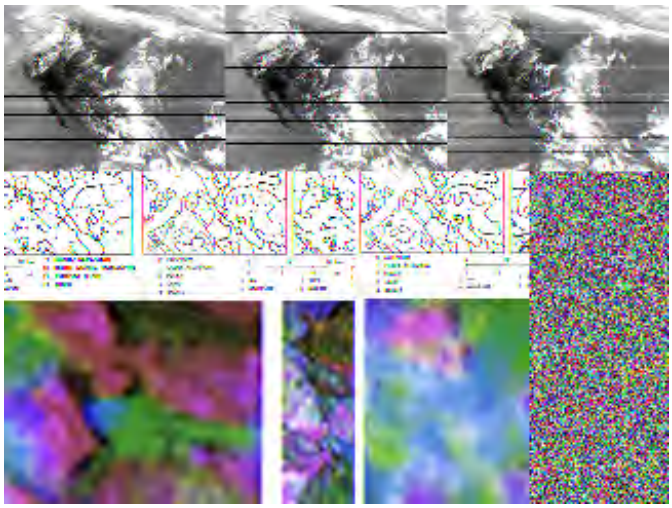
Now, consider this definitive example of a geological application of HCMM data, developed by Anne Kahle and her associates at the Jet Propulsion Laboratory.



The area examined is in the eastern Mojave Desert of California, in which Basin and Range topography similar to the Death Valley region dominates the landscape. The color image in the upper right is a Principal Components Composite, using the four Landsat Multispectral Scanner bands. Correlation of color units with rock types and deposits mapped in the field is expressed in the map at the lower right. Of particular interest is the heavy lined feature near center, which encompasses a basaltic cinder cone and associated lava outpourings known as the Pisgah Crater. The corresponding scene (upper left) constructed as a composite of HCMM Day IR (blue), Night IR (red) and Day Visible (green) shows similarities and differences. Alluvium clearly shows in blue tones, and more silicic volcanics appear in red. The lava flow emanating from Pisgah Crater appears now to consist of two units, poorly separated in the PCA product: basalt with a pahoehoe-like (ropy) texture and basalt with a (cindery) texture. Comparison of the geologic maps made from Landsat and HCMM data shows very good agreement in singling out the same units.

Sensors operating in the 3-6 µm and 8-14 µm regions have been on many of the meteorological satellites flown now for four decades. This will be considered more specifically in Section 14. For now, here is one series of images taken from the GOES-11 satellite (launched in Spring 2000) showing western North America. In this June 10, 2000 sequence, four channels are represented: 1) Visible (general cloud distribution); 2) 3.9 µm (good forest fire detector); 3) 6.7 µm (water vapor); 4) 12 µm (cloud decks).





Many meteorological satellites today have thermal sensors onboard but the most useful ones can do a general vertical temperature profile

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

>

Thermal imaging, as well as use of specialized thermal detectors, has grown into an important technology that is applied directly by users on the ground or from low-flying aircraft. The application is commonly called thermography. Using thermal sensors coupled with optical systems, one is able to "see in the dark" by detecting varying temperatures from different objects in the scene. Typical users include hunters and law enforcement officials; there are a number of military uses. Another frequent application is to check for heat loss from buildings or thermal contamination in streams.

Thermography; Night Vision Systems

We would be remiss if we didn't bring to your attention a wide range of applications of thermal remote sensing that involve aerial flights and, now routinely, ground studies with infrared instruments that are portable - even hand-held - and easy to master with little training. Most of these uses involve operating the equipment at night. Probably the most dramatic displays of this expanding technology were the graphic scenes during the Gulf War of 1991 in which CNN and other networks showed the Iraqi defense in Baghdad as a series of greenish-color live-action firings of anti-aircraft tracer flights and occasional hits by incoming missiles. Or, tanks rolling into Iraq's capital in 2003; this is a thermal IR display using a camera equipped with a sensor that responded to the thermal variations of features, from buildings to weaponry, in the scene. :

Many other applications can be cited: in industry, process control, energy audits (heat expenditure and loss), machinery function; in law enforcement, police surveillance; in firefighting, search and rescue, wild fire reconnaissance, smoke penetration; in medicine, abnormal variations of body heat related to disease and malfunctions, in environment, wildlife observation and management, oil spill detection; and much more.

All of these tasks depend on detecting temperature differences in objects that vary in their thermal response. For example, hot spots in a building could indicate inefficient energy distribution including unwanted heat leaks. Tanks and soldiers on the move can be monitored at night. Felons breaking into a closed bank stand out as warm bodies moving about in a suspicious manner. An abnormally warm area around the human throat could mean a thyroid disorder.

This now widely used technology, called Thermography, depends on scanners and cameras with detectors sensitive to emitted thermal radiation. In the 3-5 µm region, detectors made of Indium-Antimony (InSb) and Platinum silicide materials are sensitive to radiation in that spectral interval, with thermal photons knocking off electrons to create the varying signals that indicate temperature differences. In the 8-14 µm region, Mercury-Cadmium-Telluride detectors are used. Detectors that respond to quantum effect radiation require cooling. Others, that rely on what is known as the pyroelectric effect, can be uncooled; these include radiometers and ferroelectric bolometers. Depending on what is done with the signals, the imaging systems can be shown visually in black and white, in green tones, or in color (by assigning different colors to different temperatures). The sensors can detect over wide temperature ranges (e.g., from - 20 to +1100°C) while being able to measure temperature differences as small as 0.5 to 1.0 degree. We show examples of two instruments. First, is a hand camera marketed by Sierra Pacific Innovations (check their interesting Home Page) called the IRM 700.

This camera, which has a Platinum silicide detector sensing radiation between 1.2 and 5.9 µm, produces high resolution images, like the one below, to a kilometer or more range.

The second is even more exotic. Here is a set of Night Goggles which can be worn by a person directly over the eyes. It has proved invaluable to police raiding a building, firefighters, and hunters, as well as military personnel in combat.

9-19: **Think of several (imaginative) ways in which you could use Night Goggles.**
ANSWER

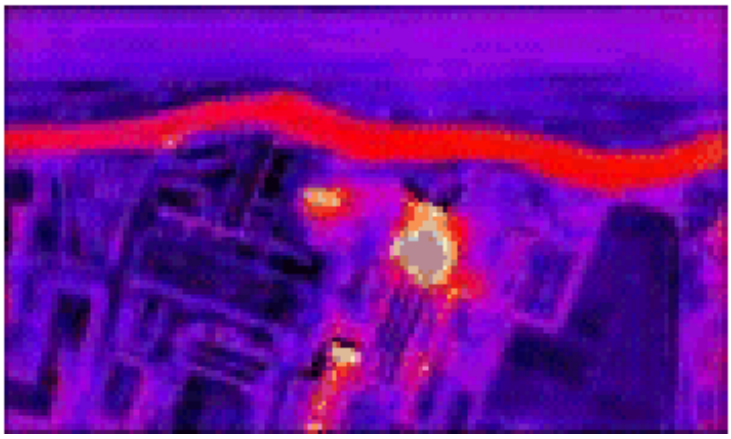
This next image shows the view through a telescopic sight adapted to thermal infrared detection of a deer (whether this is a hunter's target or the subject of a naturalist's study is not known).

We can learn more about these diverse uses of thermography - especially the Night Vision mode - just by looking at some representative images. Let us turn first to energy budget assessments, in which both aerial and ground thermal scenes look for abnormalities in heat distribution. Start with this aerial scanner night view of a residential area.

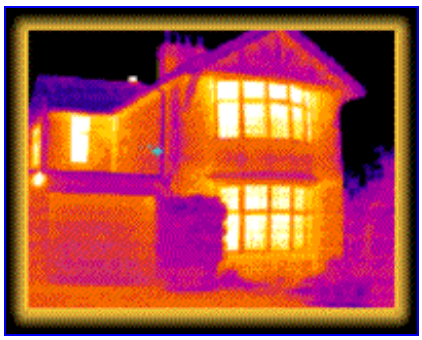
The warmer areas are the river, particularly a light-toned plume that results from industrial fluids being dumped and a cooling pond that receives water from a nearby plant (warm signature). In the residential section, houses are cool but the streets display higher temperatures since they are asphalt covered (this black tar material absorbs solar energy during the day and re-radiates thermal energy at night).

The next scene is a colorized depiction of temperature variations in a series of homes and other

buildings. The reds and white denote "hot spots" - either from small workshops or from individual houses in which excessive heat is being lost.



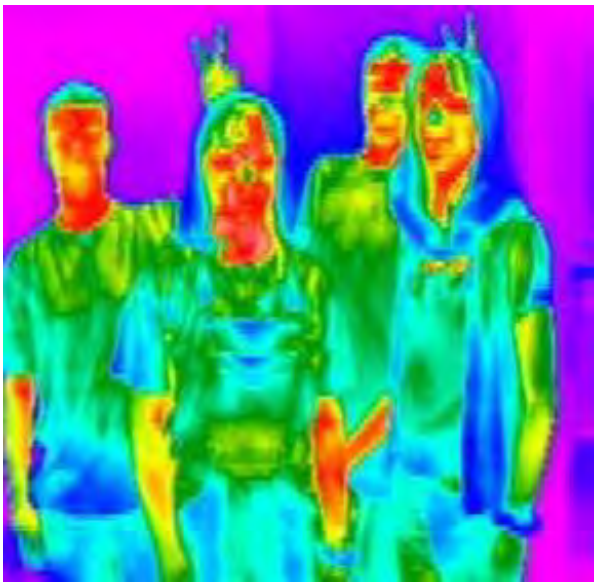
Heat variations abound in individual houses, as shown here:



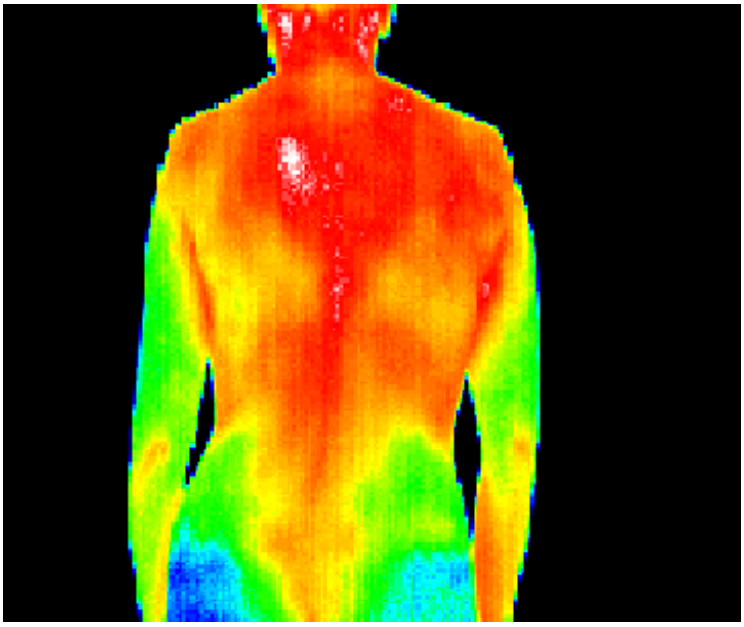
At the street level, this night scene shows people, an auto, and houses in terms of relative hotness:

9-20: **Note the house on the left, and particularly its roof. There are two bright pattern "hot spots" on or near the roof. What might they be? Account for the hot areas on the auto in the right part of the scene. ANSWER**

Individuals in everyday clothing appear as distinct temperature variants. Their faces usually are warmer (reds) than their outside clothes (in cooler greens and blues):

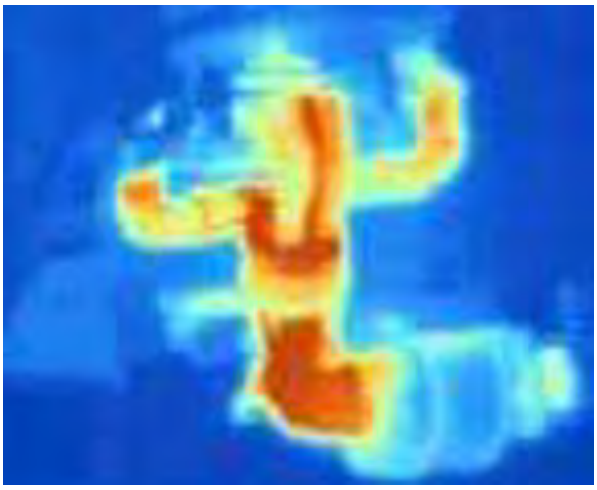


The human body, unclothed, reveals surprising differences in temperature, much being normal but occasionally with areas whose thermal departure from the norm may have significant health implications. In this case, the small white area on the back is indicative of a potential problem.

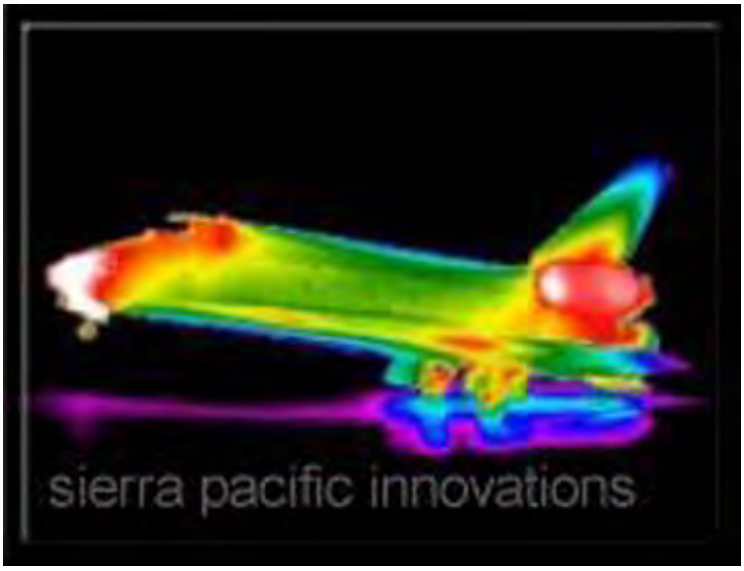


Individual inanimate objects experience temperature differences as they operate. Two examples will suffice. First, peruse this thermogram (another word for thermal image) of a motor doing its thing.

And, consider the heat distribution within this pipe fitting:



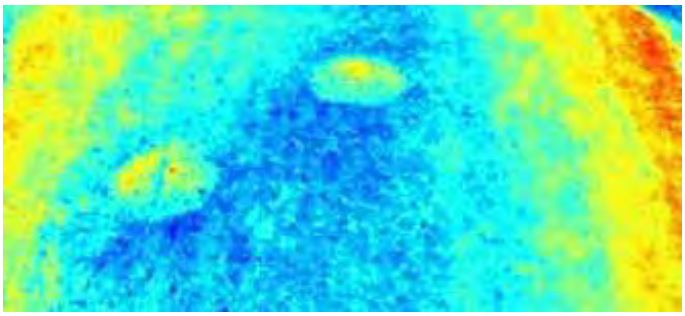
Returning to a space theme, here is a thermal snapshot of the Space Shuttle as it landed at Cape Canaveral.



9-21: **Explain the two larger areas of warmer temperatures on the Shuttle.** **ANSWER**

Finally, let's take a quick peek at two aspects of military use - here the appearance of a tank and a supply truck operating actively at night during maneuvers.

Explosive devices are a major problem in Iraq. Some are mines buried in roads. Under the right circumstances, thermal cameras can detect these if buried close to the surface:



9-22: **Here is a final question designed to test your powers of observation and reasoning: Two and a half years after the Shuttle Columbia disaster, Discovery (Mission STS-114) successfully returned to space for a two weeks mission. It landed late night (pre-dawn) at the Edwards Air Force Base in California. The left image below is an IR camera view just before touchdown. There seems to be a bright light in the Shuttle's nose. Using knowledge gained in this Section, what really is this glow. A ground picture of the Shuttle on the right affords a key clue.****ANSWER**

We can safely conclude from this review that thermal imagery opens new vistas in seeing the Earth's surface in a way largely unfamiliar to our experience. Thermal sensing is a versatile way to detect and identify features, using properties we seldom sense directly - certainly not with our eyes. There are two other kinds of thermal images lightly considered in this section: views of people and action imaged by "night vision" methods (a variant was amazingly presented as targets in Iraq during the 1991 Gulf War); and views of heat concentrating or escaping around a home being surveyed by thermal scanners for heat leaks. Hot stuff, eh!

PREVIOUS RST NEXT

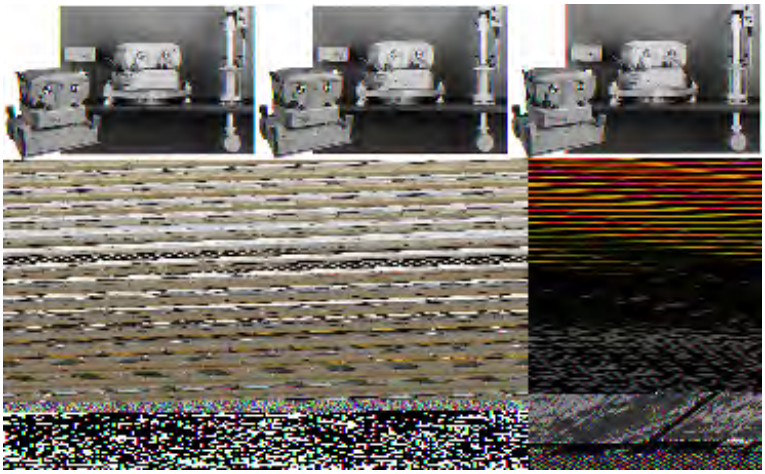
Primary Author: Nicholas M. Short, Sr.

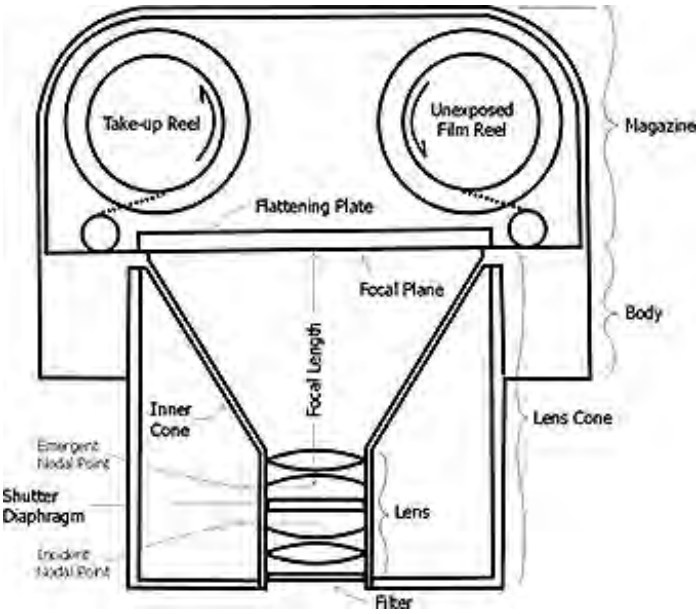
The "ancestor" or progenitor of space images like those made by Landsat, SPOT, IKONOS, et al. is the aerial photograph, in common use now for about a century. And, even today with products from the space systems competively priced, aerial photography remains a powerful and sometimes optimum tool to use for both routine and specialized applications. Two variants of aerial photos, each obtainable in black and white or color, are the vertical (sees Earth straight downward) and the oblique (sees at various angles, usually less than 45°). Photography commonly takes place in the visible but film sensitive to the very near infrared or ultraviolet allows image-taking in those spectral regions. The two main advantages of aerial photography over space imagery are 1) the mission can be flown at any time, weather permitting, and 2) resolution is usually higher (although recent Russian and American commercial products from space are narrowing that gap). Examples of aerial photos of different scales and types are shown. This section is a condensed summary of principles of aerial photography and photogrammetry (more on that in Section 11); for fuller treatment consult references given.

ELEMENTS OF AERIAL PHOTOGRAPHY

Aerial photographs have been a main source of information about what is at the Earth's surface almost since the beginning of aviation more than 100 years ago. A good review of the history of aerial photography, by Prof. Paul Baumann of SUNY-Oneonta, is found at his website.

Until space imagery,aerial photos were the principal means by which maps are made of features and spatial relationships on the surface. Cartography, the technology of mapping, depends largely on aerial/satellite photos/images to produce maps in two dimensions or three (see next Section). Aerial photos are obtained using mapping cameras that are usually mounted in the nose or underbelly of an aircraft that then flies in discrete patterns or swathes across the area to be surveyed (see page 7-1 for more details as to flight plans). These two figures show a camera and a cutaway indicating its operation:





For most flight surveys, the camera film is advanced automatically and wound onto reel spindles at a rate which is tied to the aircraft's speed.

A variant of this camera system is the multispectral camera (also discussed on page 11-1). This type uses separate lenses, each with its own narrow band color filter, that are opened simultaneously to expose a part of the film inside the camera. Here is one such camera developed for use in the Skylab space station program:

Aerial photos are taken from a variety of platforms: airplanes; helicopters; unmanned drones; balloons; kites; tall buildings. For the most common platform - airplanes - most cameras are mounted in the underside of the aircraft. Propeller or JetProp aircraft are preferred, for two reasons: 1) they fly slower, allowing easier film advance; 2) they cost less to operate. This photo shows two such aircraft used by NOAA in its remote sensing programs:

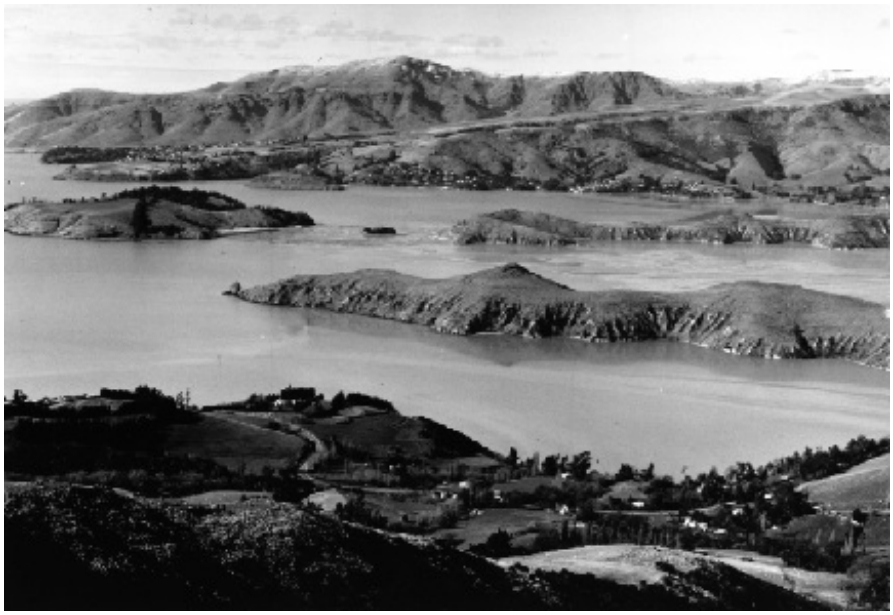
In previous sections, we have employed aerial photography to look closer at areas of which we had satellite based images (such as Morro Bay in Section 1). In fact, satellite image interpretation is in essence an extension of the concepts underlying aerial photography, taken to higher altitudes that allow coverage of larger pieces of real estate. Space remote sensing uses devices that, while much more costly to build and operate, rely on the same physical principles to interpret and extract information content.

Most textbooks on remote sensing are outgrowths of earlier texts that once dwelt dominantly on acquiring and interpreting of aerial photos. New books still include one to several chapters on this basic, convenient approach to Earth monitoring. We shall allot only limited space to explore some essentials of this expansive topic in the present Section and the next. In Section 11, we consider photogrammetry as the tool for quantifying topographic mapping and other types of mensuration. For anyone seeking more details about aerial photography/photogrammetry, we recommend consulting the reading list in the RST Overview (first page), and/or going to Volume 1 (Module 1) of the Remote Sensing Core Curriculum. Below is a recommended entry from that reading list:

Avery, T.E. and Berlin, G.L., Fundamentals of Remote Sensing and Airphoto Interpretation, 6th Ed., 1992, MacMillan Publ. Co., 472 pp.

Examples of Aerial Photos

An aerial photo is just a black and white (b & w) or color "picture" of an area on the Earth's surface (plus clouds), either on print or in a transparency, obtained by a film or digital camera located above that surface. This camera shoots the picture from a free-flying platform (airplane, helicopter, kite or balloon) some preplanned distance above the surface. Two types depend on the angle of view relative to the surface. The first, oblique photography, snaps images from a low to high angle relative to vertical. The example below is the most common type (high oblique), showing Lyttleton Harbor, near Christchurch, on South Island of New Zealand, with more detail in the foreground and a panorama with reduced detail in the background.



10-1: For the moment we shall define resolution in a photograph as the size of the smallest object whose tonal appearance is notably different from its surroundings or background; technically there is a more precise definition, given in terms as the minimum spacing between two dark lines embedded in a light background that can be visually separated. How does spatial resolution vary in this oblique photo. ANSWER

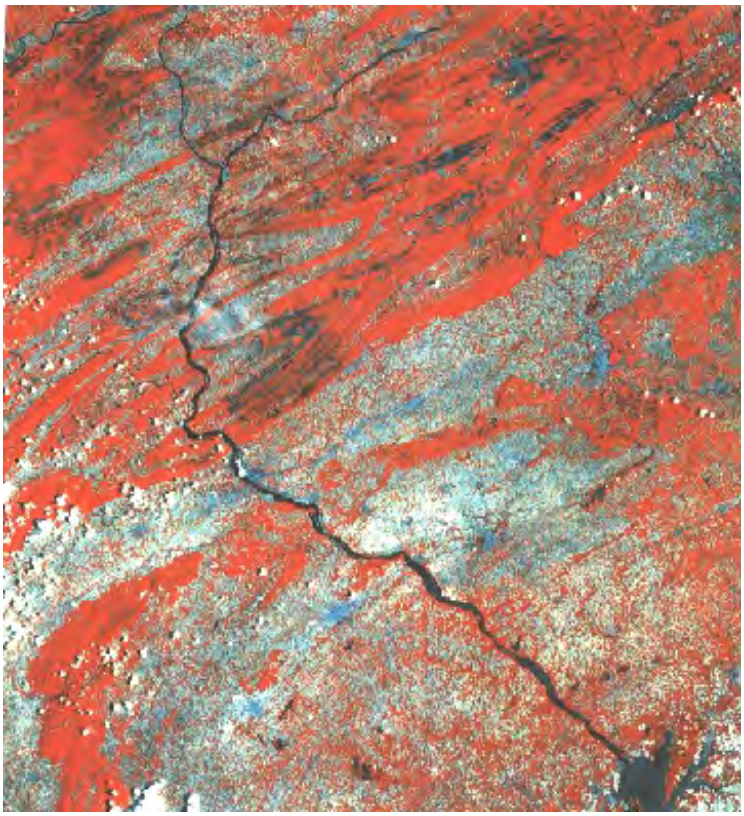
The second type of aerial photos is oriented vertically, that is, it results from pointing the camera straight down (to the nadir, at the photo center point) to show the surface directly from above.

The size of the photo and the sizes of the features represented within the photos can vary depending on the following: the camera's optical parameters, the surface area of the exposed film (frame size), the subsequent printing sizes (e.g., enlargement), and the altitude of the camera platform.

Image Scale

The ratio of the size of any object, feature, or area within the photo to its actual size on the ground is called the *scale* (defined and discussed on the third page of this Section).

We now present a series of aerial photos, acquired at different times and scales, most covering areas that lie within this June, 1977, Landsat image (original scale = 1:1,000,000) of south-central Pennsylvania, a scene we have looked at in earlier Sections, and especially during the Exam at the end of Section 1.



This scene contains heavily forested fold ridges. Some of the bluish-black areas are defoliation patches caused by the Gypsy Moth. Others areas near top right are surfaces covered with black dust from the Anthracite coal strip mining in fold valleys. Bluish areas in the wide valleys are fields still bare or with early stage growth. The Susquehanna River which empties into the top of Chesapeake Bay bisects the image. Near the left center, a blue pattern with spokes is Harrisburg, the state capital, with York below it and Lancaster to the right. Next, we show a standard medium-scale (moderate area of coverage but with considerable detail [individual buildings still visible]), black and white aerial photo of part of Harrisburg. The scale value given is that of the original photo before it was reduced to your screen size; quoting this value helps to appreciate what can be seen (resolved) at that scale, no matter what the eventual picture size becomes through enlargement or reduction.

Harrisburg (Scale = 1:100,000)

The number in the upper left corner of this black and white photo of Harrisburg is the date; on its right is the Mission number; and in center is a number denoting the flight line and particular photo within that line. Individual fields, smaller rivers, bridges, and roads are easily picked out.

10-2: One meaning of scale is this: 1 inch on the photo equals X inches on the ground. For the 1:100,000 photo above, determine how many feet are represented by an inch (on the photo, or in this case, the image on your screen) and likewise how many mile(s) extend across that inch. ANSWER

The next photo is large scale (small coverage area and high resolution for identifying features smaller than buildings, e.g., cars) and covers an area within Harrisburg, just east of the previous photo, bisected by Interstate 83. Note particularly the lake-filled quarry (left center).

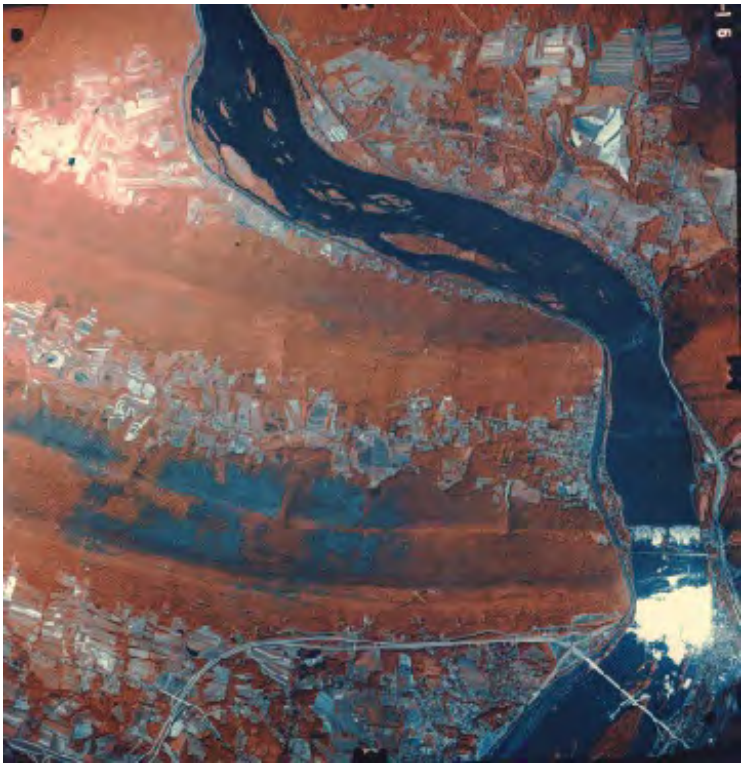
Urban Harrisburg (1:4,000)

10-3: In which other photo on this page can you find the quarry lake? For the above photo, what is/are the number of miles represented by an inch on the screen? Make an educated guess as to the effective resolution of this 1:4000 photo; how did you do it?
ANSWER

In the lower right corner of the Landsat image is an agricultural area along the Chesapeake and Delaware Canal. Its expression in a moderately large-scale, natural-color photo is shown here:

Natural Color Photo (1:24,000)

At a still smaller scale, we next show a false-color, IR image of the Susquehanna Water Gap passing through Blue Mountain just north of Interstate 81 (bottom of the picture) that, to the east, runs along the north side of Harrisburg.



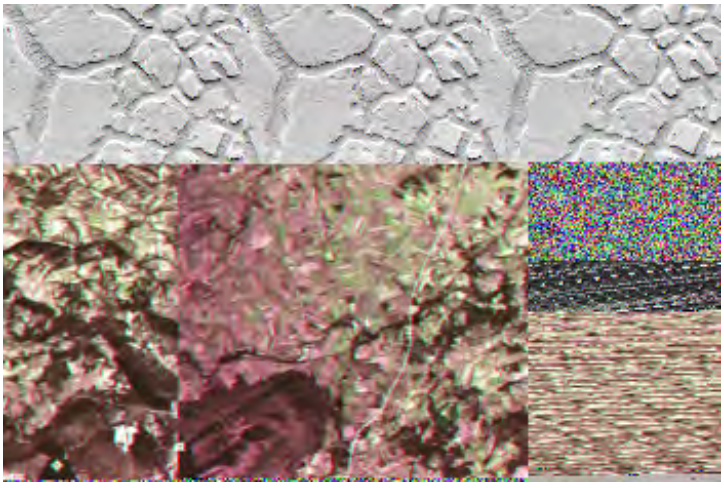
Color IR Photo (1:8,000)

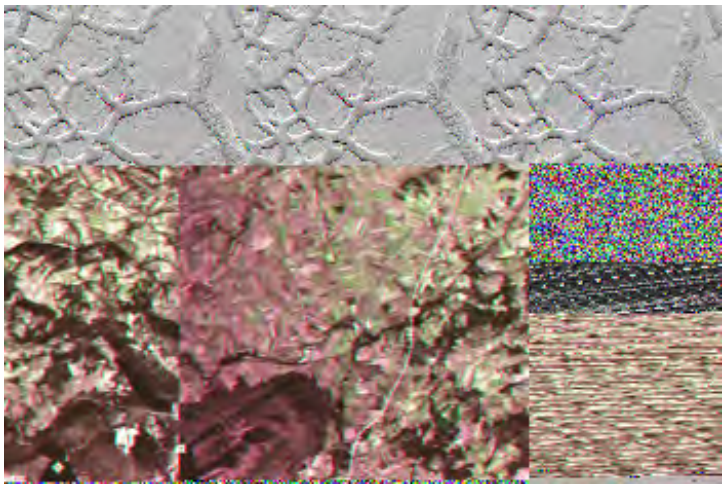
Much the same area is part of a small scale (large area coverage with reduced detail) aerial photo obtained from an RB-57, NASA aircraft, flown at an altitude near 15,200 m (about 50,000 ft) on February 5, 1974. On this date, the color-IR photo shows limited red tones from fields in which winter wheat is growing. The image is 25.2 km (15.7 mi) on a side (635 square km; 246 square mi).

High Altitude Aerial Photo (1:141000)

10-4: There is an easy way to determine whether a scale is large, medium, or small, by looking at its stated value, e.g., 1: 30,000. Propose a simple rule for this. ANSWER

A word of caution at this point. Because of shadow orientation and other factors, features in a photograph (much rarer in space imagery) that represent relief (differences in elevation), such as hills, can appear to the eye as inverted, i.e., a high appears as a valley, a valley as a hill. The best example the writer (NMS) has found is a group of mesas and troughs on Mars. On the left is the correct expression (mesas look higher); on the right is the inverted case. If you ever see an aerial photo that does not look right (expected highs show up as lows), just reorient the photo (a few people may need to reorient their brain).





Among the most obvious features in a photograph are tones and tonal variations (as grays or colors) and patterns made by these. These, in turn, depend on the physical nature and distribution of the **elements** that make up a picture. These "basic elements" can aid in identifying objects on aerial photographs.

Tone (closely related to Hue or Color) -- Tone refers to the relative brightness or color of elements on a photograph. It is, perhaps, the most basic of the interpretive elements because without tonal differences none of the other elements could be discerned.

Size -- The size of objects must be considered in the context of the scale of a photograph. The scale will help you determine if an object is a stock pond or Lake Minnetonka.

Shape -- refers to the general outline of objects. Regular geometric shapes are usually indicators of human presence and use. Some objects can be identified almost solely on the basis of their shapes: for example - the Pentagon Building, (American) football fields, cloverleaf highway interchanges

Texture -- The impression of "smoothness" or "roughness" of image features is caused by the frequency of change of tone in photographs. It is produced by a set of features too small to identify individually. Grass, cement, and water generally appear "smooth", while a forest canopy may appear "rough".

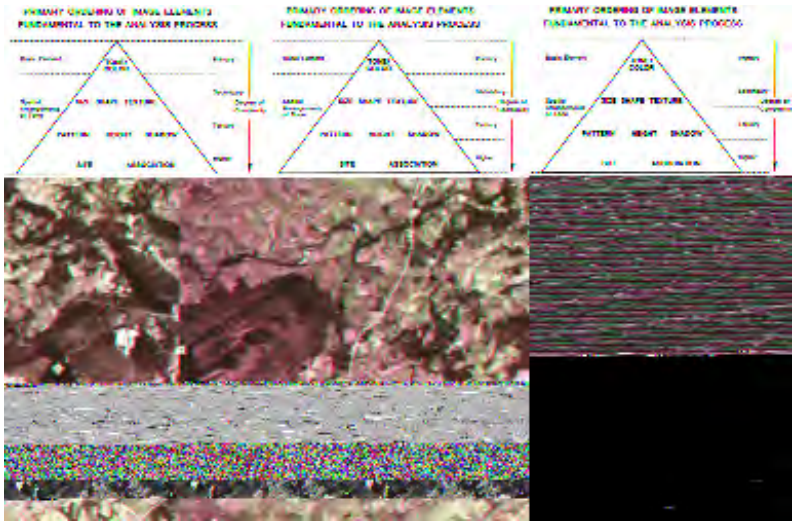
Pattern (spatial arrangement) -- The patterns formed by objects in a photo can be diagnostic. Consider the difference between (1) the random pattern formed by an unmanaged area of trees and (2) the evenly spaced rows formed by an orchard.

Shadow -- Shadows aid interpreters in determining the height of objects in aerial photographs. However, they also obscure objects lying within them.

Site -- refers to topographic or geographic location. This characteristic of photographs is especially important in identifying vegetation types and landforms. For example, large circular depressions in the ground are readily identified as sinkholes in central Florida, where the bedrock consists of limestone. This identification would make little sense, however, if the site were underlain by granite.

Association -- Some objects are always found in association with other objects. The context of an object can provide insight into what it is. For instance, a nuclear power plant is not (generally) going to be found in the midst of single-family housing.

These elements can be ranked in relative importance:



Since aerial photography is dependent on photographs, we need, at this juncture, some basic insight into how a photo is made.

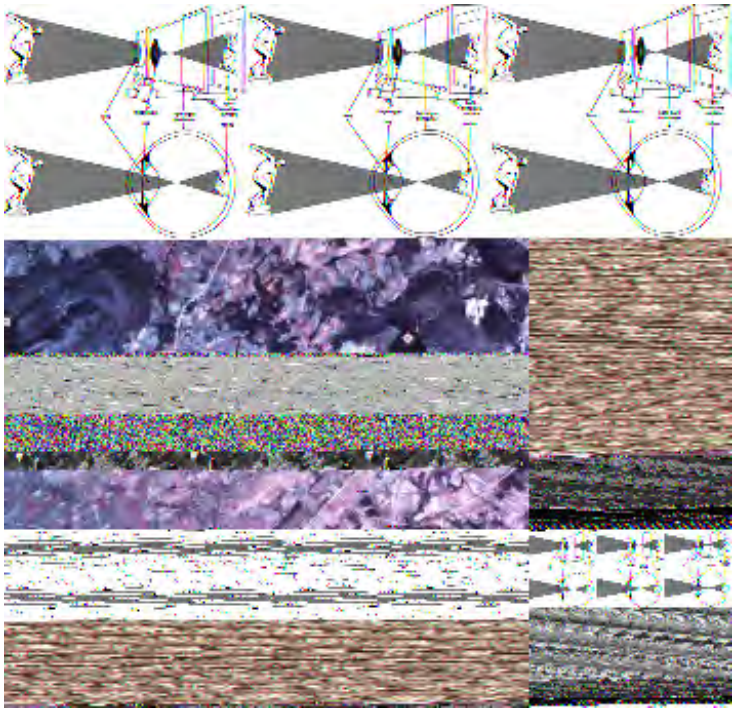
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

This page covers one topic: the photographic process is reviewed in detail. The production of color photographs is also treated in depth. For both Section 10 and Section 11, it may be helpful to review the functions of the human eye, as given in the answer to the first question in the Introduction.

The Photographic Process

Before beginning this page, a review of the answer to the first question, concerning the human eye, in the Quiz at the end of the Introduction may be helpful. With this in mind, use this diagram to compare the components and functions of the eye with that of a photo camera; note both similarities and differences:



Black and white (b & w) photographs start with exposing a light-sensitive film to incoming electromagnetic radiation (light), selected from the spectral range between ultraviolet through visible, and into the near infrared. The optical system of the camera focuses the light, reflected from the target, onto the focal plane (plane of focus). The film is held flat at the focal plane, and the light activates positions in the film area in the same spatial relation that the radiation photons had from the surfaces within the scene. The recorded exposure is a function of many variables, of which the three principal ones relate to the scene, the camera, and the film.:

1) The scene usually contains various objects that contribute their spectral character (reflected wavelengths) and intensities of the reflected radiation.

2) In the camera, we can vary the lens diameter, D, and the effective size of the aperture opening, d.

The aperture depends on the diaphragm width for admitting light. An open/shut shutter controls the duration of light admission. The optical characteristics of the lens vary the distance from the

lens to the film (focal length, f) at which the focus is sharpest. This light-gathering system is adjustable for film response (ISO, formally ASA values);

3) In the film, its properties vary, e.g., which wavelengths it is most sensitive to, and under which conditions does it develop best as a negative and then printed.

For most cameras, the four variables that we normally adjust are:

1) The focus, by moving the lens back and forth, relative to the focal plane, so that the target image is in focus in the plane of the film;

2) The F-stop, defined as f/d, the focal length divided by the effective diameter of the lens opening. Typical values of the F-number are F/1 (the lens opening is the same size as the focal length), F/1.4, F/2 (the lens opening is half the focal length), and F/2.8 to F/22. The denominator increases by approximately the square root of 2 (1.414...), so that each decrease of F/d (i.e., denominator increases), leads to a decrease in the amount of light admitted by a factor of 2. Thus the F-number increases as the lens diameter decreases, and therefore we photograph dark scenes at low F-stops, e.g., F/2, and bright scenes at high F-stops, e.g. F/22.

3) The shutter speed (typically, in a sequence from 1/2000, 1/1000, 1/500, 1/250, 1/30, 1/15, 1/8, 1/2, 1/1, 2, 4 ...., in seconds), which controls film exposure times;

4) The film speed, i.e., the exposure levels over which the film responds. The ISO (ASA) rates film properties. High ISO numbers refer to "fast" film (high speed), e.g., ISO 1000, which requires less radiation (hence, a shorter exposure time or a smaller aperture) to achieve a given response. "Slow" film, e.g., ASA 64, requires a longer exposure or a larger aperture, but provides higher resolution. For aerial film, the AFS (Aerial Film Speed) is more commonly used.

One general equation for exposure is:

$$E = \frac{s d^2 t}{4 f^2}$$

where

- E = exposure in Joules (J) mm^{-2}
- s = intrinsic scene brightness, in J mm^{-2} sec^{-1}
- d = diameter of lens opening in mm
- t = time in seconds
- f = lens focal length, mm

(see Ch. 2 in Lillesand & Kiefer, 2000). Changes in any one or combination of these variables brings about variations in photo response characteristics. These differences can be favorable and are actuated by adjusting one or more camera settings.

10-5: Given a camera in which you maintain control of the focal length, exposure time, and F-Stop (the old-fashioned or professional kind, not like those today that have automated the adjustment of these settings), and assuming it has a built-in light meter, enumerate the steps you would take in getting ready to take a picture of a) a nearby person, and b) a distant mountain range, on a sunny day and again near sunset. ANSWER

Black and white film consists of a base or backing, coated by an emulsion composed of gelatin, in which are embedded tiny crystals of silver halides (commonly, Silver Chloride, AgCl) together with

wavelength sensitive dyes. The dyes respond to radiation from segments of the electromagnetic spectrum, such as, ultraviolet, visible,and visible/near IR. Special films respond to photons from shorter or longer wavelengths; for example, X-ray film. When a number of photons strike a halide crystal, they knock loose electrons from some of the silver (Ag) atoms, ionizing them (remember Einstein's photoelectric effect; Introduction). The number of electrons thus activated depends on the brightness (intensity) of the radiation. We can control the part of the spectral range to be recorded by using color filters over the lens. These filters admit radiation from limited segments of the spectrum. This process is a photochemical reaction which conditions the halide grains for later chemical change, forming an intermediate latent image (invisible but ready to appear when we develop it).

Developing begins by immersing the film in an alkaline solution of specific organic chemicals that neutralize the electrons and reduce Ag^+ ions into minute grains of black silver metal. Chemically, today's developers are usually this (quoted from this Wikipedia website: "For black and white photography, the developer is typically a mixture of Metol (monomethyl-p-aminophenol hemisulfate), Phenidone (1-phenyl-3-pyrazolidinone) or Dimezone (4,4-dimethyl-1-phenylpyrazolidin-3-one) and hydroquinone[1] (benzene-1,4-diol). These are made up in aqueous solution with a suitable alkaline agent such as sodium carbonate, borax, or sodium hydroxide to create the appropriately high pH and with sodium sulfite to delay oxidation of the developing agents by atmospheric oxygen." The number of such metallic grains in a given volume determines the film (negative) density. For parts of the emulsion receiving more light, the density (darkness) of the film is greater. In the developing process, the ion conversion must be halted at some point using an acidic stop bath (usually a weak solution of acetic acid). Then any silver halides that remain undeveloped must be removed by chemical fixing. The most common fixing salts used are sodium thiosulfate - commonly called hypo - and ammonium thiosulfate. Volumes in the thin film that saw little exposure (fewer photons) end up with minimal silver grains and thus appear as light and clear in the film negative. The development process is controlled and modified, and hence relative densities in the negative, by changing such variables as solution strengths, developer temperatures, and times in each processing step.

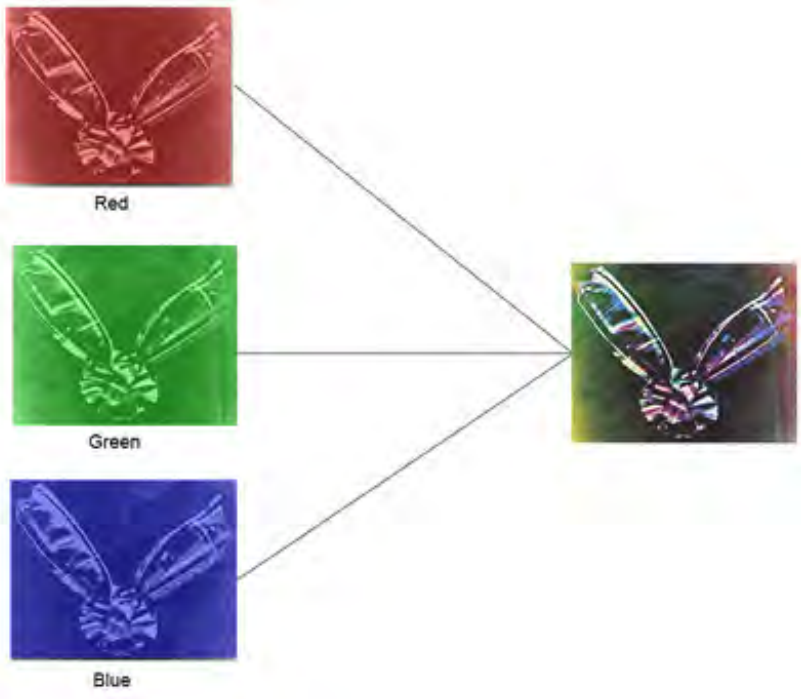
Next, one must use the negative to make a separate, positive, black and white print, in which the negative's dark tones correspond to lighter areas in the scene, and light tones (fewer Ag^+'s) to darker areas. This is known as the printing process. A print (or a positive transparency) consists of an emulsion, backed (in a print) by paper. White light is passed through the negative onto the print material. Clear areas allow ample light to pass and strike the print, which produces high densities of dark (silver-rich) tones. Thus, the initial low levels of photons coming from the target (relative darkness) ultimately produce a print image consisting of many silver grains that make the areas affected dark. Bright target areas in turn, being represented by dark areas in the negative (more Ag^+'s) that prevent light from passing, are expressed as light (whitish to light gray) tones in the print (little silver, so that the whiteness of the paper persists). Once again, we can control the relative levels of gray, or increasing darkness, in the development process by changing the same variables as above, by modifying exposure times, by using print papers with specific radiation responses, and by using filters with different spectral responses (minimizing passage of certain wavelengths) or light transmission. Thus, we can choose different average tonal levels of the print, and, more important, we can adjust the relative levels of gray (tones) to present a pictorial expression, called contrast. Contrast determines whether a scene with variable colors and brightnesses appears flat (low contrast) or presents wide ranges of light-dark areas (high contrast) that aid in discriminating features. Contrast is the ratio of density to the logarithmic value of exposure. We can plot this ratio in the Hurter-Driffield (H-D) curve, which is a straight line with a slope angle for a range of exposures but becomes curved at high and low exposures.

We can expose b & w films under a wavelength-controlling condition that converts them into multispectral images. We do this by using color filters that pass-limited ranges of wavelengths (bandpass filters) during exposure. As we explained in the Introduction, a red filter, for example, passes mainly radiation representing the red region of the visible spectrum. Reddish objects

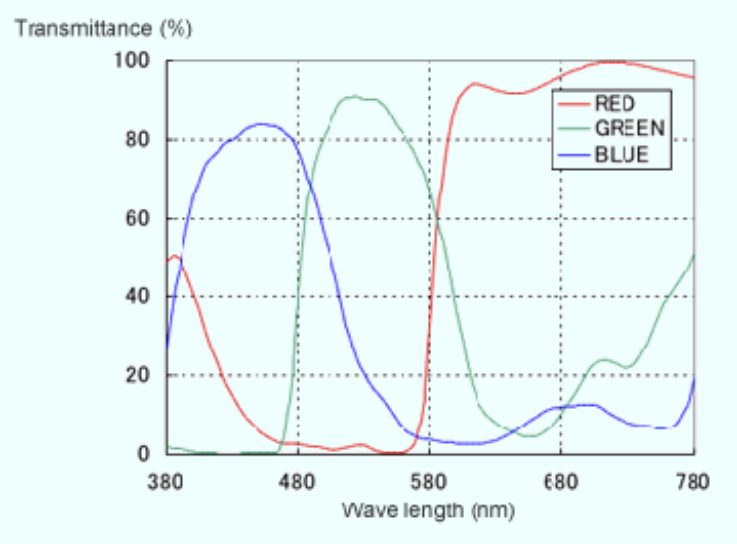
produce high exposures that appear in dark tones on a negative and reappear as light tones in b & w prints or in red on color positive film. We describe why this is so different from the response of b & w film in the following paragraphs. Green appears as dark in a b & w multispectral image representing the red region, and as dark or subdued green in a color multispectral version. We can project multispectral positive transparencies for different color bands using several color filters onto color print paper to produce natural or false color composites, as described in the Introduction.

As an aside, transparencies representing different bands can be combined in a projection system, using filters to determine the colors sought, that result in a color composite. Commercial systems are available, as exemplified by this Color Additive Viewer made by International Imaging Systems, Inc.

Below is an example of the color image produced by using a Viewer to combine three single color images:

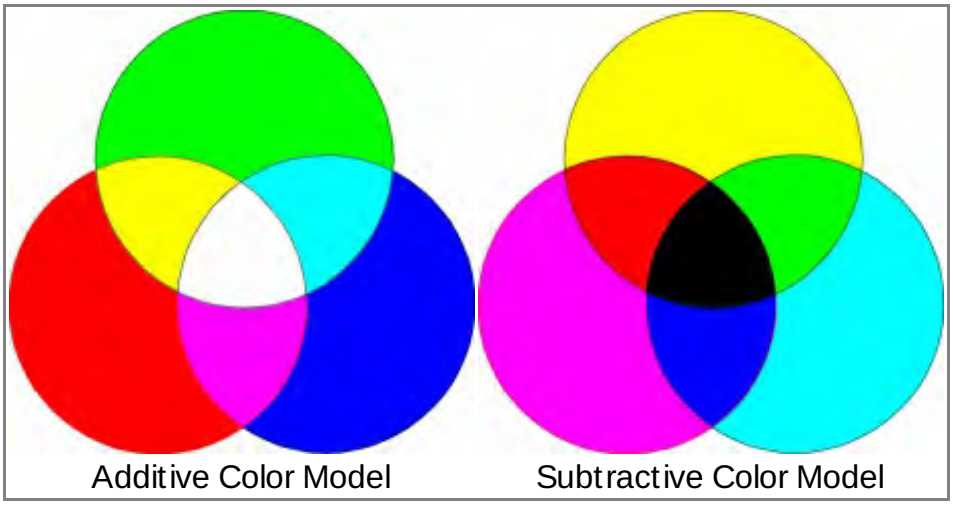


The use of filters to produce individual color photographs is one of two principal ways to make this product (the other uses multiple color-sensitive layers in the film itself). Much as does Landsat and other systems utilize filters on the sensors to subdivide the light received into wavelength intervals (the bands), so will a multiple camera array have filters of different bandpass intervals over the different lens involved. Here is a plot that shows the spectral properties of such filters:



How we use color film to produce color images involves some different concepts, although many of the same factors and mechanisms are still valid. Starting with the three additive primary colors, red, green, and blue, or the subtractive primary colors, yellow, cyan and magenta, we can make other colors by using the principles of either the color addition or the color subtraction process. Look at these diagrams:

Color Models

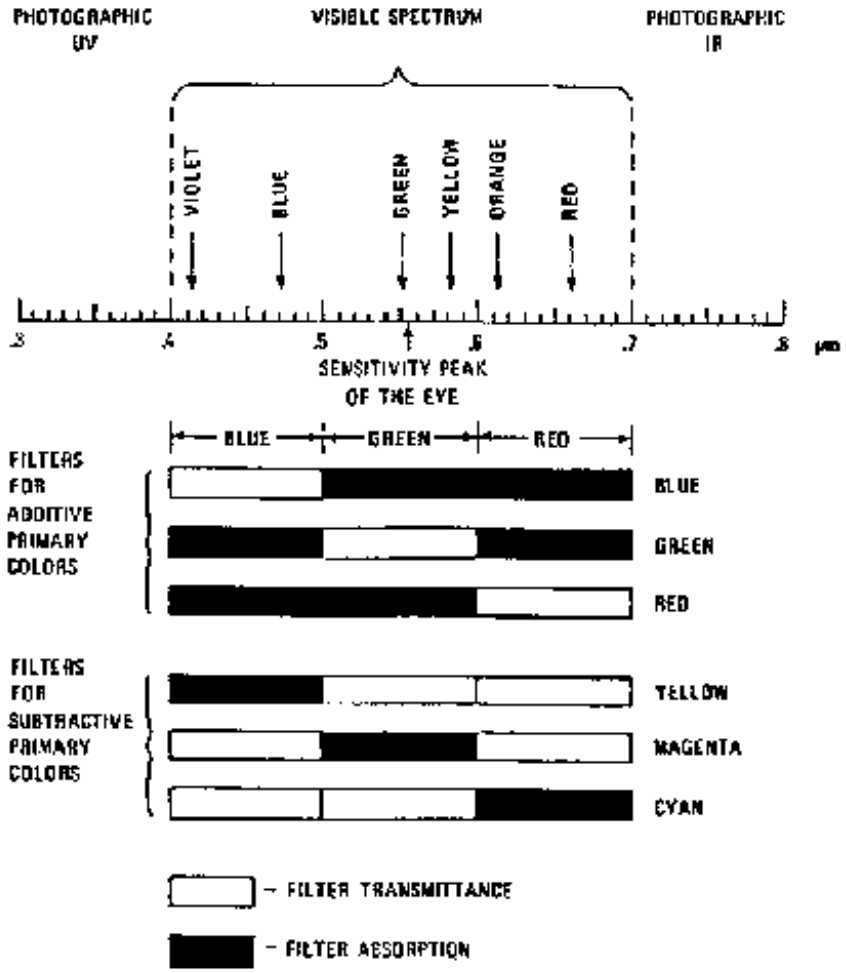


Color addition works when we superimpose the primary colors on one another. For example, if we shine a green light and a red light on the same spot on a white wall, we will see some shade of orange or yellow, depending on the relative intensity of the red and green illumination. If we add a third blue light to the spot, we will see white or a shade of gray. Computer displays work this way. Of course, the absence of any amount of all three primary colors leads to black. To create a color shade (i.e., some deviation from an arbitrary primary color value, such as pink or purple or yellow) in the photographic printing process a mix of the three primaries in some non-equal proportion is the procedure. (In a computer-generated color image, we can typically choose a

number between 0 and 255 to indicate how much of each of the three primary colors we want. If our display board has sufficient memory, we will have 2553 (16,581,375) colors to choose from.)

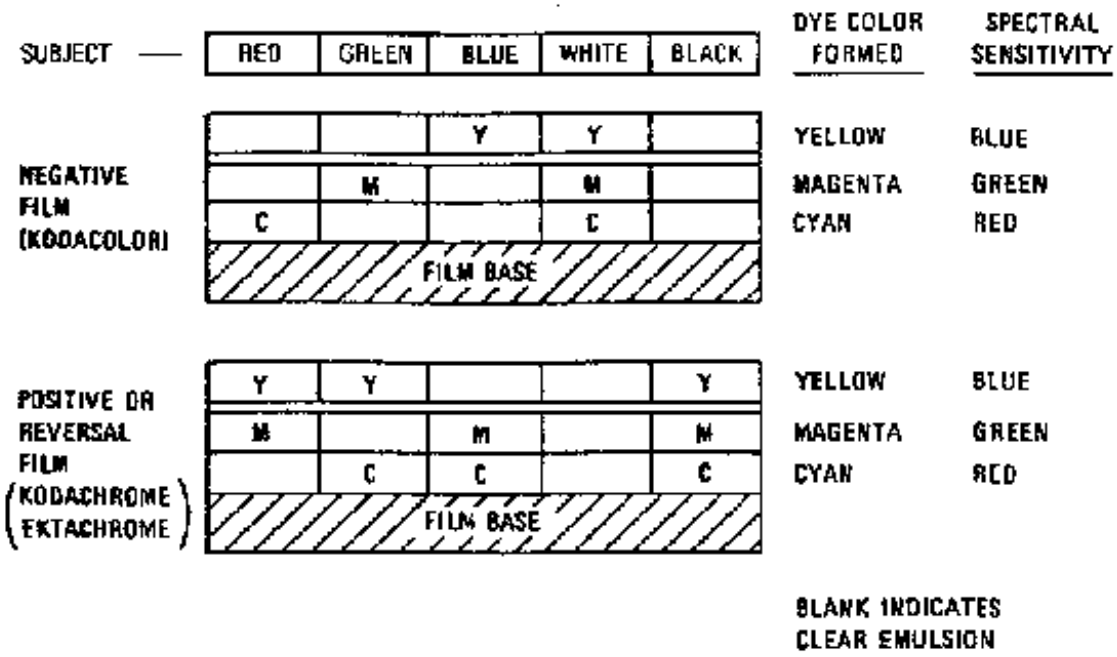
In subtractive color, we use filters to remove colors. For example, a yellow filter removes colors other than yellow, as do cyan (bluish) and magenta (reddish) filters. If one superimposes all three filters, little or no visible light gets through, so either black or dark gray results. By combining pairs of the subtractive primary colors, we can create each of the additive primary colors. Magenta and yellow produce red. What corresponds to mixing cyan and magenta; yellow and cyan?

The principles of color subtraction apply to color-sensitized film. This film consists of emulsion layers containing silver chloride treated with light sensitive dyes, each responding to a limited wavelength range. These layers act as subtractive filters during development. Thus each layer of the film responds to different sections of the scene's spectrum. These layers are stacked, respectively, as follows: a blue-sensitive layer on the top, then a yellow filter layer (to screen out ultraviolet and blue from passing into the next layers; omitted from the diagrams below), and finally, green- and red-sensitive layers.



From F.F. Sabins, Jr., Remote Sensing: Principles and Interpretation. 2nd Ed., © 1987. Reproduced by permission of W.H. Freeman & Co., New York City.

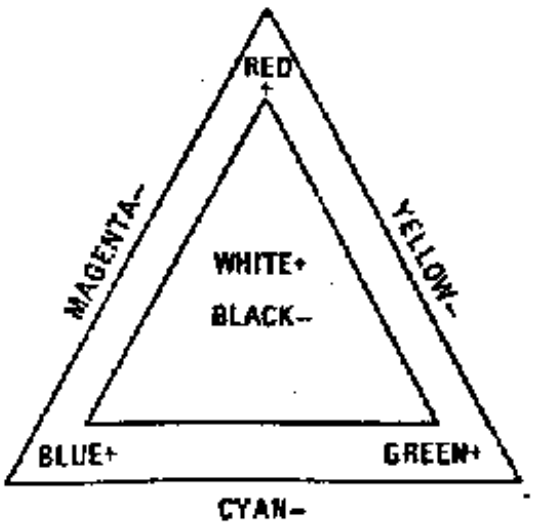
Referring to the above diagram, when a shade of red passes through a color layer sensitized to cyan (a blue-green, the complementary color to red; the sum of any primary color and its opposing complement always equals white), its absorption activates the dye/silver grains in that layer to produce, in a negative, cyan tones in areas associated spatially with reddish objects in the scene. In color film, the three subtractive color layers stack together (a fourth serves a special purpose, described below) on top of a clear base. To guide you in reasoning through production of other colors, check this schematic diagram:



From F.F. Sabins, Jr., Remote Sensing: Principles and Interpretation. 2nd Ed., © 1987. Reproduced by permission of W.H. Freeman & Co., New York City.

Thus, in a similar manner, light from a blue subject reacts with the yellow layer to produce a yellow shade (red and green make this complementary color) for its area on the negative.

10-6: To test your understanding, from the above diagram, you set up the response for green objects (magenta, a bluish-red, is a mix of red and blue). Also look at the diagram just below. No doubt you can see an obvious rule working here: consider, then state it. **ANSWER**



Additive (+) and subtractive (−) color triangle.

From F.F. Sabins, Jr., Remote Sensing: Principles and Interpretation. 2nd Ed., © 1987. Reproduced by permission of W.H. Freeman & Co., New York City.

As evident in the diagram, each primary color activates the layer containing the subtractive color opposite it. Several other rules or observations apply:

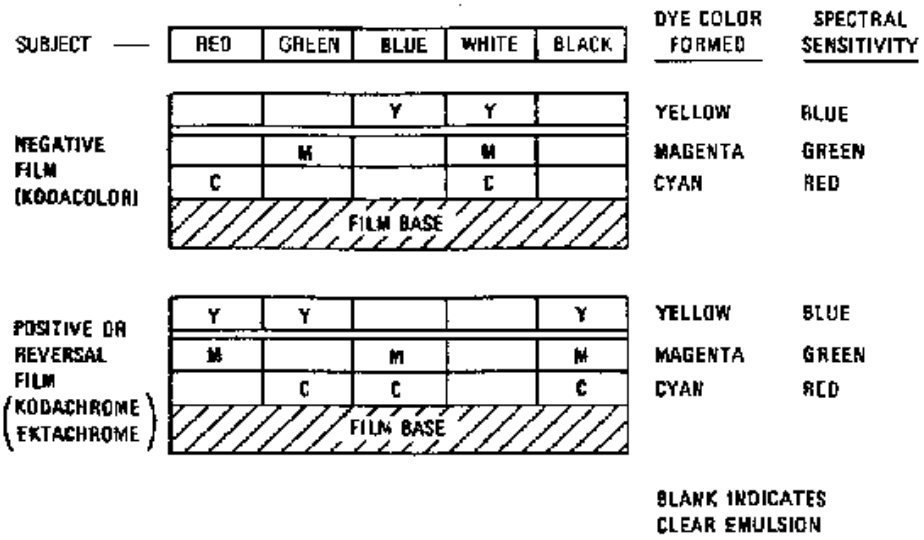
1) A given primary color does not directly activate the other two film layers.

2) Note that yellow + magenta = red. The red is common to each of these subtractive colors, with blue and green being filtered out. The same rationale applies to the other two combinations of subtractive colors.

3) White light exposes all three subtractive layers in the negative. The sum of these three layers (the center of the color diagram on the right) on a positive is black. Conversely, black (absence of light) objects produce a clear (not colored) area in the three layers of film

4) We must insert a fourth, special yellow filter layer below the yellow layer, because the dyes in the red and green sensitive layers below are also sensitive to blue, which this filter layer screens out and then dissolves away during developing.

To comprehend how to make a color print, follow this exposition, which is an exact quote from F. F. Sabins, op. cit., 2nd Ed., pp, 44-46; we reproduce one of the above figures to minimize your need to scroll up the screen:



In the negative film, "the silver halide salts of the green-sensitive and red-sensitive layers are also sensitive to blue light. A yellow filter layer beneath the upper emulsion layer prevents blue light from exposing the green-sensitive and red-sensitive layers. The yellow filter layer is dissolve are removed during film processing. On negative film the red-sensitive, bottom layer produces a complementary cyan image of a red subject. Green and blue subjects produce magenta and yellow images respectively. The white subject exposes all three layers, resulting in an image that transmits no light; the black subject results in a clear image because none of the layers is exposed. The negative film records the color of a subject as its complementary color. The image on a negative color film is projected onto photographic paper coated with sensitive emulsions that is developed to produce a color print. Positive color film records a subject in its true color, not its complementary color. Looking at the positive film part in the above diagram, a red subject forms a clear image on the red sensitive cyan-colored layer. The red subject also forms a magenta and a yellow image respectively on the green-sensitive and blue-sensitive layers. When viewed with transmitted white light, the yellow and magenta images absorb blue and green respectively and allow a red image to be projected. A white subject forms clear images on all three layers." (Similar reasoning applies to other color combinations - NMS)

Other systems of color production have been devised. One mentioned briefly here is the IHS system, in which:

- I = the color intensity or brightness,
- H = the hue, comprised of a dominant wavelength, averaged from a limited range of adjacent wavelengths
- S = saturation, the purity of the color relative to gray.

This system is sensitive to controllable transformations (computer-aided) that optimize and enhance color representations.

Conventional photography has changed significantly in the last two decades. This is largely due to the now common use of digital cameras (see bottom of Intro page 5a for description of how such cameras work). The image created can be completely devoid of the need for film or print, as it may be displayed solely on one's computer. But, computer printer pictures are one product or the digitally stored images can be printed out as actual photographs.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The concept of scale is spelled out and the factors that determine a given scale - including camera focal length and height above the ground - are considered, particularly in terms of formulae for calculating the scale when these parameters are inserted. The ways in which spatial resolution can be determined are discussed. The point is made also that large scale imagery, especially, experiences displacements and distortions increasing from center outwards.

Some Elements of Photogrammetry

Let's move now from the spectral part of photogrammetry to the spatial part. Scale, mentioned before, is just the comparison of the dimensions of an object or feature in a photo or map to its actual dimensions in the target. We state scale in several ways, such as, "six inches to the mile", "1/30,000", and, most commonly "1:2,000". These mean that one measurement unit in the numerator (photo or map) is equivalent to the stated number of that unit in the denominator (scene). Thus, 1:2,000 simply states that one of any length unit, such as an inch, in the photo corresponds to 2,000 inches on the ground or air (cloud). Or, one cm is equivalent to 20,000 cm. "Six inches to the mile" translates to: six inches in the photo represents 63,360 (5,280 ft x 12 in/ft) inches in the real world, but we can further reduce it to 1:10,560, because six and 63,360 are divisible by six. Note that if we enlarge or contract a photo of a given scale, say by projection as a transparency onto a screen, then one inch on the screen no longer corresponds to the same denominator number but now represents some other scale determined by the magnification factor. However, the effective resolution, the area covered, and the relative details remain the same.

We determine the scale of the aerial photo, expressed as its Representative Fraction (RF), by the height of the moving platform and by the focal length of the camera, according to this equation: **RF = f/H***, where H* = H - h, with H = height (elevation with reference to sea level) of the camera and h is the height of a reference point on the surface, so that H - h is the distance between the platform and the point (assuming a flat ground surface; in rugged terrain, scale in effect varies with the elevations). We can also show that RF is also proportional to resolution and distance ratios, as given by $\mathbf{RF = r_g/r_s = d/D}$, where r_g is ground resolution (in line pairs per meter; see below) and r_s is the sensor system resolution (in line pairs per millimeter); d is the distance between two points in the photo and D is the actual distance between these points on the ground (the definition of scale).

Several more questions will help to master these ideas.

10-7: A map has a scale of 9.0 inches to the mile. What is the denominator of the RF? ANSWER

10-8: Two points appear on a map 1.75 inches apart. The actual horizontal ground distance between the two points is 1108.0 meters. What is the denominator of the scale fraction (RF) for this map? ANSWER

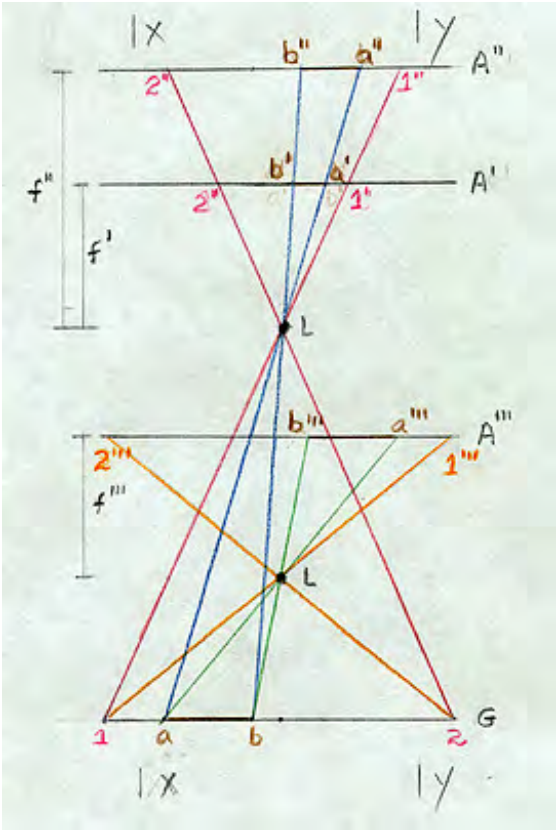
10-9: Points A and B are 2.2 inches apart on a map having a RF = 1/20000. They are 6.83 inches apart on an airphoto. What is the scale of the airphoto? ANSWER

10-10: A vertical airphoto is taken from a flying height of 5000 ft relative to the ocean with a camera having a focal length of 6 inches. The flat surface is 1000 ft above

sealevel. What is the RF for the resulting photo? ANSWER

10-11: An aerial camera has a 9 1/2 inch square film format and a 6 inch focal length lens. What must the flying height (in meters) be to obtain a scale (RF) of 1/2000? ANSWER

We can elucidate the roles of f and H* further with the aid of the next diagram, which, although not strictly correct in terms of optics and simplified to two dimensions, does allow us to visualize the effects of changing focal length and platform height:



Lines such as 1-1" or a-a' are light rays passing through the lens L. G is the ground. A' is at the focal plane (holding the film) for a focal length of f', and A" is the shift of this plane to a new value of f ". A"' is the location of the focal plane for a case in which the lens, L, is now at a lower elevation. A line on the ground, a-b, passing through the upper lens, L, is focused on plane A', such that, it has a film dimension of b'-a' (note that a - b is reversed in position but this does not matter because we can turn over a transparent negative).

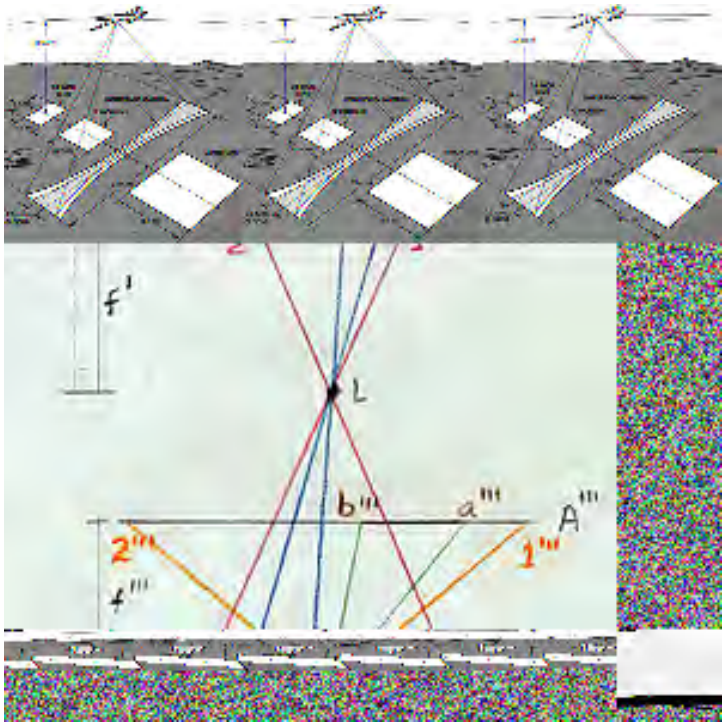
When we lengthen the focal length to f" to bring the focus onto A", b'-a' expands to b"-a". Look next at what happens when we lower the camera (and airplane) to the A"' position: a-b in this new arrangement (where L, the lens location relative to the film is the same distance as case 1, so that the focus, or focal length, is once more f ', i.e., f"' = f'), now is expressed by b"'-a"', which for these conditions is even longer than b"-a".

In these situations the frame size of the film (x-y in the two dimensional simplification) remains the same for A', A", and A"'. Therefore, when a and b (ground points) are at the A" location, the fraction of the scene imaged occupied by them decreases, that is, they occupy a smaller segment of it. In the A"' case, the size of film needed to display all of 1-2 would have to be greater (would need a larger frame), since the x-y would have enclosed less of the scene. Both 1' and 2' for the A' case and 1" and 2" for the A" case fall within the original x-y frame dimensions. Keep in mind that the dimensions shown on the line G are ground-sized, whereas those in A', A", and A"' are film-sized, reduced relative to ground distances by the scales of the photographs.

We summarize these relations in a mnemonic:

Long is large/low is large/and larger is smaller.

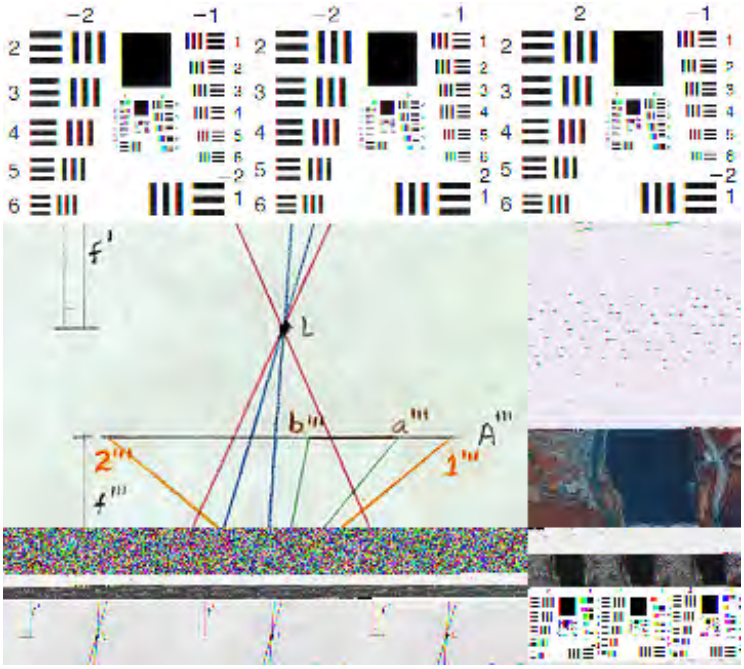
Interpret this as follows: the **scale is larger** (denominator becomes smaller) as we **lengthen the focal length** or as we **lower the platform.** A large(r) scale image covers a small(er) ground area (but has increased resolution). To appreciate how scale affects scene content, you may return to the various photos that we brought on-line in the previous page. The scale of each is printed alongside it. To further understand the role of changing the focal length, consider this figure:



In this figure, in which the airplane stays at the same altitude, the longer focal length camera would produce a smaller image but it would have a larger scale. Progressively decreasing the focal length yields ever larger image sizes but at ever smaller scales (an inch covers less and less ground "real estate"). A panoramic camera has variable scales with the largest straight down (nadir) and decreasing scale outward.

Resolution has a popular meaning but is best defined in a technical sense. We normally think of resolution as the ability to separate and distinguish adjacent objects or items in a scene, be it in a photo or real life. We specify the resolution in terms of the smallest features we can discriminate. As we look outward, intuition (or experience) tells us that we can see small objects close to us but if the same objects are far away they probably can't be spotted (i.e., are unresolved). But, contrast influences resolution. If two items are the same color, they may be hard to separate, but if they are sharply different in color, tone, or brightness, we can identify them more easily. Shape also is a factor.

So, a rigorous definition of resolution, as measured quantitatively, relies on the ability to separate adjacent alternating black and white thin lines in a target. The resolution of a film is determined in a laboratory by photographing placard-sized charts like the one shown below containing black lines with different spacings on a white background (or the reverse) *****. Then, the resolution is the smallest spacing, in which we can discriminate pairs. This, of course, varies with distance between camera and target.



The separation between light-dark lines varies in this target. The spatial frequency of these alternating lines can be given in lines pairs/unit distance (width). The sharp light-dark pattern can be approximated by sinusoidal waves of specified frequencies; a common measure is given in cycles/millimeter. Depending on relative brightnesses between the lines in a pair (or contrasting objects in a real scene), the brightness amplitude between peak and trough of the waves will vary or modulate. A technically more exacting measure of resolution is given by the Modulation Transfer Function (MTF) - this plots modulation (ratio of target amplitudes to sensor [film; electronic] amplitudes) against spatial frequency. For the human eye as a sensor, its ability to separate colors is optimal at low frequencies (wider spacings) and falls off rapidly at frequencies greater than 2 cycles per retinal degree (spacings closer than about 0.2 mm). Black and white pairs are optimally separated at 7 cycles per degree (about 0.05 mm) and fall off notably at both higher and lower frequencies. As one would expect, film MTFs are near maximum at low frequencies and drop progressively at higher frequencies.

(Note for the record: In real scenes, light reaching the film or sensor will consist of bundles of varying wavelengths (spectrally diverse) that act as a collection of different sine waves. To specify their influence on the recording medium, such as film, they can be subject to a procedure called *Fourier Analysis* which breaks them systematically into the actual wavelengths and amplitudes present. This is often applied when the spatial domain (variations of tonal patterns) of an image is being studied in scene interpretation. See pages 506-507 of Lillesand and Kiefer, Remote Sensing and Image Interpretaton, 5th Ed., J. Wiley, 2000 for a brief overview of this subject.)

In practice, we could place a resolution target in a scene (for example, painting black lines with different spacing on a concrete airport runway or road) to determine resolution for aerial conditions. Ground resolution is then the number of black/white line pairs within some width (normally one meter) that we can just discern in aerial photos taken at a particular height. Depending on the camera, film-resolving power, and platform height (the system), in the photo, the pair will either blend visually (not resolvable) or can be distinguished. We express system resolution, $\mathbf{r_s}$, (in which we combine the effects of sensor and film factors) in line pairs/mm within the print. A formula for ground resolution, $\mathbf{r_g}$, (in line pairs/meter), applicable to just separable ground lines, is:$\mathbf{r_g = f \times r_s / H}$. A typical example is a case where the lens focal length is 150 mm, the system resolution is 60 line pairs/mm, and the height is 3,000 meters, so that R_g is 3 line pairs/meter. From the relation 1 line pair/r_g = width on ground of 1 line pair, this width is 0.33 m (each line is half that value). This means that the airborne camera can resolve an object on the ground that has a dimension of 0.165 m (about 6.5 inches), if it contrasts with its surroundings, using a film of appropriate resolving power. If the aircraft were flying higher, the camera could not

detect this sized object.

10-12: Given a camera with the focal length of 9 inches, capable of a photo resolution of 15 line-pairs per millimeter, when flown at an altitude of 12000 ft, what is the equivalent resolution on the ground? ANSWER

10-13: What is the photo (system) resolution obtained when a camera with focal length of 120 mm is flown at an altitude of 6000 meters; the line-pair in a calibration target on the ground has a spacing of 4 line-pairs/m. ANSWER

Resolution in film (negatives and prints) is governed, in part, by the size distribution of the silver grains. In Landsat's Multispectral Scanner/Thematic Mapper (MSS/TM), and other electronic sensors, image resolution ties closely to the size of the pixels or to the dimensions of individual detectors in the arrays of Charge-Coupled Detectors (CCDs), such as on SPOT. At first thought, it would seem that we cannot resolve objects smaller than the ground dimensions represented in an individual pixel/detector. However, if the spectral characteristics of a subresolution spot on the ground are sufficiently different from surrounding areas, they can affect the average brightness of the pixel so that the spot is visible in the image. An example of this are roads that are narrower than a 30 m (98 ft) TM pixel, yet are quite visible in a TM image, if they are a bright and reflecting surface set within dark vegetation (such as dirt roads midst sage in Wyoming).

In an aerial photo, when we view features at ground points off the principal point (optical center, usually at nadir or normal to a flat surface), that is, along slant directions, they may appear to lean away from the center, especially if they are tall (e.g., buildings) or have high relief. This distortion is worse if the aircraft flies low to acquire large scale photos. This is one type of displacement, and is evident near the edges in the 1:4,000 aerial photo of a neighborhood in Harrisburg, shown on page 10-1. A better example is the top image in Section 11, page 11-4. There are other modes of displacement, such as apparent lateral movements of image points along slopes of differing angles.

A big plus for aerial photography: Mission controllers can fly aerial photo missions at any time during the day, but these usually occur between about 10:00 AM and 2:00 PM (in summer to avoid afternoon storms). Typically, the aircraft traverses the region to be photographed along back-and-forth flight lines and acquires pictures at intervals that allow about 50% overlap between successive photos and 20% to 50% sidelap between lines. The camera usually mounts below the plane, near its nose. Film in the camera advances automatically at time intervals that are synchronized with the speed of the aircraft. Especially in color photos, but also in black and white photos, blue and ultraviolet light that is scattered by the atmosphere may degrade the film image. We can reduce this degradation by using a haze filter that absorbs the ultraviolet and the very shortest visible blue wavelengths.

NASA has a stable of support aircraft that operate various sensors, including cameras, to gather ground reference data for remote sensing experiments (see page 13-4 which discusses this). An example of a small-scale image (about 1:150,000) obtained during a U-2 flight, which operated at an altitude of about 18,000 m (59,000 ft) over Utah (resolution about 5 meters), closes this section on aerial photography.

For anyone interested in viewing more aerial-type photos, including perhaps one of your home region, consult the on-line Net Home Page called TerraServer, sponsored by Microsoft. Aerial imagery, either individual photos or sections of orthophotoquads, of selected large parts of the United States, has been digitized from data collected by the U.S. Geological Survey. Photo resolution ranges from less than 2 to about 12 meters. Imagery taken with the KVR-1000 camera (resolution: 2 meters) flown on several Russian satellites shows regions in the rest of the world, mainly in Europe. They market these data worldwide as part of their SPIN-2 program. All photos are black and white.

Of course, the principles we have been applying in these three pages to aerial photos pertain in many respects to space imagery as well. Mapping once done primarily with air photos can now be done almost as well with space products. The main drawback is resolution, and that limitation is rapidly disappearing with the declassification of high resolution military imagery and with the ever-improving resolution capability of satellites now being flown or on the drawing boards for the foreseeable future. As we shall see in the next Section, the ability to gather data from space that pertain to three-dimensional surface variations allows the space image mapping community to essentially duplicate all the advantages once exclusive to air-based photography.

***** As as aside, the writer [NMS] played a small but helpful role in the process by which the United States government decided to lift the ban on high resolution sensors that had been imposed because of militay considerations. Around 1987, the Soviet space program - strapped for hard cash - began selling in the open world market selected photos obtained by their cosmonauts and later by sensors on unmanned military satellites. A NASA official visiting Moscow was given one such product - a false color (but with no green band) KR-1000 photo that covered part of the Oregon northwest coast extending past Portland, OR. The Russians claimed at the time that this now-commercial product had better than 2 meter resolution. Concerned NASA Headquarter people wanted that claim to be tested/verified. I was given the photo and told to "solve the problem". But there was no travel money to check the site itself. So I contacted a colleague at Oregon State University, Dr. Charles Rosenfeld, Geography Dept., and later the Commanding General of the Oregon Air National Guard] who taught a remote sensing course.

He came up with a neat solution: he took his class to the airport outside Astoria, OR. On the blacktop runway, a series of white lines 6 feet apart had been painted perpendicular to the runway length as a marker guide for landing. These were just visible with a magnifying glass on the Russian photo. This proved the veracity of their 2 meter claim.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The most ambitious effort to date to map much of the Earth's land elevations from space using radar interferometry was embodied in the Shuttle Radar Topography Mission (SRTM), flown as the main scientific experiment on STS-99 in February of 2000. Two NASA-JPL radar units, an L-band and a C-band, were mounted on a boom that extended 60 meters into space from the Shuttle cargo bay. The signals received are combined and analyzed using the interferometric method. The SRTM data have a 30 by 30 meter ground resolution and a vertical resolution that varies from 16 to 10 meters, depending on conditions. On the same mission, an X-band radar unit provided by the German DLR also gathered data that could be integrated with either or both JPL radars to add to the inputs needed to generate topographic profiles and maps. During the 10 day Shuttle mission, almost 80% of the land surface was covered, providing a near global topographic data base superior to any developed previously.

The Shuttle Radar Topography Mission

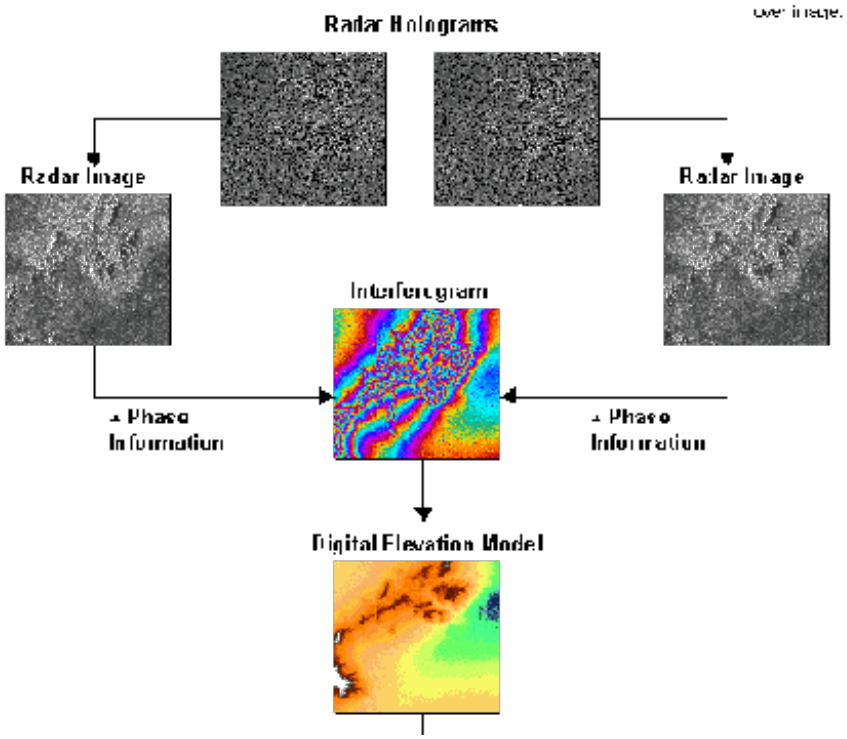
A precise overview of the SRTM program is found at the JPL video site. Access through the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes "Mapping the World", July 11, 2002. To start it, once found, click on the blue RealVideo link.

It may come as a surprise to most who've never worked on or utilized topographic data sets or maps to learn that much of the global land surface (as well as parts of ocean floors) has not been mapped at other than coarse (large value) contour intervals. Better maps at closer intervals are very much needed in many enterprises, both civilian and military. It should predictable, then, that topographic mapping from satellites can overcome this deficiency since polar orbiters and Shuttle missions pass over much of the land masses of the World. Experience with radar mapping from the Shuttle led geographers, cartographers, and other specialists to advocate a mission that could systematically map large regional terrains in need of more detailed elevation data for the land. Such a mission, covering both well- and poorly-mapped surfaces, would provide a uniform and coherent data set that would serve as a database suited to many applications.

The response from the U.S. community, through NASA and joined by a German organization, is the Shuttle Radar Topography Mission, which was successfully launched (Mission STS-99, Shuttle Endeavor) on February 11, 2000 and operated for 10 days, gathering data that covers approximately 80% of the land surface. It will take about 2 years to fully reduce all data and provide topographic maps in various formats. C-band and X-band radars (see page 8-1 for wavelengths) operate from one pair of transmission-receiving antennas in the Shuttle payload bay and a matching second pair (receiving only) at the end of a collapsible boom of 60 m length (actual distance between inboard and outboard antennas is 83 m), so that dual signal return simulates the separation needed to provide parallax-like data. Here is a view of the boom (right) center as it was extended during the mission.

NASA JPL supplied the C-band instrument pair. The data from 159 orbits (at angles to the equator that allow monitoring the Earth between 65° N and 50° S), at a Shuttle altitude nominally at 233 km (145 miles), extend for each pass over a swath width of 225 km (141 miles), so that continuous operation for 10 days provided nearly total coverage (with some duplication) of the land traversed. Under these conditions, the radar image has a ground resolution of 30 x 30 m and a vertical (elevation) resolution of 16 m (absolute) and 10 m (relative).

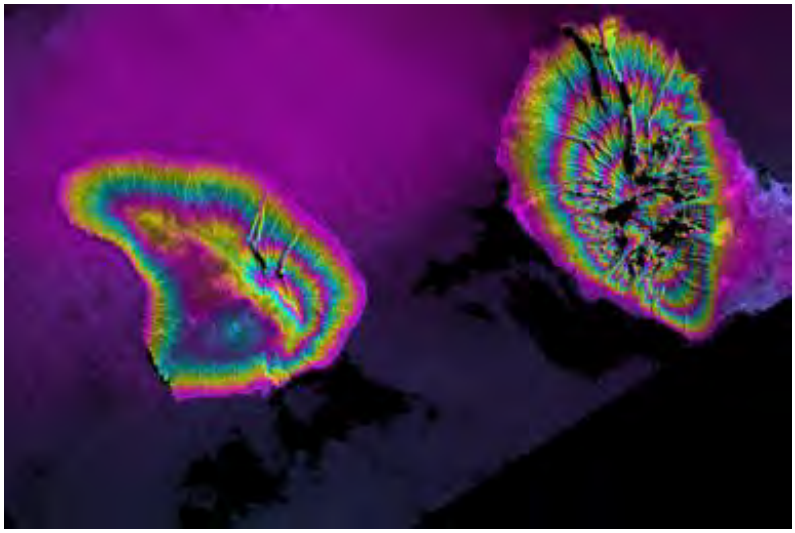
Processing the data into a stereo-mode which, through the principle of interferometry, can be converted into elevation differences and then restructured as a DEM data set by incorporating bench marks (known elevation points), is summarized in this diagram (somewhat degraded by low Internet resolution):



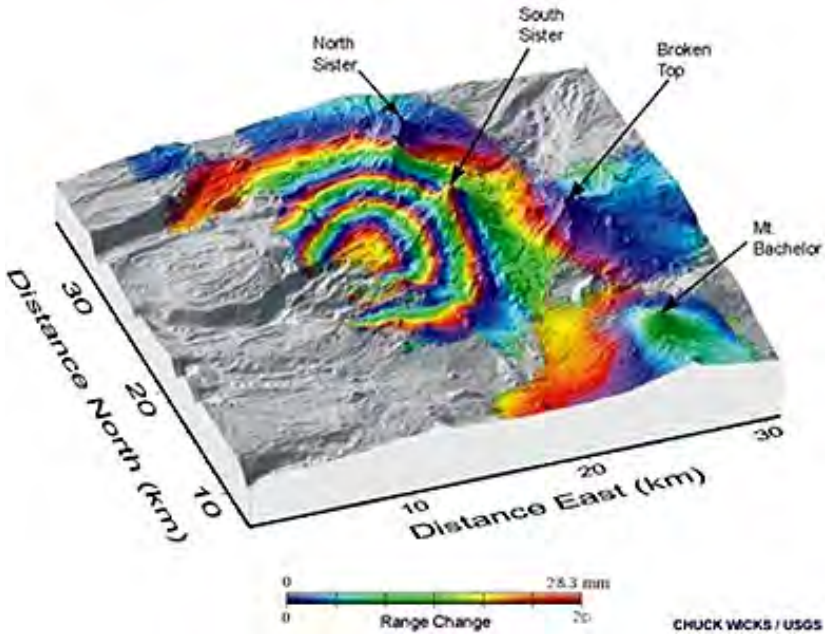
In essence, radar holograms from both antennas are generated (and can be converted to images). Then, while still in digital mode, these utilize phase information to construct interferograms (again, displayable as images). After introducing known elevation points to the data set, a digital height model (DEM type) can then be used to contruct a relief or shaded relief image. That can be combined with other data types (e.g., Landsat imagery) or converted to other map forms.

Let us look at a typical C-band image, of an area covering part of Dallas, Texas:

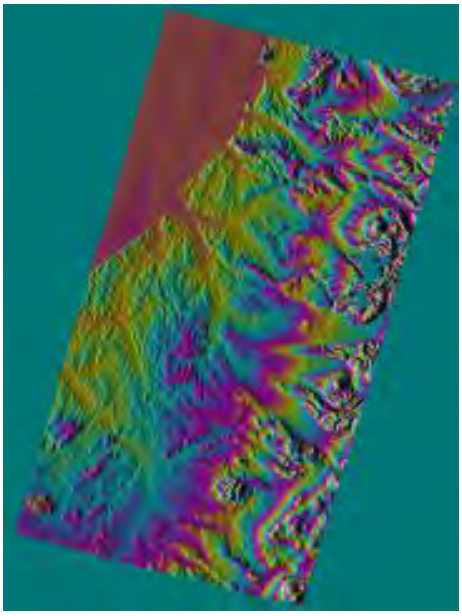
Next, examine this interferogram that shows the distinctive color banding such a data type presents. The two islands are West Maui and Lanai in the Hawaiian chain. Each cycle of colors (from pink to blue to pink) passes through an equal amount of elevation difference (400 m [1300]) in a manner similar to broad contour lines.



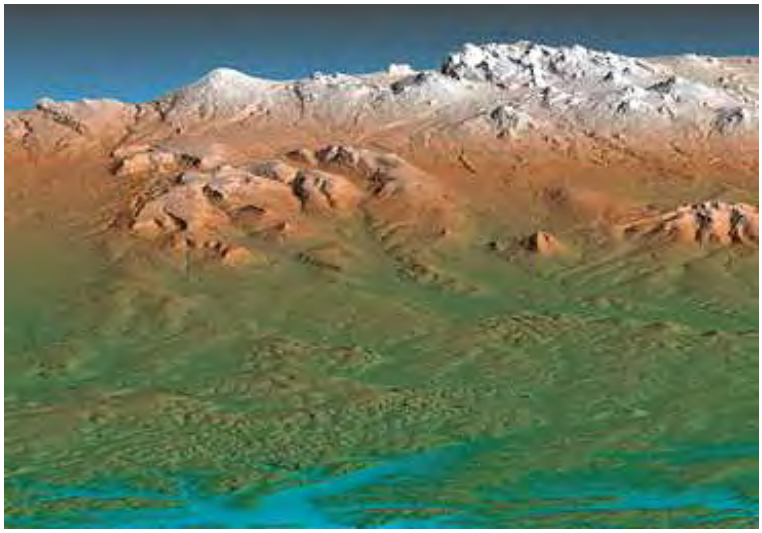
One of the practical uses of SRTM data resides in the monitoring of changes of elevation over time. This can be important as a tool for observing the swelling of a volcano with newly emplaced lavas. This causes the volcanic edifice to undergo upward swelling of the surface. This next illustration shows changes in the surface in and around the Three Sisters, volcanic stratocones in the Oregon Cascades near Bend, OR. To construct the change plot, data from several collection times over an extended period are used.



To illustrate some of the different image types derivable from SRTM data, we will display coverage of a pass across the Kamchatka Peninsula, a volcanic mountain chain in eastern Siberia. The first view is the interferogram whose color fringes suggest relief.

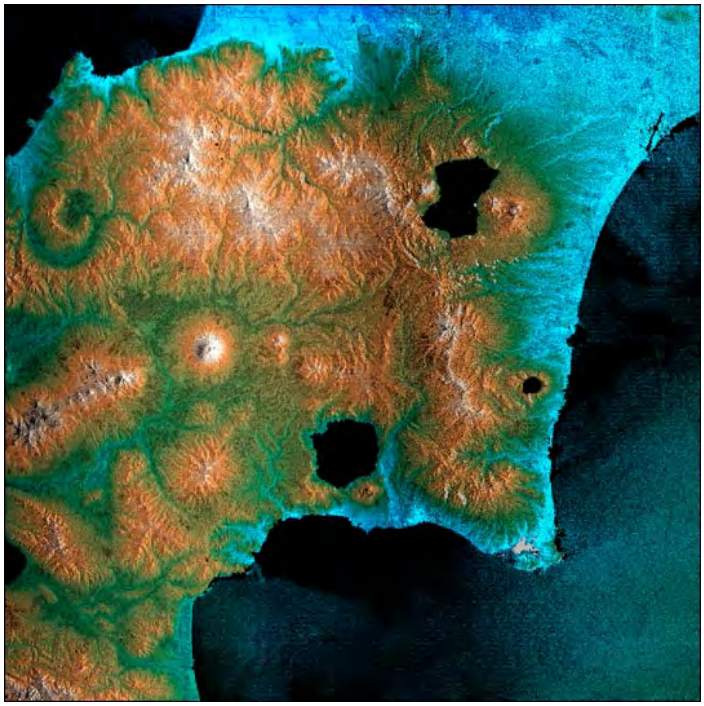


From that, a shaded relief (computer-generated artificial light imposes a pattern of shadow and light) and a shaded relief perspective view are derived:

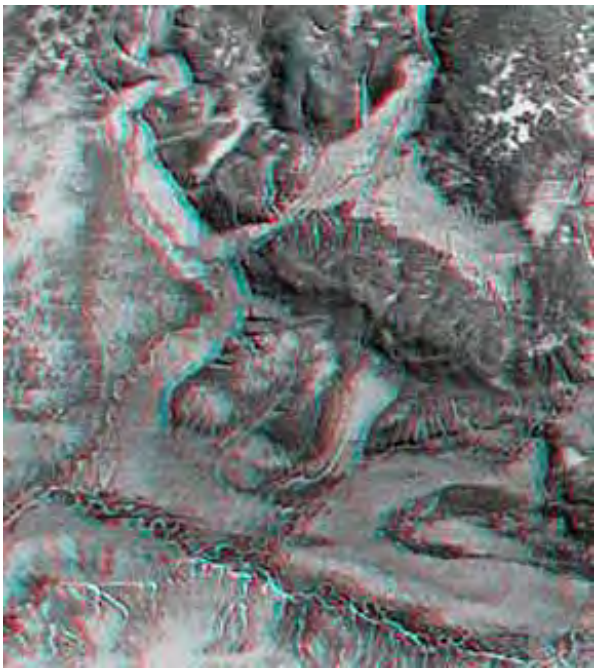


When SRTM data are combined with Landsat, these views of the Kamchatka mountains ensue:

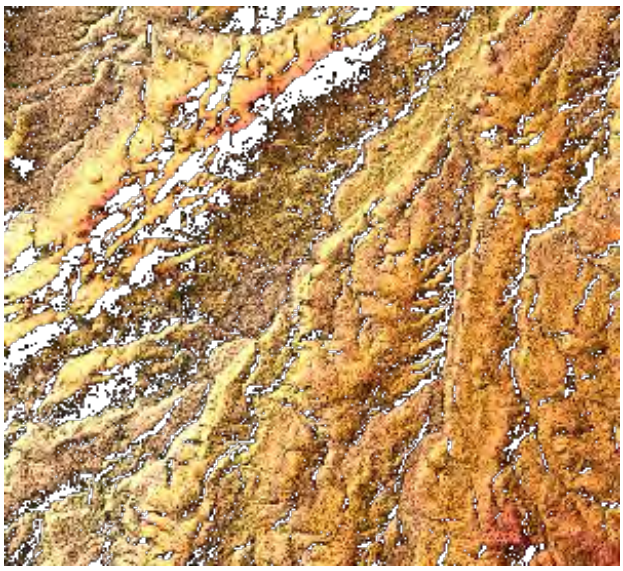
Still another SRTM example from the same part of the World is Hokkaido Island in Japan, with its active volcano, Usu.



A special type of stereo image known as an *analglyph* is made by projecting one of the image pair through red and the other through blue filters to give the resulting superimposed image. To see relief in stereo, for this segment of the Kamchatka peninsula, you would need a pair of glasses made usually of cardboard with the right eye opening covered by red cellophane and the left with blue. A few users of this Tutorial may have one from a 3-D movie experience or can make the glasses in the obvious way.

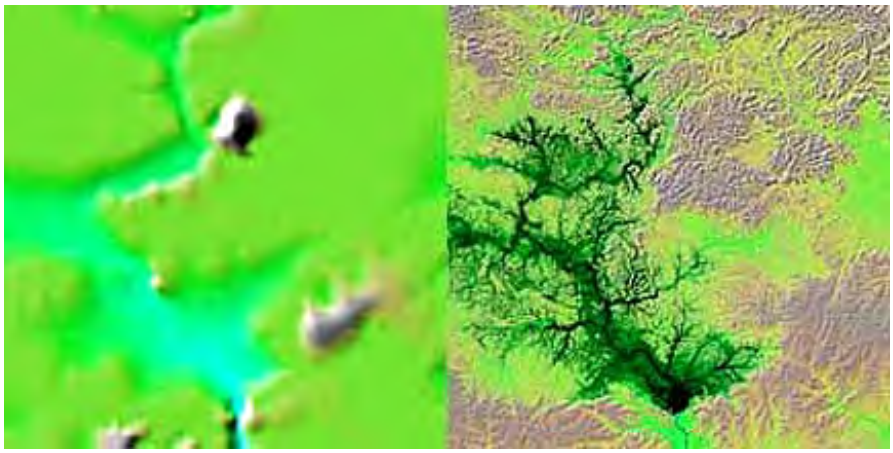


The X-band radar on SRTM is provided and managed by DLR - the Deutches Fernerkundungs-datenzentrum (the German Aerospace Center). Its antenna setup is similar to JPL's C-band system, the difference being that its antenna in the Shuttle Bay is fixed to look straight down, rather than inclined. Thus its swath width is an unchanging 50 km. Data processing to produce elevation maps is essentially the same. Here is an X-band image of part of the Kamchatka Peninsula; the highest point in the mountains is 4755 m (15690 ft):

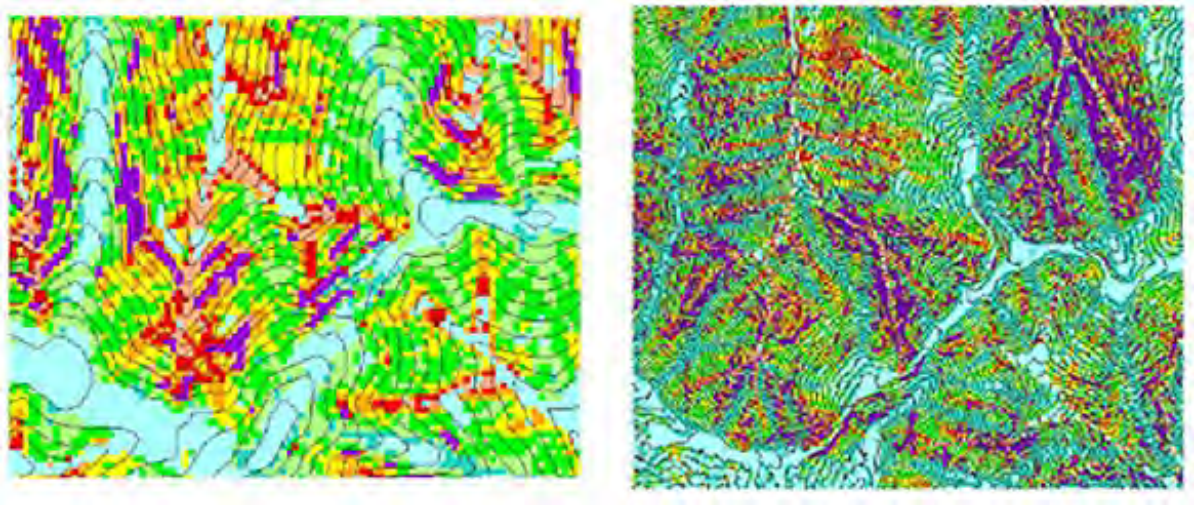


For those interested in additional information and updates and more imagery, tie into the JPL and German Aerospace SRTM Web sites.

How much of an improvement in small-scale map quality in those parts of Earth where topographic maps are lacking in details found in, say, coverage of the United States? This next image is convincing: On the left is a colorized map of the Uatuma River Basin near Manaus in eastern Brazil based on the best elevation data available; on the right is a SRTM image processed to display the optimum topographic representation extractable from the radar data. The black areas in the map on the right depict the water buildup in Lake Balbina created by damming the river.



Another example is this pair of maps of Coos Bay in Oregon. The left map is taken from the USGS's 10 meter series covering the U.S.; the right map shows the additional contour lines, at a 5 meter interval, that could be drawn with SRTM data. The coloring in both is an assessment of the potential for landslides:

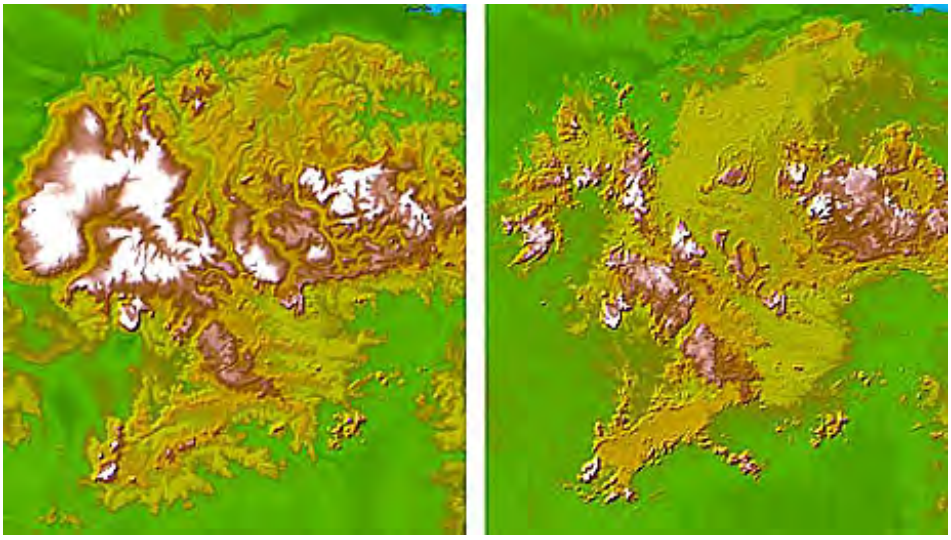


By mid-2003, the SRTM task force had completed a topographic map of the entire land surface between 60°N and 60°S over the world. This results from the efforts involved in the SRTM30 project. Although more accurate maps exist for some parts of this surface, this global map represents the best and most uniform quality map of the continental elevations now extant.

Dark green denotes the lowest regional elevations. Brown grading into white defines the highest elevations - in southern Asia and the West Coasts of North and South America.

To gain an insight into the improvements resulting from SRTM30, consider these two maps of the Guiana Highlands in the Guyana Shield of northern South America. The map on the left below was compiled from the best published data prior to 2000; its surface resolution is 986

meters. On the right is the SRTM30 version, at 30 m ground resolution and 10 m vertical resolution. This latter is far more accurate; note how the higher elevations shown in the left map are greatly diminished (i.e., were initially exaggerated) as the truer configuration from SRTM replaces the older version.



Follow-up programs are "in the works" to use radar interferometry and laser profiling to continually upgrade topographic/elevation mapping of the entire world.

Suffice to close this Section is this direct observation: Our world has always been a 3-D one; now a variety of space imaging systems that can produce end products in three dimensions will benefit cartographers and others who generate many kinds of maps that depict the true nature of the land surface.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

A TIP TO THE USER: MANY OF THE IMAGES IN THIS SECTION ARE STEREO PAIRS. SOME PEOPLE CAN SEE THESE IN 3-D DIRECTLY FROM THE SCREEN, WITH UNAIDED EYES, BUT MOST NEED A STEREOSCOPE TO VIEW A PAIR. WE SUGGEST THAT YOU PRINT OUT A STEREO PAIR AND EXAMINE THEM WITH A STEREOSCOPE (OR BY THE VIEWING TECHNIQUE DESCRIBED LATER IN THE TEXT). IF IT STILL DOES NOT FUSE INTO 3-D, THEN CUT THE PAIR APART AND MOVE ONE RELATIVE TO THE OTHER WHEN PLACED SIDE BY SIDE UNTIL THE 3-D EFFECT IS CLEAR.

Note: If you do not have access to a stereoscope, and are unable to see in 3-D without one, then you will be unable to try to answer some of the questions. Just skip them, or at least read their answers.

One major topic that is customarily included in any review of Photogrammetry (Page 10-3) is that of experiencing the three-dimensional nature of objects that we see directly with our eyes. This stereo capability is vital to our sensual power to negotiate the world about us - for example, we must be aware of the depths or distances of objects before us as we move about. It is possible to duplicate this sense of stereo using aerial or space imagery. Normally, this requires two images offset along the direction of the flight line (or through sidelap) so that each share part of the same scene that is visualized from different angles. This mimics the way in which our eyes see from slightly different angles, owing to the distance apart between the eye pair. This Section examines what stereo is and how it is produced through sets of imagery. Its application to producing topography maps (also doable using altimetry) is a main topic. The first page reviews types of maps and map projections and introduces concepts of contours as seen on topographic maps.

THE EARTH'S SURFACE IN 3-D:

DETERMINING TOPOGRAPHY FROM SPACE

Ways to Characterize the Earth's Surface on Maps

Most "look down" remote sensing focuses on the Earth's surface. The big exception is, of course, remote sensing of the atmosphere. Some surface features are readily apparent, such as tree tops, building tops, etc. On land, the real natural surface is usually soil or rock. On the oceans it is, obviously, the sea surface, although we could argue that the ocean floor is a continuation of the solid land surface. One definitive property of these surfaces is the nature of their composition, typically either rock/soil, water, vegetation, or manmade coverings. These materials are the constituents of the targets, further defined by form and use, which we try to identify using remote sensing devices, by sampling the electromagnetic radiation emanating from them.

11-1: **Such features appear two-dimensional in an aerial or space photo, although one often can get a sense of three-dimensionality from either shadowing effects or the psychology of recognizing a feature and knowing (commonly, intuitively) that it has a**

definite shape that we recognize as 3-D. Mention several advantages of being able to see a surface and objects thereon in three-dimensions. ANSWER

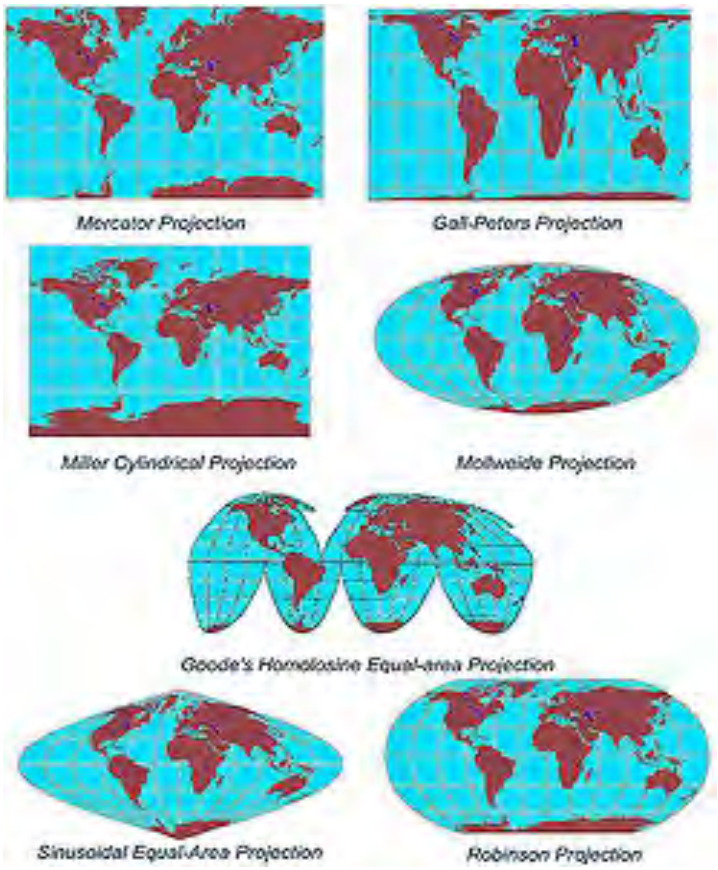
Earth's surface is not flat over large distances, and commonly, not even locally. The surface continuum of the Earth's sphere curves variably or undulates with respect to a reference shape. In fact, measurements have shown that Earth's gross shape is not a true spheroid, i.e., all points on the surface are not equidistant from the geometric center. Instead it is an **oblate ellipsoid**, in which the radius to the polar regions are slightly less than that to its equatorial belt, because its internal materials bulge outward in response to centrifugal forces developed by Earth's rotation. Large clumps of dense rock and soil located in the crust and outer mantle further modify this ellipsoidal surface.

This modified ellipsoid gives rise to the ***geoid***, a figure in which the surface broadly curves under or above the ellipsoid as gravity pulls on these masses. The geoid is defined as the shape assumed by the Earth's globe that would hypothetically occur if gravity were constant at all points on its surface. In the oceans this shape almost coincides with mean sea level (see below). Geodesy, the science of the Earth's shape, tells us where the geoidal surfaces depart from the ellipsoid. To a first approximation the differences worldwide tend to range between about -100 to +70 m (mountains and trenches are obviously much larger departures). Geodesy techniques initially included gravity measurements in the field, and during the last 40 years have depended on refinements resulting from perturbations in satellite orbits caused by gravitational variations. Even this shape still does not precisely characterize the true state of the surface because ocean levels change slightly from place to place, because of winds swelling water bodies, thermal expansion, tides, etc. Land shapes change further by tectonic and isostatic deformation and by erosion that raises, lowers, and inclines surfaces relative to the geoid . On land, in particular, these surface changes can be out of "sync" with the locations they would assume if ideally adjusted to gravity. The relation between the surfaces associated with the ellipsoid, the geoid, and the land/ocean bottom surface is evident in the next diagram. Here, a cross section of a part of the outer global interior shows the smooth curves for the two "oids" and the irregular profiles that generalize ocean floor and continental surface outlines.

11-2: The geoid is itself not the actual surface but a gravitational approximation. If the maximum relief on the Earth's true surface is set as the difference in elevation from the high point (Mt Everest in the Himalayas, at just over 29000 ft) and the low point (about 36200 feet below sea level in the Marianas trench (Pacific Ocean), express this difference in elevation (relief), and thus an up-down (radial) departure from a true sphere, as a percentage of the Earth's average radius (3980 miles). ANSWER

Cartography is the specialty that maps the Earth's geoidal surface in selected areas onto paper sheets (normally). A *planimetric* map shows locations and features on the surface that are at different heights above or below a reference (datum) plane. In their relative positions, they appear as though projected onto an X-Y (horizontal) coordinate system. Each point on the map corresponds to a point on the Earth's surface. These surface points, however, are distributed on the curved surface of the Earth. To locate the points, we must reference them to a geographic coordinate or grid system, the most common being that of latitudes and longitudes. Latitude lines run parallel to the equator, each one being a successively smaller circle as we move toward a pole. Longitude lines are meridians, great circles of equal radius that pass through the poles and are perpendicular to all latitude lines.

Most maps are geographic in nature, that is, they purport to show what is on the Earth's surface in terms of location and relative position. To get to a flat map is tricky, in that points on a sphere (the Earth) must be transferred to the flat map. Various projections have been devised to express geographic locations. A good review of map projects is found at this USGS website. A second overview has been put online at the National Atlas Internet site. Another reference website that considers map projections is produced by John Savard. A fourth source of projection information is based on material developed by Peter H. Dana, but resides on this University of Colorado website. Several of the more common projections are shown in this figure:

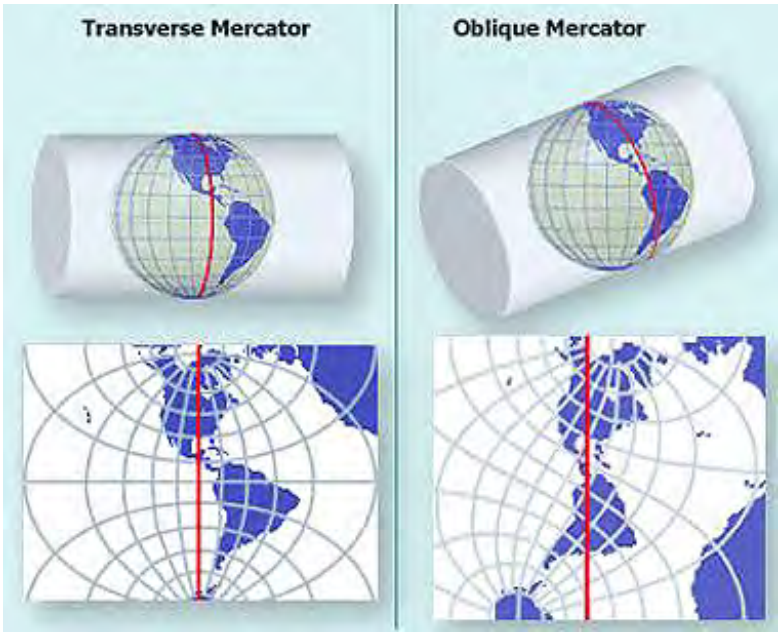


Thus, one encounters problems representing these points along a curved (topologically distorted) surface, when we project them onto a map plane. Cartographers address this problem by using a transformation between spherical and planar geometry. They create a specific projection that systematically transfers the surface points (latitude and longitude) to a reference surface, such as a cylinder, enclosing the globe, a cone, or a plane tangent to some part of it, which they can further "unwrap" into a two-dimensional plane (the map). Any transformation of an area on the sphere to a plane results either in distortion of the areal scale or of the shape. Different projections strive to minimize one or the other distortion. *Conformal* projections seek to preserve true shape: the best known of these is the *Mercator* (cylindrical), in which they space meridians equally and parallels become closer near the equator. This geometry tends to enlarge the areas of land masses at high (nearer to the poles) latitudes. *Equal-area* projections, such as the *Goode homolosine*, typically have non-straight longitudes and may have unusual outlines.

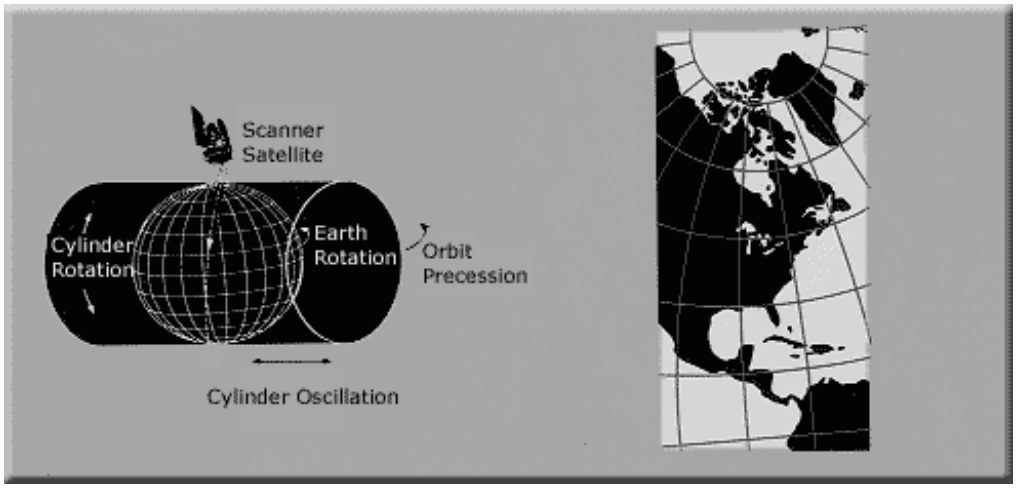
The polar projection, best for representing polar regions, centers on the north or south pole, around which is a series of concentric circles (latitudes) and straight radial lines (longitudes) outward to the outermost circle (the outline of the map). Depending on the projection type, the map scale can be constant or may vary.

11-3: What does this notation mean: lat 34°12'31" S, long. 77°03'41' E? Is the length of 1° of latitude and 1° of longitude the same at the equator and near the poles? ANSWER

Landsat and most satellites that circle the Earth in near polar orbits utilize variants of Mercator projections. The projection on the right of this next diagram is an Oblique Mercator projection. Thus, the pole to pole lines are not parallel with longitudinal lines. For Landsat, this takes into account the 11° inclination of its path with respect to the meridianal lines.



The specific projection used by Landsat is the Space Oblique Mercator system developed by John Snyder of the U.S.G.S. The Path-Row method of locating images is based on this system.



Topographic Maps and the Concept of Contours

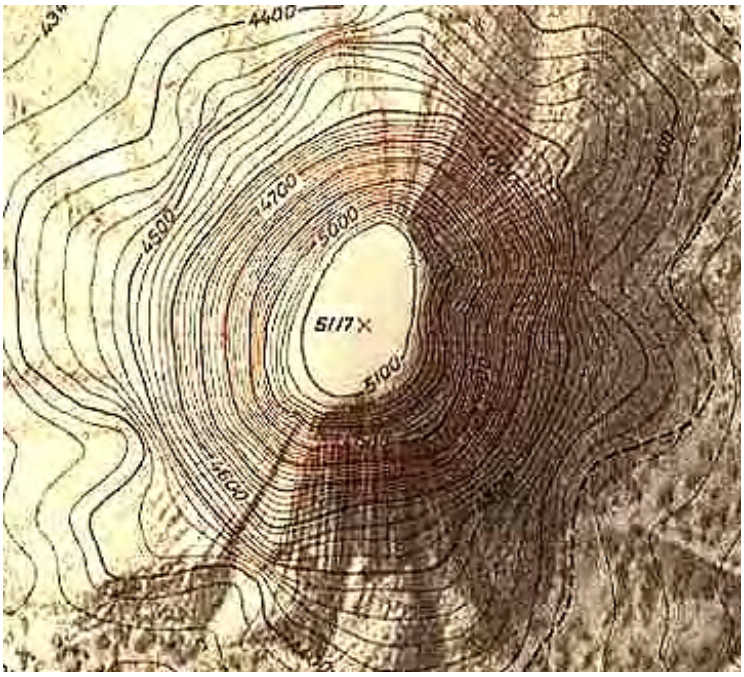
The most widely used surface-shape representation is the topographic map. In this map type, the third dimension (Z in a Cartesian coordinate system), associated with X-Y positions, plots as elevations (altitudes or heights) above or below a plane related to the spherical geometry of map projections. We choose this datum plane by convention to be mean sea level, closely approximating the equipotential surface that marks the oceanic segments of the geoid. We determine the geoid at land-sea interfaces as the average height (set to zero) of coastal tides over a long period (about 19 years).

Once tied into the geoid, surveyors take measurements to establish elevations inland (or along the ocean floor). On land, for more than a century surveyors have used theodolites (transits), which are instruments for measuring angles consisting of a telescope that freely rotates 360° horizontally and pivots from 0° to 90° vertically. Today they may use laser beams for ranging (determining direction, distance, and elevation to points being surveyed) and the Global Positioning System for location. By using chains and stadia rods, surveyors determine horizontal distances and directions (bearings or azimuths) from their instrument to a survey point or station with high accuracy. They measure lines along levels and calculate trigonometrically the vertical differences in elevation between points. In this way, they can run survey lines starting either from coastal sea level locations or from points inland of known three-dimensional position to establish benchmarks (points of accurately determined geographic position and elevation). The surveys develop nets of baselines and extensions anchored by these benchmarks using triangulation methods and other surveying procedures, and gradually the nets extend throughout a region and ultimately may cover a continent. These nets serve as control points from which we get topographic maps of various scales (and therefore details). These maps display the usual planimetric features, including latitude-longitude coordinates, specific scales, and projection type.

One of the spin-off uses of these surveys is to set up specific ground control points that are needed to properly make measurements in aerial photos, especially, as we shall see, when these are to be used in making topographic maps. These may be horizontal or vertical control points, or both. The points must be identifiable both on the ground and in the photos. As an example, the interesection of two roads will fulfill this function, being easily located on the surface and normally being readily discernible in aerial photos having wide ranges of scales (but not necessarily visible in space imagery of intermediate resolution or less).

Variations in the third dimension, from which one is able to envision changing surface curvature and to calculate relative heights of different points on the surface, can be depicted on maps by using *contour lines*. A contour line connects all points of equal elevation within the mapped area. To picture such a line, imagine that you are standing at the base of a hill that rises from a flat plane and peaks in a single crest point 100 m (328 ft) higher. If you walk around the hill base, keeping constantly at the same elevation, and trace your path somehow (such as painting the line of your trail), you would have a line, similar to a contour line on a map. Now if you move up the hill a fixed distance, say 20 m (66 ft) higher than the base, and again walk around the entire hill while maintaining the same elevation, marking your way by paint, you would have another contour line, which is a slightly different shape. Then you continue moving upslope at 20 m (vertical or elevation) intervals, each time walking horizontally around the hill, and marking your path. Finally as you approach the crest, your walk will be a minimal length while you still will return to the original starting point (closing the line). When you view these traces from high above, you will see the successive traces as a (nested) set of concentric closed lines, each interior one of lesser total length than the lower one outside. If labeled, each line has a value 20 m higher than the next outside one. If the base were assigned a value of zero, then the upward sequence would go through 20, 40, 60, 80, ..., up to the value of 100 m. The crest may be a few meters higher than 100, say 105 m, and would be plotted as a point within the innermost contour. If the base chosen lies above sea level, say at an elevation of 120 m, then the sequence would be 120...140...160...180...200...220 m (with the crest point labeled 225 m).

The contoured topographic map below is an actual example similar to what was just described. Notice that the darker brown lines repeat every hundred feet; there are 4 lighter lines in between. This spacing is called the contour interval and in this case it is 20 (feet or meters).



11-4: Imagine drawing contours on the top half of your head. Starting at the top of your ears, you can trace a circle around your head (perhaps using a scissors to cut away hair along a very thin trace path). Then trace the next contour a quarter inch higher, and so forth, until you reach the top. From above what would this circle-line haircut look like? ANSWER

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The nature of contours is examined in more detail. This is done by showing several topographic maps that show how contours express elevations. Several rules regarding interpretation of slopes and heights from contour spacing are given.

The Display of Contours in Maps

If you are not familiar with how contour lines are drawn - particularly using an interpolation procedure - we will give you a simple task of making a contour map from points that are assumed to have been obtained during a field survey. Consider this diagram (adapted from one from the U.S. Geological Survey, so that here and in subsequent maps we use English units exclusively):

The upper view is a perspective drawing of a coastal landscape, consisting of a shoreline, including a barrier spit (hooked sandbar), a river valley (with bluffs), and two hills. One hill has an inclined tabular slope, bounded by a steep cliff that grades into a terrace, and the other has a crest point, from which extend several ridges between stream cuts.

In the lower view above are a set of surveyed points with their elevations, starting from the coast at 0 ft sea level, up to two benchmarks (BM) atop the hills. This is typical raw data, from which we can draw contours. The contour-drawing procedure relies on interpolation of values set by the chosen contour interval of 20 ft between pairs of survey points. To see how, print out the above diagram. Start with the two points at 52 and 90 ft (left of panel center). Draw a light line from one to the other. Assuming equal spacing, place a short mark (say, a dashed line) across the light line, valued at 60 ft just left of the 52-ft mark, then another at 80 feet. Estimate these locations as if there were a constant slope from the 52 ft mark to the 90 ft mark. Imagine

dividing that slope into 38 1-ft units and the two short lines lie at 8/38th (60 ft mark)and 28/38th (80 ft mark) upwards from the 52 ft point.

Using the same equal spacing approach, along a path to the coast between 52 ft and 0 ft, draw short lines at 12/52ths (40 ft mark) and 32/52ths (20 ft mark). You can do the same interpolation between the 100 ft and 56 ft points (near top center), and likewise for any other pair of points. This process leads to a large number of interpolation lines of different contour (elevation) values. Now, similar to the parlor game of connecting the dots, connect the various short contour line marks with the proviso that they cannot cross one another. Your result should look like this:

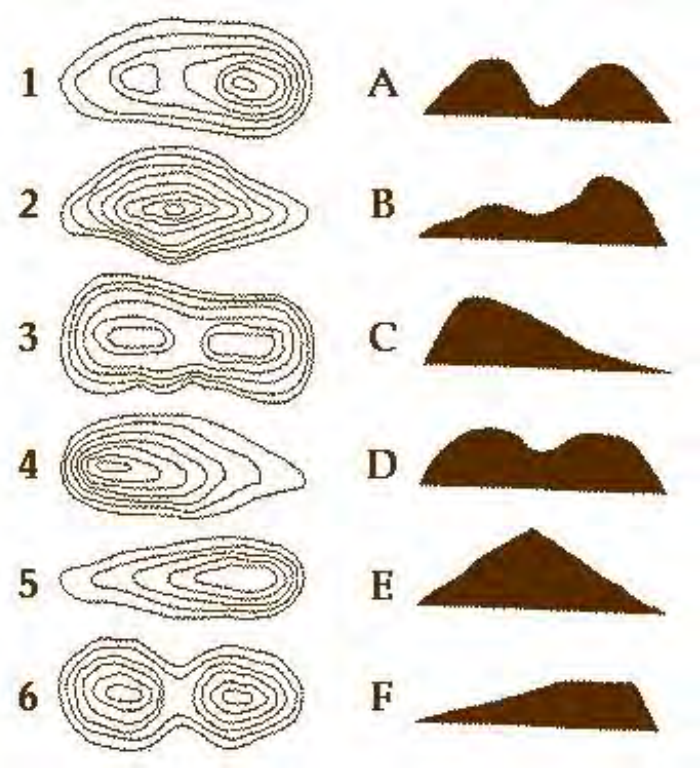
Note three other characteristics (besides non-crossing) of contour lines: 1) the gentler the slope, the wider the spacing between lines (very closely spaced lines denote a cliff), 2) where lines cross a stream, they must bend upstream (remember, a stream moves directly downslope and therefore we must walk horizontally to cross it and get to the same elevation point on the other side), and 3) by convention, we draw every fifth line in a heavier weight-for the contour interval in this case that is at 100 ft and 200 ft.

11-5: There is one topographic situation which, if represented by contours, involve some of the contours crossing over others. Can you think of what this might be?
ANSWER

Note, too, that on most maps, contours are brown. We illustrate this on this part of a larger topographic map:

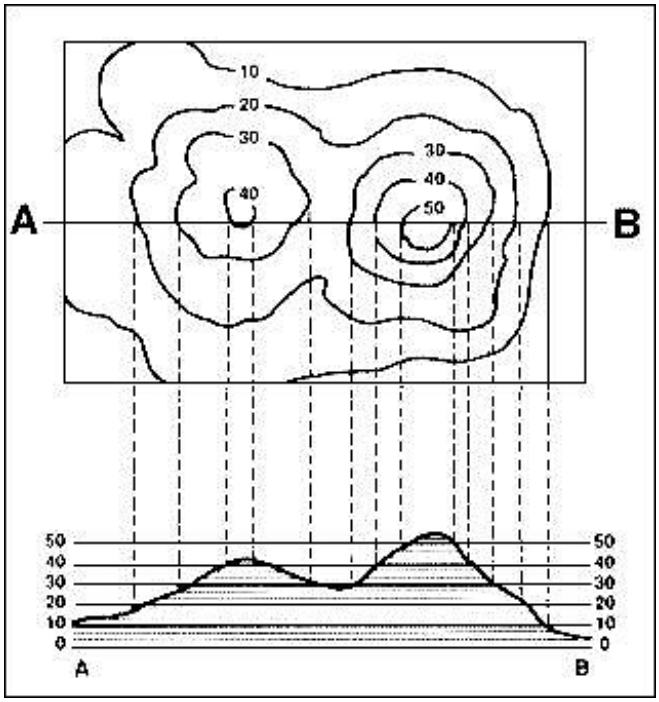
This map covers the region just south and east of Mammoth Caverns in Kentucky (below the surface in the upper left). The upper third of the map, judging from the 20-ft contour spacing and the shapes outlined by the contours, is an area of rolling hills with several hundred feet of moderate relief (relief refers to the amount of difference in elevation between two points; low relief implies relatively flat terrain, while high suggests steep slopes), with narrow stream valleys. We enhance the hilly appearance using a technique sometimes used by map makers of shading (here in uniform brown) steep slopes on one (and the same) side of the hills. In the map center are large numbers of nearly circular contours, and each grouping is quite small. These mark the sites of sinkholes. This is limestone country in which the so-called karst topography develops by solution and downward draining of the bedrock. The lower third of the map resembles the upper but with notably less relief.

By now, you should have fathomed what contour lines are and how they depict surfaces. To test your skill, consider this diagram:

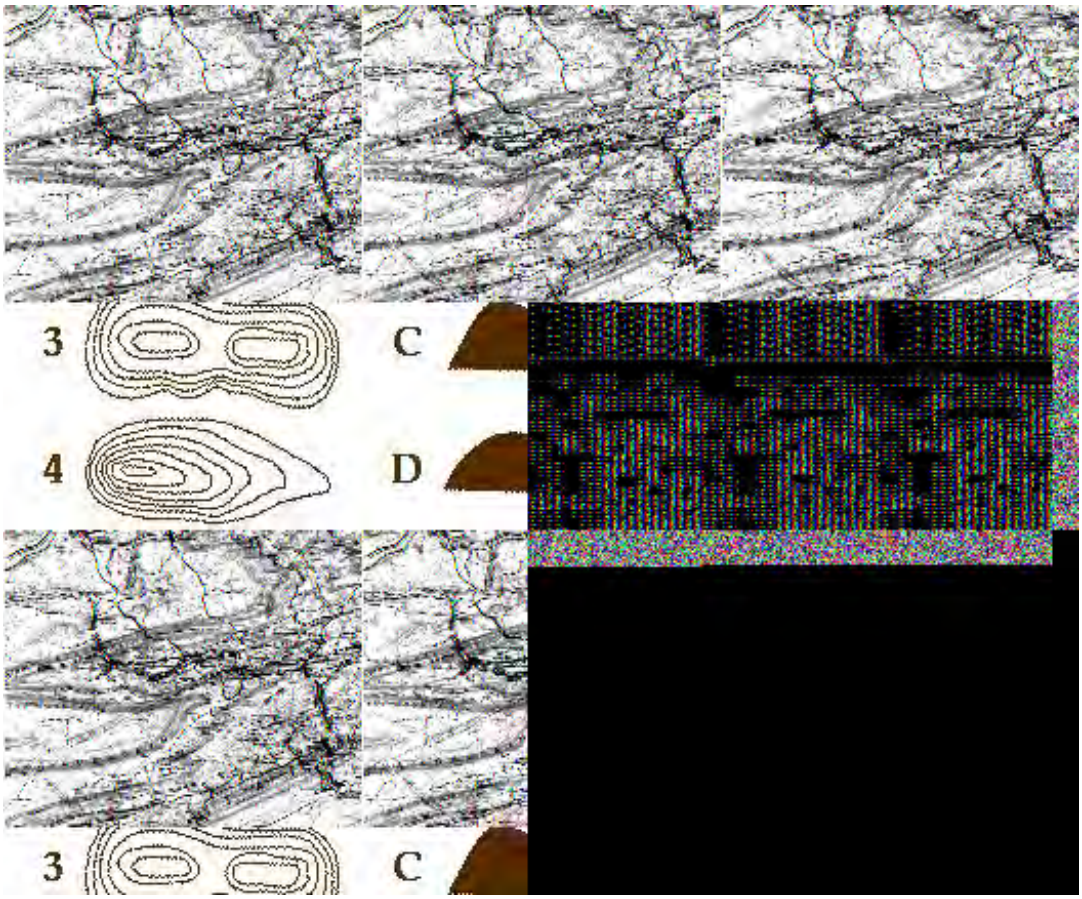


11-6: On the right is a 'side view' of hilly terrain. On the left are a set of contours. But the two don't match. Try to fit a map view from the left with the proper profile on the right. For example, 1 is associated with B. ANSWER

At the top of these silhouette-like side views is the shape of the surface. As you probably know, this shape is called a *profile*. The next illustration shows how the profile can be constructed from contours; the procedure is almost self-evident:



Now, as a more challenging exercise to test your ability to associate surface shapes by contours, look at this next map.



11-7 Try to determine which scene in Section 6, the cross-country flight, contains landforms similar to the ones you should visualize from the map. Check page 6-2 through 6-9. The map area is not in that scene but lies just to its east (next Landsat orbit to the right).. ANSWER

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Both the physical and psychological factors that underlie seeing in stereo are looked into, with attention paid to the notion of parallax. Some simple experiments that you can do to test and enhance your stereo visual capabilities will aid in fostering your skills needed to work with stereo imagery. The reason for the visual impression that objects so viewed may appear extended (exaggerated) vertically is presented in the base to height concept. The proper way to use a stereoscope is stated. Some encouragement is offered that many individuals can learn to see stereo without having to use a pocket-type or other stereoscope. The first examples of stereo pairs are displayed, with the suggestion that these be printed out for examining scenes under a stereoscope (which you may need to buy or otherwise procure.

Seeing in 3-D Stereo

Prior to the 1930s, most topographic maps were made exclusively from field surveys. With the advent of aerial photography, and specifically as aerial surveying produced stereo pairs, planimetric maps became possible and special instruments were developed that could draw contours. Ground surveying is still necessary to precisely locate calibration points within photo scenes.. To understand how we apply stereophotogrammetry to contouring requires mastery of the concepts behind stereographic viewing.

This style of viewing is as familiar as seeing. When we alternately blink our eyes, objects appear to move slightly sideways. This shift is even more dramatic when we hold an index finger about 30 cm (18 in) in front of our nose and perform the above eye activity. The finger seems to jump left and right. This apparition comes from the principle of parallax. Parallax is the apparent displacement of a viewed point or small object at a distance that results from a change in the point of observation. For a person, that change in the point of observation could simply be from one eye to the other at a fixed location or from relocating from one viewing spot to another. Our line of sight from each eye is not quite parallel to the line between our nose and the selected target, but converges from the eye pair, so that the left eye sees a bit of the left side of the target not seen by the right eye, while the right eye sees a bit of the right side, missed by the left eye. In this way our eyes send a signal to the brain which, on further processing, creates the impression of depth.

This same effect lies at the heart of stereo viewing of photo pairs, taken either from two lateral positions on the ground, or, in aerial photography, as successive photo pairs with about 50% overlap along a single flight line, or with similar sidelap between pairs from adjacent flight lines, with some of the scene in common within the pairs (see page 10-3 for further details). When we position the stereo pair properly left-right and then view them through a **stereoscope**, the eye-brain reaction is an impression of surface curvature or relief, as though we're looking down from a plane at the ground. A pocket stereoscope consists of two lenses that we can adjust along a slide bar to be as far apart as our eyes; these are placed in a raised mount (on collapsible legs) which is positioned about six inches above the central region of the stereo pair.

The sense of relief may be exaggerated relative to reality. The degree of vertical exaggeration (VE) depends on the base to height ratio (B/H), which depends on the scale of the photos. The scale, in turn, shows the actual horizontal ground distance (B) between any two equivalent points, identifiable in the two photos, and the height (H) of the camera, during the exposure of each photo in the pair. These points will, of course, not occupy the same position in the two photos because of the forward motion of the imaging platform. The vertical exaggeration also

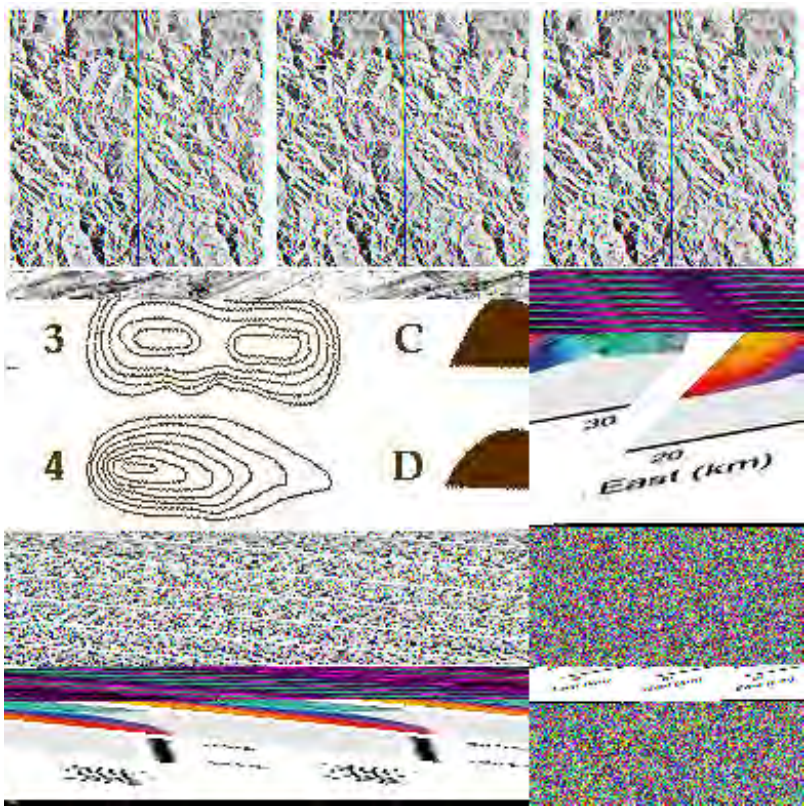
depends on the apparent height (h) of the viewer's eyes and the breadth (b) between the eye centers of the particular viewer. So VE = (B/H)(h/b). VE typically ranges between 1+ and 6+ for B/H ranges between about 0.2 and 1.2.

11-8: Calculate the vertical exaggeration for this set of conditions: Distance on ground = 1800 meters; Height of camera above ground = 4000 meters; Apparent height of viewer's eyes = 40 cm; width between eyes = 6 cm. ANSWER

There are several stereo pairs on this page and on page 11-8. If you try to print out either page to get these images to cut apart and examine in stereo, the length of the page will require several printed pages. To avoid this, we have placed some representative stereo pairs on a separate page, without text, which will reduce the number of printout pages needed. Access this page for printing by pressing on the word STEREO in blue/purple; return to this page or 11-8 by clicking on the browser BACK button.

If you have a pocket stereoscope, you may see the stereo effect by placing it, with legs extended, against the image below as it appears on the screen. This viewing usually doesn't work for most observers, so, it is probably necessary to print the pair, cut them apart, and then view them with the stereoscope. You likely will have to move one or the other laterally until the area viewed in common fuses visually into the stereo effect. (A few people can actually get a stereo effect from the images on this page as presented on the screen. But that ability is too uncommon to rely on, so we suggest instead that you go the printout route.)

We are reprinting below the first pair on the STEREO page, placing them in juxtaposition, separated enough for stereo viewing. Cut them from the printout page and position them. Staring down, try to get the stereo effect. You will probably have to shift the left laterally closer to, or farther from the right until the separation is just enough to bring about the effect (commonly, you start to see stereo but the two images aren't quite fused, so that you need to shift experimentally until the images coincide; also, when fusion occurs some people will see a third partial virtual image in the center between the two real images). This pair shows mature topography in a dissected hilly terrain, You may see the slight differences in shape (and shadowing) of the same hills in the two photos that result from the changed viewing positions. ()

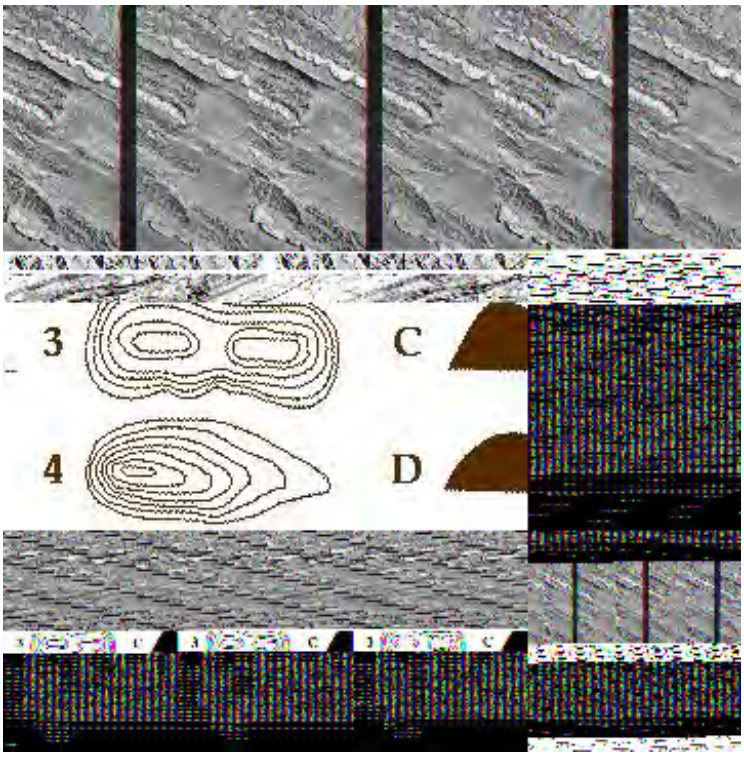


11-9: In the above stereo pair, assuming you have succeeded in viewing them

stereoscopically, where is the apparent highest peak? ANSWER

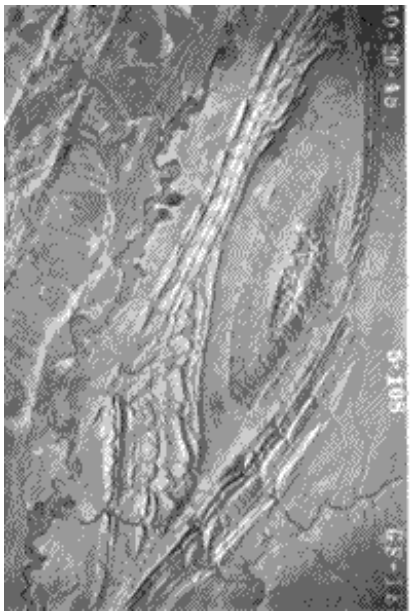
If you do not have a stereoscope, you still may be able to get the 3-D effect without one. Many people can see in stereo with their unaided eyes. To test the likelihood of doing this, contact your two index fingers tip to tip at eye level about 20 cm (12 in) in front of your nose. You may see not only them, but an illusory "sausage," about an inch long, consisting of the two finger tips, that appears to "float" between the two real fingers. Moving the fingers closer to or farther from your eyes causes the sausage to expand or shrink. This won't work if you focus directly on the fingers, but should happen if you focus on "infinity," that is, gaze at long distance, so as to focus well beyond the fingers (the "vacant stare"). If this doesn't occur for you, it means you either have some physical eye limitation, or you psychologically don't believe that you can do this. Try this natural stereo viewing either on the image pair above or the one below (print them out, if you need to).

Now practice with this second image, which you also cut out:



11-10: Which way do the strata (bands of rocks) appear to be inclined (the geologist says "dipping"), to the upper right or to the lower left? ANSWER

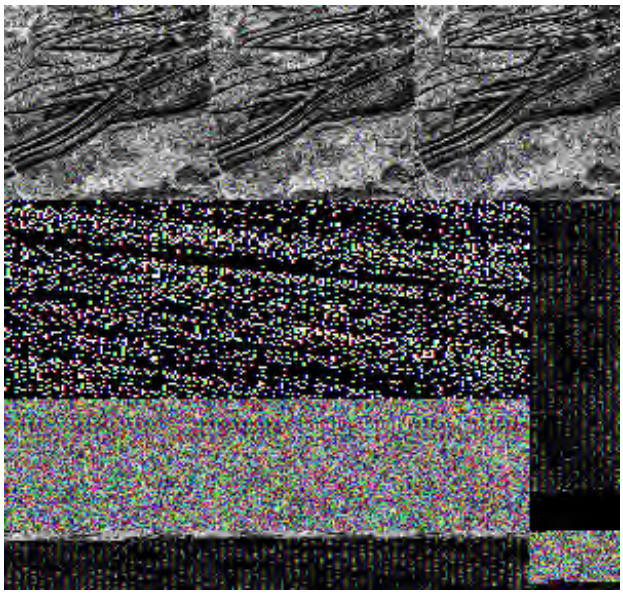
We have scanned a pair of photos of the Little Dome (Wind River Basin, Wyoming) anticline (upward arching fold) and reproduced them below at rather large size. If you have not done so yet, access them by clicking on STEREO. After printing and cutting them apart, put the left one A to the left of the right B (or bottom one) and check their stereo expression. It may show up just fine or it may look funny. The positioning relative to the sequence in the flight line must be proper for normal stereo to take place. Since there was no information available to the writer (NMS) as to whether the top photo was taken first or second, you will have to experiment with choosing which photo to place left of the other, switching until the normal effect leaps out at you. We are printing the left one (A) of this pair on this page to make sure you have the right photos. Hold one in place and move the other laterally until stereo appears. You may have to bend (curl) the paper of one to see the corresponding area in the other. Be sure to trim any borders to eliminate any white effect.

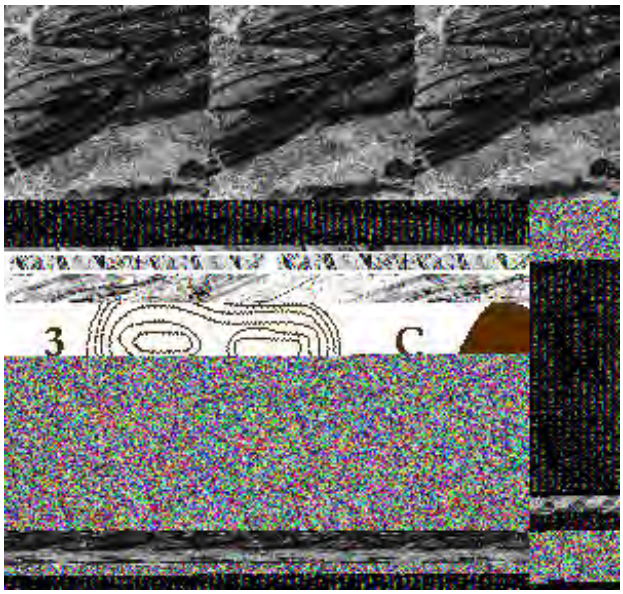


Stereo Pairs from Space

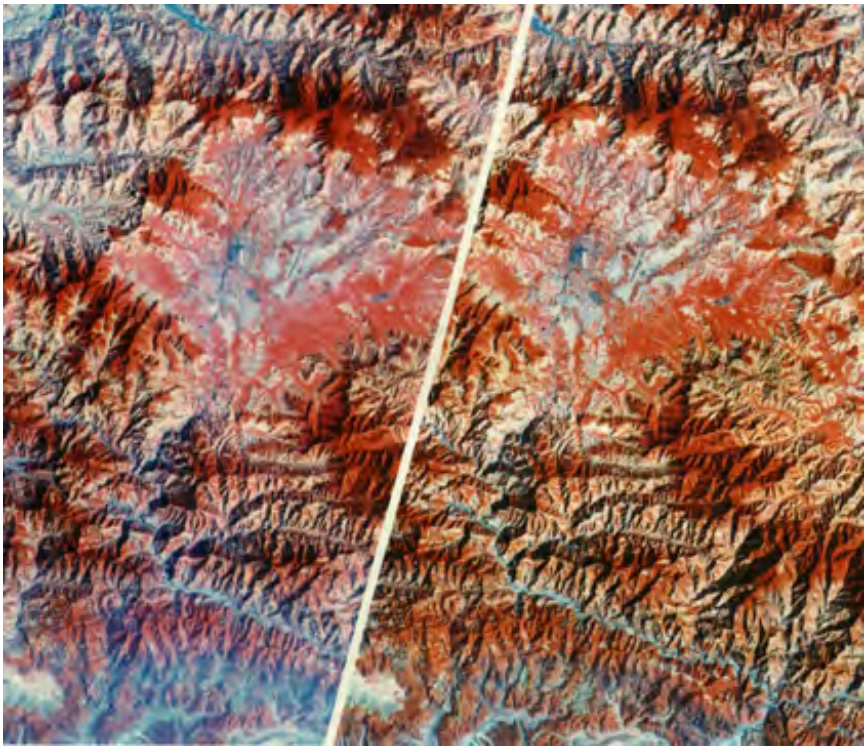
Astronauts took the first Visible-IR stereo-pair photos from space during several missions. These photos led to an argument during the first days of Landsat-1's mission as to whether the sensors should be modified to acquire stereo. Landsat was not designed to acquire in stereo, because its downtrack images result from continuous scanning. However, its orbital track results in 10% to 40% (near the poles) sidelap coverage along adjacent orbits (previous or next day), which provides limited area stereo viewing, if cloud conditions are favorable. There is a Landsat stereo pair on the linked stereo page ([STEREO]; if you haven't done so, print out this page and cut the Landsat images apart for appropriate overlaying). They consist of two complete Landsat MSS Band 7 images taken on October 10 and 11, 1972, thus just one day apart during a time of cloudfree conditions during the continued presence of a pressure High following an earlier storm. Determine where the scenes are geographically, the approximate percentage of sidelap that produces stereo, and the amount of stereo relief (governed by the B/H ratio).

This works also with Landsat images taken along contiguous orbits at various times of year. Check the stereo strips covering part of the now familiar Blue Mountain and other ridges near Harrisburg, PA which you printed out earlier (if not, go to STEREO): the top image (MSS Band 5) was taken in February (defoliate; low sun angle emphasizes relief), whereas that on the bottom was obtained the previous October (foliage still on trees). If you succeed in getting a stereo image using a printout, note the very low vertical exaggeration (about 0.3) because of the high H value (918 km).





Now, try for 3-D in this Landsat false color composite stereo pairing of images of Katmandu and the surrounding Mountains in central Nepal, taken on March 20/21, 1977. The Environmental Institute of Michigan (ERIM) prepared the mount.



Additional stereo images from SPOT are shown on page 11-9, along with the means to print them out.

11-11: In stereo, which seems to be lower: the elongate hill within the anticline or its rim? ANSWER

There are two more stereo pairs on the STEREO printout page. One pair is a color set from SPOT. It is prepositioned to achieve stereo viewing for some people (if not, you must cut and separate). The other consists of two adjacent Landsat images of parts of eastern Pennsylvania, some of New Jersey, and a bit of New York. Look carefully for the upper right part of the left scene that has a common area in the left of the right scene (use the lakes as a guide). When separated, position them properly - in this case, you will need to shift one of the pair vertically since the two scenes are offset in their relative orbital locations. Try to get the common areas to fuse stereoscopically. If you do, you will notice tht the vertical relief is quite subdued: Owing to its

high altitude, there is only a small B/H effect so that the stereo view shows little exaggeration (thus, Landsat has only limited value in this mode of viewing).

The next page continues the theme of stereo imagery acquired by satellites.

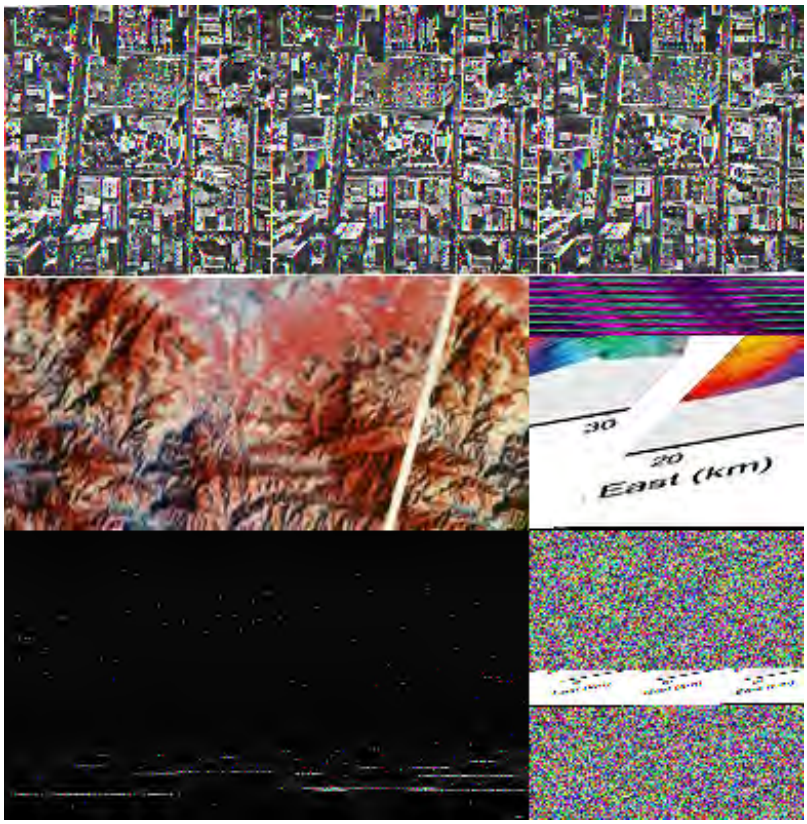
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The spatial distortion possible in an aerial photo, in which elevated objects such as buildings appear to lean outward, is first illustrated; then, you are presented with a formula by which height of an object in a photo can be calculated. A schematic diagram helps to understand the parameters in another formula that applies to a stereo pair of successive photos along a flight line. The use of special instruments called stereo plotters that allow a systematic process of determining contour lines is described as one of the conventional methods for producing topographic maps.

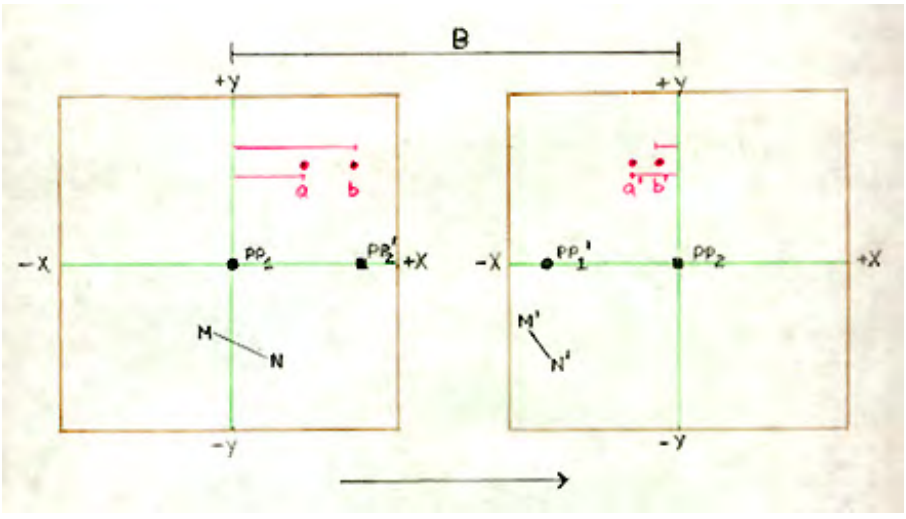
Measuring Heights from Individual and Paired Images; Contouring

We can measure heights of surface features (e.g., tall trees and buildings) and relief by various photogrammetric techniques. This information leads to expression of elevation differences as contours. Consider for example, determining the height, **h**, of a tall building that lies near the edge of a single aerial photograph. If that photo was taken vertically, i.e., looking straight down, then its center lies at the nadir (the vertical line from camera perpendicular to a surface point directly beneath [assumed to be flat]). In that instance, the nadir coincides with the principal point (p.p.) in the photo, which is the intersection of the lens' optical axis (as though extended) and the ground, so that the p.p. is also the true center of the photo. That point becomes off-center, if the photo (and optical axis) is non-vertical. If the building were to lie close to the p.p., its top and bottom would appear to coincide in the photo. If it is well away from that point at some radial distance, **r**, towards the edge, its viewed state at a slant would have the top displaced further away from the p.p. than the bottom by some amount measurable in the photo as, **d**. For an aircraft height,**H**, (calculable from the scale if we know a ground distance), the value of h is just, **h = (d/r) x (H)**. We can illustrate this approach with a large scale, low camera altitude aerial photo of the downtown area of Long Beach, California (from Sabins, 1987; courtesy J. Van Eden), in which the lateral tilt of tall buildings is obvious in the outer parts. Here, we drew d and r between the p.p. and one of the buildings near the edge (lower right); d is just the distance on the photo (inches or centimeters, or fractions thereof) from the base of the building to its top and r is the distance from the p.p. to the building base:



11-12: Using this figure, determine the height h of the building to which are drawn white arrows to distances d (photo displacement from bottom to top) and r (to building top). On the actual photo (not your screen) d = 0.5 inch and r = 3.0 inches. Scale of the photo is 1:3600. Aircraft altitude is 1800 ft. ANSWER

Elevations and slope lengths and inclinations can be determined from stereo photo pairs by application of parallax principles. The diagram below will serve to introduce the techniques involved. In this diagram, the two successive photos are shown as drawn apart to permit display of the geometric points or features that, in actuality would be common to both, as evidenced by the amount of overlap.



Stereo parallax (see also page 11-3) is actually a physical condition that refers to the apparent displacement of a point or a feature of some height in an photo or image caused by a shift in the position of observation: This will occur in aerial or space photos that have been taken in succession (such that a stereo pair has either overlap (same flight line) or sidelap (parallel flight lines). In the diagram above, note the position of the principal point pp_1 in the left photo. In the right photo, this first point is now at pp_1', and lies left of that photo's principal point pp_2. The position of pp_2' in the left photo is at pp_2, the actual principal point for that photo.

Next, consider points **a** and **b** that fall within the two photos and, for simplicity, are located on a line parallel to the flight line (arrow at bottom). In the left photo, they are located in the upper right quadrant; in the right photo, they have been shifted to the upper left quadrant and the spacing or separation between them is, in this example, less. The shifts can be measured on each photo in terms of units of x and y chosen as lengths (e.g., millimeters) in a coordinate system that places 0 (zero) at the principal points. The two red horizontal lines to **a** and **b** can be read as the x values (no numbers given in this general diagram). (Note that x and y can have positive or negative values.) Because **a** and **b** are in different locations (and different distances from the principal points) in the two photos, their separation is not the same in the photo pair (in general the parallax effect and the displacement from the principal point that increase with distance from the point, as suggested in the photo at the top of the page, will cause separations to vary).

In this scheme, the parallax P for the point **a** is $P_a = x_a - x_a'$ and for $P_b = x_b - x_b'$. Now, the height at point a, h_a can be determined if three variables are fixed: 1) the actual ground distance, known as the air base B, between the two principal points (determinable if image scale is known), 2) the height H of the aircraft, and 3) the focal length f of the aerial camera. Then, this equation applies:

$$\mathbf{h_a = H - (Bf)/P_a}$$

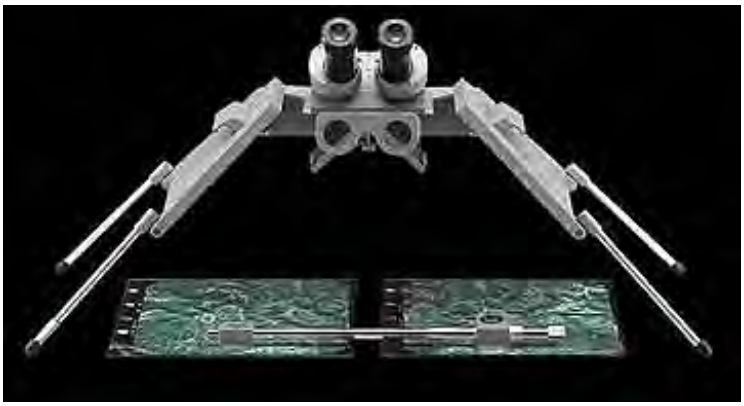
In the same way, the height of point **b** is determined. Then the relief is just the elevation difference between the higher and lower points.

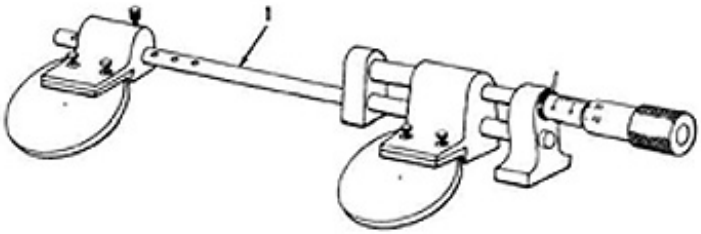
Since H and f control the photo scale, these values plus the appropriate parallaxes and the air base B are sufficient to establish the actual distance from the principal point of any location in one of the photos. For the point at **a**, the X and Y distances are given by $X_a = B(x_a/P_a)$ and $Y_a = B(y_a/P_a)$. Similar equations hold for point b. The slope angle Θ between points a and b is just the angle whose tangent is h_a/h_b, where h_a is found to be higher than h_b. The real length of the slope between a and b is determined using the Pythagorean equation; this is calculated as: $\mathbf{ab = [(X_a - X_b)^2 - (Y_a - Y_b)^2]^{1/2}}$.

11-13: For the adventurous, try this question: For the inclined surface containing separated points M and N (not on any illustration; just imagine them) on a slope whose heights or relief we want to determine, we measure the distance (in the direction parallel to the flight line) on one photo between the base (here, point N which is lower than point M) of that slope and the principal point (the optical center of a photo = nadir point if camera is on a level aircraft and points straight down). Use the above equations to figure out the heights of points M and N in the diagram, their coordinate positions (in millimeters) on the left photo, the actual length of line MN, and the slope angle. Imput parameters: H = 4000 m; B = 400 m; f = 120 mm; X_m = - 20 m; Y_m = - 40 m; X_m' = - 100 m; X_n = + 30 m; Y_n = - 70 m; x_n' = - 80 m. This problem is calculation intensive; do this question only if you are ambitious.ANSWER

It is also possible to calculate height differences (relief) between points of different elevation along surfaces from stereo pairs by using a variant of the parallax method. This formula applies: **h = (H) dP/(P + dP)**. **P** is the absolute stereo parallax. The **dP** term is the differential parallax, which is the difference between the absolute stereo parallax at the top and at the bottom (base) of the feature or point being measured in the pair. One can find the differential parallax by measuring the distance between the base and the top of the feature, or between two proximate terrain points at different elevations on a slope, that we locate in each of the two photos when the pair is in optimum alignment for stereo viewing. Then, dP is the numerical difference between the two distance values ($x_2 - x_1$), which is different in each of the photo pairs.

The photogrammetrist can optically find the values of P and dP by using special devices, such as a stereometer, a parallax wedge, or, most commonly, a parallax bar. Each uses a variant of the "floating point" method in which reference marks–dots or crosses–are visible when the aligned photos are viewed under a mirror stereoscope. A mirror stereoscope is shown in the next figure; it has a parallax bar device on top of the pair of color photos; a sketch of that instrument is placed beneath:





One places the parallax bar on the photo plane so that a fixed mark over one photo coincides with a movable mark (fused visibly in the stereo view) by turning a screw drive that translates that mark into coincidence. We then read the value of dP from a graduated scale. When we do this for a number of points, then, we can calculate heights and relief values for parts of the map. These become absolute values (elevations), if we reference the points to benchmarks.

11-13: A stereo pair was obtained from a photo mission flown at 2000 meters. We wish to determine the height of a water tower (storage tank on legs).The photo base length P between the two photos when properly placed for effective stereo viewing is 70 mm. The differential parallax dP for the tower, as measured by a stereometer, is 0.4 mm. What is the height? ANSWER

Using the above devices to specify a large enough number of elevation points in stereo photos to permit contouring is difficult and tedious. Instead, cartographers routinely apply sophisticated optical-mechanical stereoplotters. Two of these are shown in the figures below:

These devices use two or three movable projectors that pass light through stereo pairs made into glass-mounted film diapositives placed onto a tracing table (platen) which can be raised and lowered. The device produces a 3-D image or stereomodel, which an outside observer can view by some method. For black and white photos, this is done by the anaglyph method, in which one one photo is projected through a red filter and the other a cyan filter and with the observer viewing their superposed image through eyeglasses with corresponding color filters. Next a small point of light or dot is projected onto the stereomodel. When the dot appears to coincide with some part of the model surface (as seen through the viewer), that dot locates a particular elevation determined by the platen height and calibrated by control points. An experienced operator can then move the light along the model surface, such that it always remains coincident (in focus). This motion traces a contour line. The operator then raises (or lowers) the platen table a fixed amount determined by the chosen contour interval and and proceeds to contour the next elevation in the same way. This process continues until the operator traces the entire visual surface. Today, analytical stereoplotters are capable of semi-automating the contouring through computer processing of mathematically transformed data.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Elevations at individual points (commonly in systematic arrays or grids of geographic significance) as surveyed on the Earth's surface, or calculated by stereoplotters, or digitized from existing maps, can be expressed in digital numbers. The result is a Digital Elevation Model (DEM). This allows for production of a variety of maps that range from straight topographic ones to those with specialized displays such as shaded relief maps and perspective maps. The DEM database can also be "draped over" a remote sensing image (equivalent points registered) to produce views that give a 3-D appearance (usually from a perspective or oblique view) to the natural landscape or manmade features in an image.

Digital Elevation Models (DEMs) and Viewing Modes

In recent years, contour values have been digitized allowing them to be manipulated into versatile displays of topographic data as digital elevation models or DEMs. Digital terrain models (DTMs) are variants that show additional landscape attributes. We can use existing maps as inputs by tracing contour lines on a digitizing tablet or table. The data are organized into cell arrays whose X-Y positions in the rectangular grid are related to map coordinates. Each cell has a single value representing the average elevation of the land surface within it. The finest-sized cells are 30 meters on a side, associated with the 7.5 minute topographic quadrangles mapped by the USGS using the Universal Transverse Mercator (UTM) coordinate system. Only a fraction of the maps at this scale have been digitized, as yet. This digitizing is also true for 15 and 30 minute maps. To date, all of the 50 U.S. states, except parts of Alaska, mapped at 1:250,000 scale (extending over 1 degree by 1 degree in eastern states and 1 degree by 2 degrees in western states) have now been digitized at a cell size of three arc seconds. The data are stored in east-west profiles.

You can access a general review of the concepts and mechanics of producing DEMs at a site maintained by the U.S. Geological Survey. An overview of DEMs is found on Wikipedia entry.

These DEM data sets have important applications. They allow for rapid reconstruction of contour maps and for plotting elevation profiles. Also, we can easily do various kinds of photogrammetric calculations with the numbers. Image processing systems can edit and filter these data to enhance the display products and merge data sets into larger maps. Cartographers can present surfaces as X-Y-Z plots, where the Z dimension appears as regularly spaced vertical lines whose lengths are proportional to elevation. These plots are available in several formats:

Shaded relief maps, which assigns different shades of gray to slopes, depending on elevation, or a variant, which draws in shadowing to selected slopes depending on their orientation (aspect) and on sun direction and azimuth. By shifting these parameters, we create different renditions that bring to light different surface features and trends.

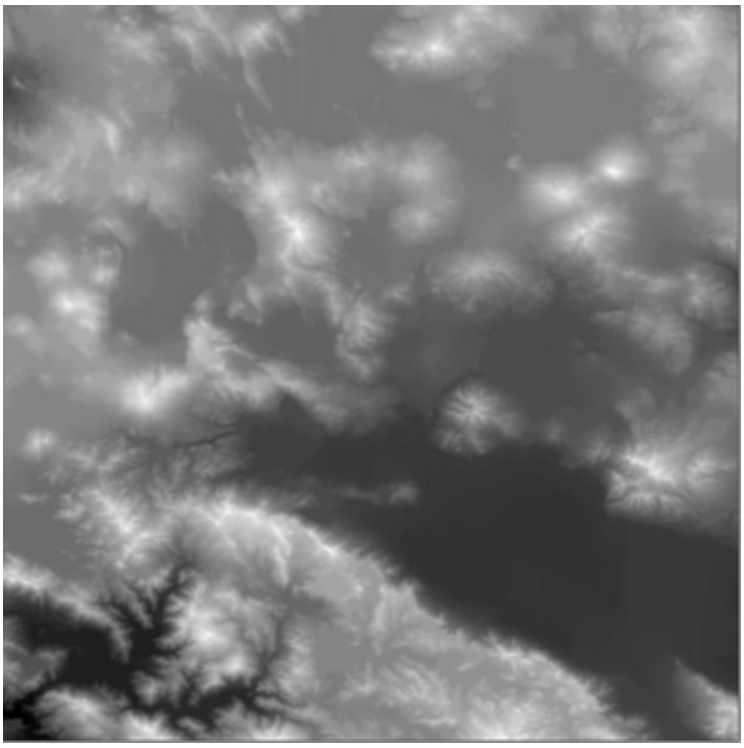
Color density slices, which color-code elevations.

Perspective Views. Since one can readily calculate geometric variables, such as orientation and height from the digitized data, the surfaces can be recreated to look like oblique photos. Then it is easy to examine such surfaces from different perspectives by rotating the view horizontally (along a vertical axis) or changing the look angle.

Draped Views. The ultimate display is usually one which registers or "drapes" a surface image, such as a Landsat scene, onto the DEM array (data cells match with pixels), causing that surface and the features on it to appear in some form of 3-D display (e.g., in a perspective view or by creating pseudo-stereo pairs as a stereo model). Likewise various thematic maps, as for example, land use, urban structure, or geology, can be put into 3-D mode.

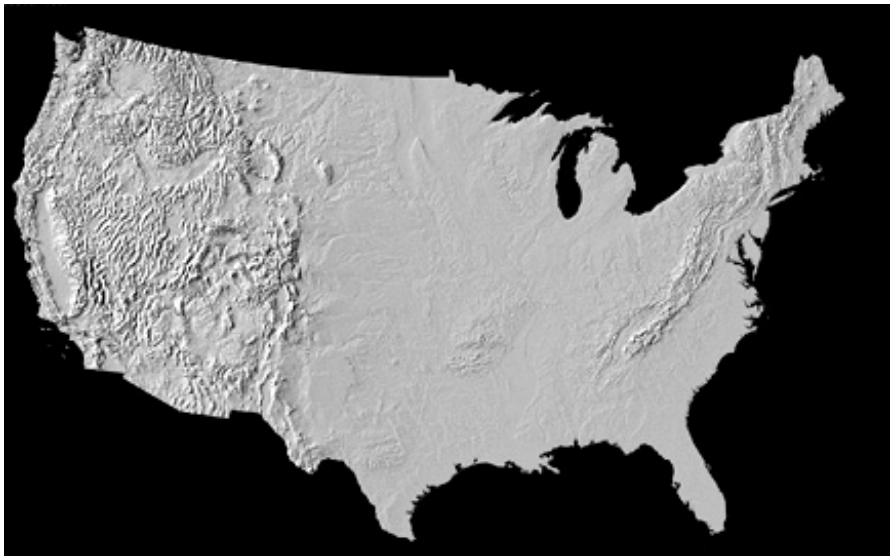
Examples of these DEM-based products abound on the Internet. We show just a few that are typical:

Check first this gray level relief map, constructed from 1:250,000 DEM data, of the Susanville area in the Sierra Nevada Mountains:

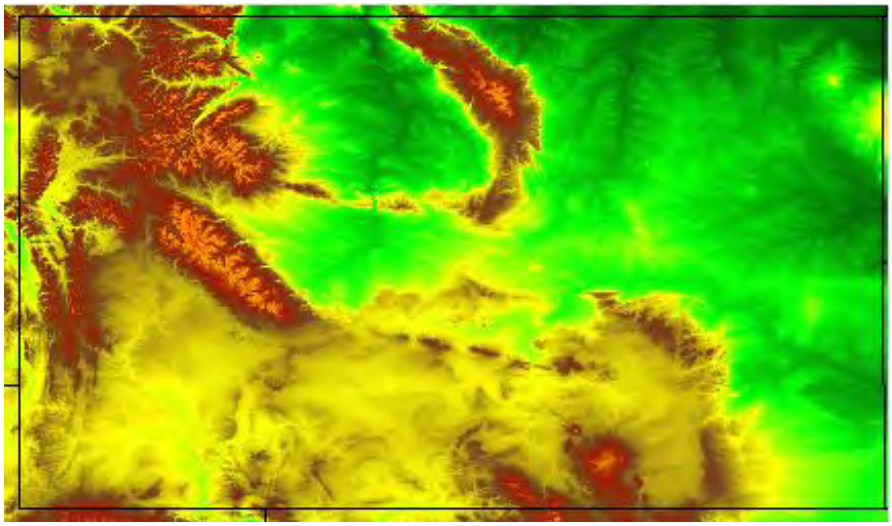


11-15: Describe what you think you are seeing in the map. ANSWER

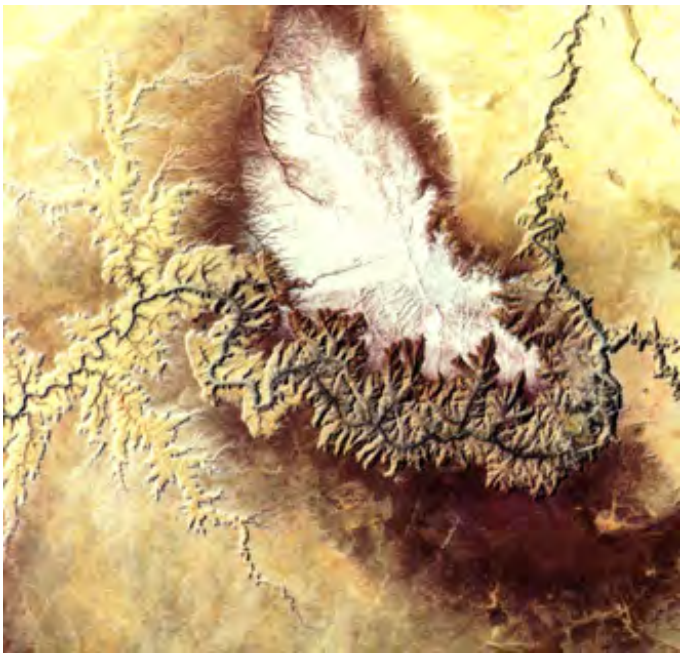
Next, examine the variant, in which shadow effects emphasize relief, by switching to a grand scale view of the entire conterminous U.S. Locate your home area.



An excellent example of a map in which the elevations are color-coded is this DEM version showing the state of Wyoming. The data source is a series of digitized 30' topographic maps. Lower areas are in green, and higher are in yellows, then brown.

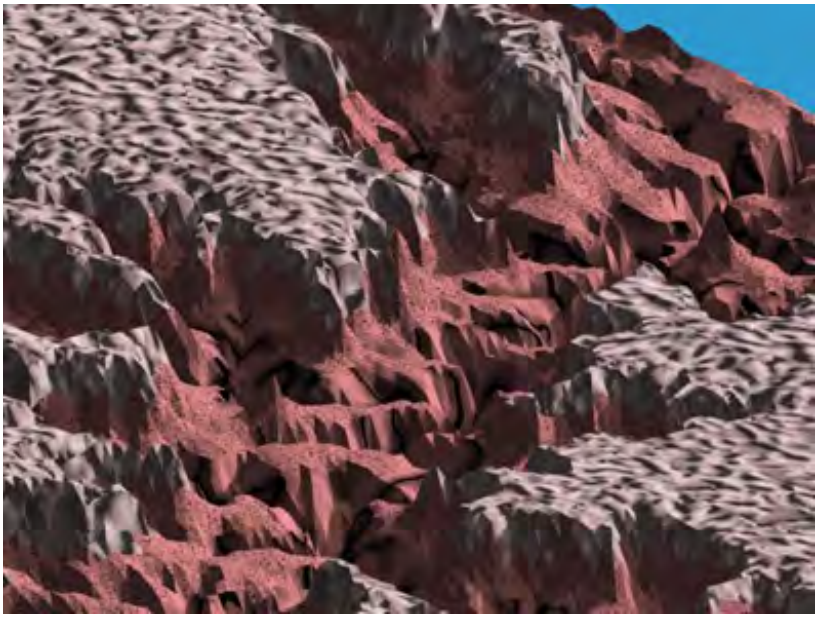


Let's now peer at a spectacular scene, looking at the Grand Canyon, as shown in a perspective view developed from DEM data and artificially colorized to resemble the rock units. But first, to gain familiarity with this incredible "ditch," carved over millions of years by the Colorado River, examine this Landsat nearly full 1972 scene. The snow-covered area is the Kaibab Plateau, some 2000 feet higher than the South Rim.



Next, a later view of the west half taken (November 1,1993) by the Japanese JERS-1:

Now let's look at the DEM rendition:



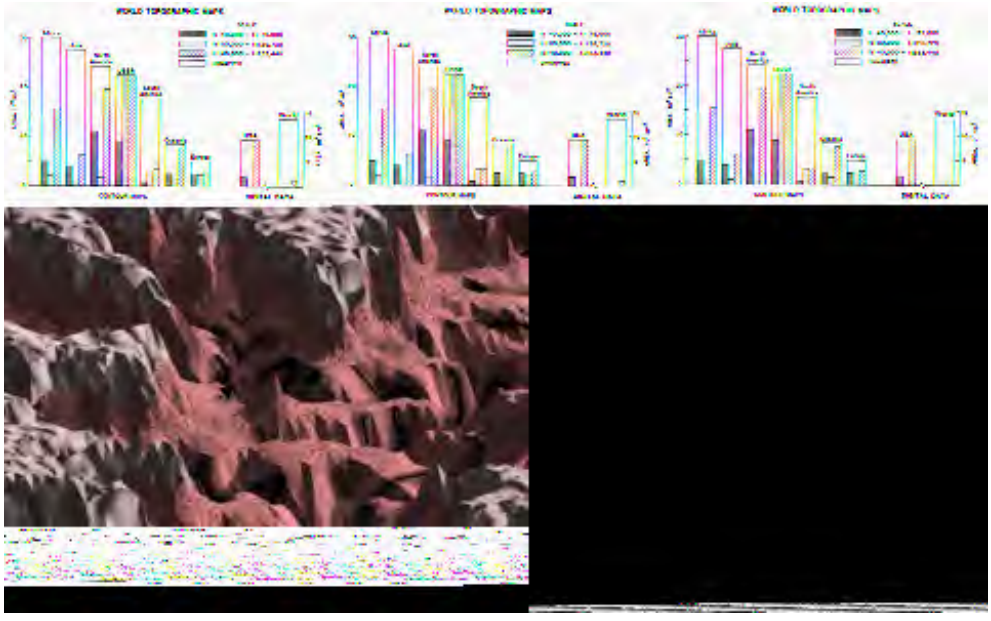
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Making topographic maps from aerial photogrammetric methods, assisted by (or in olden times, done exclusively by) ground surveying, is a slow, costly, and limited enterprise. Even into the 1990s, many parts of the world are still inadequately mapped. In recent years, the advent of global-covering satellites hosting instruments that can produce reasonably high resolution stereo imagery and can perform altimetric measurements to provide actual height calibrations has opened up the world to mapping nearly all the land surface to small- to medium-scaled levels. As a major aid in locating points on the ground with high accuracy, the Global Positioning System (GPS), which is briefly described, has evolved into a boon for surveying as well as a means for the citizen community to have the means to locate themselves as they move about the countryside (e.g., travelers, hunters).

Space-based Topographic Operations

After this preamble, we will now concentrate on the ever-expanding role of space measurements of the Earth's topography. This information has wide applications: in geology (e.g., correlating with plate tectonics effects); in geophysics (e.g., distribution of isostatic imbalances); in climatology (e.g., orographic barriers), in hydrology (e.g., drainage basin characteristics); in glaciology (e.g., ice sheet thicknesses); in ecology (e.g., ecozone disposition); and in planetology (e.g., comparative global hypsometry [frequency distribution of elevations]). Many of the operations hitherto performed by ground surveying and aerial photography can now be conducted from satellites hosting instruments suited to determination of surface variations. The end result over time will be a significant improvement in the availability and quality of maps at various scales, especially for regions of the world in which these continue to be lacking. It is surprising to realize that worldwide mapping, as of 1988, is very spotty and incomplete, as evidenced in this chart:



Programs and missions have been proposed, and some are being implemented, to apply space technology to eventually acquire digital data sets that provide global coverage of elevations and positions at several scales. These would lead to topographic maps having horizontal resolutions from 100 meters to a kilometer (cell sizes) and vertical resolutions from 1 to 10 meters (contour

intervals). High resolution regional coverage (on order of 100 m horizontal and 1 m vertical) is technically achievable even now. Similar resolutions are needed for sea state measurements.

Global Positioning Satellites (GPS)

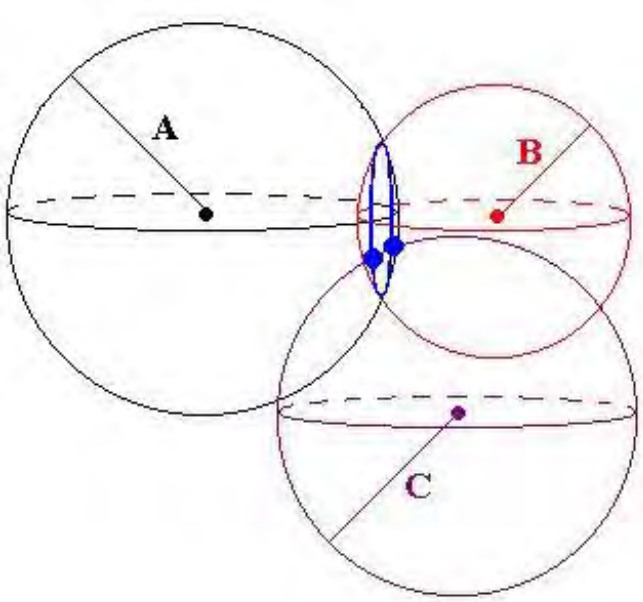
An extraordinary geographic locating technology available to everyone - including the "man in the street" - is the Global Positioning System (GPS) operated by the U.S. Dept. of Defense. U.S. government policy has made this satellite system available for civilian use - hence GPS has become almost routine in daily use by anyone. This system utilizes 24 (21 active; 3 as spares) small satellites each transmitting radio signals at the same two frequencies (L1 = 1575.4 Mhz; L2 = 1227.6 MHz). Each satellite follows a circular orbit inclined 55° to the equator at an altitude of 20,200 km. This constellation of satellites is placed, four each, within 6 orbital planes spaced 60° apart. A GPS satellite and the system orbital configuration are pictured here:

In this configuration, each satellite reoccupies the same track every 24 hrs. At any instant, a signal receiver positioned anywhere on the ground (or on the ocean or in the air [the system can function in a moving platform to yield locational information and thus is suitable for navigation]) should have at least five, and up to 8, of these NAVSTAR satellites in several orbital planes in radio "line of sight", each being at some specific different distance from this station.

One advantage of relying on these outer space platforms as signal sources is that surface obstructions that would interfere with ground-based platforms are eliminated. Accurate location is accomplished by a three-dimensional variant of the surveyor's method of triangulation (fixing on three separated points). The NAVSTAR radio signals spread out in spherical wavefronts so

that those from any three satellites (as points) in line of sight can be visualized as intersecting at the receiver. Its location now becomes a matter of determining the radii (distances) of the signal spheres. A fourth satellite aids in determining specific elevation, and a fifth further increases accuracy. The Russian space program has developed a similar system, GLONASS, now operational.

Location is established by pseudoranging, which determines distance to each satellite (which transmits its own unique identification as a pseudo-random-noise [PRN) code) by analyzing the radio signals that travel at the speed of light. The signal contains modulated time-tagged data bits that permit transmission (hence travel) times to be determined as the difference between time of origin and time of arrival. To establish ground location a variant of the classical triangulation method used in surveying is applied, as indicated by this diagram:



If signals from just three satellites are received, the location solution involves two possible points - one correct, the other false. To resolve this positional dilemma, a signal from a fourth satellite provides a solution that is unique. The position of each satellite must be closely fixed at any moment by knowing its ephemeris (orbital parameters that determine where a satellite is relative to some ground reference), which can be calculated from signals sent to a network of control stations. There must also be high precision atomic clocks (two each of cesium and rubidium) on the satellite and a corresponding clock in any ground receiver whose location is desired. Electronics and computer processing at the receiver facilitate calculation of Coordinated Universal Time (UTC) (effectively, Greenwich Mean Time) which can then be adjusted to GPS Time, so that time differences between corresponding signals from the several satellites can be determined with high exactitude.

Civilian-usable signals sent in the Coarse Acquisition (C/A) mode, confined to L1 output, are available to anyone using a receiver costing as little as $500. This setup leads to a standard accuracy of ~ 100 m horizontal and 156 m vertical. Using more expensive (up to $40,000) receivers, operating in the Precise (P) mode, L1 and L2 (the latter also is used to correct for ionospheric interference), accuracy under optimal circumstances can reach 17.9 m horizontal and 27.7 m vertical; this mode is restricted to military and approved scientific use at present. Several specialized techniques have been devised to reach this accuracy, including Code Phase Tracking (for navigation) and Carrier Phase Tracking (for surveying) which requires two receivers working in tandem; another uses differential GPS corrections. Some of the commercially available handheld GPS receivers are shown in the next illustration; receivers can be built into autos as an aid to determining location of the traveling vehicle or of any destination sought.

The upshot of GPS on the world stage is that ground survey points can be set up merely by aiming the receiver rather than resorting to theodolite transits (which, however, remain superior for very precise surveying). There are many entries that further discuss/describe GPS on the Internet; just enter "GPS" into your search program. Here are three we have found useful: (1), (2), and (3). Also, there is additional discussion of elements of GPS at the bottom of page 13-4.

11-16: Mention two civilian uses of the global position system. ANSWER

Note: A new CD-ROM serving as an educational introduction to GPS has been prepared by NASA JPL, the USGS, SC/EC, and the University of California at San Diego. This CD focuses on one powerful scientific use: as a primary surveying tool to measure small and continuing changes in reference points whose gradual separations provide vital information on movements of the Earth's tectonic plates - this is the main focus of the NASA/USGS Crustal Dynamics program. Information on how to obtain it is given at this Web Site.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

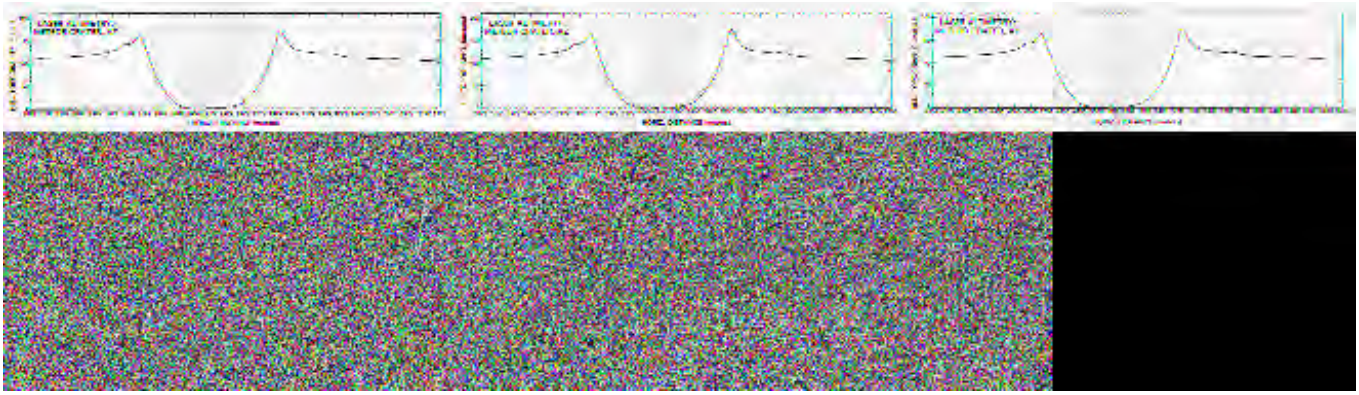
A great deal of valuable information about the (topographic state of the) Earth's land and sea surfaces can be acquisitioned by making profiles across these surfaces. This can be done by Altimetry. The conventional way in the past was simply to run survey lines across the surface. The profile serves as a control from which line of sight transits to other points results in a collection of located spots that allow contouring. Now, altimetry can be done conveniently from space using either radar (microwave) or lidar (light) pulses whose times from satellite to surface and back to satellite provide range information easily converted to distance from the satellite (whose position is accurately known at any moment) to the ground at any spot. The principles underlying this are summarized on this page.

Altimetry

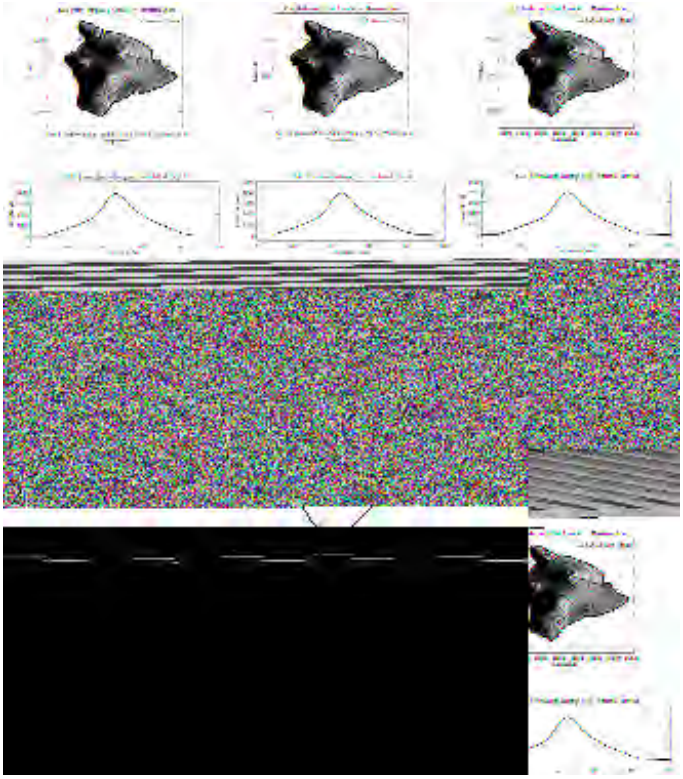
To consider altimetry, we must go from point surveying to linear surveying of the Earth's surface. Aircraft and space-based altimeter instruments yield direct measurements of elevations along narrow path footprints. We can readily plot these elevations as profiles. A series of parallel profiles serve as a framework for contouring (figure near bottom of page 11-2). Altimeters send self-generated (active) signals to reflecting surfaces and collect the reflected signals. Then they measure the total roundtrip times from the targets (solid land, tree tops, ice, or water) over which they move. The signals can be either radio pulses (radar included) or light pulses (laser).

Laser is an acronym for light amplification by stimulated emission of radiation. A typical solid laser device is a chromium-doped ruby (Al_2O_3). When an external source of radiation (as from an enclosing flash tube) acts on a shaped (usually cylindrical) ruby crystal, one end of which is silvered to act as a mirror, the chromium (Cr) ions dispersed in its lattice are excited to a new energy state (electrons raised to some new orbital level). Stimulated emission generates the laser state when the electrons drop back to a lower energy level (remember, from the Introduction, the formula: $E = hc/\lambda$). Much of the light passes sidewards out of the crystal, but light moving along the axial zone of the cylinder encounters other Cr ions as it reflects repeatedly. Multiple reflections further excite the ions and build or amplify (by a process called optical pumping) the light energy until it discharges as a pulse (controlled by the flash lamp) of intense coherent radiation at a discrete wavelength (actually a narrow frequency range; for the ruby laser the light is pinkish). That pulse sequence is collimated to form a unidirectional beam which can be aimed. Other laser materials include gallium arsenide and excited gases such as neon or helium. Both visible and infra-red wavelength light can be generated in this way.

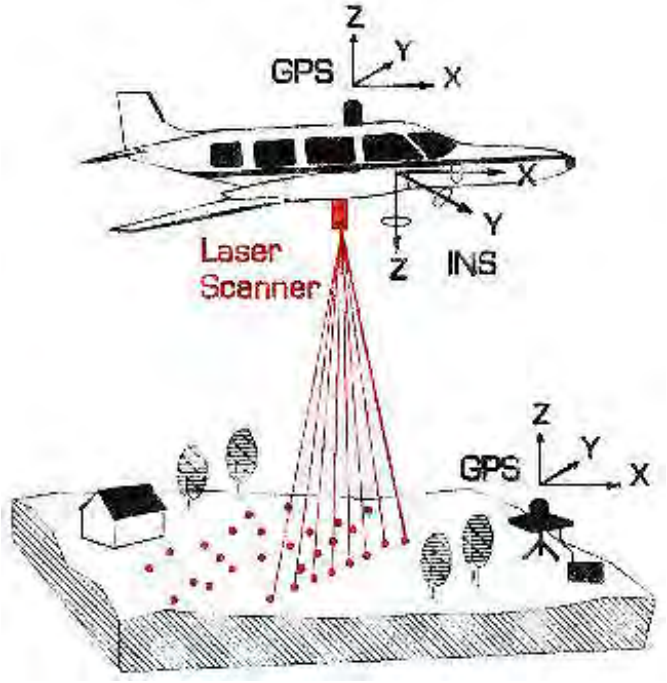
Laser altimetry has been conducted from aircraft platforms for several decades. Timing devices allow extremely precise determination of transit times, so that accuracies of a few centimeters in determining elevations (and their differences or relief) along the traverse are attainable. Aircraft should be scheduled to fly in good weather, offsetting the main disadvantage of using lasers coming from cloud interference with the light beam. An example is this profile over Meteor Crater in Arizona (see page 18-6) obtained from a NASA mission (courtesy: J. Garvin):



Laser altimeters are now being flown on space vehicles. The Orbital Profiling Laser Altimeter is a pulsed, time-of-flight optical (1.024 µm) sensor that sends 10 pulses per second (pps) in a narrow beam (footprint 30-150 m diameter; sampling in 150 to 700 m [492-2296 ft] intervals). Operated on the Shuttle Endeavor in January, 1996, it achieved a vertical precision of 0.75 m [29.5 inches]. Each laser shot fired has a dwell time of only 2-10 nanoseconds, within the 1-10 nsec resolution of the timing electronics; with this rapid return rate, ground positions are readily determinable provided space vehicle position is known. Here is a typical profile, obtained during a pass over the volcano Mauna Kea on the big island of Hawaii:



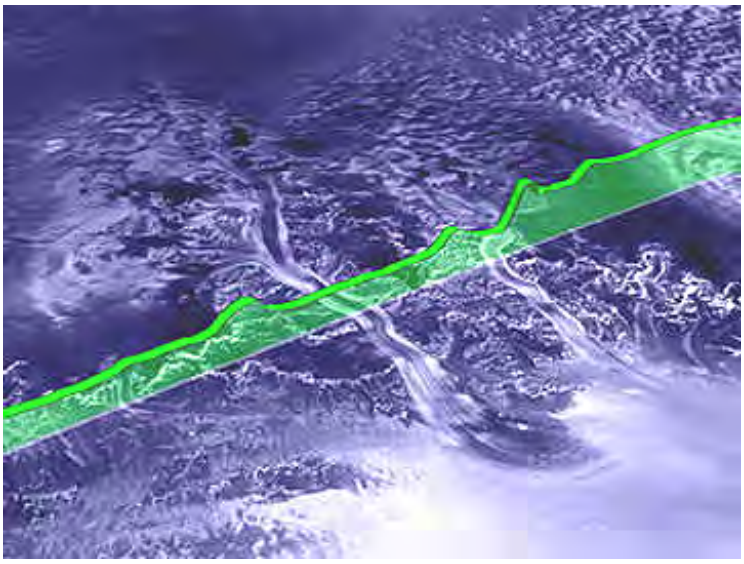
Laser beams can be directed in a scanning mode so that the entire ground surface is accessed at high resolution. The methodology is straightforward, as shown here:

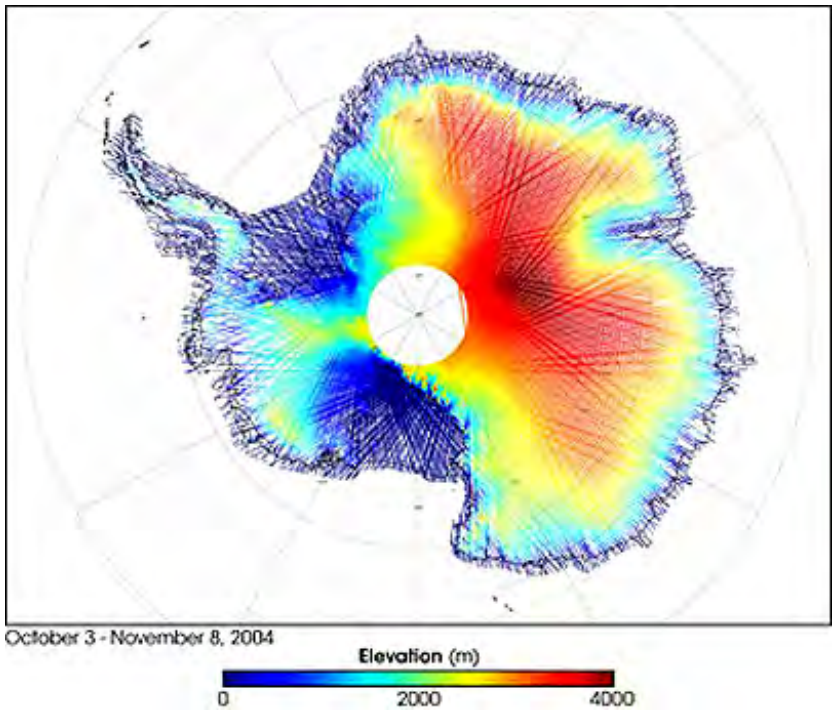


High resolution ground images are produced. These can be converted to perspective views, illustrated in this figure:

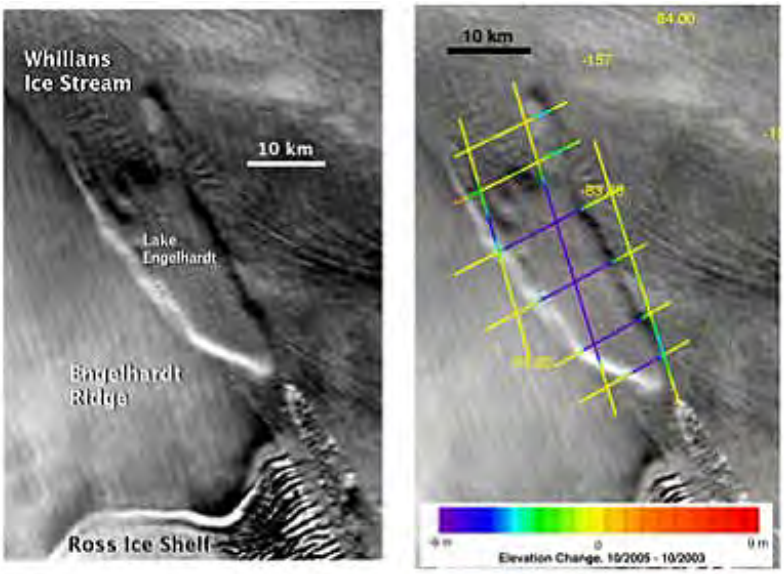
A dedicated satellite (IceSat) in the EOS series (see page 16-7), launched in the year 2001, mounts the Geoscience Laser Altimeter System (GLAS) designed to measure ice-sheet topography as well as cloud and atmospheric properties, but also surveys selected land and water surfaces. The 40 pulse/sec beam is generated from a neodymium:yttrium-aluminum-garnet (Nd:Y-Al-G) crystal that produces a two level emergy output: 1.064 µm (infrared) for surface surveys and a 0.532 µm signal for atmospheric (clouds and aerosols) measurements. Each pulse spreads over a ground spot of 70 meters (230 ft), with separation between pulses amounting to 170 m (558 ft). From its 705 km (440 miles) orbital altitude, the instrument can measure height differences on the ice of 10 cm (4 inches) precision; characteristics of the altered return pulse indicate surface roughness. More on the GLAS sensor is found on page 14-14.

GLAS has returned invaluable data on changes in the ice sheets covering Greenland and the Antarctic. This information is proving essential in estimating melting in the two largest water reservoirs on Earth. The next two images show profiles and profile lines over the Antarctic:



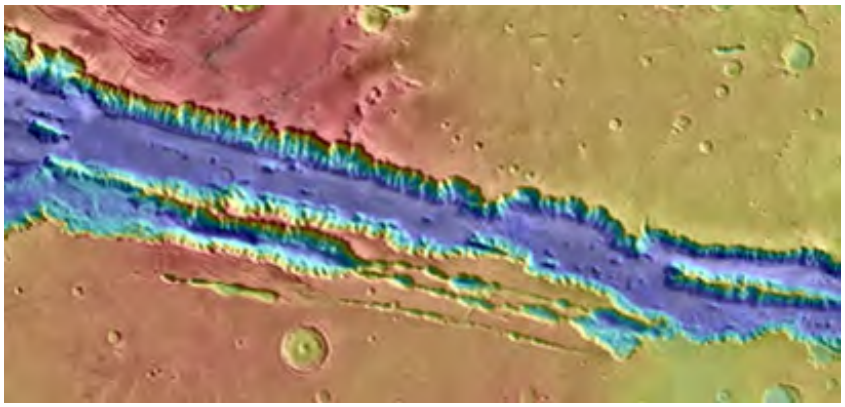


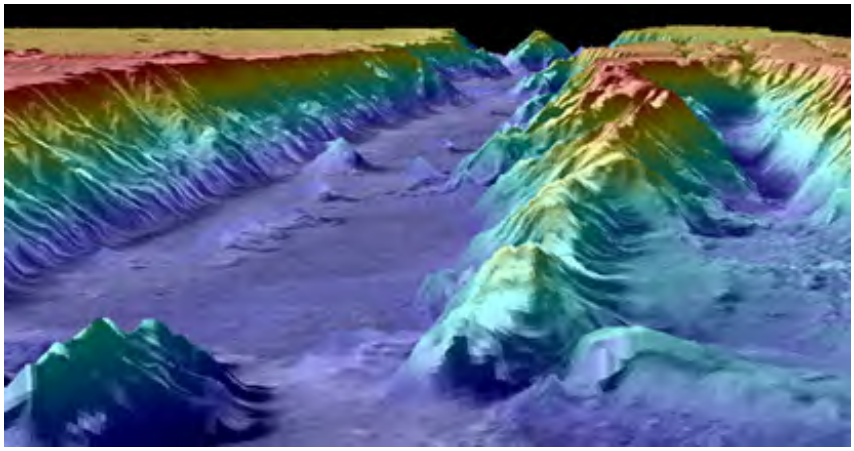
Temporal GLAS data have shown changes that have been interpreted as ice movements related to (previously undiscovered) subsurface lakes. Here is one data set:



Lasers are also reliable and very accurate tools for determining distances to satellites in orbit. Retroreflectors are small prisms (typically quartz) mounted on the spacecraft. Even at their high orbital speeds, these spacecraft can be targeted to intercept a laser beam pulse directed at them from Earth-based stations which is then bounced back by the reflectors. Along with GPS, this is one way in which the orbital position of a satellite can be precisely fixed. The astronauts in later Apollo missions left retroreflectors on the Moon's surface as a means of learning more about its orbital motions, including its recession from Earth.

The Mars Observatory Laser Altimeter (MOLA) has been launched on the Mars Global Surveyor that began to orbit the Red Planet in 1997. It has since imaged most of the planet's topography, producing stunning images (arbitrarily colored) as surface "maps" and as perspective views; two examples are shown here:

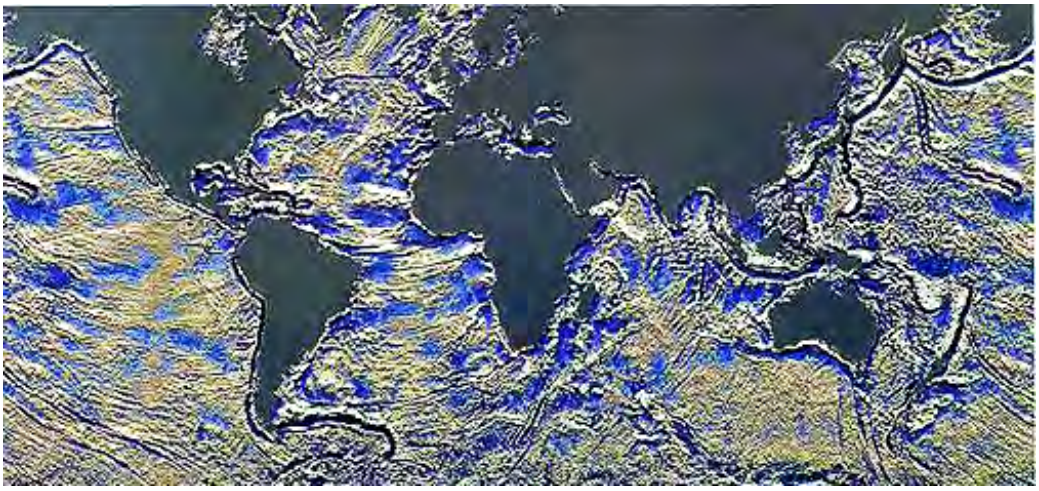




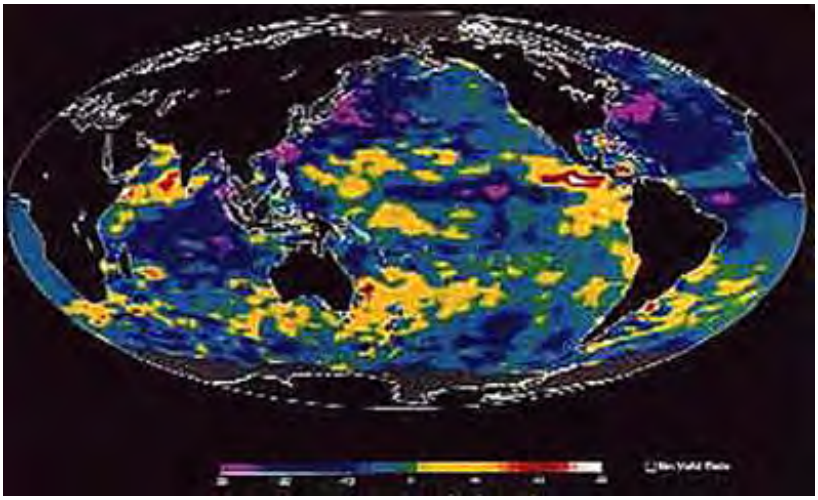
From space, real aperture microwave (radar) altimeters have served to measure various aspects of sea state, such as heights relative to the geoid, wave geometries, sea ice, and, indirectly, circulation characteristics such as currents and eddies. Like lasers, the signals are dispatched as short pulses, but at the longer radar wavelengths these signals can penetrate cloud cover. Again, using roundtrip times the instrument acts as a ranger to determine distance to target. Radar altimeters designed to secure data from the ocean surface use small antennas that dispatch long wavelength pulses through a wide beam that produces a broad footprint whose size is determined by pulse length. This pulse-limited type operates best on smooth surfaces but analysis of the degree of "stretching" of backscattered pulses (echoes) yields information on surface wave heights (roughness). For land measurements, especially on surfaces of high relief, the beam-limited altimeter requires a larger antenna capable of generating a narrow beam (hence, a smaller footprint that better discriminates changes in slope) and shorter wavelengths. The orbital position of the sensor platform must be determined with high precision and accuracy to establish the position of the geoid and local departures of the surface from it. Modifications of the signal by the atmosphere need to be accounted for, usually through corrections made from data acquired by an accompanying radiometer.

The first spaceborne altimeter experiment was conducted during the 1973 Skylab mission. Next was GEOS-3 (Geodynamics Experimental Ocean Satellite) in 1975 which measured height

differences greater than 60 cm. This was followed in 1978 by JPL's Seasat (which failed after 99 days but in the interim returned extremely valuable data over sea and land). Its altimeter, one of five instruments including a SAR, was capable of determining sea state, a term that includes measurements of wave heights, and also gross height variations of the sea surface (on which waves are superposed) as well as indications of wind speeds. Wave heights were determined with a precision of 10 cm. We have already seen one spectacular product of the correlation between altimeter data, gravitational field data derived therefrom, and ocean topography in the global seafloor map produced by the Lamont-Doherty research facility at Columbia University (page 8-7). However, it was such a triumph we will repeat it here for your closer inspection.

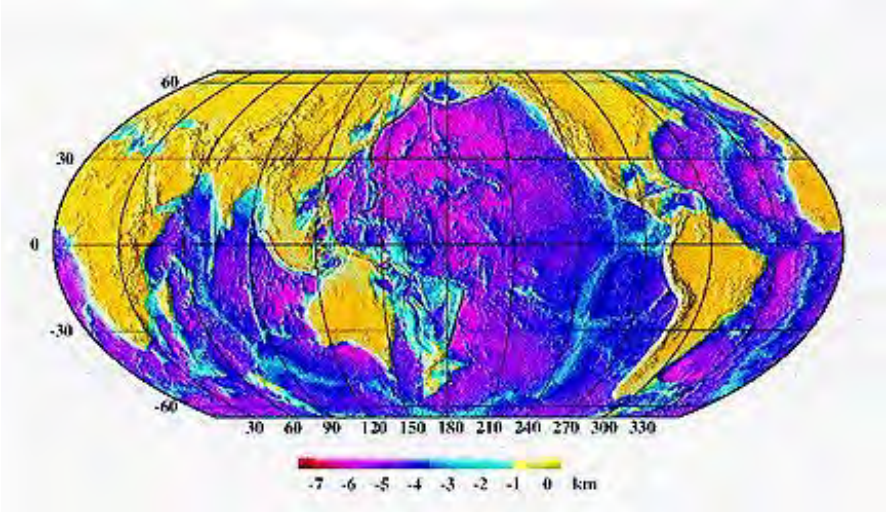


An instrument with performance similar to Seasat was mounted on the 1985 Geosat, launched by the Dept. of Defense as a geodetic surveyor. That same year planning and development was begun on the TOPEX/Poseidon mission, operated jointly by NASA/JPL and CNES (Centre National d'Etude Spatiales) which was eventually launched with great success in August, 1992. Its dual frequency altimeter (13.6 Ghz and 5.3 Ghz [to make ionospheric corrections]), and a three-channel mirowave radiometer (to make tropospheric water vapor corrections), execute an average of 50000 observations/day over a latitudinal band of ± 63.1°. Wave heights as small as 4.3 cm (1.6 inches) can be detected. TOPEX/Poseidon has produced global maps that show seasonal variations in sea level brought on by the expansion of water from changes in regional temperatures (the outer 3 meters (~10 ft) of ocean water contains as much heat as the entire atmosphere). An example appears here:



These data sets can also be interpreted to follow changes in ocean current circulation (generally poleward currents) controlled by thermal upwelling, winds, and the Coriolis force, as well as effects of tides, development of eddies, and monitoring of El Niño (the equatorial shifts of waters owing to thermal influences). All of these observations bear on major aspects of global climate changes.

More recently, altimeter data for the entire oceanic realm were acquired during the ERS-1 Geodetic mission. Coupling altimetric measurements with satellite orbital variations (changes in acceleration owing to gravitational differences from the Earth underneath) allows development of a marine surface gravity field. When that is inverted, the result is a "map" of the ocean floor in terms of depth (bathymetry) as represented by variations in water thickness between surface and ocean floor. Here is a bathymetric map from ERS-1 for the entire globe.



Space imagery can be acquired in modes suited to stereo viewing and, in principle, to analysis and contour mapping by stereoplotters. Radar images are capable of displaying the 3-D effect but this is usually not as "realistic" as aerial photography owing to layover, range effects, and other distortions (see page 8-4).

11-18: Why hasn't altimetry from space as yet replaced conventional methods for making topographic maps? ANSWER

A solid review of radar altimetry is given at this ESA website

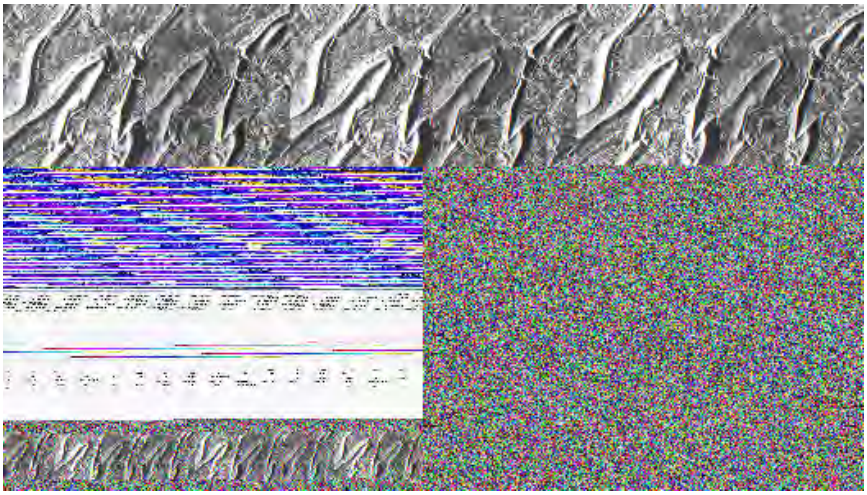
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

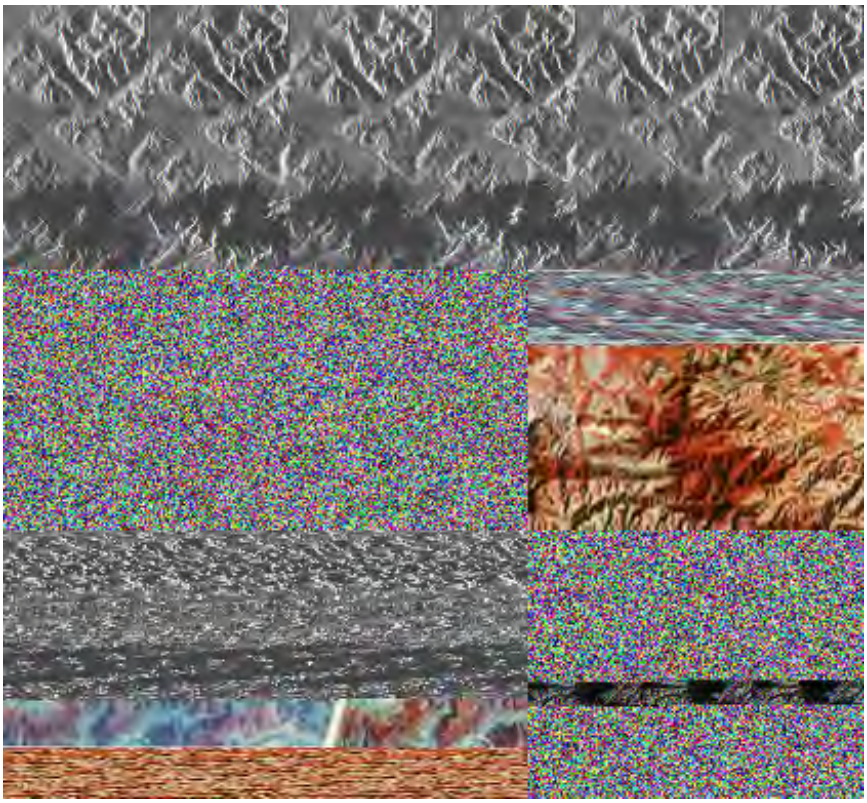
Radar stereo has been mentioned before in this Section. Again, the crux of its workability is, like aerial photography, based on getting two images that are offset. This can happen through multiple passes, the second time moving over nearly the same ground surface as the previous pass. It is also possible to make a usable stereo pair by acquiring images at two incidence angles. A stereo pairing is achievable during a single pass by having two separate but coordinated radar transmitter/receiver systems operating simultaneous from their moving platform. The radars can be of the same band or different bands. The returned signals will show phase differences that can be analyzed by a mathematical technique known as interferometry, which works well when the signals use coherent radiation. Examples are shown.

Radar Stereo; Interferometry

When flown on aircraft, a radar can point in different look directions and angles during successive passes to simulate the necessary parallax conditions for relief displacements across track. Side-looking Airborne Radar (SLAR) observations, looking out from the same side in successive passes, produce effective stereo imagery. An example showing the Appalachian Fold Belt in central Pennsylvania, using a SAR X-band system, demonstrates the quality of radar stereo.

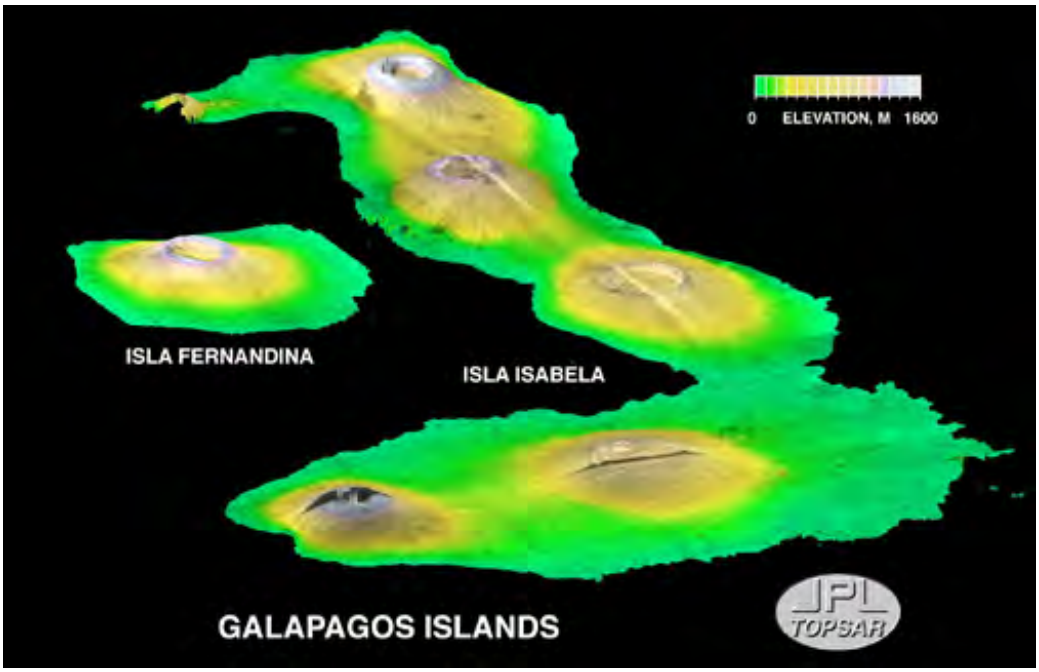


Stereo radar images from space can produce acceptable stereo models using multiple look angle (off-nadir) viewing. Vertical and horizontal resolutions are hard to achieve by optical systems from comparable orbits. We can significantly reduce the vertical exaggeration because of low B/H (base to height ratio) values. The stereo pair below, showing volcanic terrain in Chile, acquired with the SIR-B L-band SAR from two incidence angles (54° and 45°), has moderate exaggeration of a high relief scene.



11-18 Is a low VE (vertical exaggeration) necessarily a good thing? ANSWER

Multiple incidence angle imagery serves well in preparing perspective views, as demonstrated in this SIR-C oblique rendition of a group of volcanic calderas on two of the Galapagos Islands.



Synthetic Aperture Radar images (see Section 8), generated simultaneously from two antennas on the same platform, have such a small base that they do not, in themselves, function as a stereo pair. But being composed of coherent radiation, signals from each, matched to equivalent pixels in the other, show varying degrees of phase differences, indicative of partial interference. Variations in the signal travel times between two adjacent points induce the interference, and thus, show slightly different heights. When we remove Earth curvature effects and make other adjustments, residual differences become a measure of terrain elevations.SAR Interferometry is a developing technique that you can learn more about by perusing another University of Texas review dealing with this subject.

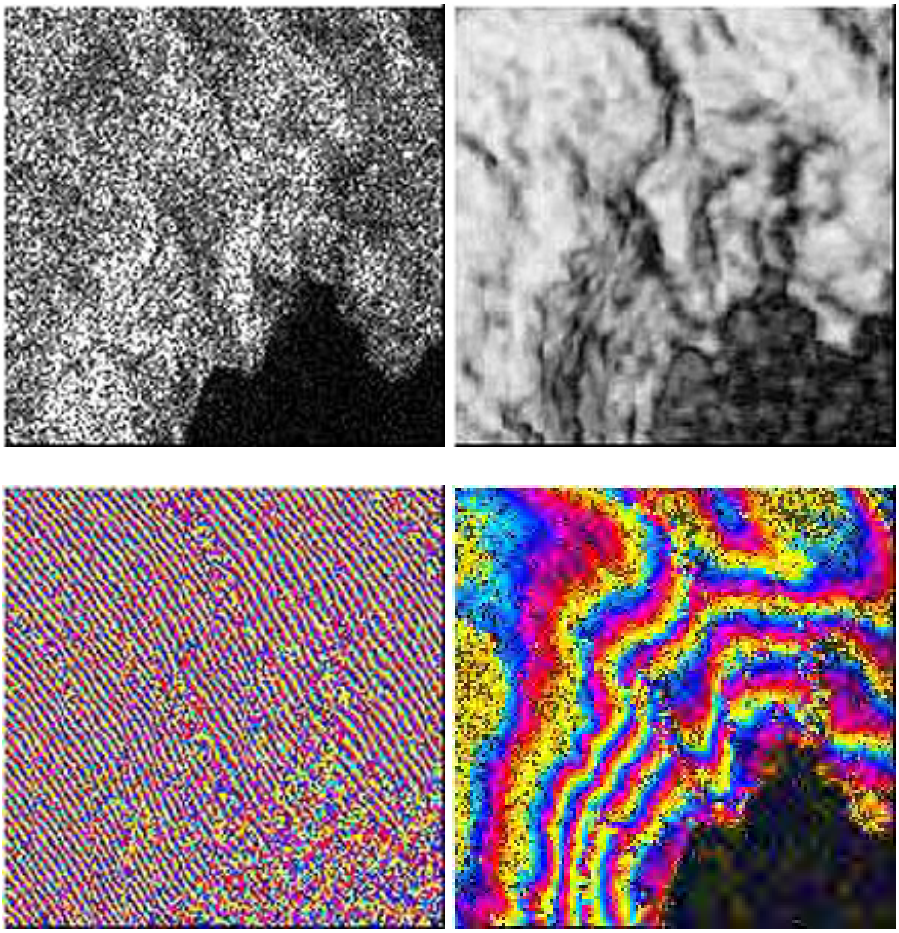
Interferometry will crop up again elsewhere in the Tutorial (for example, in Section 20 dealing with space telescopes). The principles underlying interferometry are treated in two Wikipedia websites: Basics of Interferometry and Optical Interferometry. A mathematically-based exposition of interferometric principles is embedded in this JPL site.

The next two paragraphs were extracted verbatim from a CCRS Radarsat web site:

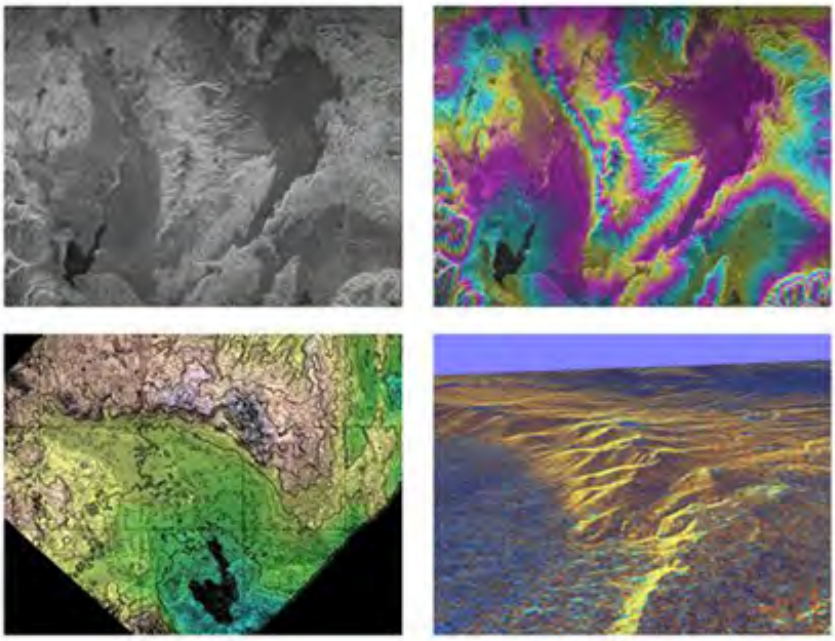
"Interferometric SAR (InSAR) techniques make use of the coherence of the radar signal and the fact that the signal phase is is equal to twice the path length between the sensor and the earth's surface. The phase difference between measurements generated from two SAR images with the sensor separated by a baseline, allows measurement of the slant range difference to fractions of a radar wavelength. The slant range difference can be geometrically related to the terrain height."

"There are two basic SAR interferometry methods. In the first, two antennas are placed on the same platform and simultaneously acquire images of the scene from two different angles. The relative phase difference may then be used to construct a digital elevation model. The CCRS airborne SAR was modified to operate in this mode. In the second, a pair of images from the same sensor are taken at different times. This is now a well-proven concept and has been demonstrated with several spaceborne SAR sensors, now including RADARSAT. For this repeat-pass interferometry, the scenes are acquired at different times, so there is a time difference as well as viewing geometry to consider. The passes must have rather similar geometry in order to allow extraction of the relative phase difference. This usually requires that the satellite be on an exact repeat orbit. For RADARSAT, this means that a candidate interferometric image pair is available for a particular site every 24 days. The difference in geometry allows the extraction of topographic information, in the same way as with the single pass airborne system."

These four images, also taken from the CCRS website, show steps involved in producing an interferogram image of part of Bathurst Island in northern Canada. Read the captions for a brief identification of each step:

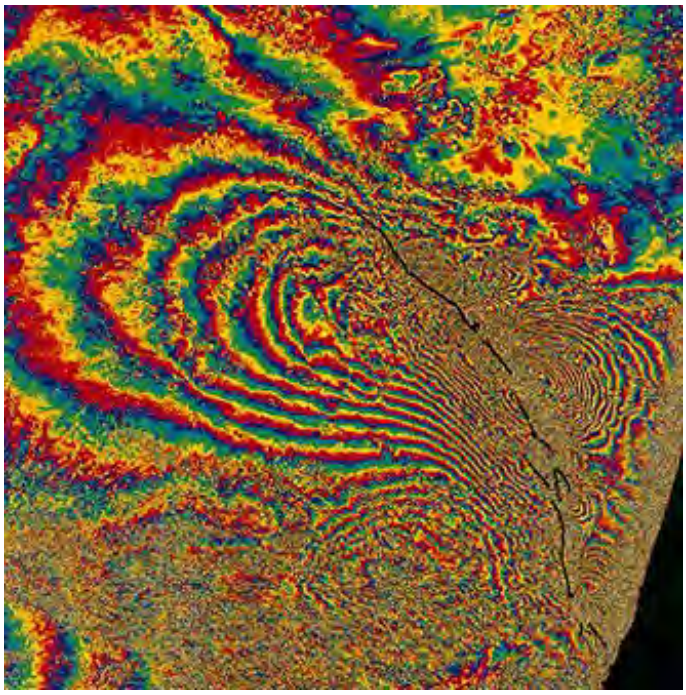


You can get some idea of these principles from this illustration showing how SIR-C interferometry produced topographic information for this Long Valley, CA test area.



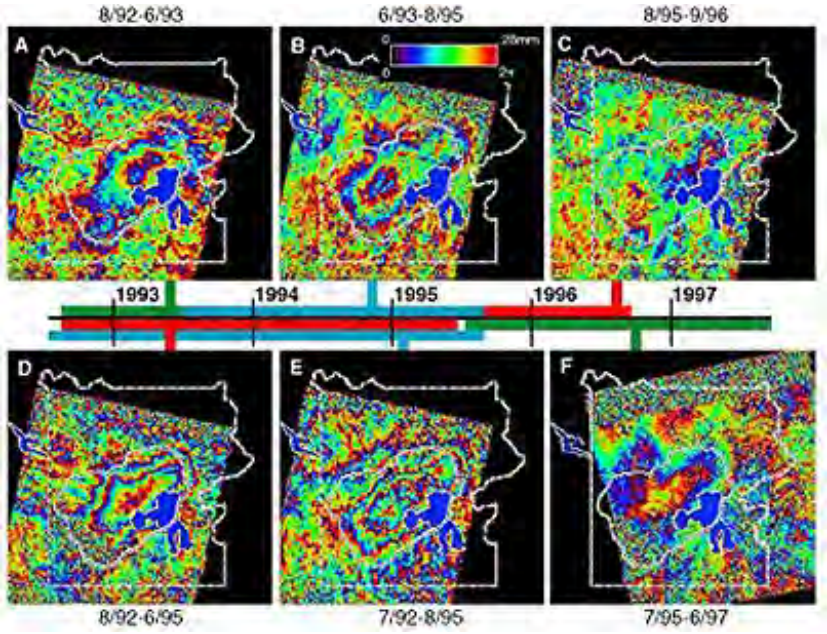
The upper left panel is a Horizontal-Horizontal polarized L-band image, in which many bright-toned patches correspond to hilly rock and dark patches to smooth valley floors. Lake Crowley (dark) is at the lower left. The ridge is part of the Long Valley volcanic caldera (inactive). The upper right panel shows an interferogram derived by combining two L-band images taken in April and October, 1992, respectively. The colors relate to differences in signal phase, caused by elevation differences. In the lower left is a topographic map derived from the interferogram. Its total relief is 1,300 m (4264 ft). The lower right panel presents a 3-D perspective of this map, made by "draping" a C-band image over the L-band topographic data.

Here is another interferogram that shows the color band pattern around the Landers (California) fault line (superimposed). Each color cycle represents a rise in elevation of 2.8 cm.

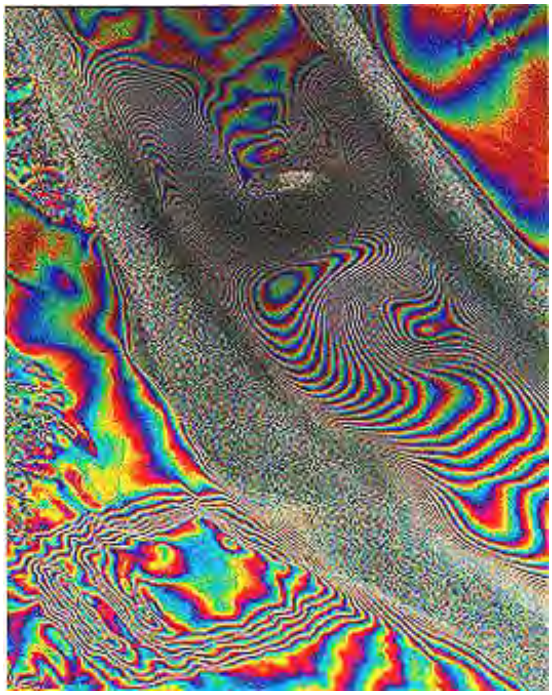


Radar interferometry using SAR data has been monitoring Yellowstone National Park. The hot region (responsible for the heated geysers) is associated with an underlying caldera - active in

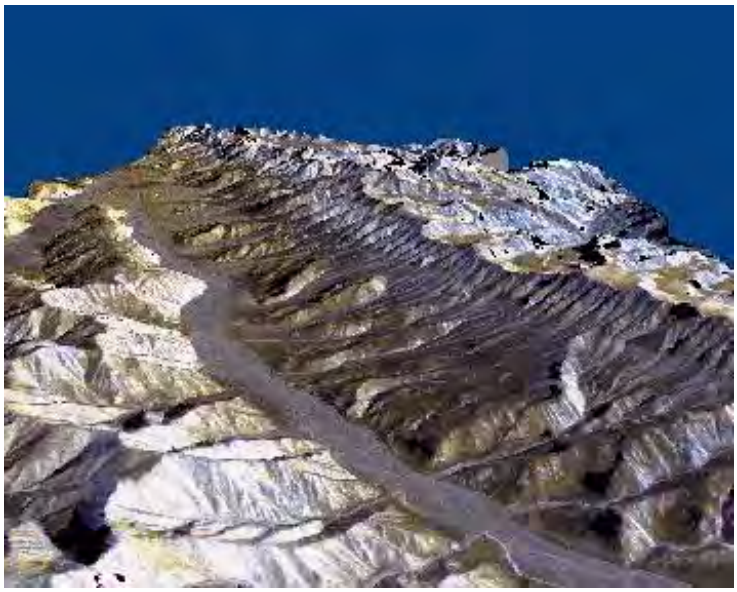
the last million years. The region is rising, as shown in the interferograms calculated for the six periods shown:



Radar Interferometry monitors other kinds of natural phenomena involving elevation changes and other aspects of motion. Here is an interferogram for a glacier in the Himalayas:



As with so many other kinds of space images, the combination of a plan view image with elevation data - both acquired by SAR interferometry - permits construction of perspective views. This one shows the Karakax Valley in Nepal:



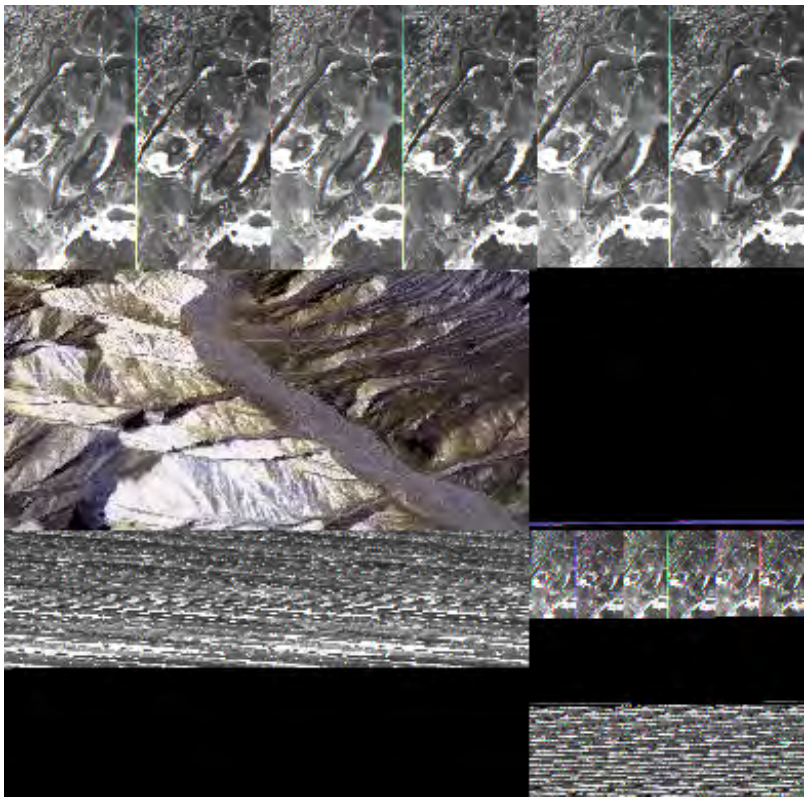
Primary Author: Nicholas M. Short, Sr.

On this page are additional examples of stereo pairs acquired by Landsat, the Large Format Camera operated from the Space Shuttle, and SPOT.

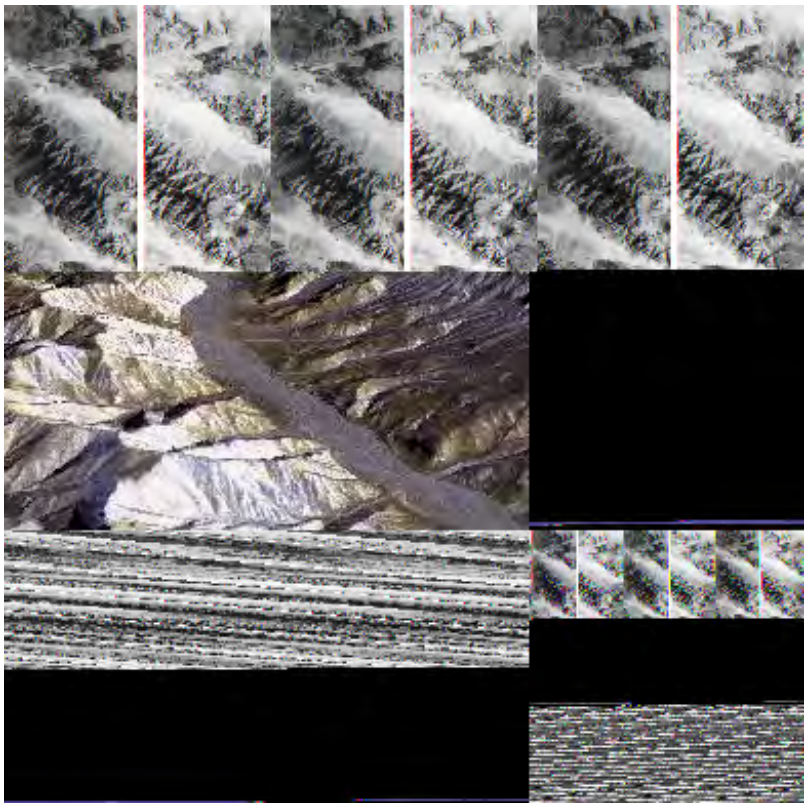
Additional Examples of Stereo from Space

The stereo pairs shown on this page can be printed out by bringing up this (STEREO2) page.

We can digitally convert single Landsat scenes (and other space imagery) into pseudo-stereo by merging the Landsat pixels with registered DEM data. Then we can manipulate that scene into view pairs having different parallax. Consider this left-right example in Morocco:

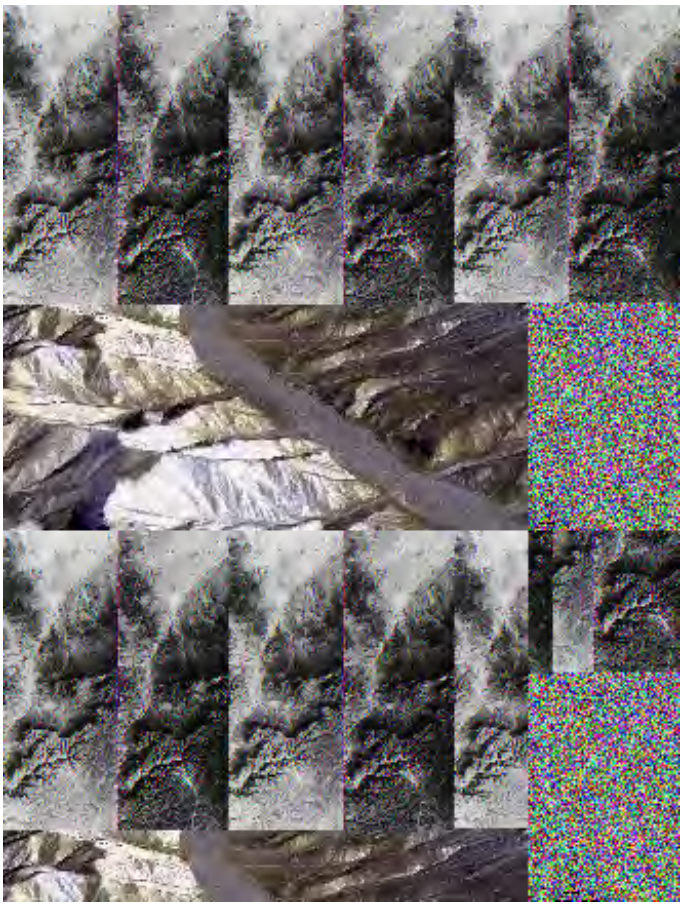


Astronauts using the Large Format Camera (LFC) (see page 12-4) on Space Shuttle Mission 41-G, in October, 1984 (and subsequently), at an altitude of about 300 km (186 mi) snapped (pre-planned and targets of opportunity [astronaut's choice of the moment]) 2,160 black and white and color/color-IR photos. The photos comprise a 23 cm by 46 cm (9.2 by 18.4 inches) area, with ground resolutions between 14 m and 25 m (46-82 ft). Taking these in quick succession allows along track forward overlap of between 20% and 80%, from which we attain excellent stereo viewing, at a B/H of about 1.2. View the two strips below, cut from an LFC pair, that cover the Panamint Mountains and Death Valley in eastern California. The European Space Agency has flown a similar Metric Camera.



Adapted from T.E. Avery and G.L. Berlin, Fundamentals of Remote Sensing and Airphoto Interpretation, 5th Ed., Fig. 5-12, © 1992. Reproduced by permission of Macmillan Publishing Company, Indianapolis, IN.

The two High Resolution Visible (HRV) scanners on the SPOT satellites (see Section 3) can tilt sidewards up to 27° in either direction. Not only does this tilting permit wider coverage during a single pass (one of the HRV scanners can stay in the vertical mode), but we can combine the off-nadir view with another view (either on- or off-nadir) taken on another date to give rise to a stereo pair. We show a plateau on the Libyan/Tunisian border below from a panchromatic mode stereo pair acquired on February 23, and 25, 1986, at an off-nadirs angle of 24°E (right image) and and 10°W (left image).



For those who want to see a striking stereo pair in color, taken by SPOT about 3 1/2 weeks apart, showing the Great Rift Valley of Kenya (note the step fracture zones), we have linked these (Make the left image on the screen the left in your stereo setup), which you can choose to print on your color printer to examine either under a stereoscope or with your unaided vision. Again, you may have to move one or both until some segment in common to each is in proper position to induce the stereo effect.

Visible-IR scanners and SAR imagers are now flying on spacecraft such as ERS-1 and ERS-2, JERS-1, and Radarsat, among others. Laser and radar altimeters are planned for future flights. One prime task is to gather global sets of data that have a wide variety of topographic applications. An ultimate goal is to fill the current gaps in mapping the Earth's land surface at more informative scales. A second objective is better resolution profiles and elevation models of sea states within the major ocean bodies.

Primary Author: Nicholas M. Short, Sr.

Since the first man in space (Yuri Gagarin of the USSR), cosmonauts and astronauts have taken photos with cameras pointed out of their spacecraft windows. The writer (NMS) recalls his conversation with Scott Carpenter (the second American in space), who was my passenger being driven to a Conference I organized where Carpenter was the keynote speaker, about that Mercury flight. He said he got so excited by the view out the window that he spent more than the allotted time for picture taking, acting really as a space tourist fascinated by the scenes below, prompting a plea from Mission Control to get back on track. The background to the long, well-organized U.S. programs to survey the Earth with a camera (which was also done on the Moon) is given on this first page. The Mercury and Gemini programs are described, with representative pictures.

THE HUMAN REMOTE SENSER IN SPACE:

ASTRONAUT PHOTOGRAPHY

Guest Writer: Dr. Paul D. Lowman, Jr.

Code 921, Goddard Space Flight Center
Greenbelt, Maryland 20771

History of Photography during Manned Space Missions

When one mentions "photograph" it is normal to understand that this word almost automatically implies "an activity involving (requiring) humans". We saw in the Introduction that humans have entered space on such vehicles as rocket-propelled capsules, Apollo modules, Shuttles, and Space Stations. From these platforms, astronauts and cosmonauts have acquired innumerable photographs as part of their scientific tasks and as "just plain tourists" looking out the window at the landscapes below.

Every U.S. manned mission since 1962 has included a look back at Earth by film photography, generally with hand-held cameras. This photography has been one of the most productive but least-known products of the U.S. civilian space program. This Section briefly summarizes the history and current status of such photography, focusing especially on geologically useful targets, i.e., land areas. We present the photography according to each program, although photography is rarely more than an incidental activity on most *manned* missions. (Women routinely fly on NASA missions, so, euphemisms for "manned" are not necessary.) We add one example from the Soviet/Russian missions.

12-1: **To get you into the mood to think "astronautical", who was the first man in space, the first American in space, the first American to orbit Earth, and the first "dynamic duo" of Americans to orbit the planet? ANSWER**

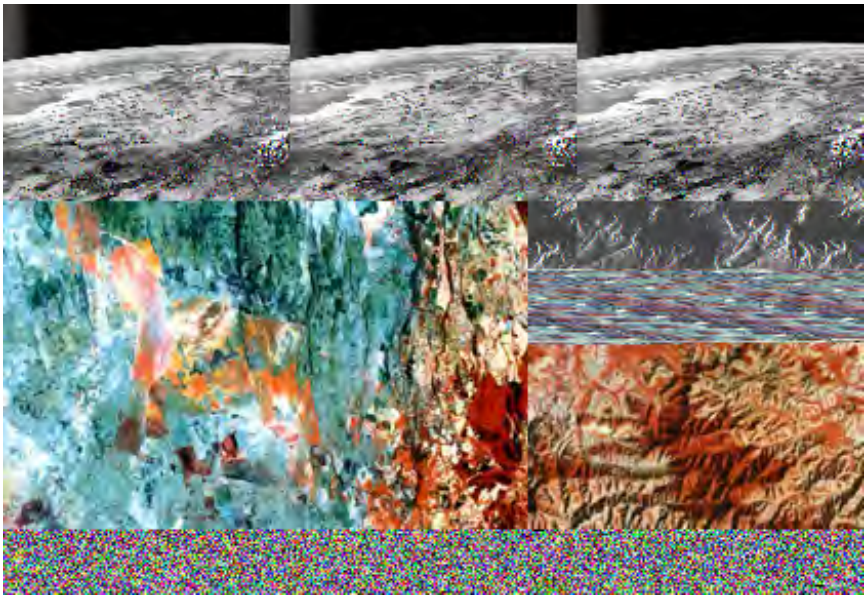
For the missions in general, the most readable and authoritative account is a book entitled, "Liftoff" (1988) by one of the astronauts, Michael Collins of Apollo 11 fame *. A general reference covering all of American space photography through 1989 appears in the journal Geocarto

International, vol. 4, no. 1, 1989. Collections of astronaut photographs prior to those shot from the Space Shuttle are the mainstays of two books, "Space Panorama" and "The Third Planet" by the author (PDL, Jr.) *.

To facilitate readability, we cite references by placing a blue asterisk near or next to individuals involved, such that the reader can access the full reference optionally. We list only the most important references this way. The literature on the subject is now quite extensive. We follow the same approach, cited by a double asterisk (**), for several tables that list photo equipment used in various missions.

Internet access to information about most of the U.S. astronaut missions, along with selected photos taken during these flights, is available at http://images.jsc.nasa.gov (select the Press Release option button).

We can trace the beginnings of space photography to automated cameras onboard V-2 rockets fired in the mid 1940s from the White Sands Proving Grounds in New Mexico. From then, through the early days of NASA, cameras documented scenes of the Earth made during various sounding-rocket and missile launches. An example is this photo taken by an automatic K-12 camera, using black and white infrared film, from a Viking sounding rocket that reached a height of 227 km (141 miles). This film was recovered from the crashed vehicle after it fell back to Earth. The scene viewed here extends southwestward across parts of New Mexico, Arizona, Nevada, California, and northwest Mexico (upper Gulf of California on the left).



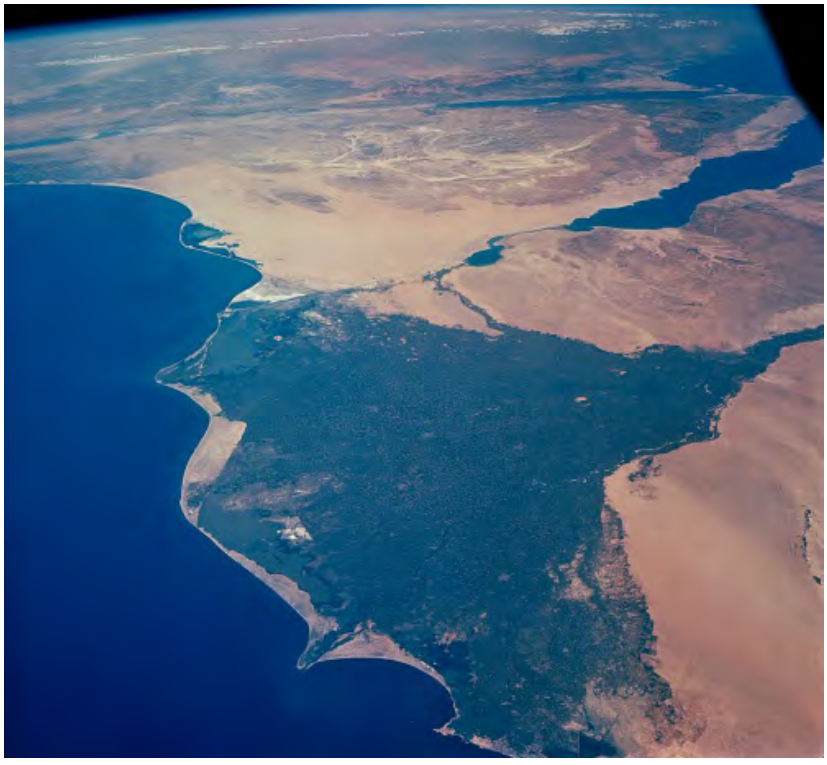
12-2: **Does the Mojave Desert appear anywhere in this photo? ANSWER**

Mercury and Gemini Photography

Astronaut photography started with the Mercury program. The unmanned MA-4 mission carried an automatic 70mm camera that returned systematic coverage of North Africa. NASA sent the photos to Morrison and Chown of Canada's McGill University for detailed analysis. All four Mercury orbital missions included hand-held cameras, primarily the precision German-made Hasselblad cameras, available commercially:

It was only in the last two missions, MA-8 and MA-9, that systematic terrain photography was a formal experiment. The astronauts for these missions, W.M. Schirra (8) and L.G. Cooper (9) had one 3 hour briefing by the writer (PDL), covering principles of geology, geology of the flight path, areas to be photographed, and a rock collection for study. Schirra's pictures, though taken precisely as scheduled, were largely spoiled by cloud cover over the Brazilian Shield, the only area we expected to be accessible within the short flight time (six revolutions). However, with more than 34 hours flight time, Cooper had far more opportunities to photograph pre-selected sites and targets of opportunity. The 24 excellent Hasselblad color photos he took, including these of the west coast of Burma (top), with its Arakan Yoma Mountains, and the Irawaddy River to the east, and the high mountains of the Himalayas (bottom photo), received wide publicity, and led to subsequent terrain photography experiments on the Gemini missions.

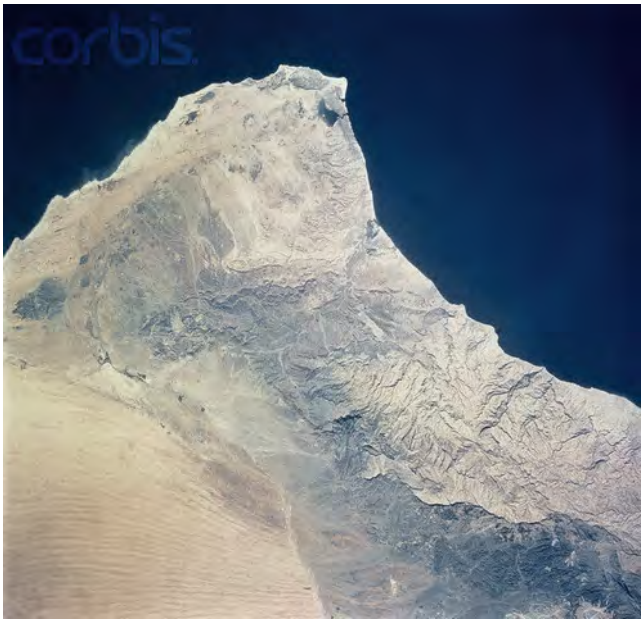
The Gemini Program (two astronauts to a spacecraft) was a major step forward in U.S. space capability. Intended solely to prepare for Apollo, ten manned missions flew in 1965-66, accomplishing orbital rendezvous, extravehicular activities, tethered flight, and many scientific experiments including 70 mm terrain and weather photography. The Synoptic Terrain Photography Experiment (S005) produced some 1,300 usable 70 mm color pictures of the Earth's surface. These were an enormous stimulus to space-borne remote sensing, partly because they were released to the public without restriction. Among spectacular photos returned from the Gemini 4 mission is this oblique panorama showing Egypt's Nile Delta in the foreground, with the Gulf of Aqaba-Dead Sea in the background:



This picture highlights two characteristics of many of the astronaut photos: 1) they are oblique shots that often contain the Earth's horizon in the background (but sometimes the curvature is from the window of the spacecraft), and 2) there is often a blue cast that overprints the entire photo, which results from blue backscattering from the atmosphere.

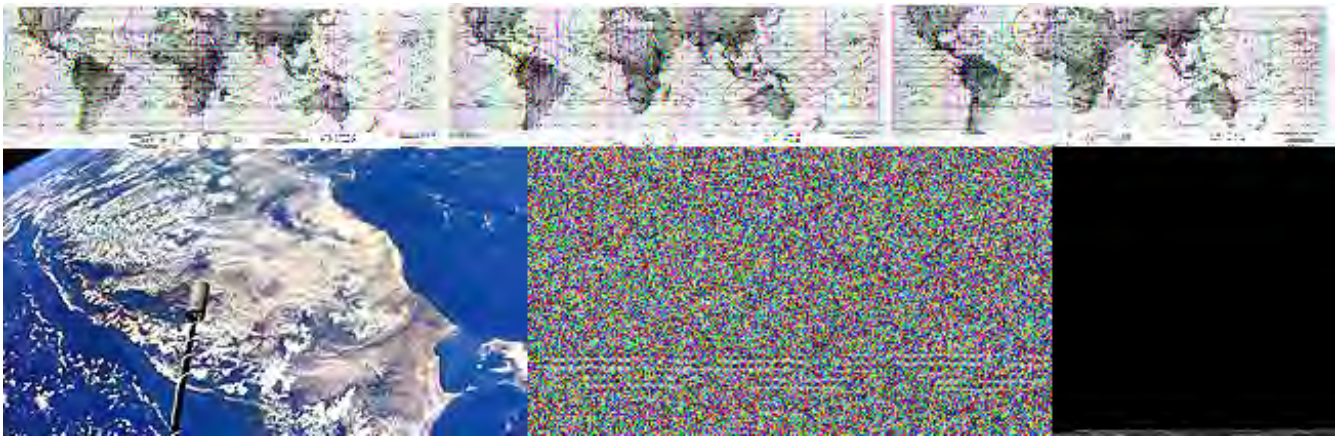
12-3: Does this picture give you a clue as to the origin of the term "delta"? ANSWER

Photographs from Gemini 4 and other missions in this series proved particularly informative when astronauts took pictures of desert scenes, such as these two:



This Gemini 11 photo illustrates how the astronauts see the Earth's curvature from altitudes of several hundred kilometers. The scene shows almost all of the Indian subcontinent; the white areas in the background near the horizon are possibly clouds atop the snow-covered Himalayas.

The flight plan (plots of orbits) for Gemini 12 is shown in this map:



One of the key photos obtained during this mission shows Oman, the Straits of Hormuz, the Persian Gulf, and part of the Zagros Mountains of Iran.

12-4: **Which way is South in this picture?** **ANSWER**

To keep space photography in perspective, we note that the U.S. Air Force conducted film photography of the Earth's surface, as differentiated from weather satellite imaging (Section 14), since 1960, under the now-declassified CORONA program. This program amassed thousands of high resolution pictures, but these and the programs that procured them (CORONA, ARGON, and LANYARD) were highly classified then. Published in *Life*, the *National Geographic*, and many newspapers, the Gemini photographs were the first real demonstration of the scientific and applied utility of space photography. They are, even today, of archival value, because in three decades since they were taken, in some places significant geomorphic and environmental changes have altered the appearance of the terrain. Africa's Lake Chad, for example, has visibly shrunk, as has the Aral Sea in Asia. Terrain photos often showed oceanographic features well, and were studied intensively by R.L. Stephenson and others. The Gemini astronauts also conducted a wide range of astronomical and other specialized types of photography.

The most important result of the Gemini terrain photography was its stimulus to what eventually became the Landsat program. Under the leadership of W.A. Fischer and W. Pecora, the Gemini pictures were the central evidence cited by the U.S. Geological Survey in proposing its Earth Resources Observation Satellite (EROS) in 1966. Interagency conflicts arose, but in due course, EROS became ERTS, the Earth Resources Technology Satellite, later renamed Landsat. A popular account of these developments has been published by Hall *). However, not many people realize that ERTS was an Apollo Program spinoff, in that the Gemini photography was an effective catalyst to stimulate satellite remote sensing, as postulated by Lowman (1996) *. ERTS undoubtedly would have developed anyway, as an offshoot from weather satellite imaging, but we suggest that Gemini advanced it by perhaps five years.

PREVIOUS RST NEXT

Primary Contact: Nicholas M. Short, Sr.

Section 12 Author: Dr. Paul D. Lowman Jr. (lowman@denali.gsfc.nasa.gov)

One of the most successful series of manned flights, insofar as photos taken, was the Apollo Program. The first six were unmanned tests but from Apollo 7 through 17 (most destined for the Moon) Earth photography was always a prime consideration during the phase of the mission when the globe was being circumnavigated. One task in particular, on Apollo 9, known as the SO65 experiment, employed an array of 4 Hasselblad cameras mounted against a spacecraft window. When any one scene was photographed, the result was three black and white photos taken through filters of different color, and a color IR film product; this multispectral arrangement helped to test ideas about multispectral space imagery that were a precursor to the ERTS program.

Apollo Photography

The Apollo Program was an umbrella label that included the Gemini, Skylab, and Apollo-Soyuz Projects. Two Apollo manned missions, Apollos 7 and 9, stayed entirely in low-Earth orbit. The remainder of the Apollo missions always started by circling the Earth and then blasting off to the Moon. Photography of Earth thus was always done, on a limited basis, before Apollos 8, 10, 11, 12, 13, 14, 15, 16, and 17 began their lunar voyages.

Apollo 7, in 1968, commanded by Wally Schirra, was the first flight test of the Command and Service Module. During the 11 days in orbit, the crew took about 500 photographs of Earth, again using color film in Hasselblad cameras. This scene looking southwest shows the Salton Sea and Imperial Valley, with the Peninsular Ranges of southern California in the background and block fault mountains in southwest Arizona and California's Mojave Desert in the lower part.

12-5: **The U.S.-Mexican border is visually evident in this photo. What defines this boundary?****ANSWER**

An Apollo 7 photo gives a very different impression of the Andes Mountains than you might have

gained from inspecting them on page 6-12:

Although of excellent quality, the Apollo 7 pictures never received detailed post-flight analysis, largely because of the tight flight schedule involving the first lunar trip in Apollo 8.

The Apollo 8 trip, famed for its broadcast of greetings on Christmas Eve of 1968 as the crew circled the Moon, was something of an improvisation to test the Saturn V launch vehicle before the Lunar Module was ready for flight. Another goal was to insure that the Soviet Union's Zond spacecraft did not make the first lunar mission (see Shepard and Slayton, 1994) *. As the first manned mission ever to reach escape velocity and thus break from Earth's gravity, it had enormous impact worldwide. The hand-held 70mm photographs, including the first color pictures of the Earth rising from the limb of the Moon, were truly historic. The astronauts took scores of lunar surface photos, and, although coverage was sporadic, compared to the pictures from the five Lunar Orbiter missions, the photos proved fascinating to the public and scientists alike.

The first manned flight of the complete stack, i.e., Saturn V booster with all components or stages, was the low-Earth orbit, Apollo 9 mission. After completing tests of the Lunar Module, the crew, commanded by Col. J.A. McDivitt, acquired more than 1,100 70mm color photos of Earth, using single, hand-held cameras. Apollo 9 included an entry into the space environment by

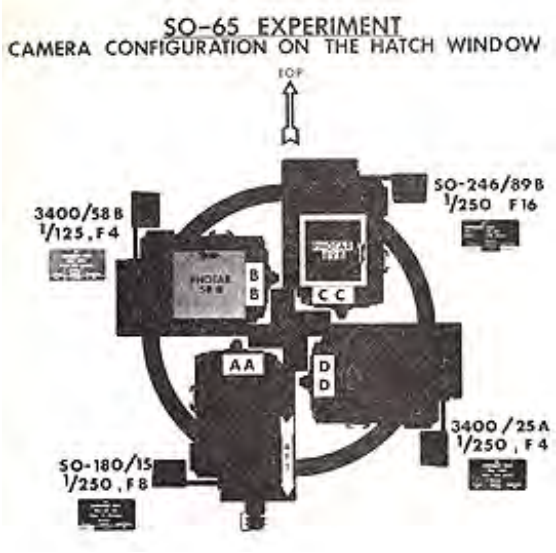
Dave Scott in his flight suit; astronaut Rusty Schweikart took this picture. In principle, the astronauts on EVA could also do photography but normally were busy with other tasks.

As an example of Apollo 9 photography, check this shot, looking north along coastal plains that extend over much of the East Coast of the U.S, with the coastal plains of North Carolina in the foreground, the Delmarva and Delaware Peninsulas in the middle, and Long Island near the top right.

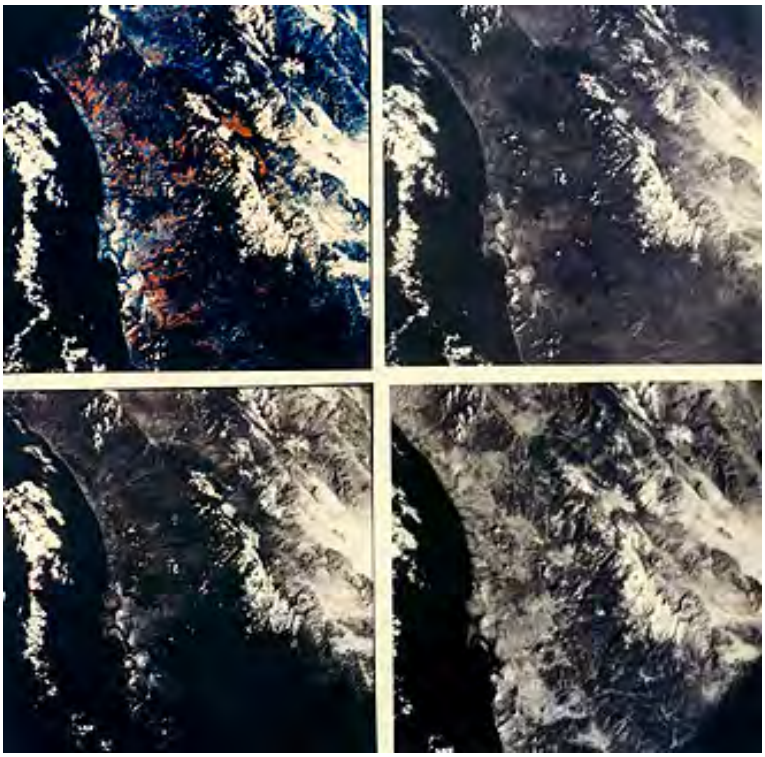
Another Apollo 9 photo covers a part of the Texas Panhandle and eastern New Mexico. The High Great Plains lies to the east, where large, irregular-shaped farms are scattered midst barren land. Almost fully barren is the fluvial sediment-covered end of the Llano Estacado (Staked Plains) in the center, bounded on the west by the Mescalero Escarpment. Little farming is done further west, on the Pecos Plains, until the Pecos River itself is reached (left side).

This Apollo 9 image shows the Ahaggar massif, mostly crystalline metamorphic rocks, in the desert of North Africa:

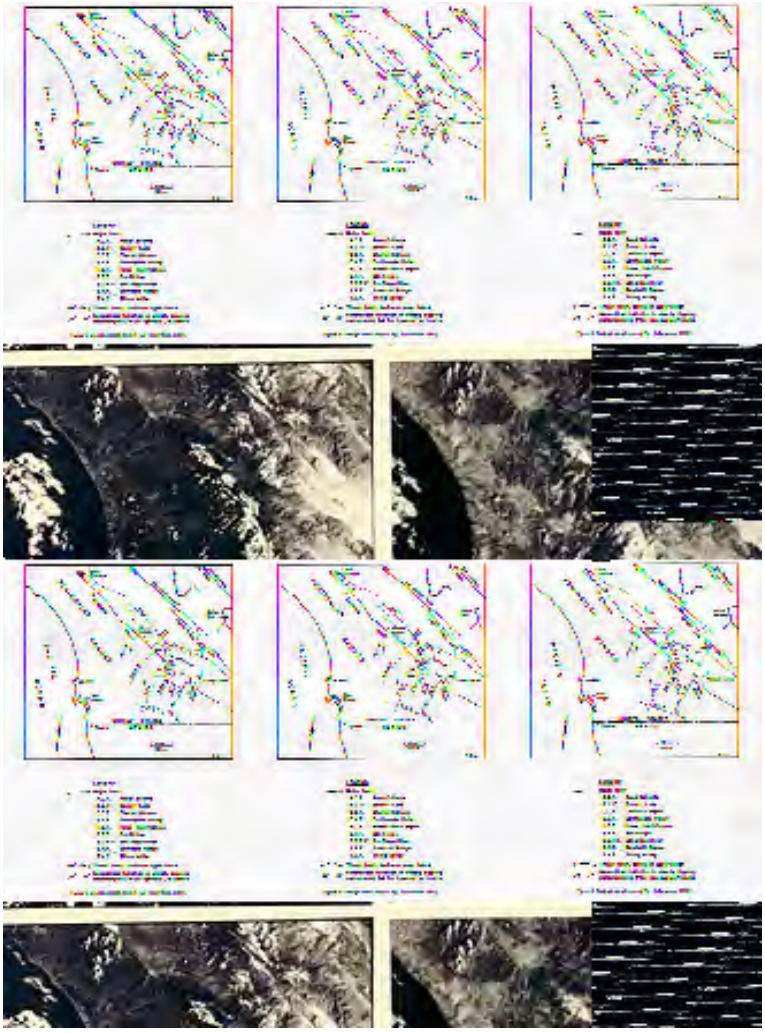
Furthermore, the Apollo 9 astronauts did a film rehearsal of ERTS (Landsat), employing coaxially-mounted 70mm Hasselblad cameras with color and black and white film, to produce multispectral terrain photographs, as part of the S065 experiment (Lowman, 1980) *. Here is a drawing that indicates the positions of the four cameras mounted in a porthole in the spacecraft.



The montage of four photos that cover San Diego and the California/Mexico Peninsular Ranges includes the following filters: upper left: false color infrared; upper right: green filter; lower left: red filter; lower right: black and white infrared.



The writer (PDL), in his role as a Principal Investigator in this experiment, used these pictures in a transparency format, along with ground truth and geological map data, to generate this interpretive sketch map:



The Apollo 9 mission was the most productive flown to that time, yielding superb pictures, some of which to this day are the best records from space of certain areas. Some people suggest that Earth's atmosphere was actually clearer in 1969 than when Shuttle missions began some years later. For example, the Amazon Basin, a "blue ocean" in the words of the Gemini 9 astronauts, was by the 1990s generally obscured from smoke rising from massive land clearing.

Before heading for its historic landing on the Moon, one of the last pictures taken from Apollo 11 showed a partial Earth view of much of the U.S. Pacific Northwest into western Canada. Thus:

Geostationary weather satellites, Galileo, and other spacecraft have returned great pictures of

the full Earth from space. But, still ranking number one (in request popularity) is this photo of our planet, showing Africa and surrounding oceans, taken during Apollo 17's return from the final manned mission to our lunar neighbor.

12-6: **Why does the middle third of Africa appear dark compared to, say, the African desert (Sahel) to the north? ANSWER**

Collectively, the deep space Apollo views of Earth have been credited as a major force in stimulating the environmental movements of the 1960s and later, by reminding us of the incredible isolation of this blue and white planet in the blackness of space.

The six Apollo lunar landing missions also returned enormous numbers of photographs of Earth's satellite from lunar orbit and on the surface. Masursky, et. al., published a good collection of these in 1978 *. Here are two Apollo 16 examples: an oblique photo (top) that shows the impact crater Aristarchus and a neighboring crater of probable volcanic origin, as attested by its lack of a central peak and terraces, and the large (Schroeter's) rille traced to it; and (bottom) a near vertical close-up of Tsiliokovsky Crater on the lunar farside.

Primary Contact: Nicholas M. Short, Sr.

In 1973, even as the lunar exploration phase of the Apollo program was winding down, the U.S. launched its first space station, Skylab, a large Earth-orbiting module that was actually the top or front stage of a Saturn V rocket. Visited a number of times by the astronauts, Skylab proved an excellent laboratory for many experiments including those involving systematic photography of Earth. In 1975, the last Apollo spacecraft (once destined to be on Apollo 18 to the Moon) was placed in orbit for a docking test with the (fully cooperating) Soviet Soyuz space station. It too witnessed considerable activity in photographic documentation of the Earth.

Skylab and Apollo-Soyuz Photography

Skylab was America's first space station. It was, in a sense, a by-product of the formal Apollo program. The manned space station was a converted Saturn SIVB stage section launched dry (without fuel; and its interior fitted for habitation and operational control). It launched unmanned on the last Saturn V sent up in 1973, and astronauts arrived later. This view shows the deployed station against a black sky background, with the extended solar panels (note that one failed to open and broke off).

The first crew, led by Pete Conrad, occupied the Skylab after rendezvous. At that time, the mission was by far the longest and most complicated of any manned flight in the U.S. space program, and was highly successful (Canby, 1974) *.

The Skylab orbit, originally planned for a 28° inclination, was increased to 50° after protests by many of the scientists involved. This higher inclination offered far more ground coverage than earlier manned spacecraft orbits that were 32° or less. As Michael Collins has pointed out *, a 50° orbit covers 3/4ths of the Earth's surface and areas with 90% of its population. The high altitude (438 km; 270 miles), also increased coverage.

As shown in the pre-Apollo table **, Skylab carried an arsenal of cameras, both hard-mounted and hand-held. During three occupations of this space station, the crews acquired tens of thousands of photographs of the Earth, in addition to many astronomical pictures and other

types of remote sensing images.

The view (top) of the South Island of New Zealand, showing the great Alpine fault which abruptly truncates the snow-covered Alps, and the scene below it, a larger-scale picture of Chicago taken by the S190B camera, are typical of Skylab imagery:

The very last Apollo mission was flown in 1975, involving the first link-up with a former Soviet Soyuz spacecraft (Shepard and Slayton, 1994) * . Although flown partly for political reasons, the Apollo-Soyuz Test Program (ASTP) mission generated a notable volume of useful scientific data, from on-board experiments and from photography. Experiment MA-136, under the direction of Farouk El-Baz, was more elaborate than previous 70mm photography. In particular, the astronauts selected targets more carefully for visual observation, drawing upon the Apollo lunar experiences, which had further demonstrated the value of the human eye. They also collected a substantial amount of ground and water truth data (see next Section). Because of the high latitude of the Soviet launch site, the ASTP mission had a high inclination, 51.8°,orbit that broadened the scope of regional coverage.

The astronauts took some 2,000 pictures, about 750 of these were of good quality (e.g., not cloud-obscured). Targets were diverse, covering geology, oceanography, and meteorology. A characteristic scene is this view of part of southwest Africa in Angola, where unique drainage patterns are controlled by broad, partially revegetated dune fields.

12-7: **For those of you who are familiar with the major types of drainage patterns, what does the pattern in the Angola image remind you of? ANSWER**

Photographs and other data from the ASTP mission were incorporated in a series of pamphlets for teachers, this being the first time educational applications were a formal objective of the Earth photography.

David Amsbury * published a general review of all pre-Shuttle Earth photography. In 1970, Kaltenbach tabulated the photographic equipment used in these missions, from Mercury through ASTP. By clicking here , you can review a series of tables from his report.

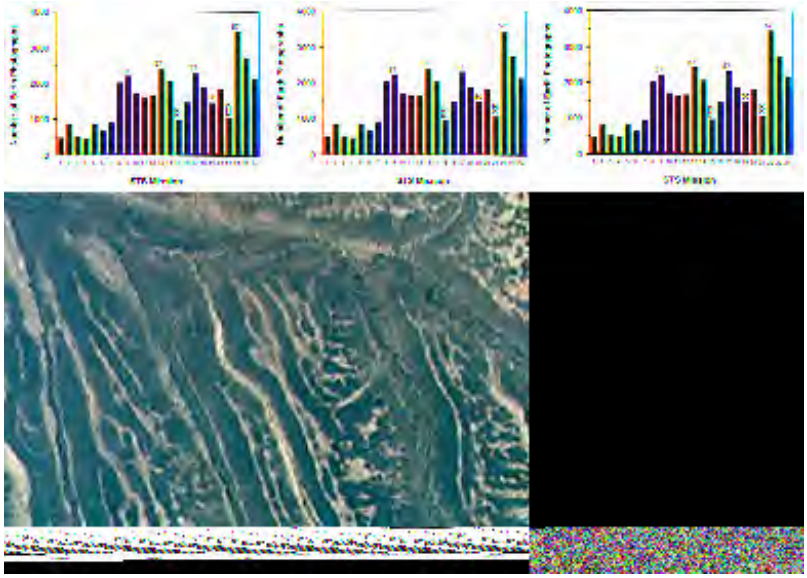
PREVIOUS RST NEXT

Primary Contact: Nicholas M. Short, Sr.

The 1980s opened a new era in manned operations in space. With the success of the Space Transportation System (STS) program, more popularly known as the Space Shuttle, whose missions are now in excess if 120, repeated opportunities have been presented to photograph new areas on Earth and to revisit areas already photographed so as to keep a running record of transient events and slower changes in both the natural and manmade status of surface features. The Russians likewise have continued their photographic programs from their MIR space stations, especially with the KOSMOS series of imagery (see next page).

The Space Shuttle Photographic Program

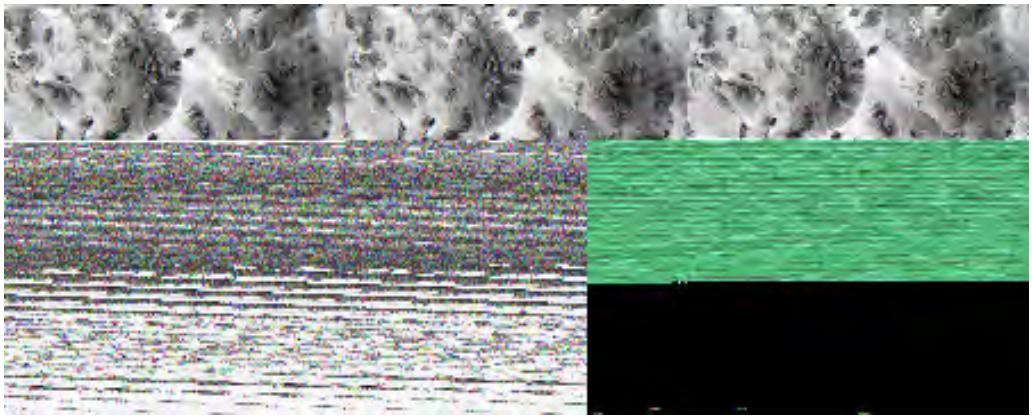
Several years passed between the last U.S. manned mission (ASTP) and the first launches in 1981 of the Space Shuttle (an informal term for the Space Transportation System [STS]). Film-based remote sensing in general advanced greatly as astronaut photography from the Shuttle became a far more extensive activity, evolving into the permanent and formal Space Shuttle Earth Observations Project (SSEOP). This plot shows the significant increase in photographs returned for just the earlier missions (current total return exceeds 250,000):



A useful summary of the SSEOP by Helfert and Wood, emphasizing photo scheduling, equipment, crew training, and the enormous problem of cataloging and inventorying, appears in the March 1989 (Vol. 4, No. 1) special issue of Geocarta International *. The recent 1996 publication * by astronauts Jay Apt, Michael Helfert, and Justin Wilkinson presents a stunning collection of the SSEOP pictures, with technical data for each.

Before summarizing the hand-held photography, we should describe the Large Format Camera (LFC). The LFC was by far the most capable camera ever flown on a U.S. manned mission. Hard-mounted in the Shuttle payload bay, this 405 kg camera, with a 305 mm focal length and a 23 cm by 46 cm format (see Mollberg and Schardt, 1988 *) was the central element of photographic experiments on the few missions in which they used it. It is probably the only space payload in which cast iron was a major constituent. From a typical altitude of 300 km (186 mi), a frame

covers ground dimensions of about 225 x 450 km (140 x 280 mi). Although quite successful, the LFC doesn't always fly, primarily because it requires dedicated payload-bay space, attitude-control fuel, and scheduled time, in contrast to the hand-held photography. This view of the Mojave Desert in California taken with the LFC during the 41-G mission in 1984, is typical of the quality achieved with this instrument:



Astronauts photographed the Himalaya Mountains (looking west) during STS-17, in which the oblique view extends from the Ganges Plains in India, across the Siwalik Hills, into the great peaks of Nepal, and beyond. Also visible is the vast Tibetan Plateau, the highest broad land mass in the world, which receives very little snow during most of the year, because precipitation falls on the mountains to the south, when monsoonal winds blow northward.

As evident from the above photo, a large percentage of astronaut photos from the Shuttle are slanted, i.e., look obliquely at an angle to the surface rather than straight down. This is evident in this next photo which shows the southwestern U.S. pointed in that direction:

12-8: **Roughly, what time of year was this photo taken? Explain your reasoning.**
ANSWER

In addition to the Shuttle photography being more advanced and varied than prior missions, the flights are more frequent, last longer, and cover more territory because of the higher orbital inclinations (up to 57°). Other major improvements:

1. Pre-flight Preparation: Shuttle crews receive extensive training in photography and in subjects such as geology, meteorology, oceanography, and ecology. Their training stresses global change, especially since we now have four decades of space photography covering some areas. Specialists prepare extensive lists of observation sites for each mission, embracing, so far, more than 1,800 significant areas.

2. Equipment: Space cameras have improved over the years, although the Hasselblad 70 mm is still the workhorse. With the wider variety of cameras and lenses, the crews can now match optimal equipment with the subject.

They still commonly use Ektachrome 64 film, which the Johnson Space Center processes. This film permits valid comparisons of Shuttle pictures with pre-Shuttle pictures of similar subjects. Also, astronauts have used an Electronic (digital) Still Camera successfully on recent flights.

3. Post-mission Processing and Archiving: The SSEOP has systematically catalogued Shuttle photographs. Archivers listed early Shuttle mission photos in printed catalogs but they stopped

this practice in favor of Internet retrieval that favors easy scene selection and rapid downloading. Addition of digital data extracted from exposed film aids in this process. The digital data help the processers reproduce better colors and detail than with the multi-generation transparencies needed to preserve the flight film.

PREVIOUS NEXT

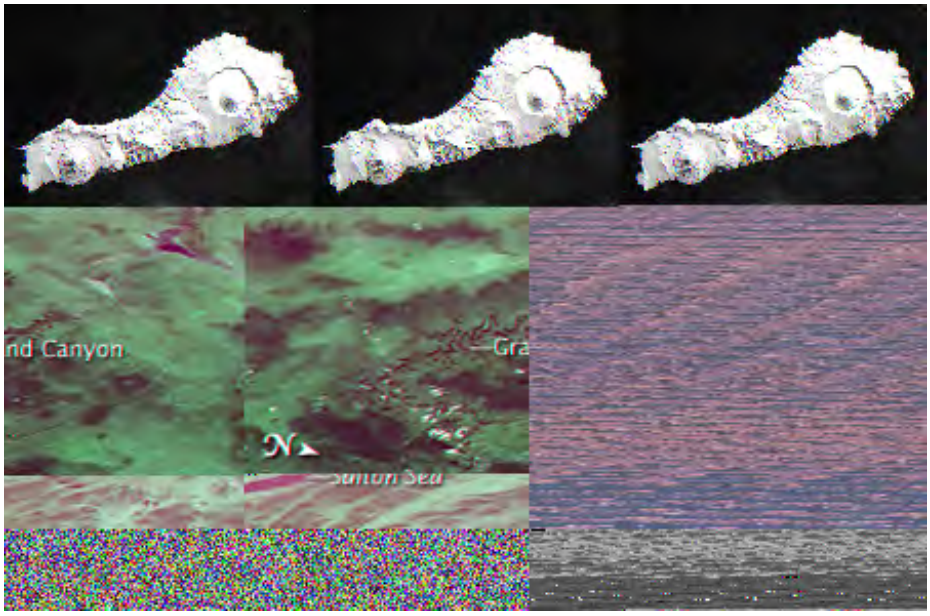
Primary Contact: Nicholas M. Short, Sr.

More examples of Shuttle imagery are displayed on this page, along with one picture taken by the MIR station cosmonauts using the KVR-1000 camera capable of 2 meter resolution and a pair of pictures from MIR taken by an American astronaut.

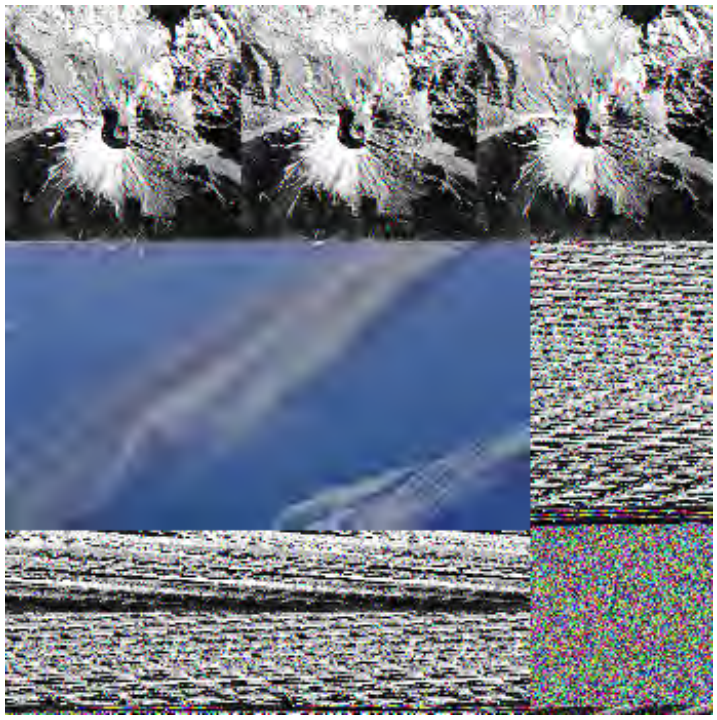
A Gallery of Photos from Space

We now present a gallery of exemplar photos, taken from representative Shuttle and other space missions. Some computer screens pair these images side-by-side but, small screens may display one above the other. Each one has a brief description.

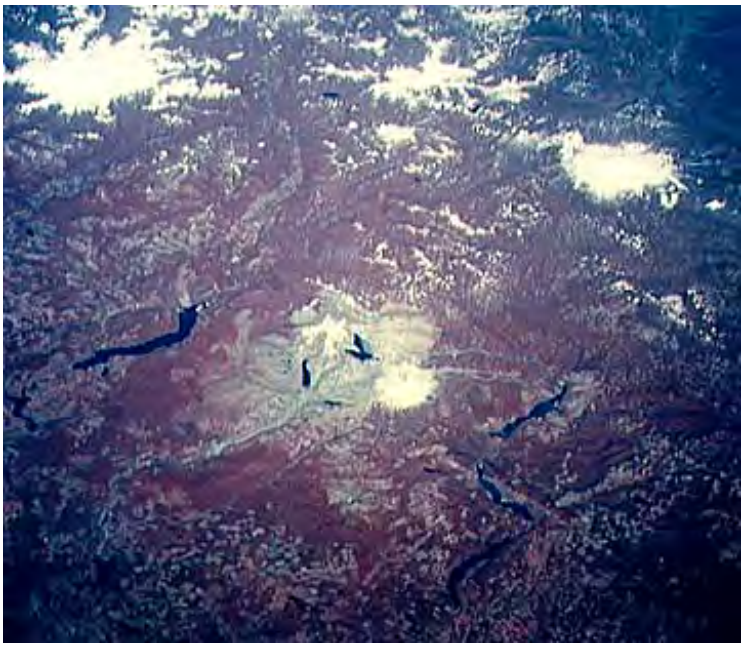
Let's start with a sequence showing active and recent volcanoes on three continents. In the first pair, on the top is a vertical photo taken during STS-59 that shows a snow-covered island in Russia's Kamchatka Peninsula (far east Siberia) that has a volcanic caldera at each end, both with a central stratocone. Below it is an oblique scene that traces a long plume emanating from the Kliuchevskoi volcano in that peninsula during an active eruption.



Mount Saint Helens, in southern Washington State, erupted catastrophically in 1980. Astronauts photographed it in 1994, with a Nikon 300 mm lens, during the STS-64 mission, clearly showing the persistent aprons of ash deposits, despite considerable reforestation, lobes of lahars (ash-mud flows), and the great gap where part of the mountain was blown away.

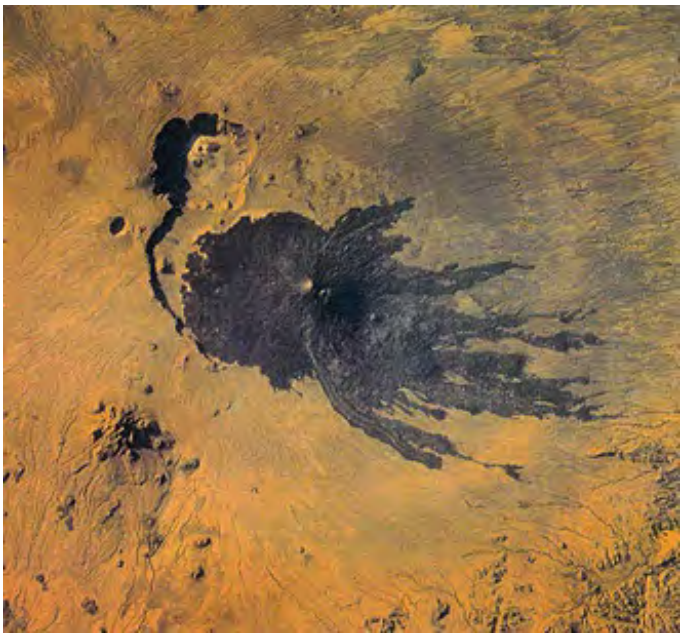


It is interesting to place Mt. St. Helens in context with two other great stratovolcanoes in the Cascade Mountains of the western U.S. These two Shuttle views, one an oblique, false-color IR picture (looking northwest) and the second, a near vertical, natural-color shot, show Mount Saint Helens (MSH) to the west and Mounts Rainier and Adams to the east, all snow-capped. Indeed, MSH is the only volcano that is significantly offset from the main line of the Cascadian volcanic activity. Note the clear-cutting patterns in the forested countryside.



12-9: **Both of the above photos have rather odd orientations. But the presence of the three great peaks allows one to tie the photos to a map atlas by the classic technique of triangulation. Try to do this. You will know you have succeeded when you find the name of the distinct large lake near Mt. St. Helens (volcano in the center in the top photo).** **ANSWER**

A volcano known as the Pic Tousside, in the Tibesti Mountains in the Sahara of Chad (Africa), is noteworthy because of the insect-like shape created by several recently erupted basalt flows:



Astronaut photography and geostationary satellites have depicted large areas of the global surface. Observe, on the top, a southwest-looking STS photo of the Great Lakes in the northern U.S. and adjacent Canada. Compare this with a scene extracted from a Nimbus AVHRR false-color image that covers all of these lakes (state and province boundaries are superimposed). Test yourself in identifying each of the Great Lakes.

We view cities in context with their surroundings in STS pictures. Below, the top view (STS-56) is Tokyo Bay in Japan. Tokyo is in the upper left, and Yokohama is in the lower left. On the right (east) side of the bay are built-up areas that include large landfills bounded by retainer walls, which hold ship docks. Compare Tokyo's visibility with the winter scene (STS-60) of Montreal in Quebec, Canada (bottom), in which cleared streets stand out against the snowscape, similar to a dusted fingerprint.

In the late 1990s, a new variant of Shuttle Photography was introduced: the EarthKam program. Here the astronauts took photos of targets that were preselected by America's K-12 grade students. The scene below is an example. It shows the Cape of Good Hope at the southern tip of Africa (south is to the right). The Hottentot Holland Mountains lie above the town of Strand. Note the column of smoke from a possible forest fire.

Another EarthKam example, also from South Africa, shows a region where this country meets Mozambique (right), along with the untouched Kruger National Park, a reserve for wild animals that is a popular "safari" tourist attraction. To the left are patches of landuse caused by recent resettlement of black Africans during the later stages of apartheid. This grouping is known as the Gazankulu Homeland, actually a series of "ghettos" that are now being rejuvenated after the addvent of self-rule in South Africa in which the dominant black population has assumed a leading role in governance.

A very large number of EarthKam images are Net accessible at the ISS EarthKam site.

Astronauts have taken some remarkable pictures from space in the dead of night. In one example, from the Shuttle, the collective lights of cities and towns around San Francisco Bay outline it in orange.

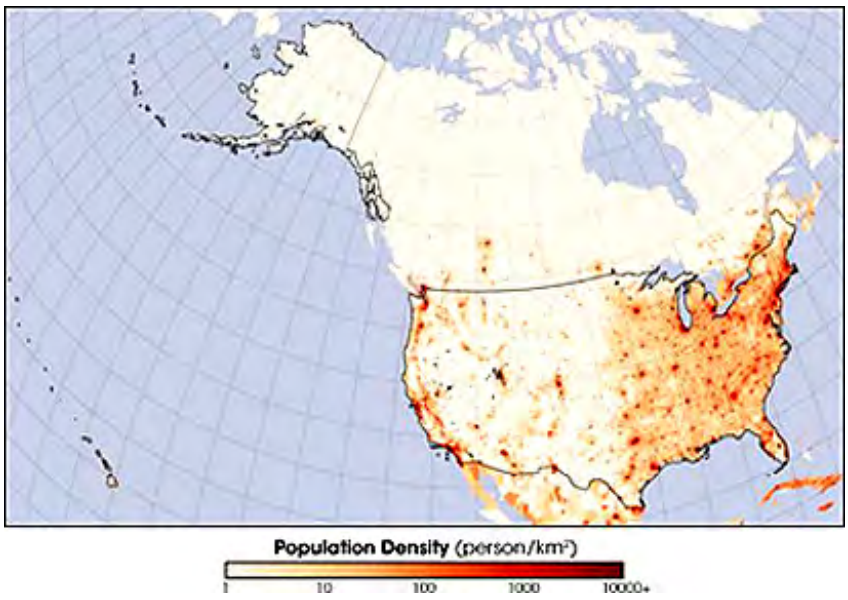
12-10: **Where is San Francisco itself in this striking Shuttle night view? ANSWER**

Although the next image is not a photograph, it ties in with the above San Francisco at night photo, but on a grandiose scale. The Defense Mapping Satellite Program (DMSP) is operated by the U.S. Air Force. One of its sun-synchronous, near-polar satellites took (from an 625 km [516 mile] altitude) nighttime images of the U.S. using a channel operating between 0.4 and 1.1 µm. Under cloudfree conditions in December 1986, the scene mosaic shows only those areas large enough to produce a concentration of light from street lamps, etc. See if you can locate the region in which you live, using the pattern of lights from metropolitan areas near you, Can you tell what accounts for the numerous linear patterns developed over most of the U.S.? How many areas of concentrated light can you identify as urban centers?

12-11: **If you live in a small to large city, see if you can locate it in the DMSP image. If you live in a more rural setting, try to find the nearest city(ies) to your residence. Many of the larger towns seem linked as beads in a rosary. Why? ANSWER**

It is instructive to compare the above DMSP image taken in 1986 with another, covering the U.S, southern Canada, the Caribbean, and Central America, and the northern part of South America, taken by a later DMSP in the year 2000. The increase in urban and town development in the western U.S. is especially noteworthy.

These images are realtime, actual portrayals of what is present at the surface of the U.S. and Canada. Satellite imagery is helpful in assessing population density and distribution. This information source has been used to provide auxiliary input into the making of a population density map for the United States (including Alaska and Hawaii) and Canada that was made public about 2 weeks (late October 2006) after the U.S. passed the three hundred million mark in its citizenry. Compare this map with the DMPS image above.



The European continent, and the Mediterranean lands especially, is heavily populated. One would expect many urban and regional centers to be notably "lit up" and thus easily recognized in night imagery. Examine this DMSP image of this part of the world and see how many specific cities you can place and identify.

Since Shuttle astronauts encircle the Earth in less than 2 hours and do this repeatedly during any 24-hour time stretch at a ground reference point, they should in this period see a number of sunrises and sunsets. This image of Europe and North Africa shows the sharp terminator marking dusk at some particular moment. Behind the still daylight area, the lights of Europe-Africa to the east have been switched on:

On a truly grand scale, the DMSP has covered the Earth under cloudfree conditions enough times now to allow production of a relative night light brightness map for all the continents of the World. Using images obtained at different times, and then mosaicked, this composite amounts to a map that also indicates the distribution of population densities. Japan seems to have the highest density, followed by Europe and the eastern U.S..

We have reproduced this image as a very large version of this global night view of the Earth's lands and seas that can be accessed by clicking here. This is a recent DSMP composite image. It is big enough for you to see details not evident or hard to discern in the above image. Note the tendency for high population densities along coastal regions.

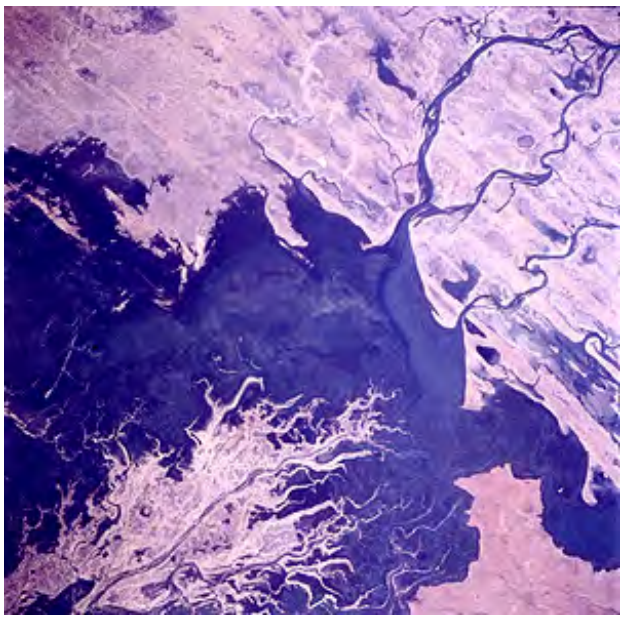
12-12: **There are several large areas on the continents that have low population densities. Give any reason(s) that come to mind for this. ANSWER**

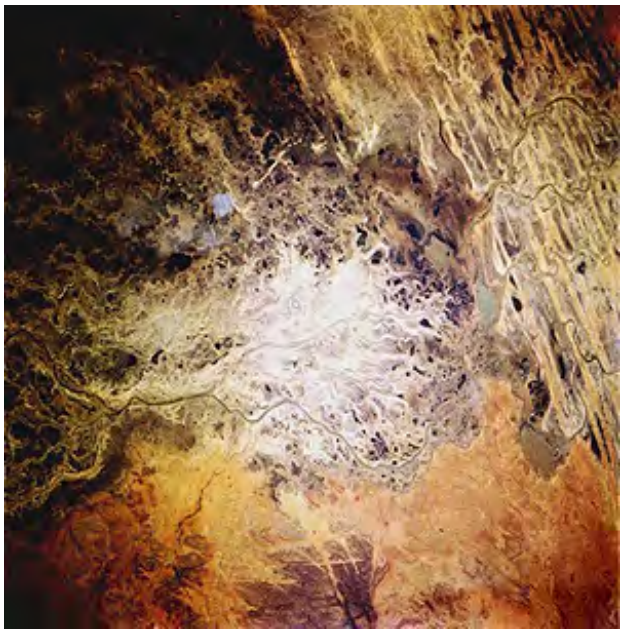
The Shuttle crews always look for terrestrial phenomena around, rather than on, the Earth. Here is a view, from STS-45, of the aurora borealis (also known as the Northern Lights) appearing as spectacular light bands, caused when electrons and other particles from the solar wind strike atoms in the outer atmosphere (see page Intro-1).

A rich source of a plethora of astronaut photography can be visited at the Johnson Space Center's Earth from Space Web site.

The Russian cosmonauts have also had a long-standing program for acquiring space photos. In recent years, they operated the Kosmos KVR-1000 camera on their MIR Space Station. Pictures, available commercially, using this camera can have spatial resolutions as high as 2 m (about 6.6 ft). An example of the sharpness achieved in such photos is this view of part of downtown Atlanta, Georgia:

Astronaut Shannon Lucid, participating with the Russians on MIR for an extended mission, took a large number of photographs of preselected targets of special interest. Two that are especially informative are scenes of the inland delta of Niger, in Africa. Almost every year there is enough rain to temporarily fill a basin along the Niger River as shown in the top photo. But, the arid, hot summer climate quickly dries this oasis of water leaving the rather desolate scene shown in the bottom photo. Note the sand dunes.





In one sense, the astronauts/cosmonauts are the "ultimate tourists" who take special pleasure and meaning from their photography activities while in space. Many of their targets are pre-selected but some are genuine targets of opportunity. Now, with the realization of the International Space Station (ISS) as a functioning platform, many new opportunities will be presented. The ISS has a Science Laboratory Module that includes a dedicated window, with special optical quality glass, through which photographs are being obtained. Here is a view of the ISS with its photo-window, taken soon after the module was attached in mid-February 2001.

And here is a photo taken in February, 2001 through this window. The scene is the ever-popular target - San Francisco - as it appears from 166 miles away.

Even more impressive is this ISS photo of Niagara Falls, near Buffalo.

A scene that has fascinated the astronauts and is thus represented by several photos is the Belasaka Estuary in Madagascar (see also page 17-4). Here is one ISS version:

The quality and sharpness of astronauts photos has steadily improved with better (now digital) cameras. The scene on the top is a November 24, 2004 ISS photo of Lake Sambhar in the Rajashan area of northwest India; on the bottom is an IKONOS view of the Glen Canyon Dam at Page, Arizona (a few miles from the Utah border) which produces Lake Powell along the Colorado River; the resolutions in both are less than 5 meters.

Here is another example of recent photo sharpness. Taken in the summer of 2008, this colorful scene shows the western end of the Florida Keys. These small coral reef islands are known as the Dry Tortugas (a tortuga is the spanish word for "turtle").

Photos both of the ISS and its crews and of ground scenes can be found at the International Space Station gallery. What comes up first is a banner and an Index arrow which when pressed gives a menu of imagery from all the astronaut programs. Click on the Station button to access ISS photos. But, if curious, click on any of the other buttons to see their contents.

The topics summarized in this Section should, first, point to the unique role of trained humans in space and, second, underscore the value of hand-held photography. A third advantage of pictures from manned missions is their low cost and ready availability compared to those from automated observation satellites like Landsat, SPOT, and similar space imagery, that generate expensive and sometimes proprietary products. Fourth, their natural color and their obliquity make many hand-held pictures generally well-suited to educational use, providing easily understood views of the Earth from orbital altitudes. Finally, the photography stands as a record

of long-term changes (going back to the 1960s) of the Earth's surface, whose appearance, as well as its physical structure are continuously changing, sometimes from minute-to-minute.

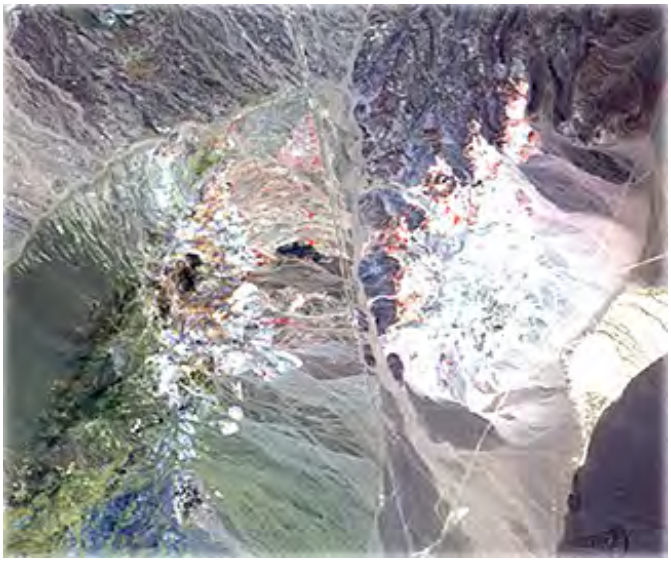
Primary Contact: Nicholas M. Short, Sr.

The page is devoted mainly to examples of AVIRIS images, plus a pair of images from two other systems. Special attention is given to one site, Cuprite, Nevada, which is a splendid case study for illustrating the efficacy of hyperspectral imaging.

Additional Examples of Imaging Spectrometer Products

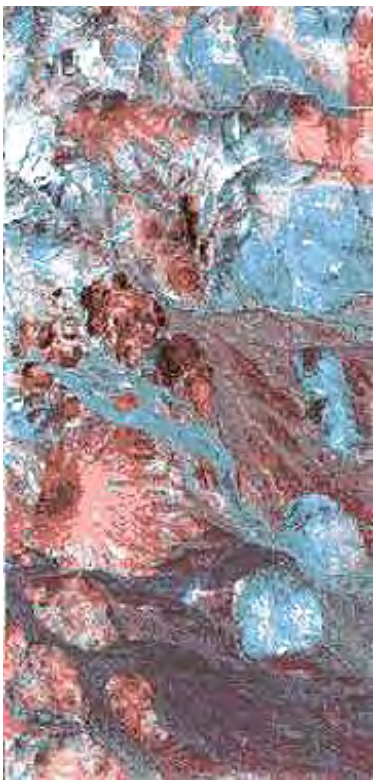
By now, hundreds of missions designed to test and use imaging spectrometers have flown, with many impressive images generated from the data. Still one of the most successful and instructive among these was the NASA AVIRIS flights over Cuprite, Nevada. The JPL AVIRIS team and the Spectroscopy Group at the U.S. Geological Survey in Denver reduce and manipulate the data. Be advised to check out the USGS page that will come up as it will offer the option of clicking on "Maps" which then brings up the Cuprite site as a choice. The images you will find at that site are superior to those below (the legends are readable), as these latter are degraded in downloading.

The Cuprite mining district lies near Tonopah, NV, in the southwestern part of the state. Gold and copper have been mined from here for more than a century. This area is a valuable geological study site to evaluate remote sensing for mineral exploration, in particular with hyperspectral data sets, because of the wide variety of telltale alteration and other mineralization. This is a Landsat view from space.



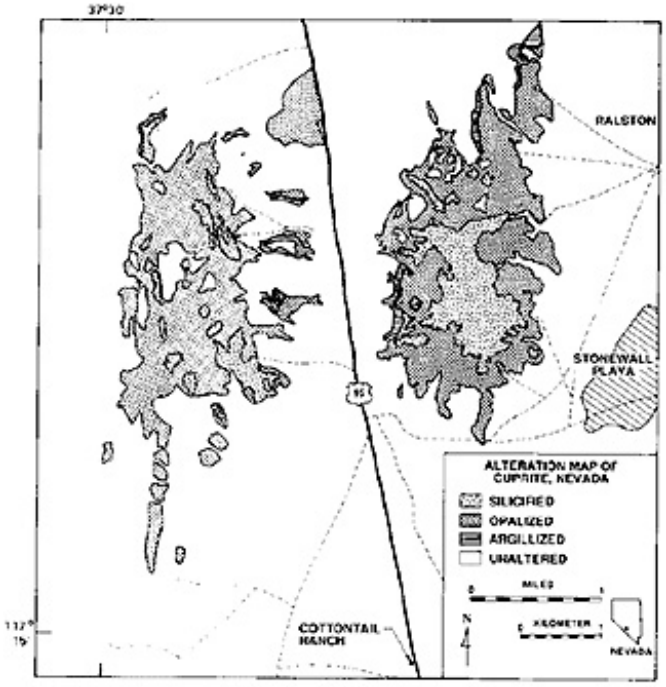
Here is a view made from three AVIRIS channels in pseudotrue color of the main area displaying significant alteration. The area shown is approximately 17 by 10 km.

By way of comparison, we got the next image from the HYDICE sensor. This false color combination consists of Red = Channel 175 (2,200 nm), Green = Ch. 125 (1,650 nm), and Blue = Ch. 50 (650 nm). The area here includes part of the AVIRIS scene but also includes other features.

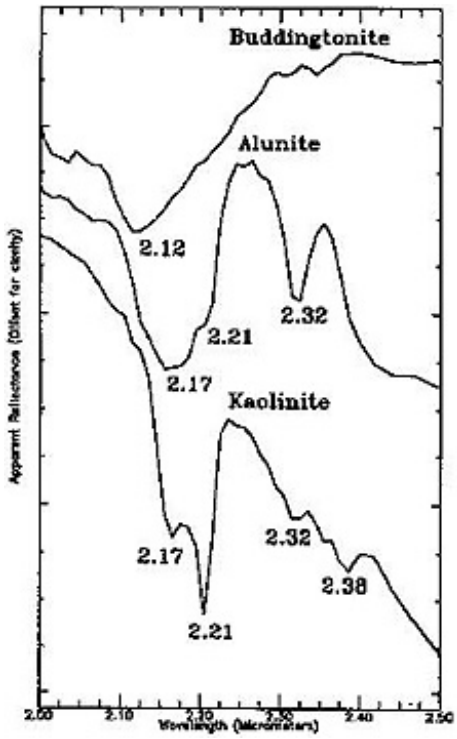


13-28: Comment on the "purity" of the above two images compared with those of Landsat. ANSWER

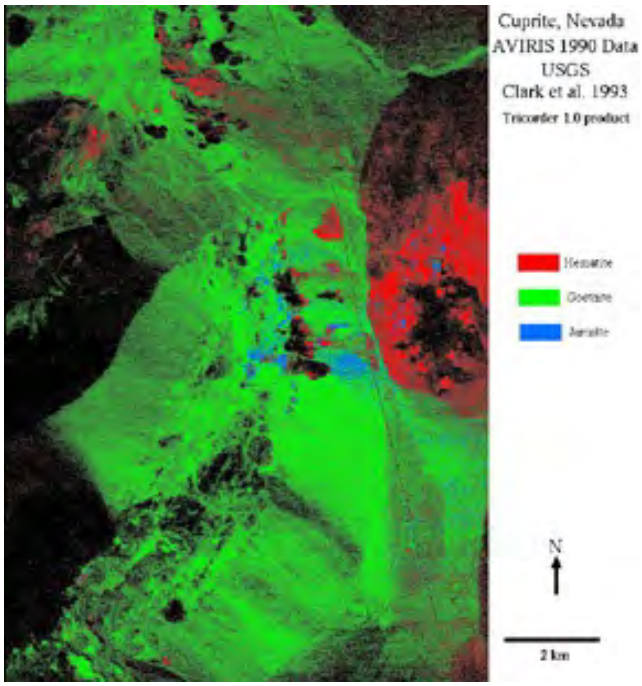
Cuprite was selected in part because previous field work at ascertained the variety of alteration minerals at that site, as shown in the map below. The alteration has affected several types of volcanic deposits of Tertiary age.



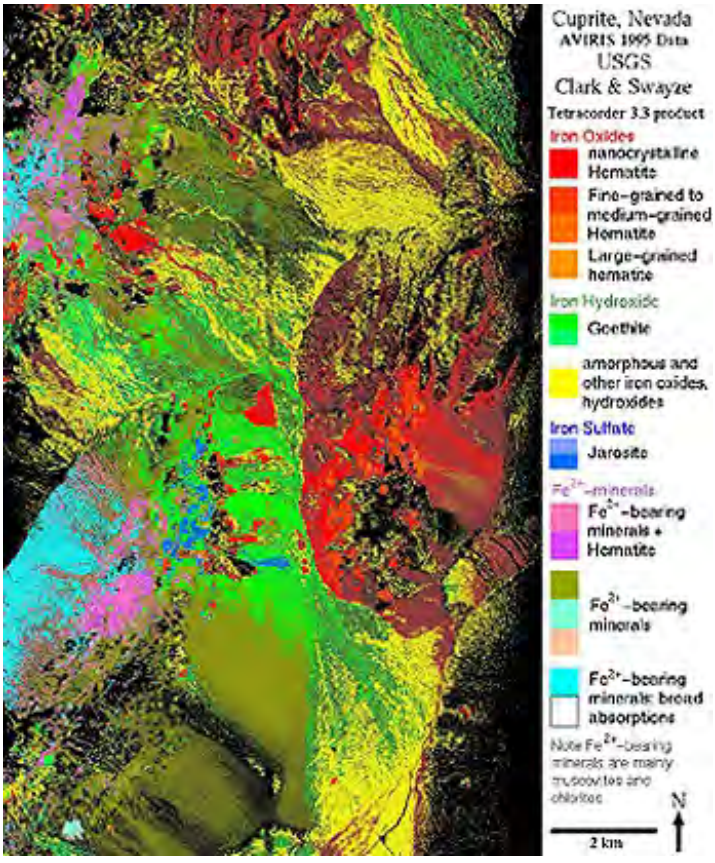
Laboratory spectra of three of the principal alteration minerals are reproduced here; spectra constructed from AVIRIS data are similar:



Here is a SWIR bands image made during the first field trials (1990) of AVIRIS in which the colors represent the iron minerals Hematite, Goethite, and Jarosite. The Goethite occurs mainly as stain or pigmentation in the alluvial materials in the valley and along slopes. (Black is unclassified).

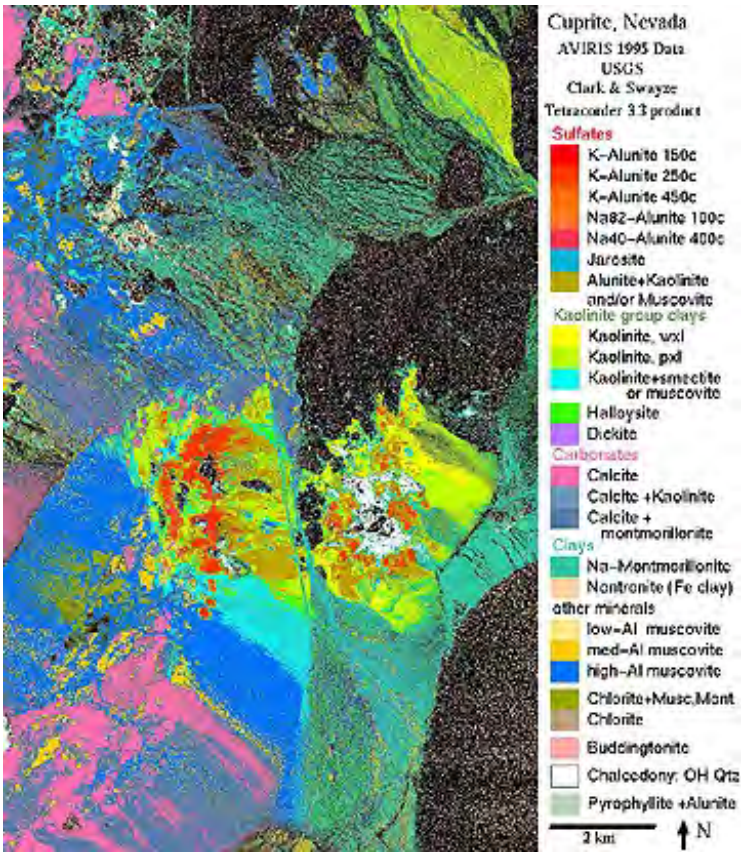


By judiciously selecting channels extending into absorption bands that are indicators of particular mineral species or groups, we have identified and mapped most of the alteration-related minerals at Cuprite. This next map extends the location of iron-bearing minerals, which have many of their diagnostic bands in the 400 to 1,200 nm range.



This map differs from the previous one (above) in subdividing Hematite and Goethite according to crystallinity and in showing other iron minerals as distinguishable chemical phases.

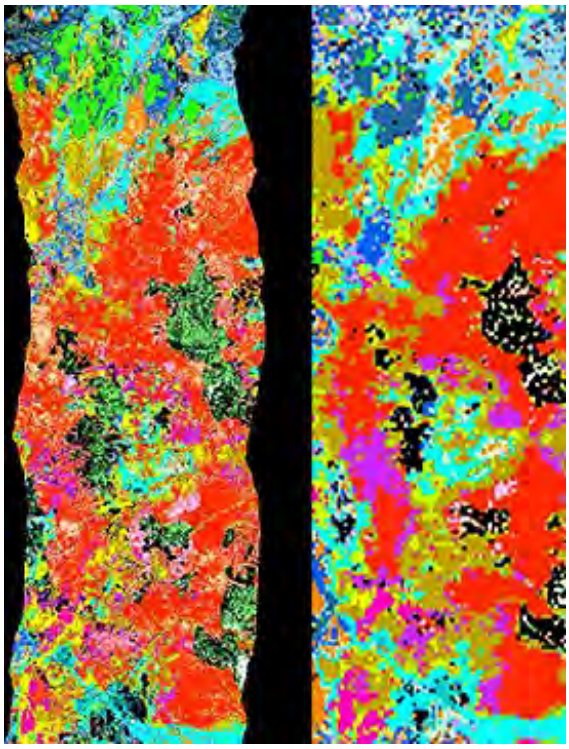
We can recognize an even more diverse assembly of minerals using bands within the 2,000-2,500 nm range. AVIRIS images have shown a large group of silicates, carbonates, and sulphates, as shown in this map.



13-29: Compare the above three images that show variations in certain elements that are then translated into their mineral variations with the natural and false color composites towards the page top. Do you see evidence in those latter two images of these mineral variations? ANSWER

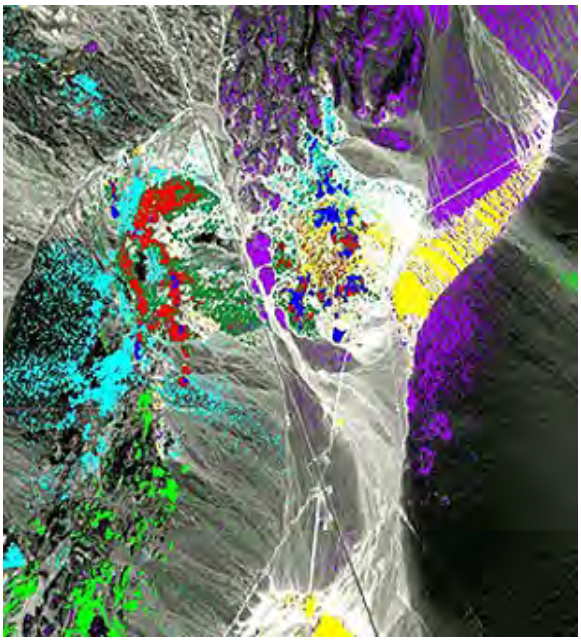
Of special interest is the mineral Buddingtonite, shown in fuchsia (a peach pink), which occurs in a few patches on the map. Buddingtonite is a rare form of the common potash-feldspar group. The ammonium ion, NH_3^+, partially replaces the potassium ion, K^+. This map is an exciting example of high resolution hyperspectral data to reveal a notable diversity of minerals in alteration zones and fresh rock, as well, at a detail that could require years of field mapping to duplicate.

The Cuprite site has now been overflown at low and high altitudes to see how height (hence resolution) affects the ability to distinguish mineral distributions. In the image pair below, the left shows a part of the Cuprite scene at 2.3 x 7 m resolution; the right, obtained from a higher altitude, is the same area at 18 x 18 m resolution. Judge for yourself from the patterns the extent of better or new information obtained at low altitude/high resolution.



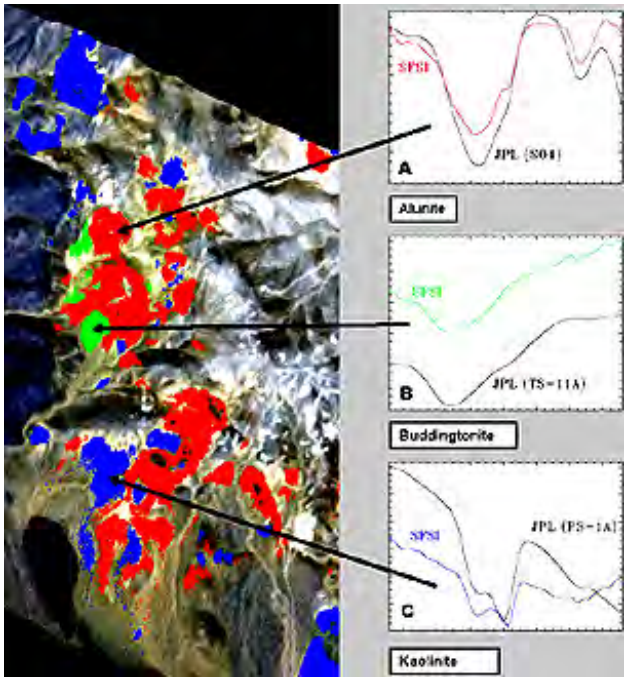
Because of the variety of alteration present at Cuprite, that site has become a standard for testing the ability of hyperspectral sensors to discriminate its characteristic mineralogy. Terra's ASTER has enough bands in the SWIR region of the spectrum to indicate how well that spaceborne sensor can pick up the diagnostic minerals. This first SWIR color composite shows that the overall anomaly at Cuprite can be spotted as evidence of an alteration zone:

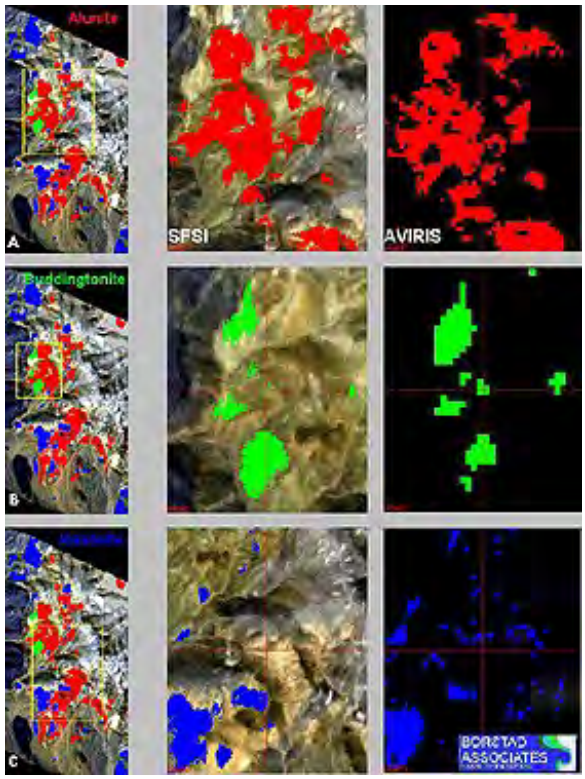
Having established this general detectability, ASTER bands considered especially sensitive to mineralogy were used to produce this image.



The resulting colors are associated with non-iron minerals as follows: blue = Kaolinite; cyan = Montmorillonite; purple = Unaltered; light green = Calcite; dark green = Alunite + Kaolinite; yellow = Dickite; red = Alunite

The great improvement in information content resulting from hyperspectral data acquired both from airborne and spaceborne sensors has generated renewed interest in using such data for both mineral exploration and other applications. A number of private companies that fly remote sensors commercially have added hyperspectral capabilities to their services; in fact, several new ones were formed to specialize in this approach. Cuprite has been used as an established test base. One company, Borstad Associates, has released their results (SFSI sensor) on the Internet. Here are two images taken from their website:

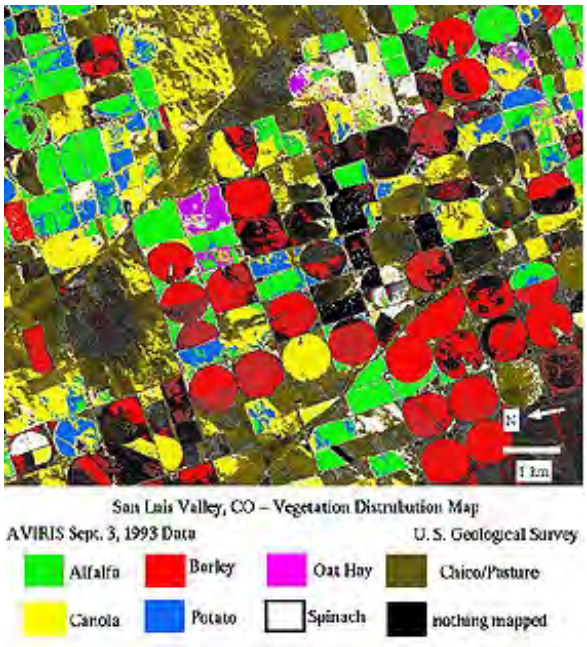




Another Cuprite hyperspectral product is this image made by the HyMap sensor:

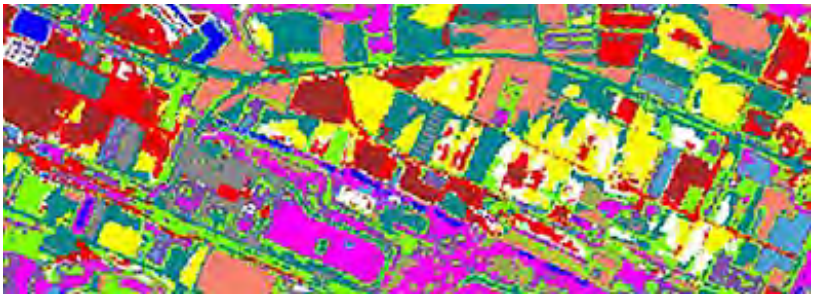
Based on the above evidence and other studies of alteration as telltale guides to ore deposits, it seems safe in asserting that Cuprite is the definitive mineralized site for demonstrating not only the value of hyperspectral remote sensing but remote sensing in general (see Section 5) as a powerful tool in searching for mineral and petroleum deposits. It seems likely that hyperspectral remote sensing for will become a prime means in future exploration programs. The strategy will be to use space imagery to look anomalies and then overfly promising ones with airborne sensors. However, if this is done systematically, most potential sites worldwide will be overflown in the 21st Century. To keep the companies viably operational, other uses need to be developed. To show that this is happening, we display several images touching on these additional uses, with minimal comment.

Determining vegetation to the species level is a major goal in remote sensing that hyperspectral sensors are helping to address. This next AVIRIS image, which we first saw on page 3-1, shows a group of crops in both circular and rectangular fields. The area is near Summitville, Colorado; a false color image of this site was included on page I-24. Note that the identified crops correspond to those shown on page 13-6, as a plot displaying spectral curves for a series of crops, first, and just below it, a continuum-removal plot of the same series. These came from the same data set used to generate this image.



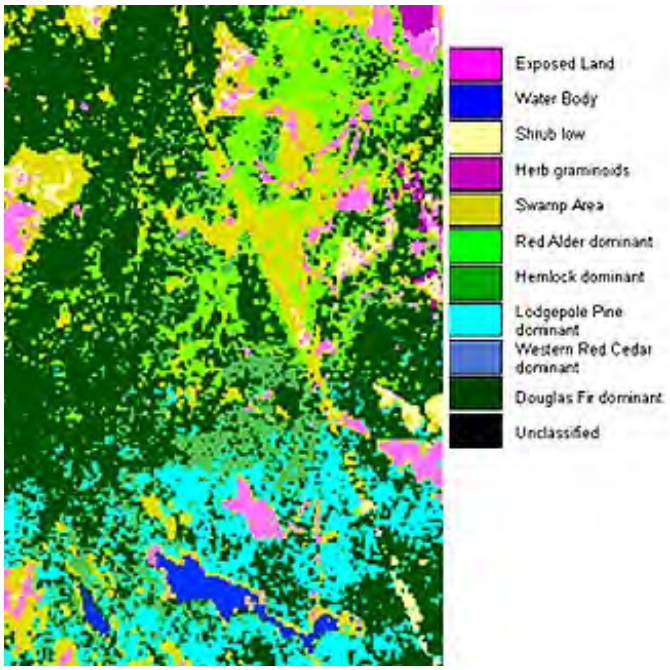
13-30: What do you think is meant by "nothing mapped"? ANSWER

Both vegetation and non-vegetation classes are present in this hyperspectral image of the Kunugawa River valley in Japan:

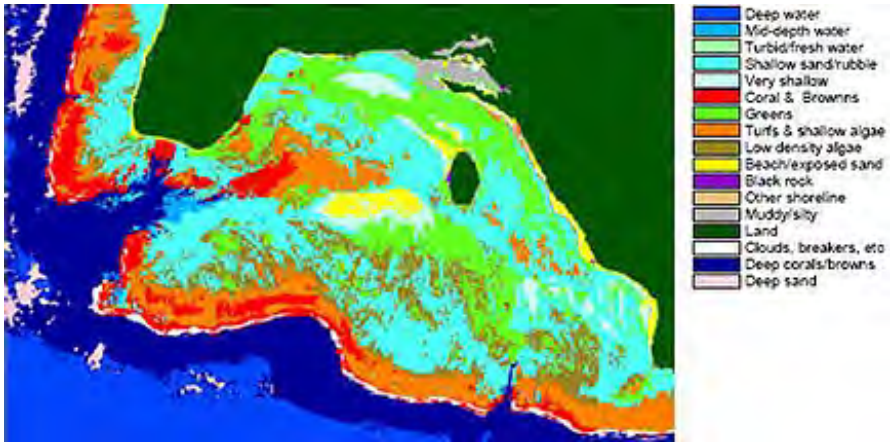


The themes shown are: Medium blue = watermelon; Dark blue = Marigold; Purple = Pumpkin; Wine Red = Grass; Green = Maize; Brown = Trees; Yellow, Dark Red, Dark Peach = 3 soil types; Dark Gray = Concrete: Red = Built up.

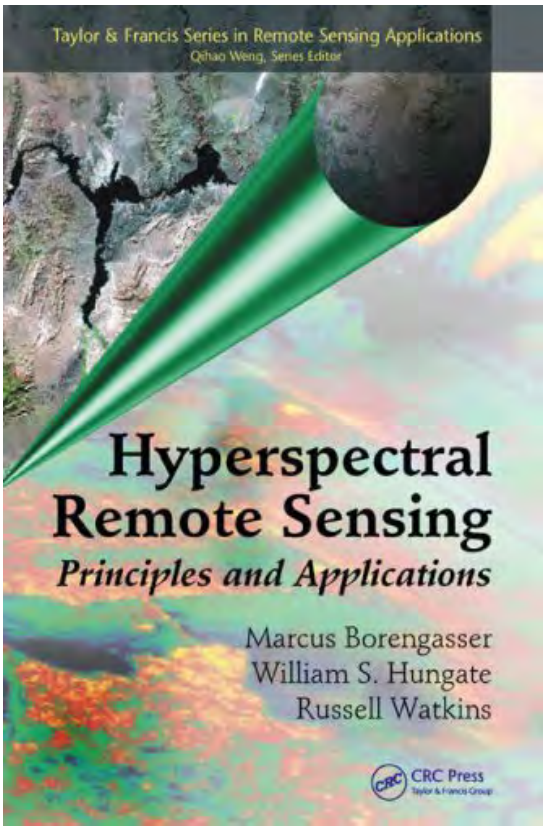
Forest types also are amenable to individual species identification:



Categorizing the classes found in shallow water off a coral reef coastline is another application of hyperspectral data:



Along with the advent of high resolution commercial satellites, hyperspectral remote sensing stands at the top of the most significant new space-centered technologies of the 21st century. As evidence of this, new textbooks devoted entirely to this subject are being published:



The chief beneficiary of this capability is the quality and accuracy of scene content classification.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

This Section will treat three apparently diverse topics, although a relationship among them will be established. The first reviews the types of ancillary data - those that provide supporting and explanatory information about what is in a remote sensing image - that are often called "ground truth". Then, the concept of the "multi-" approach is discussed, with the various forms involved explained and exemplified. The third covers the principles behind hyperspectral imaging remote sensing. On this first page, the rationale for gathering ground truth, including observations taken on site in the field, is spelled out. Some of the broad variety of surface collected data and information analysis are listed, along with examples of specific field studies that require appropriate measurements. The role of ground truth in picking out training sites and the role of returning to the field to verify classification accuracy are mentioned.

GROUND TRUTH; THE "MULTI" CONCEPT; IMAGING SPECTROSCOPY

RATIONALE FOR SURFACE OBSERVATIONS AND DATA COLLECTION

We have demonstrated so far in this Tutorial that remote sensing is an efficient way to gather large amounts of information from vast areas without being on the observed surface. But, interpreters will seldom apply this knowledge effectively unless they have first-hand familiarity with the surface of interest, or at least, with models of the surface. They gather this intelligence by several means: from circumspect field observations, judicious investigations at training sites, sophisticated measurements in the laboratory, on the ground, in the air, and from space, and, ultimately, from a rigorous mathematical analysis of the data to test for validity and correlation.

In remote sensing, ground truth is just a jargon term for at-surface or near-surface observations made to confirm their identification/classification by interpretation of aerial or satellite imagery. The term in its simplest meaning refers to "what is actually on the ground that needs to be correlated with the corresponding features in the visualized scene (usually as depicted in a photo or image)". Some purists dislike the term - "reference data" is an alternative, but this fails to convey the sense of "onsite observations and measurements". As applied generally to any planetary body, the term connotes gathering data on-site and deriving information therefrom that properly characterizes states, conditions, and parameters associated with the surface. A concise overview of "ground truth" as a term and concept, which supplements this and the next four pages is found at this Wikipedia website.

The purpose in acquiring ground truth is ultimately to aid in calibrating and interpreting remotely recorded surveys by checking realities within the scene. Since human interpreters normally experience Earth as ground dwellers, their view of the world from a horizontal or low-angle panorama is the customary frame of reference. In fact, the remote sensing specialist and the novice should retain a surface-based perspective during all phases of data collection, analysis, and applications because they will implement most interpretations and decisions dealing with natural resources and land use at the ground level.

There are three main reasons to conduct ground truth activities: 1) to obtain relevant data and information helpful as inputs and reference in the analysis of a remotely sensed scene; 2) to verify that what has been identified and classified using remote sensing data is actually correct

(accuracy assessment); 3) to provide control measurements, from targets of known identities, that are useful in calibrating sensors used on the observing platforms. The first reason is usually done prior to data acquisition, but can be done shortly thereafter (longer delays would compromise most classification efforts because of seasonal effects); the second reason is normally tied to visits to the classified scene soon after data acquisition; the third reason is less sensitive to time factors - calibration is largely independent of 'when' it is done. One goal involving reason 1 is to provide reference spectral signatures for the major classes found in a scene.

A few examples of the process of acquiring ground truth should clarify the paragraphs above. Let us assume that we have a Landsat image. We see various features of interest within the image. Some we can identify by their shapes, spectral character, or other descriptors. Other patterns are of uncertain nature. If the image is of some region near us, we can actually get into a vehicle, drive into the region, locate ourselves relative to the pattern, and look about, actually spotting image features as their counterparts on the ground, and in most instances, recognizing and identifying them. With this information, we can return to our base, get the scene data in digital format, choose all or part of the image to classify, and specify the classes of interest by having found areas in the scene that they identify with. These become "training sites". The resulting classification map shows the distribution of the features/classes everywhere within the scene or subscene we choose to examine; included are the training sites and other areas assigned to each class. We can then go back into the "field" and visit a select number of other areas to find what is there. These areas either are what we expect if the classification is a good one or they aren't what the classification process says they are - they are misclassified, that is, their identities are erroneous. By tabulating these "hits" and "misses" we determine the accuracy of classification as a percentage denoting correct identities. Most such ground truthing is "after the fact", done in a time frame after the image was acquired.

But there is also the option of realtime ground truthing. On the day of a satellite pass - assuming that the user has a vested interest in studying the scene on that date - a field crew can be somewhere in the scene. Various tasks can be performed: taking spectral measurements of several classes, determining atmospheric properties, getting actual temperatures, specifying irradiances, determining bidirectional reflectances for key features, quantitatively estimating silt loads in passing rivers, measuring soil moisture. These are some of the varied data obtainable "on the spot" or "in situ" that could aid in image interpretation.

The writer (NMS) remembers distinctly two field experiences that are classic examples of obtaining ground truth. He and a colleague took a boat onto the Potomac River a day after the October 10, 1972 Landsat-1 overflight above Washington, D.C. The purpose was to measure the amount of mud and silt carried by the river right after runoff from a big storm the previous day. This served to calibrate the streamwater status in the image acquired on the 10th, one of the prime objectives of the investigation.

The other instance involved the writer (NMS) in the late summer of 1973, while going into the field to check on an unsupervised classification made by the LARSYS (Purdue University) processing system and to designate new training sites for a subsequent supervised one. The classified area centered on glacial Willow Lake on the southwestern flank of the Wind River Mountains of west-central Wyoming. Prior to arriving onsite, I prepared a series of computer-generated printouts (long since misplaced), in which an alphanumeric symbol represented each spectral class (separable statistically but not identified). Different clusters of the same symbols suggested that discrete land use/cover classes were present. In the processing, I allowed the total number of classes to vary. Printouts with seven to ten such classes looked most realistic. But there was no *a priori* way to decide which was most accurate. In touring the area, I had to revise or modify my preconceived notions about classes. I had not considered the importance of grasses and of sagebrush, nor anticipated clumps of trees below the 79 m (259 ft) resolution of the Landsat Multispectral Scanner. After a tour through the site, I gazed over the scene from a slope top and tried to fit the patterns in the different maps into the distribution of features on the ground. The result was convincing: the eight-class map was mildly superior to the others.

Without this bit of ground truth, I would not have felt any confidence in interpreting the map and deriving any measure of its accuracy. Instead, what happened was a "proofing" of reliability for a mapped area of more than 25 square kilometers (about 10 square miles) through a field check of a fraction of that area. In this way my confidence in the capability of Landsat remote sensing data for identifying the contents of the real world was reenforced. This improved even more when, back at Goddard, I reclassified the test site - this time by a supervised method - using the field observations to provide better training sites. The resulting map appeared to increase class accuracy by about 20%.

The Table below summarizes the types of tasks and operations associated with obtaining and using ground truth data:

Table:

Role of Ground and Aircraft Observations in Supporting Satellite Remote Sensing

- Correlate surface features and localities as known from familiar ground perspectives with their expression in satellite imagery
- Provide input and control during the first stages of planning for analysis, interpreting, and applying remote sensing data (e.g., identifying landmarks, logistics of access. etc.)
- Reduce data and sampling requirements (e.g., areas of needed coverage) for exploring, monitoring, and inventory activities
- Select test areas for aircraft and other multistage support missions (e.g., underflights simultaneous with spacecraft passes)
- Accurately survey ground control points to be used in image rectification and other geometric corrections
- Develop standard sets of spectral signatures by using ground-based instruments
- Identify classes established by unsupervised classification
- Select and categorize training sites for supervised classification
- Verify accuracy of classification (error types and rates) by using quantitative statistical techniques
- Obtain quantitative estimates relevant to class distributions (e.g. field size; forest acreage)

- Collect physical samples for laboratory analysis of phenomena detected from remote sensing data (e.g., water quality, rock types, and insect-induced disease)

- Acquire supplementary (ancillary) non-remote sensing data for interpretive model analysis or for integration into Geographic Information Systems

- Measure spectral and other physical properties needed to stipulate characteristics and parameters pertinent to designing new sensor systems

13-1: Assume you are working on a Landsat scene that is close enough for you to actually go into the field to examine firsthand the features contained within it. What, in your opinion, would be the most important task to carry out 1) before the satellite takes its image and 2) after you receive and process this image? ANSWER

Examples of typical observations and measurements conducted in the field, commonly as the remote sensing platform passes over, or shortly thereafter, include these:

1. Meteorological conditions (air temperature, wind velocity, humidity, etc.)
2. Insolation (solar irradiance)
3. On-site calibration of reflectance
4. Soil moisture
5. Water levels (stream gauge data)
6. Snow thickness
7. Siltation in lakes and rivers
8. Growth stages of vegetation
9. Distribution of urban subclasses
10. Soil and rock types

13-2: Field work is expensive, and often inconvenient. Yet some kinds of data are needed in near real time. Mention three in this category from the above list. How would you go about getting these critical data values if field work is not an option? ANSWER

Ground truth activities are an integral part of the "multi" approach, discussed later in this Section. This simply means gathering the data under different conditions, such as the use of several sensors simultaneously and repeat coverage over time. We will explore the "payoffs" from this idea later in this Section.

Probably the most common reasons for conducting field activities are to select training sites prior to supervised classification or to identify key classes after unsupervised classification. The best way to collect field data, if feasible, is simply to spend a few days in the field, examining the terrain for the classification. Obviously, the scale of this effort depends on the size of the area we want to classify. One or more full Landsat scenes may require considerable travel and field time, whereas we can often examine a typical subscene (such as 512 x 512 pixels) in a day or two. If logistics or circumstances (e.g., an inaccessible foreign area or during an off-season such as winter) limit field operations, then we may use instead aerial photography, maps, literature research, interviews with residents (perhaps over the Internet), etc. In practice, to specify training sites generally means integrating the following sources of information: direct observations, photo documentation, a variety of maps, personal familiarity.

Primary Author: Nicholas M. Short, Sr.

Some words of wisdom concerning strategies for picking training sites prior to computer classification are introduced. The idea behind signature extension as an alternative to extracting signature data from training sites is considered. The problem inherent in the term "mixed pixels" is explained with an illustration; appropriate solutions are hinted at.

Training Sites; Mixed Pixels

The primary reason for setting up training sites in situ is to determine and define Land Cover/Use categories prior to classifying (mapping) a scene obtained by space observations. We must carefully select the sites in sufficient number, size and shape, variety, homogeneity, and distribution to maximize the accuracy of classification in the imagery. The categories fall into three broad surface classes: vegetated (arbitrarily, more than 40% cover); non-vegetated (rock, soil, water, works of man); and topographic types. Sensor resolution may limit class size and specification of what a class actually consists of. The total number of sites depends in part on the number of classes, their diversity, and the areal dimensions of the scene. Usually, a minimum of three sites per class is sufficient, but if the class has notable variability, we may need more. Even for an entire Landsat scene, in which we may seek 15 to 25 classes, the 30 or more required training sites will commonly occupy about 1/20th of the image's area. In most instances, locating sites depends mainly on their accessibility and our ability to pinpoint them in the imagery. Sites associated with linear features (roads) or possessing recognizable interfaces with other features (e.g., fields) are usually most effective.

13-3: Given a satellite-mounted sensor operating in the Visible-Near IR that has a spatial resolution of 500 meters. State three classes or features that will be well suited to observations at this level. Mention two that won't. ANSWER

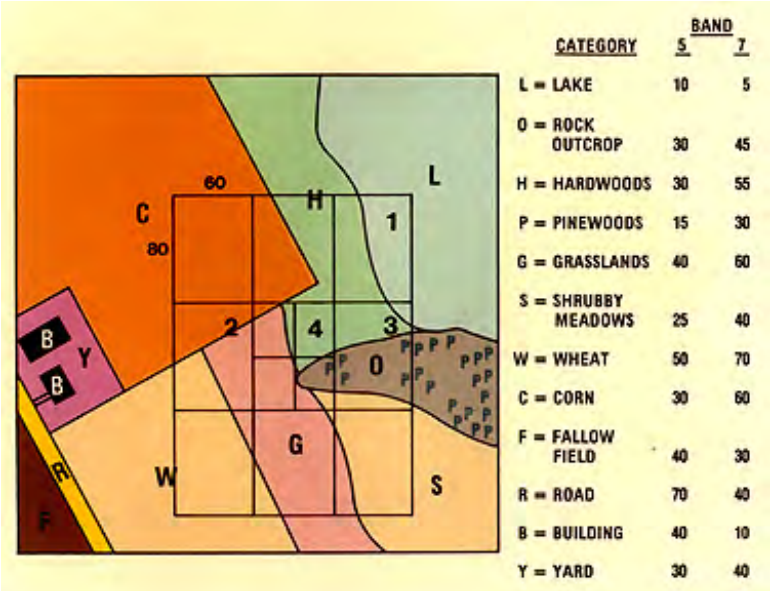
As was strongly implied on page 13-1, there are two general types of training sites: 1) those we actually visit so as to have firsthand knowledge of their suitability, and 2) those whose identities we deduce from photointerprettive experience. For cost and logistics reasons, the second approach is more common.

The alternative to depending on training sites for classification is to apply the concept of signature extension. This term refers to the assumption that we may define a single, fairly constant, spectral signature as characteristic of any class, and that this signature has broad (universal) applicability to any scene in a region, or even worldwide. As a specific example, the signature for mature winter wheat should be essentially the same for fields in the U.S. Great Plains, Argentina, the Ukraine, and Australia, provided we compensate for such variables as differing air masses, Sun position, soil types, soil moisture, etc. If that assumption proves true, then an unknown feature or class in a given scene anywhere should be classifiable by comparing its spectral properties (for a Landsat pixel, its multiband digital number [DN] values) to a data bank containing standard values for many classes. We assume the closest fit of the unknown's DN values to those of a class in the bank identifies it.

13-4: Look out your nearest window and gaze over the scene outside. Now, pretend you have a spectrometer available and wish to obtain spectral signatures for a data bank. Think of problems you might encounter. ANSWER

This approach usually works well (i.e., achieves an acceptable accuracy) for a few common features such as clear water bodies, clouds, snow, desert sand, several common rock types, certain forest types, and perhaps central urban areas. However, in most cases, signature extension is unreliable for such reasons as 1) the natural variability of most classes, 2) the "mixed pixel" problem (see below), 3) the often artificial or arbitrary way in which we set up many spectral classes (e.g., we classify rocks [name them] by mineral content and texture, which may bear no direct or simple relationship to gross spectral properties), 4) the influence of (usually undetermined) differences in atmospheric conditions from place to place and on different dates, 5) the seasonal variability of vegetation, and 6) the inability to account for and correct other variables.

A "mixed pixel" results from the fact that individual areas, consisting of several different features or classes may be below (smaller than) the resolution capability of the sensor. Consider this hypothetical "map" of a rural setting



CATEGORY	BAND 5	BAND 7
L = LAKE	10	5
O = ROCK OUTCROP	30	45
H = HARDWOODS	30	55
P = PINEWOODS	15	30
G = GRASSLANDS	40	60
S = SHRUBBY MEADOWS	25	40
W = WHEAT	50	70
C = CORN	30	60
F = FALLOW FIELD	40	30
R = ROAD	70	40
B = BUILDING	40	10
Y = YARD	30	40

13-5: In the above figure, pick out the two "worst" pixels in terms of mixture. ANSWER

In this instance, we treat each category as though it were more or less homogeneous. As imaged by a sensor whose instantaneous field of view (IFOV) (controlled by optics and sampling rates) leads to a small pixel size, if an individual pixel happens to lie completely within, or fortuitously coincides with, the boundaries of a given class, then the multiband spectral properties of the dominant material(s) in the enclosed class will determine the DNs for that pixel. It is more likely, however, that the pixel will straddle or cut across several class or feature boundaries. The resulting spectral content is then a composite or weighted average of the spectral responses from each internal class. Recognition of each feature or class becomes difficult, since there are two primary unknowns to account for - the identity of the class and its relative proportion in the mix. Mathematical methods are available to solve for these unknowns, but there always remains some statistical uncertainty. **An obvious improvement is to reduce pixel size (increase resolution), as is done in the central rectangle above, so that more pixels fall within the space occupied by a single class/feature and fewer pixels cross boundaries.** Going in the other direction, note the effect of enlarging the pixel, say, to the size of the outer boundary of the cluster of nine. The key rule in optimizing classification is to seek a resolution that approximates the sizes of the smallest specific classes whose identities we seek. It follows that accuracies of classification using, say, AVHRR images will be low unless the classes chosen are spatially large (e.g., a lake or a mountain range). Landsat should achieve superior accuracies and IKONOS even better because they can commonly pick out the smaller features as discrete entities within their pixel sizes.

Mixed pixels are usually the biggest reason for lowered success in classification accuracy. There are a variety of factors affecting the mixing. This is an important statement, so let us explore this idea a bit further. We'll consider two cases:

First, start with a small empty parking lot (for example, 50 meters on a side) covered uniformly by dark asphalt. If a lower resolution pixel (50 meters) on the remote sensor just happens to coincide with the lot as the scene is being imaged, a single uniform spatial measurement of, say, the lot's low reflectance for a given spectral spread will be recorded. If instead the lot is next to a large field of green grass and the pixel straddles the lot and the field at a 50-50 areal coverage, the measurement will be an average of the spectral response (wavelength peaks, etc.) for each class at this proportion. If the straddling were 75-25 for field and lot respectively, the measurement will differ, and so forth for different proportions. Thus the mixing varies, and the amount might not be decipherable. Now, let's make this a bit more complicated. We will populate the parking lot with crowds of people in small groups of varying numbers distributed haphazardly over the lot. The clothes colors in each group are allowed to vary. If the pixel size is still 50 meters, the distribution and color variance won't matter, as the groups contributions to the measurement are integrated. But now lets make the pixel size 5 meters instead, so that the sensor uses 100 pixels to sense the scene. In this case, the values of radiance or reflectance will vary from pixel to pixel as different groups are measured. Same scene content; different results. The 5 meter situation describes the improved information content that higher resolution affords. There is less uncertainty owing to mixing since each pixel is likely to encompass a group with its own distinctive character.

The second case is more a matter of semantics than of the physics of spectral and/or spatial remote sensing. Let us remotely sense a volcanic flow made up of dark basalt. A geologist in the field might map this flow as a single unit - basalt. Or, if it suits his/her purpose, the map might be more exacting, showing variants of the basalt that occur within a flow: the basalt could be dense, or vesicular, or of the aa or pahoehoe types, or show still other variations that are given descriptive names. In the field, on the ground, these variants can be assessed and mapped by the observer. Spectrally, in terms of material composition they are fairly similar but they are distinguishable by their shapes or extent of reflection, etc. When sensed remotely, they are probably indistinguishable to all but the highest resolution sensors. The point is this: what is being imaged, and hence mapped, depends in part on how definitive are the classes involved. If one needs only to establish the classes as trees, crops, rock outcrops, etc. the classes are "coarse" and hence the degree of spectral and spatial resolution does not have to be high. But if the types of trees (perhaps to species level), the identity of crops (corn versus wheat, etc.), the variety of rocks (limestones, sandstones, granites, etc.) are being sought, then the extent of mixing becomes a major factor. Higher spectral/spatial resolution becomes a requirement. More information - more detail - comes at a price: better remote sensors must be used.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Several kinds of errors - mainly those of "commission" or "omission" - are discussed as a basis for setting up an accuracy assessment program. Accuracy itself is defined and the point is made that much depends on just how any class, feature, or material being classified is meaningfully set forth with proper descriptors. Two factors are important in achieving suitable (hopefully, high) accuracy: spatial resolution (which influences the mixed pixel effect) and number of spectral bands involved in the classification. A case study in the Elkton, Maryland area is treated in some detail to demonstrate just what the accuracy of its classification in Landsat data sets really means.

Accuracy Assessment

Of course, in the real world many classes or features are not homogeneous, that is made of one material and formed in one definitive shape. Consider the class "field". During a growing season, the field is a mix of soil, crop(s), and some degree of moisture (ephemeral). There are many types of soils, which vary in color, composition, and texture, and crops also range in variety and density (absence = fallow field). Another class is "urban," which can include a diversity of buildings made of different materials, in various sizes and shapes; roads formed of concrete or asphalt; trees and grass, and other variables. We can often further subdivide the classes into more specific categories, such as "field of tassled corn" or "shopping center", provided they correspond closely to unique or distinctive spectral signatures, as determined in establishing prototypical training sites. This internal mix of several substances or features that are intrinsic to a class does not have the same meaning as the resolution-dependent straddle-mix of several classes described on the previous page as the "mixed pixel" problem.

13-6: Consider this class: "factory complex". In this case, it occupies a non-residential part of suburbia. Break it down into the likely components (internal mix) that are present, even though it is implied by its name to be a single unit. ANSWER

We may *define* accuracy, in a working sense, as *the degree (often as a percentage) of correspondence between observation and reality*. We usually judge accuracy against existing maps, large scale aerial photos, or field checks. We can pose two fundamental questions about accuracy: Is each category in a classification really present at the points specified on a map? Are the boundaries separating categories valid as located? Various types of errors diminish the accuracy of feature identification and category distribution. We make most of the errors either in measuring or in sampling. Three error types dominate:

I. Data Acquisition Errors: These include sensor performance, stability of the platform, and conditions of viewing. We can reduce them or compensate for them by making systematic corrections (e.g., by calibrating detector response with on-board light sources generating known radiances). We can make corrections, often modified by ancillary data such as known atmospheric conditions, during the initial processing of the raw data.
II. Data Processing Errors: An example is misregistration of equivalent pixels in the different bands of the Landsat Thematic Mapper. The goal in geometric correction is to hold the mismatch to a displacement of no more than one pixel. Under ideal conditions, and with as many as 25 ground control points (GCP) spread around a scene, we can realize this goal. Misregistrations of several pixels significantly compromise accuracy.

III. Scene-dependent Errors: As alluded to in the previous page, one such error relates to how we define and establish the class, which, in turn, is sensitive to the resolution of the observing system and the reference map or photo. Mixed pixels fall into this category.

Three examples of these errors come from a common geologic situation (also treated to some extent on page 2-5), in which we process the sensor data primarily to recognize rock types at the surface. In this process there are pitfalls:

First, what geologists in the field normally map is bedrock, but over large parts of a surface, soil and vegetation cover or mask the bedrock at many places. The geologist makes logical deductions in the field as to the rock type most likely buried under the surface and shows this on a map, thus treating these masking materials as invisible (ignored). This treatment, unfortunately, does not correspond to what the sensor sees.

Second, most geologic maps are stratigraphic rather than lithologic, i.e., they consist of units identified by age rather than rock type. Thus, the map shows the same or similar rock types by different symbols or colors, so that checking for ground truth requires converting to lithologies (often difficult because a unit may be diverse lithologically but was chosen for some other mode of uniformity). This is an "apples and oranges" situation: the remotely-sensed classified map shows rock types whereas the available reference map shows stratigraphic units.

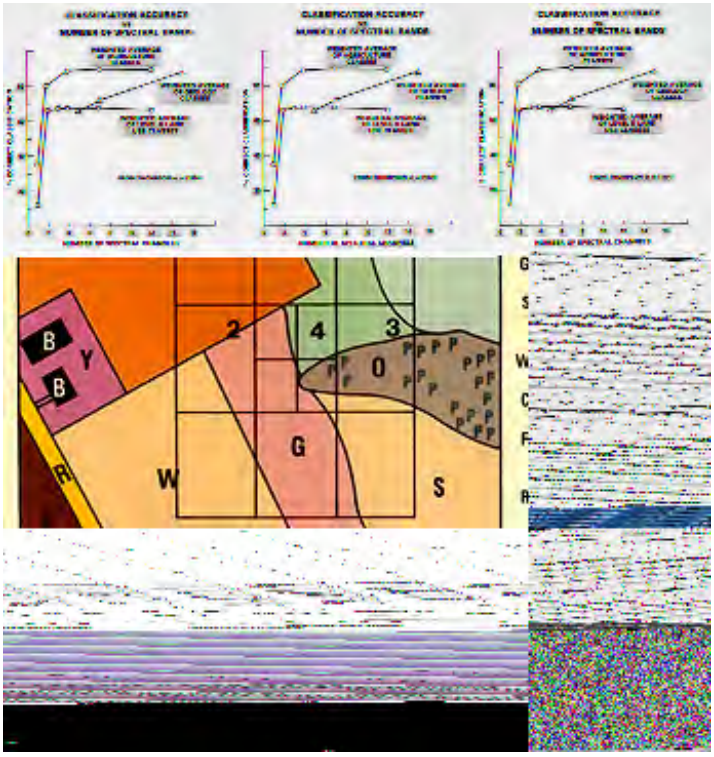
Third, we may need to consider a rock type in context with its surroundings to name it properly. For example, granite, and the sedimentary rock called arkose, derived from it, have similar spectral properties. The latter, however, typically appears in strata, because it is a deposited formation, whose spatial patterns (especially when exposed as folded or inclined layers) are usually quite distinct from those of massive granites and are often revealed by topographic expression.

Another, often overlooked point about maps as reference standards concerns their intrinsic or absolute accuracy. Maps require an independent frame of reference to establish their own validity. For centuries, most maps were constructed without regard to assessment of their inherent accuracy. In recent years, some maps come with a statement of confidence level. The U.S. Geological Survey has reported results of accuracy assessments of the 1:250,000 and 1:1,000,000 land use maps of Level 1 classifications (see page 4-1), based on aerial photos, that meets the 85% accuracy criterion at the 95% confidence level.

As a general rule, the level of accuracy obtainable in a remote sensing classification depends on diverse factors, such as the suitability of training sites, the size, shape, distribution, and frequency of occurrence of individual areas assigned to each class which together determine the degree to which pixels are mixed, the sensor performance and resolution, and the methods involved in classifying (visual photointerpreting versus computer-aided statistical classifying), and others. A quantitative measure of the mutual role of improved spatial resolution and size of target on decreasing errors appears in this plot:

The dramatic improvement in reducing errors ensues for resolutions of 30 m (98 ft) or better. This relates, in part, to the nature of the target classes. Coarse resolution is ineffective in distinguishing crop types, but high resolution (< 20 m) adds little in recognizing these other than perhaps identifying species. As the size of crop fields increases, the error decreases further. The anomalous trend for forests (maximum error at high resolution) may be the consequence of the dictum: "Can't see the forest for the trees". Here, this saying means that high resolution begins to display individual species and breaks in the canopy that can confuse the integrity of the class "forest". Two opposing trends influence the behavior of these error curves: 1) statistical variance of the spectral response values decreases whereas 2) the proportion of mixed pixels increases with poorer resolution.

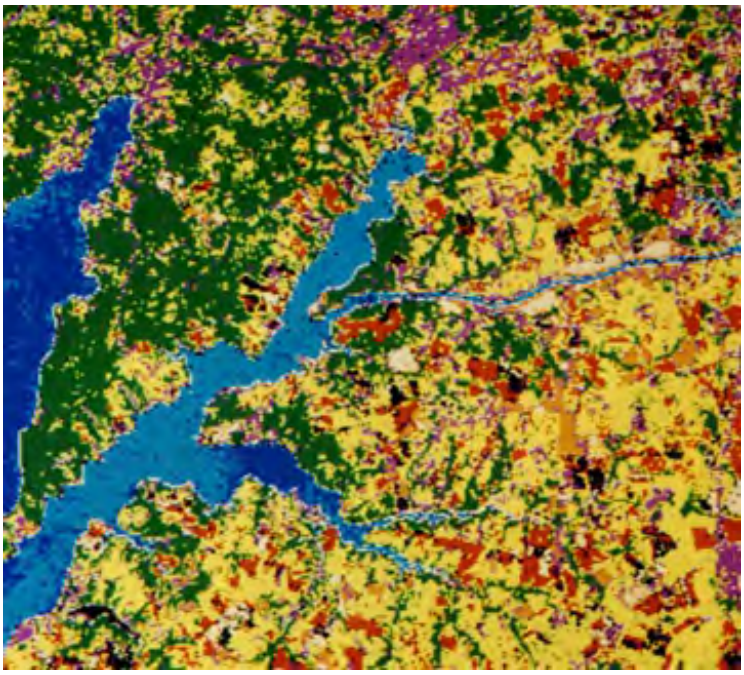
A study of classification accuracy as a function of the number of spectral bands shows these trends:

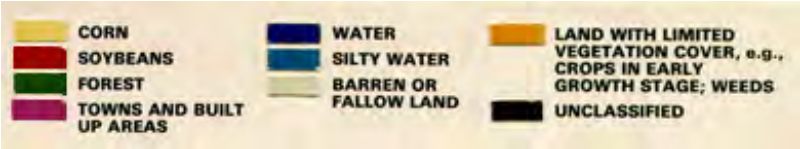


The increase from one to two bands produces the largest improvement in accuracy. After about four bands, the accuracy increase flattens or increases very slowly. Thus, extra bands may be redundant, because band-to-band changes cross-correlate (this correlation may be minimized and even put to advantage through Principal Components Analysis). However, additional bands, such as TM bands 5 and 7, can be helpful in identifying rock types (geology), because various rock types absorb certain wavelengths, which helps identify them in these spectral intervals. Note that the highest accuracy associates with crop types because fields, consisting of regularly-space rows of plants against a background of soil, tend to be more uniform .

In practice, we may test classification accuracy in four ways: 1) field checks at selected points (usually non-rigorous and subjective), chosen either at random or along a grid; 2) estimate (non-rigorous) the agreement of the theme or class identity between a class map and reference maps, determined usually by overlaying one on the other(s); 3) statistical analysis (rigorous) of numerical data developed in sampling, measuring, and processing data, using tests, such as root mean square, standard error, analysis of variance, correlation coefficients, linear or multiple regression analysis, and Chi-square testing (see any standard text on statistics for an explanation of these tests); and 4) confusion matrix calculations (rigorous). We explain this last approach using the author's study of a subscene from a July, 1977, Landsat image that includes Elkton, Maryland (top center).

I acquired from the EPA a 1:24,000 aerial photo that falls within this subscene. Starting with a field visit in August, 1977, during the same growing season as the July overflight, I identified the crops in many farms located in the photo, from which I selected about 12 as training sites. Most were either corn or soybeans, and others were mainly barley and wheat. I then ran a Maximum Likelihood supervised classification, as shown below, and printed as a transparency.





I overlaid this transparency onto a rescaled aerial photo until field patterns approximately matched. With the class identities in the photo as the standard, I arranged the number of pixels correctly assigned to each class and those misassigned to other classes in the confusion matrix used to produce the summary information shown in the Table below (the format is known as an Error Matrix), listing errors of commission, omission, and overall accuracies. Errors of *commission* result when one incorrectly identifies pixels associated with a class as other classes, or when one improperly separates a single class into two or more classes. Errors of *omission* occur whenever the user simply doesn't recognize pixels that should have identified as belonging to a particular class.

13-7: Why are the mapping accuracies for individual classes lower than that of the overall classification? ANSWER

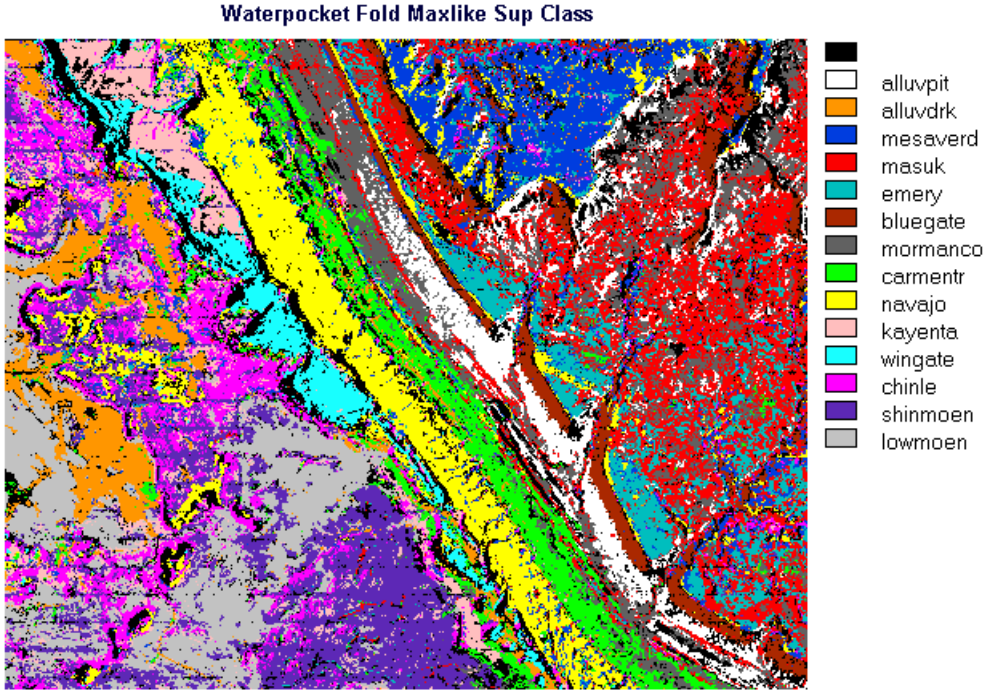
13-8: Discuss and explain what is meant by errors of commision and omission for corn. ANSWER

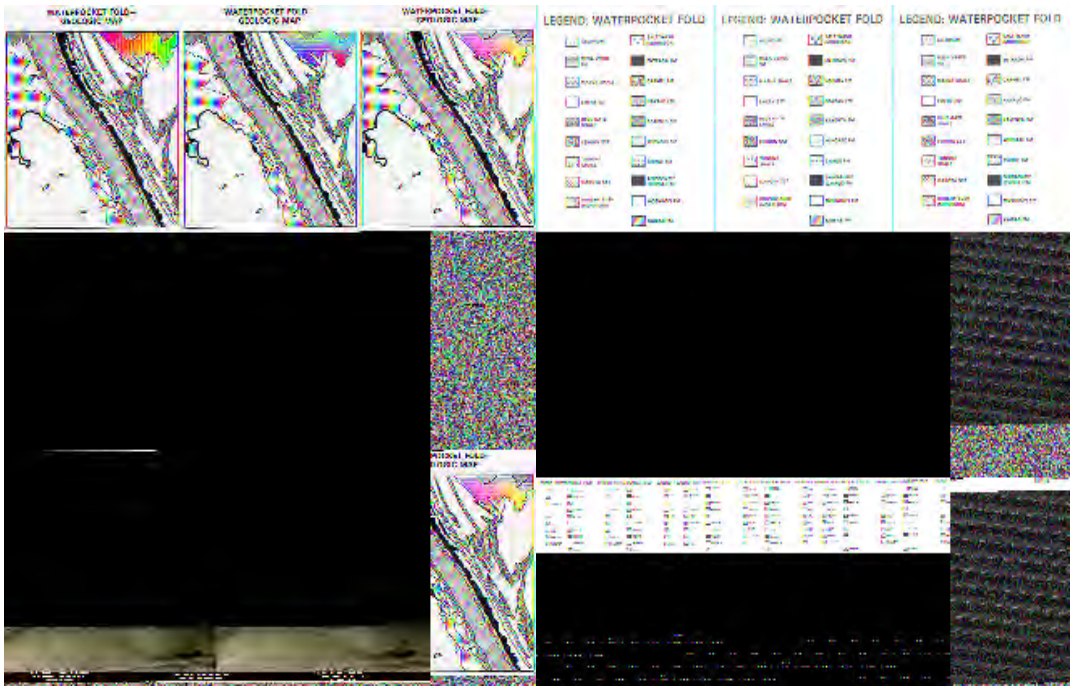
13-9: List at least five scene-dependent sources of error for the above case. ANSWER

Mapping accuracy for each class is the number of correctly identified pixels within the displayed area, divided by that number plus error pixels of commission and omission. To illustrate, in the table, of the 43 pixels classed as corn by photointerpretation and ground checks, we assigned 25 of these to corn in the Landsat classification, leaving 18/43 = 42% as the error of omission. Similarly, of the 43, we improperly identified 7 as other than corn, producing a commission error of 16%. After we determine these errors by reference to "ground truth", we can reduce them by selecting new training sites and reclassifying them, by renaming classes or creating new ones, by combining them, or by using different classifiers. With each set of changes, we iterate the classification procedure until we reach a final level of acceptable accuracy.

Another example of the accuracy of classification was considered earlier in the Tutorial on pages 2 to 5 in Section 2. Review these, and look especially at the various maps, before reading the remainder of this page. What we will now discuss is first how that classification was made and second how the errors involved depended on the manner in which the classes were chosen and defined. The writer chose to defer this discussion until the ideas put forth on the present page were examined.

Two maps in Section 2 are the key ones involved in the accuracy assessment. They are reproduced here:

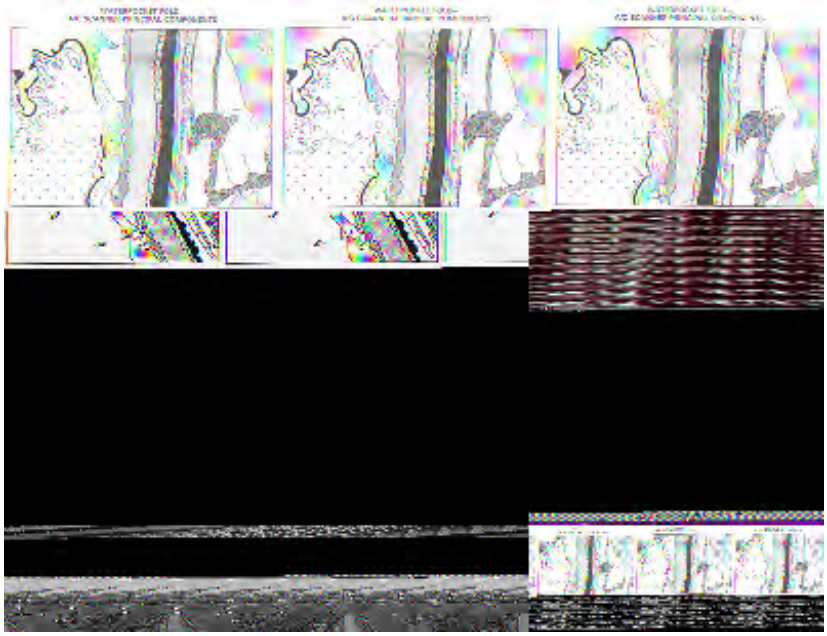




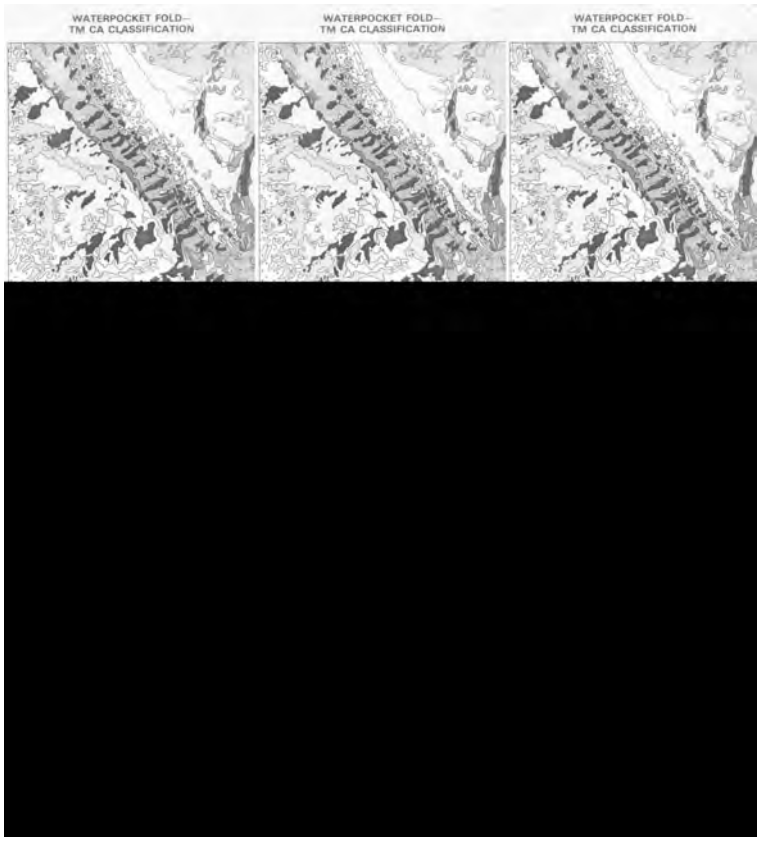
The top map was the result of a maximum likelihood supervised classification using the 6 non-thermal bands in the Landsat TM data. The bottom map was made by the writer using a published U.S. Geological Survey map for the stratigraphic formations at and near the fold, plus information on alluvium gleaned from aerial photos acquired during the mission. That map was digitized, rectified, and overlaid on the classification map within the IDIMS computer system. Then, on a pixel to pixel basis, the correspondence between the upper and lower map units was assessed, leading to an estimate of the errors of commission.

This statement was extracted from page 5 of Section 2: "The overall accuracy for the 6 band classification was 69.4%. However, maximum accuracies were 97, 94, 88, 86, and 85 percents for the Carmel, Moenkopi, Kayenta, Lower Chinle, and Navajo formations respectively, in the TM data set. The Masuk and Tununk Formations had low accuracies in the 20% range."

The supervised classification map below covers approximately the same scene but was made using selected bands from the high resolution 24-channel Bendix scanner that imaged the Waterpocket fold during a June 1976 mission directed by the writer.



Visual comparison between this map and the TM color one indicates a higher level of accuracy for the Bendix version. There are two reasons for this: First, the Bendix map is based on more (narrower wavelength) band inputs, which affords better spectral resolution. Second, the spatial resolution of the Bendix scanner data is much higher than that of the TM data. Both factors are known to improve classification accuracy. Now, consider this next map, made from a TM scene acquired in January 1976:



The accuracy of geologic units identification for this winter image was just 46.7%. This suggests that low Sun angle is another factor that affects the level of accuracy.

There are, in fact, a number of other factors on which accuracy depends. We will mention one that can be critical.

This is simply how the classes themselves are defined and set up. To elucidate: Suppose that the stratigraphic units that are being classified and mapped are not well exposed as the hard

rocks they actually are. They could be soil covered; they could be largely covered by vegetation, either everywhere or in spots. They could be exposed at high angles (cliffs). The training sites used in conducting the classification could be inhomogeneous - say, vegetated in some places but not in others. These 'real world' variables will influence the ultimate accuracy levels. The person carrying out the classification procedure could compensate for that by setting up multiple, but related, classes, such as "vegetated unit", "soil-covered unit", "bare rock unit", etc. This could improve the accuracy but the map so produced might end up as a hodge-podge of 'too many' classes.

It should not be surprising then that accuracies of classification do not usually approach 100%. The levels reached are quite sensitive to the manner in which the classes are specified and to the choice of training site locations (and by implication to what is actually there). With the advent of higher resolution sensors and of hyperspectral sensors, the overall accuracies being achieved are on the rise.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The idea of the "multi" approach to remote sensing is introduced. For example, "multilevel" refers to a wide range of different altitudes or distances from the surface being observed. This first of four pages devoted to "multi's" concentrates on different altitudes that lead to different scales and resolutions. After other examples in the next two pages, this subsection closes with a single case study illustrating how various data inputs facilitate scene analysis

THE "MULTI" APPROACH TO APPLIED REMOTE SENSING

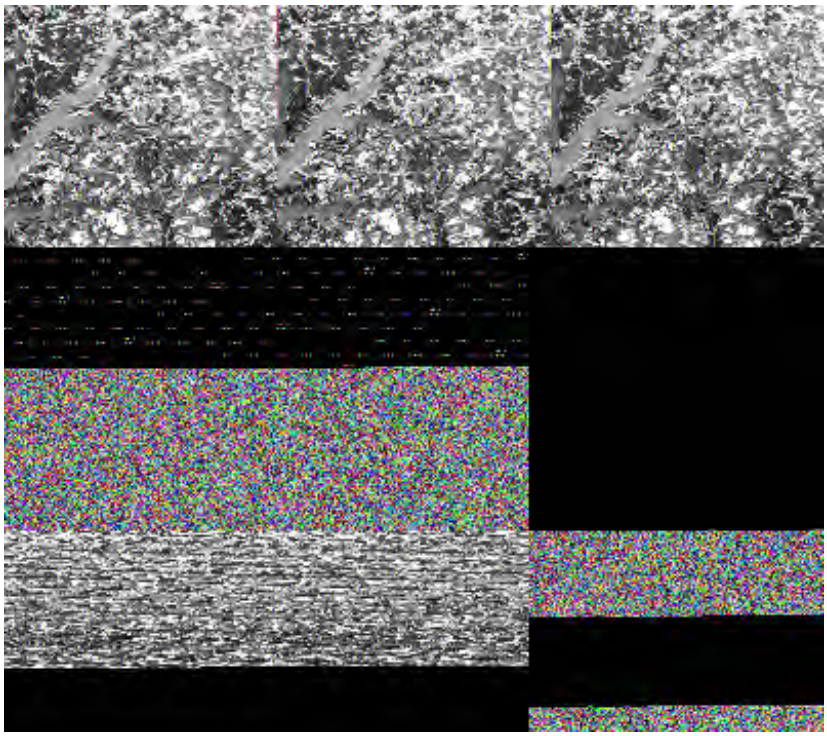
Multiplatforms and Multilevels

Ground truth activities are an integral part of the "multi" ("more than one") approach. Thus, we should procure data whenever possible from different platforms (*multiplatform*), at various altitudes (*multistage; multilevel*). (A "platform" is a synonym for any orbiting spacecraft, be it a satellite or a manned station, from which observations are made.) This gives rise to *multiscaled* images or classification maps. Ideally, we should aim to employ *multisensor* systems simultaneously to provide data, commonly at *multiresolution*, over various regions of the spectrum (*multispectral*). Often, we obtain data at different times (*mutitemporal)*, whenever seasonal effects or illumination differences are factors or change detection is the objective. Supporting ground observations should come from many relevant, but not necessarily interrelated, sources (*multisource*). Some types of surface data may correlate with one another and with other types of remote sensing data (*multiphase*).

13-15: Using your imagination and growing experience, design an experiment using as many of the above "multi's" as seems sensible. ANSWER

Many remote sensing investigations include several of the above "multi" categories but examination of both remote sensing textbooks and Internet sites generally does not highlight examples that include most of these together as applied to one study or application site. This proved the case in preparing this page. So, in order to illustrate the "multi" concept adequately, it is necessary to show images of different parts of the world that don't show the same piece of "real estate" sensed repeatedly by different systems. And, we mention now that most platforms in the last few years have multiple sensors on board operating simultaneously. The best example of that is Terra (page 16-9) which has five different but complementary sensors which often examine the same target.

Earlier in this Tutorial, there have already been individual and isolated examples of some of the "multi" types of images and photos. In the review erected on this page, we will develop one theme: agriculture, especially crop monitoring. We start with an expansion of the analysis of the Delaware/Chesapeake Bay classification shown on page 13-3, which was part of a NASA field study. Here we will follow the multi-level approach by looking successively at the farmlands around the Chesapeake and Delaware Canal, starting with a Landsat red band subscene and progressing to a low altitude photo.

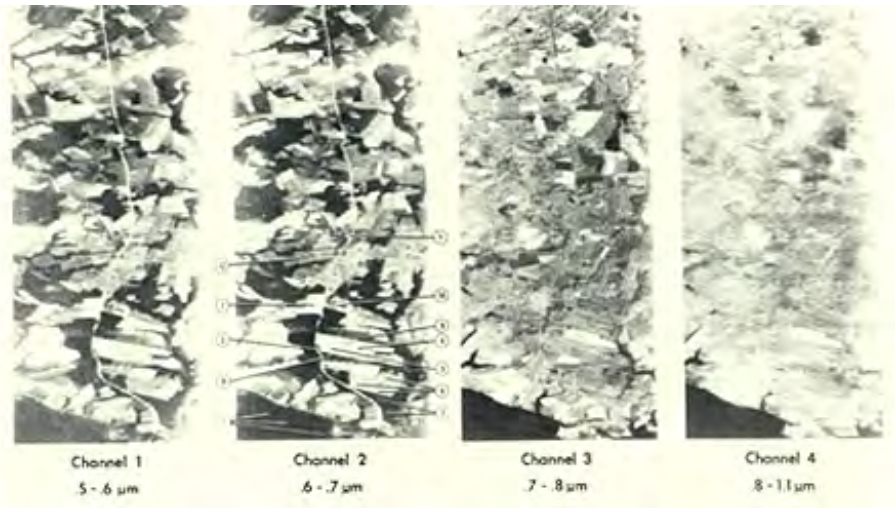


Next is a high altitude U-2 photograph of part of the above area; locate yourself using the canal.

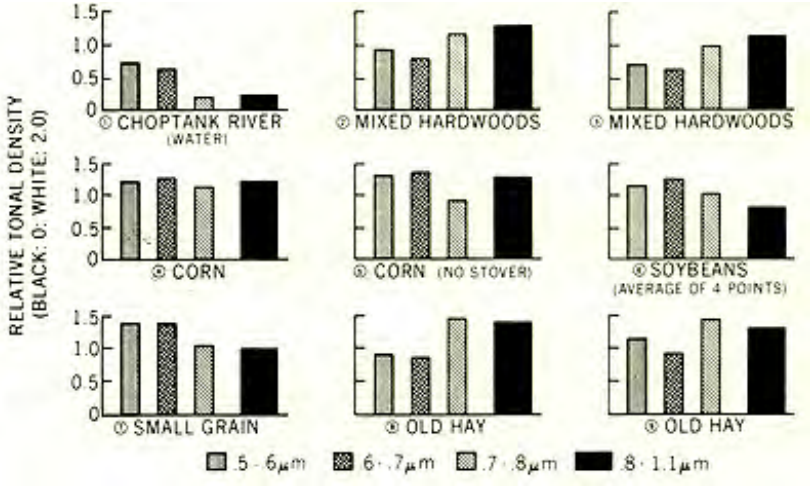
Now, to zoom in further, consider this medium altitude aerial photo which contains part of the canal.

Finally, here reproduced is a paper print of a low altitude aerial photo that was actually taken into the Delaware test site. The Soil Conservation Service's field agent has made notations showing characteristics and yield for some of the crop acreage.

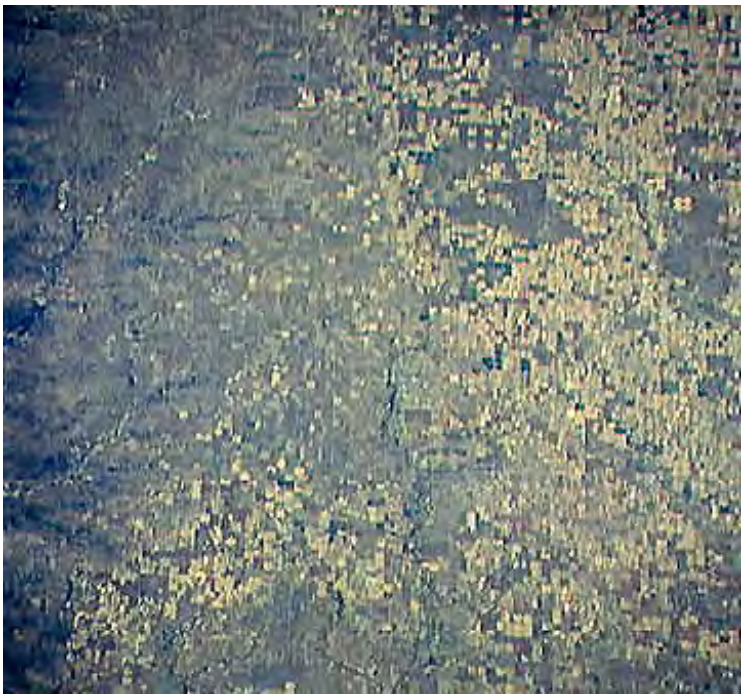
The writer (NMS) was a participant in this field study. As part of the preparation for the Landsat phase, NASA flew an aircraft mission with a multispectral scanner over fields in the Delmarva Peninsula to the south of the study site. Here are four images designed to simulate the 4 Landsat MSS bands:



From the data, an analog measure (using a photometer operating on a transparency) of the photo-density of selected fields in each of the MSS-equivalent bands led to this plot of relative darkness as a proxy for reflectance coming from the ground features and crops indicated:



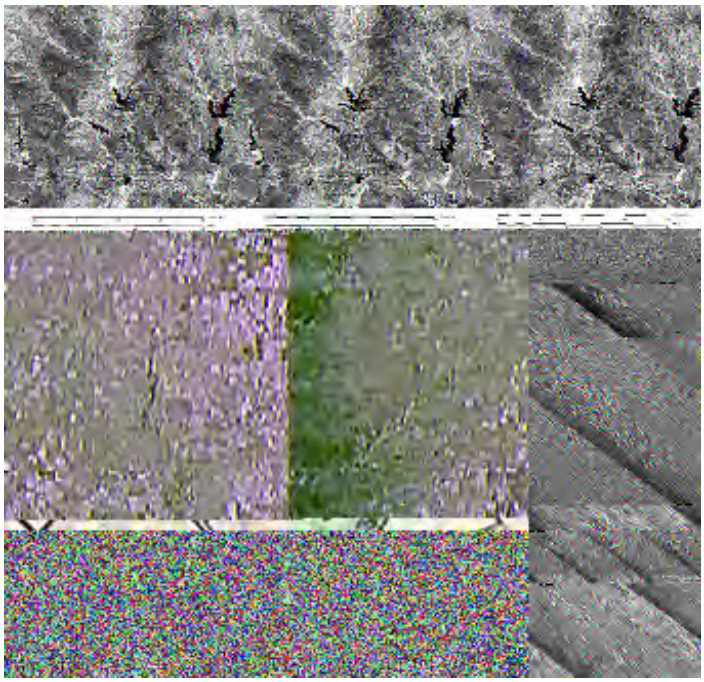
Let us turn from this specific study to some more general examples. Many of the photographs taken from the Shuttle by the astronauts have agricultural areas as their subject matter. Often these photos are not particularly good owing mainly to atmospheric problems. But this one covering the land around Enid, Kansas is one of the better.



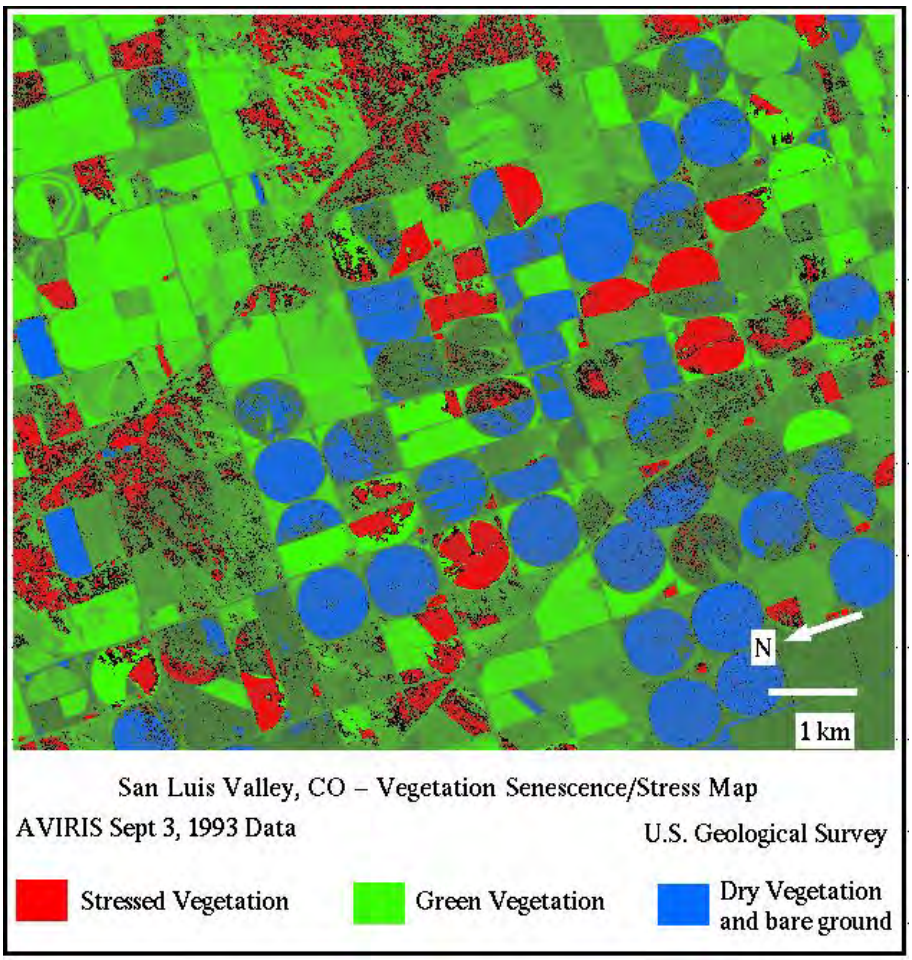
Other satellites produce excellent near-natural images of farmlands, such as this SPOT scene:

At higher resolution, here are fields in California's Great Valley near Fresno imaged by the IKONOS-2 satellite.

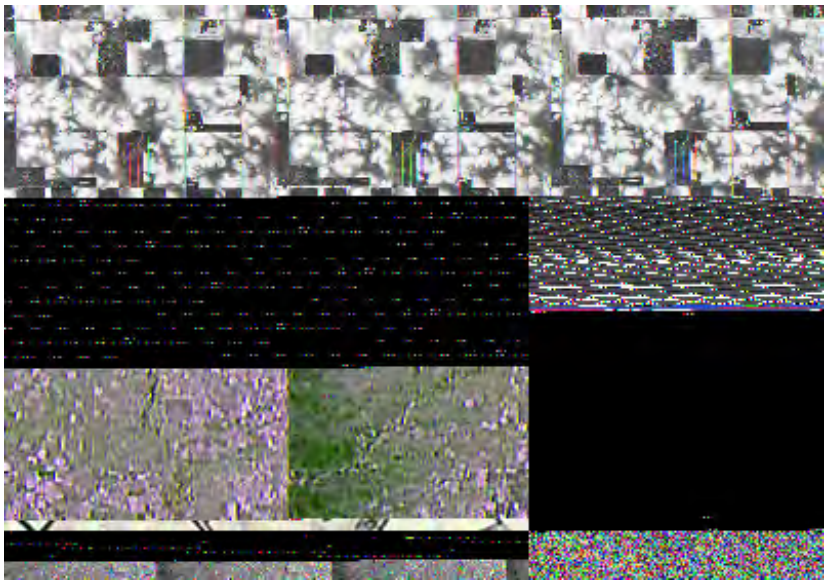
At the other extreme, the AVHRR, as demonstrated on page 3-4, is quite adept at providing small-scale, large area indications of crop and vegetation vigor, often expressed as NDVI. This next image is a black and white plot of the NDVI values (using channels 1 and 2) for the land in and around Dallas, TX. Light tones indicate high NDVIs.



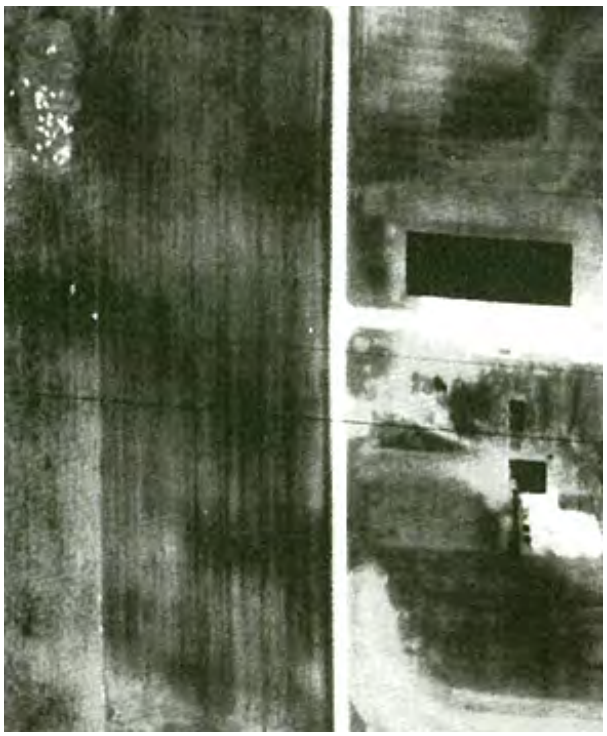
Crop stress results from insufficient soil and/or crop water (drought), improper nutrients; plant disease, insect infestations, and other factors. The next image is of cropland in Colorado's San Luis Valley. It was made by the AVIRIS sensor that will be described in detail on pages 13-9 and 13-10 of this Section. A classification of these Colorado crops is treated on page 13-10. Here we show AVIRIS hyperspectral data that use bands sensitive to crop moisture deficiency.



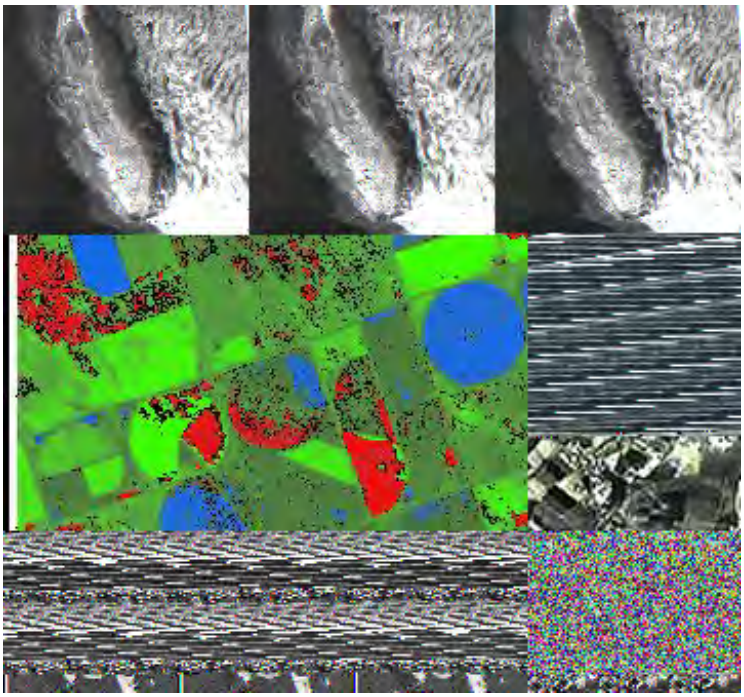
Soil moisture is one of the critical parameters a farmer needs to know in making decisions about planting conditions and need for irrigation. It is often the precursor indicator of potential or actual crop stress. This can be done through aerial photography, as shown here for some Indiana farms, but the cost of flying for specific water inventory is high.



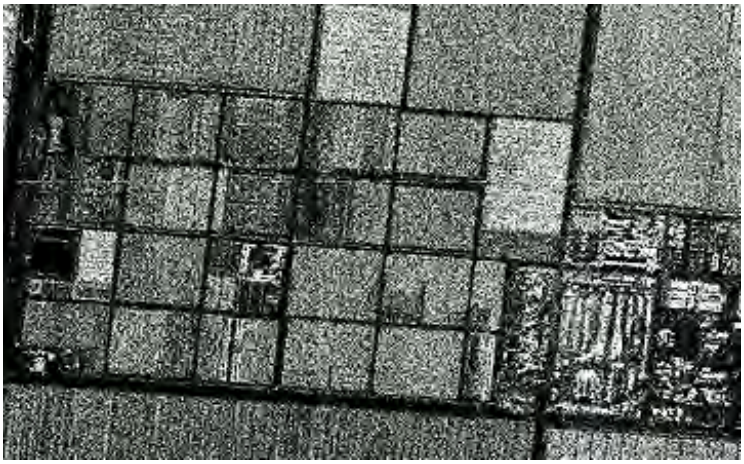
Thermal scanners are also good at detecting moisture, as indicated in this aerial image of a Wisconsin farm, taken around 9 PM at night shortly after the setting Sun. The bright spots in the upper left are a herd of (warm) cows. The black rectangle in the upper right is a sheet metallic roof on a farm outbuilding, which shows "cool" because of the very low emissivity of metal.



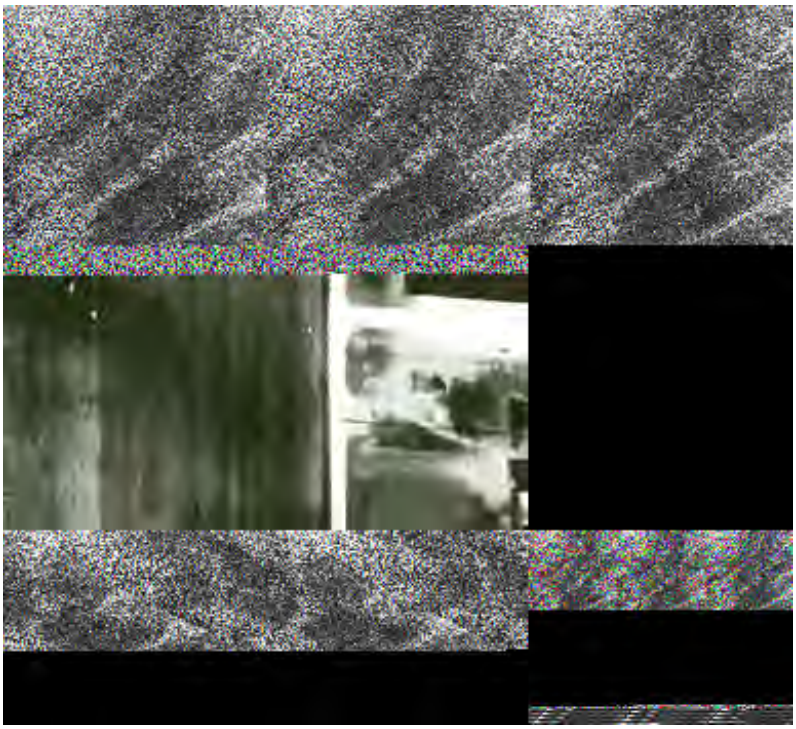
AVHRR thermal bands can also provide useful agricultural information. And so did HCMM when it operated. Here is a HCMM Day Thermal image of much of California, taken in May. Note the farm patterns in the central Valley. Note also the very dark area in the High Sierras - this is spring snow that will eventually provide water for the crops during summer meltdown.



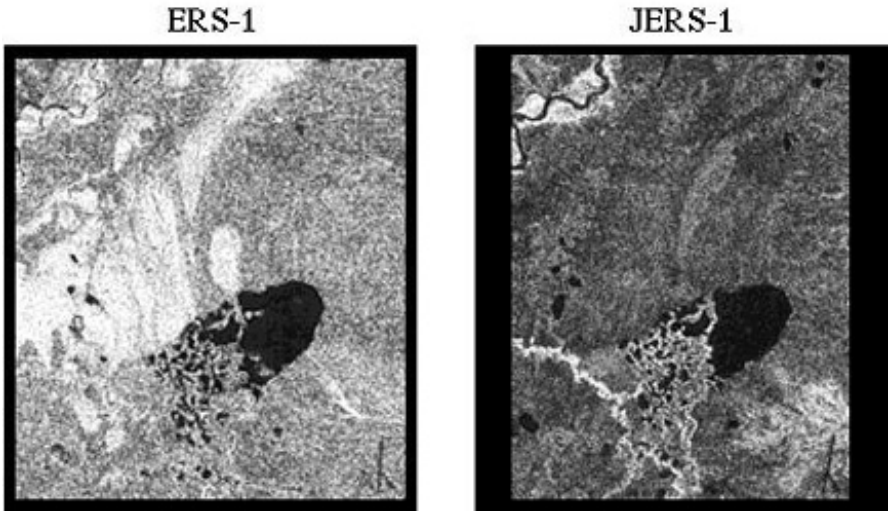
Radar is a good means of imaging farmland, as seen in this low altitude aerial radar mission over the Maricopa area near Phoenix, AZ:



Seasat radar imaged this next scene, in the Great Plains. Some fields are dark, others light, indicative of the stages of growth (light areas indicate crops that reflect more of the radar beam to the receiver). Of particular interest are the two dark patches which represent the effects of soil moisture (reduces returns) following two local thunderstorms passing over the plains.



As was put forth in Section 8, radar images made from different bands disclose information in each not expressed in the same way as in the other(s). Below is a pair of images of the Medicine Lake area in Alaska west if Fairbanks that were fortuitously imaged 18 minutes apart by two different satellites. On the left is a ERS-1 radar C-band image; on the right is a JERS-1 L band image. Note that the ERS-1 image renders some bogs in bright tones; the JERS-1 image highlights creek beds.



On the next two pages, we will finish this "multi" survey, starting with the information obtained when images from different sensors are merged and ending with multitemporal examples.

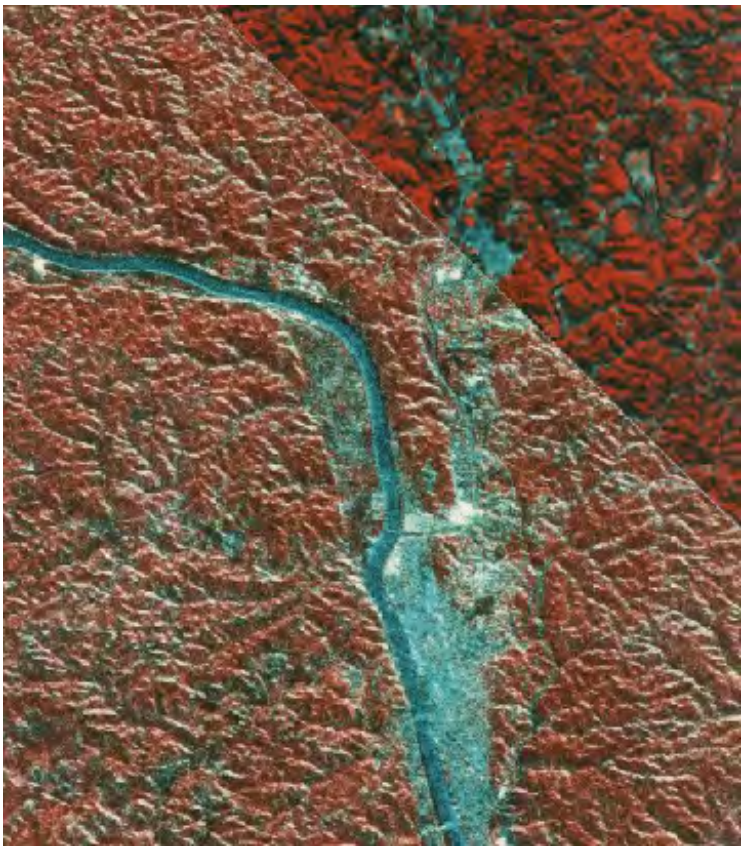
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Here we delve into the advantages of multisensors. This leads to a discussion of merging an image produced by one sensor with that of another. Both may cover the same wavelength range but differ in, say, resolution or pixel size. Or they may be quite different types of sensors, e.g., radar and visible band scanners. Merging can enhance information content but often requires considerable effort, despite usually being carried out by computer programs.

Multisensors and Image Merging

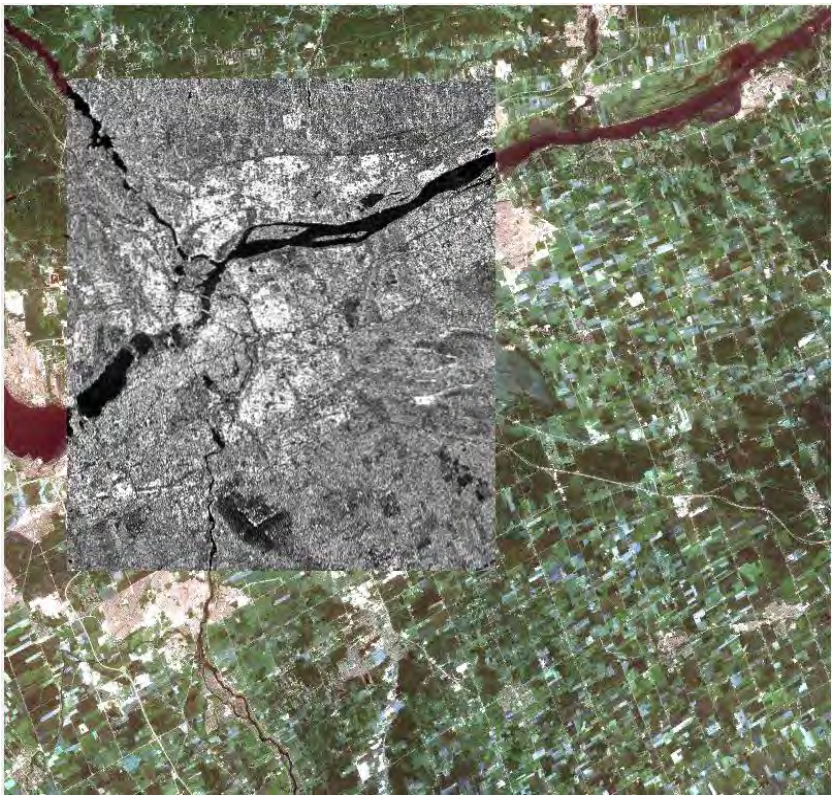
We have seen examples scattered elsewhere in the Tutorial of imagery consisting of combinations of images made by sensors operating in different modes or in different parts of the EM spectrum. One such example of a multisensor image, first shown in Section 8, is this view of the Allegheny Plateau made by combining radar and Landsat false color images:



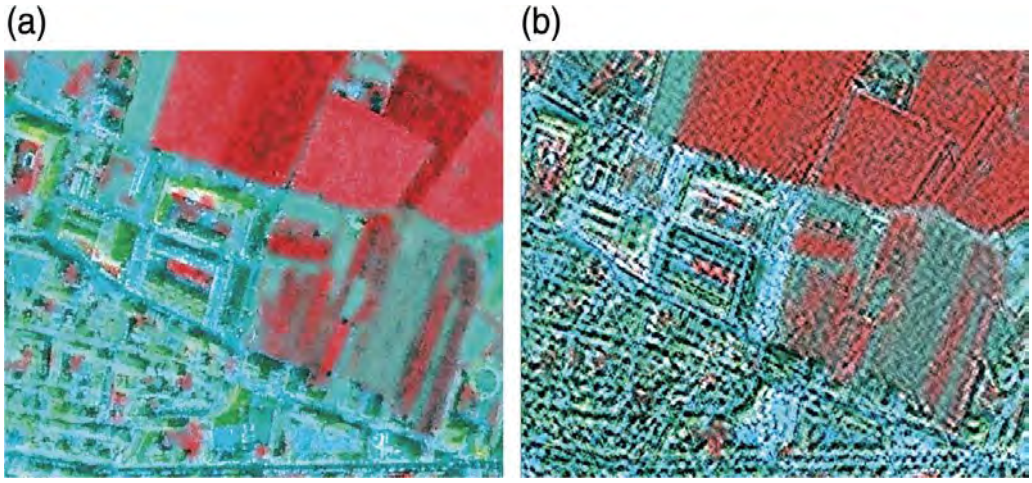
Words that naturally springs from this "multi" concept are *merge* and *combine*; *image fusion* is another term that applies to the output of overlaying one image type on another. Data acquired by different platforms, with different sensors, at different resolutions, and during different times will tend to be incompatible in some respects. The most common incompatibility is geometric: the pixel representing radiometric data in some spectral interval from a particular area on the ground or in the atmosphere will probably not be the equivalent size for the different sensors that monitor the target, be it on the Earth or a planetary surface, or the properties of the air above. In order to combine data sets from different sources, some adjustments or shifts in both geometric/geographic and radiometric values will be required. Two pixels may partially overlap; they may vary in shape. Their radiometric character may require modifications (e.g., correcting for atmospheric effects or for bidirectional reflectance). Thus, to successfully merge, both geometric

and radiometric corrections must be applied. Some form of resampling (see page 1-12) is usually necessary. Distortions must be reduced or removed. Rectification to some planimetric standard (for example a suitable map projection) has to be incorporated. And fitting or stretching one image to properly overlay another is often vital, requiring **ground control points** or tie points. These general processing steps, while vital, will remain beyond the ken of this Tutorial but the interested user should consult any of the textbooks listed in the Overview, as, for example, the latest edition of Lillesand and Kieffer's *Remote Sensing and Image Interpretation.* On this page we must be content with looking at some examples.

One benefit of merging images made by sensors that produce notably different types of images is that the new combined image can show contrasts that are informative - a form of symbiosis. This image shows that there are both similarities and differences between such types as a radar image and a multispectral (visible/IR) image:

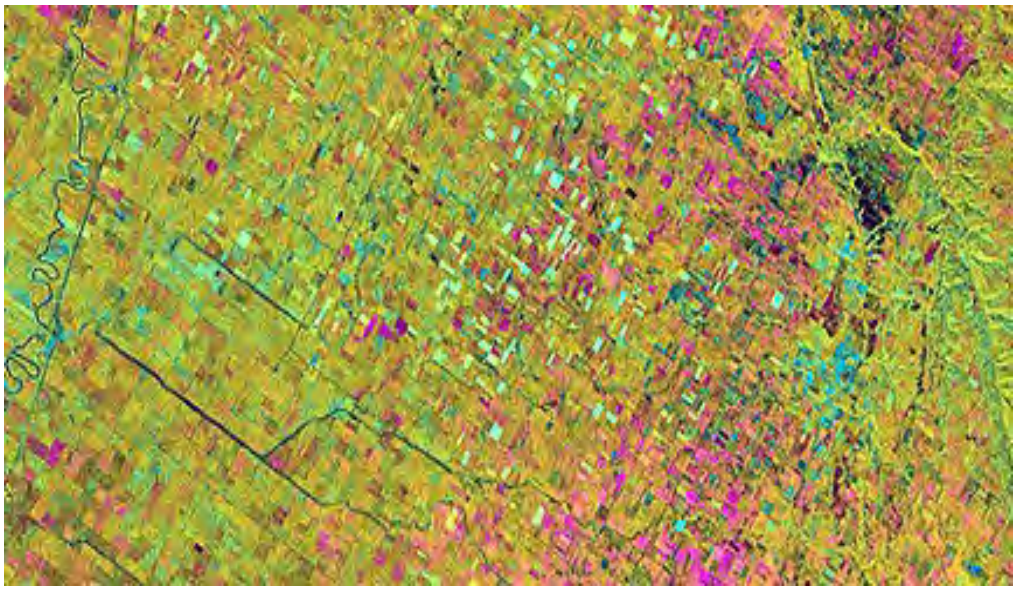


Now, lets see what combining a radar image with a multispectral image. This SPOT image was combined with a radar image. Note the effect produced.



The next image is a combination of SAR radar and Landsat-7 ETM+ data. The scene is at El Azraq in Jordan:

The following scene is a variant of the merge theme in that two radar band images from one sensor are combined with an image from another radar sensor to create this image of fields in the Red River basin of Manitoba. Two of the bands in the color composite are from the L- and C-band radar systems operated by JPL. The third is the X-band instrument developed by a German-Italian space consortium. All were mounted on the boom that extended from the Shuttle. Since the images were taken simultaneously, the time factor is eliminated. But the two instruments were not completely compatible in spatial sampling, so the pixels required appropriate algorithms to permit merging, and in some instances difference in look angles and polarization.



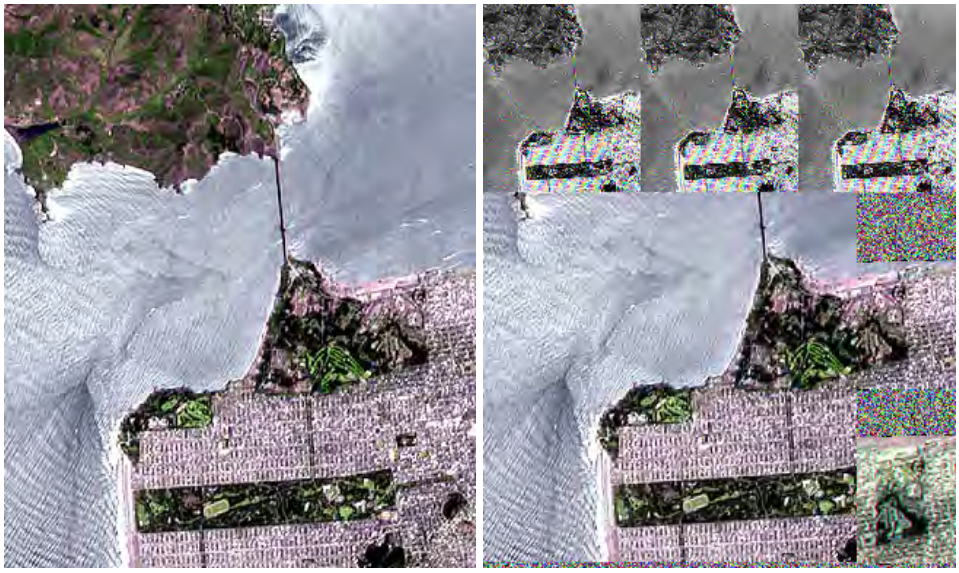
The fields were producing mostly corn, wheat, barley, sugar beets, and canola. Magenta designates bare soil; bright fields are dominantly corn; other standing biomass is in cyan.

A word of caution: Radar images of mountains or other terrain that has considerable relief will usually show foreshortening and layover effects (see Section 8). This tends to distorted a merged image that includes a radar image (see first illustration at the top of this page).

Thermal imagery also benefits from merging with other kinds of images. This next image, made by Rupert Hadyn, contains color information relating to temperature variation in the mountains of Morocco. The input data are Day-VIS, Day-IR, and Night-IR readings from HCMM, shown in a color composite that uses the IHS (Intensity-Hue-Saturation) color system. The resulting color image has been registered to a Landsat MSS Band 5 scene of the same area, so that its black and white appearance gives a sense of topographic expression superimposed by colors, such that reds are associated with the warmest night temperatures and blues are the brightest areas in the Day-Vis scene.

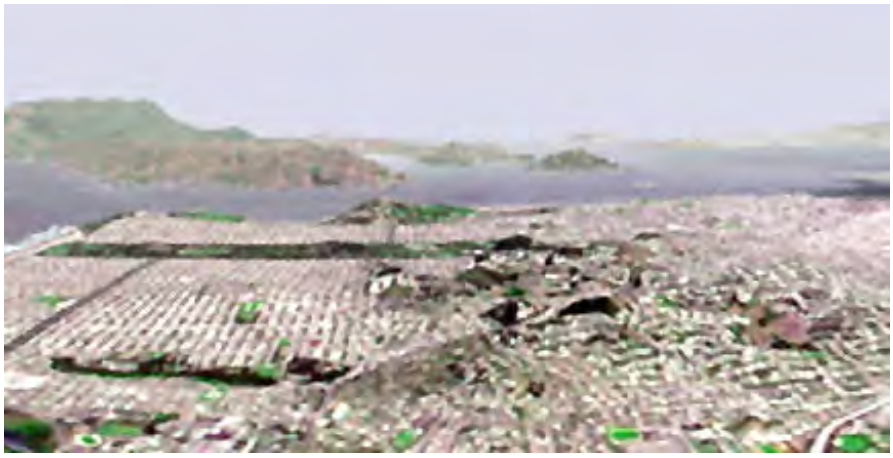
A common practice is to merge scenes that were acquired simultaneously with similar sensors that differ in their resolution; the result is a scene that seems to present more detail and hence more information. This is exemplified below in a scene made by fusing a Landsat TM color composite at 30 m resolution with and IRS (Indian remote sensing program) 5 m panchromatic image of an unnamed town, with impressive results:

SPOT provides another excellent example of a fairly simple merge of this type designed to heighten scene resolution: the 20 m HRV multispectral data with the 10 m HRV Pan(chromatic) sensor. The first two paired scenes show part of San Francisco, including the Golden Gate Park, the Presidio, the Golden Gate bridge, and Sausilito in Marin County. The image on the left is a 20 m quasi-natural color view; on the right is the black and white 10 meter resolution image of the same area. Both scenes were acquired at the same time.

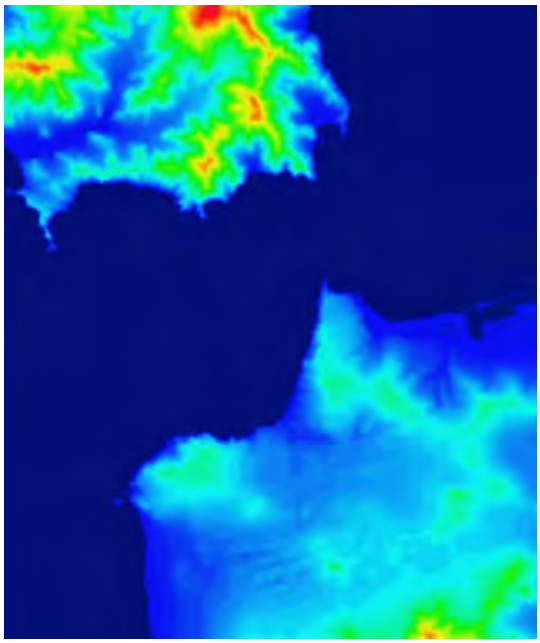


Now, in this next scene the two are merged; in the process, the higher resolution portrayal of wave patterns was blanked out, using instead the 20 m image waves:

Another manipulation involves registration of the image pixels with DEM data points, so that by using the appropriate procedural algorithm, the scene is converted to the following perspective view of San Francisco's Golden Gate:



In previous Sections, we have shown several other examples of combining space imagery with DEM elevation data to produce three-dimensional views. We can do this with the above scene. First, the DEM data are displayed in a color coded general map, with high elevations in red and lowest in medium blue (ocean dark blue). They are then input to a shaded relief map, on the right:



Such merged data can be manipulated to produce perspective views. Here is a view of the Straits of Gibraltar (Gibraltar on the left; Morocco on the right) in which the scene acquired straight down has been rotated to simulate an oblique view:

The last illustration for this topic was copied from the textbook by Lillesand and Kieffer. It shows a created color composite in which the blue and green bands are SPOT HRV bands and the red band is a Landsat TM band 5 image; the yellow is a GIS overlay. This is a sophisticated data merge indeed!

The processes described above are called "multisensor analysis." The term refers to combining data obtained by more than one type of sensor on a spacecraft or, more commonly, by sensors on different spacecraft. For example, we may image a study area at various times by Landsat, SIR-C, TIMS, SPOT, MOMS, and AVIRIS. Of course, we may independently examine each data set, and imagery derived therefrom. Or we can lay visual products side by side. From this multiset, we can interpret the scene by simply looking at the various aspects of scene content. This process is standard procedure in conventional photointerpretation.

Somewhat more sophisticated is the approach that uses each sensor data set as input to classification. Thus, we can combine visual and SWIR bands from AVIRIS with TIMS thermal data, so that 10 to 12, or more, band values contribute to the multivariate analysis that leads to a classified scene or map, likely to have improved accuracy. And, of course, we can digitize and combine other kinds of data by using aerial photography or thematic maps (described in the review of Geographic Information Systems in Section 15).

We clearly observe that, with the proliferation of sensors and their platforms (satellites) in space, the systematic combining of data, acquired over a wide range of wavelengths, scales, and temporal conditions, will result in a strongly synergistic use of the valuable data sets each operating system provides.

Merging or fusing is commonplace with Earth-oriented sensors. But it may surprise to learn that it is also used with imagery obtained by different telescopes that look at astronomical targets such as stars and galaxies (see Section 20).

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

A very common type of "multisensing" is "multitemporal", meaning that periodic repeat looks at the same ground area can be used to extract information otherwise more difficult to procure. An obvious example is in crop studies, where imagery from different times in the growing cycle can be used to better identify the crops present and estimate their biomass state or anticipated yield. Special consideration is given to one excellent example of monitoring the changes over several years: the damaging effect of the Gulf War on Kuwait's Burgan Oil Field and its subsequent recovery.

Multitemporal Coverage

Satellites are ideal for monitoring changes in accessible parts of the Earth over time. The repeat cycles of those that orbit at high angles to the equator are measured in days or a couple of weeks or so. This facilitates monitoring of crop growth and regional vegetation progression as well as drought and stress conditions. Clearcutting and environmental damage are effectively monitored. The ravages of flooding can also be accessed, particularly now that there are a number of different satellites in operation, so that the likelihood of any area being covered on a given day has increased. Of course, daily weather changes are the mission of most of the meteorological satellites. In principle, the state of a surface that is not cloud-covered is observable, albeit at usually low resolutions, by the group of geostationary satellites now in place.

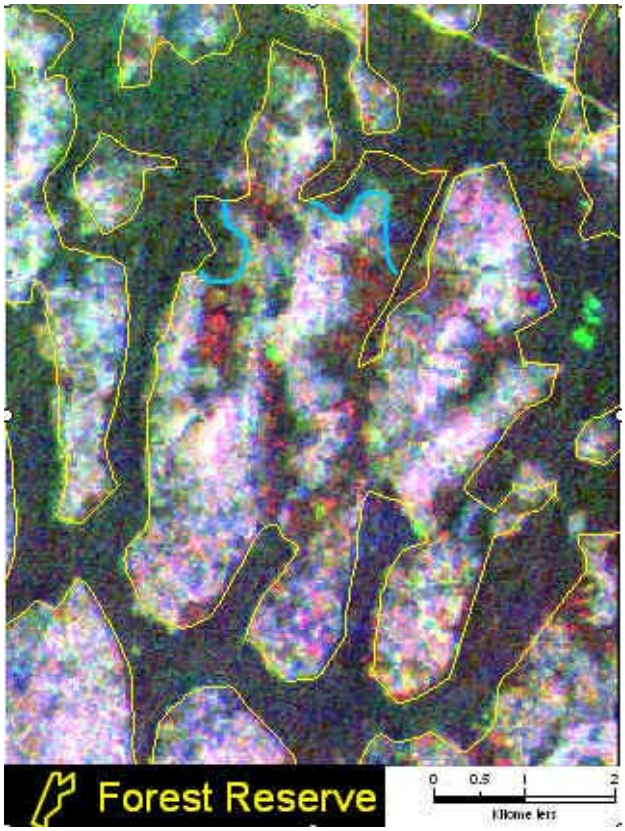
Long period changes (in years) are also suited to the steady presence of the stable of satellites in orbit. The mapping of growth of cities and population points and other land use categories is a straightforward use. The drying up of lakes (such as the Aral Sea) and the changes in areal extent of ice sheets and glaciers can be followed. Degradation and loss of wetlands generally are slow processes that still can be detected over time, perhaps early enough to reverse the trend.

We have seen several examples of **change detection** before, especially in Sections 3 and 4. Several more should suffice to document the utility of satellites for this purpose. First is a simple comparison of rain forest clearcutting in Brazil, using data from two different satellites:

Landsat TM imagery (30 m resolution)
Year 1984
Image size: circa 150 km x 150 km

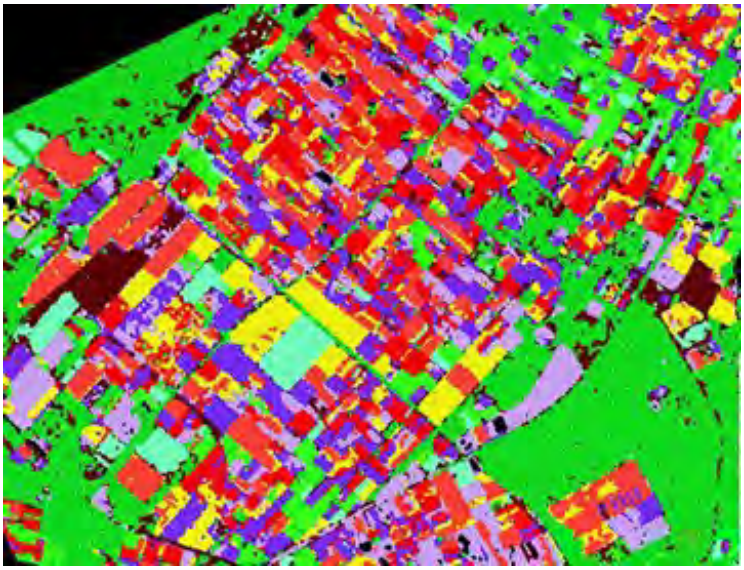
ENVISAT MERIS image (300 m resolution)
Year 2005
Image size: circa 150 km x 150 km

Here is a case of reduction in forest land in Uganda. The image format is rather novel. The colors refer to satellite data from three different years: blue = 1974; green = 1986; red = 1995

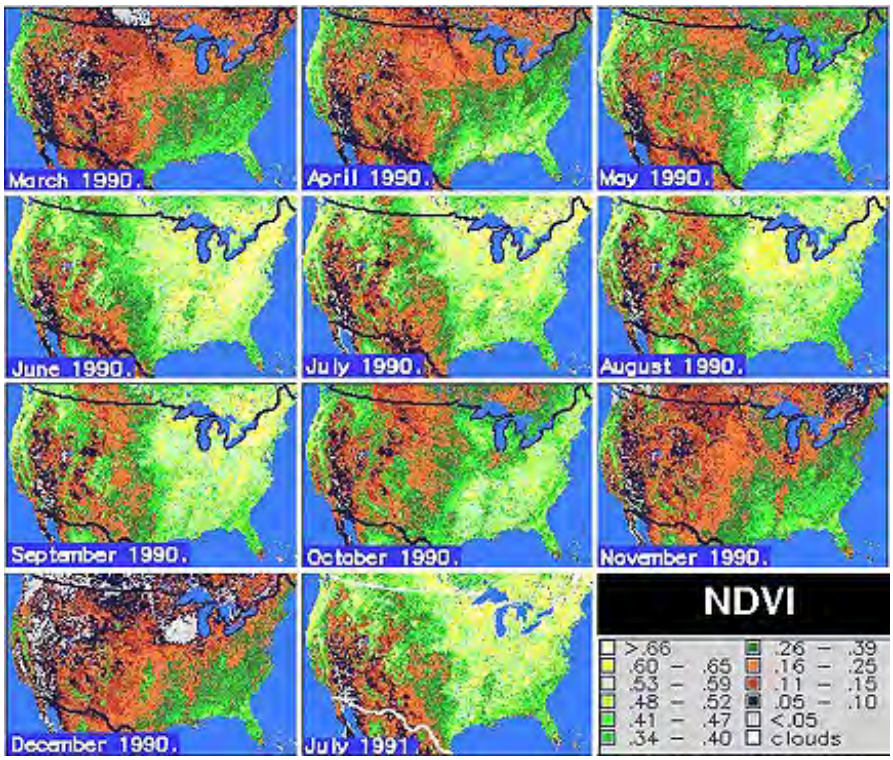


Staying with the vegetation theme, a Landsat TM classification of crops in coastal farmlands in the Netherlands was presented on page 3-2, with the comment that the single date coverage led to an accuracy of only 56%. However, when 20 dates of coverage during the growing season of 1993 were obtained with ERS-1 SAR (radar) coverage, and a supervised classification was carried out, this next image achieved an accuracy of 70% (see page 3-2 for a legend and listing of the 10 different classes).

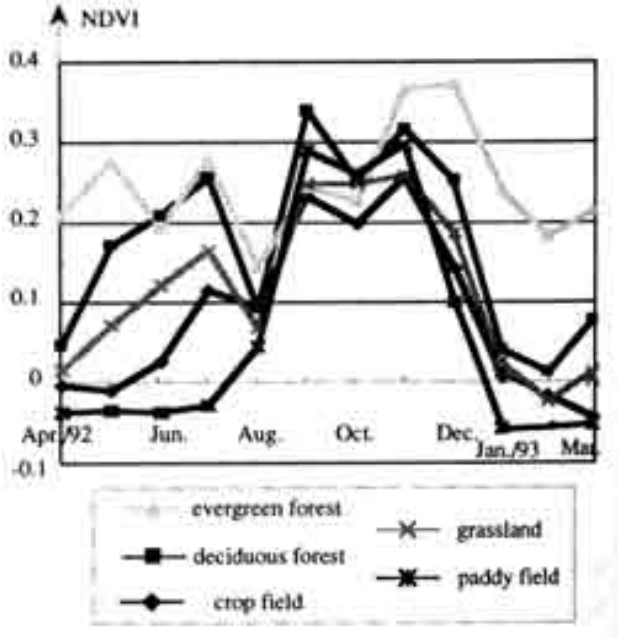
Then, when this above scene was merged with the Landsat TM classification, the image below resulted; it improved accuracy to an average of 79% for all classes.



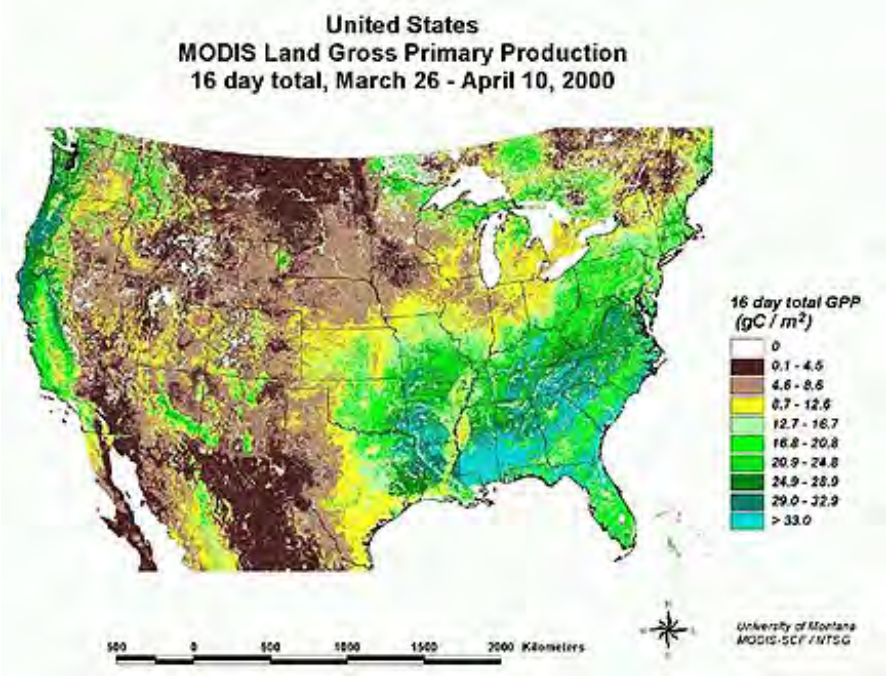
As stated before in this Section, AVHRR visible and IR data at 4 and 1 km resolutions are readily processed into NDVI imagery for regions of varying sizes up to continental scale. Here are a series of images covering NDVI variations across the 48 U.S. states on a monthly basis from March through December 1990, along with a July 1991 image that can be compared with the July 1990 image. The "greening" of the country and regression during the Fall are easily traced.



NDVI varies seasonally for different vegetation types. These plots show this effect:

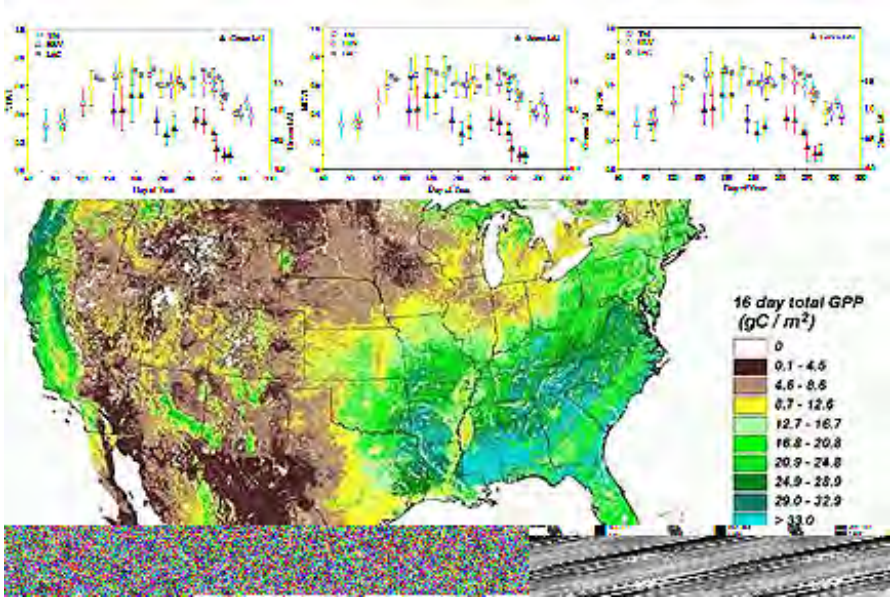


Continent-wide vegetation changes for shorter periods can be monitored by AVHRR and by other sensors, such as MODIS on Terra (page 16-9). Here the data being mapped are conversions of vegetation cover into gross biomass productivity.

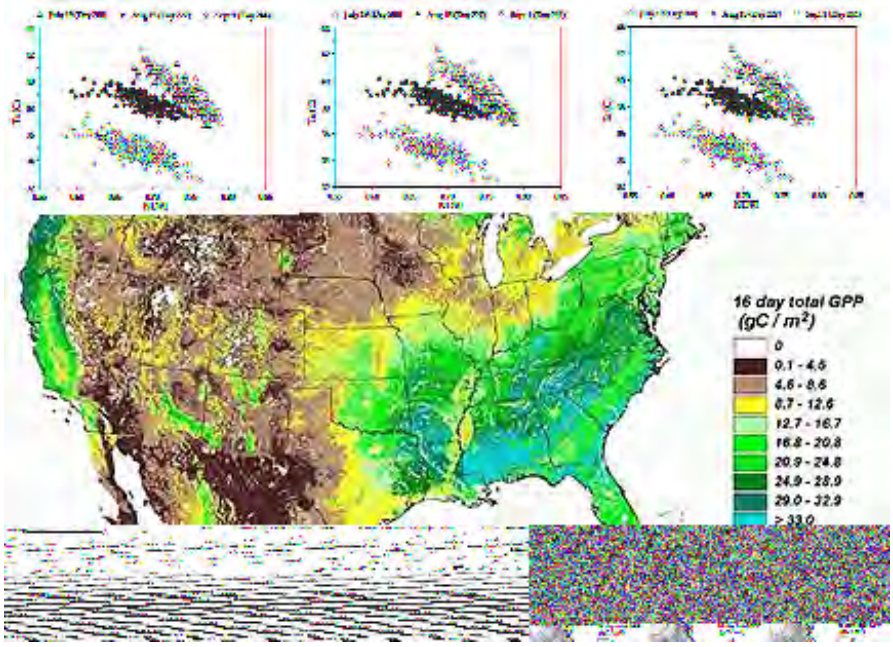


Since the early Landsat days, there have been a number of U.S. and international programs to test the value of remote sensing, coupled with GIS, and other sources of relevant information to various vegetation studies, principally of crops, grasslands, and forests. Some of these are part of the ISLSCP (International Satellite Land Surface Climatology Project) program. Currently active is BOREAS (Boreal Ecosystem Atmospheric Study). Before that was the FIFE program which followed productivity of agricultural grasslands in the Konza Prairie Test Site near Manhattan, KS. Here is a composite Landsat TM image of the FIFE site:

Many diverse studies were conducted on this site. One of interest, from which these next two graphs were extracted, was carried out by Scott Goetz. The first plot shows the variations in NDVI and LAI (Leaf Area Index) over most of the growing season, as followed by a large number of AVHRR observations and several by TM and the HRV on SPOT.



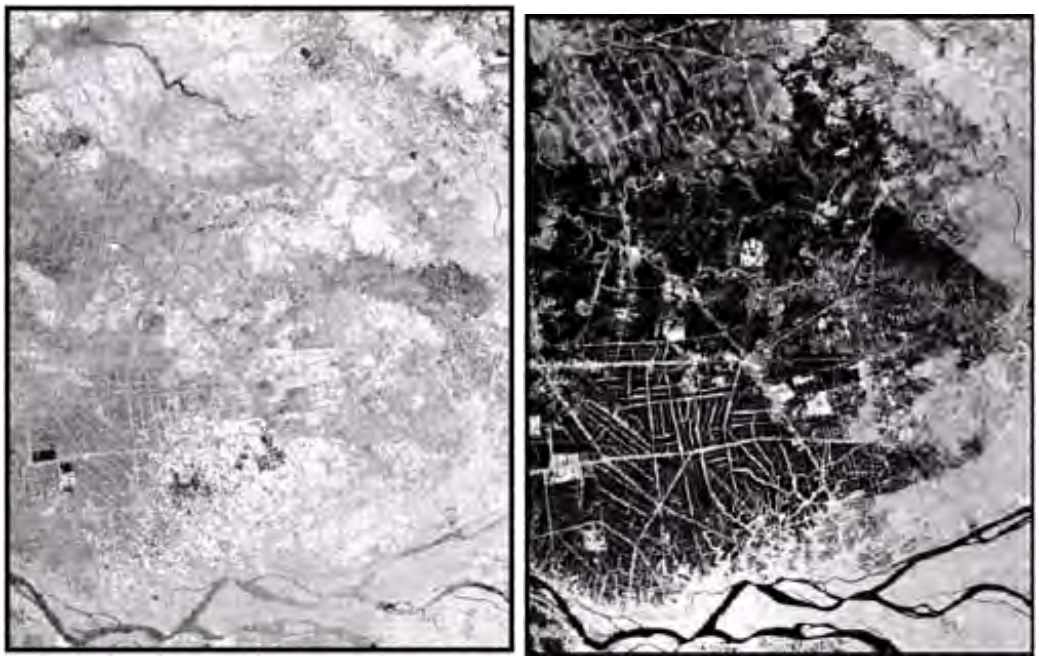
The second plot indicates the variations in temperature as determined by the AVHRR thermal channel as a function of NDVI (from AVHRR) for three dates in summer. As expected July produced the highest temperatures and the coolest were by September 1. As NDVI increased, the temperatures dropped owing to the cooling effect by evapotranspiration associated with the grasslands.



The Geographic Information Systems (GIS) approach (Section 15) is now being widely used in crop management. A search on the Internet found many references to GIS crop studies but almost none proved suitable (with good imagery and data sets) to be extracted into this Tutorial. We suggest you do your own search (using such key words as "crop identification" "crop management", "GIS", "remote sensing"). However, a good idea of GIS's potential is reviewed in this Web site covering the Bighorn Basin, Wyoming.

Remote sensing and GIS have made huge impacts on how those in the agricultural industry are monitoring and managing croplands, and predicting biomass or yields. In fact, a new term has sprung up - *Precision Agriculture* - which carries with it the essentual components of remote sensing and GIS. Keywords associated with the tasks or goals involved are: farm management; crop needs; nutrients; pest/disease control; environmental contamination; crop species and variety selection; timing of field practices; conservation; soil organic matter; irrigation; minimizing erosion. Here is a good summary Internet site that treats Precision Farming.

>In Section 14 you will see examples of flood monitoring using satellite data. We preview this common use by looking at the use of radar images of flooding in the Mekong Delta of Vietnam:



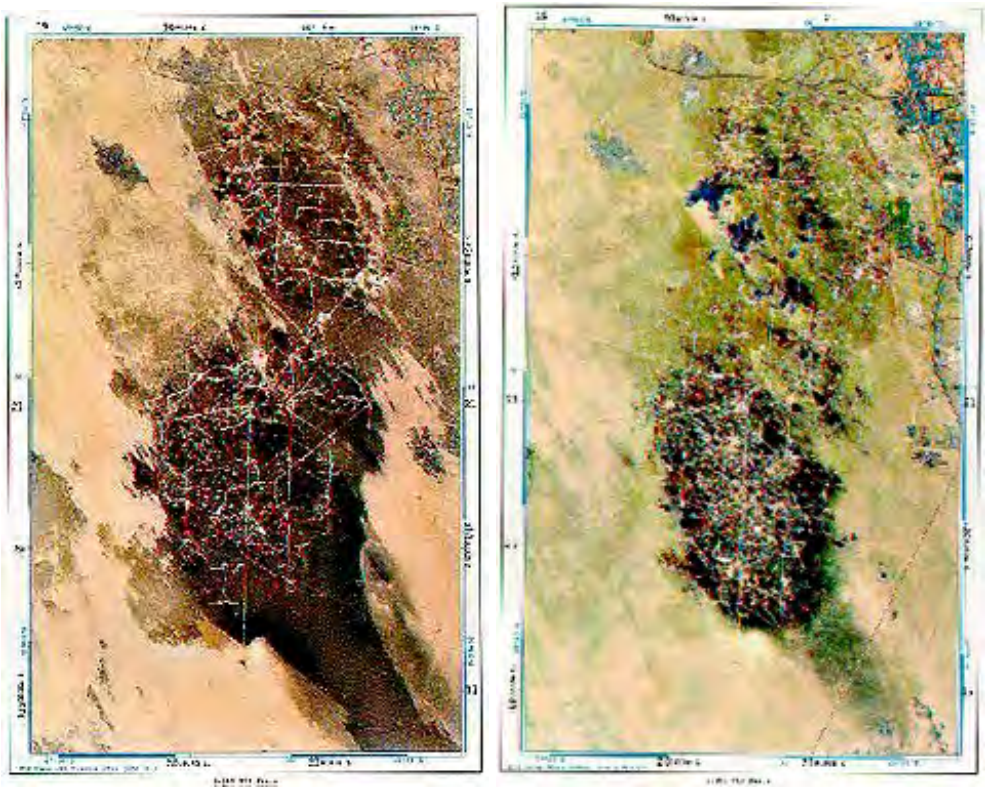
When images of the same area are taken on different dates (one at the height of flooding) and by different sensors, this color composite image (blue = ERS SAR 8/3/97; green = ERS SAR 9.7/97; red = 10/8/97) results:

The next sequence of imagery exemplifies the benefits of multitemporal remote sensing in a non-agricultural setting by following an intriguing case study of the damage and recovery of oil fields in Kuwait resulting from the 1991 Gulf War. We begin with part of a Landsat TM scene of Kuwait before the war; roads and gas-burning plumes are evident.

During the Gulf War, before they pulled out, Iraqi troops set many oil wells on fire. The most extensive ignitions occurred at the large Burgan oil field in southern Kuwait.

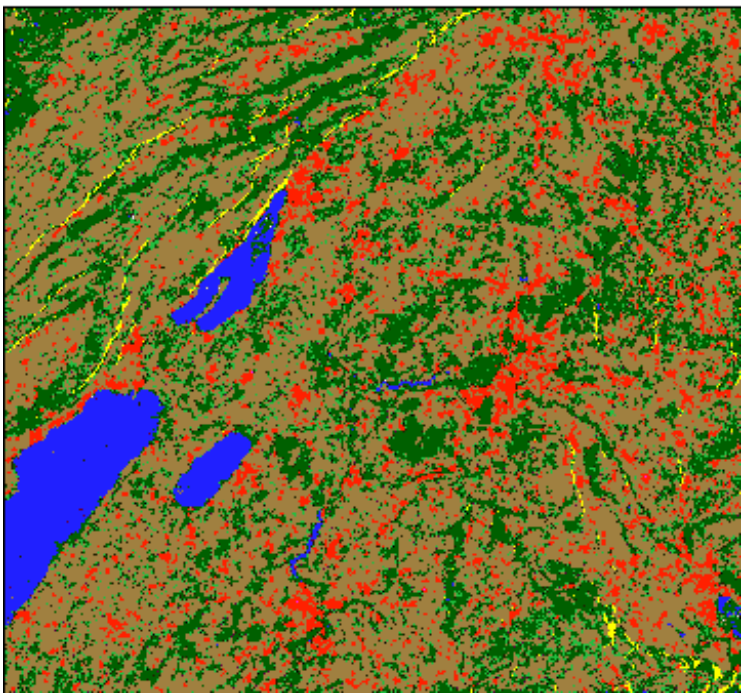
The next image is a TM subscene taken soon after war's end, with the black area being the soot and burned-out facilities at the Burgan field.

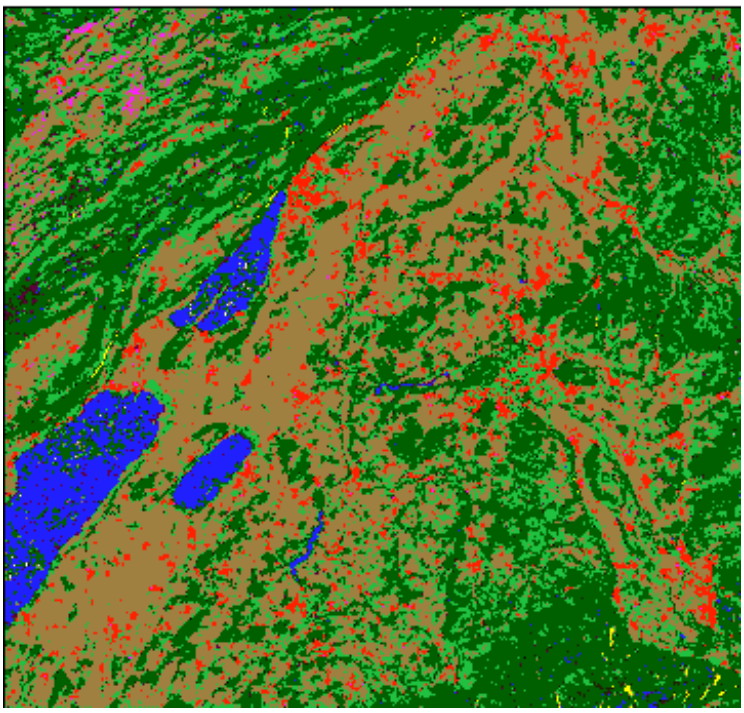
This pair of Landsat TM scenes show close-ups of the burned area on November 14, 1991 and March 3, 1995. In the 1995 image, you can see that some of the damaged area is being restored by natural and planted grass land cover.



This next image is a further enlargement of the Burgan field area in which the color patterns denote discernible changes between 1987 and 1993, brought out by a change detection merge of images from those two dates.

There can be pitfalls in using multitemporal imagery, owing to certain variables that affect the scene expression. Two obvious ones are Sun angle changes and vegetation changes. You may remember the classification of a scene near Bern, Switzerland, using radar data, presented in Section 8. We show here two other classifications, one made from July, 1995 SAR data, the other from SAR data acquired in November, 1995. The differences between the two are caused by these variables and other factors. So, remember this: A classification is sensitive to time of data acquisition.





We will now wind up these last three pages with a case study of one area and subject which illustrates how the various "multi's" work together to symbiotically enhance knowledge and understanding of a scene and a theme as examined primarily from space.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The ideas and information covered in the preceding three pages can be applied and integrated by examining just one theme: in the case study on this page, eruptions of Mt. Etna on northeast Sicily in the last 10 years. Observations have been made from a variety of platforms, some with similar sensors and others with specialized sensors. Most of what can be seen from time to time are smoke plumes from Etna, lava flows (distinctive in thermal bands), and other signs of activity. From one sensor, Terra's MODIS, a time sequence is presented that tracks smoke plume variations in amount exuded and direction of movement as determined by prevailing winds

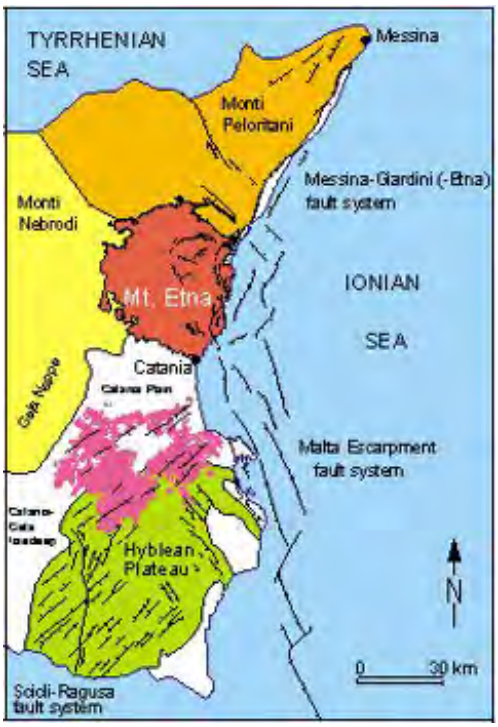
NOTE: All of the images used on this page were extracted off the Internet. In some instances, dates of acquisition were not given and in a few cases the satellite involved was not identified.

A "Multi" Case Study

On this page we will try to summarize the several meanings associated with the "multi" concept covered in the preceding three pages. We will choose a single part of the Earth where a sequence of recent natural events associated with one locality have been attracting worldwide attention. Using primarily numerous ground and space images, this event will be examined using multiplatforms, multisensors, and multitemporal coverage.

The event is the repeated series of eruptions of one of the world's most famous volcanoes - Mt. Etna near the northeast tip of Sicily where that island lies west of the Straits of Messina at the toe of the "boot" of Italy. This volcano, of the Strombolian type (the island hosting the Stromboli volcano lies just to the north), is the most active in Europe and, for that matter, is the oldest timewise in Man's historical (recorded) knowledge of active, erupting volcanoes. The earliest documented mention goes back to about 2000 BCE; Etna has erupted 192 time since. Two series of eruptions, in 2002 and in 2002, will be studied here.

First, let us establish a general geologic setting, shown in this map; the exposed volcano itself covers about 1200 square kilometers:



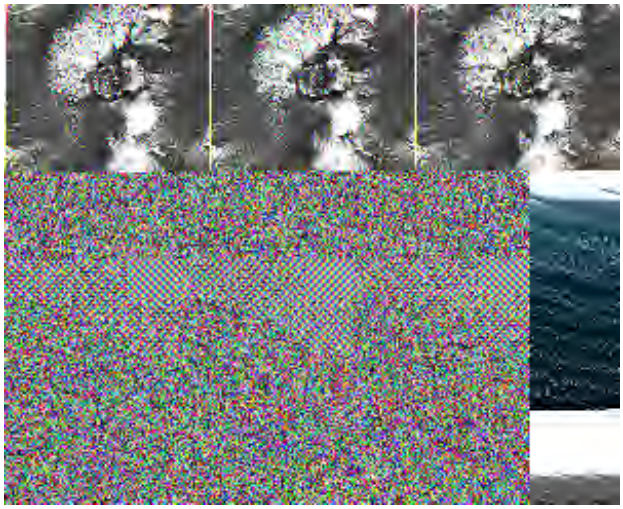
Being a part of the great alpine system of folds and faults that resulted in the northward movement of the African tectonic plate against the Mediterranean and the European plate to its north that resulted in the Alps and Apennines, the rocks exposed in Sicily are strongly deformed into folds, nappes, and faults. The dominant rock type here, and throughout much of the alpine system, is limestone. Mt. Etna seems to overlie a "hot spot" where plumes of molten rock approach or reach the surface (Vesuvius and Vulcano in southern Italy are other examples). Etna's volcanic rock is intermediate silicic (trachyandesite; some basalt) and many lava flows have aa surfaces; scoria and bombs are common. Both pyroclastic (ash) and flow eruptions are the usual mode. The lavas come from at least two shallow magma chambers, each of which seems to be fed by a deeper large magma reservoir at a depth of greater than 50 kilometers.

Mount Etna has a conelike structure but with gentler slopes than characteristic of stratovolcanoes; it resembles somewhat the shield volcano that makes up Mauna Loa on the Island of Hawaii. The summit, which has a caldera, reaches to 3350 meters (10991 ft). Of that, only the upper 2000 meters are volcanic; this edifice of interleaved pyroclastics and flows rests on a sedimentary rock base. Here is a view of this snow-capped peak as seen from near Messina, looking west-southwest.

The chances are better than even that at any visit, smoke will be issuing from the vent.

Most of the upper half of Mt. Etna is relatively barren of vegetation; its surfaces show evidence of many previous flows and channeling.

The summit itself is shown first in an aerial photo and then in a ground photo:



The next six photos were taken from the ground at various times during eruptions that took place intermittently between 1999 and the end of 2001. The first shows the ash-steam cloud at sunset.

Here is a nighttime view of crater eruption; the yellow spots are lights in small villages and individual homes.

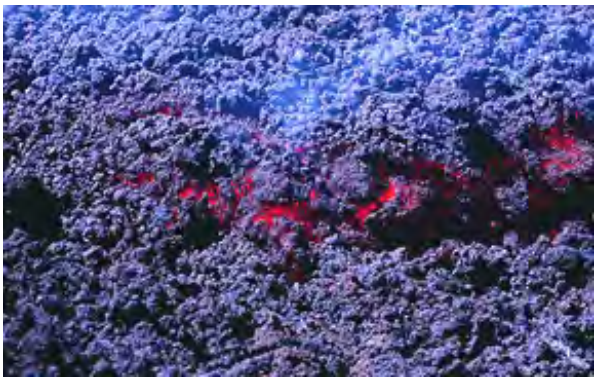
Lava flows down the sides of Etna were common.

Again, night brings on spectacular scenes such as this one showing large blobs of lava being tossed from Etna that make a fireworks-like arcuate pattern

Fire fountains are jets of lava pushed upward by gases during an eruption. The first below comes from a curtain of lava ejected along a fissure; the second is a fountain coming from a spatter cone.

This, and the preceding image courtesy Juerg Alean - Volcanoes of the World

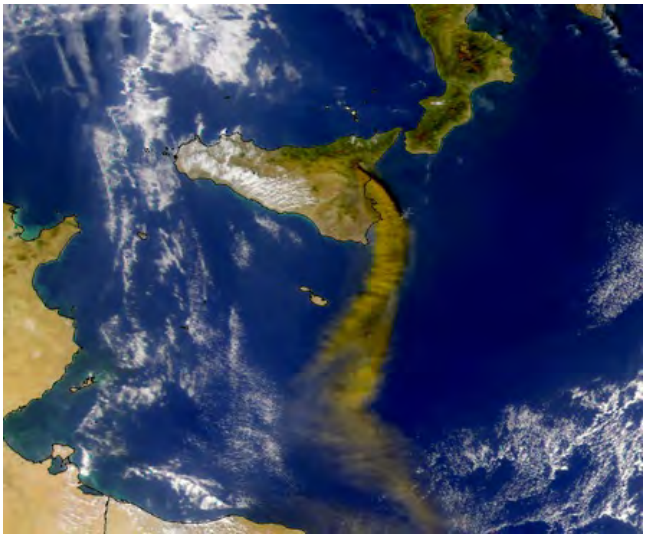
After an eruption involving lava flow, the surface is typically one of "aa" (clinker-like individual fragments), as seen here:



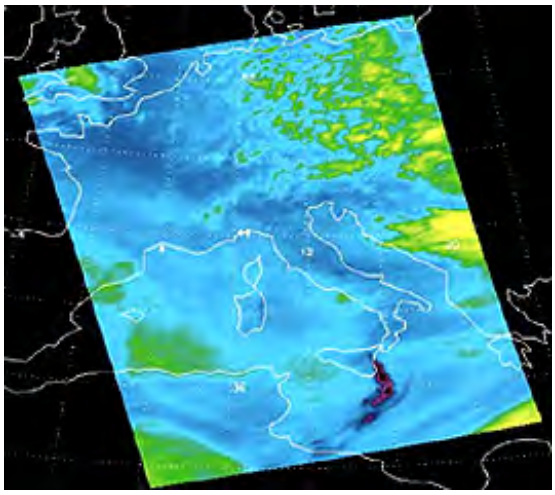
Now we are ready to study Mt. Etna from space. Most of the scenes below result from July-August, 2001 and October-November, 2002 eruptions; a few scenes were acquired in 1999-2000.

For a regional perspective, look first at this October, 2002 SeaWiFS image

This view of actively-erupting Mt. Etna was obtained by the Orbview-2 satellite



The Atmospheric InfraRed Sounder on EOS's Aqua is capable of making regional imagery; look for the black smoke plume:



The ATRS (Along Track Scanning Radiometer) can provide different aspects of a ground scene. Here the island of Sicily is dark but the plume from Etna has been colored yellow.

Or, the scene can be presented by ATRS in a quasi-true color mode

The Medium Resolution Imaging Spectrometer (MERIS) on Envisat can produce images with different colorations. The two scenes below, focusing on Mt. Etna, show differing renditions:

SeaStar, a European satellite that concentrates on oceanographic conditions, also produces adequate displays of land surfaces, here showing all of Sicily with an active Mt. Etna.

Satellites used in meteorological studies can be processed to highlight small areas. Here is a DMSP (Defense Meteorological Satellite Program) image of the area around Mt. Etna.

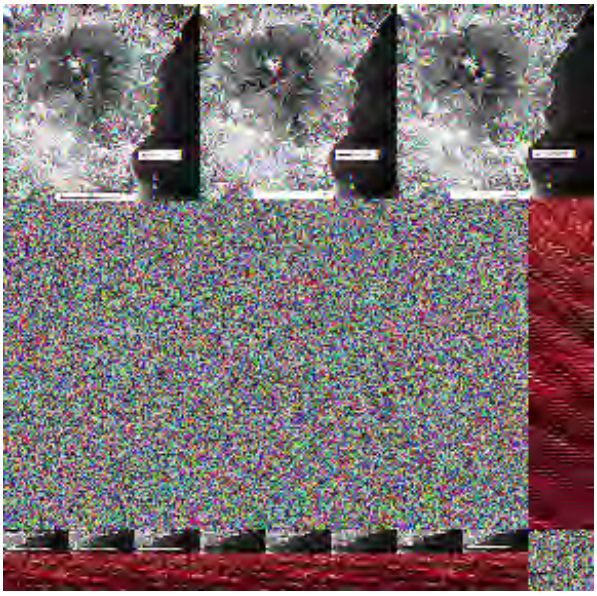
An Etna eruption was seen and photographed by an astronaut on the International Space Station as it passed just south of Sicily. The result is much like an aerial oblique photo but covers a much larger viewing area.

Here is another ISS astronaut image taken with a different camera on July 22, 2001.

On another occasion, an astronaut took this downwind view concentrating on the smoke plume itself.

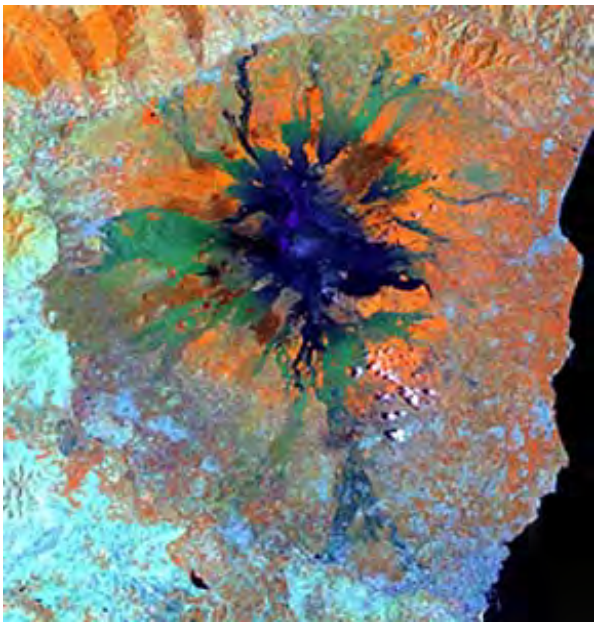
Landsats have passed over Mt. Etna many times since 1973. This is a false color composite subscene obtained by the TM on Landsat-5.

Individual TM bands can be instructive. Here is a Band 5 subscene image of Mt. Etna as imaged by Landsat-7. The light tones highlight areas where vegetation is prominent.

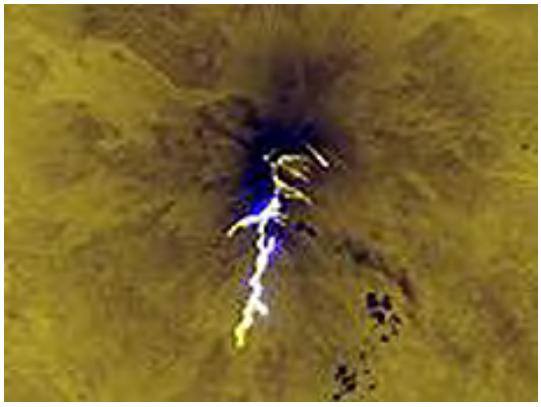


Some of the Etna images on this page attempt to render the scene in natural color. Here is a Landsat TM version:

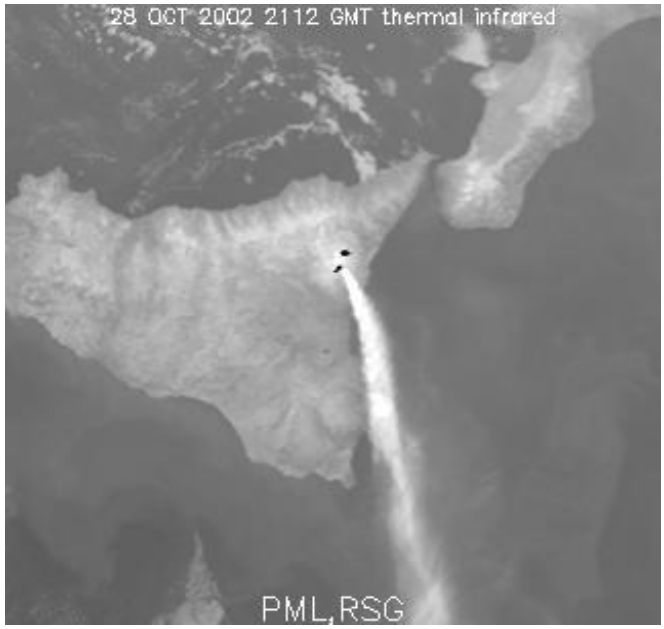
Using longer wavelength bands to include in the color composite yields colorful images, such as this one made with Bands 4, 5, and 3 as red, green, and blue respectively:

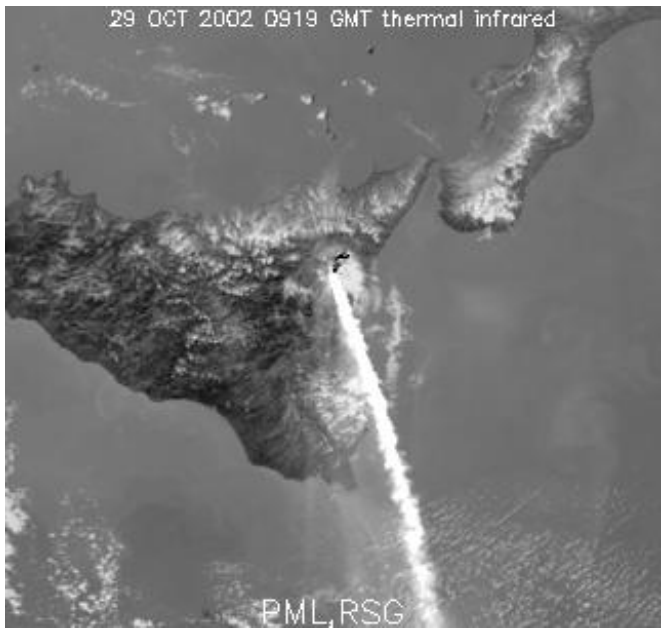


This interesting subscene was taken using the thermal band on Landsat-7 during a night pass on August 5, 2001. The hot lava flows stand out.

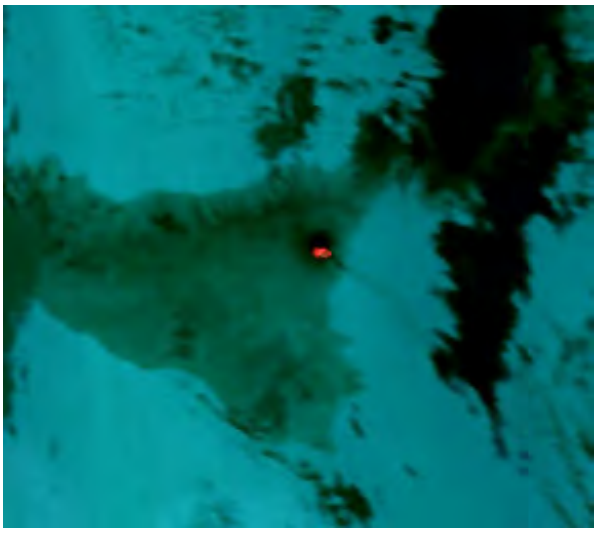


Meteorological satellites are effective at placing Mt. Etna into context. Those with a thermal channel can also pinpoint the active hot spots on the summit and flanks. These two images were gathered during the nights of October 28 and 29, 2002 respectively.



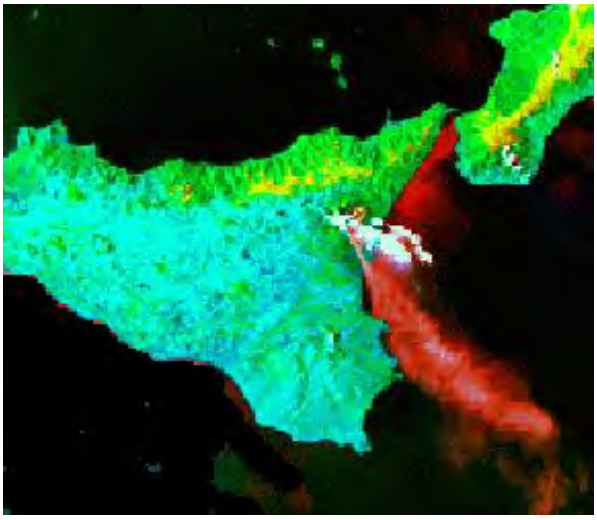


This Metsat (this term is used for any meteorological satellite whose specific identity is not given) image has been processed to show Sicily in the night as dark except for the bright orange of surface lava, with its strong thermal signal.

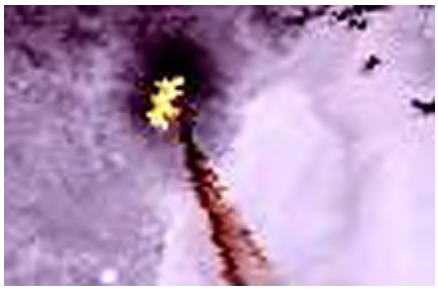


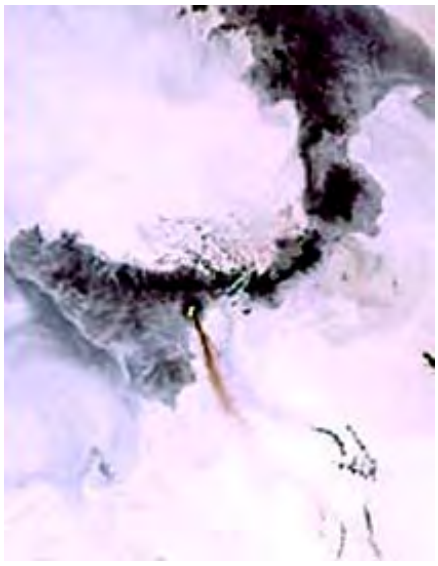
A daytime Metsat image has been reprocessed to give a color version in which the land is green and the Mt. Etna smoke plume is red.

.

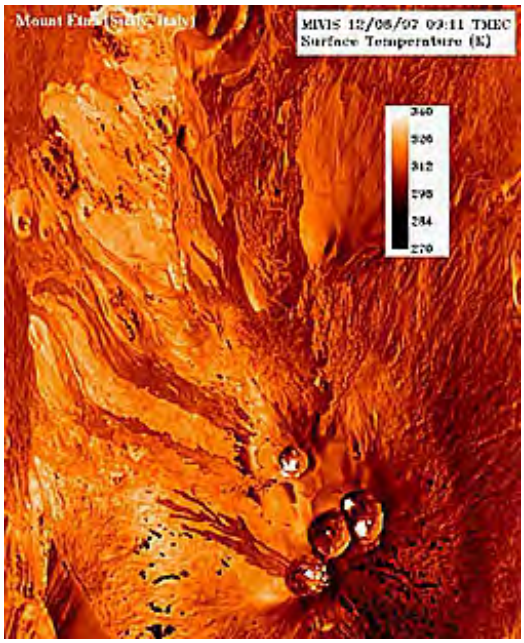


A thermal channel on the ERS-2 ATSR (described above) produced the next two images, both taken at night.



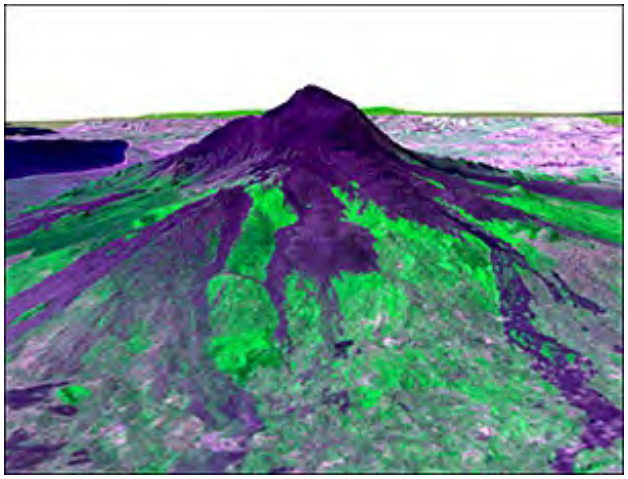


An experimental aircraft-mounted hyperspectral sensor (see following pages in this Section) called MIVIS produced this image of the surface temperatures near the summit of Mt. Etna.

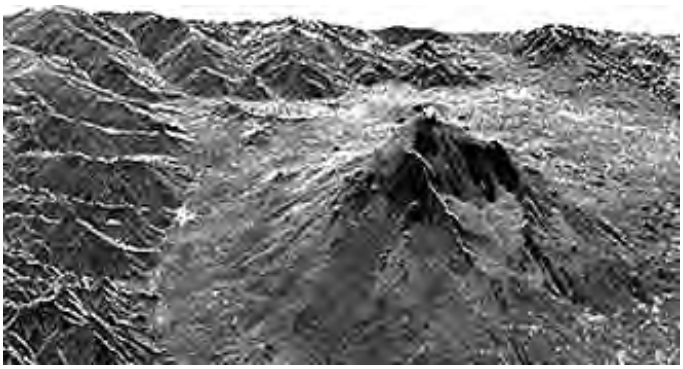


Radar is also quite useful for studying land surface conditions, such as topography (and changes thereof), around Mt. Etna. This next image is a SIR-C C-Band and X-Band color composite of Mt Etna and surrounding terrain. North is to the right.

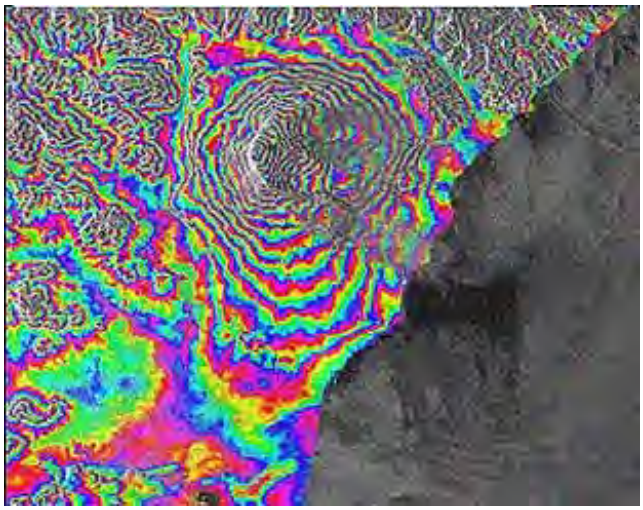
As was discussed in Sections 8 and 11, perspective views of surface topography, in this case Mt. Etna as imaged by SIR-C, can be produced using either elevation data from maps or from onboard altimeters.



The SAR radar on ERS-2 provided both scene and altimetric data needed to generate this perspective view:

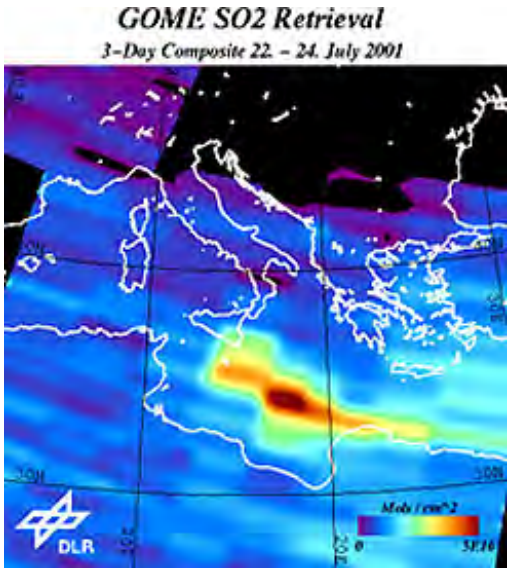


As was explained on pages 11-8 through 11-10, radar interferometry is capable of indicating changes in surface heights or elevations. Here is the interference ring pattern that was determined from ERS-1 data acquired in 1995. This suggests some ground swelling.



One of the hallmarks of the smoke plumes from most Mt. Etna eruptions is a relatively high SO_2 content, mixed with steam and ash. This is an ERS-2 image of an eruption on July 21, 2001 at Mt. Etna. Bands 7, 5, 2 (RGB) were combined to make the image.

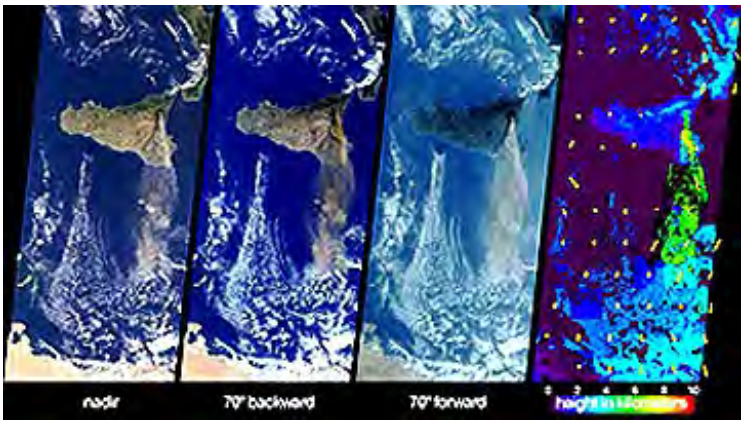
On board that spacecraft is the GOME (Global Ozone Monitoring Experiment) sensor which can provide quantitative measurements of SO_2 content. The map below shows the distribution of SO_2 summed over the period between July 22 and 24, 2001.



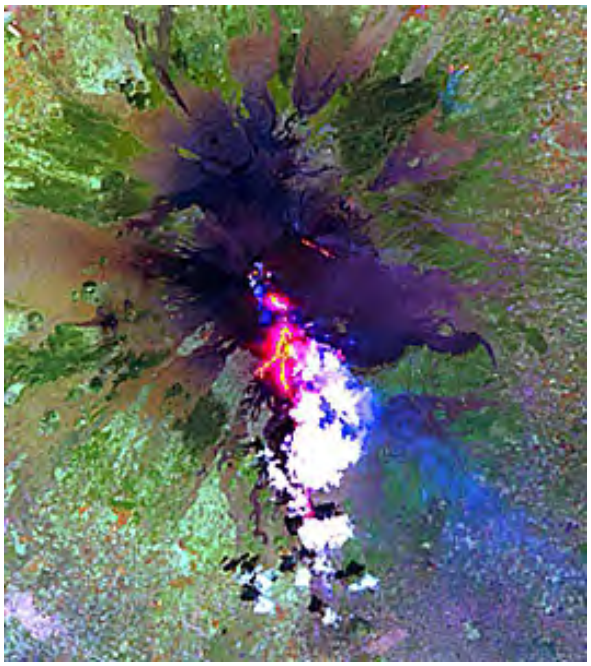
Another ESA satellite, PROBA (Project for Onboard Autonomy) monitored Mt. Etna during the first stages of eruption in 2002. The next two images show both the CHRIS and the HRS sensor products obtained then.

We come now to images generated from several sensors on the EOS program's Terra and Aqua satellites. These satellites are described in detail in Section 16. They are among the most versatile earth-observing satellites now in orbit.

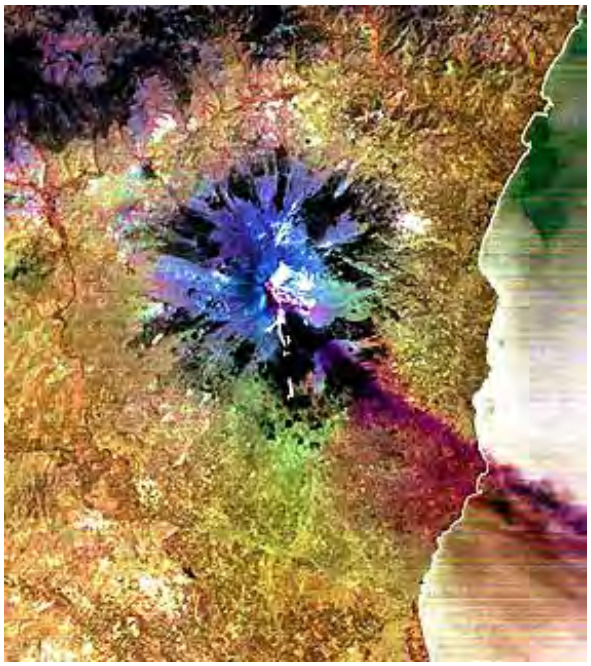
The first image from Terra was made by the MISR sensor which is capable of looking at the scene simultaneously at several fixed angles.



An earlier monitoring of this eruption, in early August of 2001, produced this striking view of the hot lava fields, as imaged by Terra's ASTER.



In this next ASTER image, one of the thermal bands is included in making the false color composite.



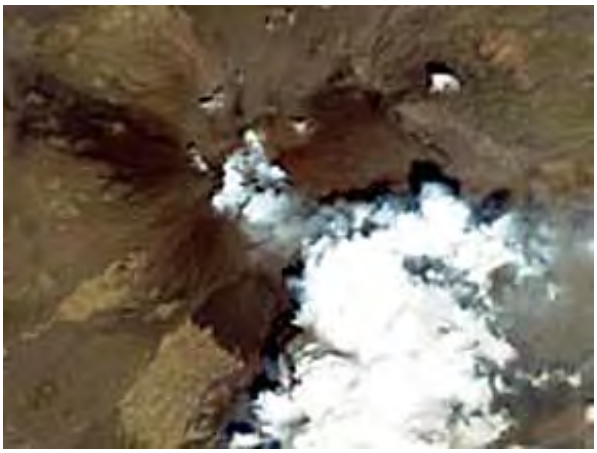
ASTER can pinpoint the hot lava extruding from its sources high up on Etna. This set of views

shows a color composite for context and two thermal band views showing these hot spots:

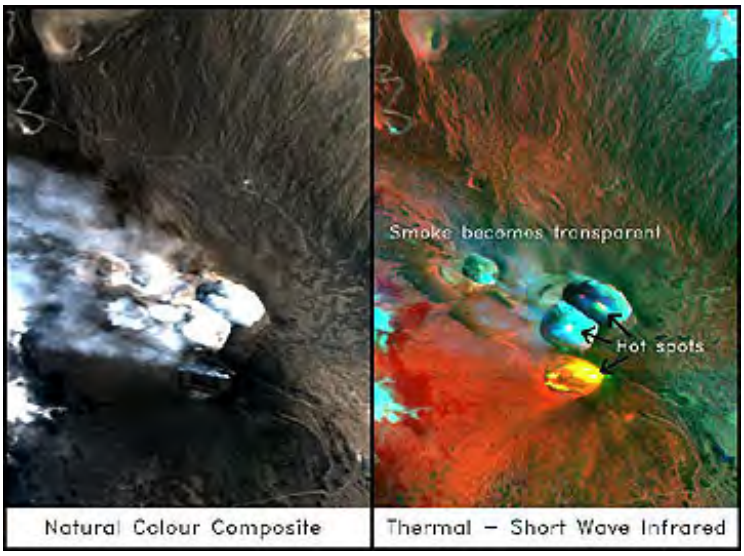
We will now use the MODIS sensor on Terra to do a time sequence study of the changes in smoke plume directions over a short span between November 3 and October 28, 2002. See the captions for details:

In late October of 2002 Etma was imaged at high resolution during a pass by Digital Globe's Quickbird-2. The view shown here is somewhat degraded so as to fit the screen. The image downloaded from the DigitalGlobe site, as shown with many megabytes, is remarkable for its detail - comparable to high quality aerial photography.

Space Imaging's IKONOS has obtained a number of Mt. Etna images, but only one was found through the Internet. Here it is, at a resolution of 4 meters.



Later on in this Section we will concentrate on the relatively new technology of Hyperspectral sensors. For now, we show a scene (shown again and explained on page 13-10 made of Mt. Etna by a European hyperspectral system called DAIS (see page 13-9).



A thorough review of Mt Etna's behavior is posted in the April 2003 *Scientific American* article entitled "Mount Etna's Ferocious Future".

Etna then is one of the more active, and often spectacular, erupting volcanoes in the world. Its repeatability and its photogenic appearance from space, air, and ground have guaranteed frequent coverage, as the above images confirm. However, most recorded eruptions cannot be marked as violent (like Krakatoa or Thera), yet at least one having that savage nature did occur in the past and some volcanologists predict another catastrophic eruption in the future.

We alluded to the other famous volcano in Italy - Vesuvius, perilously close to Naples. Its role in destroying ancient Roman towns in the infamous eruption of 79 A.D. was covered on page 4-5. Here we re-examine Vesuvius just to compare it with Etna. This still active stratocone lies just inland between Naples and Pompei. The next two images establish the geographic context of this volcano:

Much more detail is evident in these ASTER true and false color images. They show both a partial outer rim and a small volcanic cone with the depression at the top of Vesuvius.

In 79 A.D., everyone in towns like Herculaneum and Pompeii were buried alive by these "nues ardentes" (hot ash that hugs the surface, moving downslope much like an avalanche). Both towns have been exposed as they were by archeological digs that removed the ash cover. But today much of Herculaneum has been rebuilt over unexcavated areas.

We take leave now of this part of Section 13 that has dealt with ground truth and the "multi" concept to advance to an introduction to hyperspectral remote sensing after some principles of spectroscopy are espoused.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Measurements with radiometers and spectrometers are carried out primarily to accumulate specific spectral signatures of various materials and features that help to build up a spectral data "bank". This can be done under controlled laboratory conditions, leading to "representative" signatures for the particular material. Signatures gathered in the field, from portable instruments on tripods or mounted on "cherry pickers", will include more natural conditions, such as atmospheric effects and solar illumination. Prototype sensors as well as operationally-proven instruments are also flown on aircraft; the data are used either as direct inputs to a remote sensing application or as support for space-mounted sensor studies. Other kinds of useful information about a scene being studied can be monitored by sensors housed on the ground; the data are relayed intermittently or continuously from Data Collection Platforms. The nature of the Global Positioning System (GPS), based on satellite triangulation, is once again discussed.

Field Instruments and Measurements; Data Collection Platforms; GPS

Another major phase of ground truthing involves getting detailed and more complete spectral signatures using spectrometers, spectrophotometers, and radiometers in the laboratory, the field, or from aircraft or manned spacecraft. Although subject to the caveats mentioned earlier about signature extension, these controlled measurements are invaluable in defining reference signatures and designing sensors to better discriminate a greater variety of classes. Signatures obtained in the laboratory are normally obtained for single materials (pure substances).

A typical lab spectrometer generates illumination that irradiates a sample mount with diffused light scattered from a metal-plated hemisphere that integrates over an area measured in steradians. Light reflected at an angle from the sample surface collimates as a beam through a hole in the hemisphere. A chopper (rotating signal divider) interrupts the beam by alternately passing and blocking the light that goes to a grating and/or prism, which disperses the radiation into its spectrum. A second beam from a reference source of known and constant intensity reaches the dispersing element during the blocking phase. The dual beam signals then go to a detector that scans them (i.e., moves through the dispersion angles) to measure reflectances (as intensities) as a function of wavelengths. The signals are amplified and then plotted on an X-Y recorder. We display these signals as the ratio of sample reflectance to the reference output, to derive total reflectance (specular and diffuse components). Some instruments can vary the angles of incidence and observation to derive bidirectional reflectance, in which intensities vary with the angles selected (displayed in a series of curves).

Operating the spectrometer in a near vacuum eliminates the effects of absorption bands in the atmosphere. Thus, a series of fixed angle reflectance measurements on some typical rocks from Wyoming (which you first saw on page 2-1), made using a lab (in-doors) spectrometer, appear below for your re-examination. The absorption bands at 1.9 µm and 2.3 µm are associated with mineral constituents and pore water rather than atmospheric gases.

In recent years, the term hyperspectral has come to apply to spectral curves that are either continuous over a broad range, such as the Wyoming group, or consist of a large number of individual narrow-wavelength (high spectral resolution) channels that are so closely spaced that they constitute an almost quasi-continuous spectral curve. Until the last decade (see AVIRIS described later in this Section), it was technically very difficult to operate a spectrometer from fast-moving air and space platforms, because the instrument was unable to dwell on a small target (IFOV) long enough to scan the full spectrum. This limitation is the main reason Landsat, SPOT, and other sensor systems have had to use broad wavelength bands that integrate the variations in spectral intervals into single values for the reflectance ranges they represent.

Using spectrometers and related instruments in the field has the advantage of looking at surfaces that contain the mix of components that make up the classes of usual interest. We can often obtain spectra for each component and then back away to get the full mix. This collection of component spectra assists in interpreting the mix response. Also, illumination conditions from solar irradiation, plus diffuse skylight and multiple reflections from ground surroundings that can contribute 10% to 25% of the total, during sunny or even overcast conditions, are best in the outdoors.

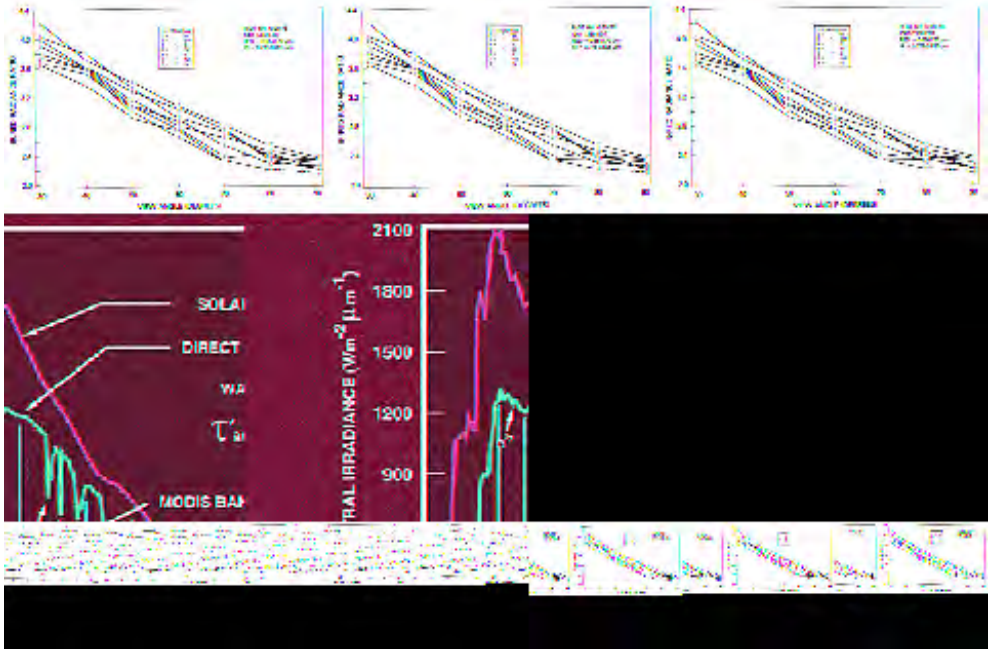
13-10: Suggest three other advantages to acquiring spectral data in the field. ANSWER

One of the simplest devices is a hand-held two- or three-band radiometer, such as the one held by the writer (NMS) in this illustration:

The instrument has its own portable power source and recording system. The spectral bandwidth is typically 0.05 to 0.10 µm. Common channels are in the green, red, and near-IR. These are especially pertinent to calculating Vegetation Indices (page 3-4).

13-11: Specify narrow band wavelengths for vegetation detection using a two or three band radiometer (if you have forgotten what vegetation signatures look like, check Section 3). ANSWER

We can take readings at various look and solar angles, and on different dates and times of year to provide records of spectral variations in the same test areas. These variables can have a pronounced effect on the character of the spectral response, and hence the interpretation of changes, as indicated in this plot of *bidirectional* reflectances, showing the variation of IR/red-band radiances as a function of view angle and azimuth (compass) direction. Each plotted curve, here for a vegetative target, results from the operator changing to a new compass direction and taking a reading at the same angle.

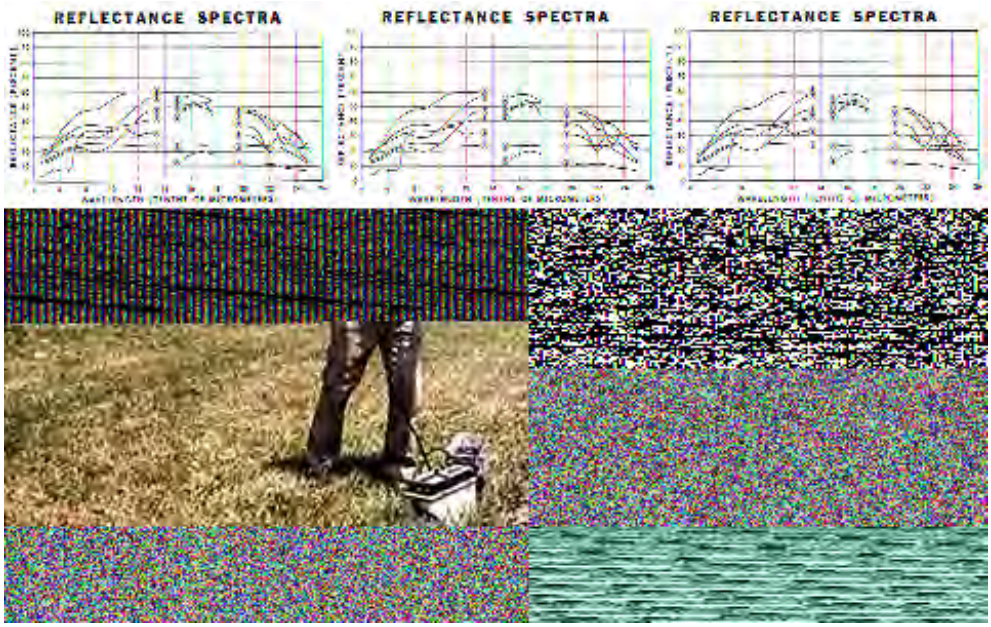


13-12: Formulate a generalization about the implications of these curves. ANSWER

In general, spectra of rocks show much less bidirectional variability than do vegetation spectra. For forests, irregularities in canopy shape, leaf or needle shapes, and species mix can have notable influence on response as a function of viewing and illumination geometry. Portable field spectrometers are now in common use. The next image shows a typical setup for a portable

reflectance spectrometer developed by the Jet Propulsion Lab (JPL) in the mid-70s:

With this instrument, an optical head gathers the reflected light and passes it through a filter wheel operating between 0.4 and 2.5 µm onto a cooled detector, made of lead sulphide (PbS). The backpack contains a power source, amplifier, and recording (analog to digital) assembly. After the ground target scan (of 30 seconds or less duration) is complete, the operator quickly repeats the scan on a a flat reference plate made of a high reflectance (near white) material. Sometimes a black plate is used as well, to fix both ends of the reflectance range. Over a brief time span, the scene lighting remains about the same, but over longer periods, changes in clouds, sun angle, etc., may cause variations in spectral response. However, the spectra can be normalized by dividing the target readings with the reference values. The assumption here is that variations in irradiance over a series of paired readings taken minutes to hours apart cancel, since differences in the illumination conditions affect both target and reference in the same way at each sampling time. For example, if at time 1, at some wavelength, the target reads 30 and the reference 90 and at time 2 they read 20 and 60 (percent), the normalized spectra are both 0.33 (33%) in reflectance units. In the next image, we show field (in situ) reflectance spectra of rocks (1-5) and ponderosa pine (6) acquired by this instrument:



13-13: These spectral curves show much less structure (peaks and troughs) and less amplitude (intensity) than the laboratory-produced curves for the Wyoming rocks (above). Explain this difference. ANSWER

JPL engineers, leaders in developing ground truth instruments, developed a portable field emission spectrometer that uses argon-cooled, mercury-cadmium-telluride (HgCdTe) detectors to sense thermal IR responses in the 5 to 15 µm spectral region. In keeping with the trend toward using linear array multispectral systems (e.g., SPOT), instead of scanners with filters (Landsat), some field spectrometers today use a fixed grating that spreads radiation over a range of angles onto array detectors made of indium-gallium-arsenic (InGaAs) alloys capable of more rapid scanning and greater sensitivity. For further insights into field spectrometry, consult the review found on the Home Page prepared by Analytical Spectral Devices, Inc., a company founded by Dr. Alexander Goetz (formerly at JPL), a leader in this field. One of their most versatile field systems is this spectrometer/backpack combo that scans between 0.35 and 2.5 µm, with spectral resolution from 3 nanometers at the low end to 10 nm at the high.

Another approach is to operate a spectrometer from a truck in which we mount the sensor head on a movable cherry picker, as illustrated below. This allows us to vary the Instantaneous Field Of View (IFOV) height, so that we can examine larger surface areas.

Often the user obtains the most valuable supporting data from sensors mounted on aircraft that fly over study areas. In remote sensing programs administered by NASA, investigators specify test sites for acquiring ancillary data or system developers do to test prototype (breadboard) instrument and sensor designs, proposed for future missions. Such research missions help to

determine the spectral and spatial resolutions, the signal-to-noise (S/N) response, and the time of day and year that optimize detecting and identifying. We show here part of an earlier fleet of airplanes used for these purposes:

The large plane on the left is a Lockheed Electra that operates up to 7,600 m (25,000 ft). On the right is a C-130, which can be flown at higher altitudes (9200 m or 30,000 ft). The two center planes are RB57F's that fly up to 18,500 m (60,000 ft). The jet near the hanger has sensors for lower-altitude missions.

NASA has also used aircraft that fly above the bulk of the stratosphere at altitudes from 60,000 to 70,000 ft. First were two U-2's decommissioned from the military. In the early 1990's, NASA contracted with Lockheed, the U-2 builder, to developed a larger version of this aircraft, dubbed the ER-2 (ER stands for Earth Resources). One of these appears below in flight across a desert terrain.

Among the complement of aircraft-mounted sensors are one or more film camera systems (including multiband arrays), multispectral scanners (including those using charge-coupled detectors [CCDs]), thermal-IR scanners (such as the Thermal IR Multispectral Scanner [TIMS], see page 9-7), microwave sensors (including radiometers and scatterometers, and multiband radar), and special request equipment not routinely flown.

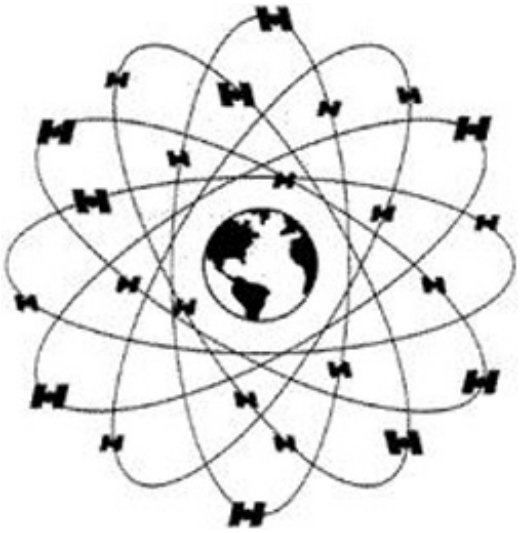
At the other extreme, investigators often need to collect continuous data on the ground at widely separated stations or over extended time periods, often in inaccessible areas. Costs from repeat trips and other factors may preclude sending field parties after initial visits. With such requirements, we set up automated, remote-sensing, sampling sites, at which we measure several defining properties constantly or at fixed intervals.

To accomplish this during the Landsat program, NASA and other organizations, including Principal Investigators in research and verification programs, deployed Data Collection Platforms (DCPs) to measure certain properties on site, coding the results and transmitting these by radio whenever Landsat or some other satellite was in line of sight, and then relaying the data to appropriate ground stations for processing. Typical remote field measurements include: 1) stream heights and velocities; 2) silt loads; 3) snow pack densities; 4) meteorologic parameters; 5) point source pollution; 6) seismic disturbances, and 7) surface tilt on volcano slopes. Some investigators now commonly use networks of remote stations in the U.S. and other countries, such as this example of a stream hydrograph and its transmitter:

13-14: Cite three other plausible uses for DCPs - these need not be directly pertinent to Landsat-type observations. ANSWER

In performing field studies and measurements, it often proves difficult to know exactly where one is located. In the early Landsat days, finding one's self at a site where something in the imagery is being checked out was generally an approximation at best. With the advent and full implementation of the Dept. of Defense (DOD) Global Positioning System (GPS) program in the late 1960s, a new means was developed to locate a point on the Earth's surface to an accuracy within 70 m or less, and to just a few meters when special adjustments were made. Refer to page 11-6 for an earlier discussion of aspects of GPS that we supplement with the material below.

When all the needed satellites were launched over an extended period, this configuration was achieved:



There are six near circular orbits around Earth, each 60° from the orbits on either side. The

orbital altitude is approximately 11000 nautical miles (20,200 km). Each orbital path contains 4 properly spaced satellites, for a total of 24 (currently 27; 3 are spares). Each satellite makes two complete orbits in a day - passing an area of the globe once every 12 hours. The orbital mechanics are organized so that from any point on Earth a minimum of three, and usually 4, satellites from several orbital planes are in positions where they can read radio signals from a ground transmitter/receiver.

Communication with a GPS satellite by radio allows the ground-satellite transit to be a simple ranging action. That is, knowing the speed of radio waves (EM, the speed of light), the system on the ground measures the precise delay time in receiving the sent signal. Adjustments for various factors improves the quality of location. Since the particular satellite's position is also known, the length of the signal vector is established. This in effect results in a virtual sphere of space defined by the vector length. That signal is also sent simultaneously to two other "in view" satellites, yielding two more vectors (radii) and two more spheres. Location now becomes a straightforward problem in triangulation. The intersection of the three spheres produces two common points, one of which makes sense - and is selected as the appropriate location - if the user has a good idea where he/she is. The returned signal from a fourth satellite helps to reduce or eliminate clock errors (these are related to the atomic clocks onboard each satellite which are not exactly synchronized; a method called Differential GPS, which uses a fixed base station receiver elsewhere to make further corrections can improve accuracy of location to 3 meters).

Presently, the NAVSTAR (Navigational Satellite Timing and Ranging) system is fully operational. Location or geographic position is reported either in latitude-longitude (down to seconds) or in Universal Transverse Mercator (UTM) coordinates. Survey teams supporting ground truth measurements use setups similar to this:

GPS is finding its way into wide civilian use. For example hunters in a game forest may wish to know where they are (lost?). The instrument below is hand-held. GPS systems are also being installed in automobiles so that any driver can almost automatically establish the car's location and can also use a complementary system that determines the best routing to destination.

This closes the first part of this Section. The next three pages explore the meaning and uses behind the "multi" concept.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Laboratory and field spectrometers have proved the value of obtaining a "continuous" spectral curve covering extended segments of the EM spectrum. But for most of the early remote sensing era, technology prevented using spectrometers on moving platforms - aircraft, and certainly not spacecraft. All that changed by the mid-80s when Charge-Coupled Devices and very fast signal readout systems developed to a level where spectrometric instruments flying above the Earth could contain enough channels, each of narrow bandwidth, that could be rapidly sampled and reset so as to function as a working spectrometer. This has opened up a truly powerful new approach to remote sensing, called hyperspectral spectrometery, which already has proven to be a superior identifier of materials found at the Earth's surface - rock types, minerals, plant species, etc. can often be identified specifically. The first truly successful model, AVIRIS, flew on aircraft. Its end products were both spectral signature plots, that closely resembled those produced on the ground, and a large number of images made from any of the plus 200 individual channels (detectors) comprising the system. Each such product was valid for the small "piece of real estate" (the pixel) examined at any moment; the resulting array leads to images of a much larger area.

HYPERSPECTRAL REMOTE SENSING

Imaging Spectrometers

Almost all sensors that are multispectral in function have had to sample the EM spectrum over a relatively wide range of wavelengths in each discrete band. These sensors therefore have low spectral resolution. This mode is referred to as broad-band spectroscopy. Spectral resolution can be defined by the limits of the continuous wavelengths (or frequencies) that can be detected in the spectrum. In remote sensors an interval of bandwidth of 0.2 µm in the Visible-Near IR would be considered *low* spectral resolution and 0.01 µm as high resolution. (The term has a somewhat different meaning in optical emission spectroscopy, where it refers to the minimum spacing in µm or Angstroms between lines on a photographic plate or separable tracings on a strip chart.)

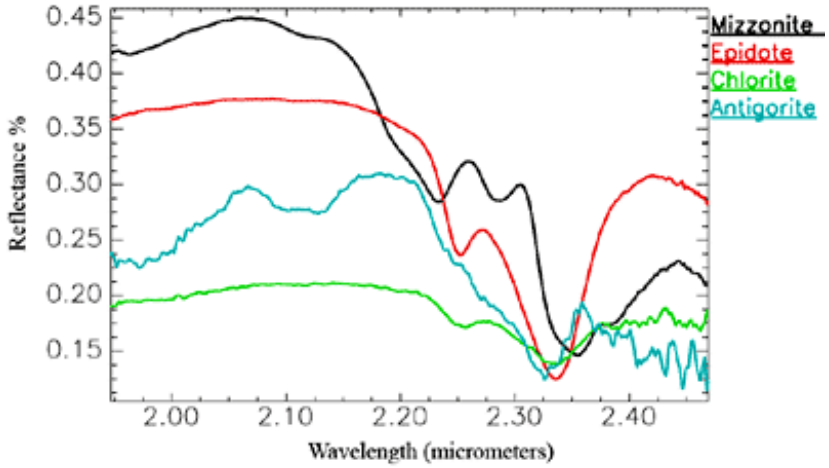
It is now possible to operate remote sensors that can have high spectral resolution. Terms referring to high spectral resolution systems is *hyperspectral remote sensing*, *hyperspectral spectroscopy* , *imaging spectroscopy*, and *narrow-band imaging.*

Since higher spectral resolution leads to much more definitive information on the composition and certain physical properties of different materials and objects, being able to build and fly sensors with such resolutions has been a goal since remote sensing was in its infancy. Beginning in the 1980s, Dr. Alexander F.H. Goetz and his colleagues at the Jet Propulsion Laboratory began a revolution in remote sensing by developing a powerful new instrument called AVIRIS (for Airborne Visible-Infra Red Imaging Spectrometer).* This instrument took advantage of new detector technology to extend ground-based spectrometers into the air on moving platforms. Thus, the distinct value of obtaining hyperspectral curves has made it possible to acquire detailed data on the materials and classes present on the Earth's (or other planets) surfaces. Essentially these curves are continuous spectral plots that measure reflectances from the ground, water, or the atmosphere in the wavelength region responding to solar illumination (Visible-NearIR-Shortwave IR). The plots also record the fine details of absorption phenomena.

With these hyperspectral curves, it is practical now do rigorous analysis of surface compositions

over large areas. Moreover, data can be displayed either as spectral curves with detail similar to those on the preceding page or as images similar to those obtained by Landsat, SPOT, etc. With spectral curves we capture the valuable information associated with diagnostic absorption troughs, and with images we get relatively pure scenes, colorized (through color compositing) from intervals that represent limited color ranges in the visible or in false color for the near-IR (NIR).

Images can be constructed to show variations in reflectance covering very narrow spectral intervals (e.g., 0.01 micrometers). This means that color composites made from bands that lie astride, or near, significant spectral indicators, such as absorption troughs, can contain color patterns closely tied to specific features. This next diagram - a plot of spectra for four common minerals - helps to elucidate how one would go about making informative images from hyperspectral data.

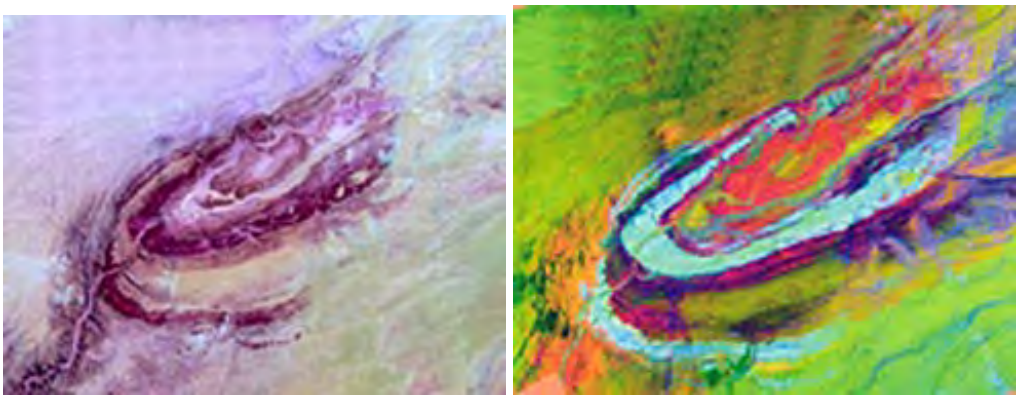


Each mineral has an absorption band in the 2.30 - 2.36 µm interval. The low point of the band varies by about 0.02 µm - just the band width that AVIRIS reads - from each other (Epidote and Antigorite almost coincide). Constructing a 3-band RGB composite using three contiguous narrow bands from the 0.06 µm interval would likely discriminate the four minerals if all are present (this singling out would be better if the minerals were spatially separated rather than co-mixed). Note that even better discrimination would result if one of the colors was used for a narrow band between 2.0 and 2.2 µm.

As an example of narrow band imagery, consider this pastoral scene produced as a true color image from data obtained by another imaging spectrometer (AVIRISwiss'91).

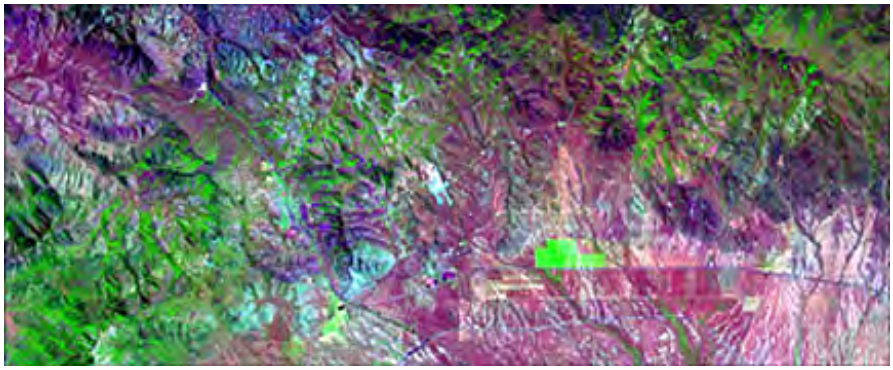
Your reaction to this image is that it looks almost identical to a typical image made from Landsat and other broad band satellites. This is true, largely because broader spectral intervals generally have about the same levels of response (e.g., reflectance), so that if one takes a narrow band

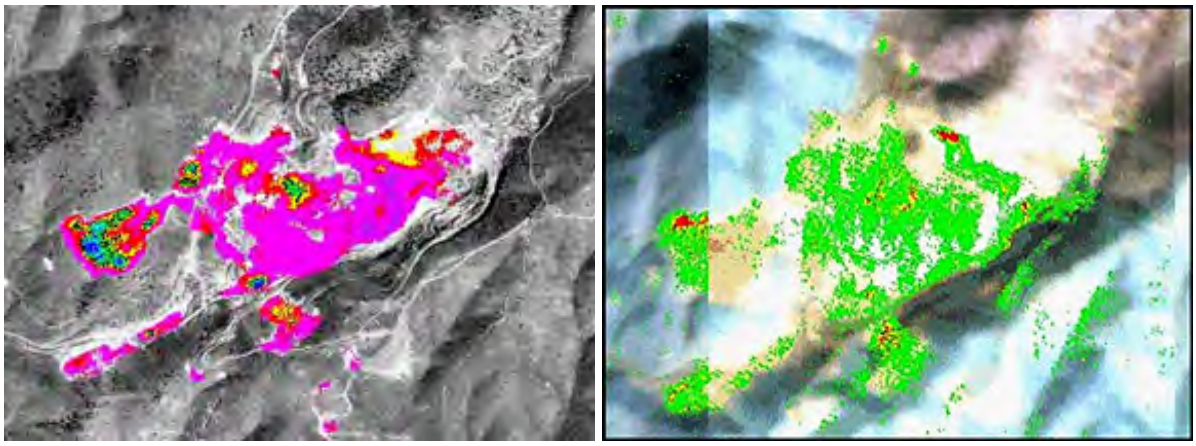
sample to make a color composite, it has about the same contribution as the broad band in which it is included, except if there is a particular absorption trough that is sampled instead. Images made from combinations of bands that are closely sampling the fine structure of a spectral curve will usually be more sensitive to the diagnostic deviations that broad bands tend to miss. Here is an example that makes this point:



Both images show a closed anticlinal fold in the Namib Desert. The left image is a natural color composite made from Landsat data. The right image was acquired with an imaging spectrometer in which 3 narrow hyperspectral bands were combined as a color composite. That image seems to show much more information, although an interpretation was not offered on the website from which both images were extracted.

Examples given on pages 13-8 through 13-10 will help to substantiate these ideas. For now, we will give one concrete example of how narrow band spectroscopy can be quite powerful in identifying materials at the "species" level. AVIRIS was flown over a test area - the Dragon Mine in the East Tintic mineral district of Utah. Various clay minerals are associated with different ore minerals. Kaolinite and a close relative, Halloysite, are two principal species; the precise location of each would be an aid to propecting for the ores elsewhere in the scene. These two minerals have similar spectra but with one diagnostic absorption band located about 0.3 µm apart. First, we show an image made from AVIRIS bands within a 2.0 - 2.4 µm interval. Below that are two "maps" showing the spatial dispersion of Halloysite (purple-red) and Kaolinite (green) in a part of the scene:



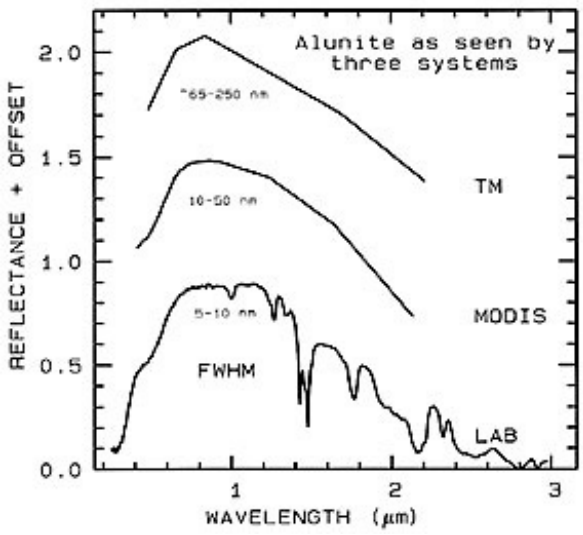


Thus, hyperspectral remote sensing in this case is able to recognize and locate minerals so similar that they could probably not be separated in the field (unless one has a spectrometer handy; small portable spectrometers are available commercially).

13-16: So, once again, in your own words: What is/are the main advantage(s) of hyperspectral remote sensing? ANSWER

As experience accrued through operating AVIRIS, instrument designers built other imaging spectrometers, using different spectral dispersion devices, detector sizes, and other variables. We shall discuss the basics of instrument design later, along with a list of many of these operational spectrometers. As a result a large number of service organizations have emerged or expanded to offer hyperspectral remote sensing from aircraft as a commercial product. By the turn of the millenium, imaging spectrometers had entered space. At one point, the Space Shuttle manifest included a spaceborne version of AVIRIS, known as SISEX, but it fell victim to technological and budgetary constraints. A HyperSpectral Imager (HSI) was launched in 1997 onboard the Lewis satellite, but that system failed to reach a functional orbit.

The obvious improvements in information content gained by operating a spectrometer from air/space platforms are evident in this plot of the spectral response for the mineral Alunite (potassium-aluminum sulphate). Neither the TM nor MODIS, with their broader spectral bands are capable of picking out the diagnostic absorption band that identifies Alunite. The bottom plot has this information; curves produced from imaging spectrometer data would point to the band, so that the spectral data for that band could be used to make images or otherwise analyze for Jarosite locations within a scene containng this mineral.



Before describing these various imaging spectrometers and examining imagery and applications associated with them, we need to take a more extensive look into the fundamentals of spectroscopy than we presented in the Introduction of this Tutorial. To do this, we provide, in

the next three pages, a condensation of a valuable review produced by a leading spectrometrist, Dr. Roger N. Clark of the U.S. Geological Survey.

---

* Note: A benchmark paper at the outset of imaging spectroscopy's appearance as a viable remote-sensing tool is: Goetz, A.F.H, G. Vane, J.E. Solomon, and B.N. Rock, 1985, Imaging Spectroscopy for Earth Remote Sensing, Science, v. 228, pp. 1147-1153.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

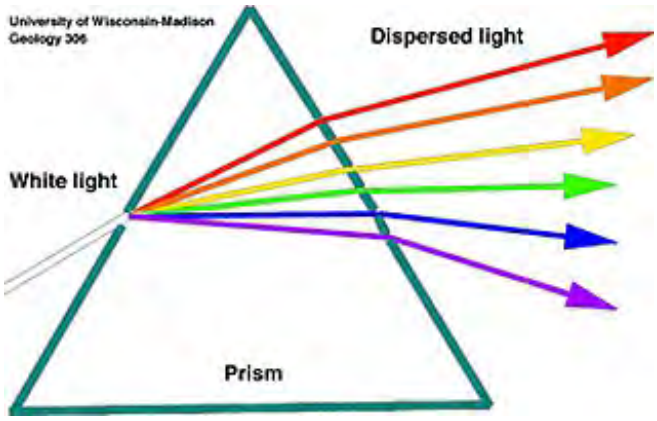
This, and the next two pages, deal with some of the fundamentals of spectroscopy, that branch of physics that considers the subdivision of the electromagnetic spectrum through dispersion into (continuous or discontinuous, depending on mode of excitation) wavelengths whose intensities vary in a characteristic pattern for each individual kind of material tested. The spectral variations depend on atomic and molecular types and their bonding within a given material. Radiant excitation of the atoms in the atomic structure give rise to emission at various energy levels, each corresponding to a discrete wavelength. Thus, when a substance is strongly heated or subjected to electric energy, its atoms undergo changes in electron energy levels (jump from one orbital shell to another) that upon returning to initial state give off (emit) radiation at characteristic wavelengths distributed discontinuously on a recording medium (such as a photographic plate) after being dispersed by a prism or diffraction grating. Absorption of radiation is another aspect of spectroscopy. Spectral curves that represent a continuous wavelength-dependent response from a material either irradiated (as by sunlight) or emitting because of internal heat content show rises and falls that result from intensity differences (for example, radiance variations that relate to transmittance, reflectance, and absorption). Examples of spectral curves for similar or dissimilar materials are covered on this page, with attention to factors that account for differences between compared curves.

Principles of Spectroscopy

Much of this and the next two pages (until AVIRIS is discussed), has been condensed and reworded from a thorough review by Dr. Roger N. Clark of the U.S. Geological Survey, entitled *Spectroscopy of Rocks and Minerals, and Principles of Spectroscopy*, which you can access in its entirety at a site on the Internet. We took several of the illustrations in this summary from that source. Also, it may help to refresh your memory of the theory of spectroscopy covered on page I-2a in the Introduction. We start this page by discussing several basic concepts associated with the type of spectrometers used to analyze light for its constituent wavelengths; spectrometers used in chemical analysis are capable of separating and identifying specific wavelengths that are discrete and often unique to substances (such as elements) that give off EM radiation when excited.

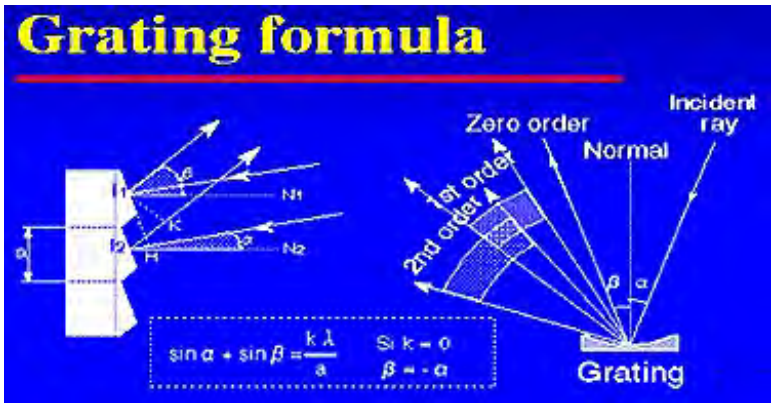
The vital component of any spectrometer is its spectral dispersing device. These instruments have a physical means of spreading radiation composed of differing photon energies (thus a range of frequencies or their inverse, wavelengths) laterally onto a stretched linear display.

The simplest device is an optical glass prism whose cross-section is a triangle. A white (polychromatic) light beam arrives at one side of the prism.



At this glass surface, the rays with different wavelengths bend according to their response to the refractive power of the glass. The degree of bending varies with wavelength, which means that the index of refraction also differs for the range of wavelengths. For example, in crown glass the index is 1.528 for violet and 1.514 for red. If the beam strikes one side of a right-triangle prism at an angle, the light rays slow and bend (refract) to differing extents. When they emerge at the opposite side, they bend again. The net effect is to spread the (visible) light rays according to their effective indices of refraction into a continuous geometric color pattern (see pages I-2 and I-4). Light consisting of just a few discrete wavelengths again will bend differentially to emerge at specific angles characteristic of each wavelength.

A diffraction grating disperses light according to a different mechanism. The grating can be metal or glass, on which are ruled (cut) fine grooves of straight, parallel lines that are extremely close-spaced (e.g., 15,000 per inch), so that the spacing between pairs is roughly equivalent to the wavelengths in the Visible-NearIR region. Each line, analogous to a slit, causes polychromatic light to diffract (bend) at angles that depend on wavelength. For the array of closely-spaced lines (each pair separated by a distance, d), we apply the Diffraction equation: **$n\lambda = d \sin\Theta$** , where n refers to simple integers (1,2,3...) that establish orders (overtones), λ is wavelength, and Θ (Theta) is the diffraction or bending angle for that set of conditions. The diagram below shows a close view of a few of the lines in the grating. The formula in this diagram differs from the more general one above but is related ("a" is the line spacing; "k" refers to the order of spread)



Thus, a light ray striking the grating undergoes spreading at various angles, according to the wavelengths contained within it. Some wavelengths have larger numbers of photons (more intense) than others, so that we can recast a plot of reflected light energies as reflectance versus wavelength, hence yielding the type of spectral curves we have examined in the previous two pages.

An emission spectrometer, used to analyze material compositions, takes light of discrete wavelengths, representing excitation states of different chemical elements, through a slit and then a prism or diffraction grating onto a recording medium, such as a photographic plate. The wavelength-dependent bending reproduces a series of lines (repeated images of the slit) at varied spacings, whose wavelengths, we can measure, and thus identify the particular elements in the sample. The sample is usually heated in a flame or electric arc to force electrons into

higher energy states, because light of given wavelengths is emitted according to the quantized energy when the electrons transition to lower states. Remember Planck's equation **E = hf**, where f is the frequency, discussed in the tutorial Introduction.

Spectroscopy as it applies to remote sensing is the science of measuring the spectral distribution of photon energies (as wavelengths or frequencies) associated with radiation that may be transmitted, reflected, emitted, or absorbed upon passing from one medium (vacuum or air) to another (material objects). In much of the Tutorial so far, our concern has been with reflectance spectroscopy. The exception - Section 9 in which objects emit radiation in response to the temperatures they assume when heated (usually after absorbing radiation from the Sun). Heating will always cause an object or material to experience some sort of radiative emission. High temperatures, such as brought on by solar fusion or by high voltage electrical excitation, produce diagnostic spectra that identify chemical elements and other substances when the emitted radiation is dispersed through prisms or gratings. Typical emission spectra for different elements can be displayed in spectrograms such as those shown here (in absorption spectra within the Visible region the lines appear black against a colored background):

Imaging spectroscopy is the special case in which spectral characteristics and variations of one variable tie to two additional variables (the spatial dimensions, given by x and y positions), to generate color composite images (pictures), ratio and principal-components images, and classification maps. In particular, images that represent the effects of diagnostic absorption bands can be produced to show specific spatial variability of certain material features that one or more such bands discretely identify.

When illumination (either polychromatic, like sunlight, or monochromatic, such as a laser beam) strikes a material, the electromagnetic radiation will likely partition into one or more components that behave differently. Some of the radiation directly reflects. If the material is transparent, the bulk of the radiation passes through but undergoes a change in direction according to the differences in indices of refraction between the material(s) and the external medium (usually air; or water). If the material is translucent or, more commonly, opaque, fractions (wavelength-variable) of the radiation may penetrate. When it penetrates, its rays undergo refraction, but some rays ultimately reflect. Of the fraction absorbed, some converts to heat, so that the temperature of the object rises, causing an increase in emissions, detectable as thermal radiation.

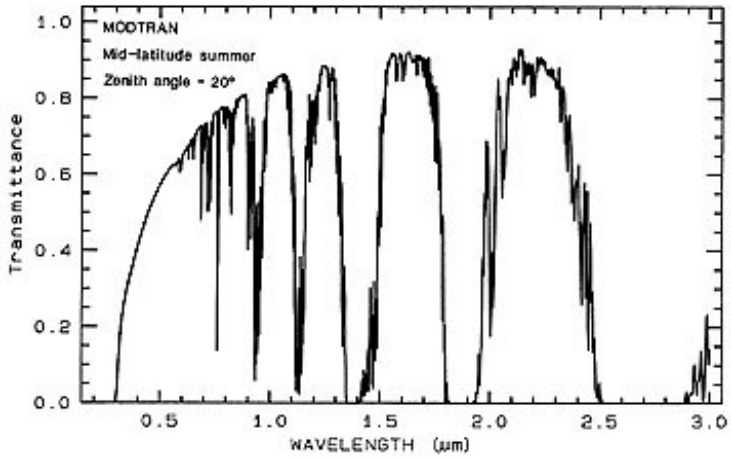
If the material is granular or polycrystalline, light that reflects will strike a number of surfaces (associated with grain or crystal boundaries), meeting individual surfaces at different angles of incidence and thus scattering the radiation at various angles. The light may bounce around, or back and forth, from several such surfaces before finally leaving as scattered beams. If the object's surface is smooth (mirror-like), a significant fraction reflects at an angle related to the angle of incidence. But, most surfaces tend towards some degree of roughness, so that the percentage or proportion of light reflected directly to the observer (an eye or an instrument)

notably decreases. The fraction of radiation that is absorbed also controls the degree of scattering.

For a medium such as air, which also contains CO_2 and other gases, water (usually as vapor or tiny droplets), and particulates, some scattering occurs. Hence, we have blue skies from a high degree of scattering at blue wavelengths, and red sunsets from additional scattering at longer wavelengths, leaving the red radiation less affected, i.e., transmitted as red.

13-17: What is the difference, if any, between the processes that give a blue sky and a red sunset? What color is the sky as seen from the Moon's surface? ANSWER

But, air transmits or absorbs the bulk of the radiation, as indicated by this figure, with most of the minima (absorption bands) due to the presence of carbon dioxide, water, and oxygen.



In the Visible-NIR range (VNIR), water ice and dry ice (solid CO_2) give characteristic spectral curves, as shown here:

Over most of this range, the dry ice remains highly reflective but displays a prominent set of absorption bands around 2 μm. Water ice reflects at shorter wavelengths, but its reflectance diminishes beyond about 1 μm. Note that there are several broad absorption bands that reduce reflectances to very low values. Liquid water tends to absorb well over most of the range, reflecting slightly more in the greens and blues.

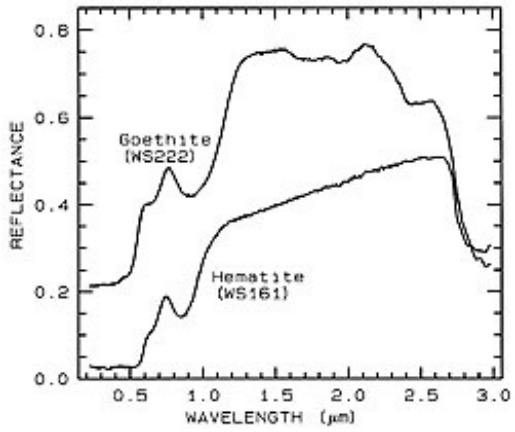
13-18: From the above two curves, beyond the visible what wavelength regions should be avoided in seeking information about the spectral characteristics of materials? ANSWER

While moderate to high reflectances are needed to produce the light tones in pictorial images, the wavelength-dependent absorption bands are the features in a spectral plot that commonly aid in identifying the materials that have bands (narrow to broad) that center at specific wavelengths. In the visible region, the reflected wavelengths control the observed color(s). Thus, a bright green color of an opaque material implies strong reflectance at green wavelengths and near total absorptance in the reds and blues. If, instead, there is also notable red reflectance, yellows to oranges would result. An absorbing medium affects the intensity of incoming radiation according to Beer's Law:

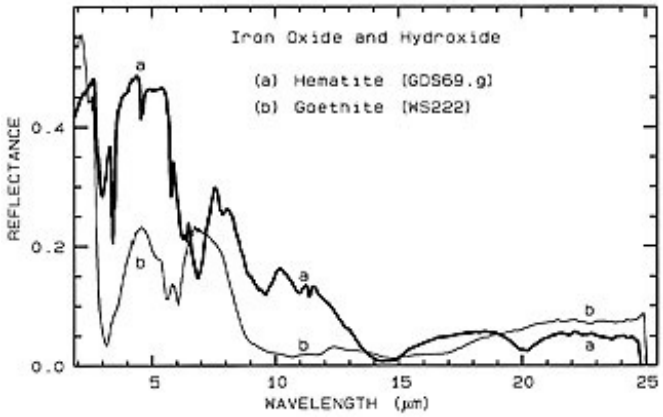
$$I = I_0 e^{-kx},$$

where I_0 is the intensity of the incident radiation, e is the natural log base (2.71828...), k is a constant that depends on absorption as a function of the complex index of refraction (which takes into account the role of the extinction coefficient, K), and x is the depth of penetration. The bandwidth and depth of any given absorption depends on many factors, one of which is the spectral composition of the illumination.

To illustrate how we can use absorption bands plus reflectance levels to distinguish chemically similar materials, we now look at the spectra for two minerals: Hematite (Fe_2O_3) and Goethite (FeOOH). The first spectral curves, obtained with a spectrometer in a laboratory environment, cover the spectral intervals (ranges) between 0.3 and 1.0 µm (VNIR, for visible and near-infrared) and 1.0 to 2.5 µm (SWIR, for Short-Wave InfraRed).



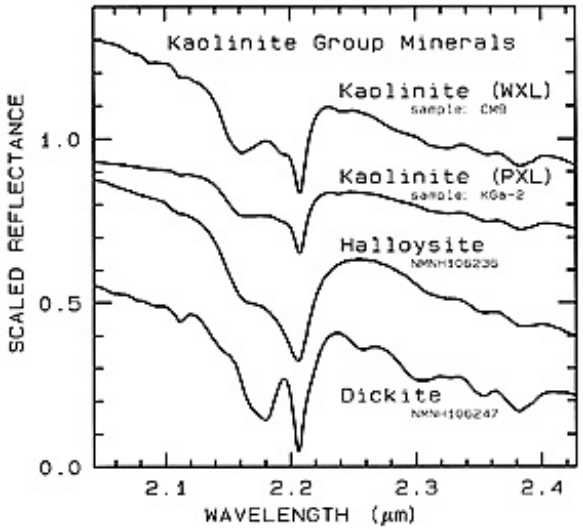
In this plot, the Goethite curve has been raised about 0.2 reflectance units above the Hematite curve to allow comparison. We use this offset procedure to separate multiple spectra that have similar reflectance levels. In this case, the two curves nearly coincide, if we place Goethite at its actual values. The absorption band at 0.86 µm for Hematite and 0.90 µm for Goethite permit us to distinguish the two spectrally but only with a high resolution spectrometer. These two minerals would not differentiate, if a broad-band system, such as the Multispectral Scanner Band 4 on Landsat, were the observing sensor. The OH molecule in the Goethite has a slight influence on its curve at about 2.4 µm. When we examine these two minerals in the mid-IR (thermal) range, a notably different response clearly separates them.



Here, there is no offset, so the overall reflectance of Hematite is greater. Note the dual narrow absorption bands for Hematite around 3 µm. A paired absorption band for Goethite near 6 µm is distinct from the single band for Hematite at 7 µm. At longer wavelengths, including two in the 8-14 µm interval available to thermal sensors, Hematite shows several shallow absorption bands (sometimes called "troughs," as contrasted to "peaks"). As we shall see later with several more examples, these MIR spectra can contain varied and definitive absorption features of considerable utility in distinguishing between and within classes of materials. Unfortunately, no one has developed yet a fine resolution instrument like AVIRIS (operating between 0.4 and 2.5 µm) for air/space platform use.

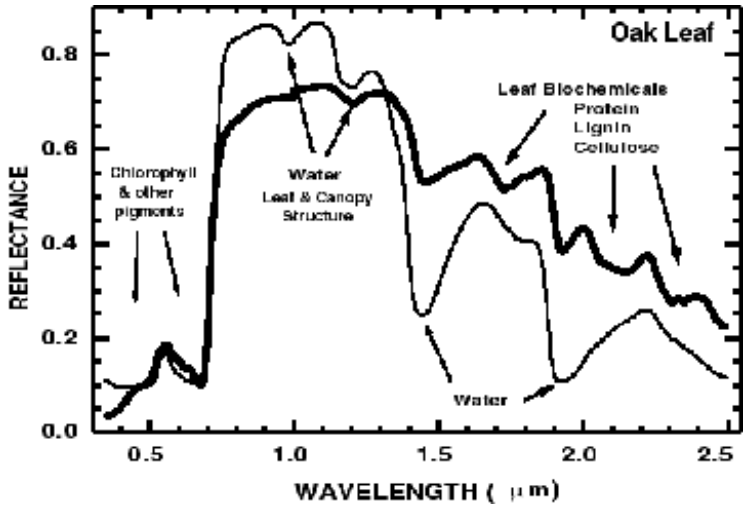
13-19: Disregarding possible atmospheric interactions, what narrow wavelength band is best for separating and mutually identifying goethite from hematite? ANSWER

One more example emphasizes the power of detailed spectra in discriminating similar materials. We present offset curves covering a portion of the SWIR range for several minerals in the Kaolinite family of clays.



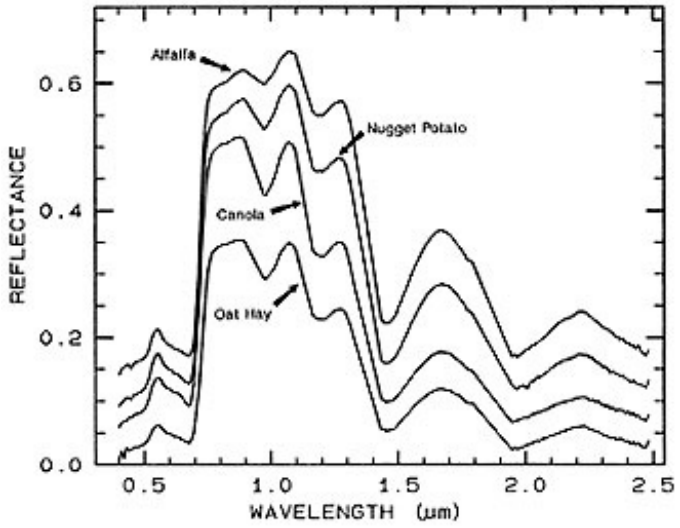
WXL refers to well-crystalized, while PXL refers to poorly-crystallized. Although the gross expression of the absorption features is very similar among the four samples, slight shifts in the absorption band near 2.2 µm and other fine curve shifts suggest that under exceptional circumstances we could distinguish the members of this clay group, but only with difficulty by AVIRIS.

Absorption bands play a key role in defining the spectral curves for organic matter, such as vegetation. Consider this general curve (lighter line width) depicting the VNIR spectrum for a healthy oak leaf.



At longer wavelengths, its pigments and cellular matter absorb light. Water bands also have a notable effect. Chlorophyll absorption dominates in the visible region, removing red and blue reflectances, leaving green as the dominant spectral wave range. The sharp rise in reflectance at 0.7 µm, continuing well beyond 1.1 µm, is largely the result of the walls of multiple cells reflecting the light. The second curve, rendered in heavier line weight, describes the spectrum of an oak leaf that is now dried and brown.

The next illustration shows four spectral curves, each for a particular vegetation type, and each, of the upper three, offset by 0.05 units from the one below. These were sampled from the hyperspectral image of Summitville, Colorado, displayed on page 13-10. In general, these plots are nearly identical, with variations mainly in the depths of individual absorption bands.



Although difficult for the eye to detect and distinguish, there are real differences in equivalent absorption bands that allow us to make separations. The absorption band at 0.7 µm is a case in point. We need to use special processing to single out small differences.

13-20: If the spectral curves for these four vegetation types are so similar, how can we hope to distinguish them in the field? ANSWER

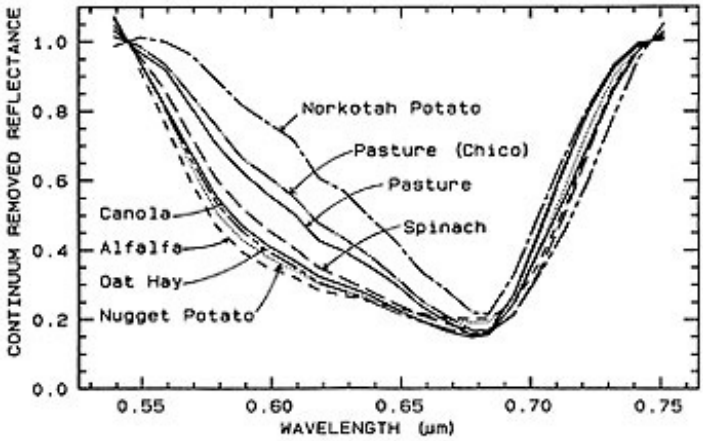
A procedure that facilitates making this distinction is *continuum-removal*. The continuum consists of the so-called, "background absorption," which is in essence an extrapolation of the baseline of the general curve (fits a smoothed curve to the general trend so as to extend across the base of absorption bands). This local reduction specifies the continuum and is determined by mathematic manipulation of absorption coefficients by a subtraction process.

The depth of an absorption band, D, is:

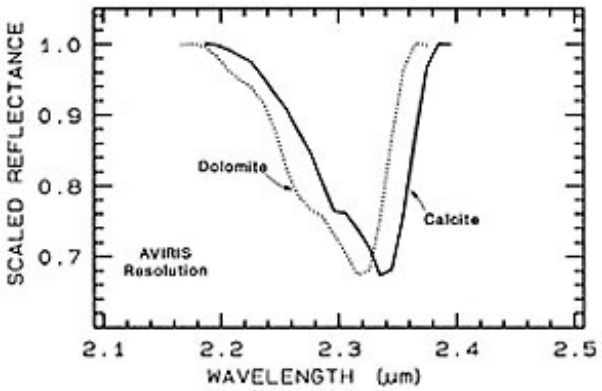
$$D = 1 - R_b/R_c$$

where R_b is the reflectance at the bottom (trough center point) of a band and R_c is the continuum base.

The result for the above four vegetation plots (and several other crop types) is a set of continuum-removal curves that show slight to moderate differences in relative reflectances at a minimum centered on 0.68 µm. At least four of these crops appear distinguishable by their separations in the 0.56 to 0.66 µm interval.



The technique can work particularly well in picking diagnostic bands for minerals that are very similar in crystal structure but differ in substitution of one chemical element (usually as an ion) for another. The common minerals, Calcite ($CaCO_3$) and Dolomite ($Ca,MgCO_3$) have a prominent absorption band near 2.3 µm, which reaches about the same depth in spectra of each species. The continuum-removal diagram for both shows that Dolomite reaches its trough point at a slightly lower wavelength.



This approach to absorption band analysis has proved to be a powerful tool for enhancement and separation of small but often significant differences that allow us to properly identify materials (those belonging to related groups and those unrelated but with absorption bands that tend to coincide).

13-21: Thematic Mapper Band 7 covers the spectral range 2.08 - 2.35 µm. Any chance that dolostone (main mineral is dolomite) can be distinguished from limestone (calcite)?
ANSWER

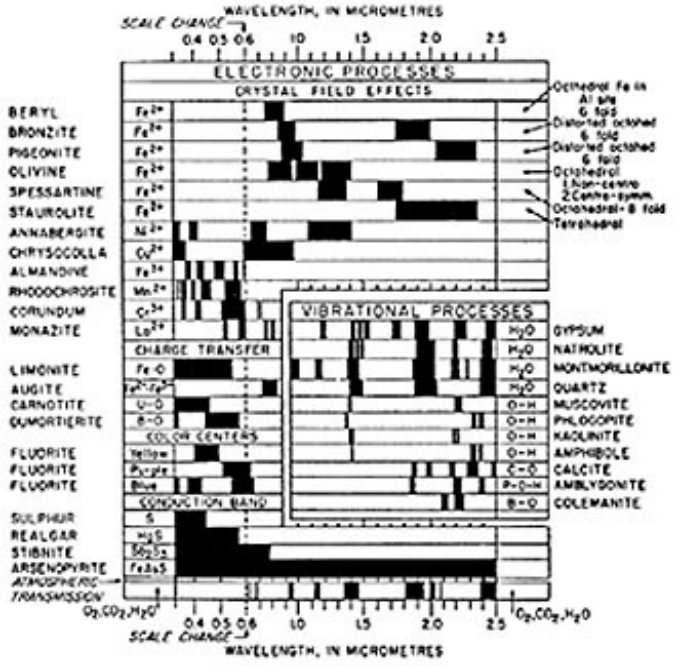
Primary Author: Nicholas M. Short, Sr.

In reading a spectral curve for its information content, the most useful feature is the wavelength location and depth of the diagnostic absorption bands. Several atomic mechanisms are responsible for the absorption of radiant energy that is then expressed by these bands. These include Crystal Field disturbance, Charge Transfer Absorption, Conduction Band shifts, Color Center absorption, and Vibrational processes.

Absorption Processes

In the spectral (signature) curves we have encountered throughout this Tutorial, the parameter given on the ordinate is usually reflectance. In much of the UV, the Visible, and Near IR spectral regions, some of the radiation is absorbed and some reflected. It is the latter that is represented by these curves. But absorption determines how much of the incoming photon energy, at different wavelengths, is subtracted from the outgoing portion of this energy, that we measure as reflected radiation.

The causes of absorption are multifold and different ones may manifest in the same material. We now examine briefly each of the more common processes, which depend on electron behavior in an atom's electronic shells and on bond characteristics during interactions with photons (see review of energy level diagrams on Page I-4 (in the Introduction)). For some of the discussion we will use minerals as examples, since they have been much studied in regards to their behavior under absorption conditions. As a visual guideline to several of these processes, scan the chart below, and return to it as needed during the ensuing treatment. Note that the dark, black bars represent principal absorption bands in various mineral species.



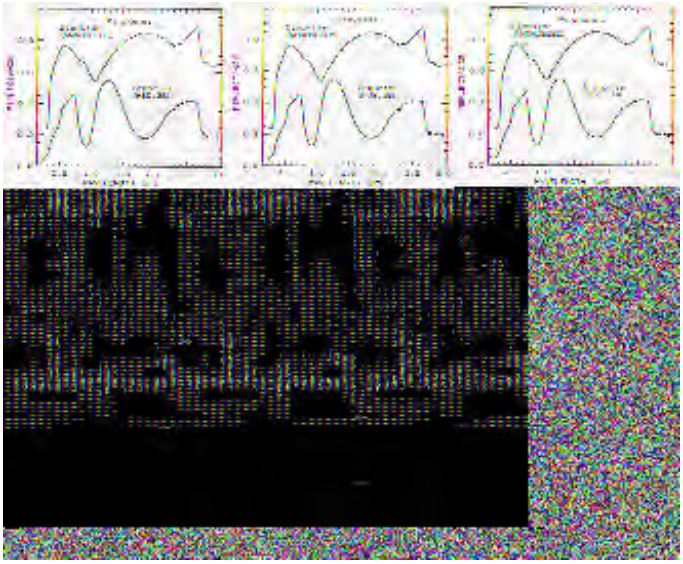
Electronic processes involve absorption of photons with specific energies (hence, wavelengths) that cause an electron jump from a lower energy state to an electron shell at a higher energy state. If the electron returns to a lower state (lower energy shell), it may emit a photon. For shared electrons between atoms, the energy jump spreads over a range of values, producing "energy bands".

A common electronic movement results from Crystal Field disturbance. A crystal field describes the effects of perturbing the "d" orbital shells of a transition metal (Fe, Cr, Ni, Co, Ti, V, etc.), distributed within the atomic lattice of a crystal held together by ionic bonding. This metal cation interacts with the electric field imposed by surrounding anions (negatively charged) or dipolar groups (ligands). The charge symmetry is thus distorted.

Five d orbital shells, with different spatial configurations, are available for occupancy by electrons. These orbital shells have different energy levels (i.e., are split). For a jump to occur, the amount of energy is quantized, that is, requires a photon of a particular energy that equals the difference between the specific levels of the higher and lower states. An electron that jumps may remain for a short period in the higher state (metastable), producing a characteristic absorption, or it may revert rapidly to a ground or lower level. The wavelength of the incident electromagnetic radiation that caused the jump correlates with the difference in the orbital-shell energies. Crystal field theory is important in determining the color and magnetic properties of minerals and other substances containing transition elements.

Iron-bearing minerals, such as Pyrite (FeS_2) and Magnetite ($Fe_3 O_4$), are telling examples of how transition metals influence their spectra. Many minerals contain divalent iron (Fe^{2+}) and trivalent iron (Fe^{3+}). Thus, the valence state, the ionic coordination number, the site symmetry, the ligand geometry, and other factors determine the permissible energy increments.

The influence of iron is evident in this next spectral plot, through parts of the Visible-Near-IR and Short-Wave-IR ranges of two pyroxenes. Diopside ($CaMgSi_2O_6$) contains almost no iron. Bronzite ($[Mg,Fe]SiO_3$) has Fe but no Ca. The presence of Fe^{2+} causes two absorption bands, near 1 and 2 µm , to deepen and shift notably towards lower wavelengths.



A variant of this iron influence is Charge Transfer Absorption. Here, photo absorption brings about a relocation of an ion (commonly, of the transition elements) to another position in the ionic or ligand crystal structure. The result is a series of deep absorption bands, many occurring in the ultraviolet wavelength region and extending into the blue and green. The reds of Hematite (see plot on the previous page) result from this type of absorption, in which the spectrum now shows a strong reflectance peak around 0.75 µm , set apart by absorption on either side.

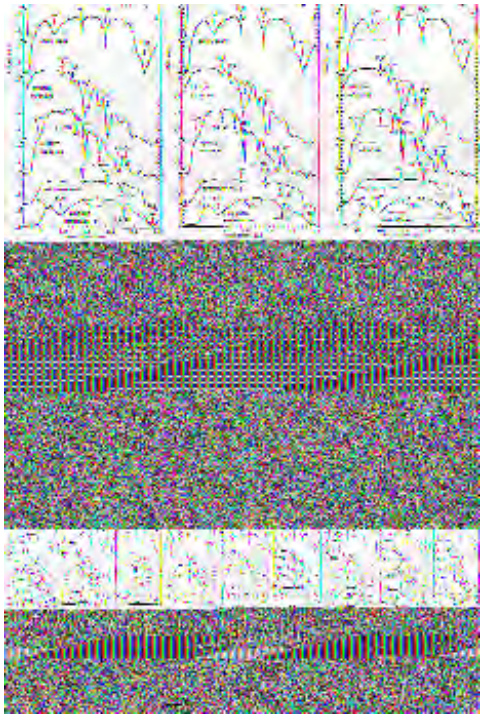
A third process uses Conduction Bands, where electrons move freely through the crystal lattice at a higher energy level (the conduction level) but tend to stay attached to their individual atoms at lower levels (Valence Bands). Photon excitation raises the level through this Band Gap. Not typical of metal-ion behavior, this process is characteristic of semiconductors.

Some minerals, including Fluorite (CaF_2) show a range of characteristic colors, controlled by absorption at Color Centers. These centers can develop from crystal structure defects, or commonly, from impurities. Such anomalies cause absorption at certain wavelengths, leaving one or more reflectance peaks that combine to give the observed colors. Fluorite is normally purple to blue, but green, yellow, and brown varieties are possible. The phenomenon contributes to this mineral's electron jumps, in response to ultraviolet radiation, in which the return to the ground state gives off a color glow or fluorescence.

Separate from electronic processes are the vibrational processes. These involve the bonds in a lattice or a molecular compound. At certain energy levels, the atomic units held by bonds (usually covalent) are set into motion, either as a back and forth vibration and/or as a rotation. The frequencies that these molecules absorb depend on the strengths of the bonds and the masses of atoms or ions participating in the movement. Solids experience smaller vibrations than liquids or gases. Given a molecule composed of N atoms, there are 3N - 6 normal modes of vibration. Each of these constitutes a *fundamental* vibration. There can be additional vibrations at lower frequencies (fixed multiples of the fundamentals), which comprise *overtones.* These overtones normally are weaker.

The majority of vibrational absorptions occur in non-metallic materials and appear throughout the infrared (short and mid-wavelength ranges). They tend to produce rather diverse and complex spectra. Some spectral troughs come from fundamental vibrations, while others come from overtones. In some instances, two or more vibrational modes are possible at the same frequency(ies).

The figure below shows mainly vibrational absorptions for several silicate and sulphate (Gypsum and Alunite) non-metallic minerals. Some of the absorption features are due to water (H_2O) and others to hydroxyl (OH) molecules bonded to iron (Fe) or aluminum (Al). Not shown are Calcite and Dolomite spectra that show multiple absorption troughs in the 2.0 to 3.0 µm interval.



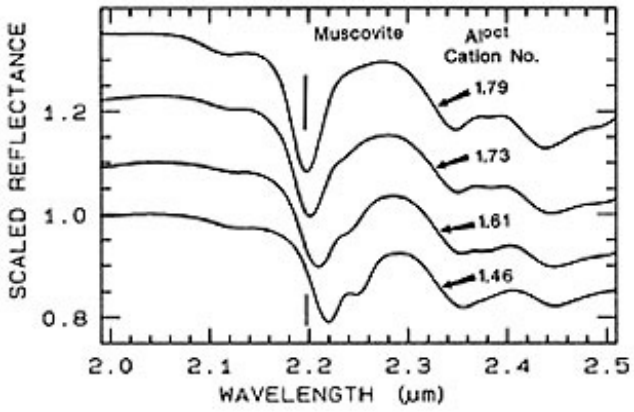
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

A series of spectral curves, produced under laboratory conditions, shows the conditions and factors that cause changes in heights, depths, and wavelength shifts of individual absorption bands or overall increase or decrease in reflectance. Grain size is one such factor; mixing of inert or active materials (multiphases) also brings these changes. Within the crystal lattice of most materials, ion substitutions or impurities will likewise affect the curves. The "mixed pixel" effect (page 13-2) in which several objects, classes, or features are present in the target being sensed is a big concern when using a spectrometer or multispectral scanner in the field. Data reduction and processing is discussed briefly on this page.

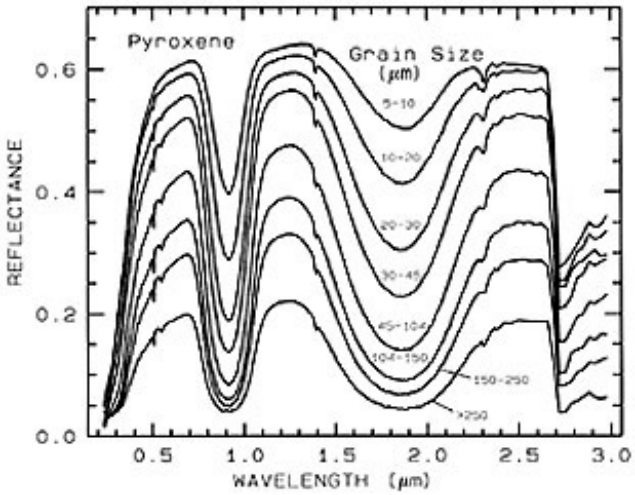
Factors that Modify or "Confound" Spectral Curves; Data Analysis

So far, the spectral plots displayed on the preceding page are mostly pure phases, irradiated under laboratory conditions. Even under ideal conditions, though, for the same species, notable physical and/or chemical variations can exist, e.g., ones related to color or from foreign coatings (such as iron rust). So, a single spectrum, while containing the dominant and diagnostic absorption features for a species, is a standard that we must generalize. Straightforward chemical changes can significantly modify the spectrum. Many minerals, for example, allow substitutions of one element or ion state at lattice sites by other elements, so that the formula for such materials is variable. Take, for instance, the progressive substitution of aluminum in certain octahedral sites that occurs in the common mica, Muscovite. Note how this affects the principal absorption trough at 2.2 µm. Again, the curves are offset.



13-24: Devise a rule for the shift of a diagnostic absorption band with increasing amounts of Aluminum substitution. ANSWER

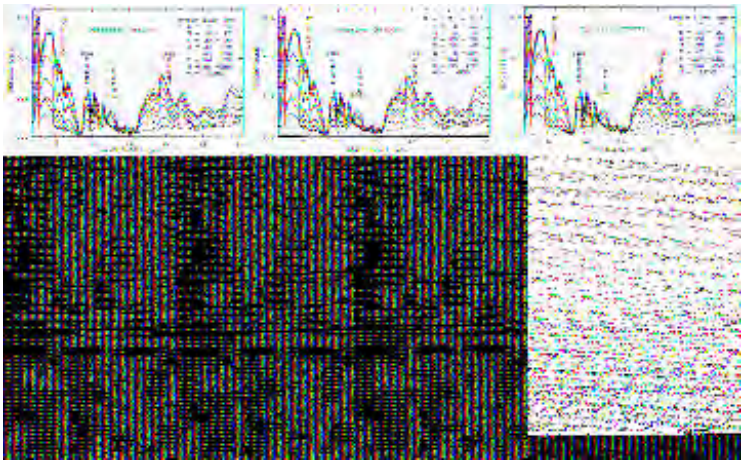
The size range of individual mineral grains or crystals can have a major effect on the characteristic spectral curve of the same species. This usually influences the total reflectance but can also modify the depth of an individual absorption feature. Consider the effect of grain size (in µm) for finely ground samples of the same specimen of a pyroxene species (there is no offset applied to these curves).



As the grain size enlarges, the reflectance diminishes, although the absorption troughs remain relatively constant. Two factors control this effect: the larger grains allow more absorption and the smaller grains provide a higher proportion of surface area available as reflectors.

13-25: Explain the effects of grain size change on the spectral curves. ANSWER

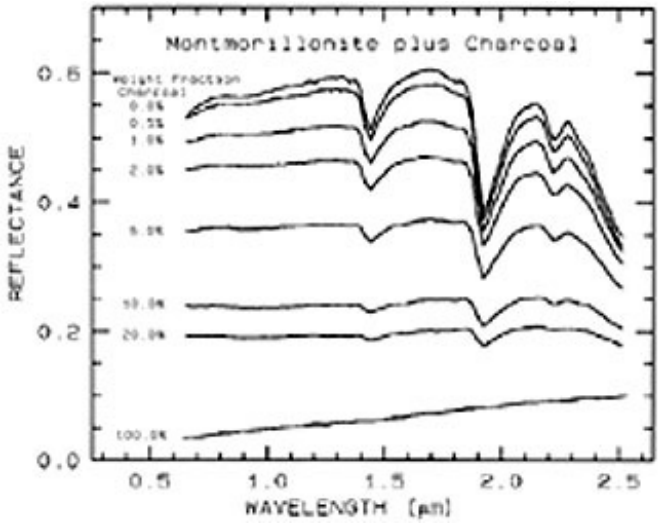
Rather unexpectedly, there is a broad reversal in the relative reflectance as a function of grain size when we examine the same material in the mid-IR region, as shown here:



As we saw earlier in this section, mixed pixels are the rule for most space images and become ever more troublesome as resolution becomes coarser, so that more diversity is expected as the ground area represented by a pixel increases. This problem can be a problem also when we acquire a single spectral curve that represents all the features and classes normally present in a pixel representing Landsat Multispectral Scanner or Thematic Mapper resolutions. Thus, for TM the spectrometer's Field of View would enclose a 30 meter square (this is hard to do from a cherry picker). The resulting spectral plot shows a variety of absorption troughs, some of which we can match to individual phases (classes), whereas others may lie in nearly the same wavelength positions, but are really other bands that are present in common with the phases. The reflectance at any given wavelength is also a composite average of the proportionalized phases contributing to the mixture. Because of the operational difficulties in encompassing large areas on the ground from a near surface position, to get a representative spectral curve one could point the spectrometer's view at a slant rather than straight down. Another way: take close-up spectra of each individual feature or class within the scene, weight the proportions of the various features, and calculate a composite spectral response (signature) curve (possible, but sometimes impractical). The best case scenario for obtaining relevant field spectra is where the scene within a pixel area contains only a few classes, these are small, and all appear close together (to illustrate: suppose the scene contained sagebrush, shortgrass, and bare soil;

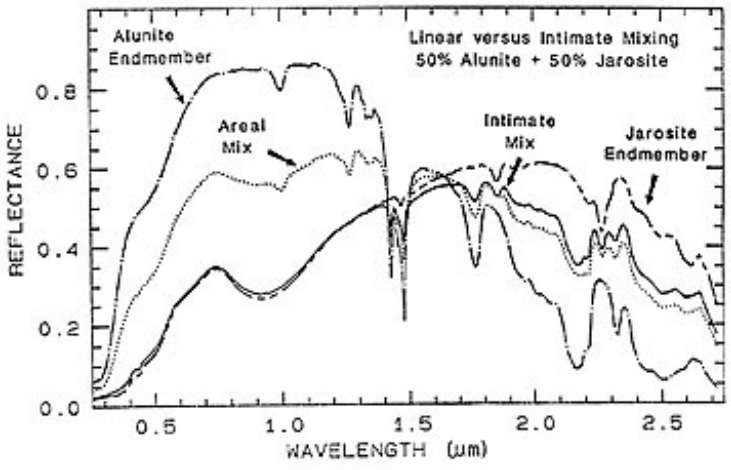
somewhere these would be together, allowing a spectral reading of their integrated response as though they could be described as "grassy brushland".

To illustrate this for two simple cases, inspect these curves. The first is just a series of a two-phase mix, in which the clay mineral Montmorillonite is progressively diluted with charcoal, which has a spectral curve with no absorption troughs and an overall low reflectance.



The reflectances are absolute (no offsets). The result of adding the black charcoal to the whitish Montmorillonite is to reduce the depths of the absorption troughs. Above 20% charcoal, the toughs nearly disappear due to the dominance of charcoal in reducing reflectance.

Now look what happens when we mix two very similar minerals, Alunite (a potassium sulphate also containing aluminum) and Jarosite (the same, except iron replaces the aluminum) in two different ways: Areal, which we compute as a linearly equal combination of the two end members, and Intimate,which we combine and thoroughly intermix and then irradiate under laboratory conditions. The darker (yellow-brown) Jarosite has a lower reflectance continuum at the shorter wavelengths but becomes higher at longer wavelengths. In a 50-50 mix, this darker phase predominates, so that the reflectances are (similar to the effect of charcoal) controlled almost entirely by Jarosite.



From the above considerations, we learn that practical remote sensing, which depends upon imaging spectrometers and hyperspectral data acquisition is especially sensitive to many variables, including foremost, the mixed pixel problem. Because the spectral curves that we can derive from sensors, such as AVIRIS. vary "all over the place," from pixel to pixel, as the combinations of surface features change in a scene, we need some method(s) of reducing the data to meaningful components that we can assign to their proper classes. We must simplify complex spectra, containing contributions from a variety of phases (in a single 30 meter square pixel soil, water, vegetation, buildings, etc. can mutually occur), so that we can extract the

identity and proportions of each. Fortunately, with effective mathematical models and computer processing, we can do this task with surprising effectiveness.

This subject of hyperspectral data reduction and analysis is specialized and somewhat complicated, so that we offer only an outline of the approaches used.

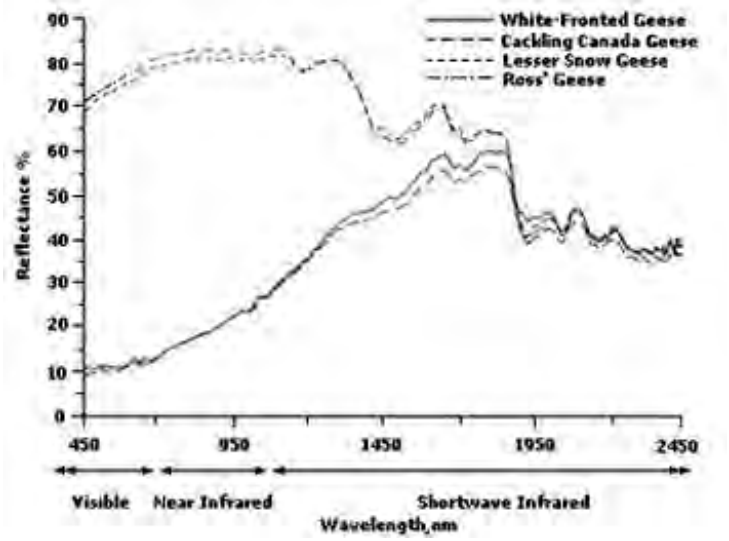
The starting point (and the key) is to develop a Spectral Library. This library consists of literally thousands of individual spectral curves, obtained by spectrometers applied to discrete materials (as pure as possible,) and classes in laboratory and field settings. This library has built up over the years as a data collection developed from many observations by numerous groups. Best known in the U.S. is the library assembled by the Spectroscopy Lab of the U.S Geological Survey in Denver. (For more information, check their Home Page).

From such information, identification algorithms have been devised to quantitatively analyze hyperspectral data, acquired during aircraft overflights. The U.S.G.S. has put together such a program known as Tetracorder, which contains various routines that systematically reduce the raw data to specific identifications. It convolves individual fits of library curves to the hyperspectral data obtained from a mission. In some instances, it can analyze a complex (composite) curve by appropriate algorithms that break it into a set of "end member" curves, each representing a material or class that is a component of the curve. It calculates band locations (indexed to the wavelength at the bottom of the trough) and trough depths, using a continuum base as reference. The weighted fit leads to correlation coefficients that a least squares routine numerically fixes. Band ratios can also prove helpful.

Since spectral bandwidths may be broad, analysis of many individual channels (for AVIRIS, 224 spread continuously from 0.38 to 2.5 µm) is customary. However, an analysis need not sample all the channels for certain problems. If the task is to identify a group of minerals in an alteration zone in a mining district, located in a vegetation-sparse region, the critical interval between 2.1 and 2.5 µm containing a number of diagnostic absorption bands may suffice to identify them. For some identifications, data from two, three, or a few more channels may be sufficient to identify a particular class in compositional detail. From such data, one option is to devise correlation coefficients relating composition to wavelength using least-squares statistics.

Of course, it is usually not expedient to plot spectral curves for every pixel in a mission data set. Instead, the spectral data applicable to training sites containing the classes of interest serve as clues to the individual materials present in any given pixel. If the spatial resolution is high (a few meters), then commonly we assign a dominant class to each pixel. In the real world that we want to picture, that dominant class may extend over many pixels, leading to a hyperspectral image that has similar characteristics to typical Landsat scenes. But, the ability to fine-tune the class identities, and to record separately some of the nature of the mix, are the great advantages that emerge with imaging spectroscopy - certainly a significant improvement in sensor technology for remote sensing.

The writer (NMS) cannot resist closing this page with a *non sequitur*, the oddest spectral curves I've ever seen:



What these curves tell me: 1) at wavelengths encroaching on the thermal region, all the geese have about the same temperature; and 2) two of the birds are all white (Ross's Goose; Snow Goose) and two are much darker (the White-fronted Goose and its similar relative, the Cackling Goose). As a birdwatcher, I've seen all but the Cackling Goose - this last is rare in Pennsylvania but a few show up each winter.

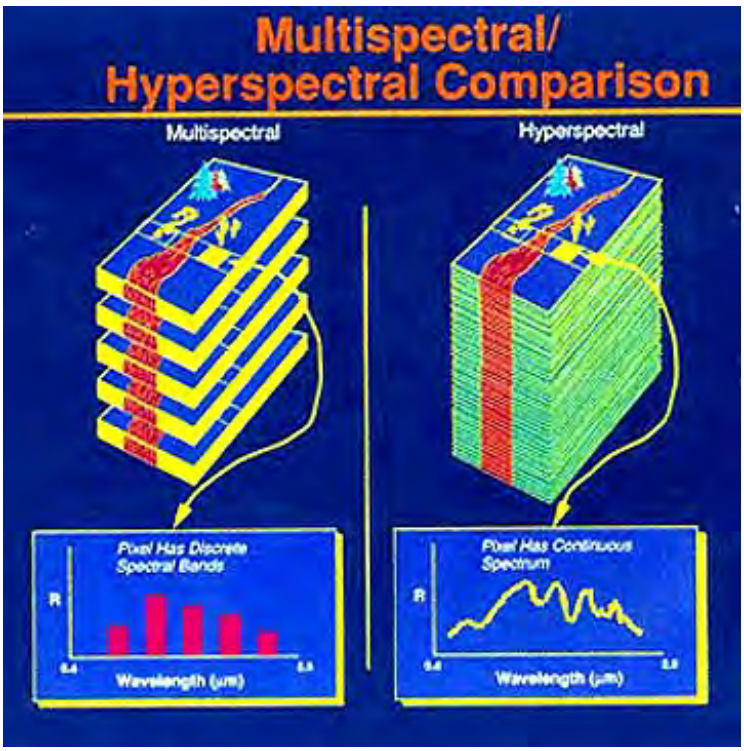
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The advent of Charge-Coupled Devices and the Pushbroom technique for retrieving the detector charges to create signals are discussed, establishing a basis for building AVIRIS. A diagram relates the resulting signal sequence to a sampled pixel; a second illustration shows in an image cube format the final reconstructed image and variations in signal strengths along sample rows and columns. Several other hyperspectral instruments are mentioned.

AVIRIS and other Imaging Spectrometers

The advantage of a spectrometer operating in space over a multispectral scanner, which samples only bands of extended spectral intervals, as hinted at before, is that it samples essentially the full continuum of wavelengths (actually, with present technology, very narrow individual bands that are wavelength contiguous). This illustration helps to clarify the point just stated:



A stationary reflectance spectrometer, looking through a collimating lens at a ground scene, breaks the light into wavelengths emanating from the fixed view. When such a spectrometer is flown on an aircraft or spacecraft, a problem with recording the light arises because the scene moves past the lens (or the spectrometer swings across the scene) at high speed. Older types of detectors didn't have enough time to record the ever-changing field of view, that is, the electronics couldn't sample the dispersed light fast enough to resolve it into the closely-spaced wavelengths needed to construct a spectral curve. Instead, they recorded the light as broad bands, in the manner of Landsat-type scanners.

The technology for a scanning spectrometer that could sweep across moving terrain, while sampling at narrow wavelength intervals, had to await a breakthrough. This breakthrough came with the advent of Charge-Coupled Detectors (CCD's). A CCD is a microelectronic semi-conducting metal chip that detects light. Radiation produces an electron charge on the chip in

proportion to the number of photons received, which is governed by the intensity and exposure time. The charge must be rapidly removable, resetting the CCD for the next influx of photons, such as those coming from the next part of the scene..

A chip is extremely small (tens of micrometers). It is made of some light-sensitive material, such as the silicon used in computer microchips. Normally, we mount hundreds of them on an insulator backing, in a line or a two-dimensional array. Consider a linear array of, say, 2,000 chips per inch. If we allow light to enter through lenses and strike the array simultaneously from one end to the other, then each chip will receive light from a small part of the scene. Adjacent chips will get their light from adjacent ground or air locations. Thus, the instrument samples a line of finite width and length on the ground or in the air.

Each chip, accruing a charge representative of the arriving photon batch, is a pixel that defines a spatial resolution, which depends on the chip size and the height (distance) of the chip array above (from) the scanned scene. After an instant of dwell time over the scene, all of the chips discharge sequentially (producing a signal stream for a recording device) to reset for the next batch. The chips have a high signal-to-noise (S/N) ratio, which allows, in the brief moment, enough buildup of electrons, coupled with gain amplification, to yield usable signals. After a single exposure, and the spacecraft moves on a small distance, the process repeats for the next line of the ground scene. The advance of the linear array, similar to the forward motion of a wide "pushbroom", generates a succession of varying electronic signals that the instrument converts into an image, in which the range of gray levels relate to the signal strengths.

A variant of the linear array is the two-dimensional array, which receives and records light from rectangular ground scenes on multiple lines of chips. Or, another mode uses a rocking or rotating mirror to sweep across the array.

A sensor that uses CCD elements can be multispectral if it uses several arrays, each dedicated to a wavelength band, for which a bandpassing filter determines the bandwidth. SPOT's sensor uses CCD arrays and red, green, and infrared filters to create its multispectral data set.

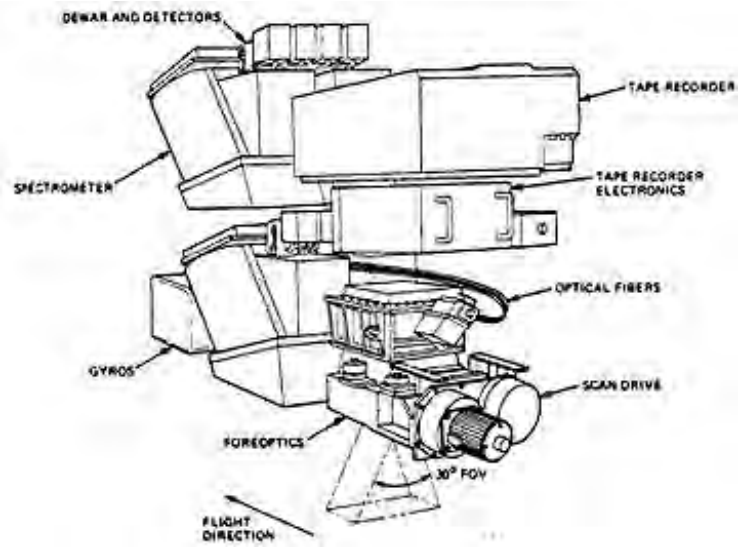
But, these broad-band sensors do not provide hyperspectral data, that is, they do not sample the spectrum in narrow intervals. To accomplish this, the sensor detector must consist of many parallel rows of chips, each dedicated to a narrow wavelength interval, that the recorder can sample (CCDs discharged) extremely rapidly. Imagine a two-dimensional array that is several hundred chips wide and 200 or so long. Let the light enter the sensor through a telescope or focusing lens, impinge upon a moving-mirrored surface, and then pass through a diffraction grating, which disperses the light over a spectral range in the direction of the array length (which

is also the same direction of forward motion of the sensor). At one instantaneous position of the mirror, the light from the ground activates the first pixel chip in the array width-wise, and at the same time, does so for the other wavelength arrays (the spectral dimension). The recorder electronically samples lengthwise the first chip in each line. Next the mirror moves widthwise to the next ground spot and the light from it spreads its spectral dispersion lengthwise. The mirror continues progressively to complete the equivalent of one ground sweep. While this happens, the sensor moves on to look at the next ground position and the whole scanning-dispersing process repeats.

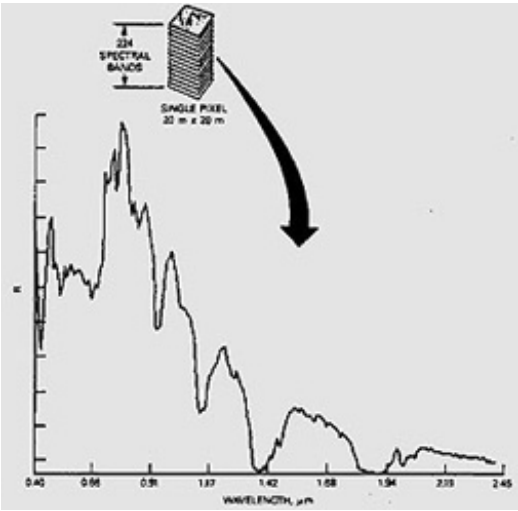
As the instrument proceeds along its flight path or orbit, the final result is a vast collection of data that has both spatial and hyperspectral inputs. From the data set, we can construct images using individual narrow spectral bands associated with small plots on the ground. Or, we can tie spectral data for any pixel position across the width to the wavelengths sampled lengthwise, to plot a spectral curve for that piece of the surface. With special modifications, we can image the atmosphere, if that is the target.

This, in a general way, describes how AVIRIS and other hyperspectral imaging spectrometers operate. JPL's AVIRIS uses diffraction gratings with two sets of CCD arrays, one with silicon chips to sense in the visible range and the other with Indium-Antimony (InSb) chips for wavelengths in the Near-IR to Short-Wave-IR range. A refrigeration unit cools the detectors with liquid nitrogen for optimum performance. There are 224 detectors (channels) in the spectral dimension, extending over a range of 0.38 to 2.50 µm. This arrangement leads to a spectral resolution for each chip of 0.01 µm. By choice, the units that the remote sensing industry adopted for reporting hyperspectral data is the nanometer (nm); 1000 nm = 1 µm. The resolution stated this way is 10 nm, and the range of coverage is 380 to 2500 nm (0.38 - 2.5 µm). AVIRIS gathers its light through a 30° field of view, sending diffracted light to the 614 individual CCD's in the width (across flight) direction. An oscillating mirror scans the scene and sends the incoming radiation to sweep across the array. The spatial resolution derived from this depends on the platform height. A typical mission, mounting AVIRIS on a NASA aircraft (ER-2), produces a spatial resolution of about 20 meters, but that can be improved to five meters by flying at lower altitudes, which, of course, narrows the width of the ground coverage.

This is a picture of AVIRIS:

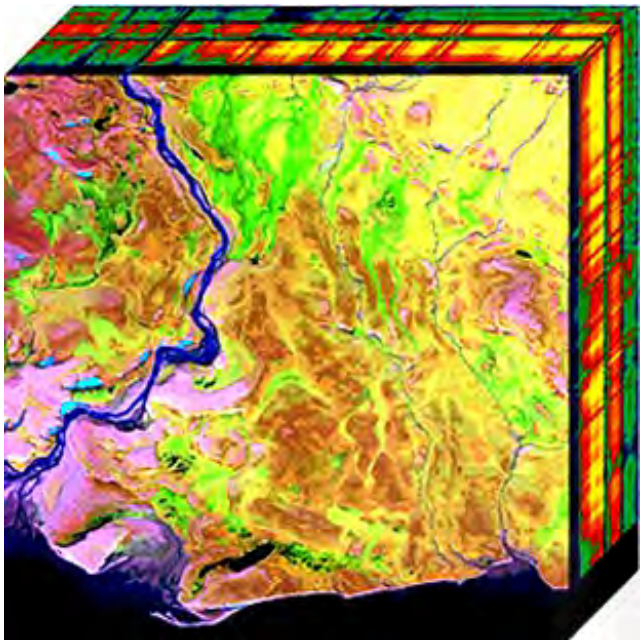


And this is an example of an AVIRIS spectral curve for a single pixel:



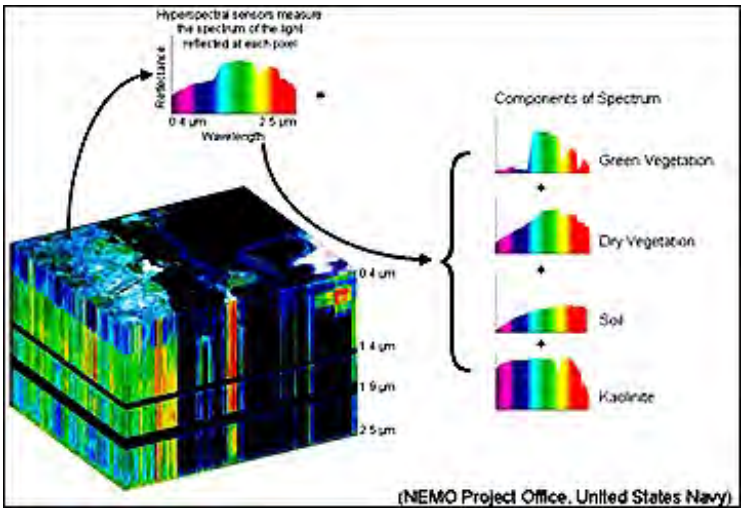
The stacking under the pixel at top is meant to denote the multiple 224 10 nm channels, whose values we plotted to attain this curve; abscissa increments = 0.15 µm.

Another way to visualize the relationship between an image developed in the usual color compositing manner, but using just three 10 nm data values at different wavelengths, and the spectral variations over the interval sampled is to depict this as a Hyperspectral Cube:



Here, the front face of the cube is a color image made from the reflectances associated with three narrow spectral bands in the visible region. On the top and right front sides are the external edges of the thin planes representing each narrow band. The top corresponds to the low end of the spectrum and the bottom the high end. The reflectances for those pixels located along the top and right lines of the cube have been color-coded to indicate various intensity ranges. Black through purple and blue are assigned to low reflectances. Yellow through red and then white denote high reflectances.

Here is another hyperspectral cube, with a twist on the information presented:



At the top is the spectral curve for a single pixel. On the side are four more spectral curves. These were derived from the cube data for pixels in which the stated class name dominates. The top pixel is a mixed pixel composed of weighted contributions from each of the classes that are present in the ground correspoinding to the pixel.

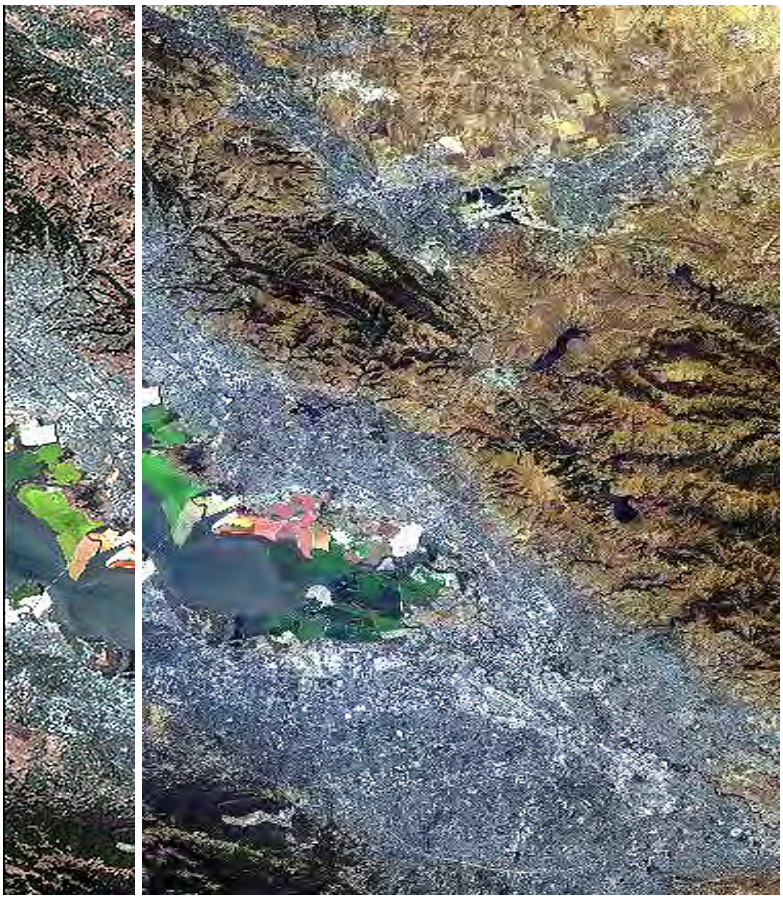
We will show some selected AVIRIS results on the next page. Since AVIRIS, many other imaging spectrometers have been constructed and put online. You can choose a list of most of these, in tabular form.

One of these is HYDICE, developed by the Navy, for aerial use. It has 210 channels, each with a spectral resolution of about 10 nm, extending from 413 to 2,504 nm. It uses a prism as the spectral dispersion device. The spatial dimension is defined by a row of 320 pixels. When flown at low altitudes, HYDICE yields images with resolutions approaching one meter.

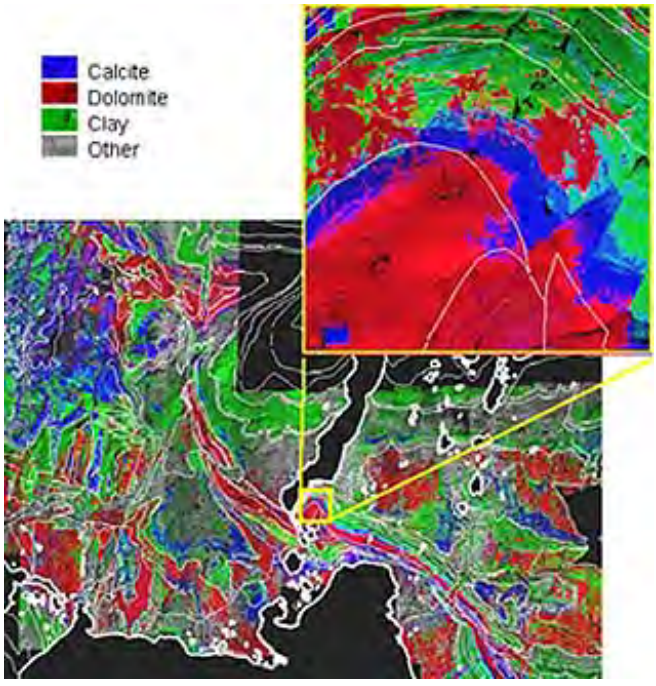
Another instrument, developed in Europe, is DAIS, the Digital Airborne Imaging Spectrometer, which uses a diffraction grating to obtain spectral coverage between 450 and 2,500 nm. Its 72 channels collect radiation in three spectral intervals: 400-1200 nm (bandwidth 15-30 nm); 1,500-1,800 nm (bandwidth 45 nm); 2,000-2,500 nm (bandwidth 20 nm). The gaps coincide with atmospheric absorption bands. As separate sensors using the same optics, it also has a single broad-band channel between 3,000-5,000 nm and six channels that operate within parts of the 8,000-14,000 nm interval. These bands provide important information on thermal emissions from ground objects. Flown low, it provides one-meter resolution imagery for strips a few kilometers wide.

For most systems, the diffraction grating or the prism accomplishes the dispersion, but other techniques of separating the spectrum include interference filters, acoustical-optical filters, liquid crystal tunable filters, Michelson interferometers, Fourier Transform interferometers, and multi-order etalons. Chips made of mercury-cadmium-tellurium (MCT) or platinum silicide are sensitive to certain usable wavelength intervals.

Hyperspectral sensors are beginning to make their way into space on unmanned satellites - part of the trend towards operating these powerful imagers that can yield much more information than those now on Landsat, SPOT, and others. As introduced in the Overview (see page I-2), EO-1, launched in November 2000, had both a standard 9 band sensor (ALI) and a 220 band hyperspectral sensor (Hyperion). We show a Hyperion flight strip across San Jose, CA below on the left and the larger ALI image on the right which includes the strip area (on its left):



The Canadian Centre for Remote Sensing has developed an airborne hyperspectral instrument called Probe-1. Here is the rock units map produced from several of its bands for an area in northern Canada (Arctic):



In 2001, ESA launched its experimental Proba satellite which included CHRIS (Compact High Resolution Imaging Spectrometer) that images 18 by 18 km targets at 18 meter resolution. CHRIS can use its CCD sensors to produce 200 narrow bands. An example of a CHRIS image, with bands chosen to produce a false color composite, is this view of the Ardennes in Belgium.

More information on Hyperion and Proba is found in the Introduction, page 24.

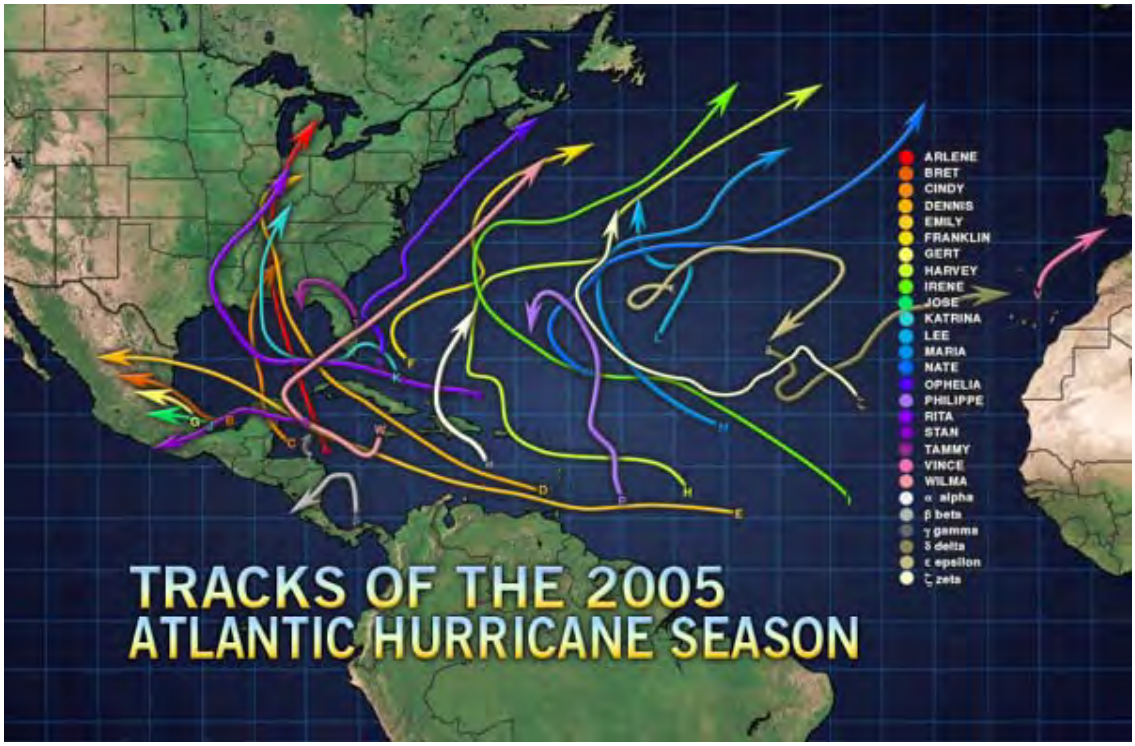
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

This page, added on to the hurricane review on page 14-10, surveys the history and effects of Hurricane Katrina, what is now cited as inflicting the costliest natural disaster ever to happen on the mainland United States. Hurricanes Rita and Wilma, occurring later the same year (2005), are also examined. The role of remote sensing is emphasized, especially as it pertains to the types of damage produced by this devastating storm.

Hurricanes Katrina, Rita, and Wilma

The year 2005 witnessed an extraordinary hurricane season. As of December 31, 27 storms ranked from tropical depression to fullblown hurricanes (9 of those) had developed in the Atlantic and were given names (the male/female naming system has been exceeded, so that the secondary system of Greek letters is now in place). Here is a map showing the names and tracks (paths) of these storms. For comparison below it is a map for 2009 hurricanes - a more normal year:

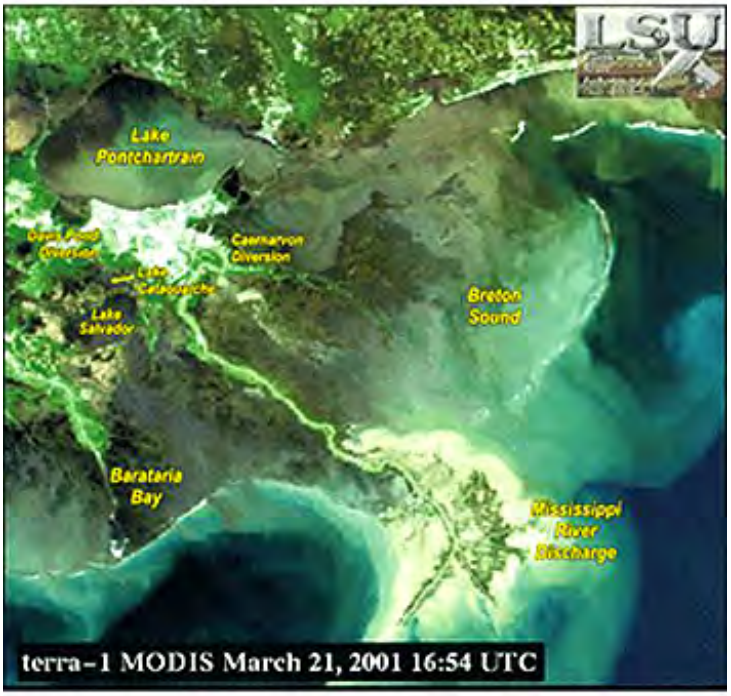


For 2005, the devastation to Atlanic and Caribbean islands, the Atlantic Coast line, and states and countries in the Gulf of Mexico is greater than any past time for which records are kept. The cost in 2005 U.S. dollars far surpasses any previous year. By far the worst destroyer is Hurricane Katrina - now rated as the worst natural weather disaster to ever beset the United States.

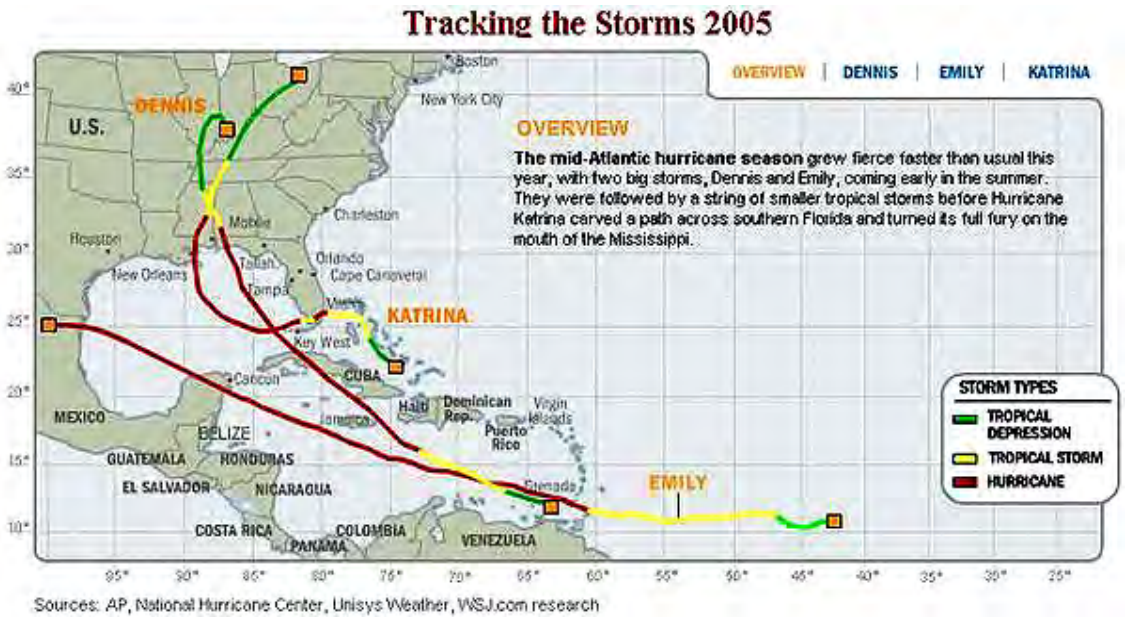
Before you proceed, check out these two websites: Audio-Video review of the 2005 hurricane season, made by NASA (note: the musical sound track may sound loud) and National Geographic summary of Katrina.

On August 29, 2005 and for days before and after, meteorological remote sensing experienced one of its "finest hours" as a monitoring system when the southern U.S. was subjected to arguably the worst natural disaster in North America on record. Hurricane Katrina (the second time this name was use; one called Katrina struck Central America in 1999) slammed into the Gulf Coast as a Category 4 storm (but later analysis has determined it to be a high Category 3 at New Orleans) which caused extraordinary havoc in and around the Crescent City. The death toll in Louisiana exceeded 1700 - most at and near New Orleans. Damage greatly exceeds $100 billion dollars ($55 billion covered by insurance companies), making it even more costly than Hurricane Andrew. Most of New Orleans (around 75%) was underwater for nearly a month after storm surges from Lake Pontchartrain breached several dikes, and anarchy was unfettered for several days. The effects of this massive storm, whose large size (width) accounts for huge area of devastation, have already exceeded the Category 5 Hurricane Camille in August, 1969 - reputed then to be the worst storm catastrophe as judged by the material destruction to hit the U.S Gulf Coast. Towns including New Orleans, Biloxi, and Gulfport may require years to rebuild.

On this special add-on page we will review the story of this national tragedy. First, let's establish the setting through these two images. The top one is an astronaut photo of the Gulf Coast of Louisiana and Mississippi (New Orleans is just at top center edge. The one below is a MODIS image that shows most of the areas hardest hit by Katrina; New Orleans is just below Lake Ponchartrain:

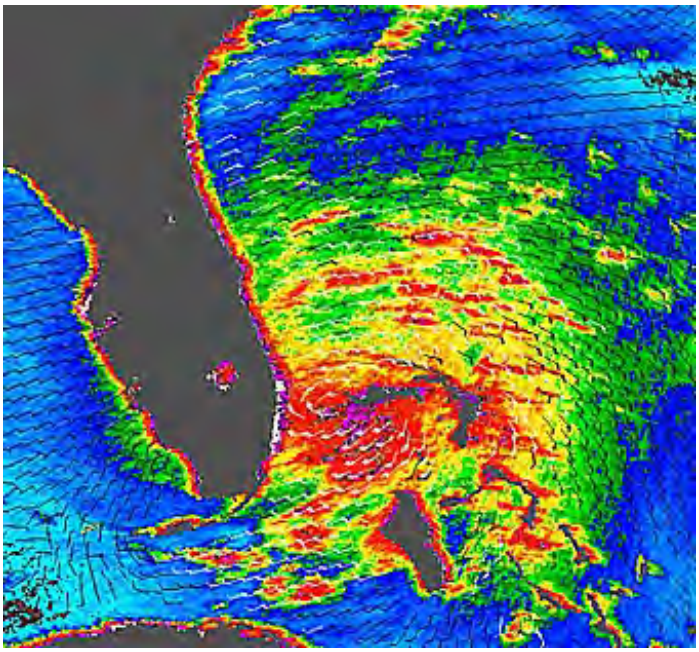


By the end of August of 2005 three significant hurricanes - Dennis, Emily, and Katrina - had hit Florida and/or the Gulf Coast. Here is their tracking history



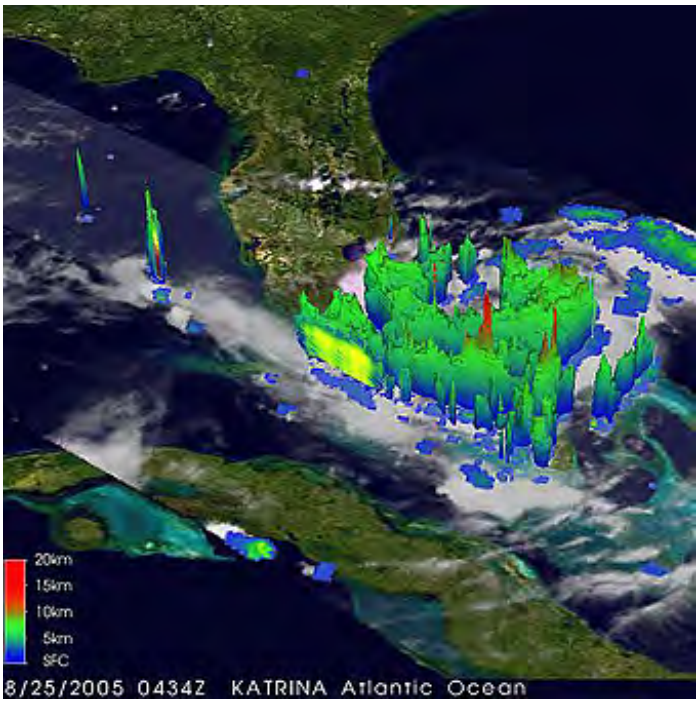
Katrina, by far the worst of the trio, began as Tropical Depression 10 by mid-August in the tropical zone within the south-central Atlantic. (Already, with the main hurricane season still ahead through October, there have been an abnormally high number of such tropical disturbances, some developing into hurricanes such as Emily and Dennis.) That depression apparently dissipated but may have re-emerged as Tropical Depression 12, located several hundred kilometers southeast of the Bahamas. The space image below shows this depression, now named Katrina, in this area just before it became a Category 1 hurricane:

The new hurricane side-swiped the Bahamas to slam into South Florida near Miami. As it passed the Bahamas, the Quickscat satellite obtained this picture of the increasing wind velocities:

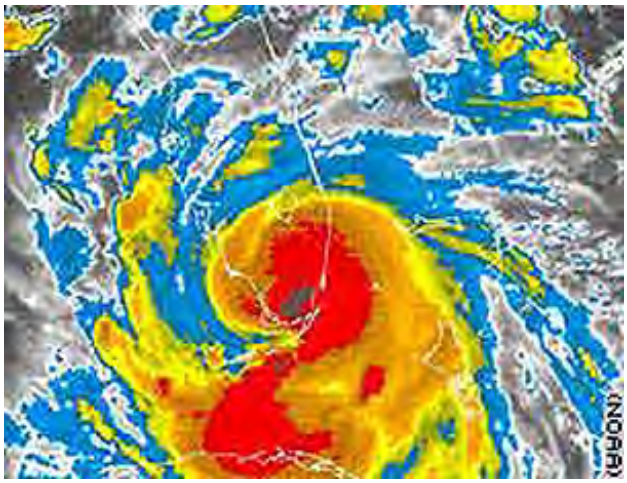


At this stage, the hurricane had yet to develop a distinct eye:

Cloud height data from TRMM is shown in the next image:

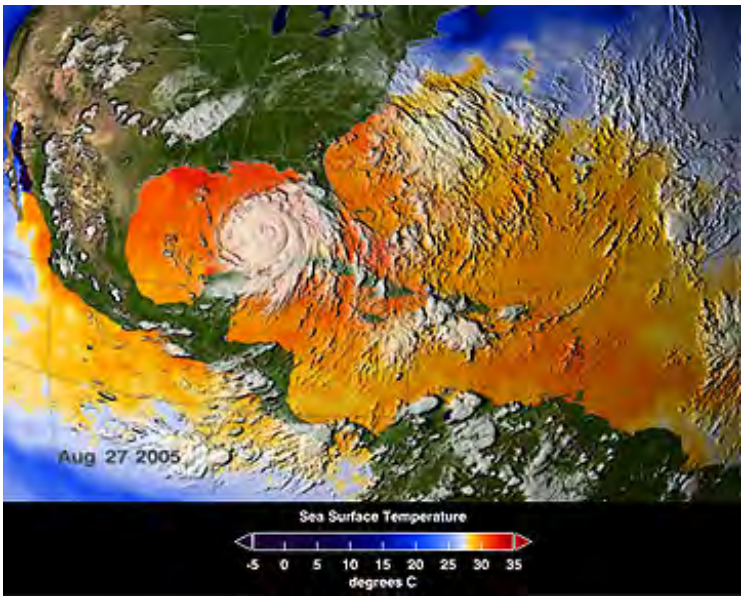


This NOAA satellite image, colored to show general ranges of wind speeds, shows Katrina had developed an eye which passed just north of Miami on August 25. Wind damage was moderate but widespread flooding inundated whole towns. Damage costs amounted to less than $2 billion; 12 lives were lost mainly by drowning.

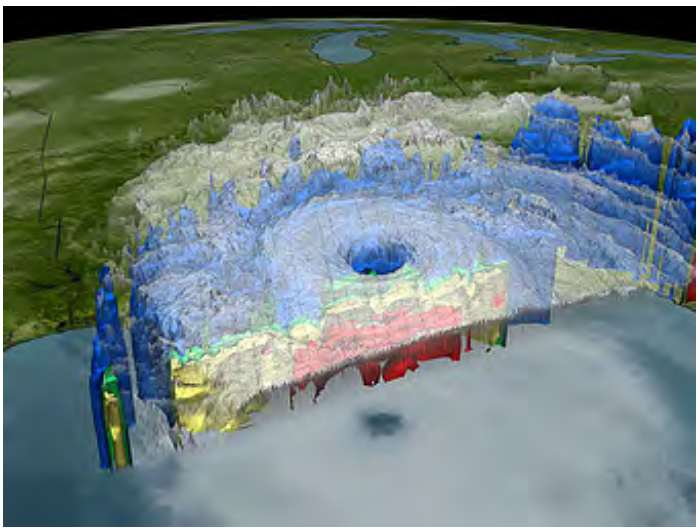


For a while Katrina looked poorly organized and weakened as it entered the Gulf of Mexico. At first it seemed headed westward towards open Gulf water.

However, Sea Surface Temperatures (SST) were quite warm in the Gulf and steering currents began to divert Katrina's path towards a northwest, then a north trajectory.



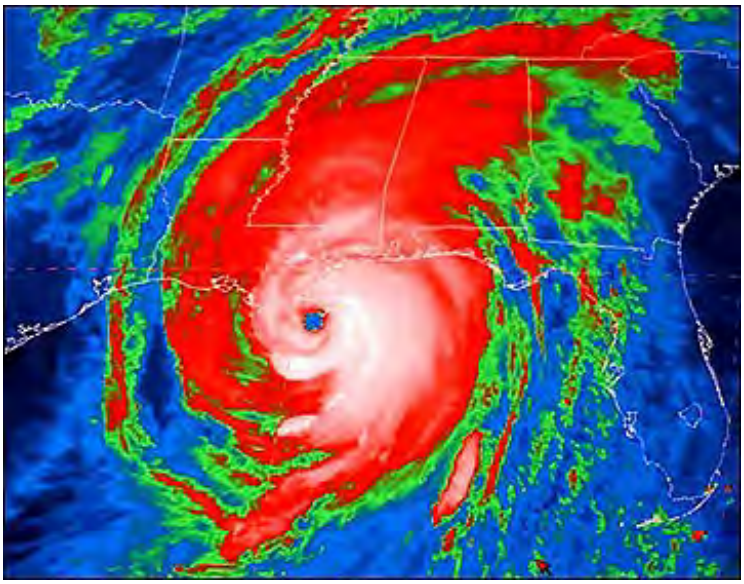
As it moved northward, Katrina quickly intensified into a Category 2, then 3, then 4 strength hurricane. When it hit the New Orleans area, it finally had diminished to strength 3. Rainfall seen by TRMM indicated that it would be a floodmaker when it struck land.



Residents from west of New Orleans and east through Biloxi and Gulfport in Mississippi, and further east through Mobile, Alabama and Pensacola, Florida were first warned to "batten down", and then urged to begin a mass evacuation. More than three hundred thousand escaped New Orleans, spreading out northwards.

Finally, nearing land, the size of Katrina expanded and its winds topped 170 mph, making it for a short time a Category 5 hurricane. At one stage, its internal pressure reached 902 millibars, 4th lowest recorded from Atlantic storms. Evacuation on Saturday and Sunday accelerated. Here is

a NOAA image of the storm at this stage.

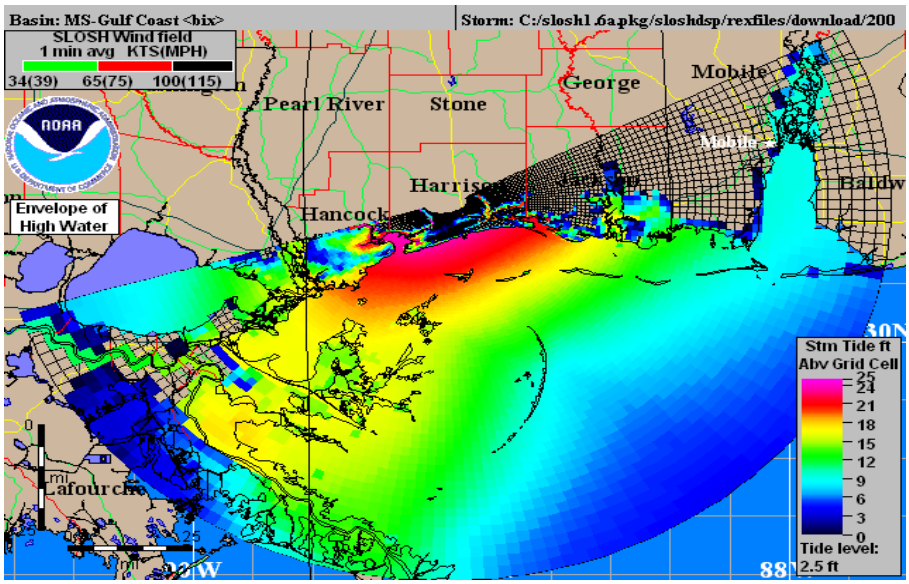


The clouds of Katrina were impressive. The top of the next pair shows Katrina's cloud deck near its eye. The bottom indicates the front of Katrina as it passed inland over sugar cane fields:

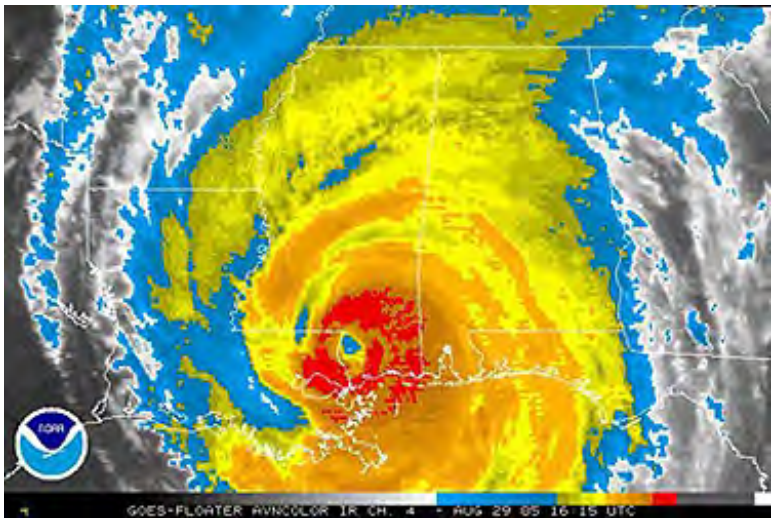
No series of still pictures can give a "you were there" idea of the storm's fury. Those who were there attest to its frightening destruction. Here is just two that show something of the wind and rain effects on a stretch of the Gulf Coast.

Hurricane-force winds can cause direct damage, knocking over trees, smashing windows, and ripping off roofs. But they also can whip up water, creating what is known as a storm surge. Surges from Katrina established records: in several places these reached 10 meters (34 ft) and averaged 5 meters (16 ft) over the affected parts of the Gulf Coast. This map shows the surge

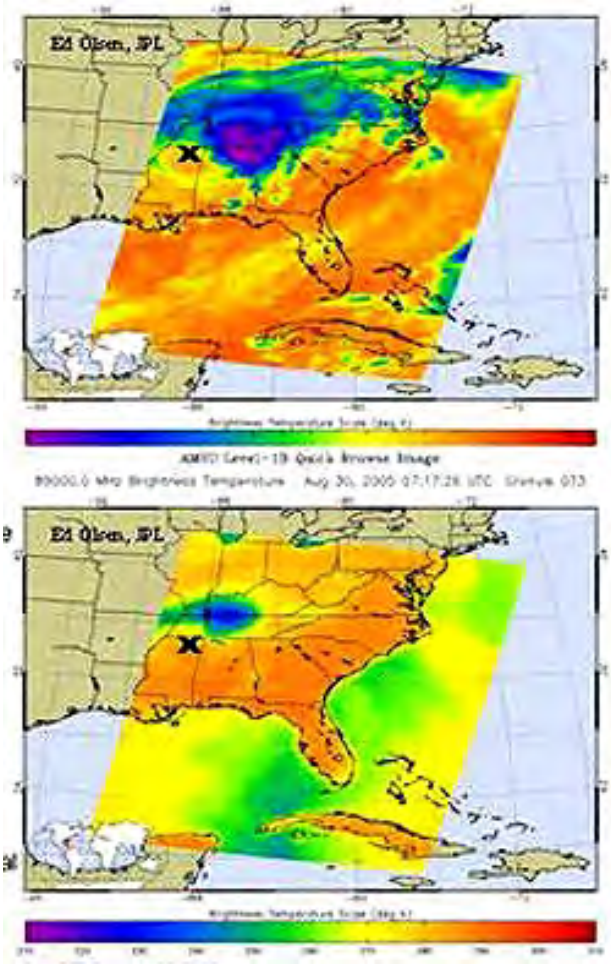
distribution, and a picture beneath it indicates how water is raised up into destructive waves:



The storm, as it passed onto the Mississippi Delta, diminished to a category 3 and steering currents moved it just east of New Orleans, where it made landfall in the morning hours of August 29. Thereafter, it moved inland, constantly weakening, into the Tennessee and Ohio Valleys, into western New York and then Canada. The high level of damage it inflicted on the southern coast covered 237000 square kilometers (90000 square miles).

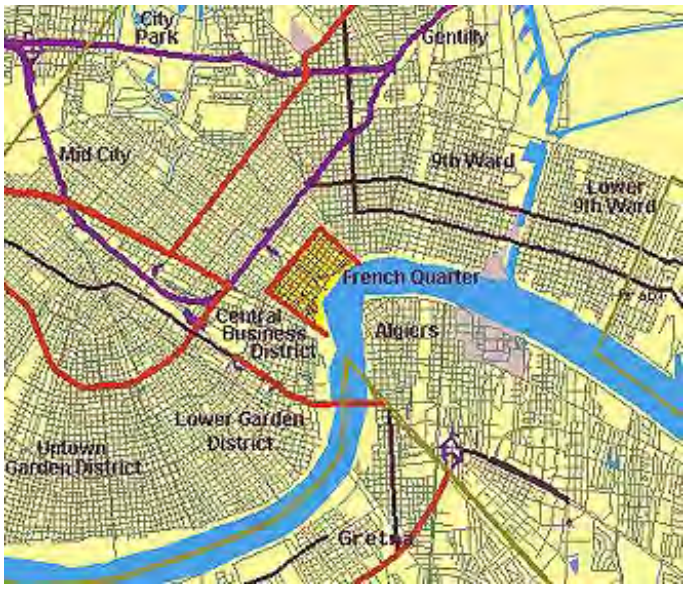


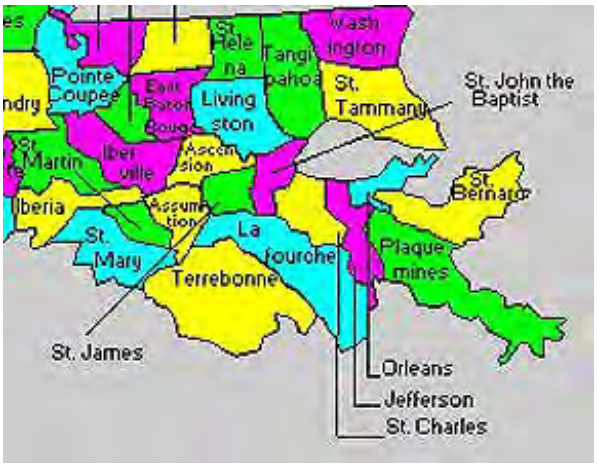
Observations made by the AIRS instrument on Aqua (see Section 16) show the temperatures of the cloud deck above Katrina as it began to dissipate. The oranges in the images indicate still considerable moisture causing the heavy rains in the central U.S. The blue shows lower temperature that correlate with the "drying out" of the central storm, hence less rainfall.



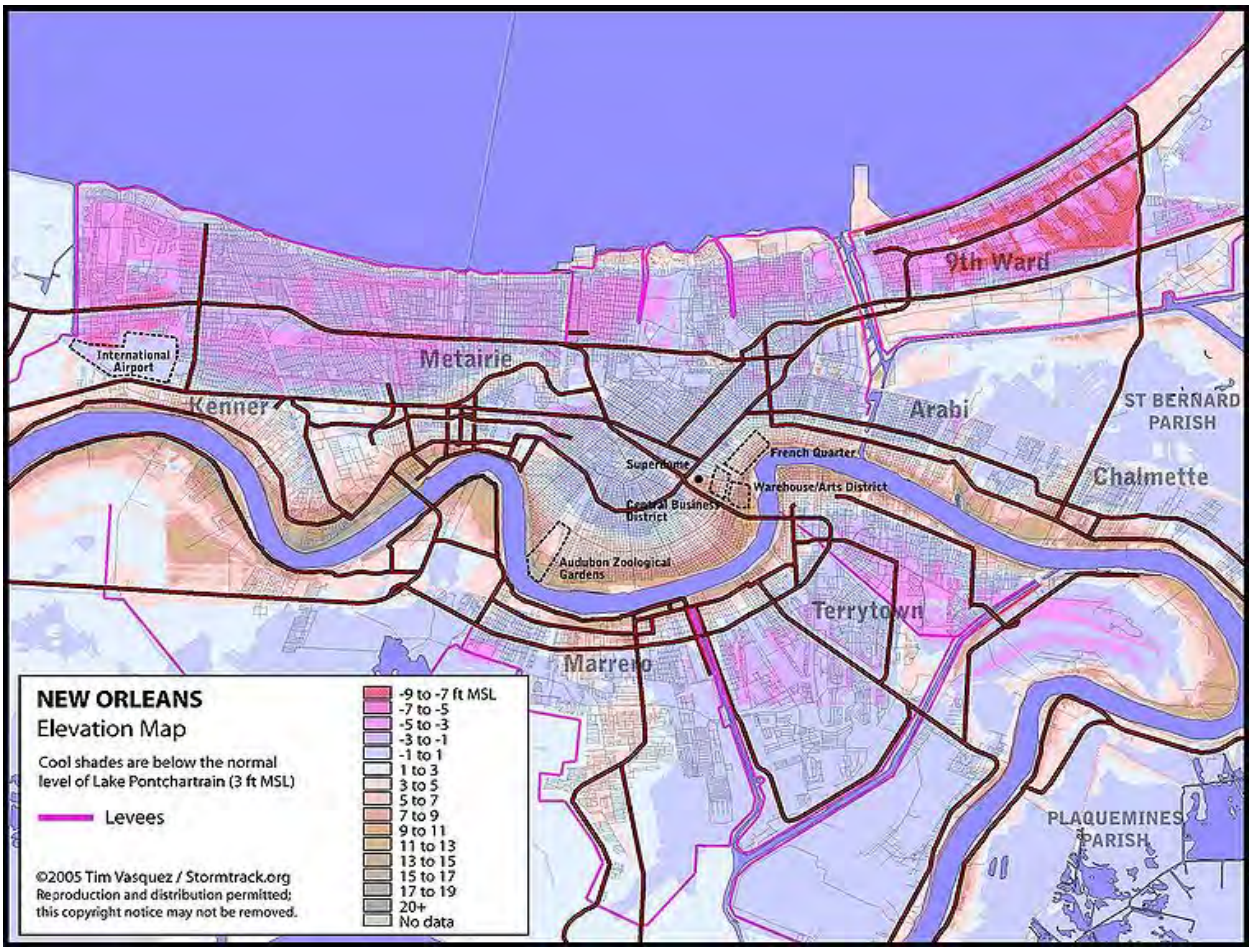
The greatest urban calamity in American history resulted from the passage of Katrina on shore. Let us set the stage for what happened to New Orleans by first reminding you of how this city appears from space (also, refer to the write-up on page 4-2). Below the space image are three maps - one giving a regional geographic perspective; the second, the central parts of the city;

then a map of the Parishes (Counties) around New Orleans.





Because much of New Orleans is just below sea level, over decades a system of levees had been built up. These levees appear as maroon lines in the map below:



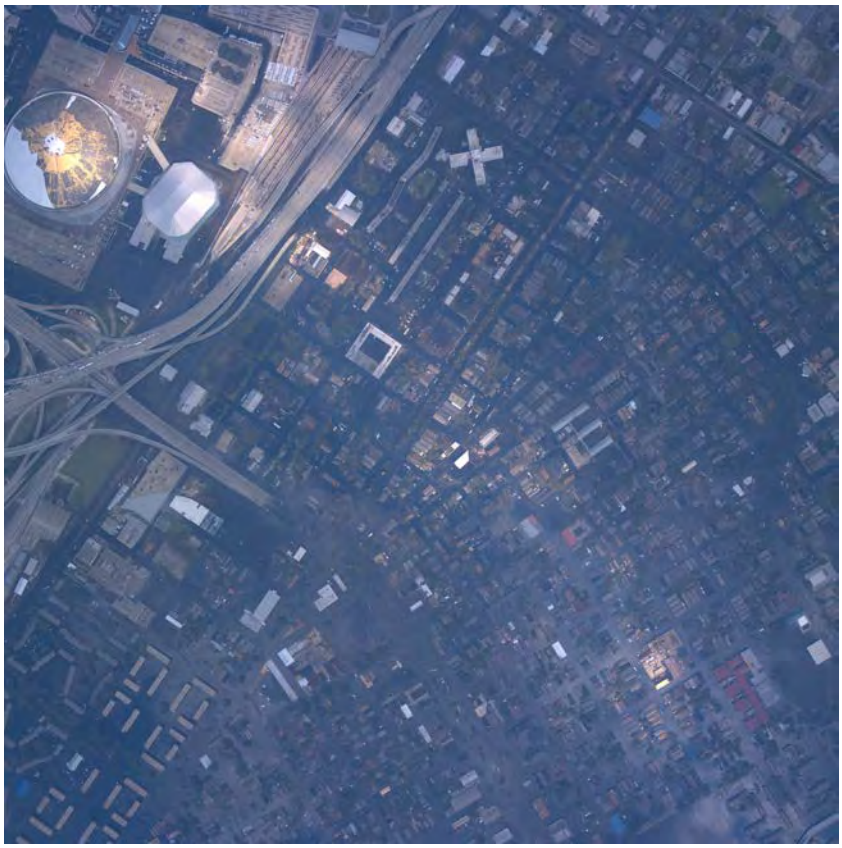
The writer (NMS) remembers one special feature of New Orleans from his first visit in 1952: Because of its sea level relation, and the water table in the region being so near the surface, most of the graves in New Orleans cemeteries cannot be below ground; families owned small masoleums in which individual members were entombed:

At first, the remaining residents of New Orleans - called "The Big Easy" or the "Crescent City" by the tourist interests - believed they had "dodged the bullet". Word came soon that Gulfport, Biloxi, and other Gulf towns had experienced heavy wind and water damage. There the storm surge reached 8.7 meters (29 ft), largest on record from an American hurricane.

The storm surge also affected Lake Pontchartrain. The levees coming off the lake (actually a lowlands connected to, and drowned by, the Gulf waters) started to break, spilling the top 2/3rds of a meter (two feet) of its water into New Orleans (remember, much of the city is in a topographic "bowl" just below sealevel), and eventually submerging up to 80% of its area to depths of 1 to 6 meters (3 to perhaps as much as 19 feet). These pictures show some aspects of this levee collapse, which occurred in three ways: 1. topping the structure, 2. breaching it with breaks, and 3. undermining its base.

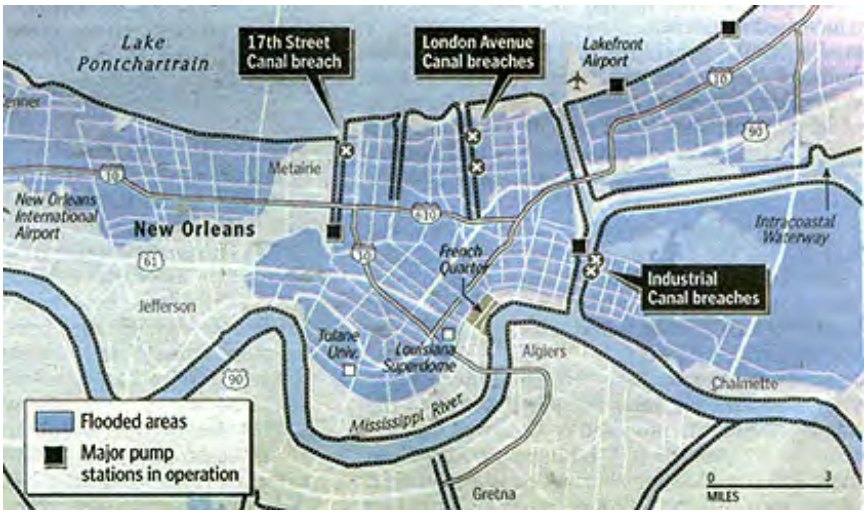
Hundreds of photos of flooding in and around New Orleans have been published:

This aerial photo indicates the extent of flooding near the center of the city; note roof partly off the Superdome.:



This satellite image indicates how widespread the flooding was. The red arrows points to areas especially hard hit.

This map shows the extent of maximum flooding and the dikes that were breached (largely a failure of inadequate levee construction which eroded away its material as storm surge effects pressed against the structures). Note that neither the dikes along Lake Pontchartrain or the Mississippi were breached directly, only those along the canal system within the city.



Courtesy: Washington Post.

The next two higher resolution (10 meters) images were obtained from NASA's EO-1 satellite. In both, flooded areas are indicated by black tones in the streets. The first image shows that the somewhat higher land in central (downtown) New Orleans remained dry - never flooded. This includes the famed French Quarter, which can trace its origin to the founding of New Orleans in 1718 when French pioneers built the first settlement on the highest ground adjacent to the Mississippi River (even in its early decades, the citizens experienced enough flooding to erect protective earthen levees). The second shows how north-central New Orleans was flooded by failure of the east 17th Street Canal dike but the west dike held up keeping that part of New Orleans from flooding.

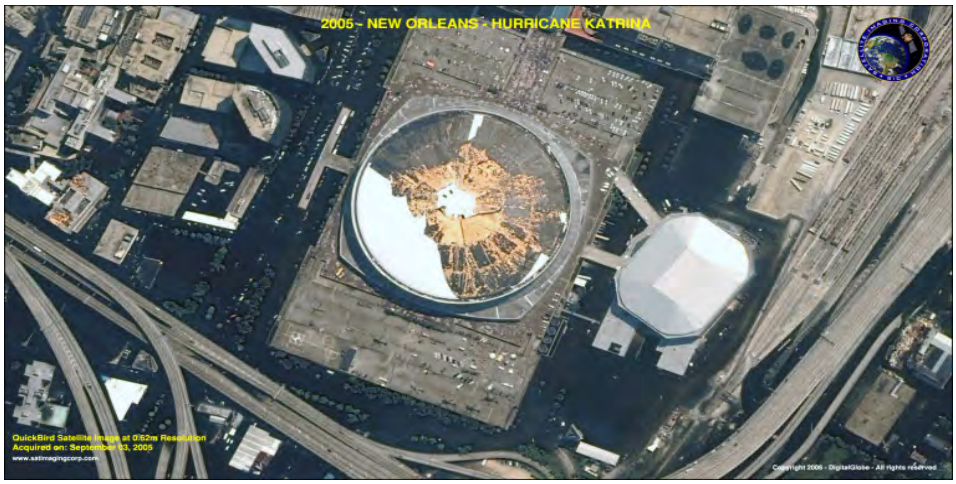
Experts had known for years that a powerful hurricane could breach one or more of the (obviously too low and outdated) levees raised against Lake Pontchartrain and the Mississippi River. Their recommendations to upgrade these levels were ignored (in part for unwillingness to commit funding). **What happened in the case of Katrina was that counterclockwise winds on the west side of the hurricane's eye (which passed just to the east) pushed water from Lake Ponchartrain against the levees from the northeast as a storm surge that topped many of the levees**. Water from Lake Borgne, which lies in a swampy area, was particularly effective in inundating the Lower 9th Ward.

The Lower 9th Ward suffered the greatest damage, as shown in this aerial view and a ground photo. The Lower 9th has only been partially reoccupied since Katrina.

This cross-section explains why the French Quarter (St. Louis Cathedral, on left) and adjacent areas did not flood - they were five feet or more above sealevel. The French Quarter is where it is because the original settlers chose the high ground :

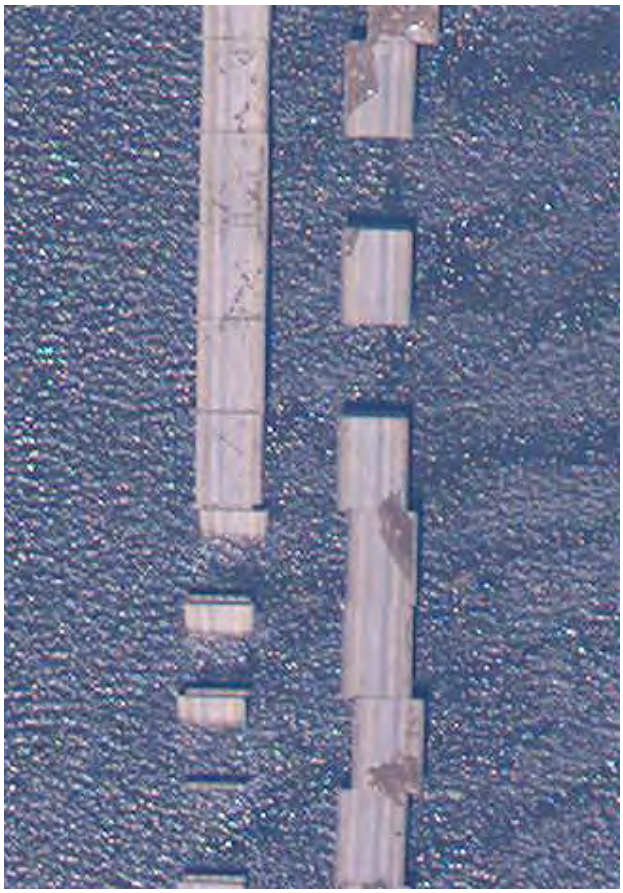
After a "mandatory evacuation" order was given on Sunday, August 28, a quarter of a million people left town. But more than 130,000 citizens of New Orleans stayed behind, some choosing not to evacuate but most without the means (such as autos) to escape on short notice. As waters rose, most of these were driven from their homes, or had to be rescued by air, or drowned. Over much of the first week, the search and rescue operations proved woefully inadequate (the "blame game" fingers all levels - federal, state, and local - for poor execution of existing plans. At the federal level, FEMA (Federal Emergency Management Agency) shouldered much of the criticism, particularly against its top management which included key decision makers with little emergency handling experience.

The famed Superdome in New Orleans was opened to shelter about 15000 people. That dome had its outer cover pealed off and some panels broken, causing roof leaks. The Superdome soon became uninhabitable and its refugees (and many at the Convention Center) were moved from New Orleans by bus. For several days, before National Guard and federal troops arrived to support the beleagured town police, looting, shooting, and other forms of anarchy ruled the passable streets. Here is the Superdome as seen by the Quickbird satellite and from the ground.



The New Orleans Convention Center also served as a refuge. Both large facilities failed as shelters, largely because of inadequate bedding, food, and toilet provisions. As help finally arrived, and some semblance of disaster management took hold, these two buildings were evacuated. In time more than 400000 citizens from the most affected areas on the Gulf were moved to Texas and Florida and some to many other states. Tragically, as recovery operations hit their stride, it was discovered that several hospitals and nursing homes contained scores of dead who were overlooked during the critical evacuation hours.

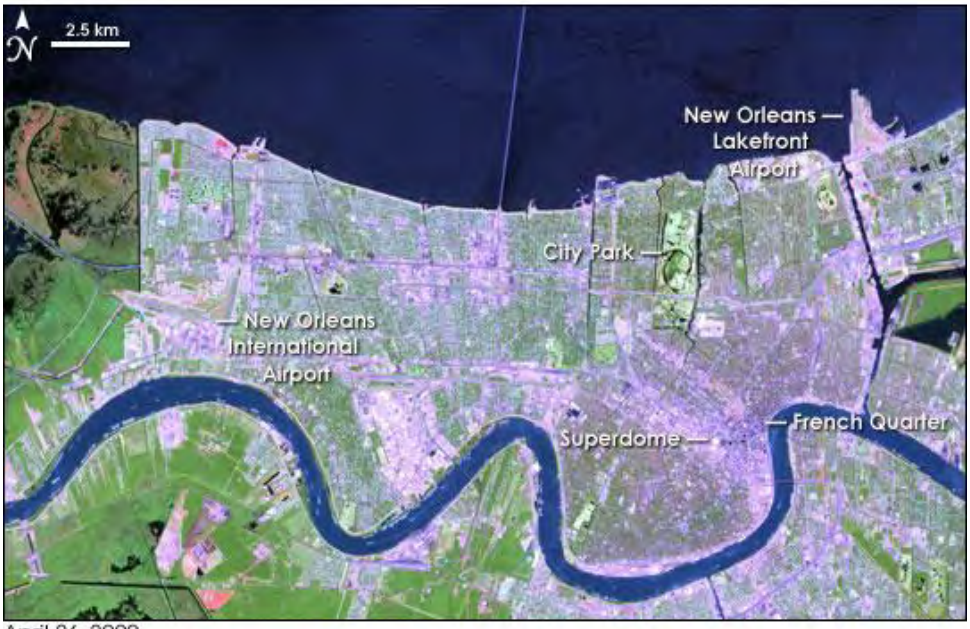
Surprisingly, the Lake Pontchartrain Causeway - longest of its kind in the world - sustained little damage: a tribute to its excellent engineering (in the open waters of the lake the storm surge effect was reduced). But the shorter (12 km; 8 miles) Twin Spans Bridge, part of Interstate 10 from Slidell, LA to New Orleans, being in a more confined part of the Lake, was heavily damaged with sunken or separated span blocks, and will need nearly complete rebuilding.



To the east in Mississippi, the double span bridge on Highway 90 between Bay St. Louis experienced a west-leaning downdrop of most individual segments of the concrete bridgeway:

Landsat-7 checks in with its own partial scene view of flood and preflood imagery of New Orleans.

August 30, 2005

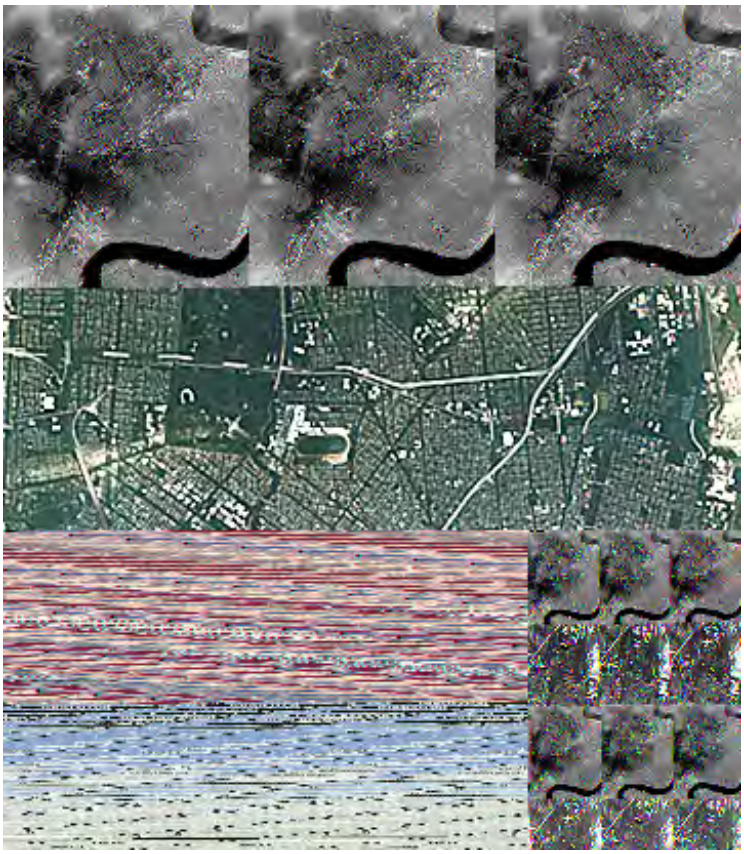


April 26, 2000

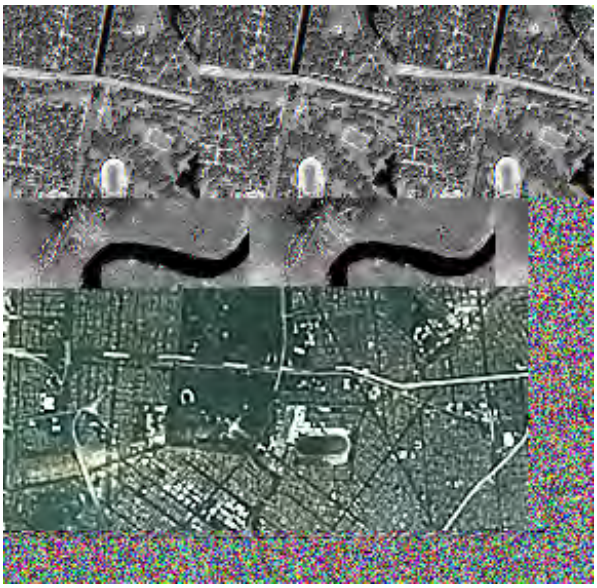
A SPOT-2 false color image taken at 20 meters on August 31 shows that most of the flooding came from failure of levees associated with Lake Pontchartrain. Areas close to the Mississippi River were largely spared.

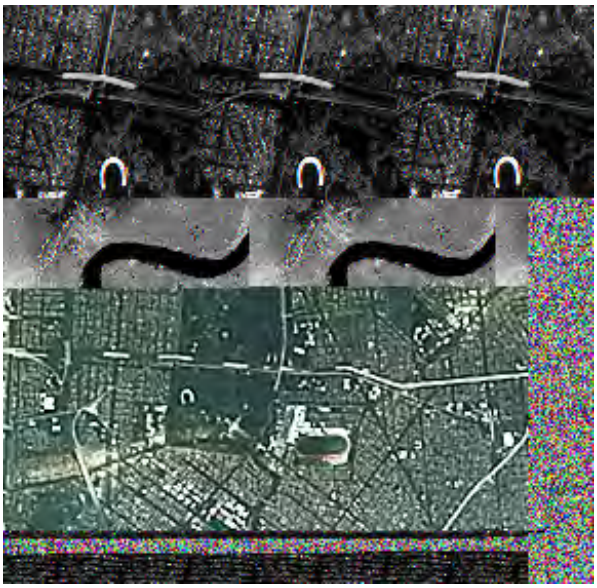
Digital Globe's Quickbird satellite obtained the first high resolution views of flooding in New Orleans. The top image is one taken in March of 2004, the bottom is the same area seen on August 31 right after the hurricane, with flooded ground appearing black.

The Orbimage satellite produced several black and white images that really show the boundary between flooded residential areas and the downtown business district and industrial areas along the Mississippi River:



Orbimage-3 obtained a view of Tulane University before the flood, and again after the campus was almost totally inundated.





New Orleans could not start rebuilding until the floodwaters are drained off, after which electricity, sewerage, and water systems have gradually been brought back up. This lovely city has suffered a catastrophe but the townsfolk and the federal government have vowed to restore most of the city. New and better floodwalls will need to be built and adjustments to the Mississippi River channel system will be designed to restore some of the protective delta wetlands.

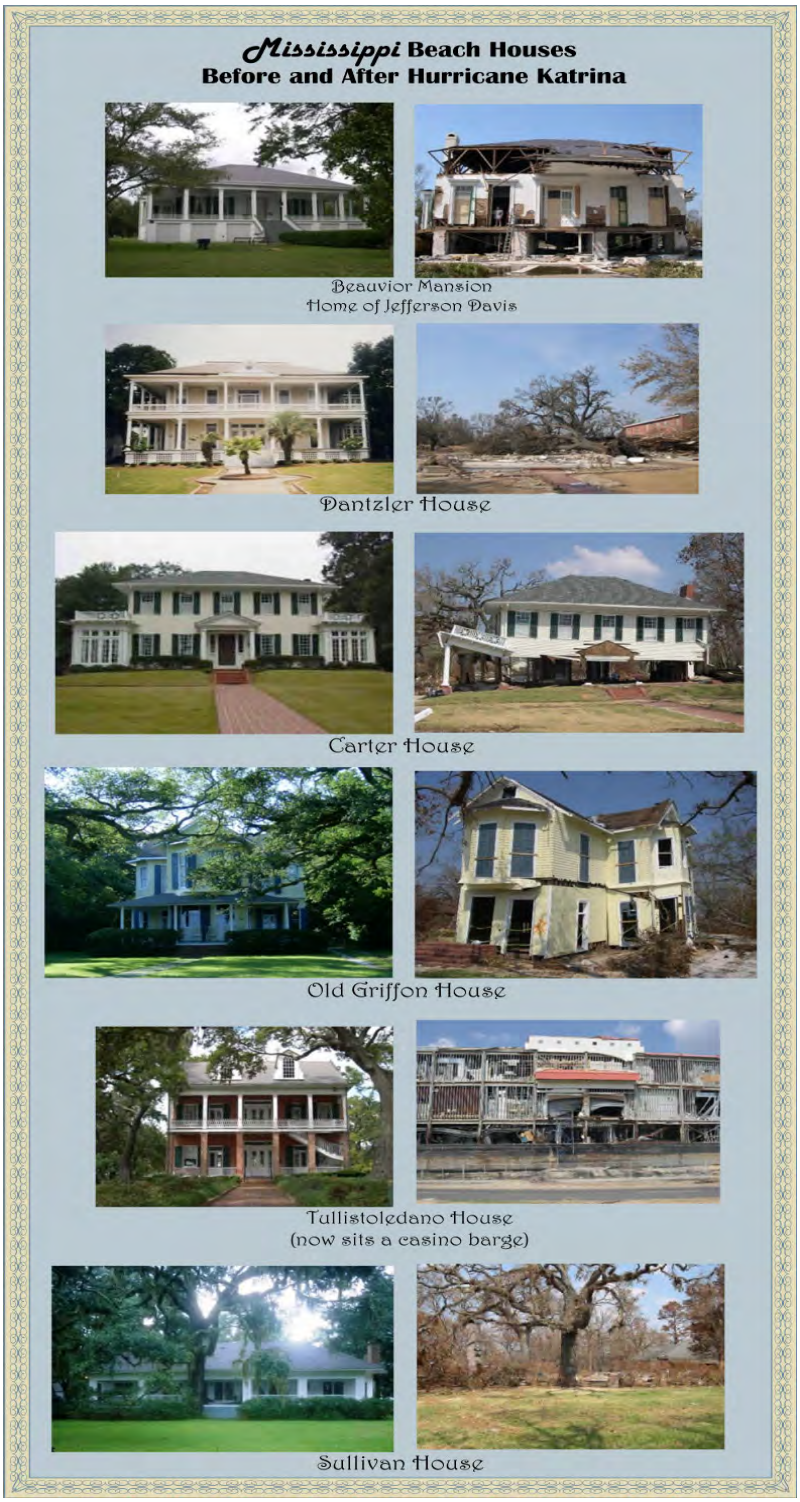
Now lets take a tour of some of the damage along the Mississippi coast. As one would expect, the delta country in Louisiana sustained huge damage. Here is an aerial view of destruction in Plaquemines Parish south of New Orleans:

Eastward, Mississippi state experienced severe damage that extended tens of miles inland. Here is a Quickbird image of part of Biloxi.

August 31, 2005

April 12, 2005

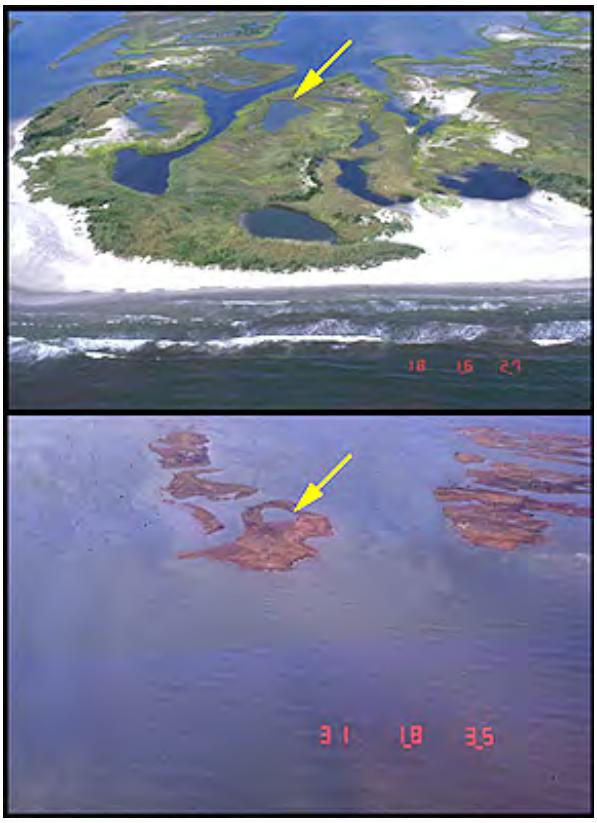
Among the casualties of Katrina are many of the antebellum (pre-Civil War [1861]) homes in Biloxi and other Gulf cities.



Among Gulf cities receiving extensive damage was Gulfport, in Mississippi:

The rest are ground and aerial views along the Louisiana and Mississippi coastal areas; read the captions for details.

The delta wetlands at the mouth of the Mississippi and many offshore islands were either washed away or flooded. For example, Chandaleur Island, off the Mississippi state coast, about 160 kilometers (100 miles) east of New Orleans experience this near total destruction:



In addition, some land segments at the Mississippi's mouth had homes on them, many to house the families of workers on the offshore oil drilling platforms (a large number of these were severely damaged). This curved land within the distributaries saw complete destruction of this pre-hurricane housing:

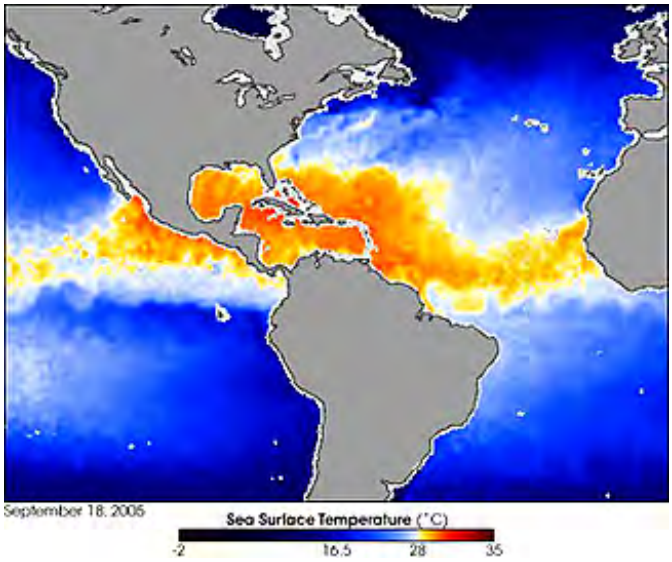
As of September 20, the area of flooding in New Orleans was, through a massive pumping effort, reduced from 80% to 40%. By month's end, almost all water was gone - well ahead of schedule - only to experience some return in the worst hit 9th Ward when Hurricane Rita struck (see below).

In retrospect, these facts or strong surmises can be recorded here: 1) Katrina is now called the worst natural disaster to hit North America in terms of area affected and dollars needed to recover; original estimates of insured damage at $25 billion proved quite low, new estimates range beyond $70 billion (compared to $43 billion for an inflation-adjusted Hurricane Andrew) and will probably rise another $120 billion when other sources of recovery funding (federal; private)are factored in; experts in disaster assessment conjecture a real total cost as nearing $200 billion - enough to fund NASA for more than 10 years - when lost job income, business failures, and industrial shutdowns are included; 2) up to a half million refugees were spread over many states following near-complete evacuation of New Orleans and other areas; unfortunately, for the desire for full recovery, many displaced citizens as of early 2007 are defiantly contemplating not to return; 3) full reopening New Orleans took months just to get the situation

in control before years of permanent rebuilding; however, about 180,000 were told they can return by early October, 2006 to those areas that avoided flooding; 4) authorities knew that a Category 4 or 5 hurricane would likely flood New Orleans, yet did almost nothing to improve its protection with better levees; 5) the Search and Rescue phase and evacuation/refugee care efforts were slow, especially since the pre-hurricane studies had predicted the problems so that a quick response should have been in place; the Afro-American community bore the brunt of this failure; 6) the disruption of offshore Gulf oil production (more than 140 platforms damaged and 12 destroyed by Katrina and the later Hurricane Rita [see below]) and the shutdown of the many refineries along this part of the Gulf Coast have accounted for part of a sharp increase in U.S. gasoline prices and a period of scarcity (mainly in states east of the Mississippi where most of the pipelines were directed)

Mayor Ray Nagin of New Orleans was the first to forecast a possible death toll rising to 10000. Mercifully, this death toll has proved to be much lower, with the full number in New Orleans not determined until all the water is pumped out. As of November 15, 2005 the four state toll stood at around 1900, far more than the 246 who died when hurricane Camille struck the same area in 1969 (although 113 of those died from floods in Virginia as the hurricane tracked north). New Orleans and surrounding areas of Louisiana have counted more than 1700 fatalities; Mississippi suffered 226 dead, Florida 12, and Alabama 2. This is thus the third most deadly hurricane in U.S. history (first: Galveston, TX in 1900, more than 6000 dead; second San Felipe-Okeechobee FL in 1928, 1826 dead in U.S. and 312 in Puerto Rico).

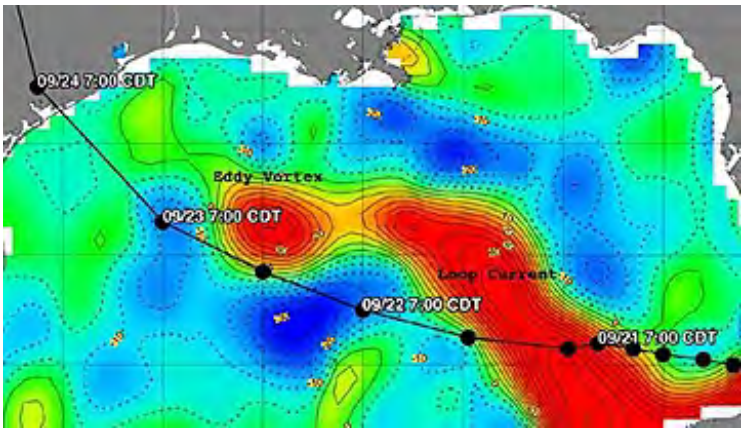
As if Katrina weren't enough, another Category 5 hurricane developed in mid-September. Hurricane Rita began in the western Atlantic, moved past the Bahamas, passed over the Florida Keys, and entered the warm waters of the Gulf of Mexico where it strengthened from Category 1 to 5. It is now rated as the third most powerful Atlantic hurricane on record. Even in September, the waters of the Atlantic were unusually warm.



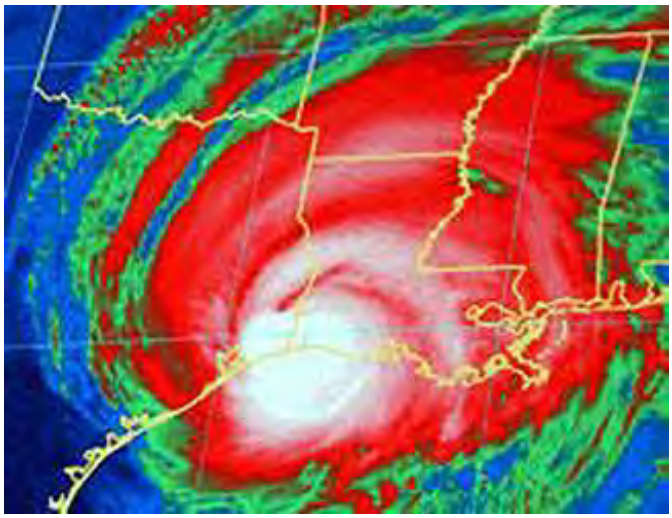
As Rita moved into the central Gulf of Mexico, its huge size (more that 400 miles in diameter), its winds reaching 175 mph, and its low pressure (884 millibars) indicated it could be a rival of Katrina or even worse.

The entire central Gulf seemed threatened. Mandatory evacuations from north of Corpus Christi, Galveston and Houston, Beaumont, and western Louisiana were ordered. Three million people went northward. Officials were determined not to have a repeat of the Katrina fiasco. Massive traffic jams occurred in east Texas. Concern for much more damage to the Gulf oil refineries sent prices rising once more.

But as Rita swung northward, one natural condition favored less of a potential calamity. The hurricane moved over pockets of cooler water, as shown in this Aqua AMSRE image.



This loss of thermal energy weakened Rita so, while still large and ferocious, when it came on shore in the early morning of September 25 about at the Texas-Louisiana border its winds had dropped below 135 mph.



Still, this was a destructive hurricane that caused widespread damage. The next two images show a coastline just inside of Louisiana and the flooding of Cameron, LA. Extensive damage occurred in Beaumont and Port Arthur, TX, in Jefferson Parish, LA, and other small cities and towns. Parts of New Orleans were reflooded when wind-driven storm surges and rainfall caused several levees to be topped. But, while damage overall may exceed $8 billion, loss of life was minimal (3 direct; 24 turning an evacuation bus fire) and the gasoline refineries were largely spared.

One longer term destructive effect from Katrina and Rita might not come to mind until one is prompted to think about the topic of "fallen trees". An inventory of downed trees in the Gulf Coast caused by these two hurricanes disclosed that about 320 million trees were affected. The areas of major damage appear in this map:

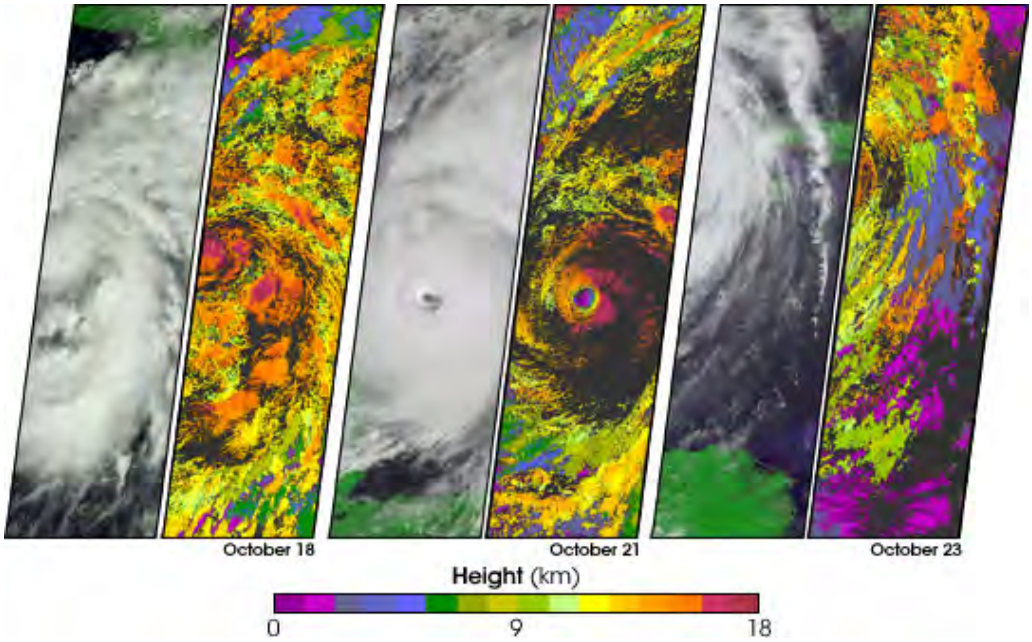
This aerial photo illustrates the haphazard way many of these trees were downed:

Most of the trees affected were pines and some hardwoods. These trees have been decaying rapidly, putting CO_2 into the atmosphere (equal to about all the carbon dioxide removed from air over the U.S. by normal growth processes). For the short term, trees that have taken root since the hurricanes are of species not favorable to the desired ecology. It will take time and effort to restore the balance by reintroduction of pines and other evergreens.

On October 15 a third major hurricane started to organize. Of the three discussed here, this one in some ways was the most bizarre. It started in the Carribean as a tropical depression that in just a few days grew into a Category 5 hurricane. In fact, for a brief time, Wilma became the most powerful hurricane since modern instrumentation came into use to determine strength as its barometric pressure dropped to 884 millibars (26.04 inches). For 2+ days, it struck the Mexican Yucatan and sat over Cozumel and Cancun with almost no forward motion. Then as predicted, upper atmospheric steering currents caused an abrupt 90° eastward turn that headed it towards southern Florida. It completely crossed that state, wreaking havoc as a Category 3, then moved rapidly up the western Atlantic paralleling the American coast.

The track of Wilma appears below.

This set of 6 image panels shows three visible images of Wilma on different dates and for each a colored coded image indicating cloud heights.



The insured damage from Wilma exceeded $10 billion dollars. Considerable damage was first produced in Jamaica, then much flooding and wind damage in the Yucatan, followed by severe flooding in Cuba. Wilma entered Florida on October 24 as a Category 3 and remained at that level when exiting the east coast around Fort Lauderdale. Structural damage was not that great - although quite noticeable - but thousands of trees were uprooted throughout southern Florida and for the first two days up to 6 million people were without power. Here are two views of damage; location is given in the captions.

Since Wilma tropical storms Alpha (the National Weather Service went through all its named storms for the year and elected to use the Greek alphabet for subsequent storms) has come and gone with only rainfall effects. Then on October 27 tropical depression Beta become a Category 1 hurricane that struck Central America. In mid-November, the third such storm,

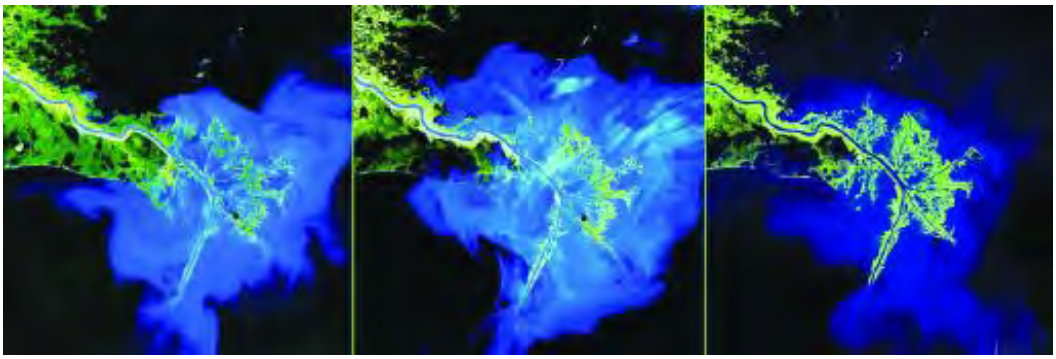
Gamma, grew in the western hemisphere and moved onto Honduras with heavy rains. To everyone's surprise, on December 30th Tropical Storm Zeta formed in the eastern Atlantic.

As the stories of Katrina, Rita and Wilma unfold, some very difficult questions need to be answered. Most stem from the Katrina fiasco. Along the Gulf Coast, one insurance institute estimate places the number of homes that may not be salvagable and must be torn down as approaching 150000 (mostly middle class/poor in New Orleans) - a staggering number which may diminish if/when experts come up with innovative solutions to restore water-soaked lower levels and disinfect disease-carrying mud and waste. Some have even raised the trenchant question of whether the present location of New Orleans should be abandoned. The city has continued to sink (ground water withdrawal plus compaction without replensishment of soil from river flooding - New Orleans does have periodic small floods from rainstorms) even as sealevels are rising nearby.

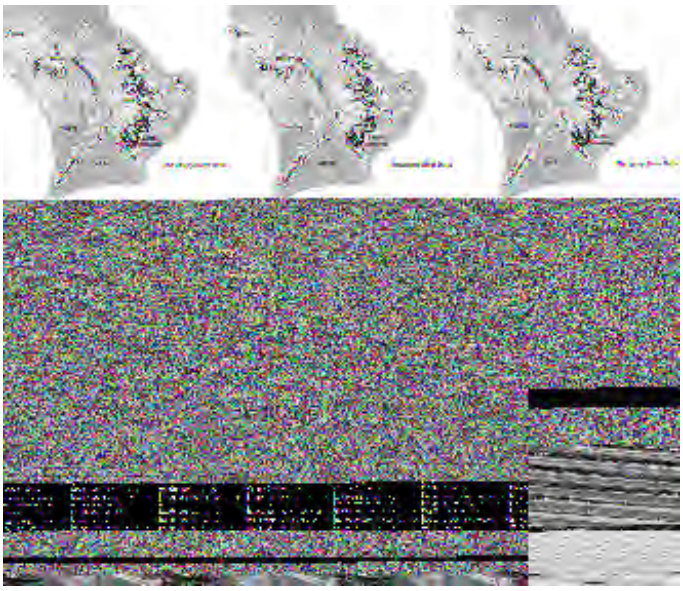
What is needed to improve New Orlean's chances from inevitable future hurricanes are two obvious lines of defense: 1) higher and more secure levees, with possibly more canals (the present ones go back to the 1840s), and other engineering adjustments such as pumping stations that have their own protected power sources; and 2) a potent program of land reclamation that produces more barrier islands, delta extension, water-attracting swamps, marshes, and bayous; this would also entail redirecting the Mississippi River without compromising its key role in shipping. Another important modification: make the several gasoline and other petroleum product refineries much more immune to storm damage, and further protect the pipeline systems that transport these energy sources to the midwest and east.

For decades now, scientists and engineers have been warning about the potential for catastrophe along the Gulf Coast. Plans to give back to the Mississippi its natural ability to deposit sediment along the delta to build its shoreline into the Gulf and extend the areas of wetland, which together would provide more protection for New Orleans to the north, have been proposed but not acted upon. The history of the delta is well known. Its principal distributaries have shifted significantly over the last 10 million years, as seen in this map:

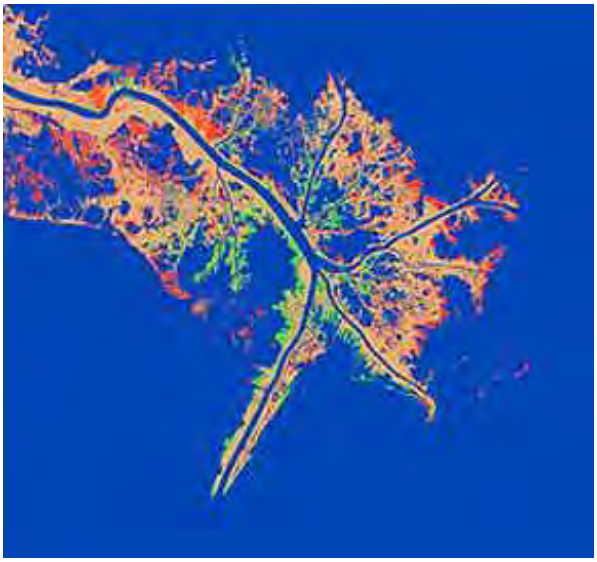
Enough information from ground and aerial surveys, and satellite observations, have allowed maps of the present delta distributary complex to be produced. Look closely - there are discernible changes (see caption for dates):



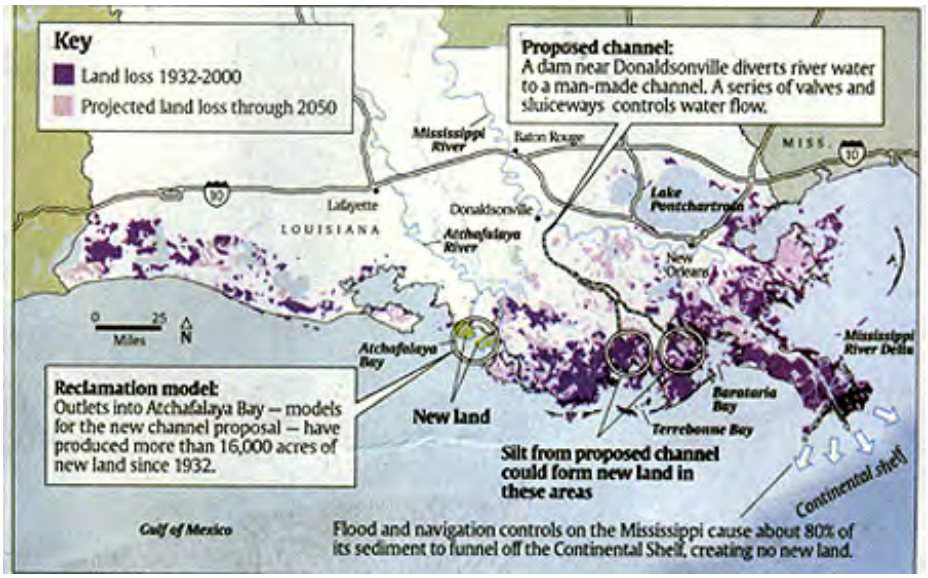
Satellite imagery has been used to determine where and how much land was removed by storm surge/tidal erosion caused directly by Katrina. This map shows this land in black:



The Bird's Foot distributaries are constantly being modified. Some of this is natural; some is caused by the confinement of the Mississippi upstream by levees which affect the river's ability to spread its sediment load in its normal way; some results from hurricane action. This image of the delta, constructed from multiyear satellite observations, shows land eroded in red and deposited in green:



All told, more than 600,000 acres of wetlands have been lost in recent decades. Erosion rates as high as 40 acres in a day have been recorded. This map shows the extent of this loss and possible ways of eventually adding rather than eroding land:



From David Merrill, USAToday

Engineering the Mississippi flow patterns to optimize wetlands and shoreline protection is one of the main solutions. Building better dikes, levees, canals, and pumping stations are also important in protecting lowlands. It has long been known that New Orleans is highly susceptible to flooding. Proposals to provide major new components of a protection system (for a Category 3 hurricane; a Category 5 was considered too unlikely and thus too expensive) have been made for decades to state officials and more importantly to Congress and Federal agencies. Although some monies and authorization were provided, these fell far short, so that only token work to improve some of the facilities has been done.

The U.S. Army Corps of Engineers has now spent about $20 billion on shoring up the levee system in New Orleans. It is expected to be completed by 2011. This system should withstand a category 4 hurricane, but if the eye passes just to the east, flooding from Lake Ponchartrain remains a threat.

In hindsight, the crux of the problem was failure to recognize the "Achilles heel" of the New

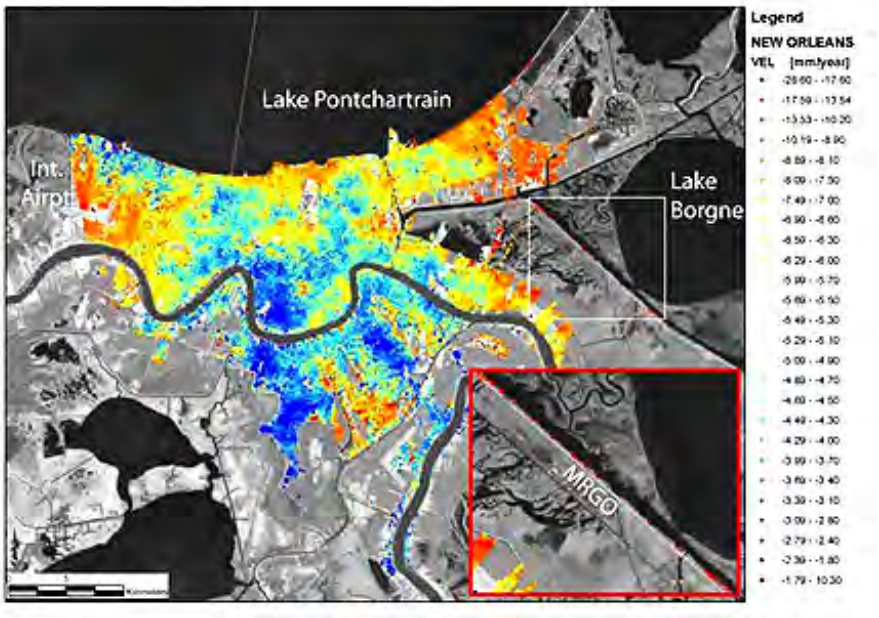
Orleans dike system. The dikes were anchored by steel pilings driven to a depth of 33 m (20 ft). At that depth the local soil was a form of organic debris plus mud (Peat) which was very weak and porous. Water driven by the storm surges that seeped to this depth eroded through the peat, sapping (undercutting) the levees above so that the concrete floodwall within the levees were made unstable and then collapsed. Particularly vulnerable was the Industrial Canal, affected by the Lake Borgne storm surge. Warnings given about the peat layers as potentially unstable had been given but went unheeded.

The blame for the Katrina disaster therefore can be spread to many; most of the failure is tied to political decision-making (rejection or minimalization of critical funding) that turned a blind-eye to the dangers forecast by the engineering/science communities. These shortcomings extend over several administrations. But, the latest example is typical: in 2005 Senator Mary Landrieu of Louisiana has pushed for $98 million dollars for short term action to improve levees by the U.S. Army Corps of Engineers, the Bush Administration countered with $22 million, and Congress settled on $42 million. Little of this has been actually allocated. Compare these numbers with sound estimates that the total dollars needed to really protect New Orleans and other areas in that part of the Gulf would probably be about $9 billion. Now, reflect on the huge price tag of as much as $150 billion to recover from Katrina and Rita and recall the phrase "Pennywise and Poundfoolish".

Perhaps the best example of "I told you so" is to be found in the October 2001 Scientific American article by Mark Fischetti entitled "Drowning New Orleans" in which his scenario proved to be remarkably prescient when compared with the actuality of Katrina.

In preparation for the future, the Corps of Engineers has strengthened and raised the key dikes throughout New Orleans. Other "preventitives" are being emplaced or considered. But so far nothing significant has been done to prepare the city for a Category 5 hurricane. A much improved evacuation plan is being implemented.

Unfortunately, two critical conditions are getting worse. Sealevel continues to rise (some protection can be provided by restoring wetlands). But, studies in 2005-06 have revealed that parts of New Orleans and surroundings are *sinking* at rates from 6 to 29 millimeters per year. Exact causes are yet to be discovered. Here is a map of relative sinking published in 2006.



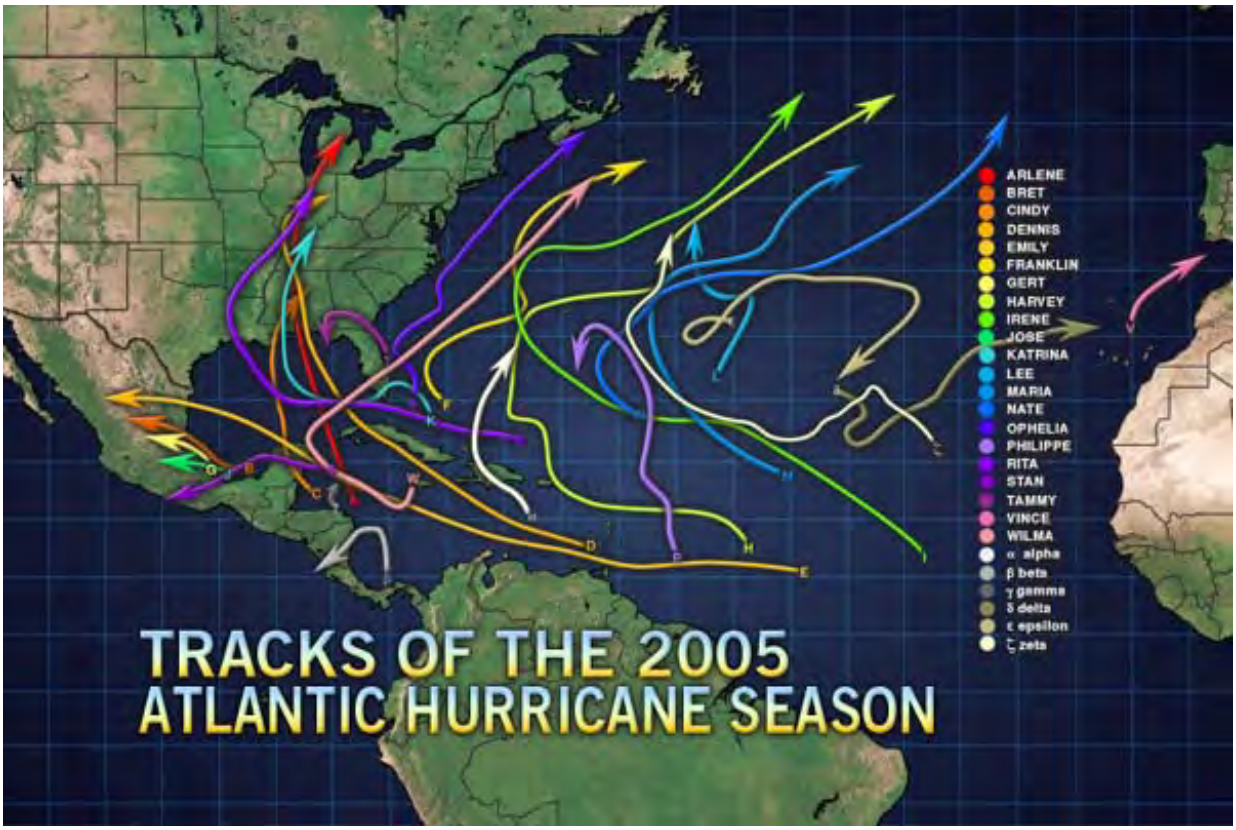
On a more general note, a rise in sealevel of 3 meters (10 ft) during the 21st Century (global warming due to a combination of natural interglacial temperature increase and gas emission hothouse effects) would notably modify the Gulf Coast, drowning many cities (if these cannot be protected) as shown in this map:

As this paragraph is written on August 28, 2010 - the 5th anniversary of Katrina - these general observations seem appropriate: New Orleans has "bounced back" (largely recovered). Tourism is at new levels of popularity. The economy is stable. As a blessing from the attention Katrina forced on the city, the problem it had before with subpar schools has been turned around through a highly successful charter school program that has raised student performance levels above the national average. The city's morale was given a big boost by the victory of the New Orleans Saints (football) in the Superbowl. On the negative side, the city population remains about 100,000 below pre-Katrina days. And the citizens of this Gulf Coast region continue to look with distaste and disfavor on the shortcomings of both federal and state government responses during the first days and weeks following the hurricane.

Looking back at the Katrina disaster, let's be grateful that worst case scenarios were somewhat WRONG. Lets hope that applies to the future. Let's put faith in Americans' ingenuity to restore much of the water-logged properties in New Orleans. Above all, that city must be renovated beyond its previous qualities. What other city east of the Mississippi (besides Charleston, SC) retains such a modern remnant of a glorious past - symbolized by the ever popular French Quarter, which deserves future visitors. Bring back Jazz to its native home!

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.


ARLENE
BRET
CINDY
DENNIS
EMILY
FRANKLIN
GERT
HARVEY
IRENE
JOSE
KATRINA
LEE
MARIA
NATE
OPHELIA
PHILIPPE
RITA
STAN
TAMMY
VINCE
WILMA
α alpha
β beta
γ gamma
δ delta
ε epsilon
ζ zeta
TRACKS OF THE 2005
ATLANTIC HURRICANE SEASON

Grace
Claudette
Danny
Henri
Ana
Fred
Ida
Erika
Bill
TRACKS OF THE 2009
ATLANTIC HURRICANE SEASON

This page deals with two major episodes in the 1990s that made weather history. The first was Hurricane Andrew in 1992 which struck land below Miami; before it dissipated to the north, it caused $30 billion dollars in damages - the costliest natural disaster in American history. The second was the infamous "Storm of the Century" in March 1993 which buried the northeast U.S. under a massive snowfall. Satellite images give some idea of the intensity of both these events. Hurricanes are by far the most costly (in terms of scope of damage) of storms. Tornadoes, also very destructive, can sometimes be captured (actually, the unstable air mass that spawns them) in satellite imagery and damage assessment made through space imagery after the storm has passed. Lightning strikes, another storm hazard, happen with high daily occurrences and are most frequent in central Africa and South America, as determined from data obtained by sensors dedicated to their detection.

Hurricanes & Tornadoes: Hurricane Andrew; 1993 Storm of the Century; Lightning Discharges

Satellites are now the primary tool used by meteorologists to track and forecast hurricane behavior. The GOES geostationary satellites are particularly adept at monitoring multiple storms visible simultaneously in a global hemisphere. Here is a September 3, 2008 GOES image acquired at 1:45 PM EDT that shows five tropical cyclones. Gustav and Ike both slammed into Gulf Coast states, causing billions of dollars in damage:

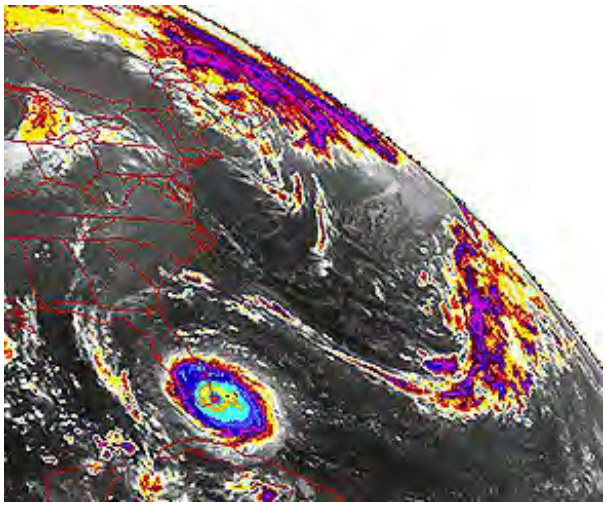
We look next at several special-interest stories dealing with hurricanes and violent storms that

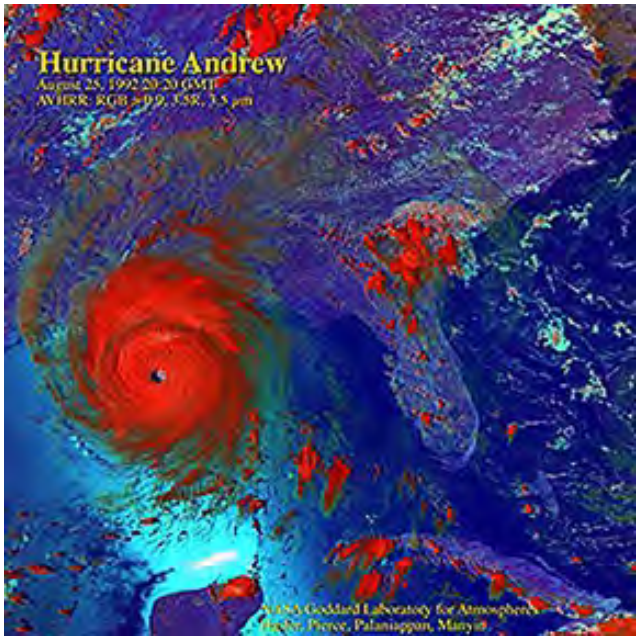
occurred either in the U.S. and/or the Caribbean and Central America. An excellent review of major hurricanes in these regions is found at this NOAA site. Also, you may wish to review the discussion of hurricanes near the bottom of page 14-1d. As a more general examination of cyclone/hurricane development, consider this online Web site: Earth Observatory Hurricanes.

In late August of 1992, Andrew, a category 5 hurricane, ripped through the Bahamas, slammed into southern Florida (its eye or center passed over Homestead Air Force Base), and moved into the Gulf of Mexico to hit Louisiana. Here is a montage of Andrew on different dates as seen with a metsat imager:

Although rather small in diameter, this intense storm was, up to that time, the most costly in U.S. history - estimates as high as $25 billion in damage - even though it took just 43 lives (a tribute to the early warning efficacy of Metsat monitoring). Winds in excess of 240 km/hr (149 mph) flattened entire housing developments. Hardest hit were homes in communities about 25 miles south of downtown Miami:

This devastating hurricane got a lot of attention by a variety of sensors from above and on the ground. Consider these images:

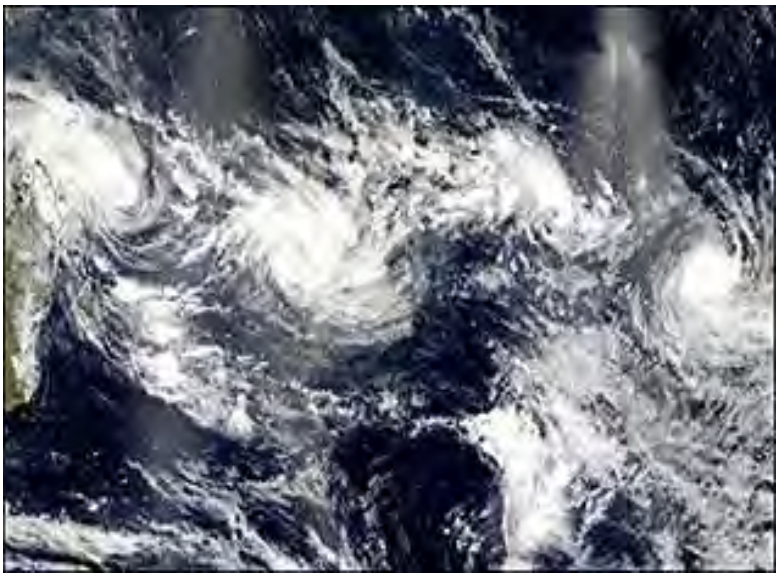




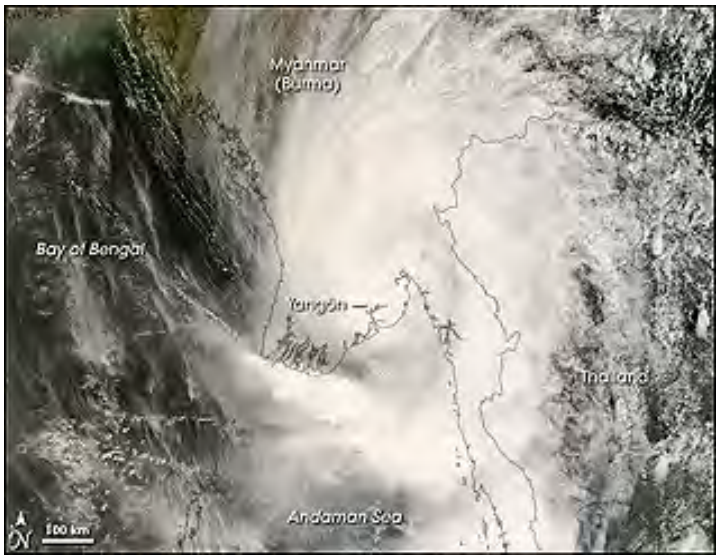
The top image from NOAA-7 shows the hurricane as it strikes land in Florida, with colorized tones representing higher intensities. In the middle is a GOES-7 full-Earth color view obtained on August 25, 1992, in which Andrew is seen approaching New Orleans. This perspective indicates scale, showing the hurricane, while powerful as organized, was still just another mass of clouds of no greater extent than some other storm systems. But, on close-up (bottom) in the AVHRR color version (RGB = 0.9; 1.5R; 3.5 µm), the perfection of the eye and the well-developed structure of this counterclockwise-spinning, low-pressure system are obvious.

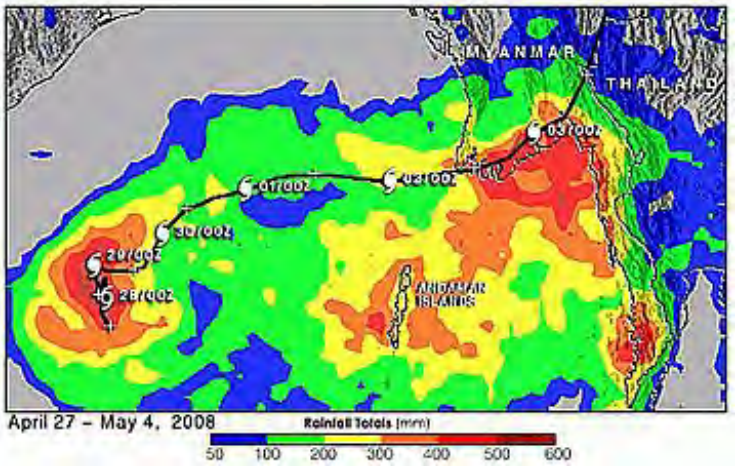
14-25: **What do you think happened to Andrew after it made its second landfall around the Mississippi Delta? ANSWER**

The Terra and Aqua programs are the subject of Section 16. These two satellites are not designed specifically as Metsats but they do gather considerable data pertinent to weather and climate. Here is a Terra MODIS image of 4 simultaneous hurricanes (called typhoons or cyclones in the eastern hemisphere) crossing the western Indian Ocean on February 12, 2003:

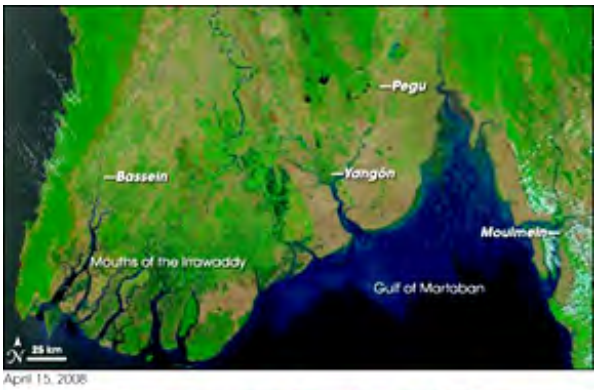


The most destructive cyclone in modern times was Cyclone Nargis, which passed over Myanmar (Burma) on May 2 and 3, 2008. Estimates of those killed range up to 128000. Large parts of villages and towns in the Irrawady Delta were obliterated by winds (the cyclone produced winds up to 160 mph, making it a category 4 storm) or washed away by storm surges up to 10 meters (33 ft) high at the coast. Here is a metsat image of Nargis, and a rainfall map of southern Asia for a week ending May 3:

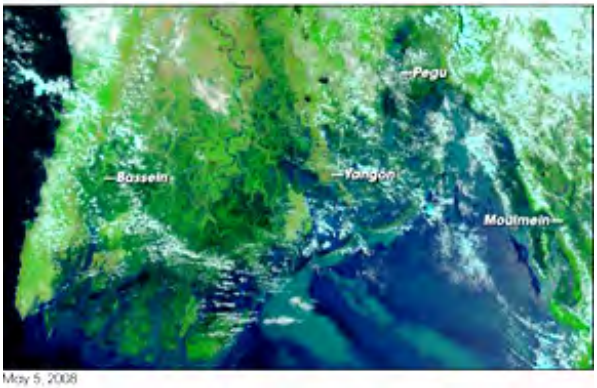




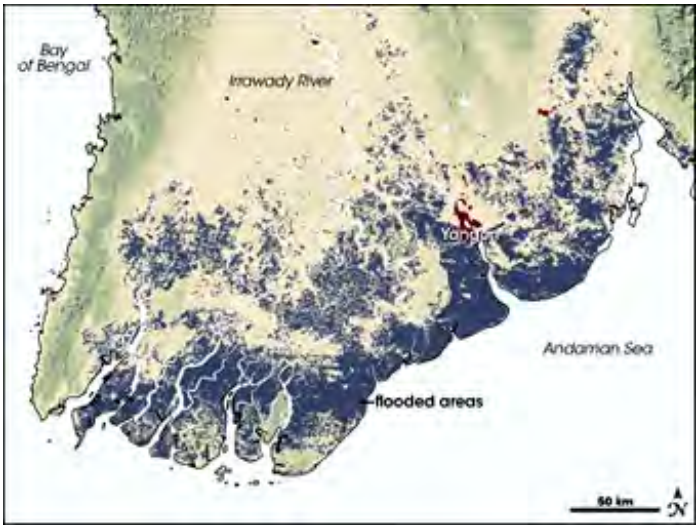
MODIS has produced these images of the floods in Myanmar; a preflood image is shown for comparison.



April 15, 2008



May 5, 2008



The capital of Myanmar, Yangon (used to be known as Rangoon) received extensive flooding:

May 7, 2008

November 27, 2007

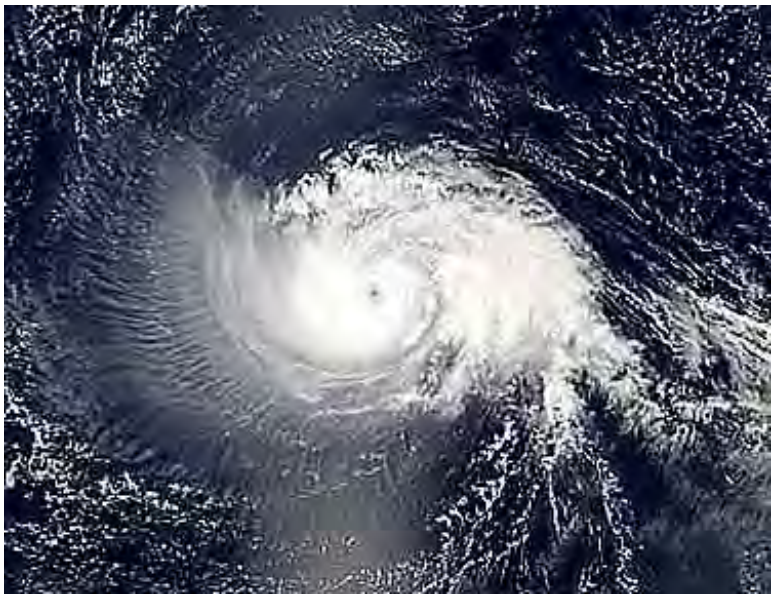
Entire villages were obliterated, as exemplified here:

May 7, 2008

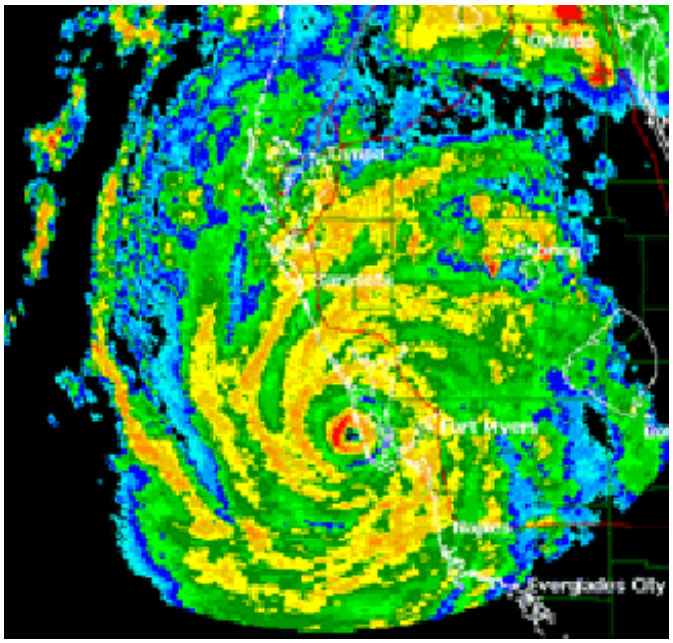
May 3, 2002

The Nargis disaster is a prime example of how satellite imagery can monitor a catastrophic event in near real time.

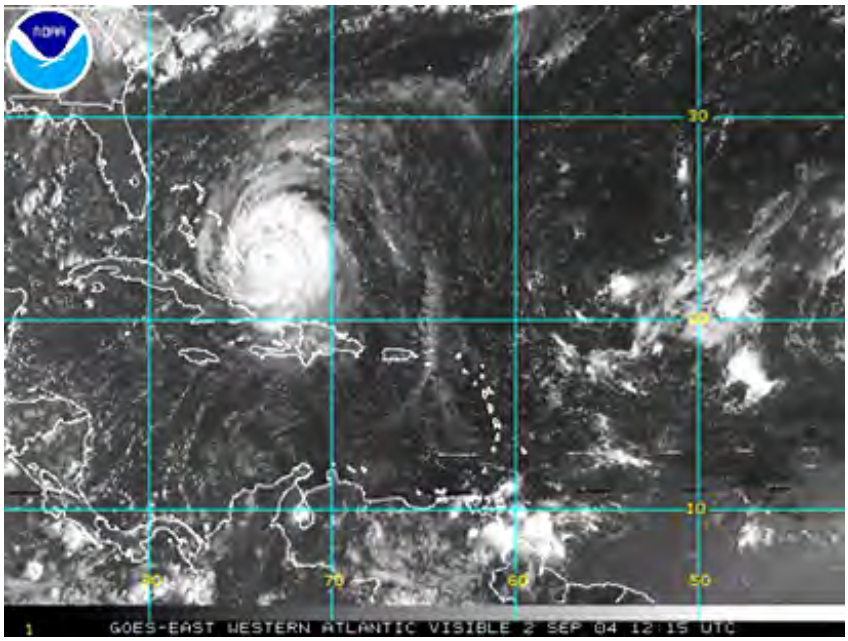
MODIS has proved a valuable adjunct to hurricane monitoring since its field of view can encompass an entire storm. In September of 2003, Hurricane Isabel reached a category 5 status (winds in excess of 146 mph) as it approached the U.S. Here's how it looked on September 9:



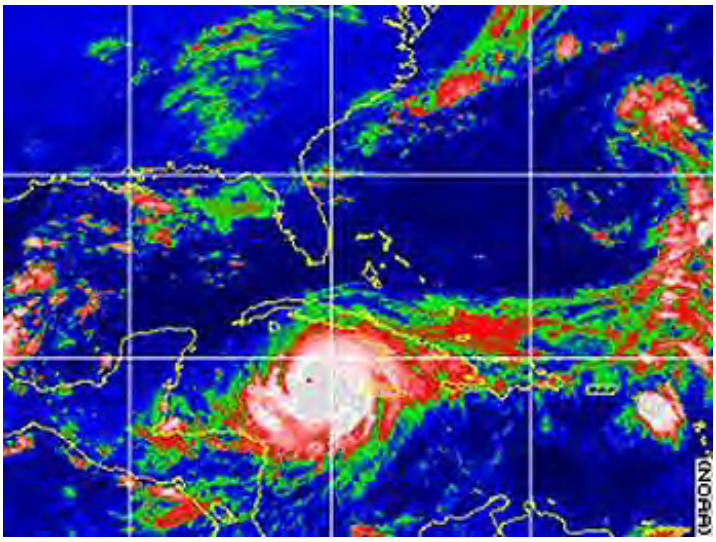
The hurricane season of 2004 was one of the most active and destructive in recent decades. First, in August, Hurricane Charley (Category 4) swept up to the west coast of Florida, going inland below Tampa and crossing the state:



Then, in early September, Hurricane Frances damaged Caribbean islands and the Bahamas before touching land near Fort Pierce on Florida's east coast, crossed the state, and then swung north into the Appalachians. Though its initial winds made it another Category 4 hurricane, these subsided as it struck the U.S. Most damage was done by extensive flooding. Here is a NOAA satellite view of Frances after its march across the Atlantic:

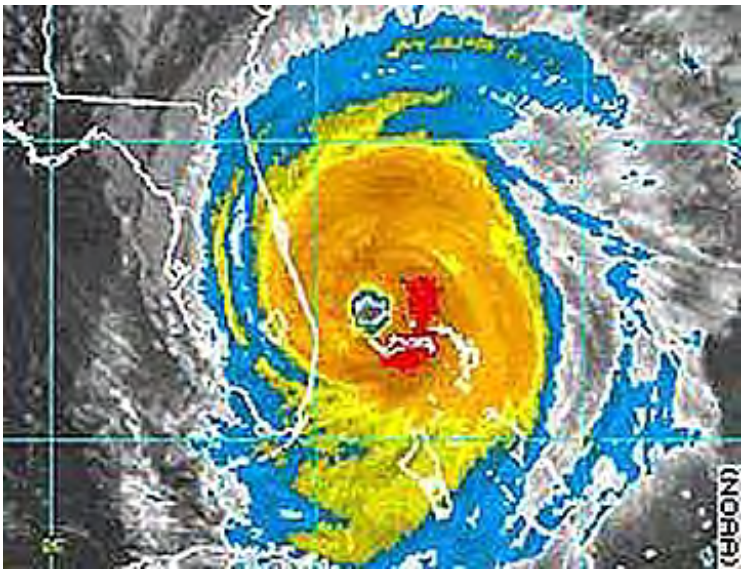


Shortly thereafter, Hurricane Ivan began its assault as a Category 5 hurricane by striking and devastating the island nation of Grenada (85% of the homes there were destroyed or damaged). Then after glancing blows against Jamaica and western Cuba, it has moved against the Gulf Coast east of New Orleans, causing extensive damage in Mississippi, Alabama, and around Pensacola, Florida.:

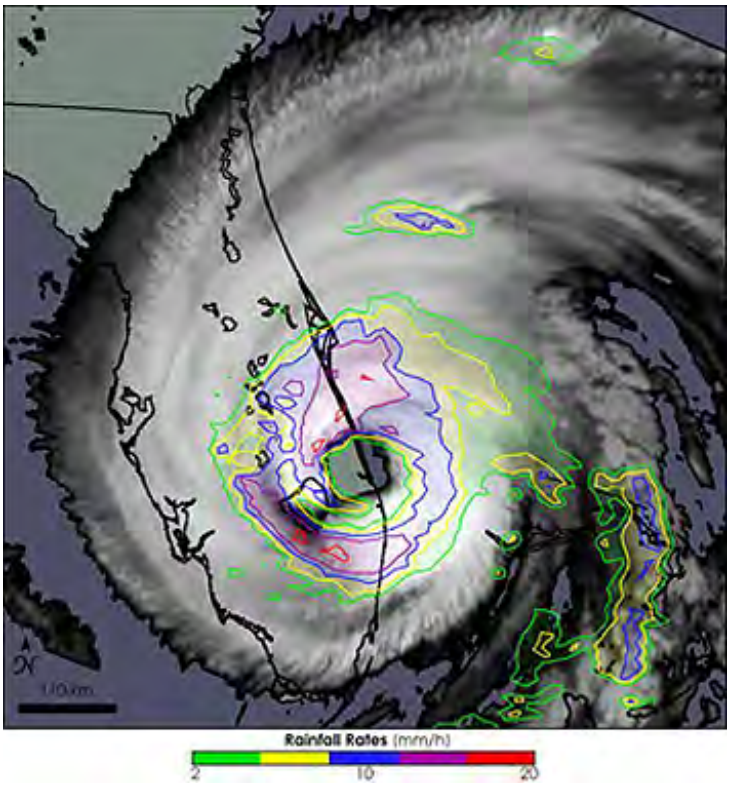


Tornadoes and swirling squalls with high winds did a great deal of damage in waterfront marinas and homes near Pensacola, shown here in two IKONOS "before (right) and after (left)" scenes:

Then, on September 26 a fourth hurricane, Jeanne, a Category 3 with 115 mph, struck Florida near Fort Pierce and Vero Beach, making almost exactly the same landfall as Frances.



The TRMM provided a measure of the amount of rainfall actually falling, as contoured in this image:



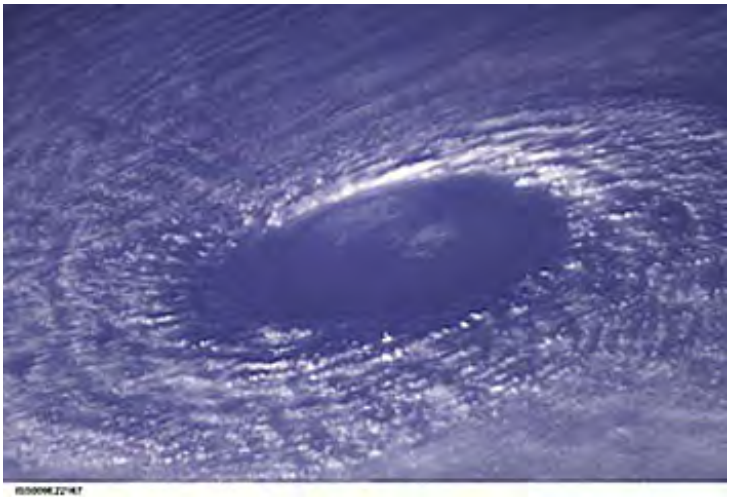
Especially susceptible to hurricane damage are trailer homes anchored on a cement base. Such homes are usually the most readily damaged by the high winds and flooding, as seen here in this Fort Pierce trailer park:

Never before have four major hurricanes struck Florida in one season. The combined dollar estimate for these events is likely to exceed 35 billion. The combined death toll exceeds 100. But before reaching the U.S. Jeanne caused similar devastation, even as it was then only a Tropical Storm, in the Caribbean. At least 1500 people drowned in and near the city of Gonaives in northern Haiti, as homes there were beset by flooding (landslides also contributed to the death toll).

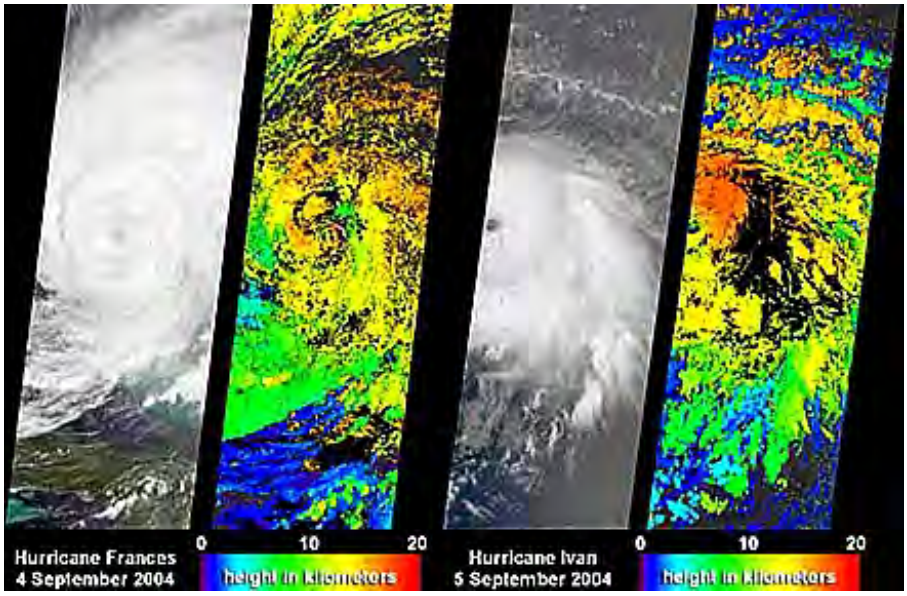
One of the greatest values of meteorological satellites lies in both their realtime tracking capabilities and in their acquisition of data helpful in forecasting their future paths. Here is the prediction of Ivan's course as of 5 AM September 14. In fact, the early morning of September 16 Ivan's Eye struck at Mobile Bay.

Look at Ivan (the Terrible) as it is about to enter the gap between western Cuba and the Mexican Yucatan and into the Gulf of Mexico where it remained a Category 4 hurricane until it came ashore in Alabama, moved up the Appalachians (drenching my house in Pennsylvania with 7 inches of rain), went to New England and - most unusual - turned around, went south over the ocean to Florida, crossed it and finally blew itself out in Texas.:

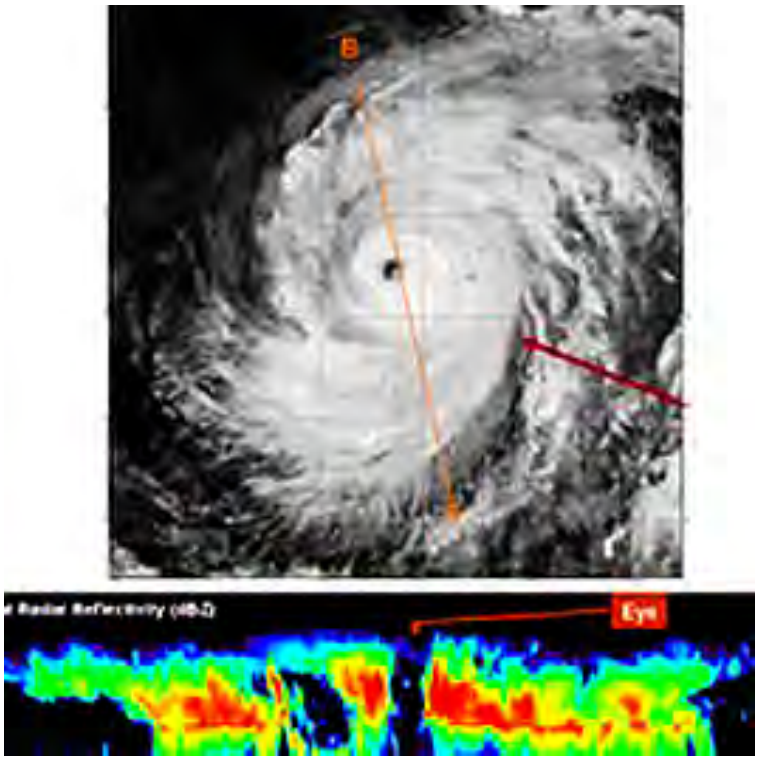
Astronauts on the International Space Station used their digital camera to photograph in detail the Eye of Ivan:



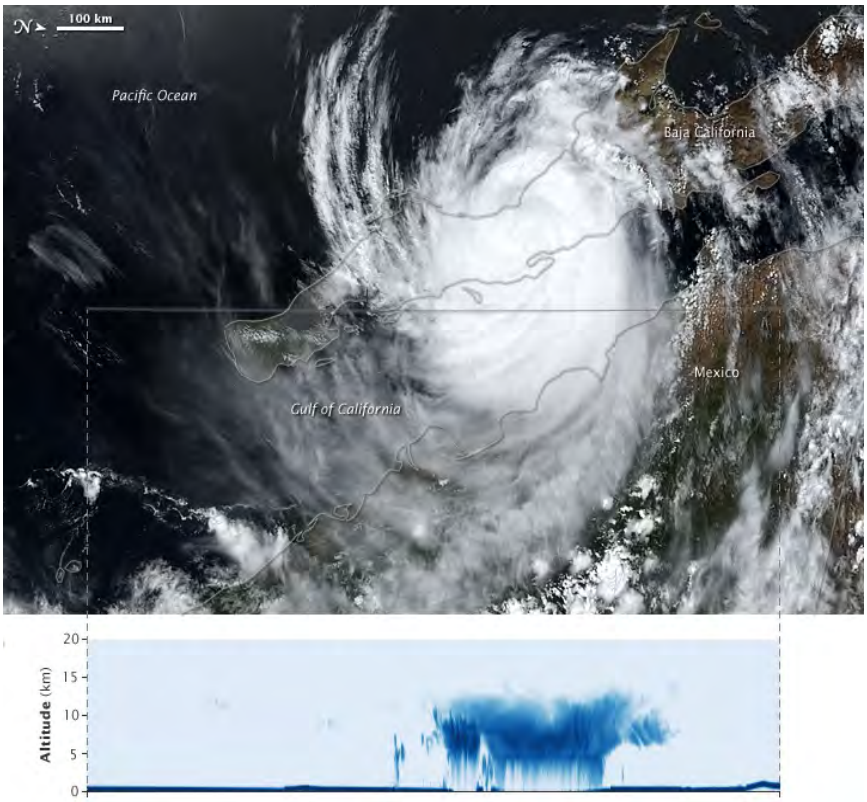
Research satellites such as Terra and Aqua are adding significant knowledge to understanding hurricane characteristics and mechanisms of development. This Terra MISR image (see Section 16-10) shows how measurements from space can be converted to, in this case, cloud heights in different parts of Hurricanes Frances and Ivan.



One goal in using satellites to monitor hurricanes is to use space observations and measurements to improve estimates of intensity (Category levels). This can be done when profiles of wind speeds are obtained. This is possible from Cloudsat data, as illustrated here for Hurricane Ileana in 2006 (the top image is from Aqua):



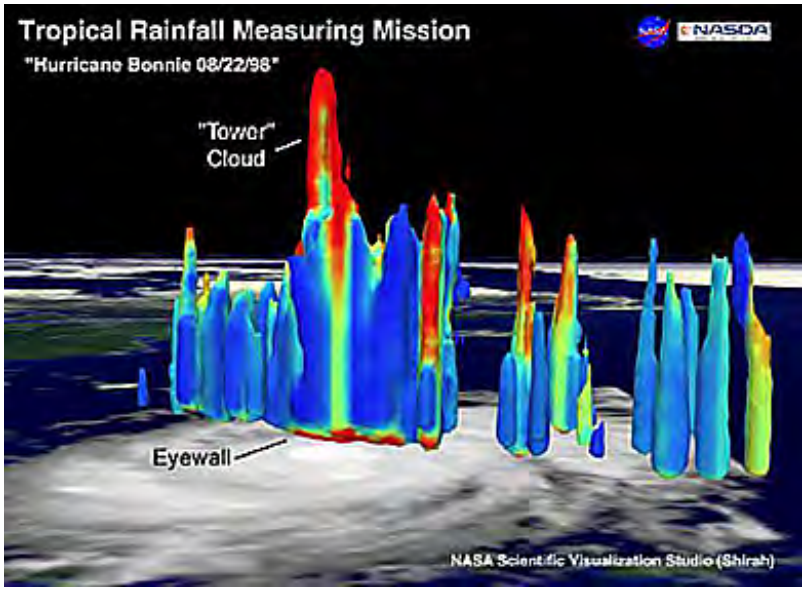
Cloudsat obtains its radar images by bounding radar waves off of fine-sized ice particles in the rain sheets. This type of image appears at the bottom of this view of Hurricane Jimena off Baja California in late August; a MODIS image of that hurricane is shown above:



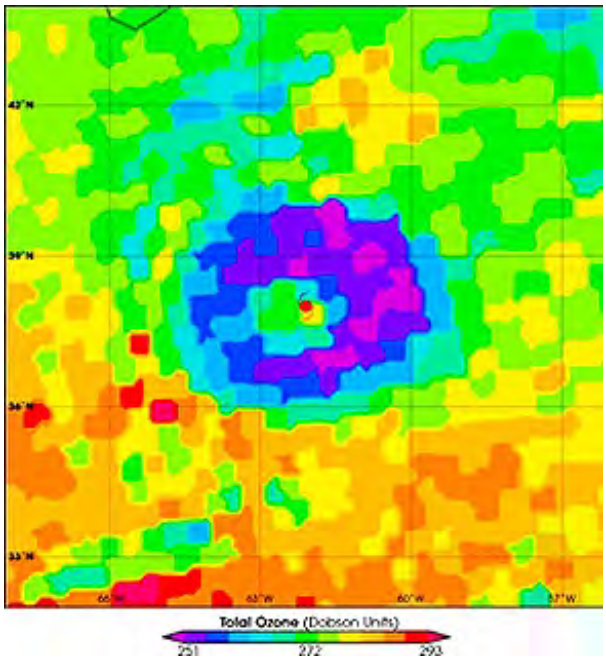
This next MODIS image is a *rarety*, maybe even a first. Here is an ISS astronaut photo of an unnamed hurricane in the South Atlantic moving off the coast of Brazil in late March 2004. Conditions are very unfavorable for hurricane development below the equator in that ocean. Records indicate that this may be the first such hurricane ever to form in the South Atlantic, which accounts for why no name was given it (no list has ever been needed).

14-25a: **How does this hurricane differ from the ones shown above (hint: think southern hemisphere behavior of the Coriolis force)? ANSWER**

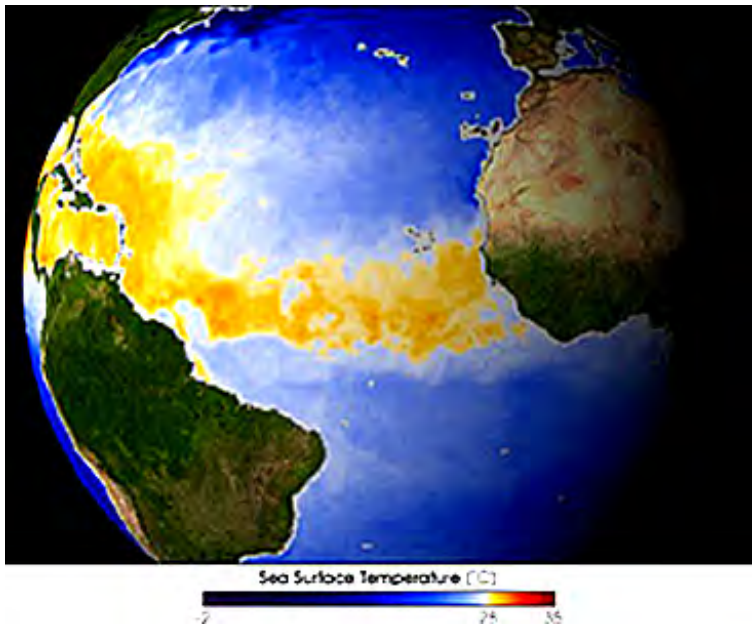
Another sensor on TRMM, which can get vertical profile data, has added to our knowledge of hurricane structure. Convective clouds, called "hot towers" can rise from the main cloud deck of a hurricane, approaching the top of the troposphere, releasing latent heat enroute. This leads to cooling and condensation to produce columns of rainfall. In the 1998 Hurricane Bonnie, the towers reached to an altitude of 18 km (11 miles) as shown in this illustration:



A recently determined phenomenon in hurricanes is a drop in total ozone above the heavy wind/rain segments of a hurricane. An Earth Probe TOMS produced this ozone map for Hurricane Erin:

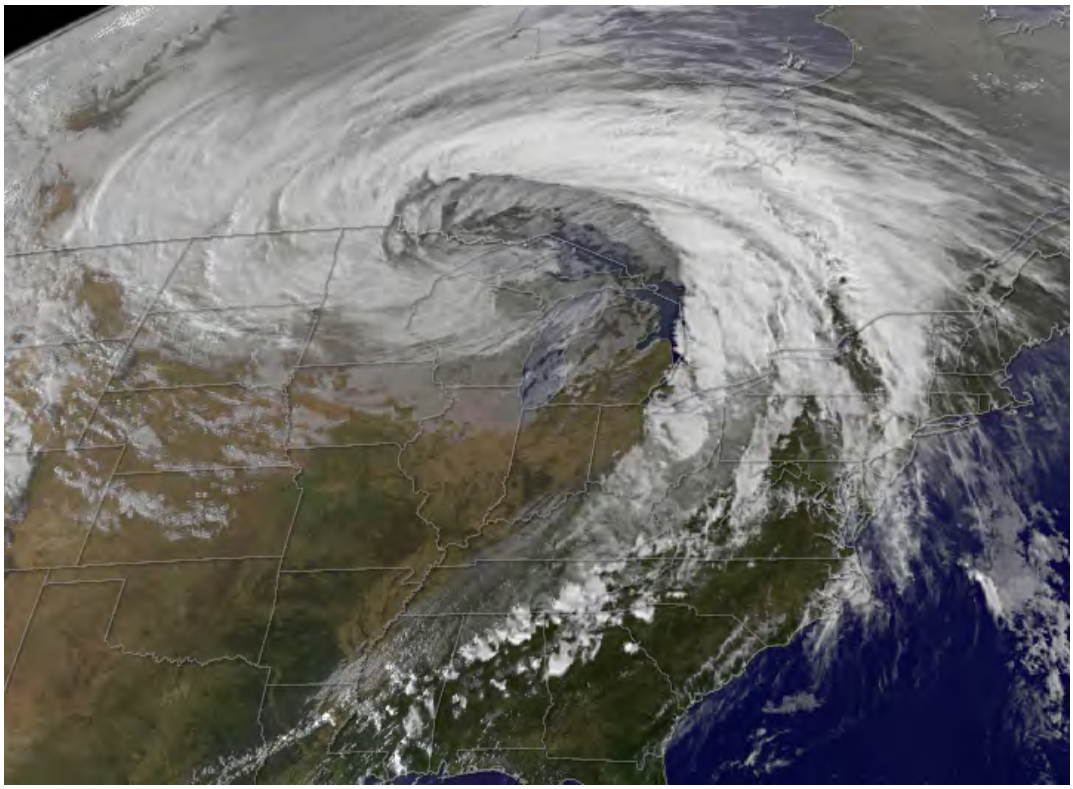


Satellites that measure ocean water temperatures give forewarnings of thermal conditions in tropical waters that control and breed hurricanes and tropical storms. This image, made by the AMSR instrument on Aqua, shows the heating of Atlantic equatorial waters in the early summer of 2003:

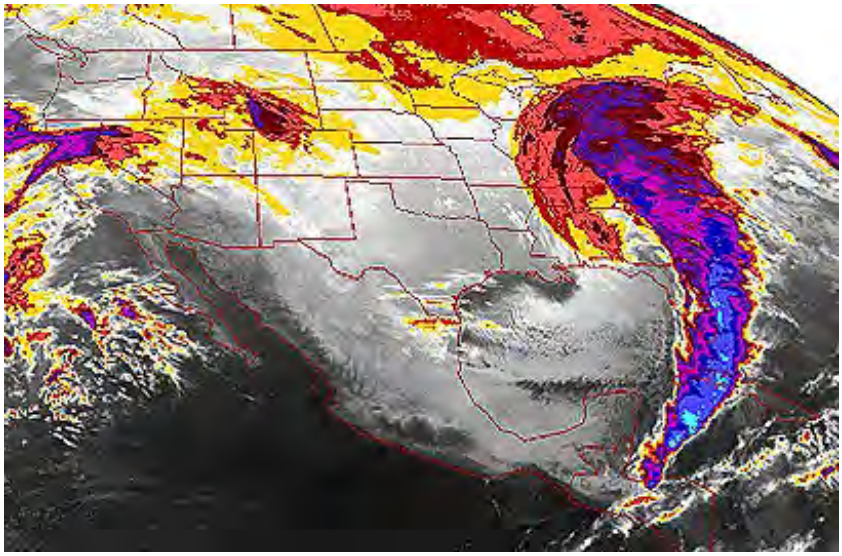


Since this page was prepared (with the assumption that after Hurricanes Andrew and Ivan were about the worst to hit the U.S.), a monster hurricane, Katrina struck first Florida and then the Gulf states with a devastating below. So much information needs documentation here that a separate page is warranted. Access it by clicking on page 10a.

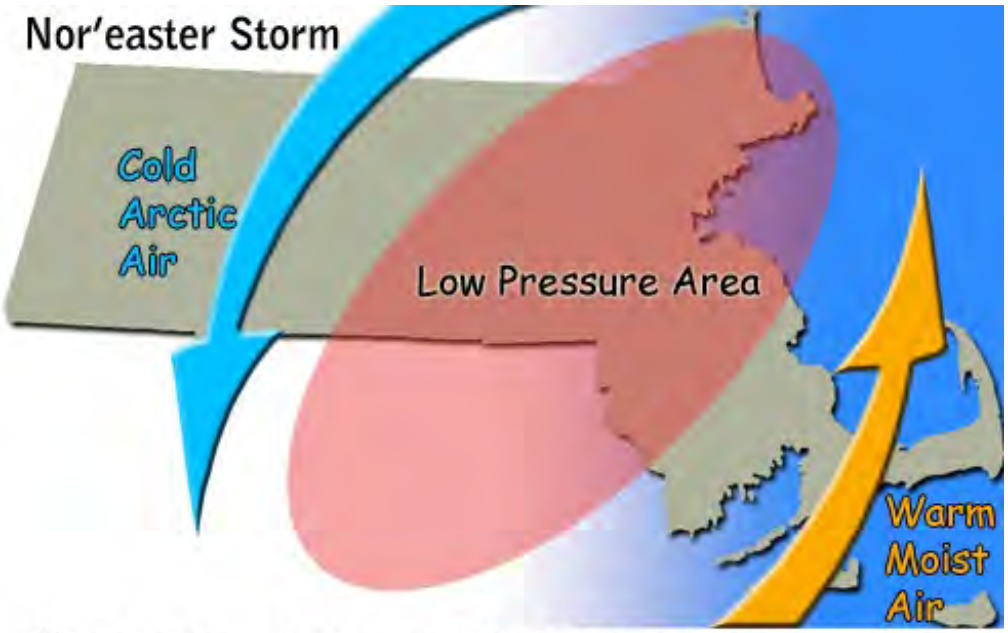
This next image, made by a NOAA GOES satellite, shows what appears to be a well-formed hurricane in the midwestern United States. It is not a true hurricane but is a monster low pressure storm that produced 30 tornadoes and extensive rainfall on October 26 and 27, 2010. At 5:13 p.m. CDT, the weather station in Bigfork, Minnesota recorded 955.2 millibars (28.21 inches of pressure), the lowest on record in North America; this measurement corresponds to the pressure seen in a Category 3 hurricane.



Leaving hurricanes - perhaps nature's most spectacular storms - we turn to land-based storms that can be quite destructive. On March 13, 1993, the "Storm of the Century", shown here in this GOES-7 image, with more intense clouds (color-enhanced), struck the eastern U.S. with massive snow falls and high winds.

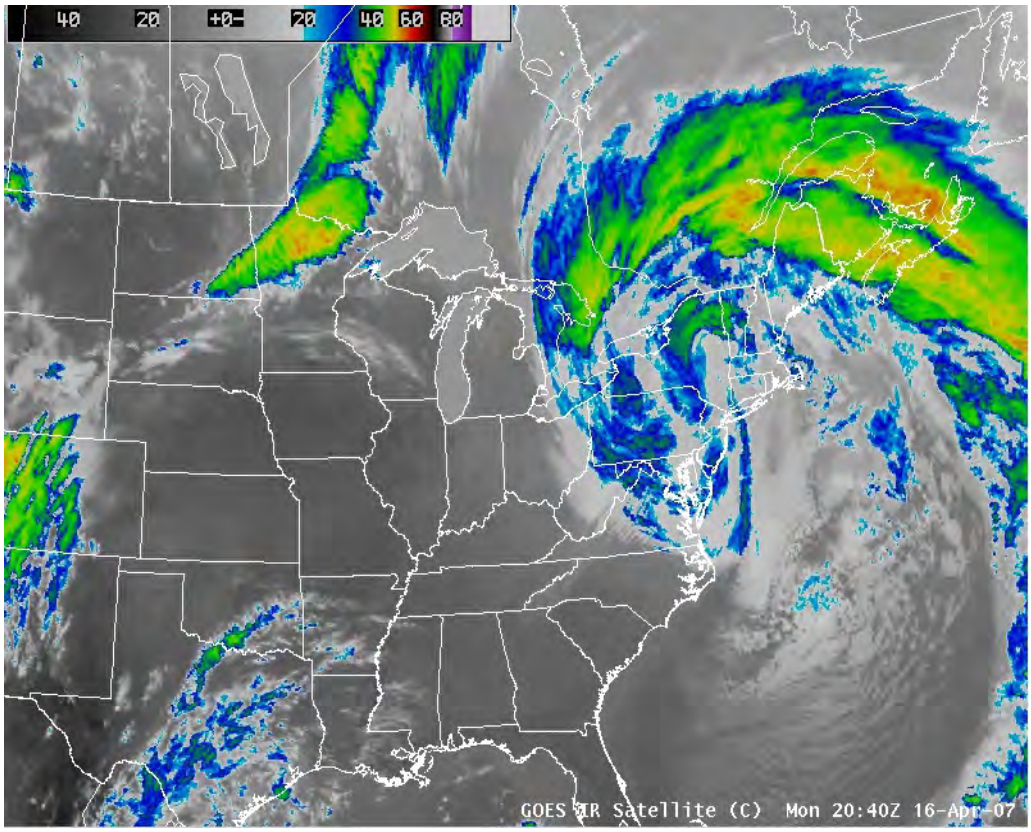


One class of snowstorms (or if too warm, extensive rainfalls) is the so-called "noreaster". The term has two meanings: first, this storm is especially common in the northeastern U.S.; second, the winds tend to be coming from the northeast, i.e., blow towards the southwest. This is illustrated by this schematic for a noreaster over Massachusetts:

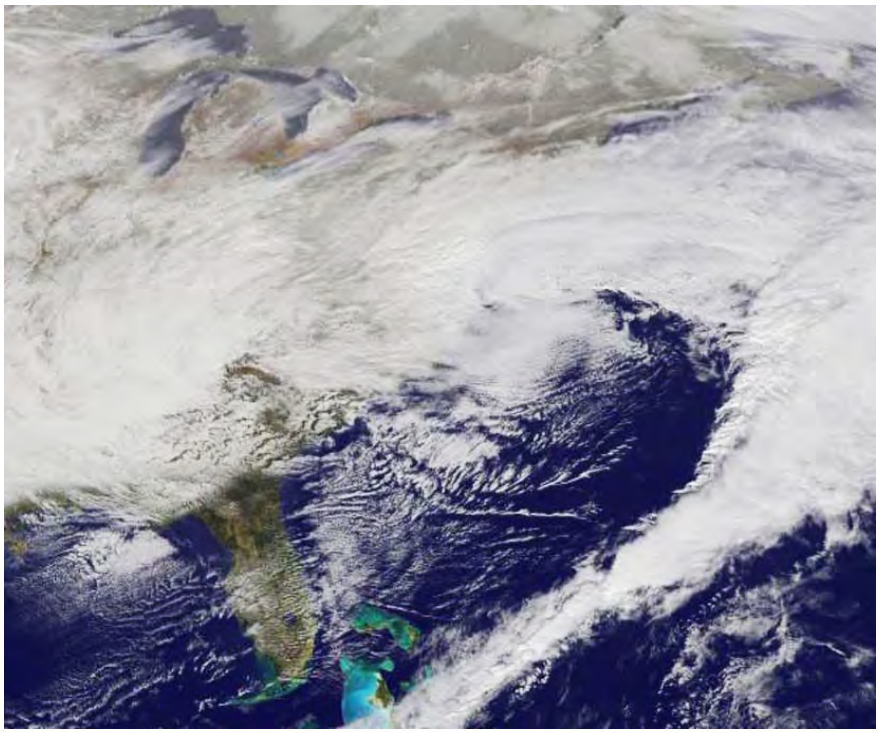


*Storm not drawn to scale. Usually, a nor'easter will encompass much of the east coast. I have drawn this to show you it's effect on Massachusetts.

Noreasters usually originate either in the South Atlantic or come across the southern United States from the west. They track from south to north along the Atlantic States. The wind pattern shown above results from their centers being offshore to the east.



Some noreasters are huge. The storm of February 8, 2010 covered most of the eastern U.S. into the Carolinas. It dumped more than 20 inches of snow on Washington, Baltimore, and Philadelphia. Here are two satellite images:

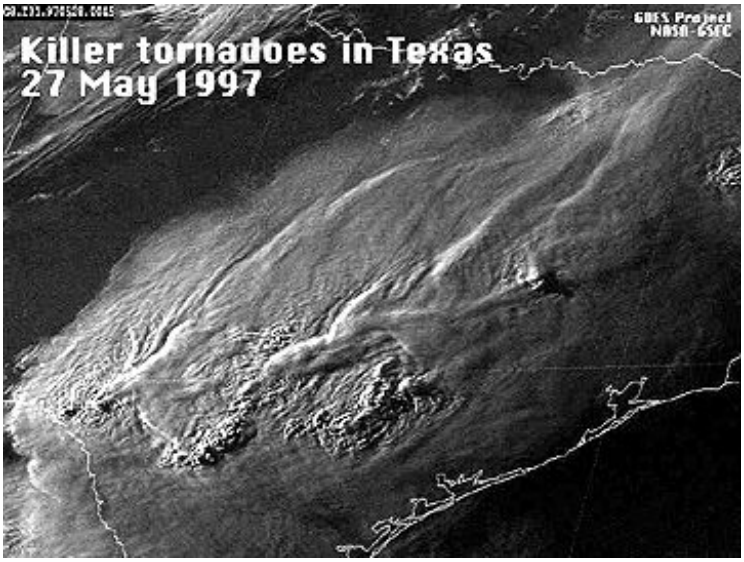


Snowfalls are usually widespread. But under some conditions a blanket of snow covers only a rather limited region. This happened on February 26-27, 2004 when a foot or more of snow was dumped by a clipper storm passing over much of North Carolina. This is how it looked the day after, as seen from MODIS. With such imagery one can delineate snow accumulation boundaries with high accuracy:

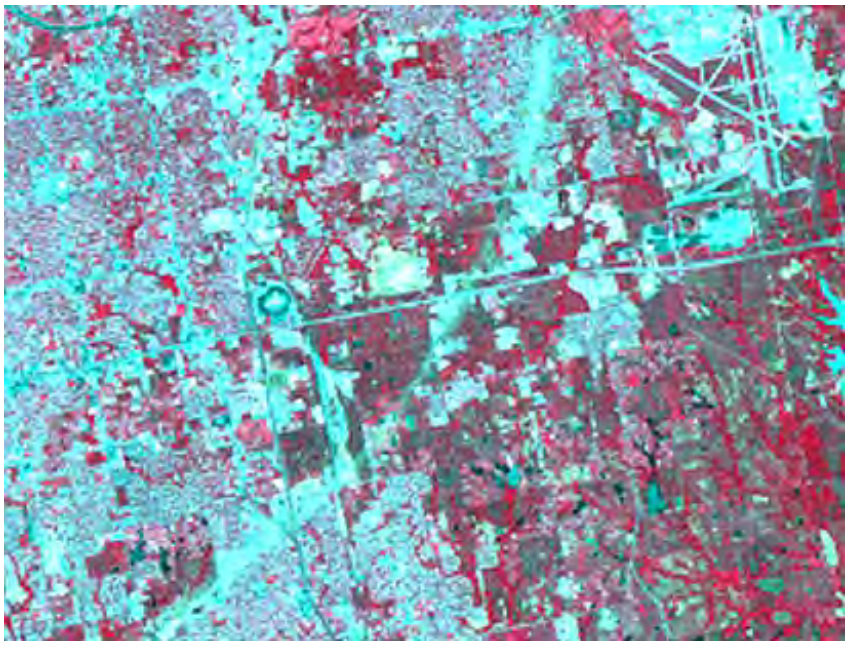
Metsats are excellent monitors of less violent, more conventional, but still interesting weather data. We are all familiar with the passage of cold fronts, followed by clear weather as a pressure high moves in. Commonly, moist, warm air is encountered by the front and develops into a broad area of clouds and often stormy weather. This next image shows the sharp leading edge of a front now off the Atlantic coast of the U.S. and out to sea. At its back edge is another sharp boundary beyond which cold, dry air is moving in.

Tornadoes, while much more localized than hurricanes or subcontinental storms, can do incredible damage to the areas on which they touch down. Viewed at close distances, most tornadoes have an ominous, sometimes terrifying look, that helps to explain why they are so feared (add to this the knowledge of the horrendous damage they can cause). Here is an example:

Tornadoes are low pressure systems, and can generate wind velocities that can exceed 200 mph. Severe ones will flatten buildings (tornadoes have a special fondness for trailer parks), uproot trees, and carry objects for hundreds of feet to even miles from their spot of origin. The storms that brew them are effectively tracked by ground radar and by satellites (usually geostationary). On May 27, 1997 several lines of storm clouds bearing multiple funnel clouds crossed central Texas with deadly results. Here is a GOES-8 image of these advancing fronts.



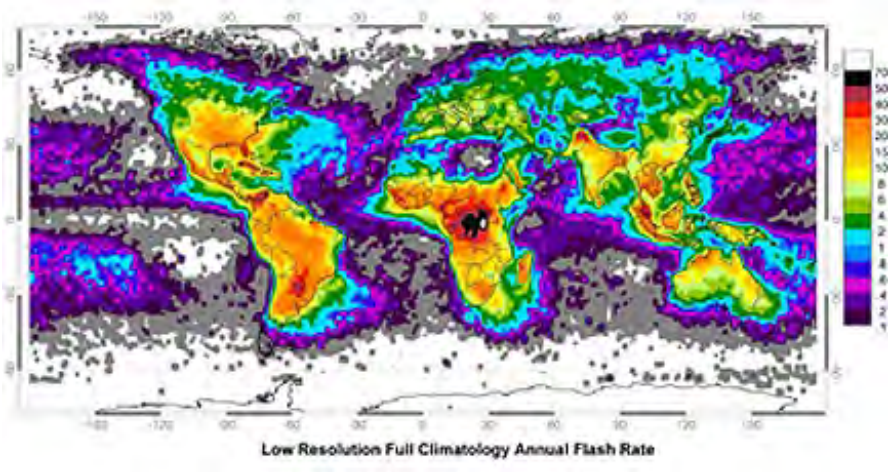
When a storm has passed and the sky clears, a passing satellite can often see the actual damage if conditions are right. This was the case three days after a killer tornado touched down on May 5, 1999 within Oklahoma City, OK., cutting a path through city and suburbs and just missing the airport. Look for the diagonal blue strip in this IRS-C LISS false color composite - this marks the zone of heavy destruction.



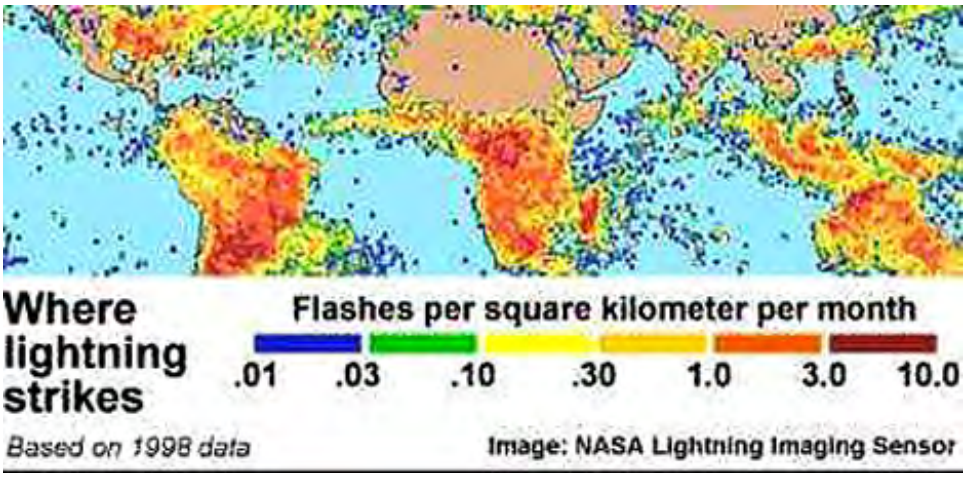
Now that IKONOS and other high resolution sensors are operational, it is feasible to inspect damage from tornadoes in considerable detail - equivalent to flying an aerial photo mission. A tornado touched down in Fort Worth, TX, ripping off roofs from warehouses and other buildings, as evident in this 1 m resolution IKONOS image.

The first tornadic storm of 2002 to produce multiple deaths (4) occurred on April 28 around La Plata, Maryland, about 40 miles southeast of Washington. This may have been an F-4 tornado (the second highest wind category) and as such was the most powerful ever to hit Maryland (and one of the strongest ever to hit the East Coast region). Within just two days, the narrow path of destruction (that was, however, 6 miles long) in which several hundred buildings were damaged was imaged by EO-1, as seen here (the pink color was caused by destruction of vegetation [reducing the natural green input]):

Still another hazard, and sometimes a killer, associated with usually severe storms are *lightning strikes*. Worldwide thousands of individual lightning discharges, including dramatic bolts, occur each day.This next map shows the geographic distribution of the frequency of strikes averaged over 5 years (1995-2000) of data collecting by NASA's Optical Transient Detector and by the Lightning Imaging Sensor on TRMM. Note that the central part of Africa has the most lightning strikes; almost all of South America is prone to frequent electrical storm activity.



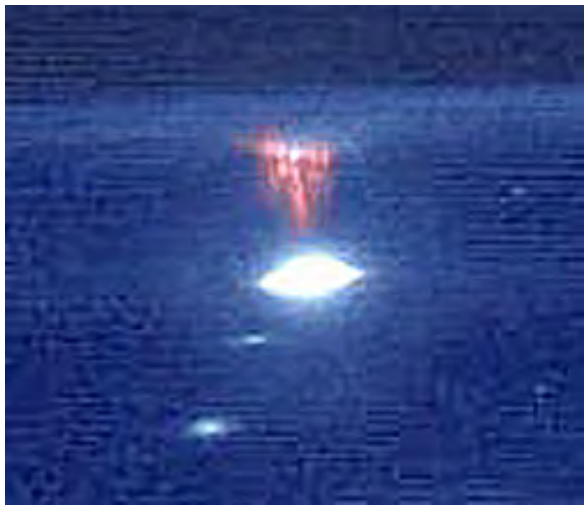
A different perspective occurs when one looks at the distribution of lightning strikes on a monthly basis, as plotted in this global map.



Still, when compared with the previous global map, similarities are evident.

A less well known phenomenon is the occurrence of "sprites", a term applied to electrical discharges in the upper atmosphere (ionosphere) at altitudes from 60 to 100 km. These develop

almost coincidence with ground strikes. The Republic of China's ROCSAT has an instrument (ISUAL - Imager of Sprites and Upper Atmosphere Lightning) that concentrates on studies of this type of discharge (dissociation of rarified gases accompanying the lower atmosphere lightning strikes). Here is an ISUAL image of the near surface lightning and above it a reddish glow from ionized nitrogen



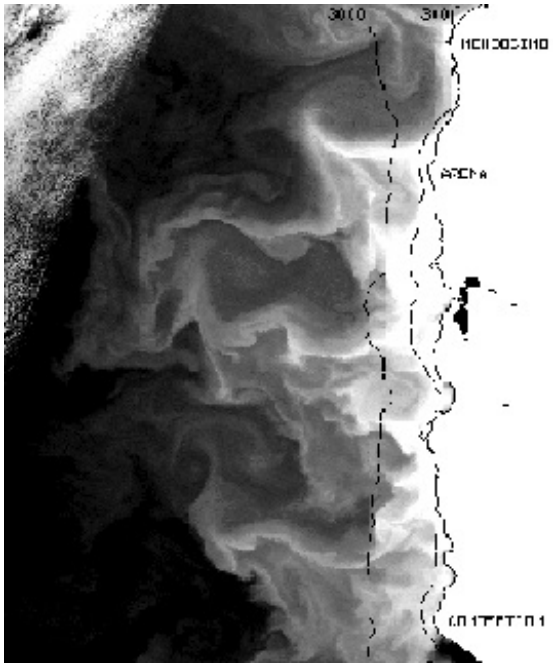
PREVIOUS RST NEXT

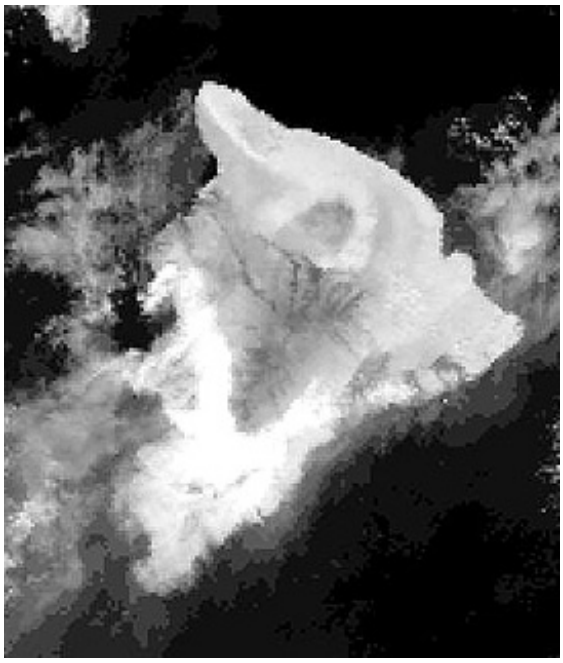
Primary Author: Nicholas M. Short, Sr.

Most meteorological satellites designed to monitor the atmosphere do not provide "good looks" at the ocean, largely because they are purposely lower resolution systems that need to monitor atmospheric properties at gross scales. Their thermal bands do broadly indicate temperature differences in oceanic water that give some information on current flow. The NOAA satellites, for example, produce valuable thermal data that can be plotted at regional to global scales. But, other aspects of the marine surface waters including sediment patterns, sea state, wind flow, wave patterns, planktonic life, etc. are of vital interest to oceanographers. A number of satellites dedicated to gathering oceanic data have notably increased our understanding of these great bodies of water; some of this information bears directly on weather and climate control. This page describes some general types of observational data from such satellites and shows the value of continual monitoring by focusing on the El Nino and La Nino thermal events in the Pacific.

Oceanographic Observations

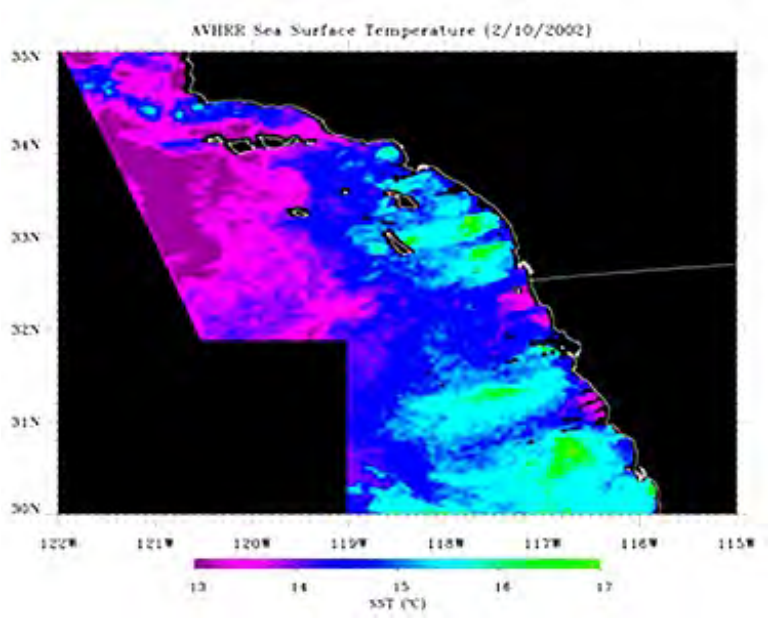
We now transit from viewing atmospheric weather systems from space to observing the oceans, which contain about 360 million cubic miles of saline water (97% of all surface/near surface water; the rest is classified as fresh). Much of the oceanic information, as gathered from space, comes as secondary information from Metsats, although a growing number of satellites, and a few Shuttle missions, are/were specifically dedicated to collecting oceanographic data. The kinds of data acquired by the sensors include the following: sea-surface temperature, oceanic-current patterns, formation of eddies and rings, upwelling, surface-wind action, wave motions, ocean color (in part indicative of phytoplankton concentrations), and sea ice status in the high latitudes. Coastal and shelf waters adjacent to continental margins generally show considerable variation in near-surface temperatures. Some variation is due to inflow and mixing of river waters but ocean currents and upwelling also modify the patterns. Look at these thermal-IR images, made from NOAA AVHRR data, of part of the California coastline from Mendocino, south to Lompoc (top), and the big Island of Hawaii (bottom), in which offshore warmer waters are displayed in lighter tones.



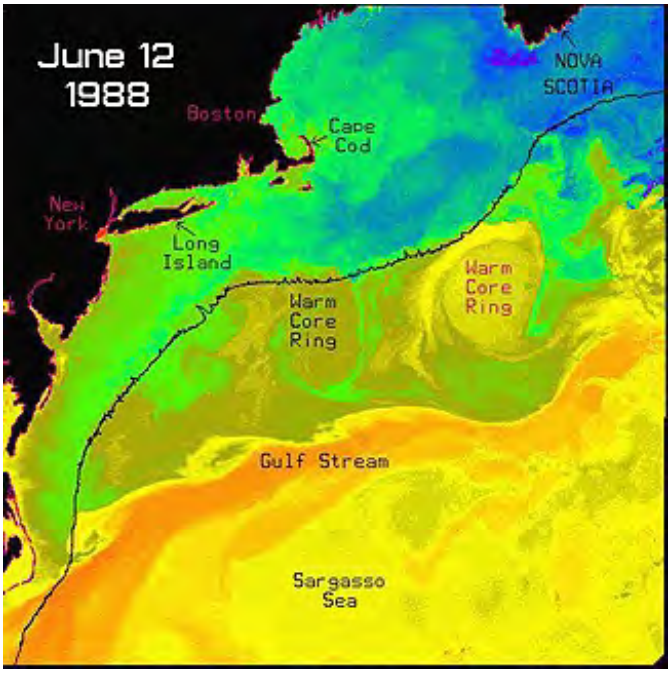


14-26: **Comment on the thermal patterns in both of the above images. ANSWER**

In recent years these temperature maps are now routinely rendered in color, as this map of the coastalwaters off southern California aptly illustrates:

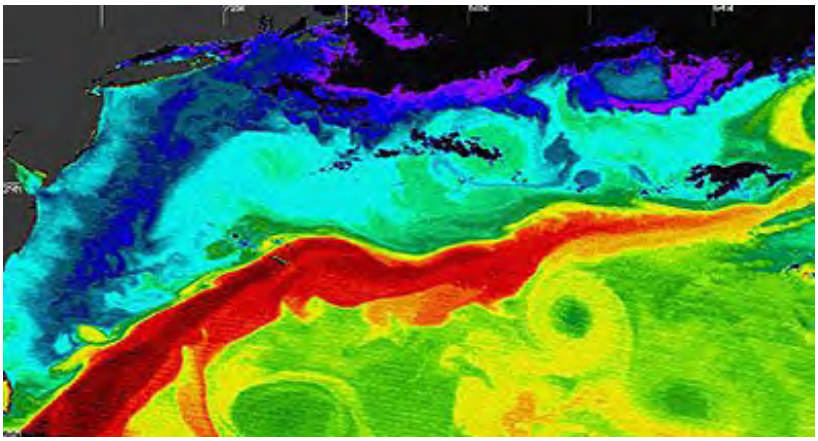


Ocean currents, such as the Gulf Stream off the eastern U.S. coast, and the Pacific current, off the west coast, result from redistribution of warm water that collects in tropical regions and flows towards cooler zones at higher latitudes. A color-coded rendition (below, top) of part of a Day-Thermal HCMM image (page 9-8) shows the well-defined Gulf Stream off the North Carolina-Virginia coast. Paired with this image is one of the East Coast showing surface temperatures calculated from algorithms that process multi-channel data obtained by NOAA-14's AVHRR.

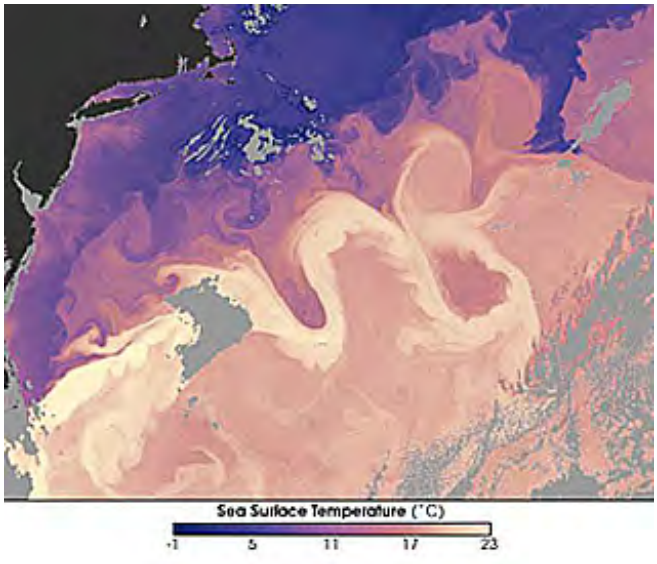


That second image also shows the Gulf Stream, which, in places, breaks into warm core rings, i.e., some meanders in the current get pinched off (cold core rings also occur). Temperature values for the colors include: orange = 25-28 ° C (77-82 ° F); yellow = 23 ° C (72 ° F); green = 14 ° C (57 ° F); blue = 5 ° C (41 ° F).

A recent image from Terra's MODIS, using bands at 11 and 12 µm, shows how sharp the temperature contrast can be between the main Gulf stream (red) and surrounding waters:

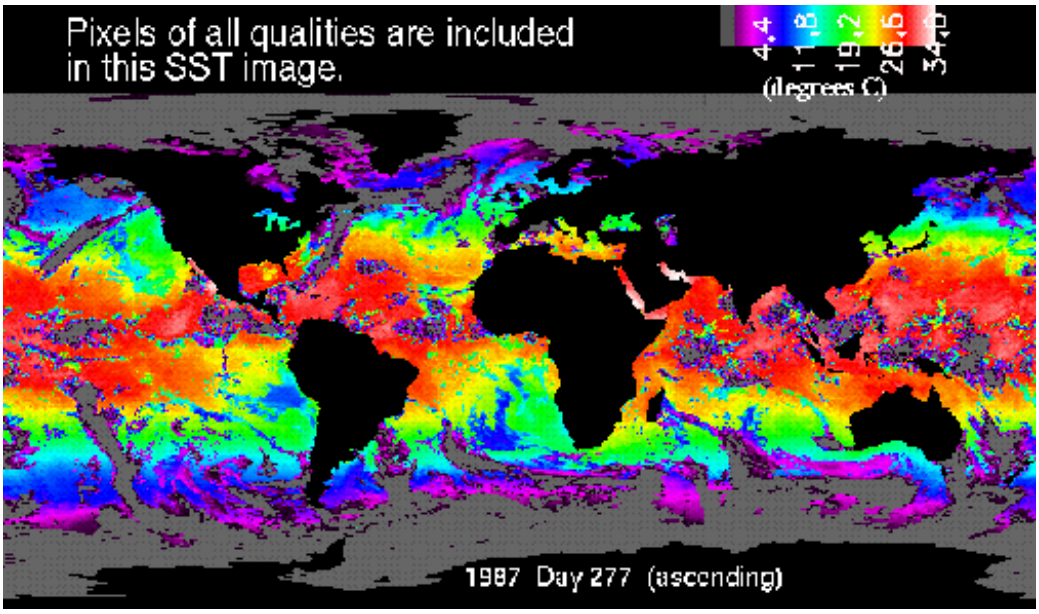


Another MODIS image of the warm Gulf Stream emphasizes its tendency to meander as it moves northward:

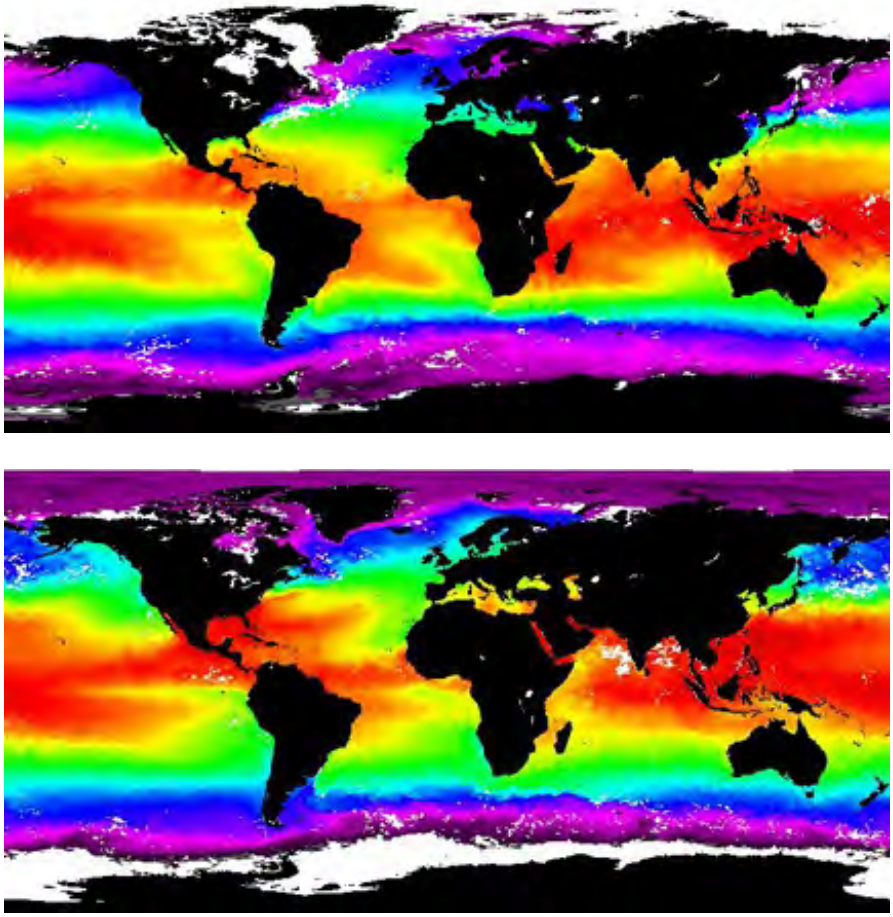


14-27: **Birdwatchers today are taking a new kind of bird trip: the pelagic (off-shore ocean) boat trip to see many varieties of sea birds. Off the Virginia and North Carolina coasts, these trips often venture as much as 60 to 80 miles seaward from the coast. Can you surmise as to why? ANSWER**

Daily, Metsats also routinely procure global observations of temperatures in marine waters (known as Sea Surface Temperature or SST). Here for example is a map of SST made in late September of 1987.

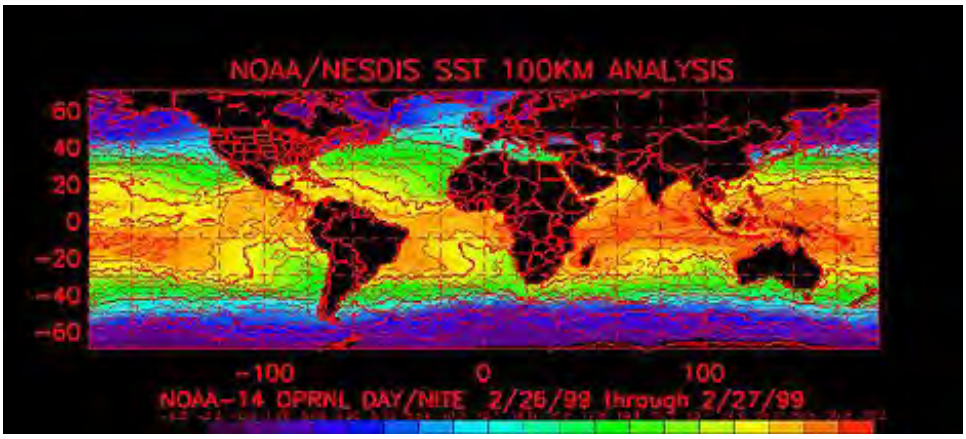


We can integrate SST values into calendar intervals and thus compare them month by month or between equivalent periods in years. Below are SSTs for the months of January (top) and July (bottom) in 1993, as determined from NOAA AVHRR data.

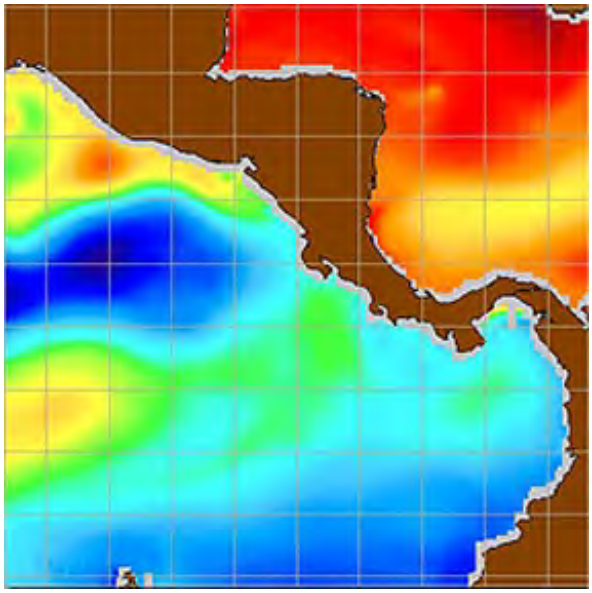


Again, reds and yellows show warm water, and blues and purples depict cold water. At first glance, we don't see much seasonal variation, when we view it worldwide, although close inspection reveals real differences. In general, the oceans tend to maintain their average temperatures with notably less variations than the atmosphere above them.

Sea surface temperature distribution on a worldwide basis can be obtained from NOAA satellite data on a near real time basis. Here is a plot for February 26-27, 1999:

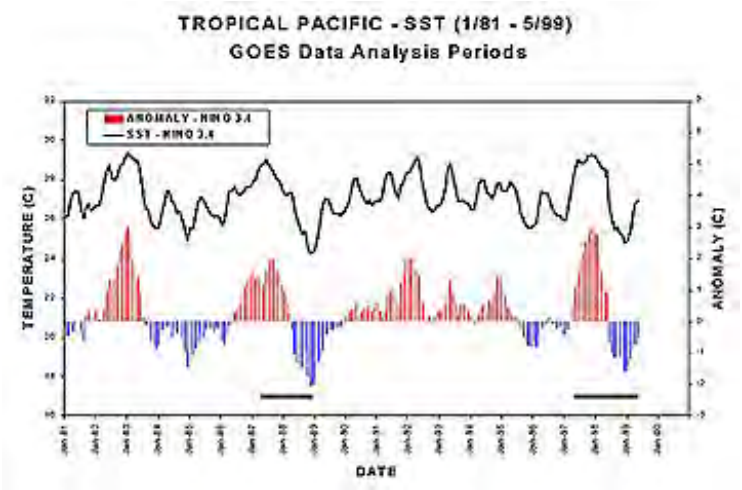


Orbview-2 has a thermal sensor. Here is a view of Central America with the warmer Gulf of Mexico to the upper right and the more variably warm to cool eastern Pacific Ocean to the left.



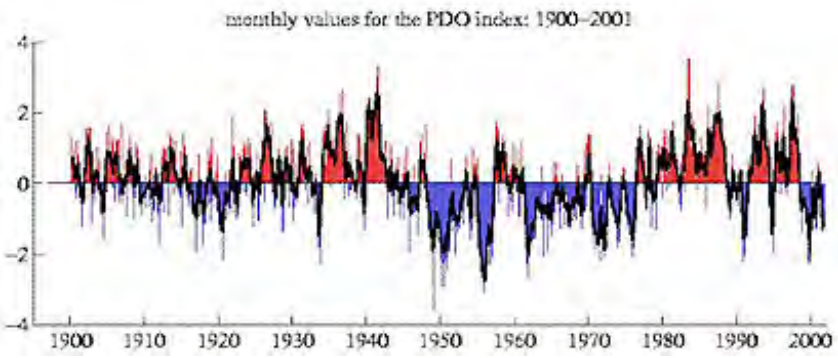
El Niño

Ocean surface temperatures have been measured since the late 19th century. It is now well known that the values of these temperatures, and their changes from year to year, are key factors in the global and regional climate and weather patterns. Note the variations from the average of temperatures in part of the Pacific Ocean (this large body of water is the main influence on global meteorological conditions although the Atlantic has a strong effect on Europe and Africa).

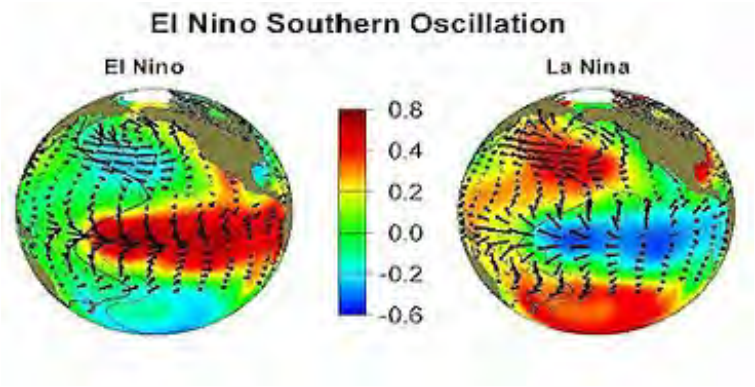


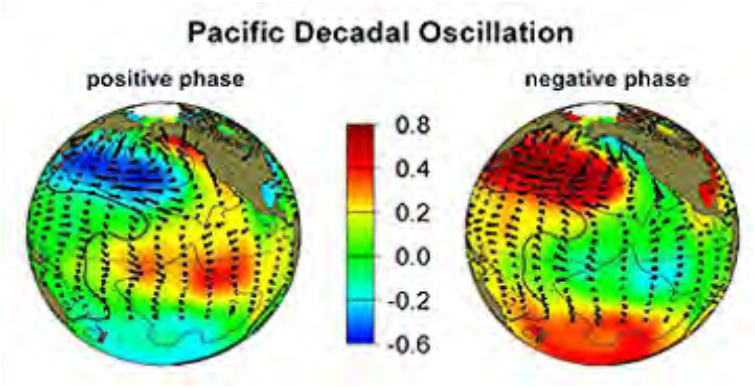
These "ups and downs" appear to follow specific trends. Changes can be recognized to define

notable variation on a 6 to 18 months timeframe. This next plot reveals a time frame for temperature values in the Pacific Ocean that seems to be on the order of 20 to 30 years (indicated by stretches of dominantly red [hotter] or blue [cooler]):

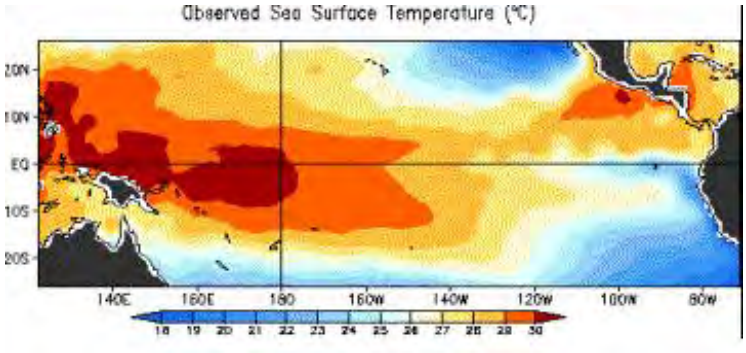


The short period variations refer to a condition in the Pacific Ocean called "El Niño" (Spanish for "the little child" namel, Jesus Christ - the term was applied by South American fishermen to refer to warm waters around Christmastime). The long period pattern is called the Pacific Decadal Oscillation (PDO). These two illustrations indicate a general or averaged temperature anomaly (either hotter [red] or cooler [blue]) in the Pacific Ocean for "El Niño" (also called ENSO for "El Niño" Southern Oscillation) and for the PDO.

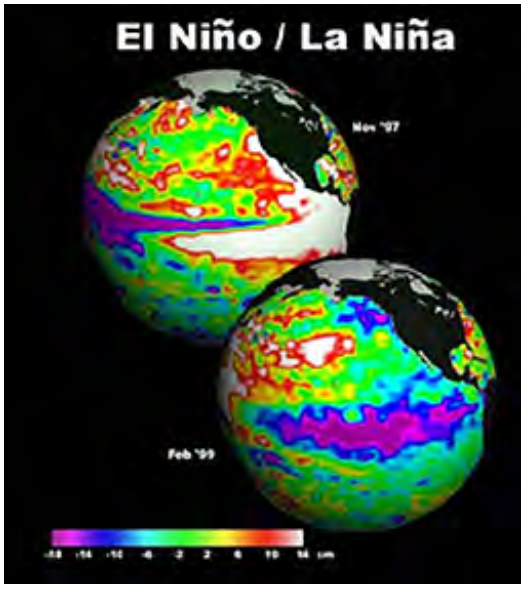




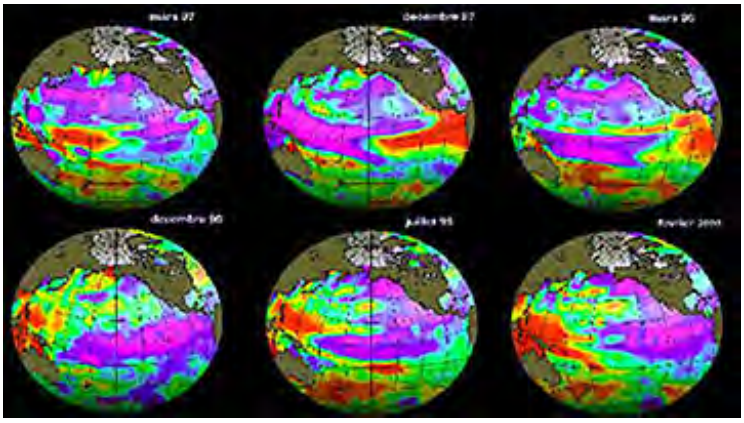
Here is a similar diagram for what is called La Niña, which describes the condition of now colder water off South America, with the previous hot water having migrated westward towards Australia.



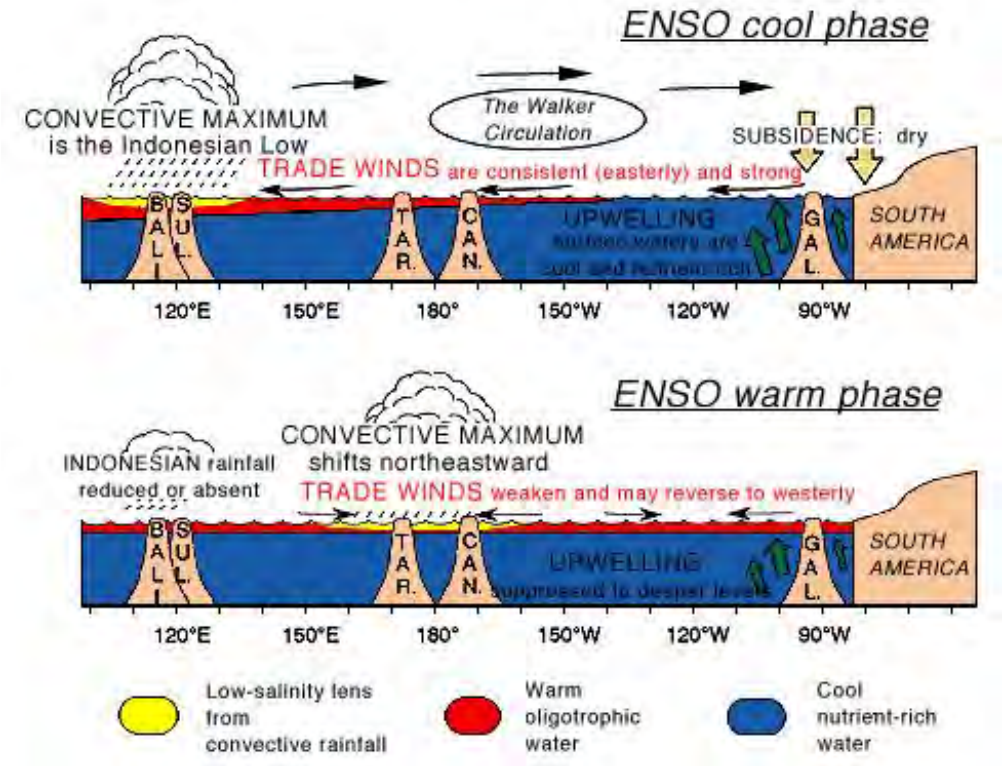
Generally, El Niño and La Niña's alternate during several years (but short periods of so-called "normalcy" also are interspersed). We summarize the temperature anomalies of the two phenomena with this well-known illustration:



Between times of maximum anomalous temperatures, the temperature distribution in the Pacific shows considerable locational variability, as depicted by these measurement made by Envisat's thermal sensor:

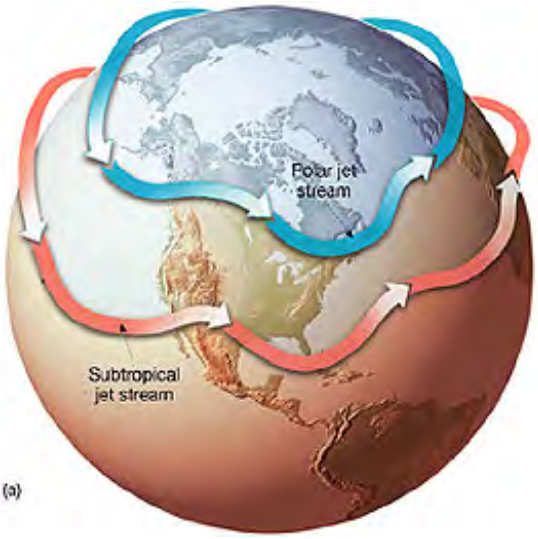


An El Niño event results from changes in atmospheric pressures in the eastern Pacific Ocean that cause the normally westward flowing trade winds to reverse direction, which, in turn, diminishes or reverses an upwelling of cold water off the South American coast and displaces the Peruvian current. The surface waters there become warmer (by as much as 8° C [15° F]) leading to increased southern-hemisphere summer-storm activity. This next diagram offers more insight into this behavior:



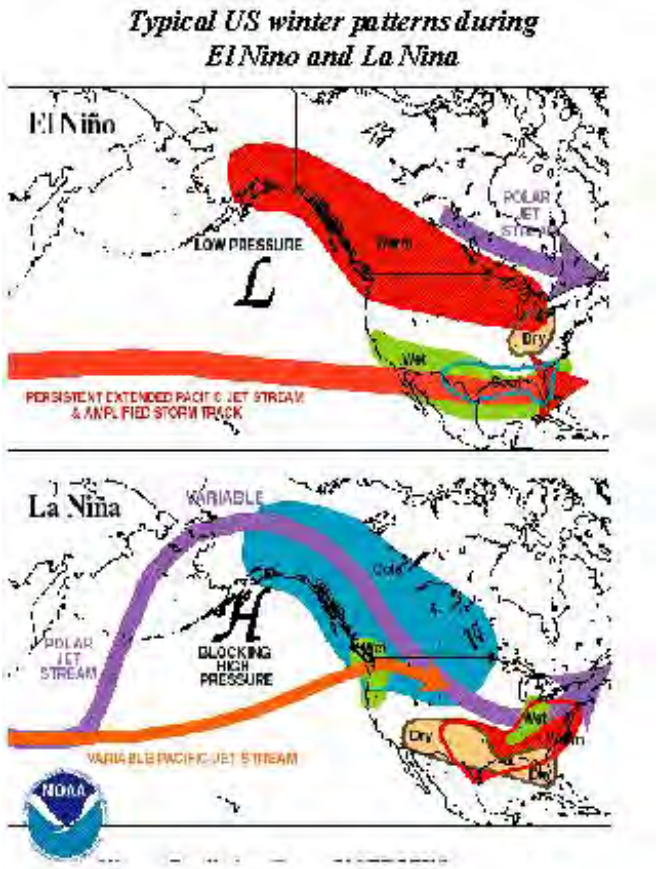
Above the North American continent, an El Niño can greatly perturb normal weather patterns, causing abnormal rainfall in some parts of the country and droughts elsewhere. Ferocious storms are more frequent and hurricanes may increase or decrease from normal numbers, depending on the effects in the Atlantic and Pacific Oceans. An El Niño usually precedes a La Niña, essentially a reversal of conditions off the western South-American coast, in which colder water than usual comes to the surface.

El Niño's effect on continental weather, such as over North America, is essentially that of its repositioning of the Polar and Subtropical Jet Streams. As a reminder of their general locations around Earth, we return to this illustration from the meteorological Tutorial at the beginning of this Section:

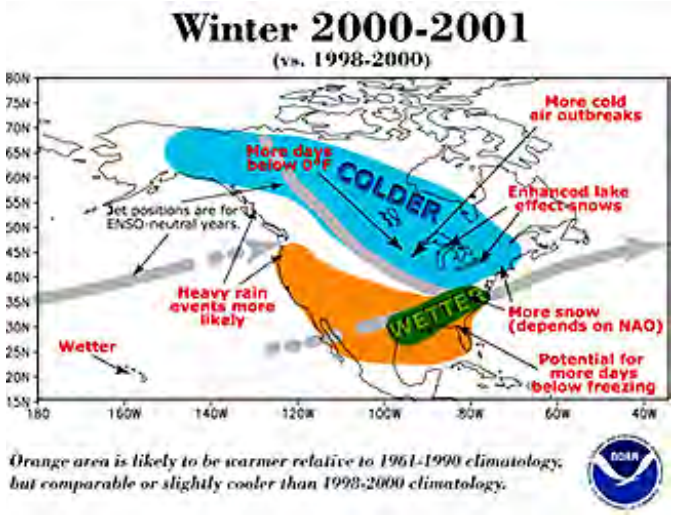


This easterly flow of air both steers weather systems and serves as a "barrier" to air masses. For the Polar Jet Stream, the air to its north is usually cooler, or even cold, relative to the air to the south. The Subtropical Jet Stream helps to position moist air to its south and drier air to its north. The two streams twist (meander) and shift (towards the north or south) in both short time spans (weeks) and longer (seasonal). Thus, one's local climate at some time during a year will be determined in part by the position of the relevant Jet Stream overhead. The effect of an El Niño is to warm or cool the air that then flows laterally to perturb the usual position of the Jet

Streams so as to make any given region warmer or colder and wetter or drier than normal.



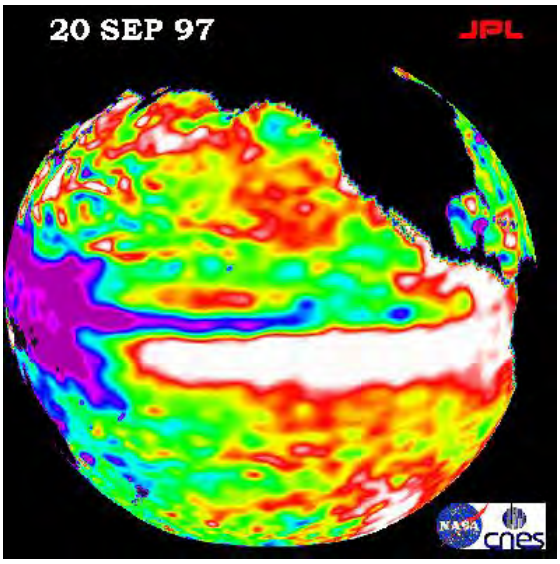
This ability to follow and interpret the Jet Stream shifts resulting from El Niño and La Niña is a key input to prediction models that forecast expected weather conditions for the coming season(s). Here is the forecast for the Winter period 2000-2001; in fact, the actual weather history was similar to this model.



Satellite observations have been instrumental in identifying and understanding the "El Niño" weather phenomenon. Let us now show several examples of this use during one particular time span. These utilize data from TOPEX/Poseidon (next page) which uses radar to characterize sea surface elevations that rise when water is warmer and fall when cooler. The overall El Niño condition was particularly active in the early 1980s - publicity about which made the general public aware of the phenomenon. It was active again in the late 1990's. Experts in early 1997 predicted a very strong El Niño condition for the latter half of that year into 1998 - and this event indeed occurred. Its disaster phase may have started in 1997, with Hurricane Pauline on October 9, the strongest to hit the west Mexican coast in decades (devastating Acapulco). Events into

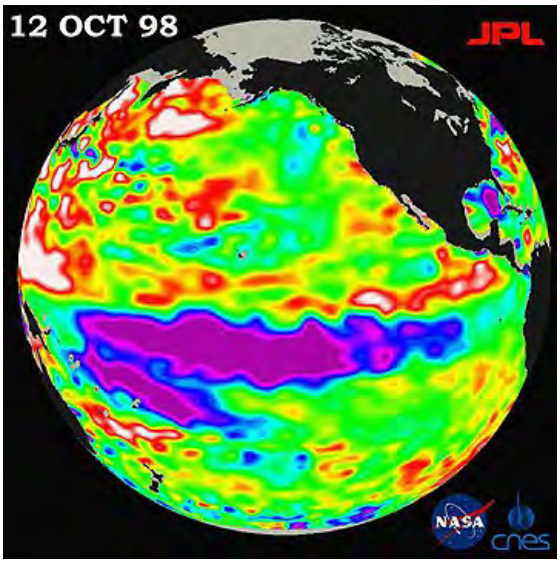
January 1998 bear out the forecast, with heavy rainfall in the Southern and Western U.S., ice storms in New England and Canada, and abnormally balmy weather in some places.

Onset of marine warming began to appear by the time this TOPEX/Poseidon image was taken on September 20, 1997. It shows a broad, elongate band of very hot water stretching westward from Peru across the Pacific.

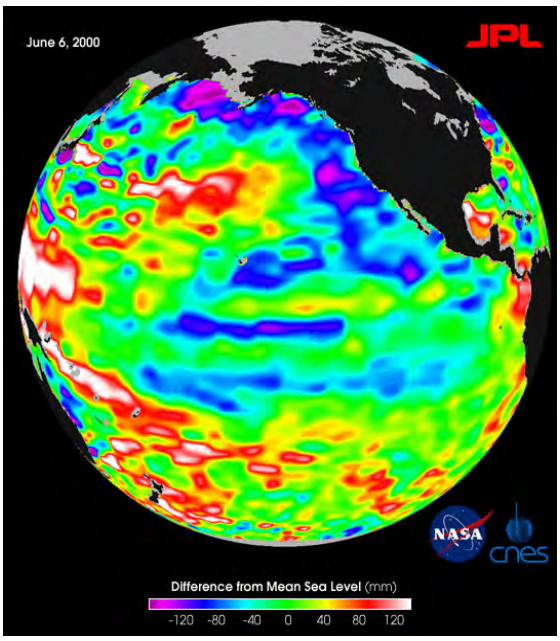


Although regional in extent, this water-temperature perturbation can decidedly influence weather over much of the Earth, as the conditions in the western Pacific also become disturbed. The combination of warm Pacific Ocean temperatures and shifts in the atmospheric jet streams means that certain areas of the world with typically dry conditions instead get heavy rain and potential flooding, while other regions can be drought-stricken. An El Niño occurs every two to seven years, reversing normal weather patterns throughout areas of Canada, the United States, Mexico, South America and as far away as Africa. The previous major El Niño occurred in late 1991 through mid-1992, with a smaller and less destructive one recorded in 1994-95. The weather system can last as long as eighteen months, depending on how rapidly the ocean temperatures cool and return to normal. El Niño has a profound effect on global weather systems. This idea is supported by the recent summary of 1997 worldwide temperatures: that year was the warmest twelve months in terms of average temperature maxima since records for the entire Earth started in 1860. Some of this may also reflect a significant contribution from global warming.

As the summer of 1998 moved into fall, with several major hurricanes including Mitch which killed more than 10,000 people in Honduras and neighboring Central American countries, a transition began, in which the earlier El Niño gradually changed into a La Niña. By mid-October the central band of cold water had largely replaced the equatorial belt of warm water off South America while shifting the warmer water towards Australia, setting up the conditions associated with a La Niña, as shown in the Topex/Poseidon illustration below. The purple band denotes cooled (denser) water some 18 cm (7 in) below normal heights.



After two years of wild La Niña weather in the U.S. and elsewhere - droughts in the Southwest and Southeast in 2000 and severe conditions in the Midwest and record year round heat over much of the country - at last in June of 2000 this phenomenon appears to be waning, and probably disappearing. Note that the large purple patch in the above image has now dissipated into several smaller, discontinuous blue patches in the June 2000 scene below.



At present there is an informative Web site sponsored by NOAA that offers an overview of El Niño and daily to monthly reports on their stage of development and related activities. You can access this site through the link NOAA-El Nino. You can access data on El Niño, as monitored by TOPEX-Poseidon (see next page), through this JPL site

14-28: **What can we do now to stop or control the Niño(a)'s? ANSWER**

PREVIOUS RST NEXT

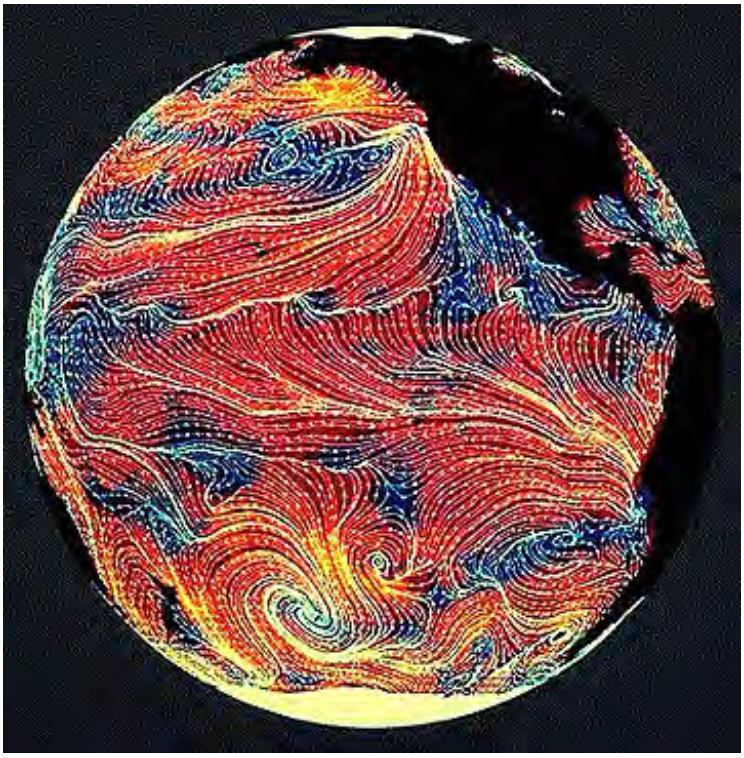
Primary Author: Nicholas M. Short, Sr.

Other satellites that have provided oceanographic data are Seasat, ERS-1 & ERS-2, TOPEX-Poseidon, QuickScat, CZCS, SeaWiFS, MOS, ADEOS, SMOS, and Aquarius. The first three are radar-based observers. Each of the eight provides important information on various aspects of sea state conditions including wind direction and velocity which helped to forecast wave conditions. The most advanced of these was ADEOS, a Japanese-directed satellite that had many sensors; unfortunately, a malfunction put it out of commission after only two months.

Seasat; ERS-1 and ERS-2; TOPEX-Poseidon, Jason-1; NSCAT; Quicksat (SeaWinds); ADEOS; SMOS; Aquarius

This page considers a group of satellites dedicated to water studies - mainly in the oceans but several also look at the land.

Temperature variations are major factors in the development, strength, and directional behavior of moving atmospheric gases (winds). Their prevailing motions change over time, but tend to follow certain preferred paths in various parts of the world. Wind directions are determined indirectly by relating them to the patterns of waves they produce, especially in the open seas. Data from the scatterometer on Seasat helped to generalize wind patterns over the Pacific Ocean, shown in this image:

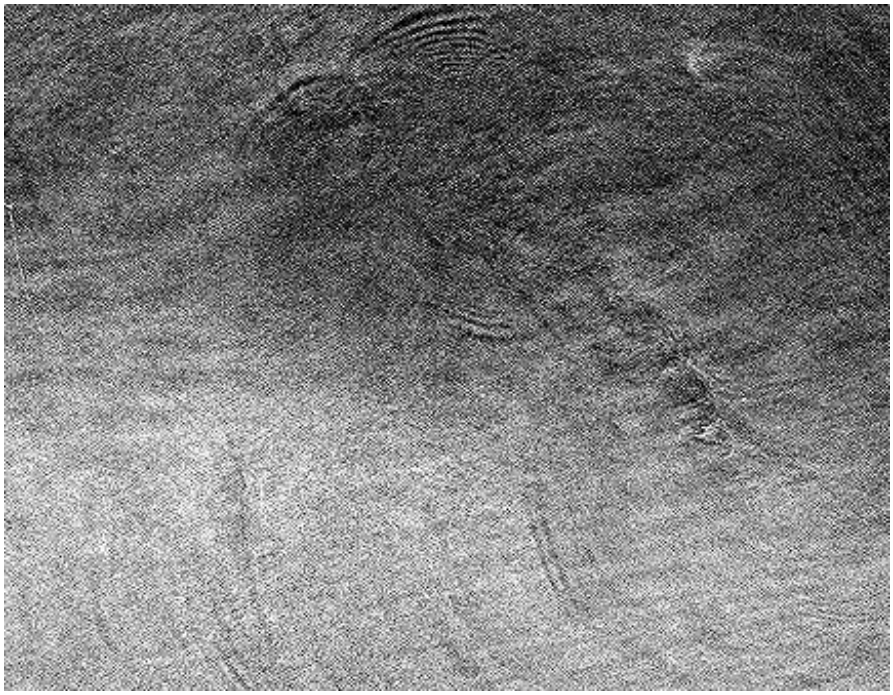


14-29: **Why do we want to know about wind patterns? ANSWER**

Seasat was the first U.S. satellite that, as an original intention, had as its primary mission the monitoring of oceanic surface phenomena, such as sea state (surface wave parameters, including wavelength, period, and height), surface wind fields, internal waves, currents and eddies, and sea ice characteristics. It was launched on June 28, 1978. The radar system on Seasat was described on page 8-6. Radar images from Seasat over land surfaces proved so

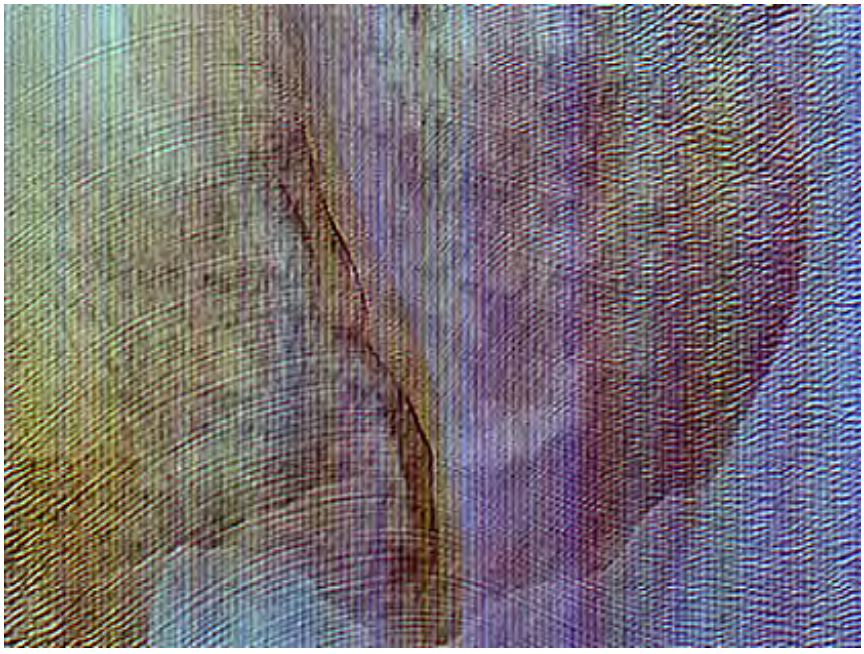
invaluable that the impression may be that the marine applications were subordinated. Nevertheless, the system performed well over water as intended and scientists gleaned much information about the oceans during its 98 days of service in 1978, before an electrical short circuit disabled it.

Operating at a high depression angle (67 to 73°) to optimize wave detection, the L-band radar on Seasat (with its 25-m resolution) provided some exceptional images of ocean waves. Here is an enlarged image of the North Sea to illustrate this (the image requires careful inspection to see the faint but widespread wave crests):



14-30: **How many directions of advancing waves can you detect in the above image?**
ANSWER

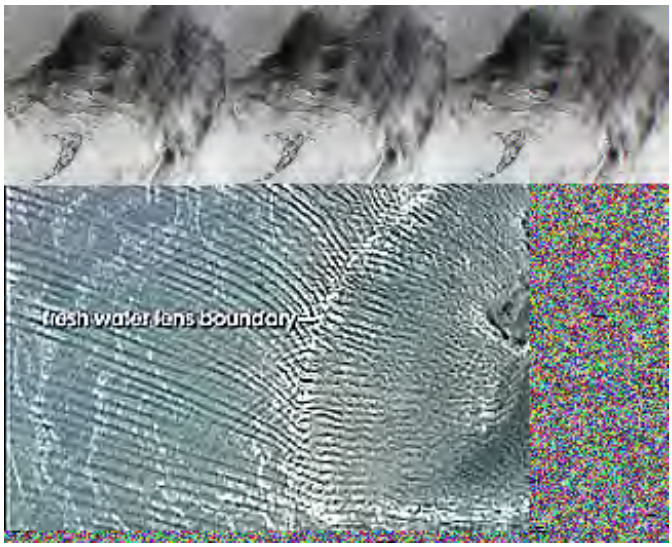
Intersecting waves are strongly expressed in this Terra ASTER (sensor) image of waters in the Bay of Bengal east of India.



Sometimes conditions are favorable for showing prominent waves and their interactions. A reprocessed astronaut photo taken from the International Space Station contains several sets of conflicting waves and an oceanic stream current passing through them off land in the Gulf of

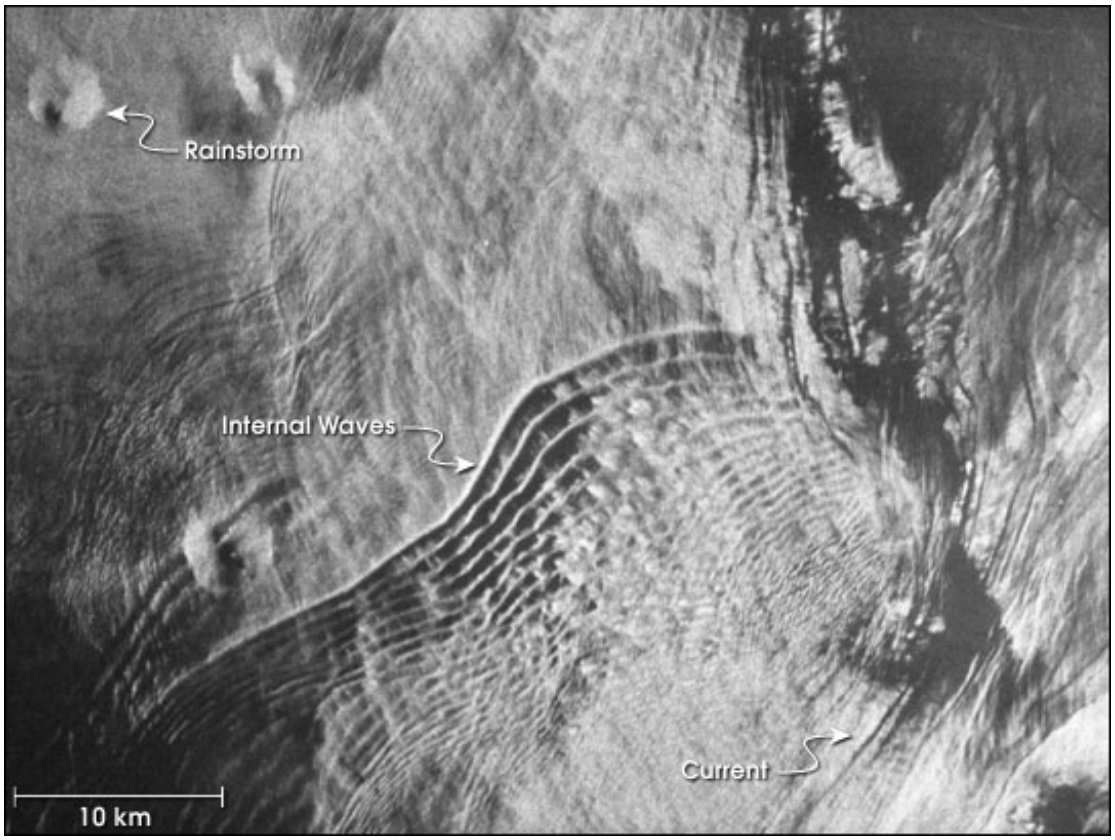
California:

Surface wave expressions related to variations in bathymetric depths are evident in this Seasat radar image (north to upper left), showing the Nantucket Shoals off Rhode Island. Nantucket Island is at the bottom left.



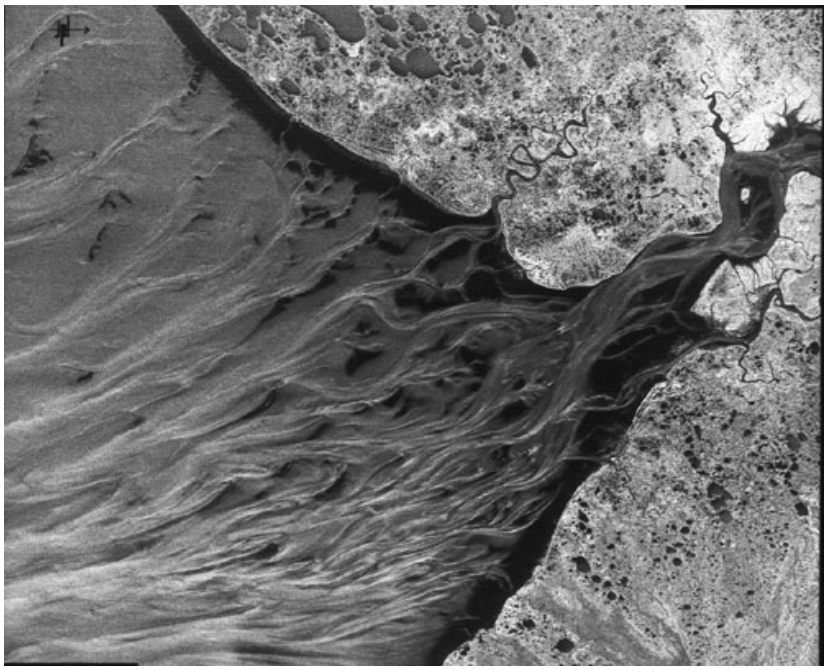
Internal waves, such as those shown in the Sulu Sea (Pacific Ocean), result when a subsurface layer of water with different density and temperature is moved by currents against the upper layer. No surface crests are formed but the motion produces an interference effect visible as dark bands.

Seasat was effective in using radar to image internal waves:

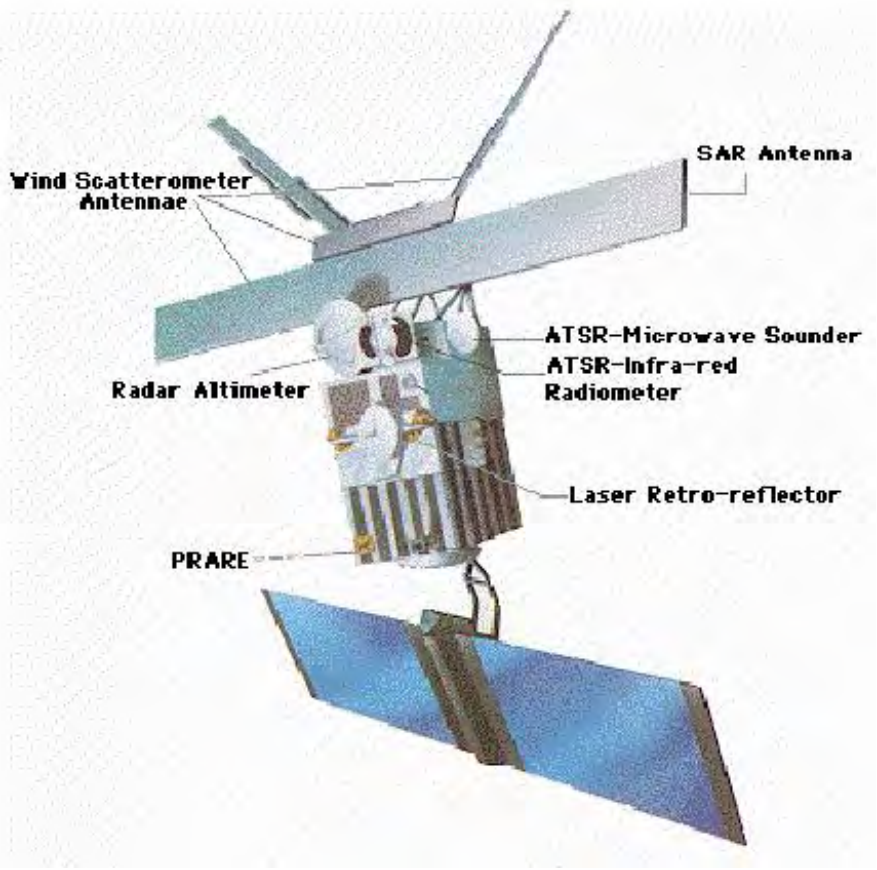


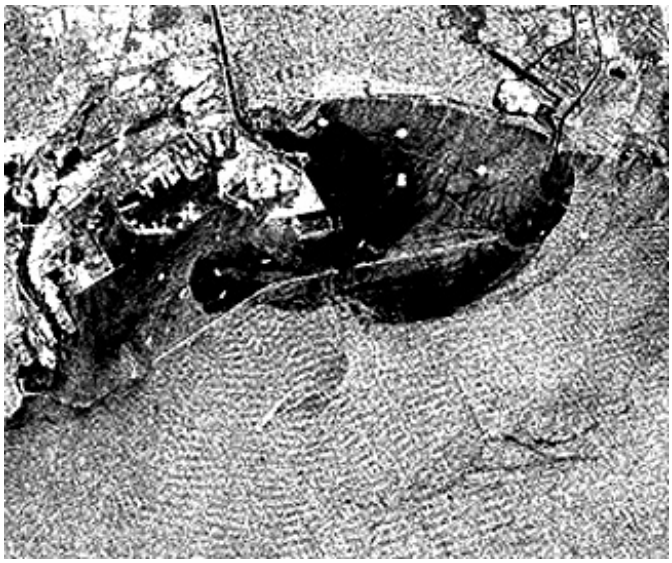
An interesting phenomenon known as "wake effect" is evident in this SeaWiFS (see next page) image of the Windward Islands in the eastern Caribbean. Prevailing winds from the east encounter topographic highs (mostly volcanic peaks) and are disrupted and slowed down. This leaves calmer surface winds, and much reduced wave heights, on the leeward side. Sun glint helps to emphasize the contrast.

Inserted at this point - somewhat out of context - is this Seasat image of fluvial (Kuskokwim River) current effects in marine waters off the west coast of Alaska:

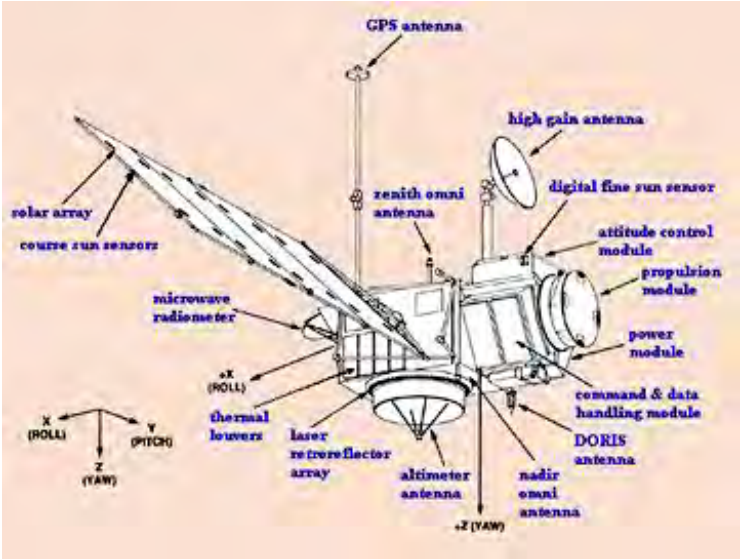


The European Space Agency launched two satellites hosting a SAR system capable of both oceanographic and land surface imaging. ERS-1 was placed in orbit in 1991; ERS-2 in 1995. Both were described in Section 8 on radar. Here we show the ERS-2 spacecraft and its instruments and two radar images that show off its marine observation capability:


SAR Antenna
Wind Scatterometer
Antennae
ATSR-Microwave Sounder
ATSR-Infra-red
Radiometer
Radar Altimeter
Laser Retro-reflector
PRARE



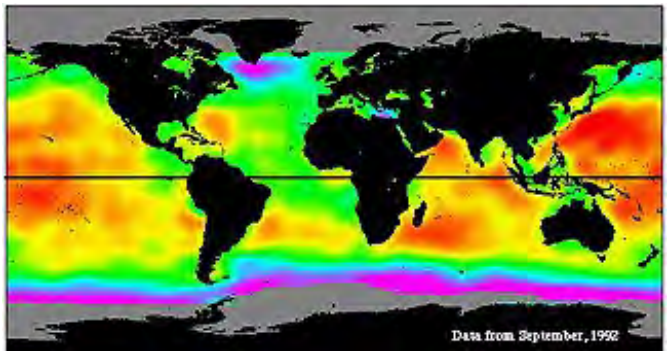
We first described the TOPEX-Poseidon (T/P) mission, run by JPL and launched on August 10, 1992, in page 8-7. We suggest you access this (outstanding tutorial) prepared by the TOPEX-Poseidon team, which explores the kinds of information that radar altimetry and scatterometry can acquire. Here is a sketch of T/P's instruments:



TOPEX-Poseidon uses two radar altimeters to measure distance from the satellite to the sea's surface (to a precision of 4.3 cm). From such data, global maps are derived to show Dynamic Ocean Topography (rises ["hills"] and depressions ["valleys"]), Sea Surface Variability, Wave Heights, and Wind Speeds. A separate microwave radiometer on this mission determines Precipitable Water Vapor, which allows for corrections in the pulse transit time to improve distance accuracy.

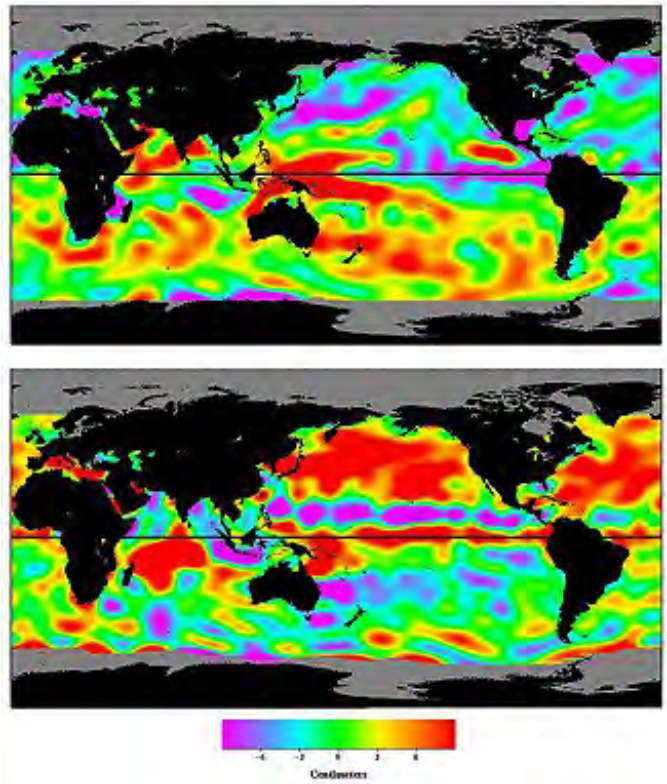
14-31: **How would T/P determine wind speeds from the observables? ANSWER**

The first TOPEX-Poseidon maps we examine show ocean topographic variations. Below is a plot of large-scale variations during September 1992 in ocean elevations relative to the Earth's geoid. The data used to construct the map come from other sources, such as gravitational perturbations of satellite orbits. Note that as much as 150 m (492 ft) of relief exist on the seas between the Atlantic Ocean (lower, green) and the western Pacific (higher, orange-red).

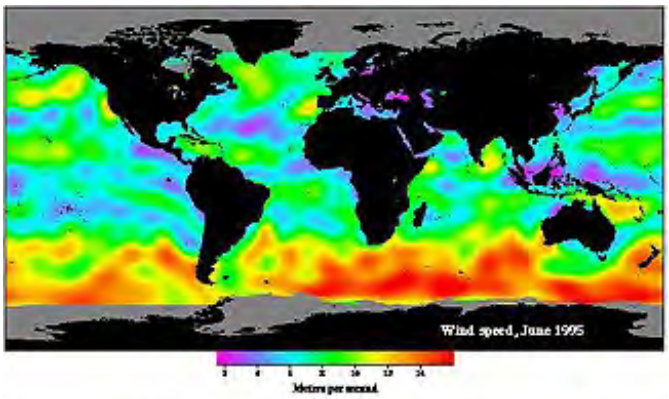


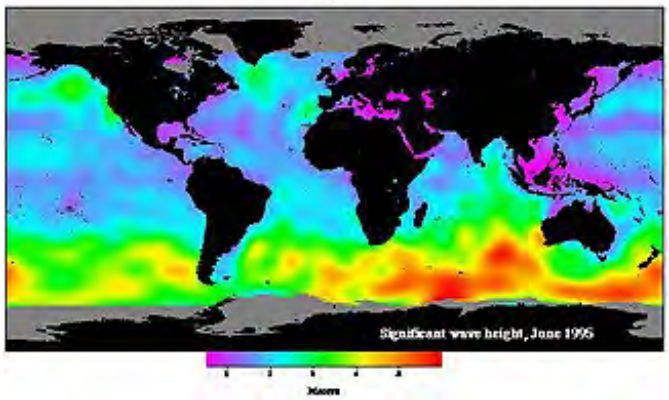
14-32: **Where are the oceans the highest?** **ANSWER**

Seasonal differences do occur because of departures from the norms in height, as water expands or contracts in each hemisphere. The pair of maps below show Sea Surface Height (SSH) variation in the northern Fall (upper) and Spring (lower), in a plot centered on the Pacific Ocean. The differences in elevation lead to redistribution of water through thermally-induced currents.

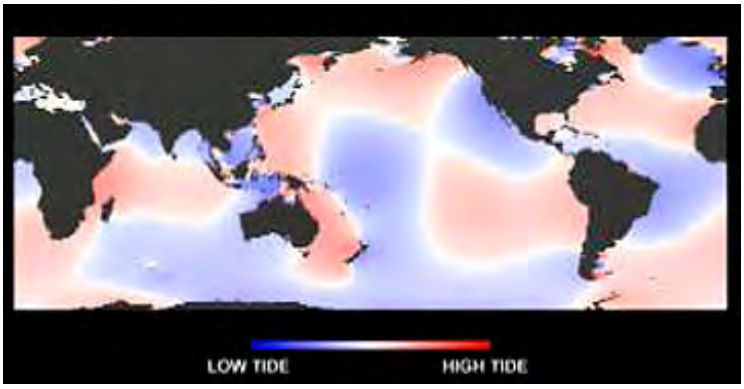


The next pair of maps are plots of wind speeds and wave heights during June 1995. There is some correlation between these parameters: greater speeds tend to cause higher waves.

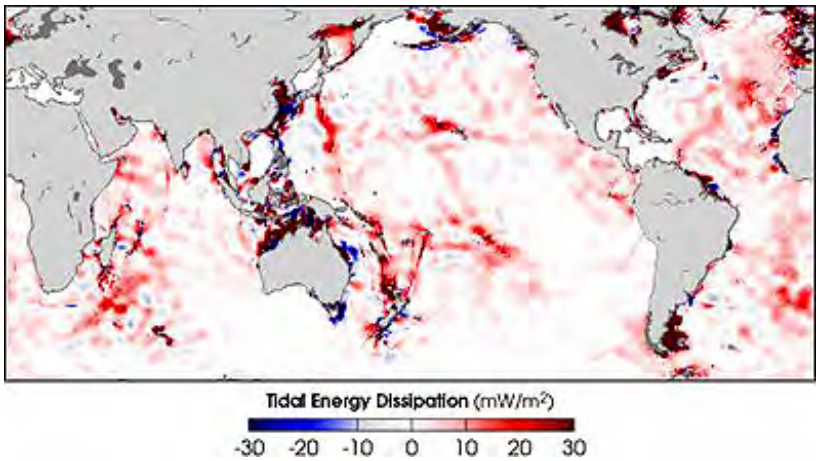




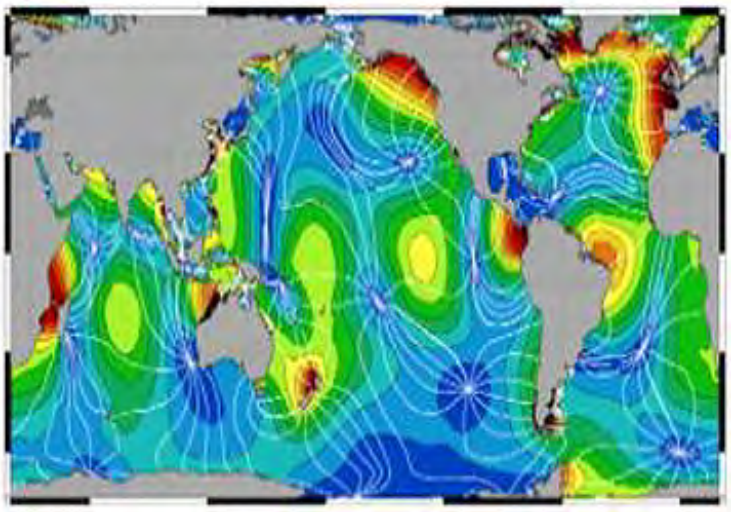
TOPEX-Poseidon also has shed new light on the oceans tides. There has been an ongoing mystery as to balancing the energy provided by the Moon's gravitational attraction, which produces the tides, and the dissipation of that energy. What was known is that much of the energy goes into setting up surficial ocean currents that carry water from higher areas to lower. Ocean heights measurements by T-P have now better fixed the areas of the seas that are higher and lower than mean sea level. This image shows a general pattern:



This image shows more details, representing a data sets for six years of observations. The so-called tidal energy dissipation thus displayed is affected in part by variations in sea surface heights which establish gradients that cause water flow that influences tidal rises or falls (see page I-1b.

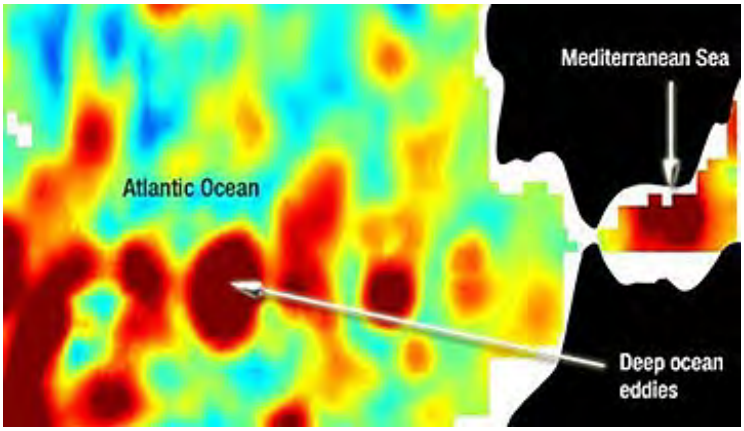


From this can be derived this broad outline of the current flow lines outward from the highs:



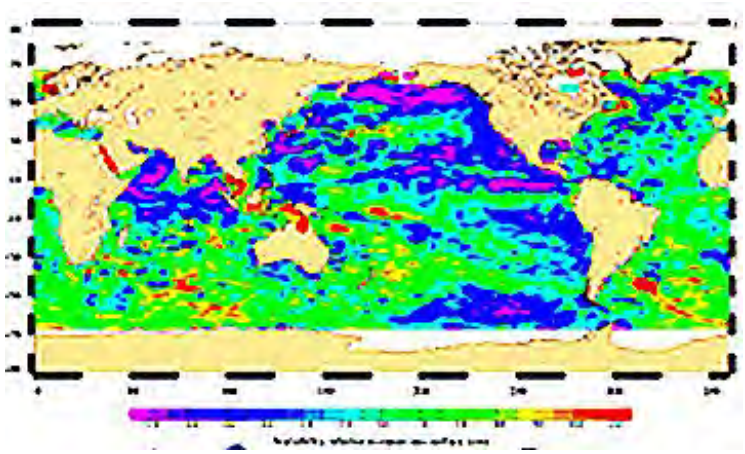
The shortfall of about 30% for the accounting of energy balance in tidal energy distribution has from the T-P observations now been explained by work done in changing ocean volume below the surface. The configuration of the sea floor contributes to this.

One subsurface feature that T-P and subsequent satellites such as Jason and Envisat could detect with their radars is thermally-active whirlpools or mega-eddies (about 100 km wide; several can coalesce) that have maximum spiral circulation at depths below 600-1000 meters (but much less near the surface). The effect diminishes surfaceward but can be identified because surface waters are pushed upwards, so as to have a positive topographic expression. These mega-eddies have 2 to 4 times the salt concentration of normal ocean waters; the resulting increase in density accounts for their activity. Here is an example in the Atlantic (off the Mediterranean Sea where such a feature is commonplace since that Sea has higher average salt concentrations.):

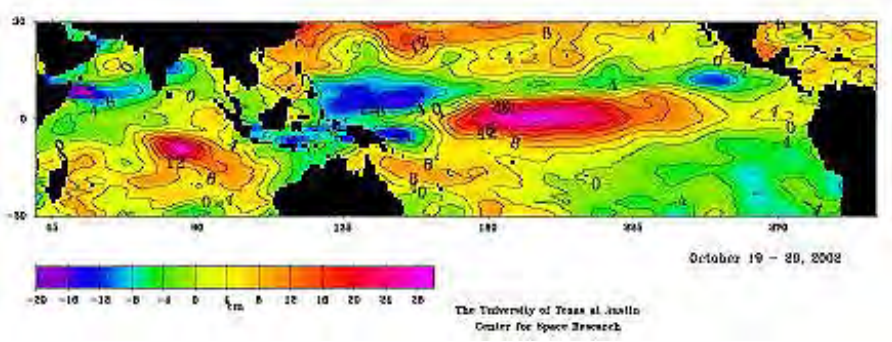


Topex-Poseidon was decommissioned (ceased data gathering) on January 6, 2006.

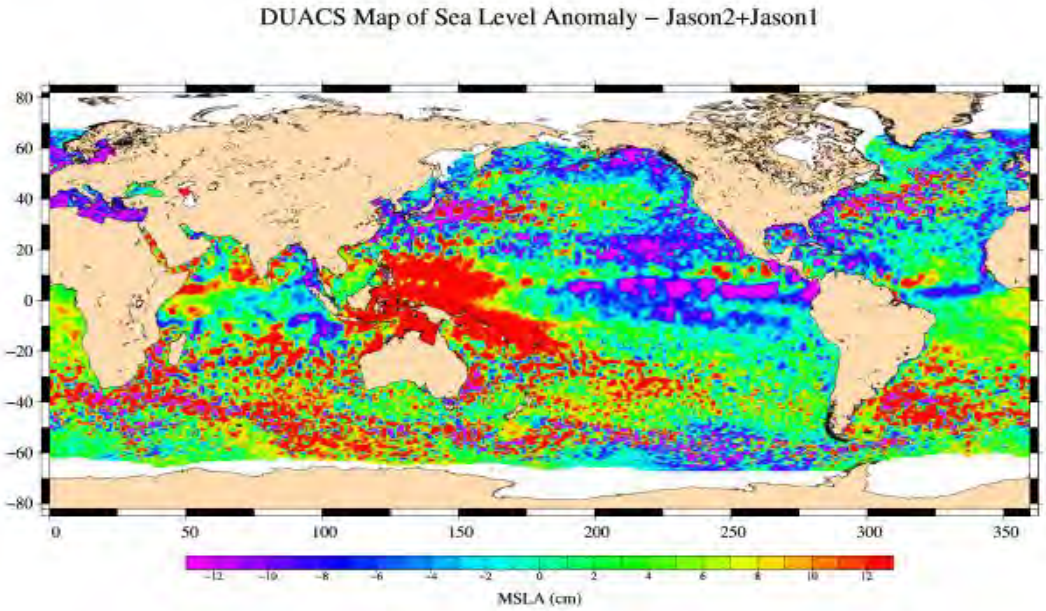
A follow-up to TOPEX-Poseidon, named JASON-1, is a component of the EOS program (see Section 16). Operated jointly by NASA JPL/CNES, this spacecrafts sensors include C and Ku band radar altimeters, a microwave radiometer, and a Doppler radar. Again, sea surface heights (SSH) are the main oceanographic phenomenon being measured; for Jason, differences in SSH as small as 4.1 cm can be determined.



Data from Jason-1 are now being processed routinely by various government and private centers engaged in marine studies. This map of global tropical zones was produced by the Space Center at the Univesity of Texas-Austin.

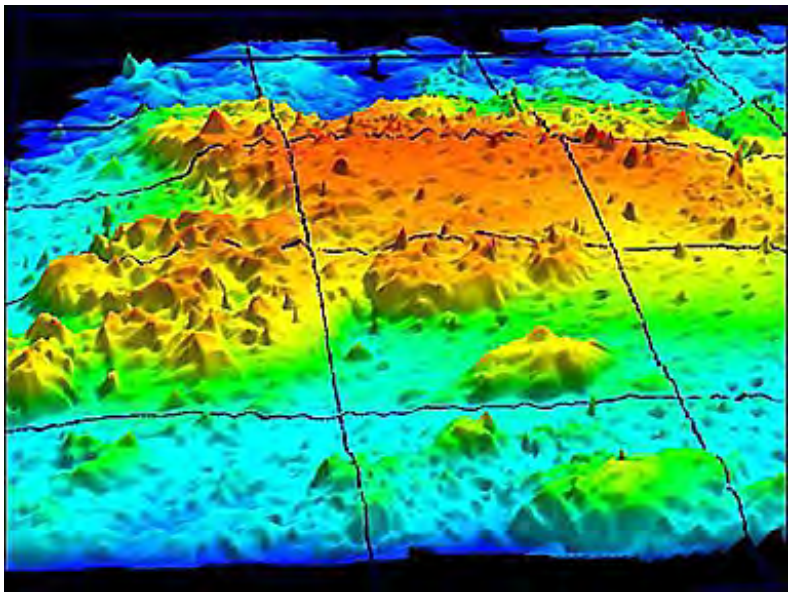


Jason-2 was launched in June of 2008. A map using both Jason-1 and Jason-2 shows variations is sea surface heights to a sensitivity of +/- 3 cm.:

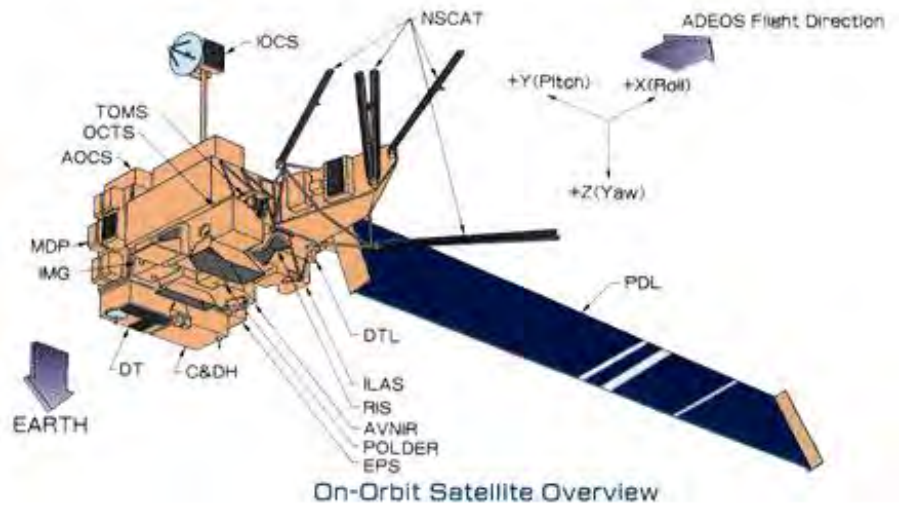


One aspect of oceanographic studies and management that has received considerable attention lately is the state of health of corals - the animals (polyps) that make the foundation of coral reefs which support a wide variety of biota. Various satellites are contributing to an organized monitoring program designed to gather long term data and to "flag" potential local to worldwide conditions that threaten coral populations. Landsat 7 is a mainstay of this effort. Here is a Landsat false color view of part of the Florida Keys (see also page 8-6) which illustrates the monitoring capabilities from satellites.

Lidar mounted on aircraft have proved capable of determining the shape, and to some extent, depth of coral reef complexes off island shores. Here is a relief image of Johnson's Reef off St. Johns Island in the American Virgin Islands:



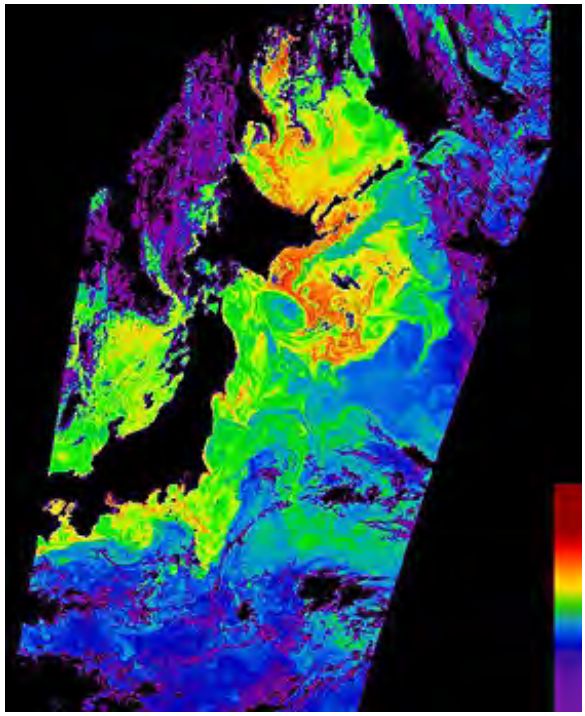
Turning now to the Japanese ADEOS (Midori): ADEOS was launched from Japan on August 16, 1996, and operated successfully for 9 months. The view below shows the sophisticated nature of this satellite:



On-Orbit Satellite Overview

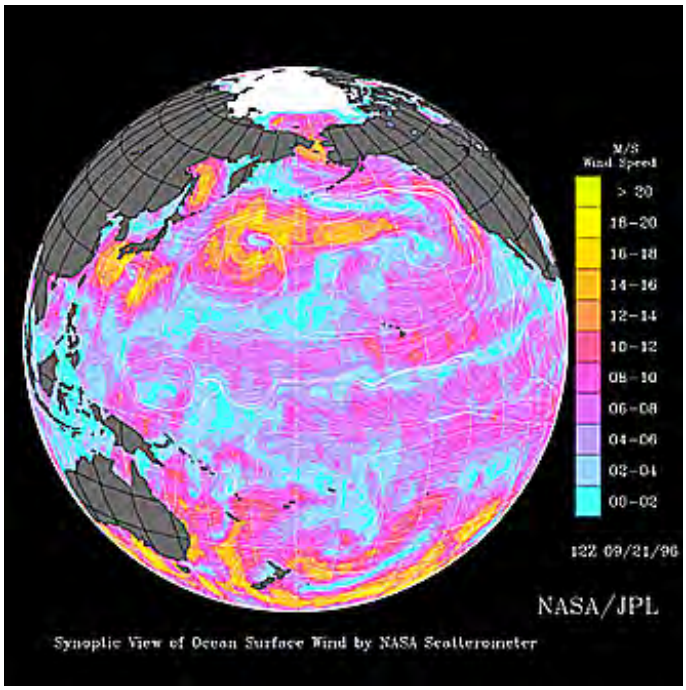
Here is a Vis-NIR imager (called AVNIR) image of some of the island reefs in the Great Barrier Reef of northeast Australia.

ADEOS also produced some interesting ocean color images, none more so than this view of the sea around Japan itself, here with the natural colors modified to the full range of blues through reds to emphasize currents and other water circulation patterns:

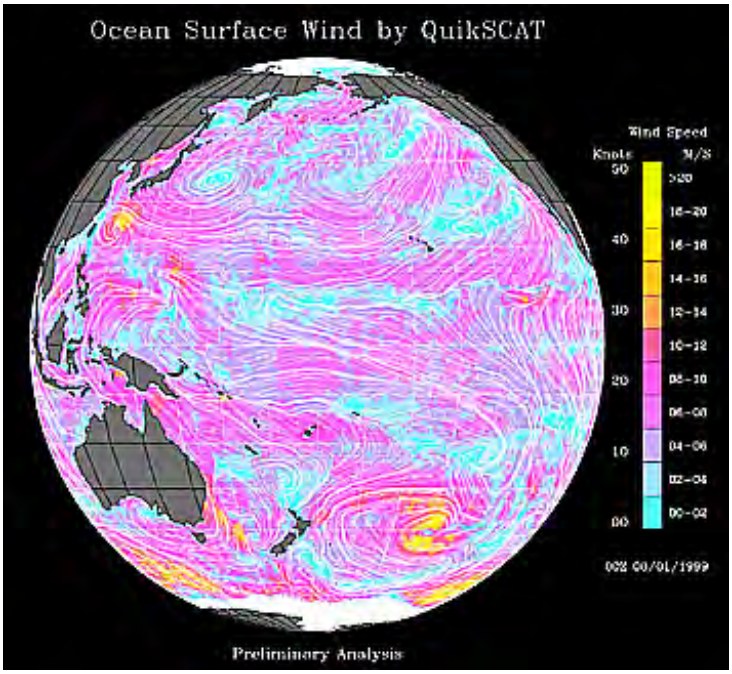


The Jet Propulsion Laboratory (JPL) developed, in cooperation with Japan's NASDA, a follow-on instrument to the Seasat Scatterometer, called NSCAT (NASA Scatterometer) which first flew on ADEOS. NSCAT was a microwave radar scatterometer that transmitted pulses at 13.995 GHz and measured their reflections (backscatter) from ocean ripples and other surfaces. Six antennas provided eight beams that extended over two wide bands. The returned signal is subjected to Doppler processing. In general, rougher seas correlate with greater wind speeds; in turn, rough seas increase backscatter. Operating at 50 km spatial resolution, the system distinguished and recorded 268,000 wind vectors, derived from wave-induced backscatter data.

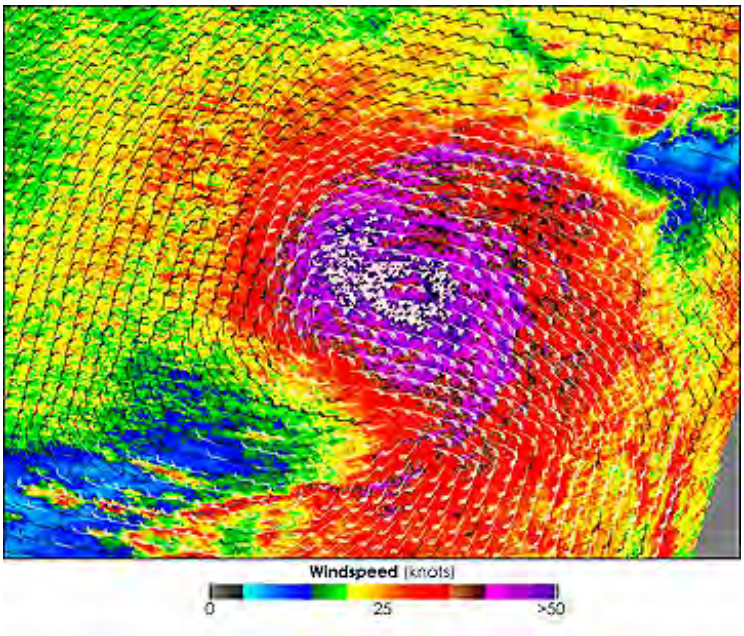
The principal use for NSCAT was in determining wind-direction, from which is derived useful information relative to ocean waves and global climate patterns. The instrument could also operate over land (see page 3-5) and produce sea-ice images (page 14-14). Here is a global, wind-vector map, color coded for grouped intervals of different velocities.



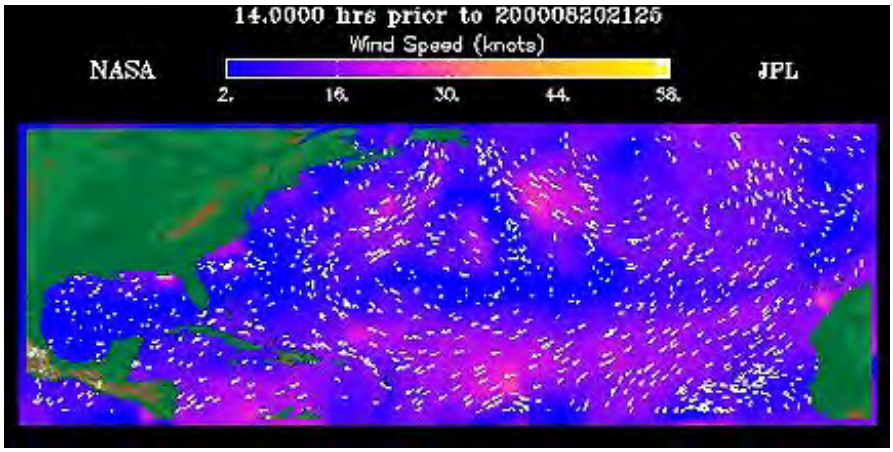
On July 19, 1999 NASA JPL launched QuickScat, a satellite whose prime sensor (SeaWinds) is a radar with 25 meter resolution. Its primary mission is to provide near-realtime measurements of surface roughness that translate into wind velocities.



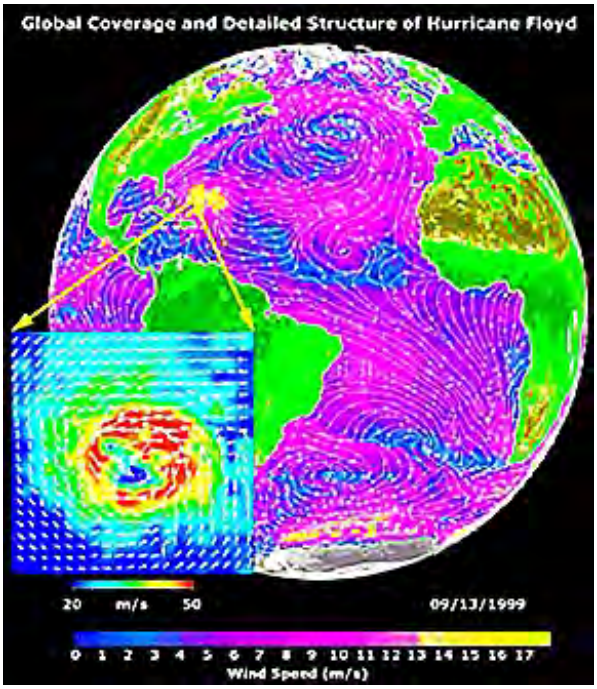
A detailed map of wind speed intervals associated with Typhoon Nesat in the Pacific, on June 6, 2005, shows the details obtainable with the SeaWinds instrument



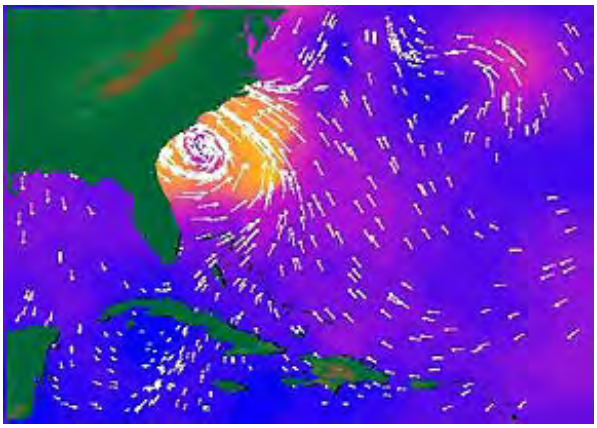
Below is a map of Hurricane Alley in low latitudes of the Atlantic, on which color shading indicative of wind speeds and vectors showing prevailing wind directions are plotted.



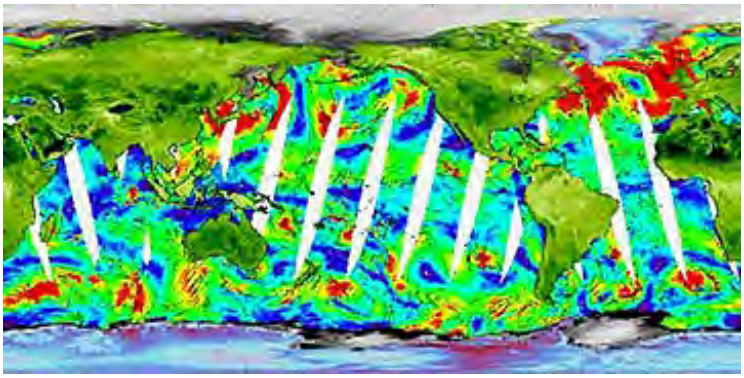
And here is a QuickScat SeaWinds map that includes an actual hurricane, Floyd, a destructive one that hit the southern U.S. in September, 1999; the immediate area of the hurricane appears as an expanded inset:



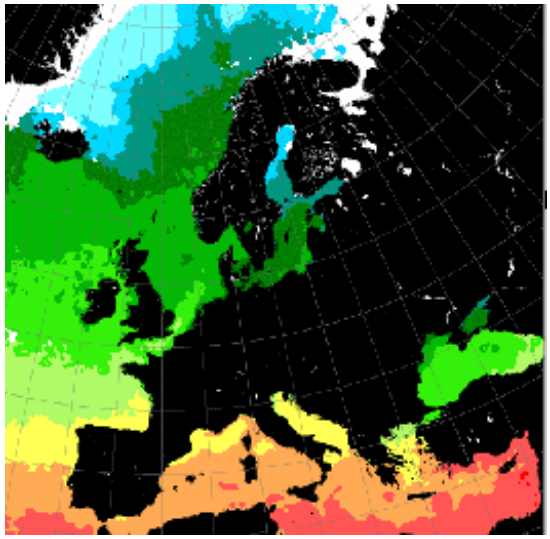
That localized area is shown in more detail in this SeaWinds map of wind flow in the Gulf of Mexico and the Atlantic during Hurricane Floyd on September 16:



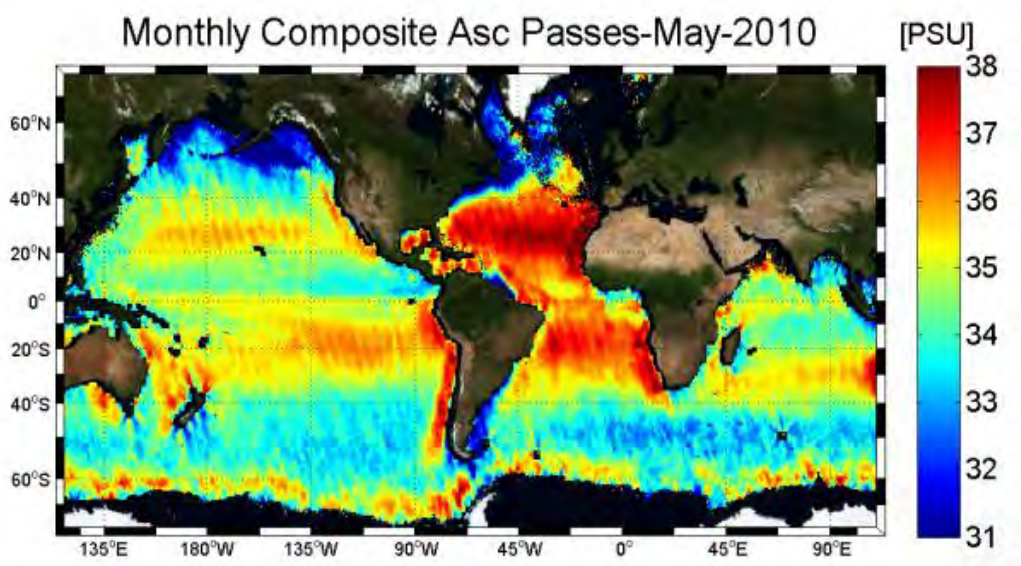
NASA sponsored another SeaWinds radar scatterometer as an instrument package onboard ADEOS-2 (renamed Midori-2) launched on December 14, 2002 by the Japanese NASDA program. Primarily an ocean surface monitor, this sensor can identify sea ice and can, like its predecessors above, determine wind velocities. Here is the first returned data set acquired on January 28-29, 2003:

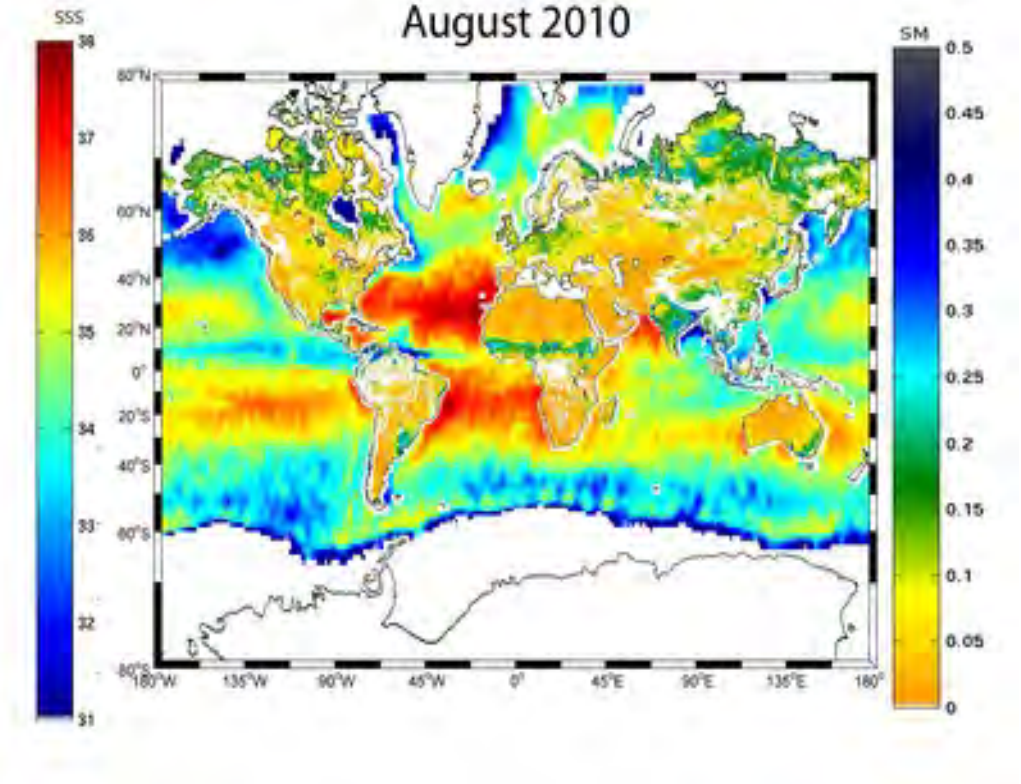


The Japanese have also launched (1987 and 1990) two Marine Observation Satellites (MOS) that include a scanning radiometer (MESSR), a Visible-Thermal instrument (VTIR), and a microwave unit (MSR). These sensors were turned on over both land and sea; they are no longer operational. This is a Sea Surface Temperature map of European waters:

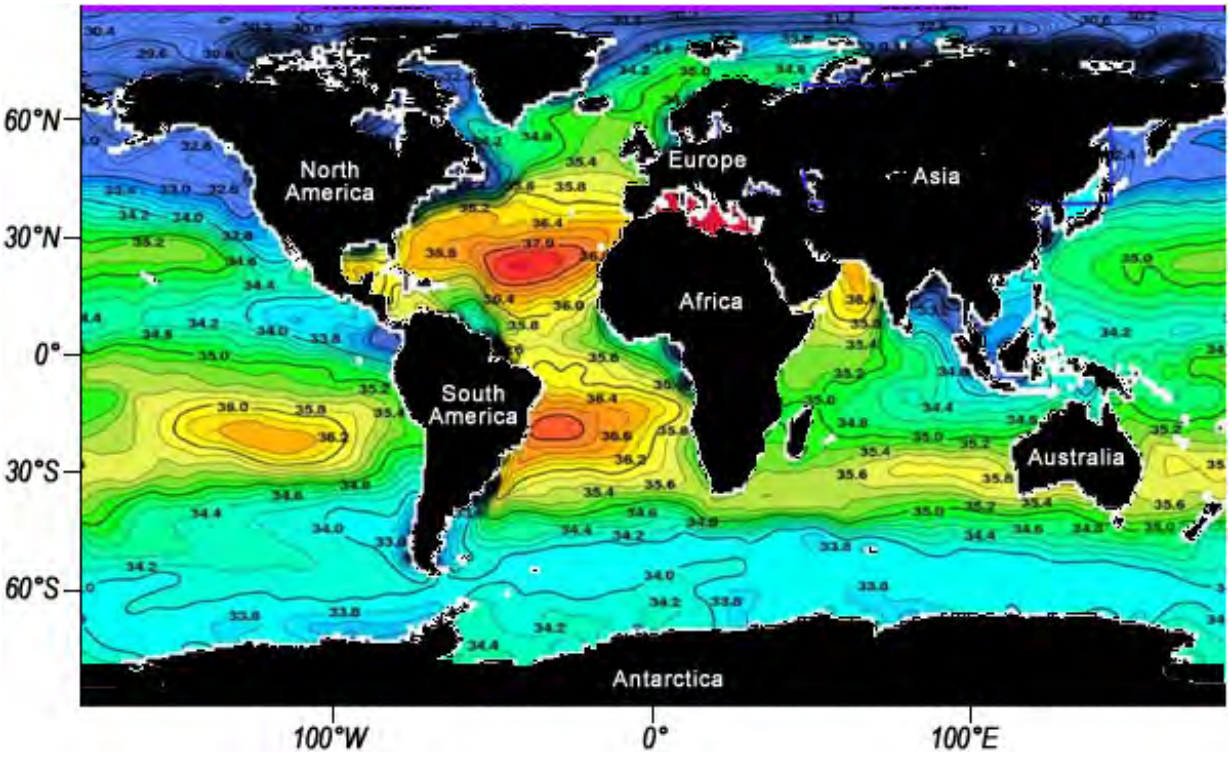


The European Space Agency launched the Soil Moisture and Ocean Salinity (SMOS) satellite in November of 2009. Soil moisture measurements are important in monitoring drought conditions. Ocean salinity is a major factor in generating and controlling ocean currents. Here are two data sets released in 2010:

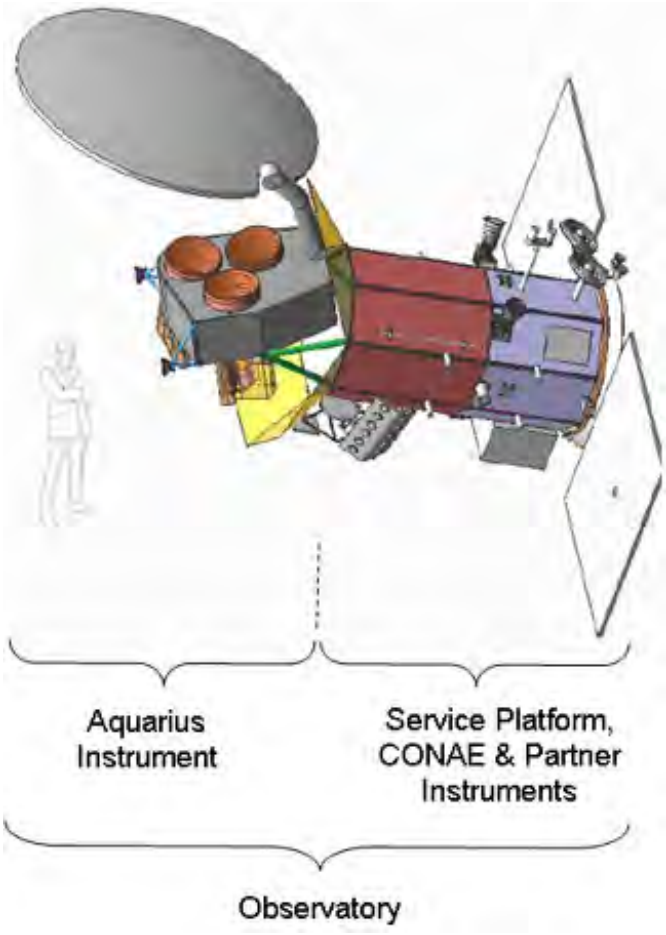




Other ocean-observing satellites are on the drawing boards and at least one is scheduled for launch. Aquarius, one planned to launch in May of 2010 and now postponed into 2011, is a joint enterprise between NASA and Argentina's space agency. Its purpose is to measure both short and long term variations in Sea Surface Salinity (SSS). Although these variations are small, they have distinct influence over such marine activities as evaporation and ocean current behavior. Prior to Aquarius, measurements of SSS are made from ocean buoys, of which there are only a few hundred emplaced. This is a typical global SSS distribution map.

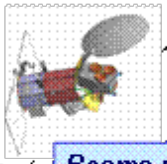


Here is a view of Aquarius, a chart showing its mission operation, and a table related to its instrumentation:



Mission Design and Sampling Strategy

Sun-synchronous exact repeat orbit
6pm ascending node
Altitude 657 km



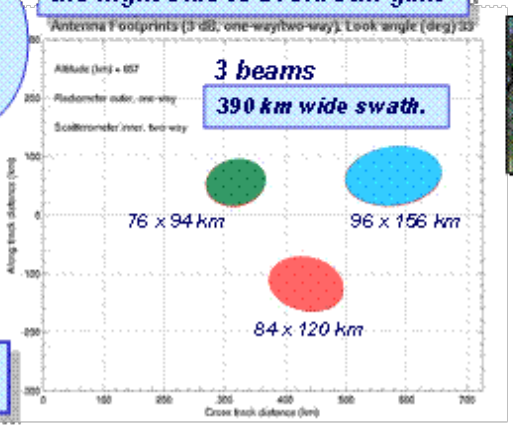
- *Global Coverage in 7 Days*
- *4 Repeat Cycles per Month*

Surface Validation

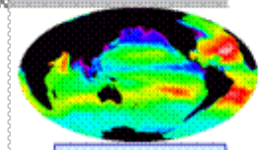
Beams point toward the night side to avoid sun glint

In Orbit Check out

Launch 2010



Aquarius Ground System



Salinity Data

Aquarius/SAC-D Science Instruments				
Instrument	**Objective**	**Description**	**Resolution**	**Source**
Aquarius:	Sea Surface Salinity (SSS)	Integrated 1.413 GHz polarimetric radiometer, 1.26 GHz radar, 390 km swath	3 beams 76x94 km, 84x120 km, 96x156 km	NASA
MWR: Microwave Radiometer	Precipitation, wind speed, sea ice concentration, water vapor	23.8 GHz and 37 GHz; Dual polarized; 390 km swath	40 km	CONAE
NIRST: New IR Sensor Technology	Hot spots (fires); Sea Surface Temperature	Bands: 3.8, 10.7 and 11.7 µm Swath: 180 km;	350 meters	CONAE
HSC: High Sensitivity Camera	Urban lights; fires, Aurora	Bands: 450-900 nm; Swath: 700 km;	200-300 m	CONAE
DCS: Data Collection System	Environmental data collection	Band: 401.55 MHz uplink	2 contacts per day w 200 platforms	CONAE
ROSA: Radio Occultation Sounder for Atmosphere	Atmosphere temperature and humidity profiles	GPS occultation	Hor: 300 km Vert: 300 km	ASI (Italy)
CARMEN 1: (ICARE and SODAD)	ICARE: Effect of cosmic radiation on electronics; SODAD: Distribution micro-particles and space debris	ICARE: Three depleted Si and Si/Li detectors SODAD: Four SMOS sensors	I: 256 channels S: 0.5 µ at 20 km/s sensitivity	CNES (France)

The chart shows that there are a variety of instruments on Aquarius. Some are "piggybacked", onboard to measure phenomena on land and sea not related to SSS. The SSS is measured using two microwave radiometers, one operating at 1.413 GHz and the other at 1.26 GHz. These give temperature information that depends in part on variations in the dielectric constant which is an indirect measure of salinity. The MWR gives information that pertains to sea state, including wave heights and wind action. This allows correction of the SSS data which are sensitive to such variables.

Information derived from the Aquarius experiment will be of significant value in determining the role of ocean water on short-term weather effects (such as precipitation) and long-term climate changes.

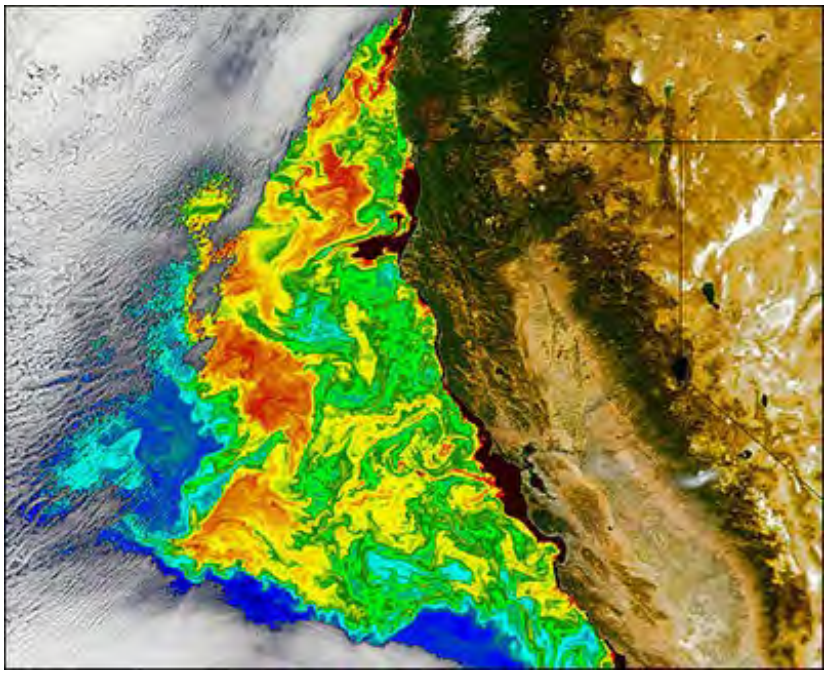
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

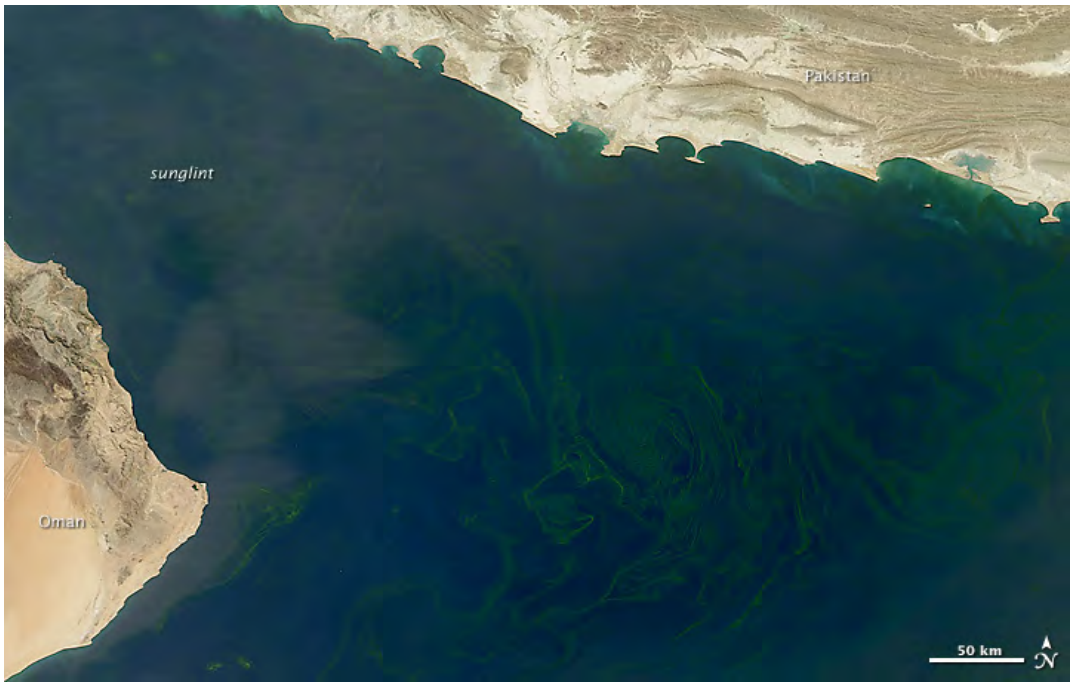
The oceans teem with life but of course fish and even whales are not resolvable by the usual satellite sensors. But sealife has to eat and at the base of their foodchain is microscopic plant and animal life known as plankton. Algae, in particular, are fundamental food (even for whales). Being marine plants, they experience the same responses to light owing to chlorophyll content (absorbing in the blue and red). Satellites can determine plankton content by analyzing the color variations (mainly in the visible) of seawater inasmuch as much of the color changes correspond to presence of life. Two satellites, the Coastal Zone Color Scanner (CZCS) on Nimbus 7 and SeaWiFS provide valuable data on ocean color. This translates into practical information for the fisheries industry. Satellite sensors also are very effective in scoping the patterns of sediment distribution, using mainly the Vis-NIR bands.

CZCS and SeaWiFS

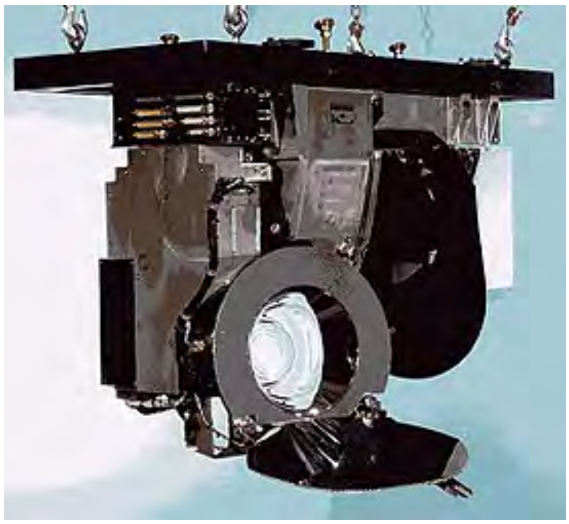
The world's oceans (and its freshwater bodies, as well) are replete with life. Central to the marine foodchain is phytoplankton - microscopic plants that photosynthesize chemicals in sea water. This process depends on the plankton content of chlorophyll-a, a pigment that strongly absorbs red and blue light. As plankton concentrations increase, there is a corresponding rise in spectral radiances, peaking in the green. Upwelling masses of water (usually associated with thermal convection) containing phytoplankton take on green hues in contrast to the deep blues of ocean water with few nutrients. Plankton occur over the entire ocean expanse but tend to concentrate off continental shores because of the abundance of nutrients coming from the land. Here is a map of plankton distribution off the California coast:



Although often subtle in appearance, phytoplankton blooms can be seen in natural color images. This MODIS image of the Arabian Gulf shows a bloom consisting mainly of Noctiluca miliaris, which forms there mainly during winter months.



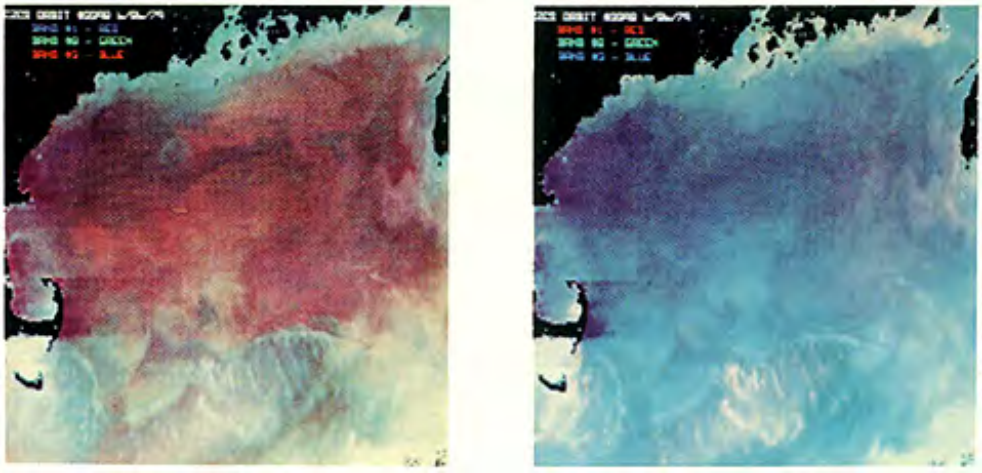
The Coastal Zone Color Scanner (CZCS) is a sensor specifically developed to study ocean color properties. These properties can be related to organic content, such as plankton, as well as sediment. CZCS launched in October 1978, as part of Nimbus-7's instrument complement and continued to operate until late 1986. It sensed colors in the visible region in four bands, each 0.02 µm in bandwidth, centered at 0.44 (1), 0.52 (2), 0.57 (3), and 0.67 (4) µm. A fifth band between 0.7 and 0.8 µm monitored surface vegetation and band six, at 10.5-12.5 µm sensed sea surface temperatures. Here is the first CZCS before fitting onto the Nimbus spacecraft:



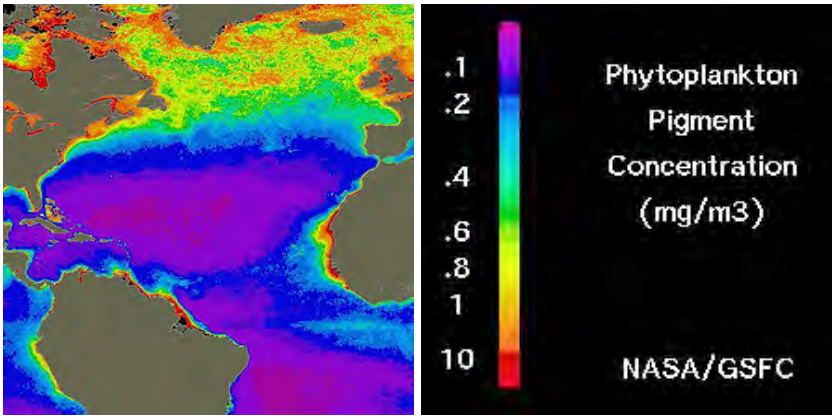
In monitoring ocean color, band 1 (blue) measured chlorophyll absorption; band 2 (green), tracked chlorophyll concentration; band 3 (yellow), was sensitive to yellow pigments ("gelbstoff"); and band 4 (red), reacted to aerosol absorption. With data from those bands, we calculate the chlorophyll variation, which correlates closely with relative abundance of marine phytoplankton, as the ratio of band 1 to band 2 (for phytoplankton concentrations less than 1.5 mg/m^3) and band 2 to band 3 (>1.5 mg/m^3)

A good primer for how to use CZCS and similar data sets is found in this (somewhat outdated) Goddard User's Guide. Additional information is available by clicking on Ocean Color. Some of the more informative CZCS scenes are found in this Goddard review.

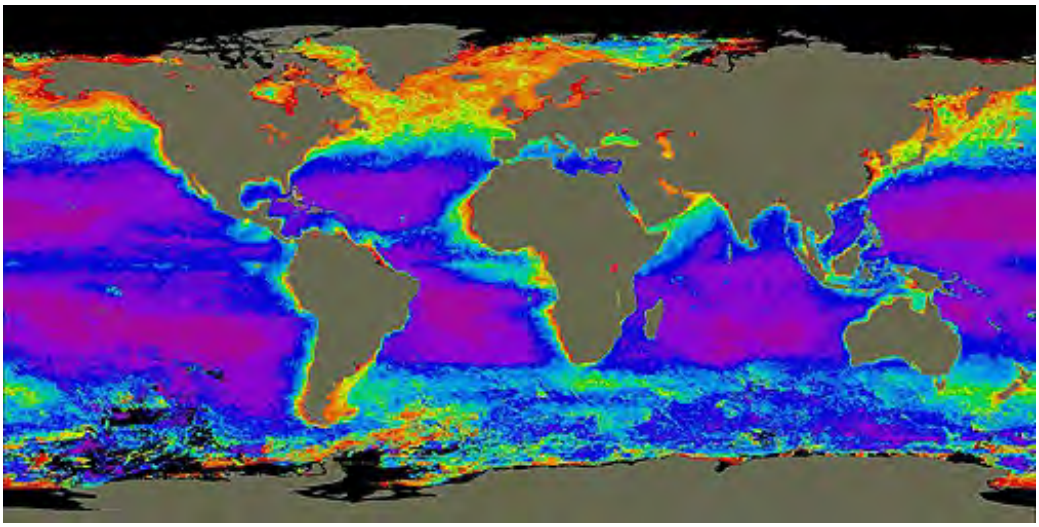
The next pair of images are CZCS color composites: the left one emphasizes chlorophyll-enrichment (in reds) in the Georges Bank off the New England coast (BGR = bands 3,2,1) whereas the right one simulates natural ocean color (BGR = bands 1,2,3)



Using data from all relevant bands, here are chlorophyll concentrations spread over the Fall to Spring season in 1978-79 in the Atlantic Ocean:

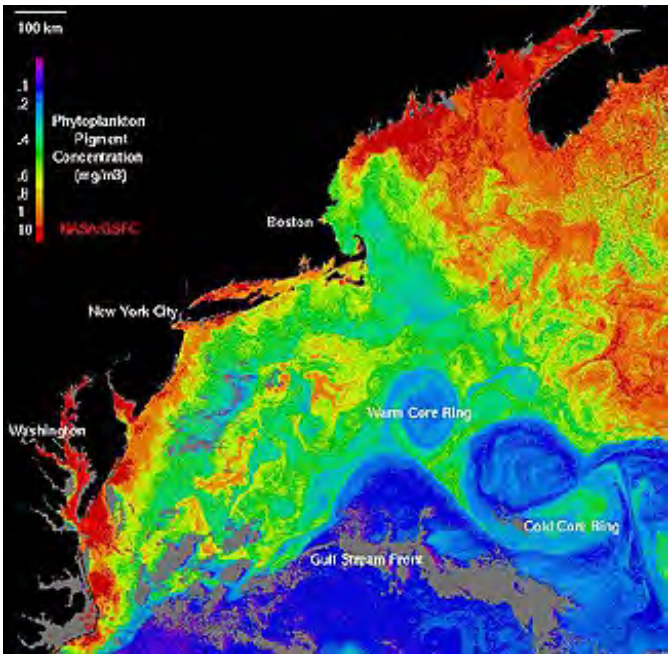


The distribution of chlorophyll on a global scale averaged between 1978 and 1986 appears here:

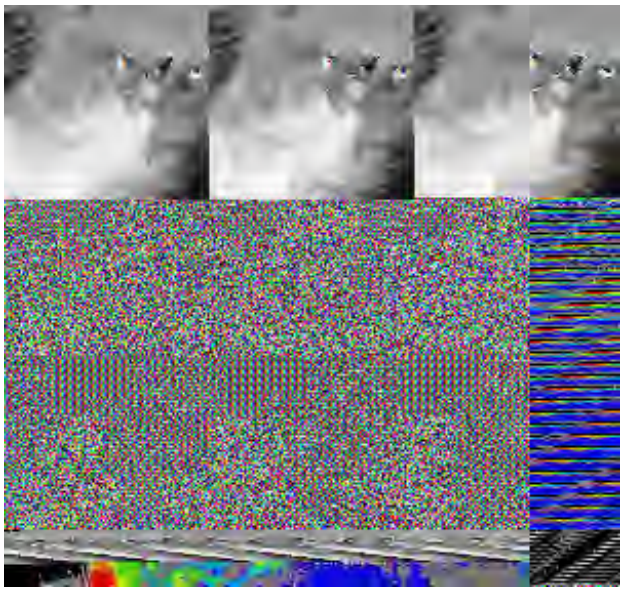


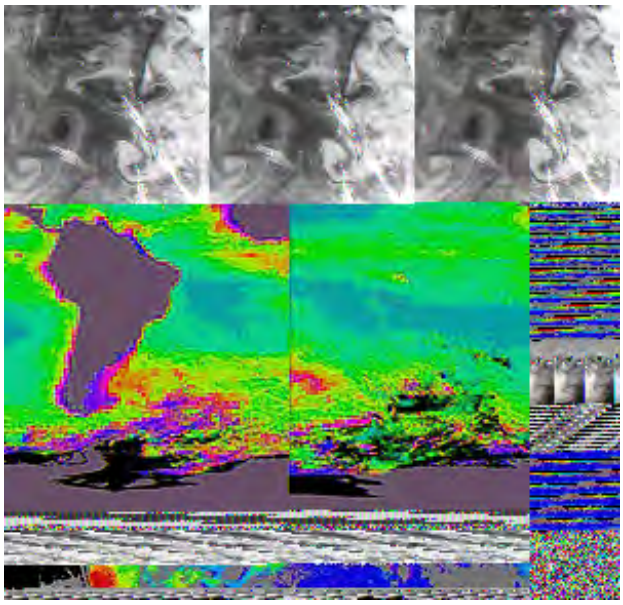
14-33: **These chlorophyll distribution maps show concentration gradients that seem to run counter to logic (or intuition), in that the lowest amounts of chlorophyll (hence plankton) are in the equatorial and subtropical zones. Those zones, being warmer, should support greater quantities of plant life (as do the tropics on land). What gives here? ANSWER**

We can correlate thermal data from CZCS with chlorophyll, as demonstrated in this plot:



In the color coding, blues correspond to the lowest levels of phytoplankton and reds to the highest. Thermal data as such are not directly contributing to the color composite of this image but notations about warm and cool areas in the water result from thermal data from another satellite. Note the eddies or rings. Phytoplankton tends to concentrate along the edges of warm core rings (which rotate clockwise) but concentrate centrally in cold core rings (counterclockwise motion). The motions in these rings are analogous to circulation around atmospheric high- and low-pressure systems. Warm core rings can extend over several hundred kilometers, as seen in this HCMM Night-IR thermal image (top), which shows such "gyres" in varying stages of development and coherence in Atlantic waters near the Canary Islands off the African coast. Cold core rings are usually less well shaped but are probably displayed in this June 19, 1976, Landsat-1 band 4 (green) image (bottom) of waters off the southwest coast of Iceland; concentrations of phytoplankton are associated in part with the lighter tones that may also be tied into sediment "murkiness".





14-34: **Which way do the eddies seem to be rotating in the above two images? ANSWER**

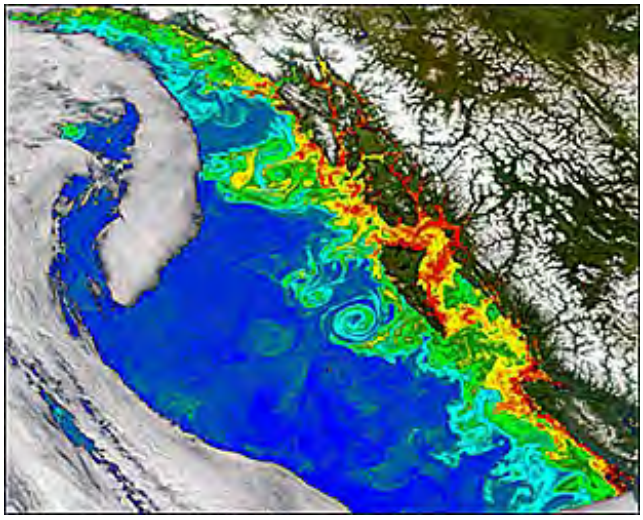
These observations types - ocean color and thermal patterns - aid in locating conditions where large schools of fish are likely to live, so commercial and sport fishermen actively apply them to locate the best current fisheries.

The follow-on to the CZCS is NASA Goddard's SeaWiFS (Sea-viewing Wide Field-of-View Sensor), launched successfully on August 1, 1997. It is the prime sensor on a commercial satellite, Orbview-2 (originally named SeaStar) operated by the Orbital Imaging (OrbImage) Corporation. Once again, the sensor system monitors ocean color variations, especially those caused by concentrations of plankton and other sealife that strongly moderates chlorophyll response, detectable spectrally. Thus, the prime objectives are: 1) to quantify ocean plankton production; 2) to determine observable couplings of physical/biological processes; and 3) to characterize estuarine and coastal ecosystems. We show the prime sensor here:

The SeaWiFS sensor consists of eight channels at: 412, 443, 490, 510, 555, 670, 765, and 865 nm (nanometers: 1µm = 1,000 nm), each with bandwidths of 20 or 40 nm. The instrument can swing ±58° off nadir. From an orbital altitude of 705 km (438 mi), spatial resolution in the Local Area Coverage (LAC) mode is about 1.1 km (0.68 mi) (the optimal resolution is 0.6 km at nadir), and in the Global Area Coverage (GAC) mode, it is 4 km (2.5 mi). Swath widths are: LAC = 2,801 km, and GAC = 1,502 km.

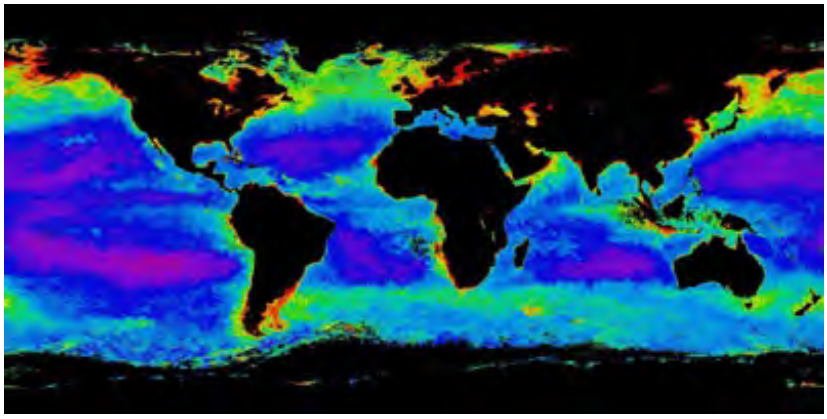
Like CZCS, SeaWiFS produces regional scale images in which eddies and circulation patterns are evident. In this view of the western North American continent, marine eddies have formed off the

British Columbia coast around Queen Charlotte; to the west is still another eddy-like pattern made by clouds.

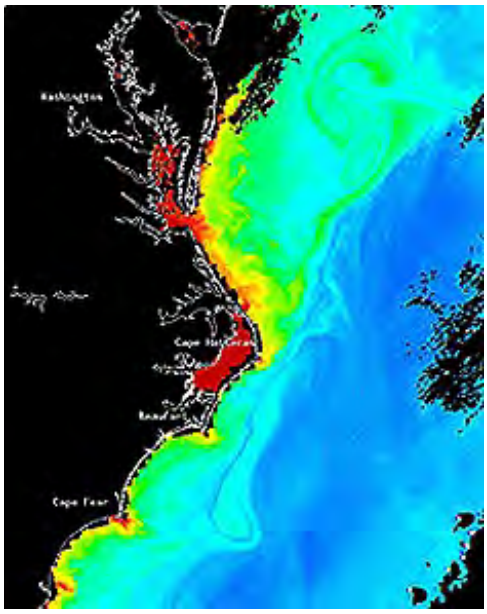


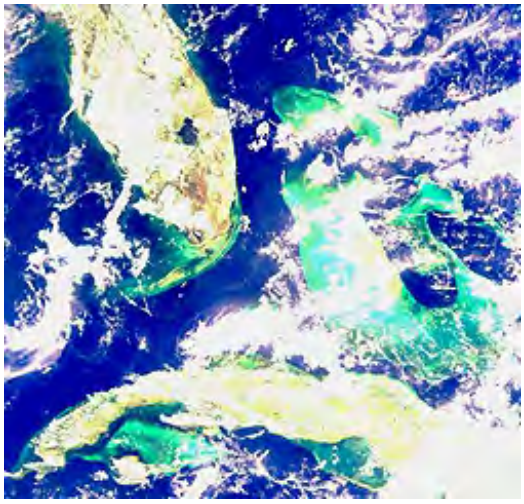
One job that SeaWiFS does well is to monitor sedimentation patterns. Below is an image of the western Gulf of Mexico, showing sedimentation either in brown or in green (much lower concentrations of sediments, but enough to modify water color to lighter tones). (Tie this image with the one four illustrations down.)

But the main job of SeaWiFS is to monitor organic activity in the oceans. Here is a characteristic regional map, showing SeaWiFS-derived chlorophyll content for the globe using data recovered from September 4 through November 20, 1997, using the same color coding as CZCS:



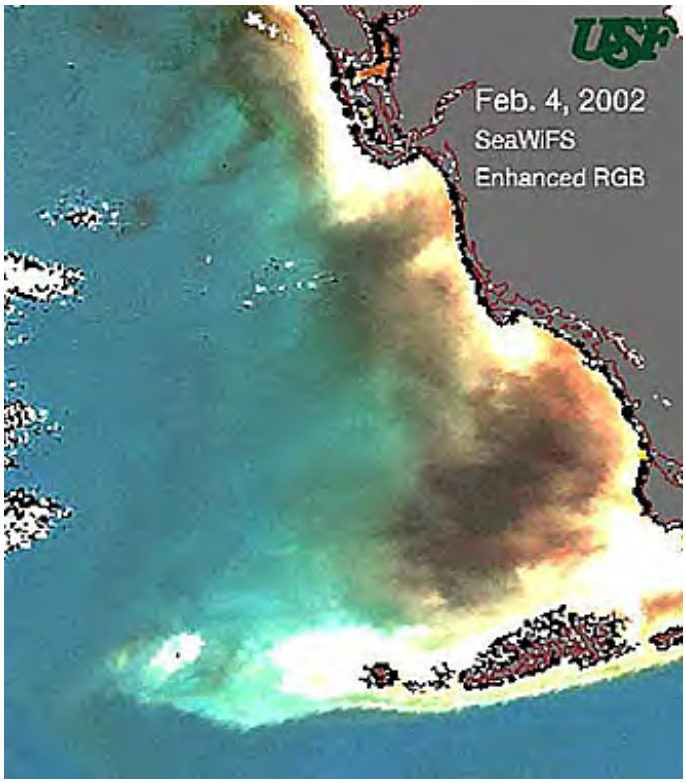
The next two images show (top) chlorophyll mapped off the eastern U.S. coast on September 30, 1997, and (bottom) ocean color off southern Florida on September 25, with greens denoting high phytoplankton concentrations.





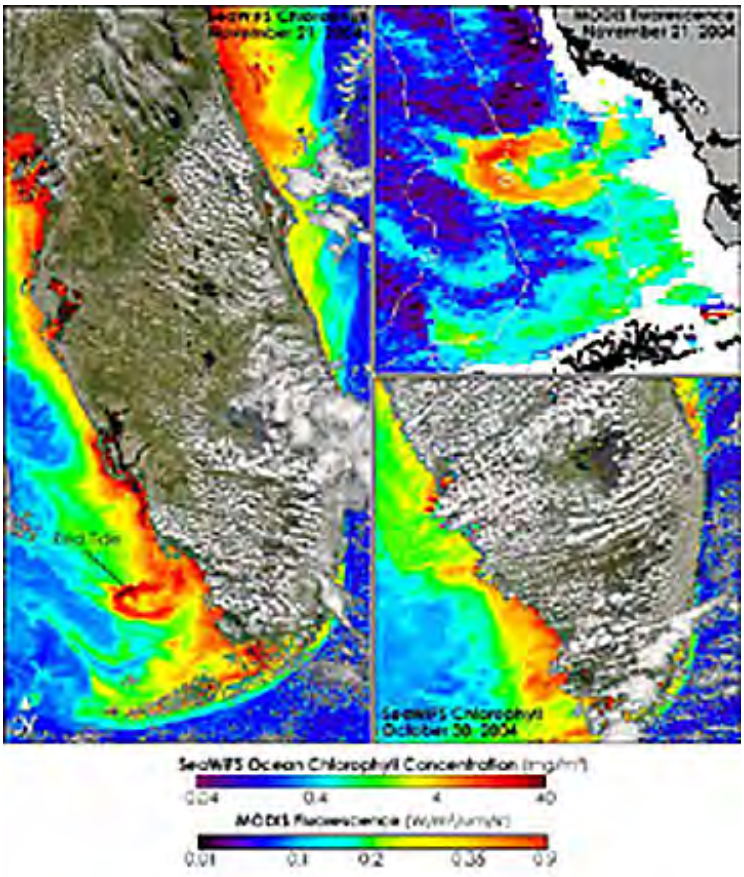
Chlorophyll response is also strong in floating algae. The Red Tide is a much dreaded form of algae that secrets and expels a poison that can be fatal to fish and shellfish such as oysters. This SeaWiFS image of the northeastern Gulf of Mexico shows (in light blue-green) a nearshore concentration of this algal bloom on March 1, 1999:

Another bout with the Red Tide occurred north of the Florida Keys in 2002, as seen in this SeaWiFs image that renders the red algae in almost natural color;

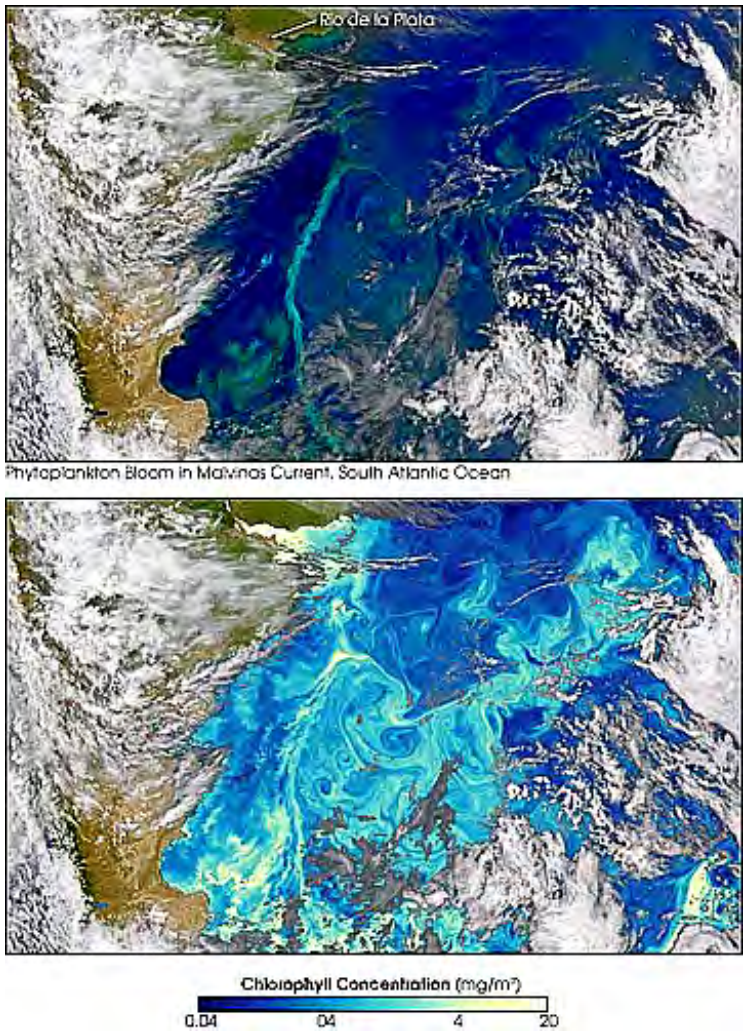


In Winter of 2002, another algal bloom of a different nature affected the eastern Gulf of Mexico north of the Florida Keys. The seawater became black for a while, then reverted to a brownish-green. The bloom extended over a larger area than customary. From a boat, the water indeed appeared black, which occurs when the concentration of both bloom and dead organics becomes high. Fish within the bloom were not killed but nevertheless left the area until the algae died off. Here is a SeaWifS image of the bloom at its height:

The Black Tide condition was bad in late 2004 extending into 2005, Both SeaWiFS and MODIS were effective in monitoring this situation:

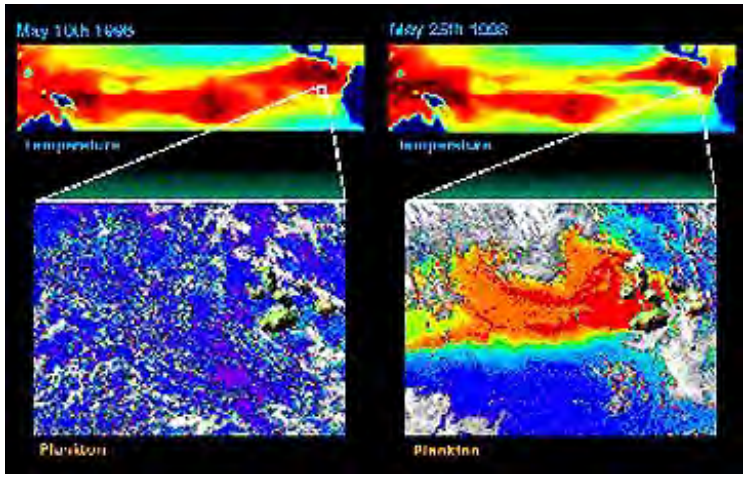


Over large areas, patterns of chlorophyll concentrations can be quite eye-catching. The next two SeaWiFS images are a natural color and a chlorophyll map of the southern Atlantic Ocean of the Brazilian and Uruguayan coasts; the Malvinas current forms a narrow line running N-S offshore.



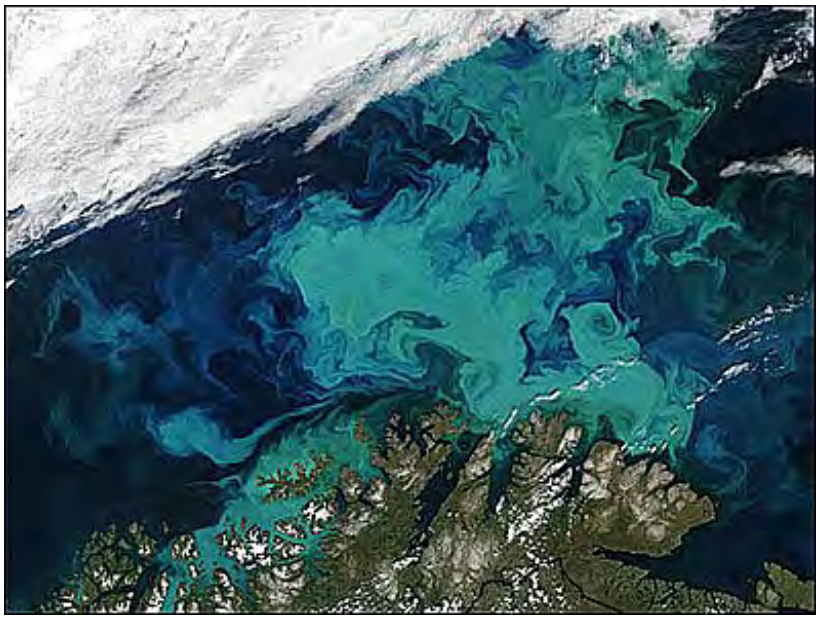
Phytoplankton Bloom in Malvinas Current, South Atlantic Ocean

Plankton concentration (as revealed by its chlorophyll signature) is very sensitive to water temperatures. Thus, significant temperature changes in tropical waters associated with the El Niño event during the late 1990s can be the controlling factor in plankton distribution, as indicated by these two SeaWIFS images in May of 1998, in which plankton appear in profusion (expanded right inset) as temperatures drop in the area it represents:

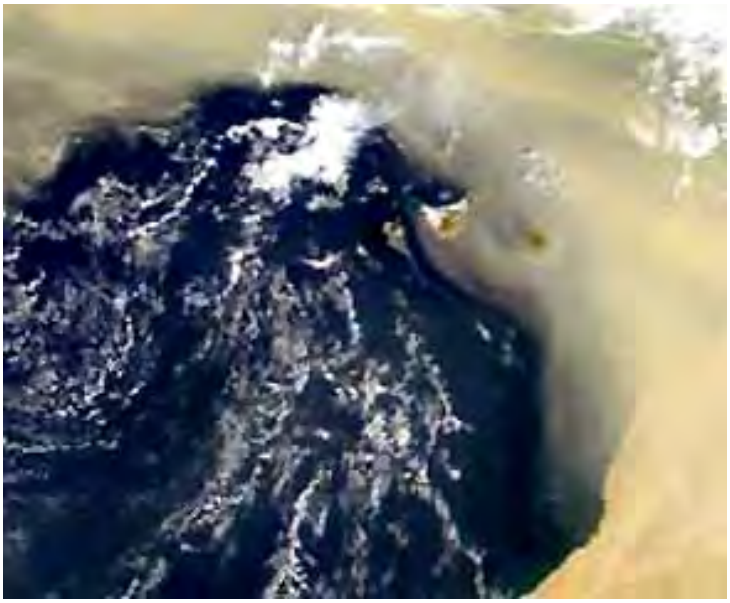


14-35: **Imagine you are a commercial fisherman. How would you use satellite imagery to find the best fishing grounds? ANSWER**

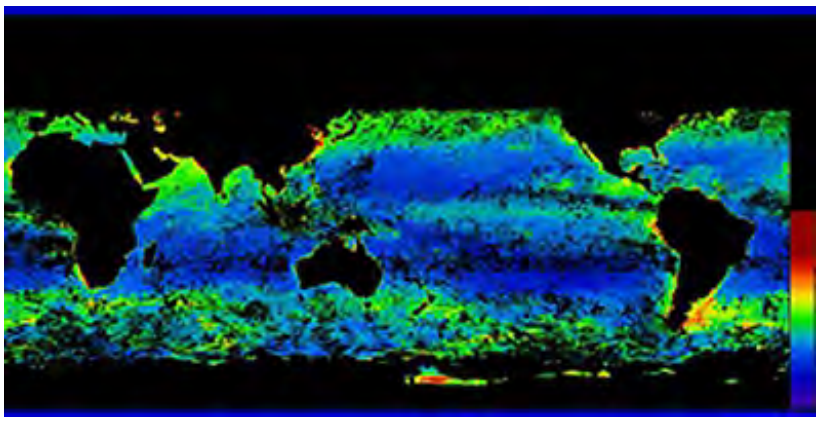
The MODIS instrument on Aqua picked out a large phytoplankton bloom off the northern coast of Norway on July 19, 2003. Good news for Norwegian fishermen! Most of the blue-green color in the (near true color) image below was found to be due to Coccolithophore plankton whose carapaces (shells) are composed of chalky $CaCO_3$.

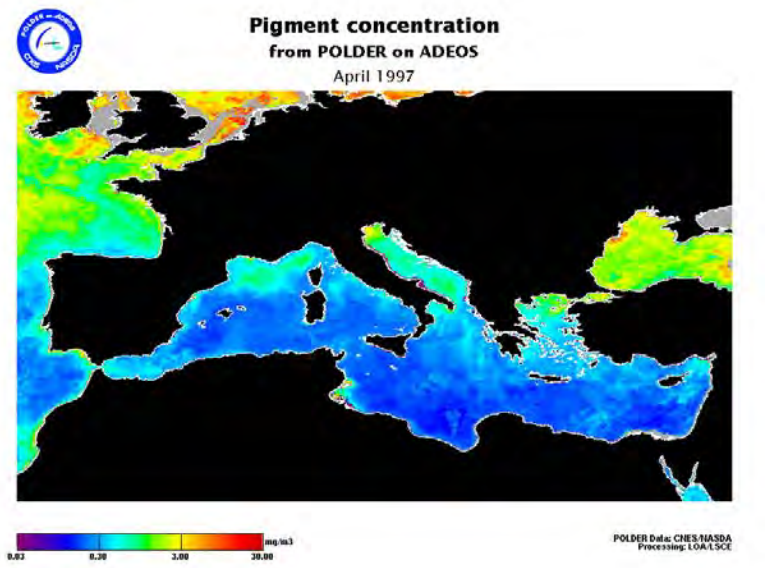


We can also use the sensor data from SeaWiFS to map land surfaces on a local to global scale. We showed a global, true-color map of Earth near the end of the Introduction Section. This next SeaWIFS image depicts still another use for the meteorological/oceanographic class of satellites by monitoring a dust storm coming off the northwest Sahara desert of Africa as it blows out to sea:

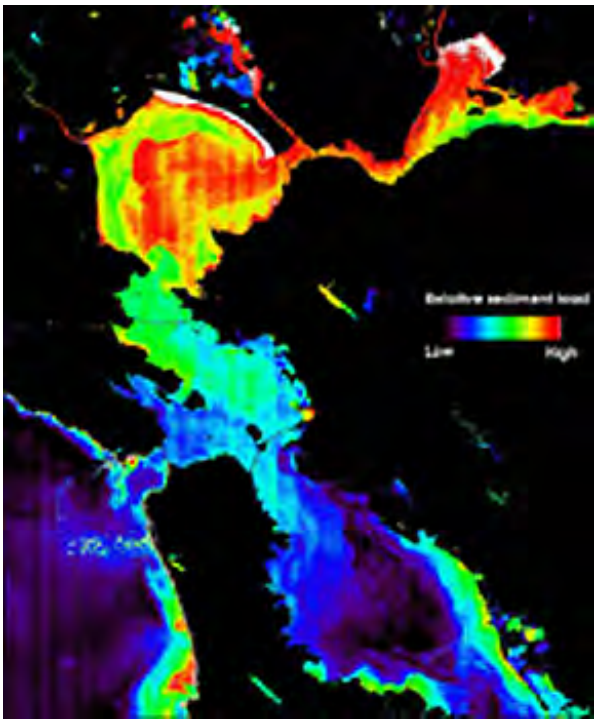


ADEOS, that ill-fated but most promising multisensor satellite, had two instruments that could perform ocean color and phytoplankton assessments. First below is a January 1997 global image made by the OCTS (Ocean Color and Temperature Scanner) and beneath it is a April 1997 view of the Mediterranean made by POLDER.





As seen in several images elsewhere in the Tutorial, sensors operating mainly in the 0.4 to 0.7 µm spectral interval are capable of monitoring the distribution of sediments (current-driven) in circulating or even in standing water. The density of sediment in the waters very near the surface can be quantitatively assessed. In practice, on-site measurements can provide calibration points to determine densities. Here is an example of the determination of variations in sediment density in the San Francisco Bay and off the coast of the Peninsula there, as measured by Terra's ASTER (see also page 16-10 for complementary images).



The higher densities in the northern S.F. Bay and adjacent San Pablo Bay are due to sediment being carried into these waters from the Sacramento River.

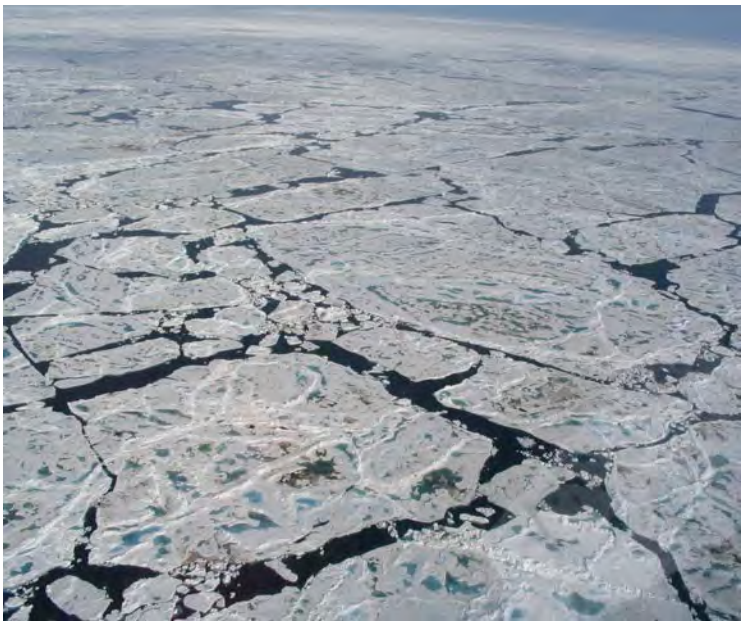
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Sea ice is both an oceanographic and a meteorological feature. Satellites are almost the only convenient way to monitor ice accumulations in the polar regions. Satellites with high resolution sensors can also detect icebergs. Generally, any spectral band in the visible is suited to spotting ice but thermal imagery and radar (especially at night) also will find it. Some examples of what can be seen from space substantiate the utility of satellites for picking out ice distribution - of vital concern to those who must travel in shipping lanes where passage may be impossible or where dangers lurk. (The big "what might have been" question: could satellite observations have saved the Titanic?)

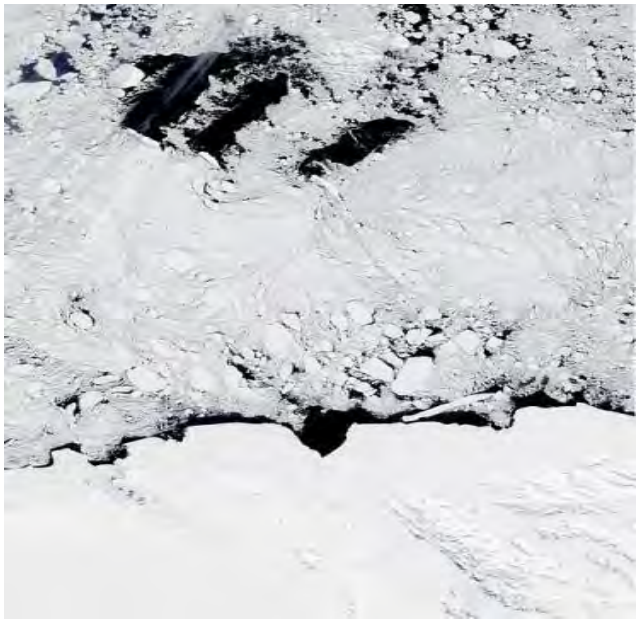
Sea Ice Monitoring

Our final look at satellite use in oceanographic monitoring concerns sea ice, which normally occurs year round in the polar regions but to varying extents. When one thinks of sea ice, these two scenes are typical:



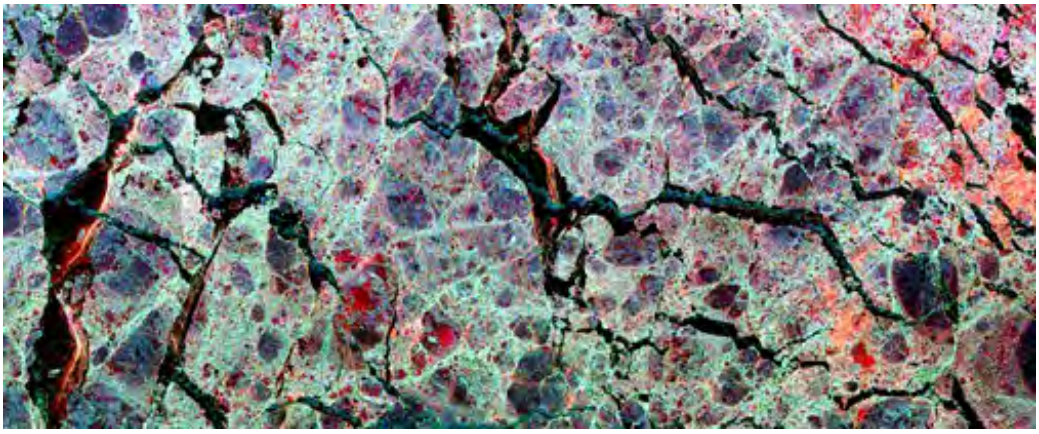
These are ice flows (blocks of ice of varying sizes separated by open water. When the open

water occupies narrow cracks, these are called leads (top and center images below); when the open water is present as elliptical to irregular, wider stretches, they are called polynyas (bottom image):



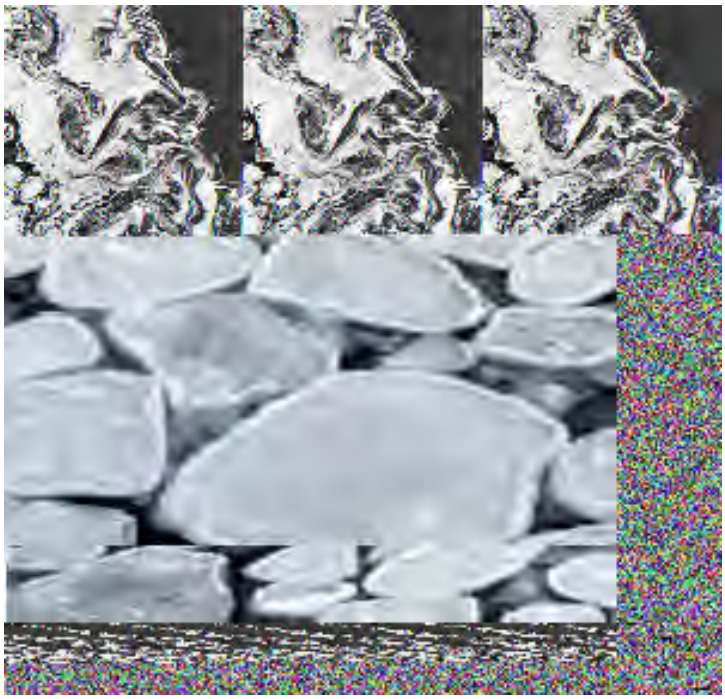
Sometimes the ice floes have rounded edges and are called pancake ice:

Sea ice is easy to monitor from space as long as the scene is largely cloudfree when a satellite passes overhead. Polar orbiting satellites repeatedly pass near the poles on a daily cycle. Visible and radar imagery is effective in observing sea ice on a continual basis. Radar at bands that penetrate clouds (e.g., L-band) is now operational (Canada's Radarsat, for example) to monitor shipping lanes subject to ice hazards. Below is a SIR-C multiband color composite that shows ice in the Weddell Sea, off Antarctica, south of the Atlantic Ocean. Open water polynyas show as darker tones.

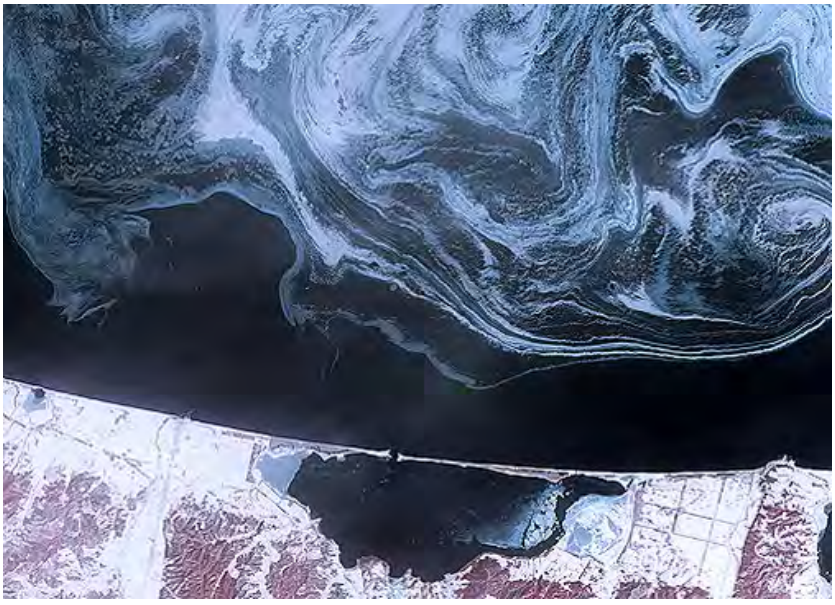


14-36: **Why does this ice show a blocky or patchwork pattern?** **ANSWER**

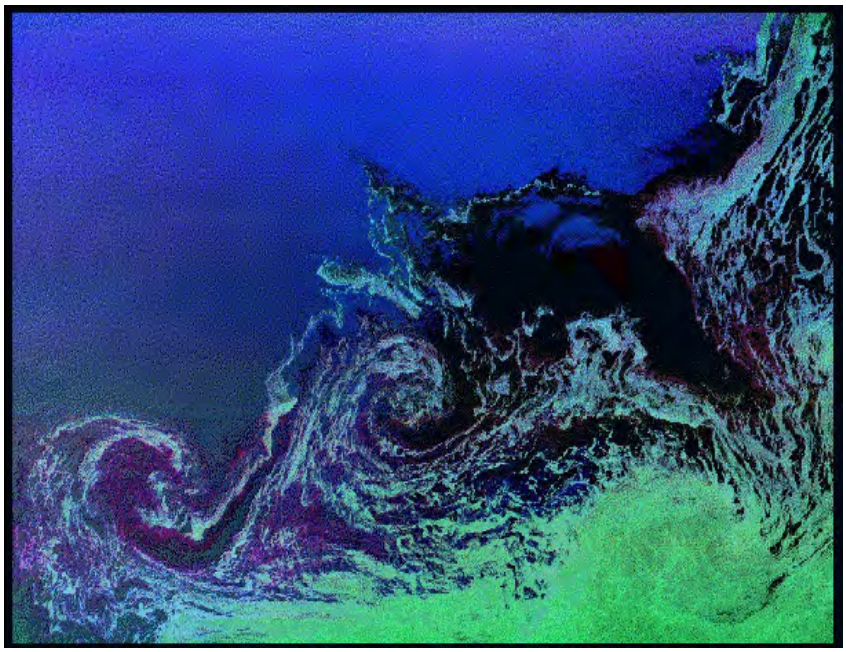
Sea ice may retreat during the summer but reform in new patterns as winter ensues, as shown in this Landsat image off the Greenland coast:



As the ice forms seasonally, the growing pack can take on wispy patterns, as seen in this Envisat image:

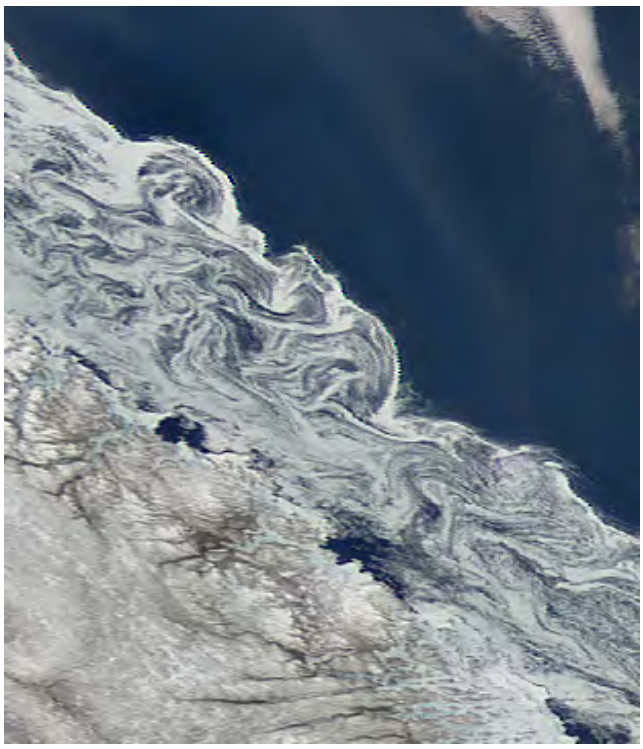


Offshore eddy currents can cause the forming ice to swirl:

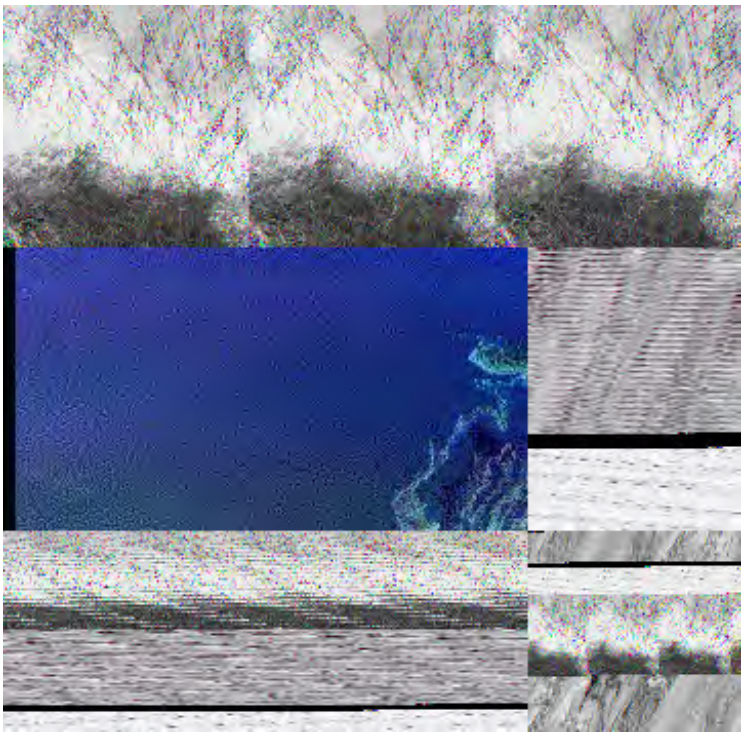


14-37: **Which way are the ice eddies swirling? ANSWER**

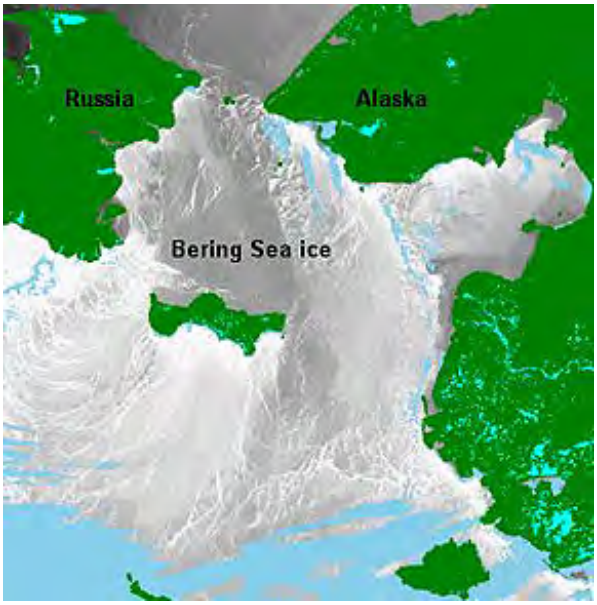
This next image is ice swirls off the Labrador coast, as seen by SeaWIFs:



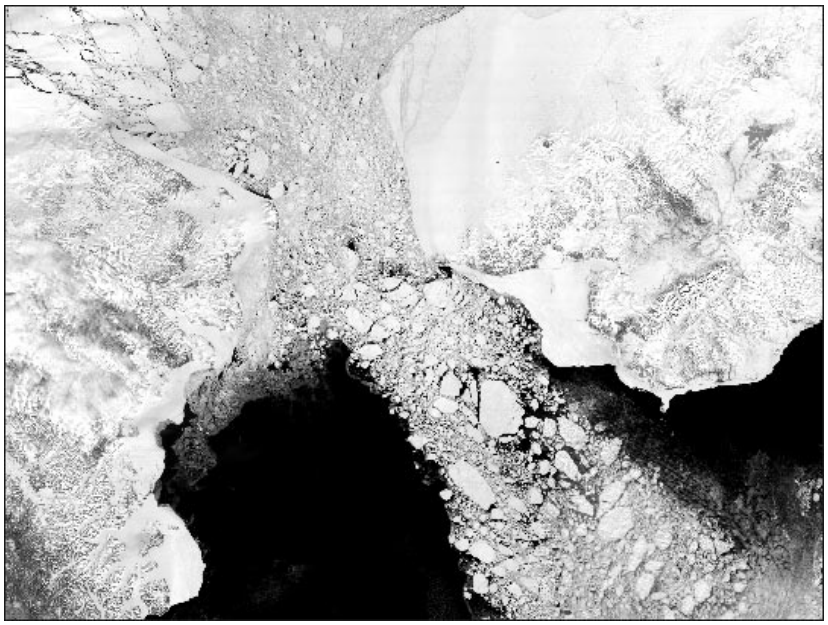
We see the nature of ice packs in the Arctic, over a large area, in this HCMM-Visible image of the Chukchi Sea in the Bering Strait between Alaska and Siberia. The ice in this scene has a network of cracks, the so-called leads, which open during summer breakup and refreeze when conditions demand.



The MODIS instrument on Terra has also observed this region:

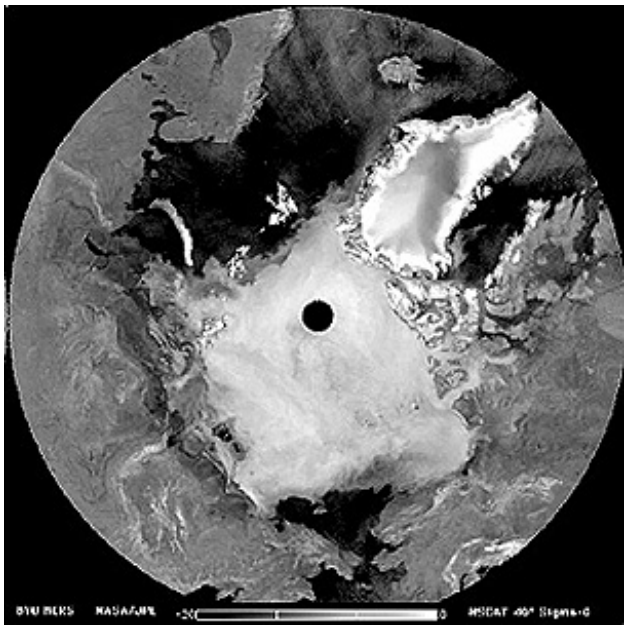


Sea ice distribution over a large region is effectively being monitored now by the MODIS sensor(see Section 16). Here is a view of ice passing through, and blocking, the Bering Strait between Siberia and western Alaska.

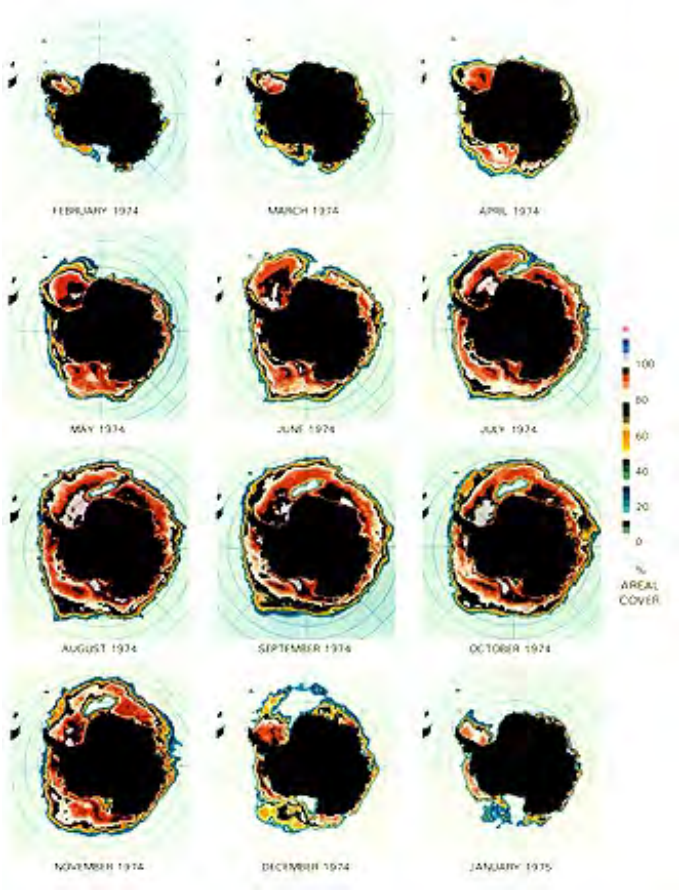


The polar regions of the Northern Hemisphere are covered by sea ice, ice caps or sheets (e.g., Greenland) and snow, as depicted by this composite image constructed from Radarsat SAR images:

Below is another view of the north polar regions, showing the prevailing ice cover, made by NSCAT (the NASA Scatterometer).

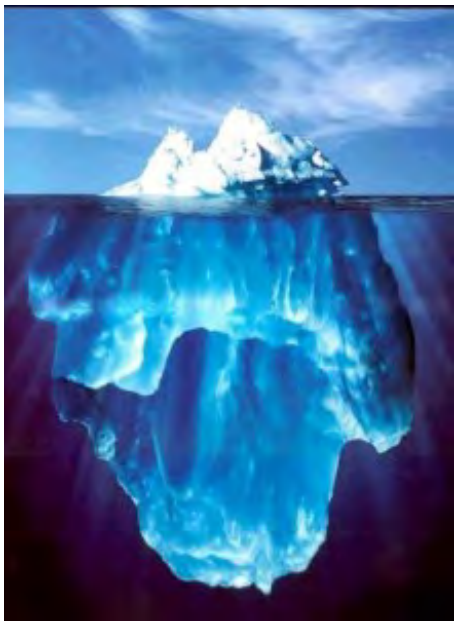


Next, we show the growth and shrinkage of the ice fields surrounding Antarctica as sensed by ESMR that appeared earlier in this section. Below these images are SMRR images that show changes in the ice cover for four years, during the 1978-86 period, in the top diagram.

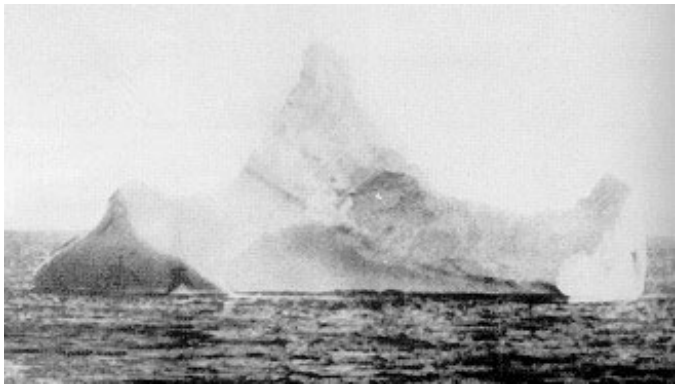


14-38: **In the lower of these two diagrams above, why is ice minimal in December? In the upper diagram, what is ice-covered besides the Arctic Ocean?** **ANSWER**

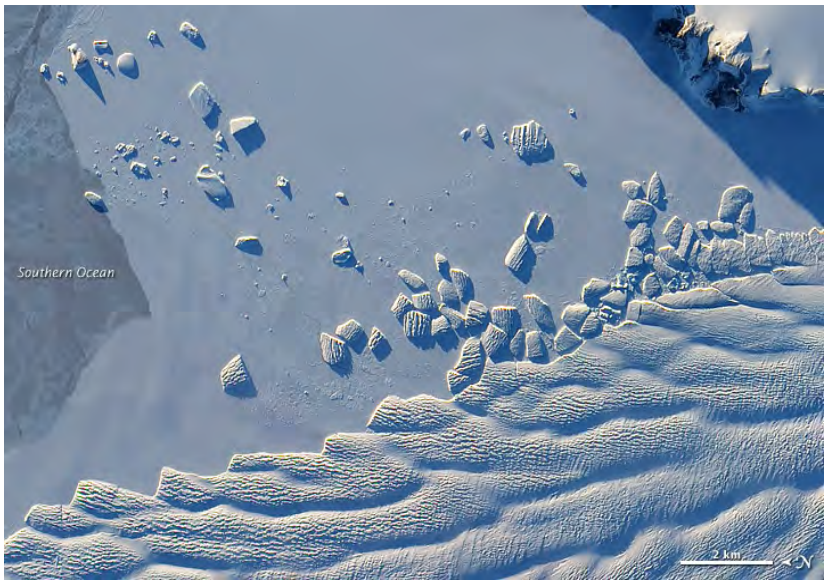
An important topic for this page is the formation of icebegs. Almost all icebergs come from glacial ice reaching the sea, breaking off, then being carried by currents into open oceans, where they can become serious hazards for shipping (remember the Titanic!). Numerous small blocks of ice are typical but some can be large and therefore dangerous. Because the density of ice is less than that of water (arbitrarily set at 1.0), it floats. But only 1/7th of the ice is above water as shown here:



The most famous iceberg in history was the one that sank the Titanic in 1912 (more than 1500 people perished in that disaster). Someone on one of the rescue boats arriving at the scene to pick up the 700+ survivors took a photo of the iceberg shown below, almost certainly the culprit in the sinking since red paint from the ship was noted above its waterline:



Surprisingly few satellite images showing icebergs are found on the Internet. Here are icebergs forming by splitting off the Matusevich glacier (note its crenulated surface) in the Antarctic:

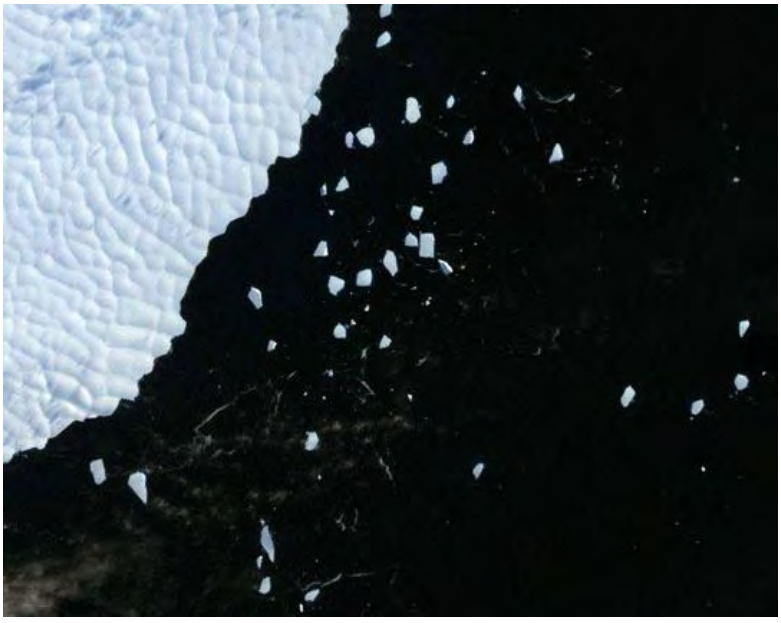


As the Bear glacier in the Denai Peninsula of Alaska reaches an ocean inlet, ice breaks off to form small icebergs that could become a hazard if they reach the open sea. This IKONOS image has high enough resolution to show crevasses formed by differential movement of the ice mass:

Sometimes large sections of icesheets break off and become tabular icebergs. Here is a segment broken from the Mertz glacier in the Antarctic that has entered the open sea.

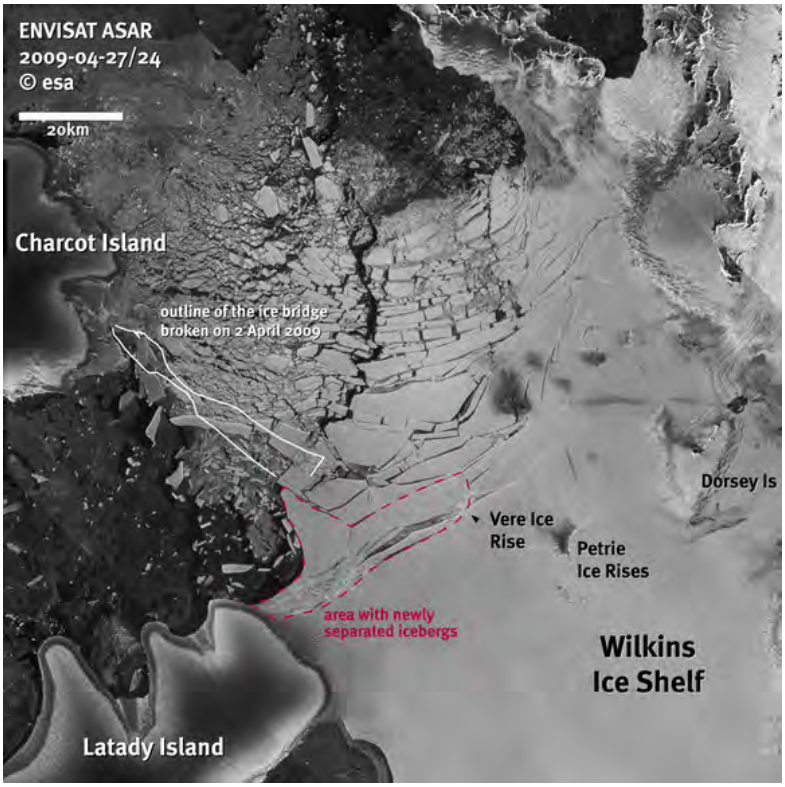
Because icebergs are evident in satellite imagery, one would expect that there would be at least one satellite dedicated to monitoring potential iceberg hazards. Not yet. Radarsat and Envisat have been used for this purpose but do not provide systematic surveillance. Still, consider the next group of images as examples of what can be detected.

The ice shelf surrounding the Antarctic continent extends sometimes 100s of kilometers from shore. Periodically parts of its edge will break loose and float free in the southern ocean. In 1995 a large (greater than Rhode Island in size) raft of ice, assigned the identifier B10, separated and starting moving north. It later broke again into two sections. Small pieces continue to break off (calve) as icebergs. Here is a Landsat image of B10A showing the sheet and its offspring 'bergs:

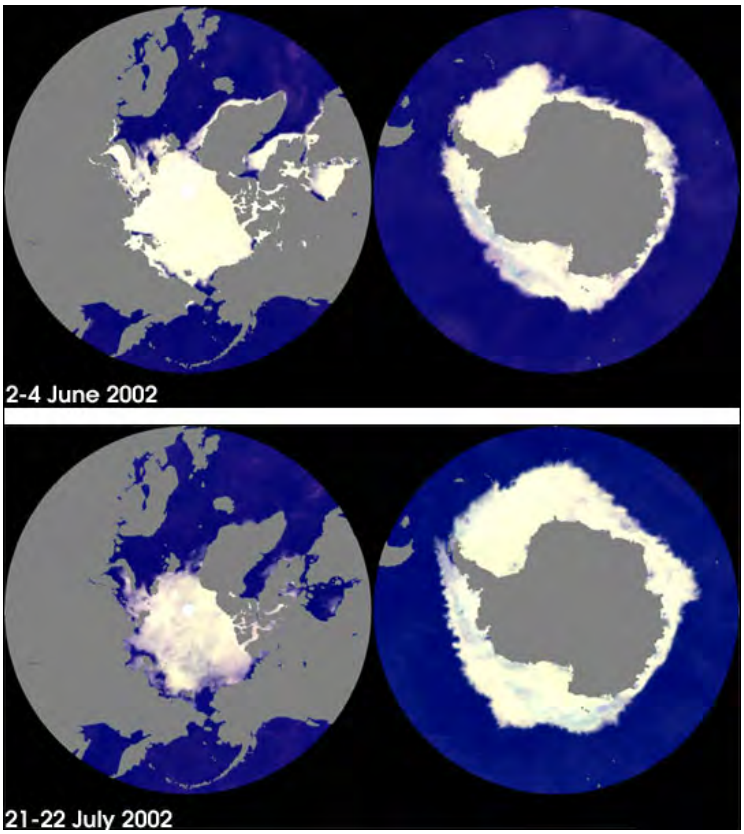


In early 2002, another great slab of ice from the Larsen ice shelf, between the Bellingshausen and Wendell seas, broke off and is now cruising the sea off Antarctica. Here is a scene during this rupture stage, imaged by the ASAR radar on the Envisat (see Overview) launched in February 2002.

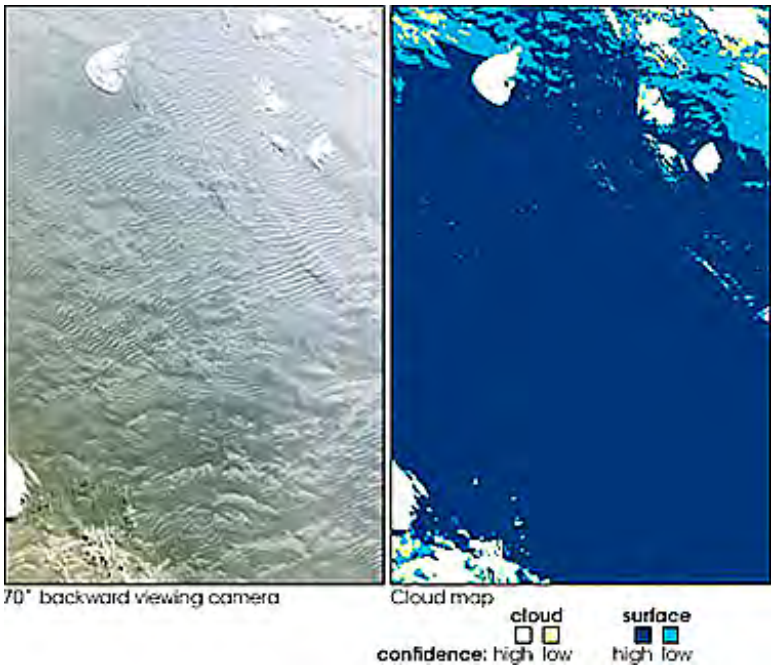
As temperatures continue to rise slightly in the Antarctic, more ice from the shelves is breaking away. Consider this image of the Wilkins Ice Shelf, as imaged by the SAR on Envisat:



While some changes in sea ice (and ice cap) cover, both in area and thickness, may be progressing towards lower overall areal extent because of natural/man-induced global warming, there are as well normal seasonal fluctuations. Here are two sets of images made by the AMSR-E sensor on Aqua (page 16-11) of Arctic sea ice (left) and Antarctic shelf ice (right) during two periods in 2002 between June and July:



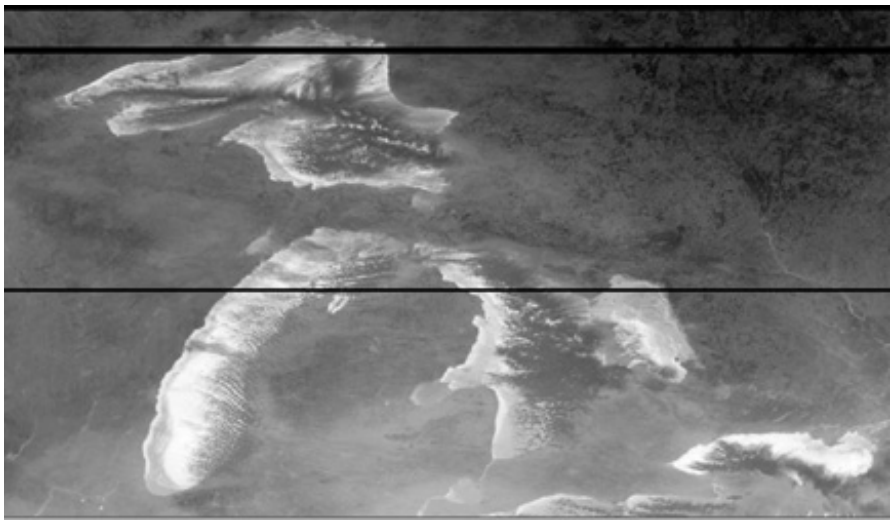
Both ice and clouds appear white in some images. But using different bands, a distinction can be made. Here is an image of Antarctica. The question arises: is this all clouds or is there a solid surface within the scene. The image on the right was processed using selective bands. It brings out the few clouds present and discloses the other patterns to be ice markings on the continental ice surface:



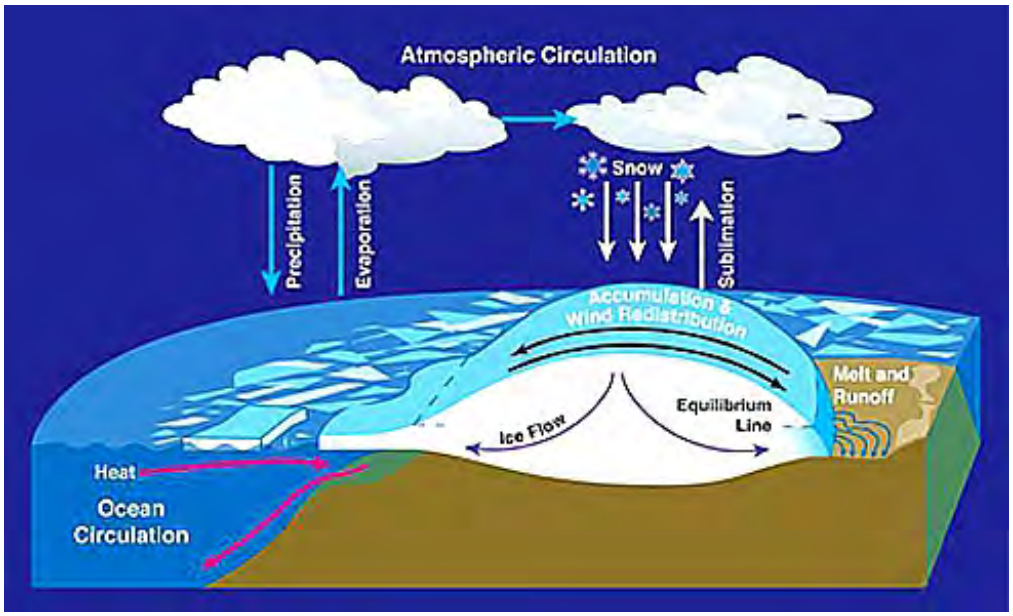
In the northern hemisphere, the ice shrinks in extent as the summer progresses. But in the southern hemisphere, where it is winter at this time of the year, the ice shelf around the Antarctic increases.

In the U.S., winter ice on the Great Lakes becomes a major impediment to shipping and usually some or all of the lakes are closed to normal travel. The AVHRR on NOAA satellites daily monitors the status of ice cover, giving results like this thermal IR image, made on January 31, 1996 (ice is light-toned because its temperature is above that of the land that is experiencing a cold-air outbreak from a Canadian airmass that has cooled land surfaces to lower temperatures

than the ice; clouds appear very dark):



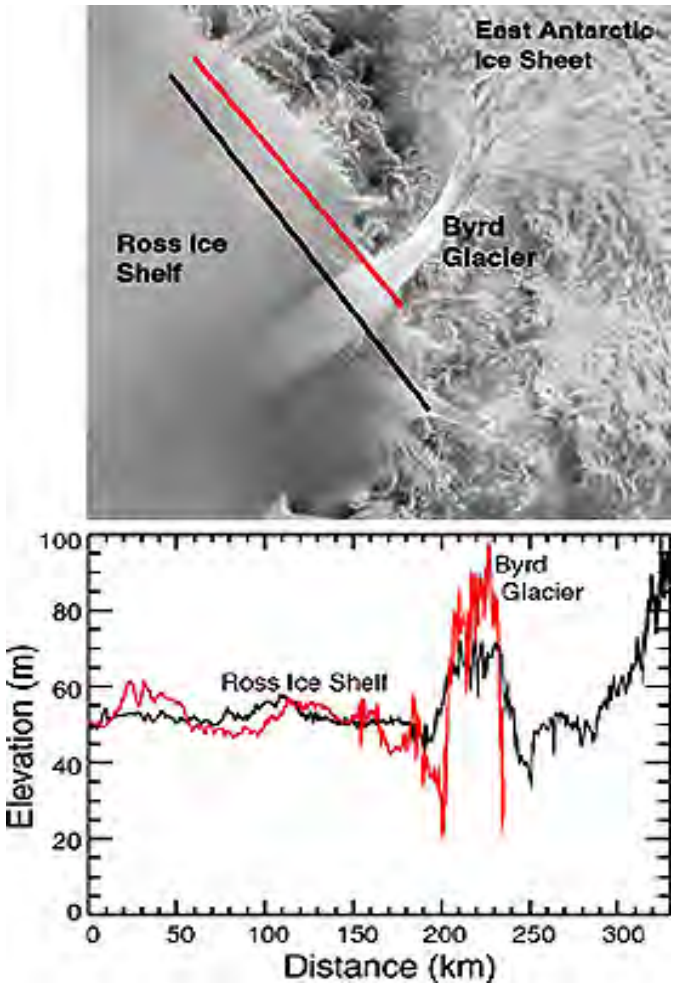
The latest addition to the "stable" of ice-monitoring satellites is ICESat, which stands for Ice, Clouds, and land Elevation Satellite (check out its Home Page). Part of the ESE (Earth Sciences Enterprise; see page 16-6) series, the primary mission is to map changes in elevation of major concentrations of ice on land surfaces (Antarctica, the Arctic islands, Greenland) over time to gain precise data on melting or accumulation of ice mass which can have a major effect on changing sealevel worldwide. It also surveys ice cover in the Arctic Ocean. This simple diagram shows the ice model used:



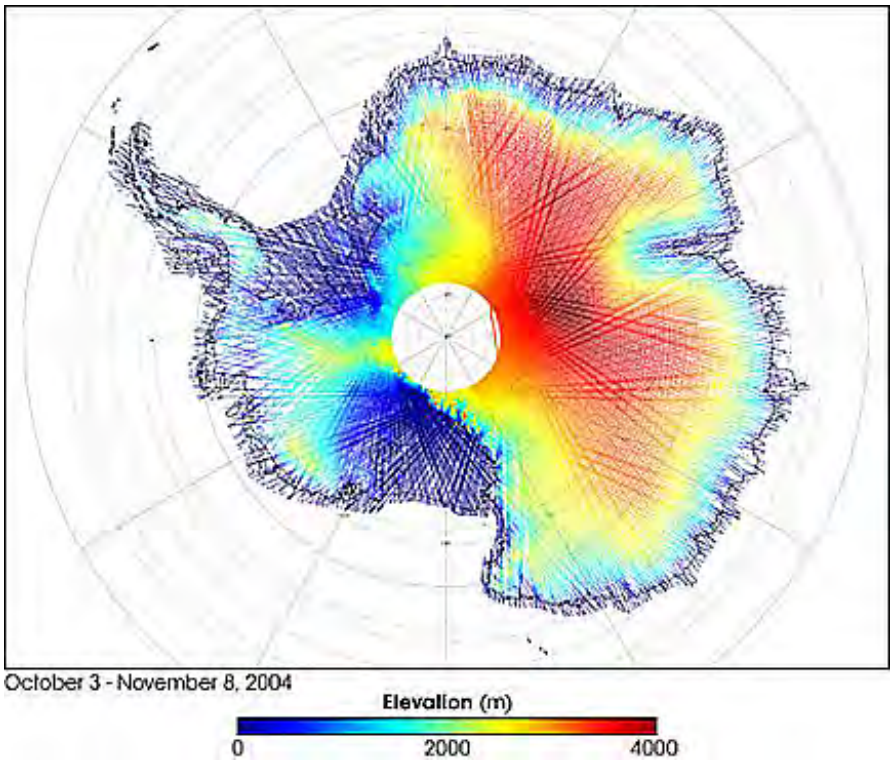
The ICESat orbits at a 600 km altitude such as to cross the polar regions frequently over short time spans. The satellite's position is accurately established by radio linkage to Global Positioning Satellites (GPS series). The prime ICESat instrument is GLAS (Geoscience Laser Altimeter System). The laser used is the Nd:YAG type; this uses a Yttrium Aluminum Garnet (a manmade mineral with a specific crystal structure) in which some of the Yttrium is substituted for by Neodymium (Nd^{+3}) (said to be "doped"). In the mode used, this laser produces signals at two wavelengths (green: 0.532µm, used to determine aerosols and IR: 1.064µm [twice the first], used to measure elevations. The pulses are sent in bursts of 40 seconds (traveling at the speed of light), each covering a circular surface footprint 75 meters wide and pairs spaced 175 m apart, and they require 5 nanoseconds to travel from satellite to Earth and back. Very slight differences in travel time allow surface elevation changes of 10 cm to be measured accurately.

ICESat was launched on January 13, 2003 and required a few months to refine the operation of GLAS. But now a number of laser-determined elevation profiles have been produced. Here is one

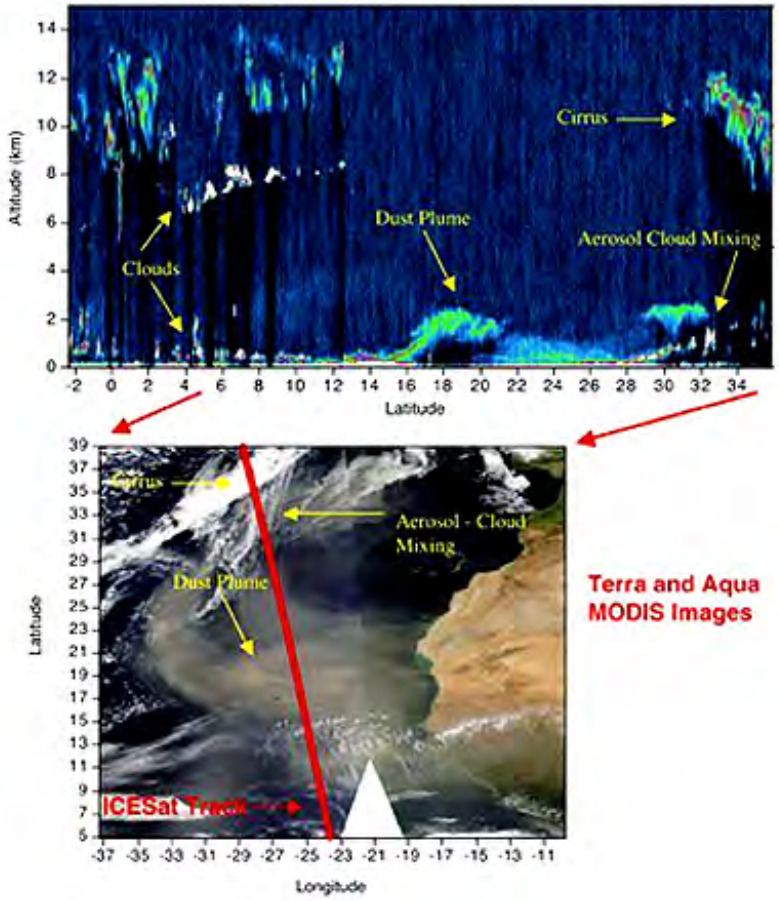
from the glacier off the Ross Ice Shelf in the Antarctic:



The GLAS sensor has now obtained enough topographic measurements to produce an elevation map of all but the central part of the Antarctic:



This next illustration shows off the ability of the 0.532 µm laser pulses to obtain information about clouds and dust in the atmosphere.



Although its primary mission is to measure ice surface changes, the ability to determine heights from space has other applications, one of which is to determine canopy heights for forests, as presented at the bottom of page 3-5.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

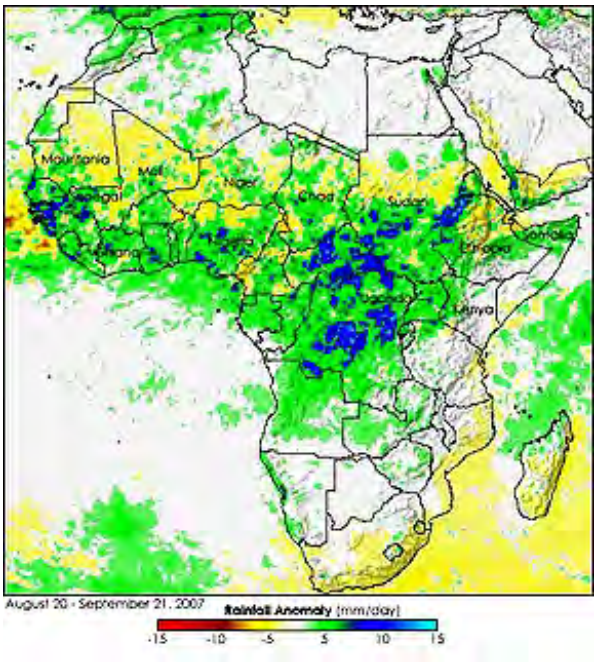
This entire Section has dealt with water - sensing its presence and its activities using spaceborne sensors. Water is ubiquitous - occurring in all the major spheres of the Earth: its outer rock shell; its surface; in its oceans and rivers, in the atmosphere, and in living matter. We close with an examination of some of the practical uses of remote sensing in finding and tracking the movements of water on the land. The first examples examine drought-driven changes in water volume and extent of large lakes. Attention is then focused on water as snow and water in streams. Satellite data are excellent forecasters of floods and imagery is capable of seeing the extent of flooding over wide areas. Both Metsat and Landsat type imagery are effective monitors of snow cover and rivers out of their banks - the wide areal coverage suits this kind of use.

Hydrologic Applications: Monitoring Drought and Snow Cover

These views of frozen water carry us naturally into the last topic - hydrology - in this Section. In this context, hydrology refers to water distribution and management on land surfaces. We discuss here three of the four principal uses of satellites for hydrologic applications: drought assessment, runoff prediction, and flood monitoring/damage assessment. The fourth topic - drainage basin characterization - we do not consider.

First, several statistics: 1) There are approximately 2 million cubic miles of freshwater throughout the world, mostly in the top 173 m (570 ft) of the crust (both in and on); 2) most freshwater (70%) resides in Antarctic and Greenland ice; 3) there is ~30 times more freshwater as groundwater than in rivers and lakes; and 4) Lake Baikal (see below) has 20% of all lake freshwater. One added fact: water consumption (humans and animals) is about 400 billion gallons/day - about 65 gallons for each higher order creature on Earth.

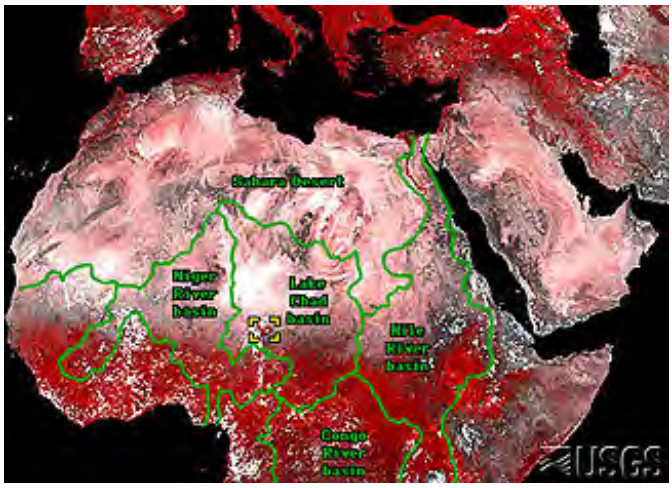
A primary use of meteorological satellites is simply to measure the regional and continental distribution of normal, excessive, or sparse rainfall. TRMM is especially suited to that task. As an example, here is a six weeks summation of rainfall over most of Africa in August- September 2007. The rainy season centers around late northern hemisphere summer, at which time the amount of rainfall will determine whether drought or good harvest conditions will prevail for the year. 2007 has proved to be a period of better than average precipitation:



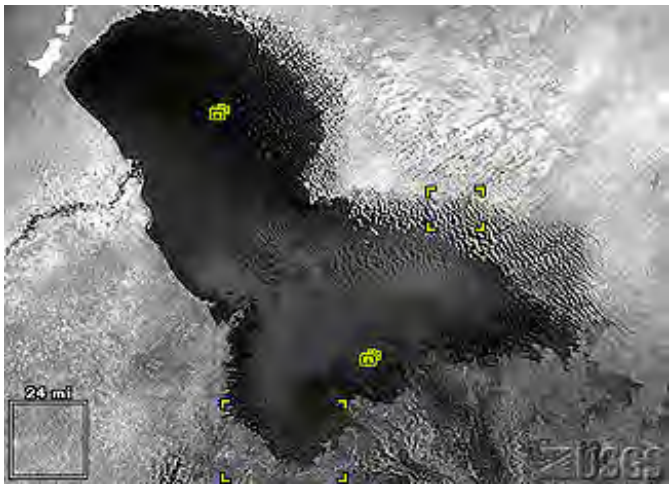
We touched upon the use of space imagery to monitor vegetation stress and other effects of drought on page page 3-4. These tend to be seasonal or short-termed over years, with eventual reversal of the climatic conditions that bring about water shortages which threaten crops and habitats.

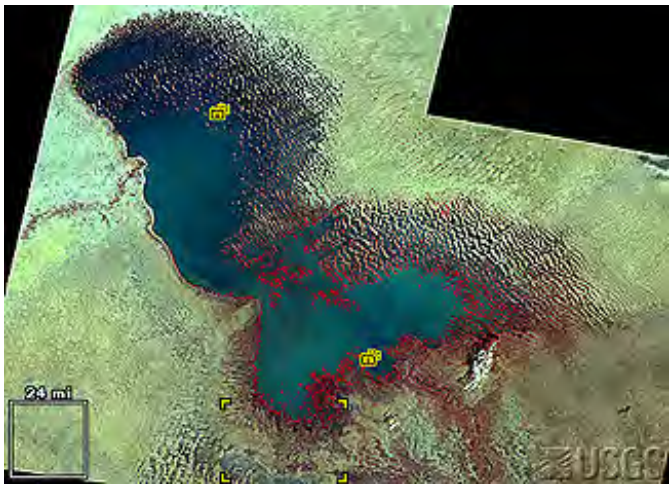
One significant long-term change is the gradual reduction or disappearance of lakes. Lakes tend to be ephemeral when considered at geologic time scales. However, notable shifts in climate - such as is now much talked about in terms of global warming - can sometimes destroy major bodies of water in a century or less. In the U.S., the Great Salt Lake (see Overview) has experienced some fluctuations of size and water level in the 20th Century. Drought often accompanies increased aridity and desertification which pose a threat to lakes that have survived intact since their host areas were first settled. Two drastic examples will illustrate this.

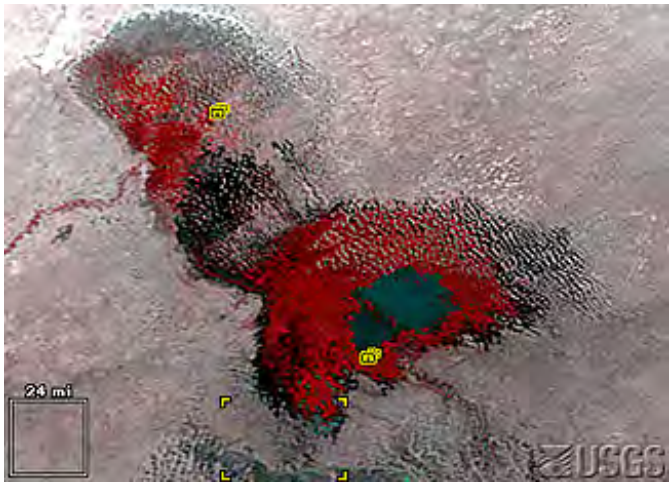
The first is at the southern end of the Sahara Desert which now appears to be enlarging at the expense of once vegetated central African lands. Lake Chad, which lies at the triple junction borders of the nations of Chad, Niger, and Nigeria, was once the 6th largest freshwater lake in the world. This AVHRR image locates it within the yellow brackets (lower left of Lake Chad Basin label):



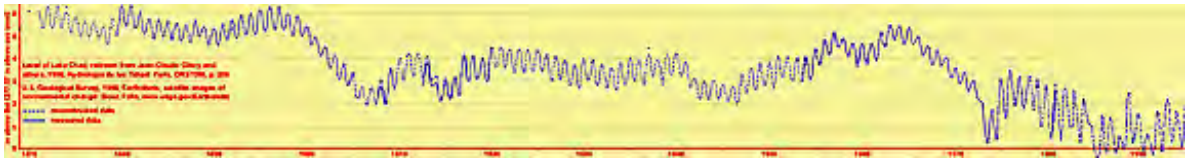
This body of water has been essential to nomadic peoples living around its shores. But its area and volume have been rapidly diminishing over the last half century, as the vegetation line pushes south. This image - actually a photograph taken by astronauts - shows its state in 1963. Below that are two Landsat images - the top acquired in 1973 and the bottom in 1997.





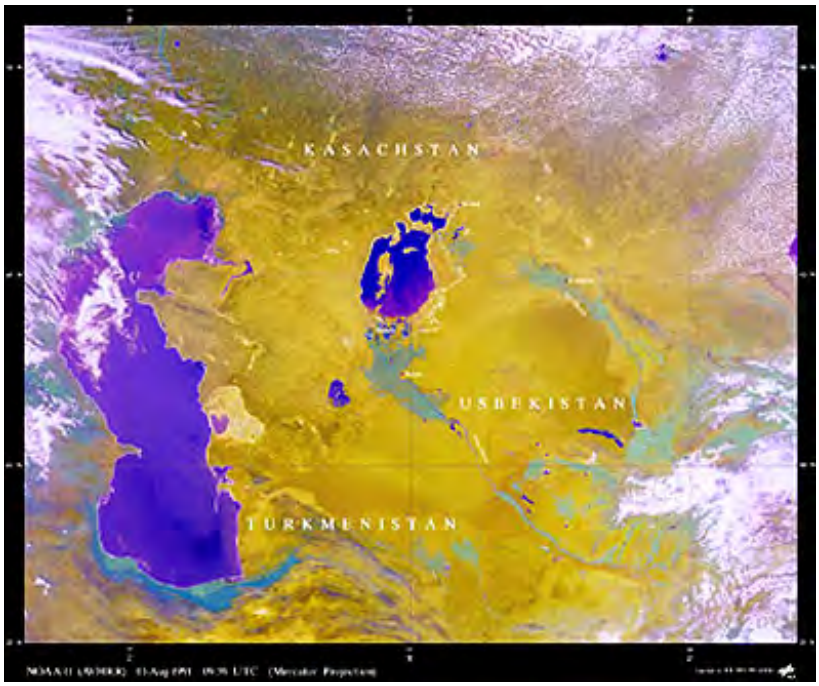


These views almost speak for themselves. In 1973, the southern part of the lake began converting to marshlands. By 1997, almost all of the open waters of Lake Chad had disappeared as both marsh and landfill replaced them. A time-sequential plot of water levels over a 120 year span, made from data processed by the U.S. Geological Survey, shows (unfortunately, with information in red too small to be legible here) the pronounced changes since 1960. Unless there is a miraculous reversal in climate and rainfall, the future for this lake is obvious: *Chad has been Had*.

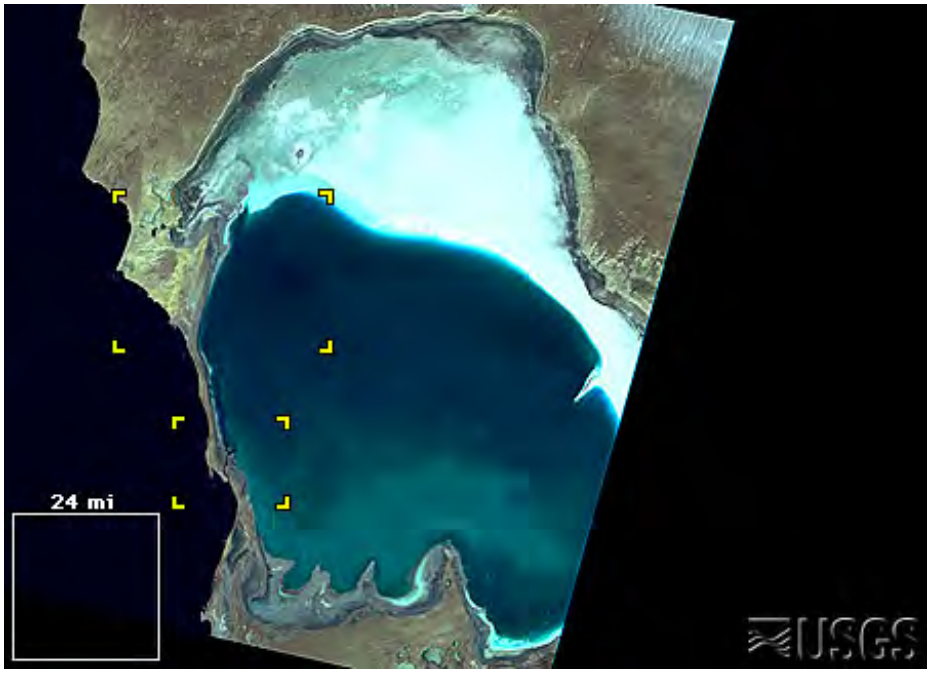


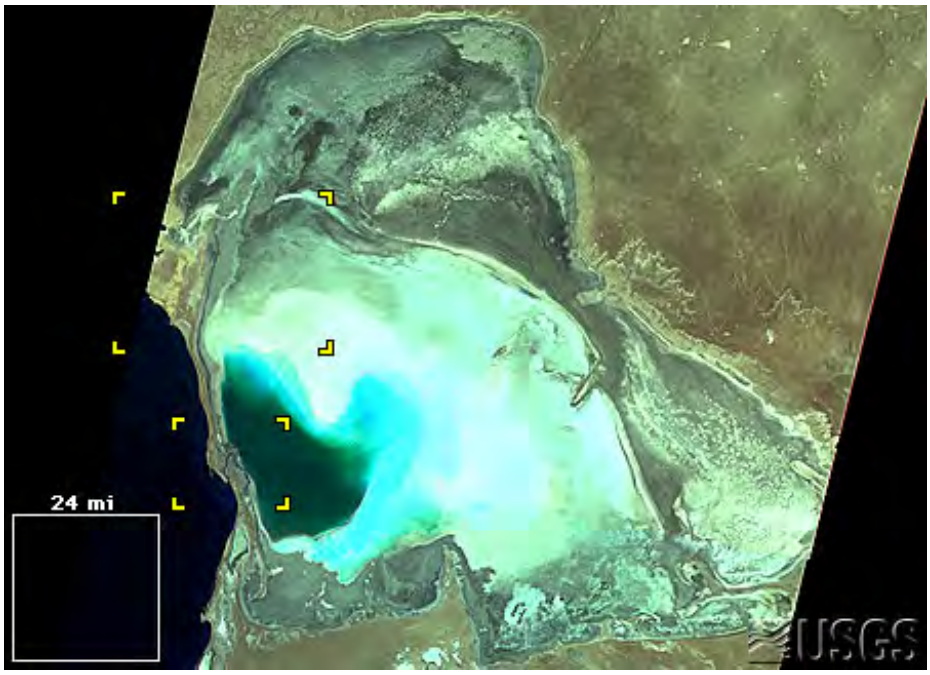
A comparable situation faces the both the Caspian and Aral Seas in southern central Asia, as seen in this AVHRR image. The somewhat saltier Caspian Sea (the larger body of water near

the left edge of the view below) lies within Turkmenistan and abuts parts of southern Russia. The Caspian Sea holds the distinction of being the largest lake containing near-freshwater in the world. Lake Baikal in southeastern Siberia is the deepest (1.7 km or 5712 ft).

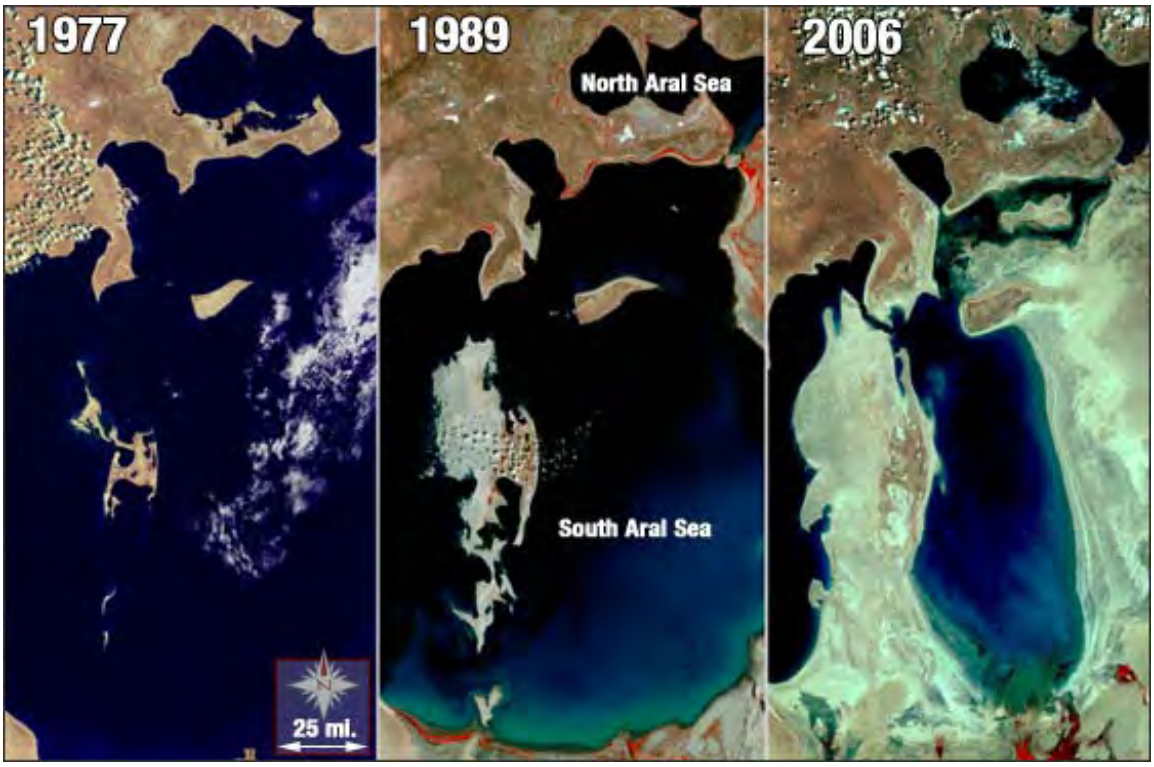


The Caspian Sea is slowly losing water. The best indication of this is the Kara-Bogas Gol (Gulf) located on its eastern shore near the center. It is nearly filled in the AVHRR image. But this is now a rapid process. Two Landsat subscenes, in 1972 and 1987, show its progression in recent years, as salt is filling its interior.:

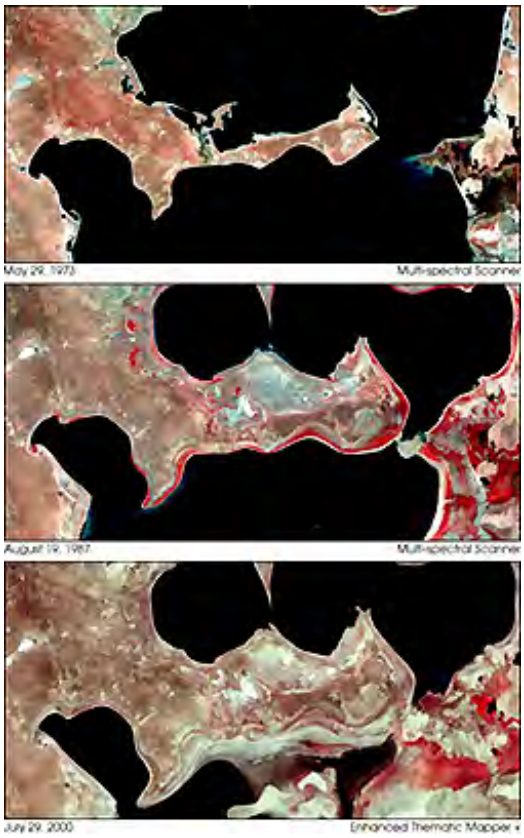




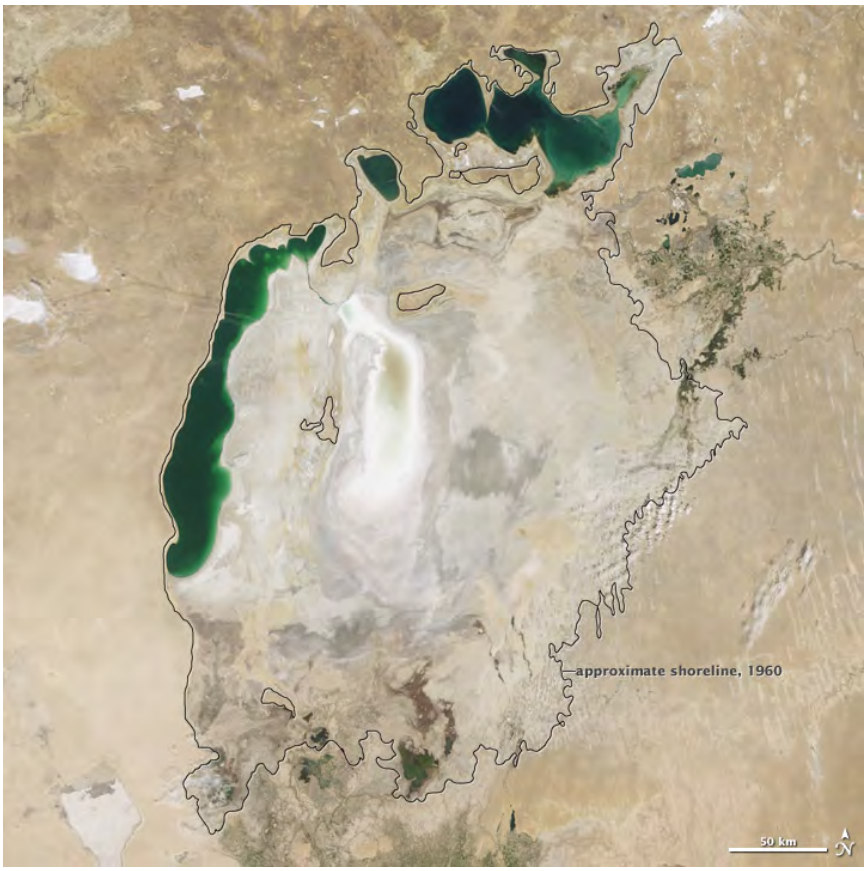
The Aral Sea in Uzbekistan has lost 50% of its area and 75% of its water volume since 1960. Some of this loss may be due to decrease in annual rainfall in that part of Asia, but the diminution is largely caused by diversion of water from rivers feeding this inland lake for use in agriculture elsewhere. This has impacted the means of living by residents around the sea: loss of suitable conditions for cotton growing and great dropoff in fish supply. The changes this has wrought are evident in the next two views made by Russian satellites. The first shows the Aral Sea in the mid-1970s. The second consists of images taken on three separate years:



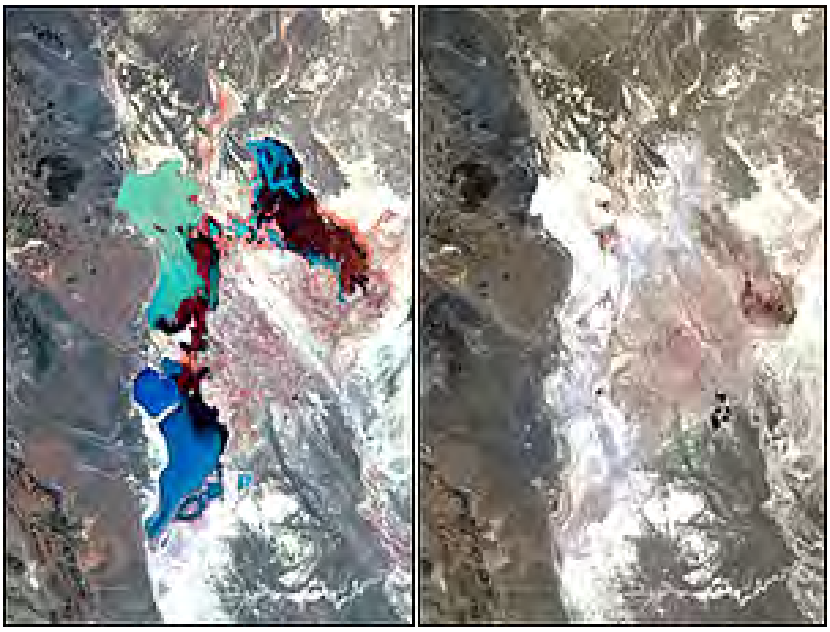
The reduction in water-covered area is obviously matched by a gain in land surface. The next three images, taken in 1973, 1987, and 2000 by Landsats, clearly reveals the growth in land area along a part of the Aral Sea.



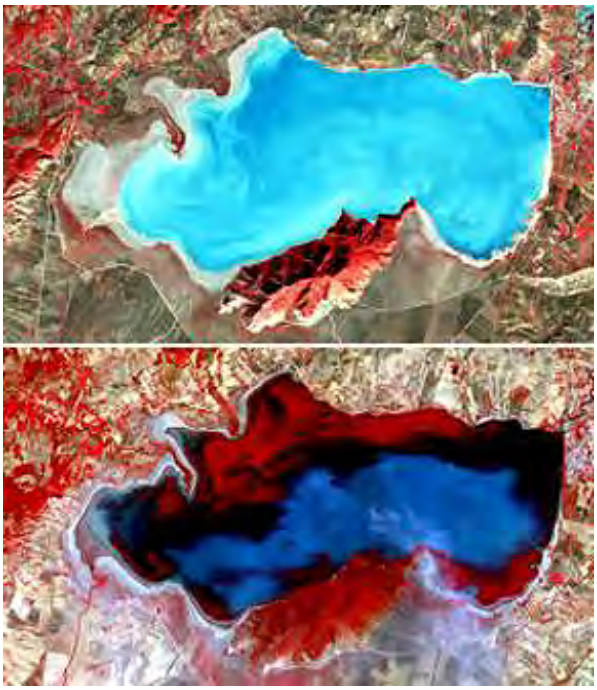
The trend towards shrinking of the Aral Sea has caused alarm in the Soviet spinoff countries of Kazakhstan and Uzbekistan, within which this body of water lies. It is an important resource whose loss would hurt fishing and other user groups. So, a plan underwritten by the World Bank to replenish water in the Aral Sea by damming its main outlet on the south has been executed. However, by August 2009 little has been accomplished; a continuing drought in the region has led to almost a complete drying up except for part of the northern Aral Sea.



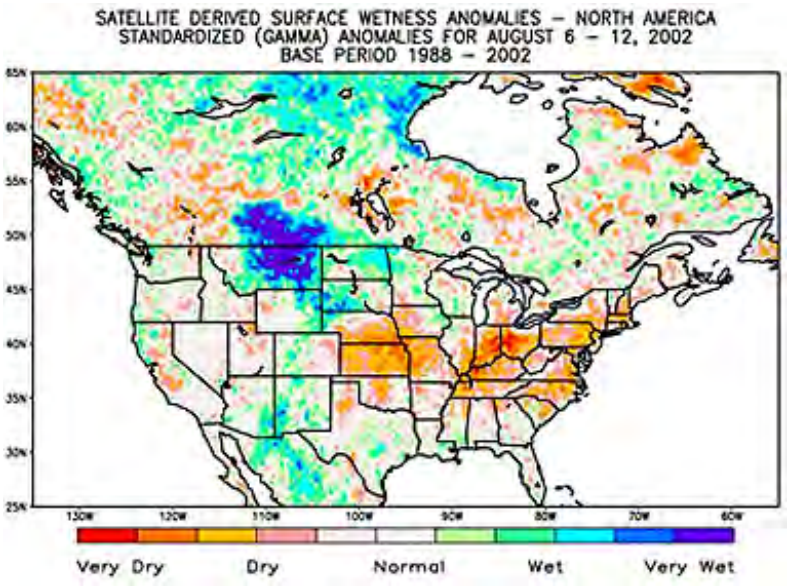
Further east, along the border between Afghanistan and Iran, the Hamound Wetlands were still in good shape in the early 1970s, with natural vegetation and some cropland benefiting from the water carried in by the Helmand River. The extended drought in that part of the world converted this wetland to a wasteland of mud and salt. These Landsat images show this abrupt change:



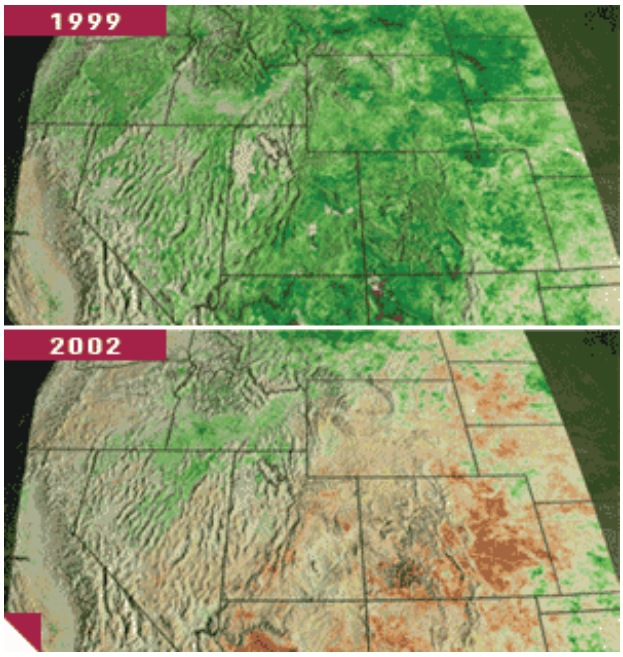
Space imagery is also of significant value in monitoring changing conditions within lakes, as we saw elsewhere in the Tutorial. Here are two ASTER images of Lake Ichkeul in Tunisia. The top one (November 14, 2001) displays the lake filled with sediments during a winter rainy season. The bottom image (July 29, 2005) examines a now clear lake that is being encroached by algae (red)(Utah Lake in the Overview-1 has a similar dual set of conditions):



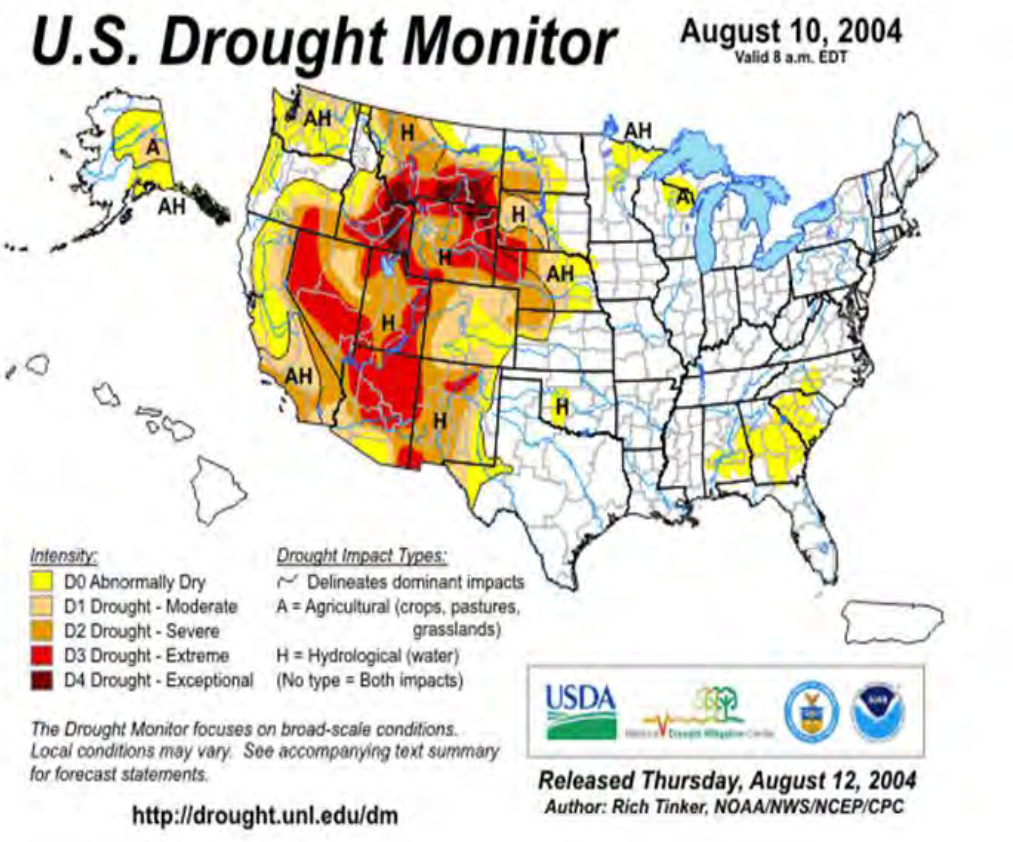
As discussed on page 3-4, the North American continent has been experiencing droughts of varied severities in the 1990s into the 21st Century. That occurring in the Summer of 2002 has been especially severe and widepread. Much of the East and a large percentage of the West were most seriously affected. One consequence: a large number of wildfires burning over 4 million acres. This map shows satellite microwave measurements converted into surface wetness across the U.S. for a 6 day span in August. This does not disclose the full extent of the drought (see page 3-4) because it reflects rainfall over just one week but it does help to assess the patterns of soil moisture that will influence longer term conditions that contribute to the water deficit underlying the on-going drought. For this week, the rain shortfall has been greatest in the eastern half of the country but in prior weeks the West has had similar deficiencies.

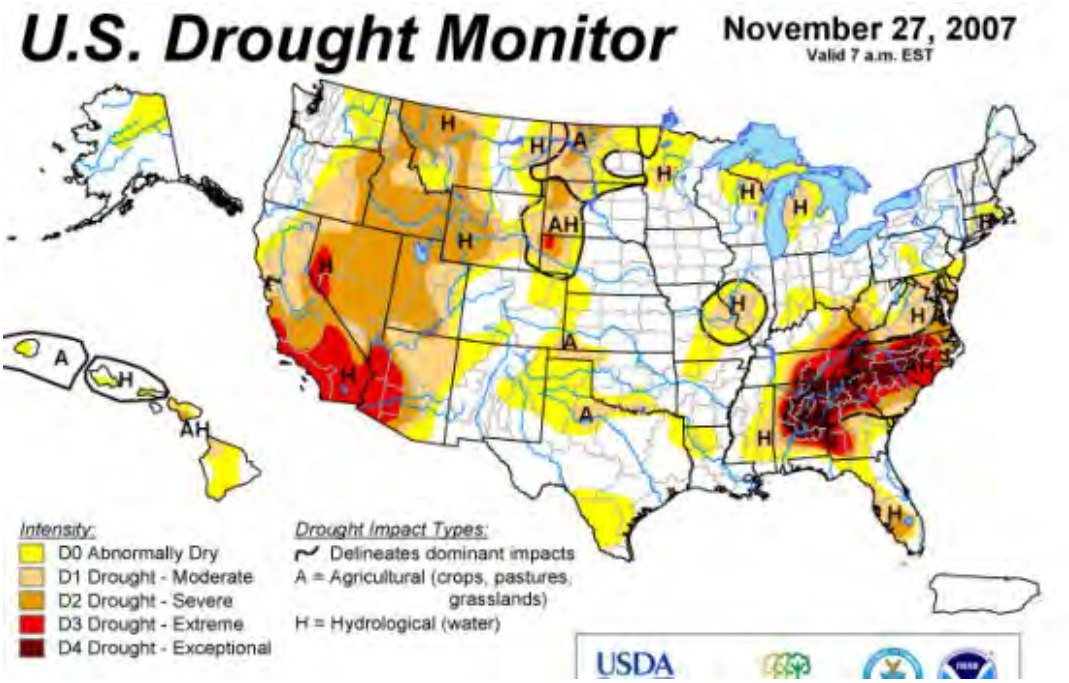


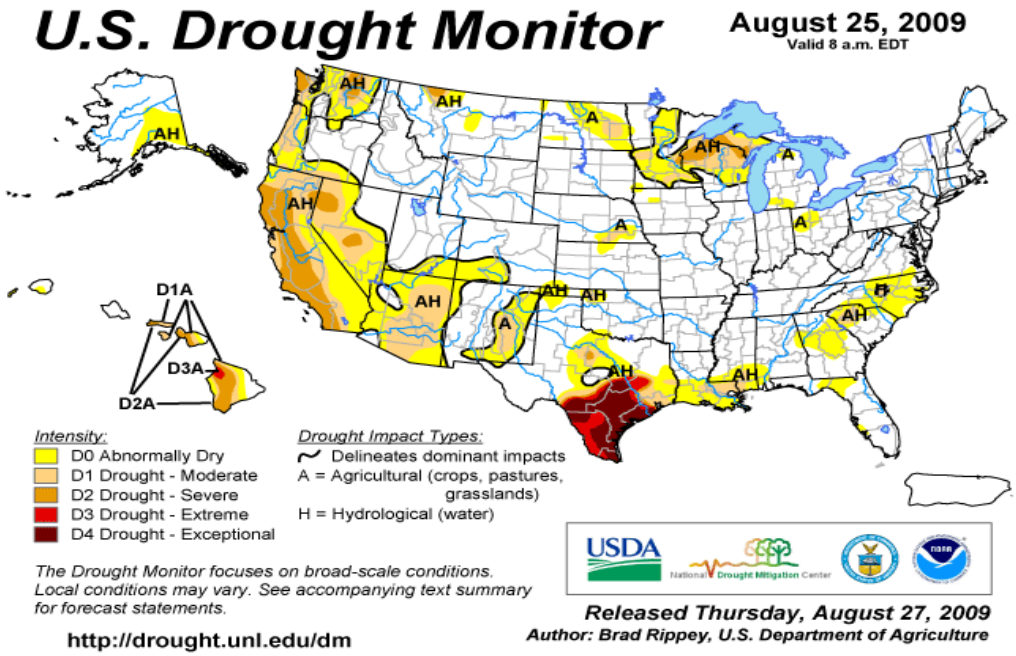
Thus, parts of the (48 contiguous) United States has been experiencing significant and worrisome drought conditions since the 1990s, largely caused by reduced rainfall and warmer temperatures. This pair of satellite images shows vegetation calculated as NDVI for parts of the western U.S. under normal rainfall conditions (top) and in a moderate drought state (bottom):



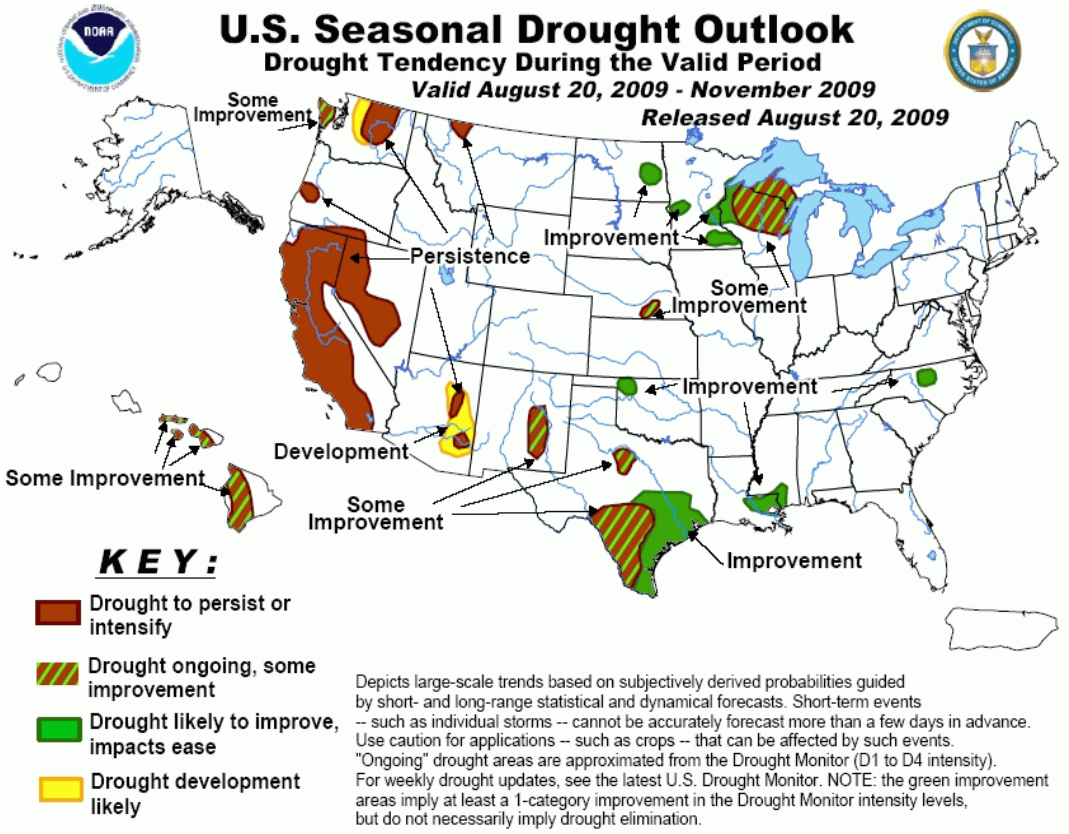
NOAA and the U.S. Dept. of Agriculture for decades now have been using satellite imagery routinely to publish maps showing the degrees of drought in the 48 contiguous states. Drought conditions vary from year to year and even seasonally. Here are three maps for 2004, 2007, and 2009:



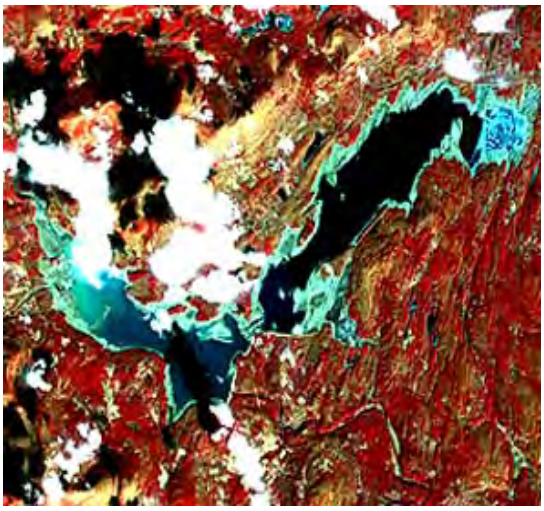




As one might expect, these same agencies make predictions of conditions at least three months in advance. This map shows anticipated drought conditions from August to November 2009, as a forecast:

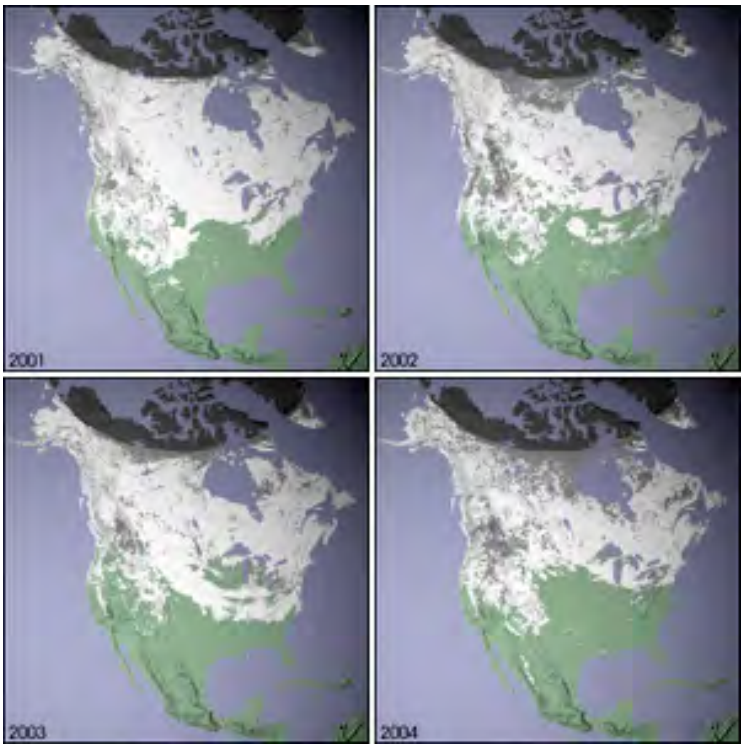


In much of the northeastern region, this drought has become troublesome and serious to severe right after the turn of the millenium. New York City and the surrounding metropolitan areas were forced to apply use restrictions as their major sources - reservoirs - shrank to as low as 10% capacity by volume. The next pair of images, acquired by the ASTER sensor on NASA's Terra, shows a reservoir in the Catskills to the northwest on September 18, 2000 and again on February 3, 2002, when the volume has been reduced to 58% of normal.

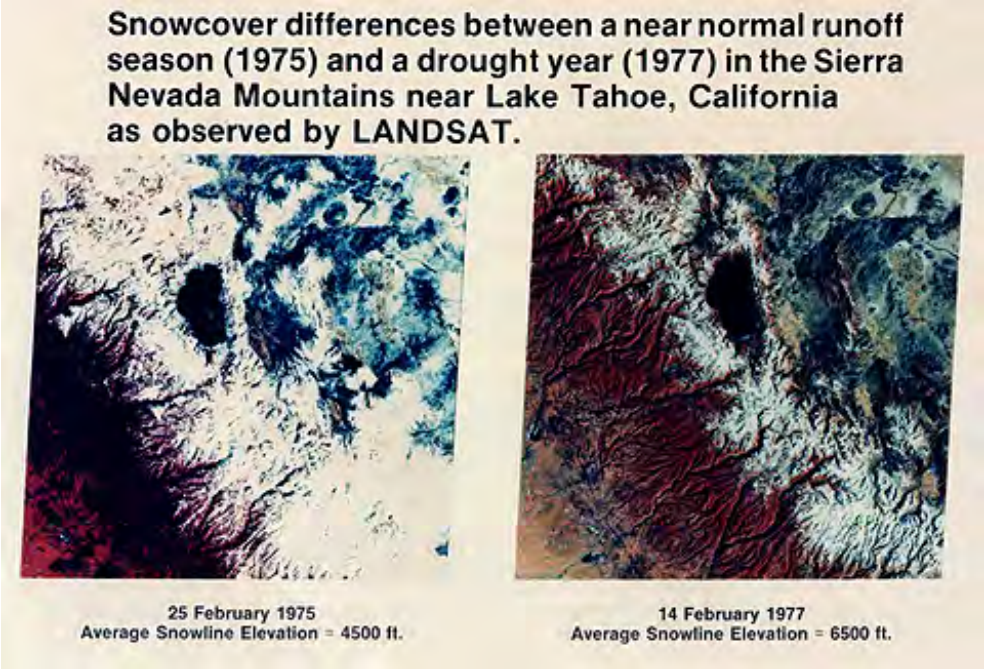


Obviously, at the other extreme Metsats can provide near real-time indications of weather conditions that portend severe storms and heavy rainfall. Over a longer period, satellite imagery can show potential flooding from spring thaws and image interpreters can estimate expected quantities of water runoff by monitoring snow cover over large regions. Satellite observations of surfaces blanketed by snow (in the U.S. principally in mountains and high prairies) suffice to measure the areal extent of the masses likely to melt. Meteorological satellites are particularly adept at recognizing snow - it appears bright white and usually continuous over large areas (clouds are either discontinuous or a dull gray). This is a typical Metsat view:

These four MODIS images show the extent of snow cover in different years:



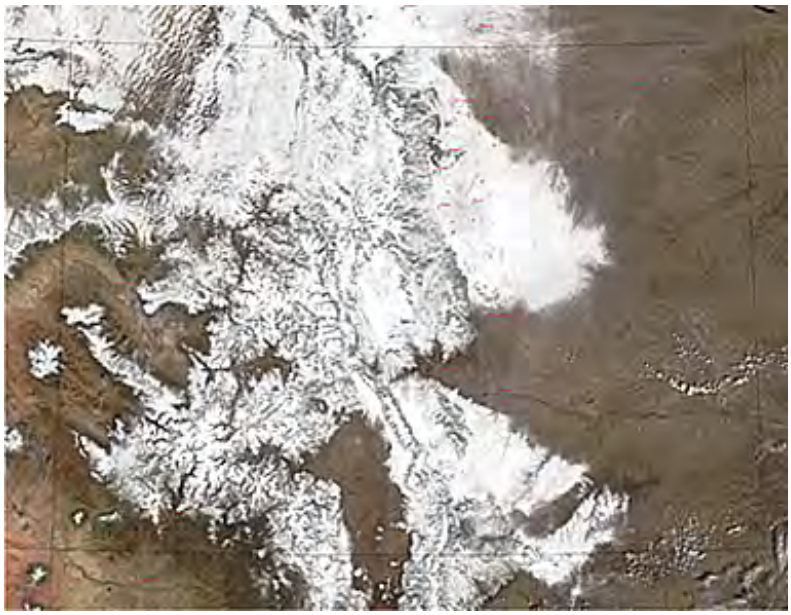
Seasonal variation in snow cover from year to year in a smaller region is illustrated by this pair of Landsat-1 images of the central Sierra Nevada highlands.



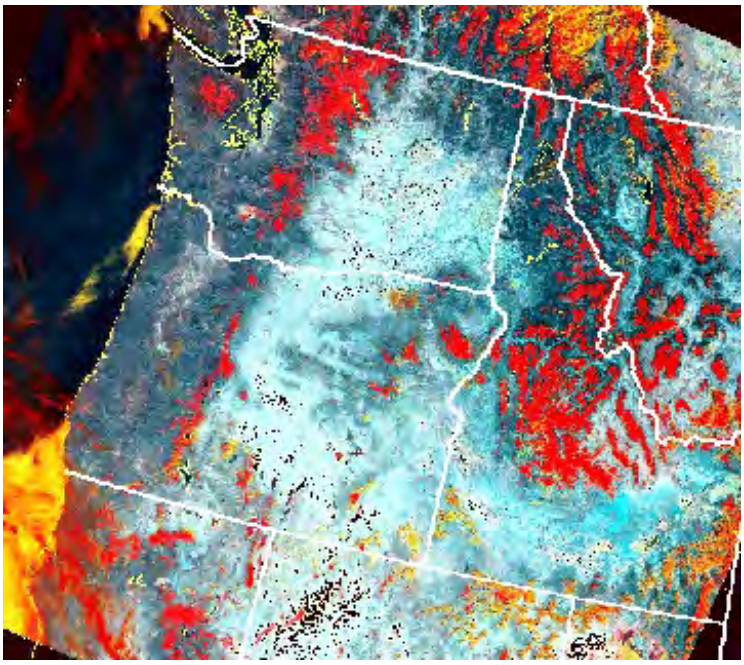
However, we must determine thickness variations and packing densities from onsite ground measurements, in order to estimate the volume of runoff. This information is important not only for flood warnings and control but also to estimate water supply from reservoir fillings, river channeling, and aqueduct retrieval.

14-39: **Make an estimate, using the above Landsat image pair, of the amount of increased water one could get from the melting of the 1975 snowpack in the Sierra Nevada compared with 1977, assuming the snow is melted entirely.** **ANSWER**

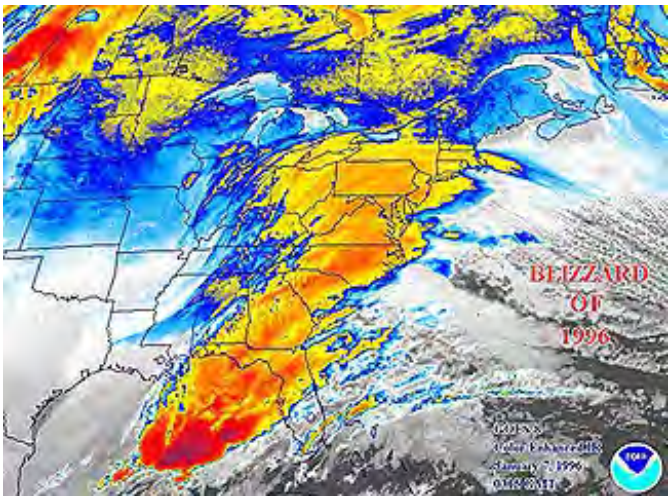
Similar estimates are made routinely to forecast runoff from the Southern and Central Rocky Mountains. This MODIS image shows the distribution of a record snowfall on March 22, 2002 in the Colorado Rockies. Most of the white is snow (a few clouds are in the scene).



Routinely, images from the NOAA and the GOES satellites are applied to determining the extent of snow cover. We can separate snow from clouds by differences in spectral absorption in longer wavelength bands. Both are highly reflective in the visible and photographic IR. The Thematic Mapper band 5 also can distinguish snow/clouds, and thermal responses vary, as well. Here is a color composite made from three NOAA-12 AVHRR bands (1,3,4 = RGB) showing snow in April 1995 in the northwestern U.S. as red:



Major snow storms can cause heavy cover that leads to extensive flooding, if a rapid thaw occurs. Look at this GOES-8 coverage (top, below) of a massive storm that hit the northeastern U.S. on January 7, 1996. The bottom image taken after storm passage and return of clear skies shows the extent of snow cover.



14-40: **Below (south) of the edge of the snow cover, the land is shown as green. What happened to it during the storm? ANSWER**

The MODIS sensor on Terra caught a distinctive snow pattern in the central U.S. following a "clipper" storm on December 2, 2006. The snow is deposited in a narrow band that developed rapidly across Missouri into northern Illinois. This resulted when cold polar air met warm Gulf air (moisture-laden), making snow, with the front being guided by the jet stream moving eastward.

We are accustomed to think of snow as always white in appearance. But snow can take on other colors, notably blackish, to brown or tan. This ASTER image of the snow capping the San Juan Mountains of southwestern Colorado shows the snow to have a brownish color. This is real

and not a lighting artifact. It is caused by a strong coating of dust blown into these mountains from the desert to the west. The reflectance of the snow is reduced, hence it tends to absorb solar heat and thus melts sooner than normal.

PREVIOUS NEXT

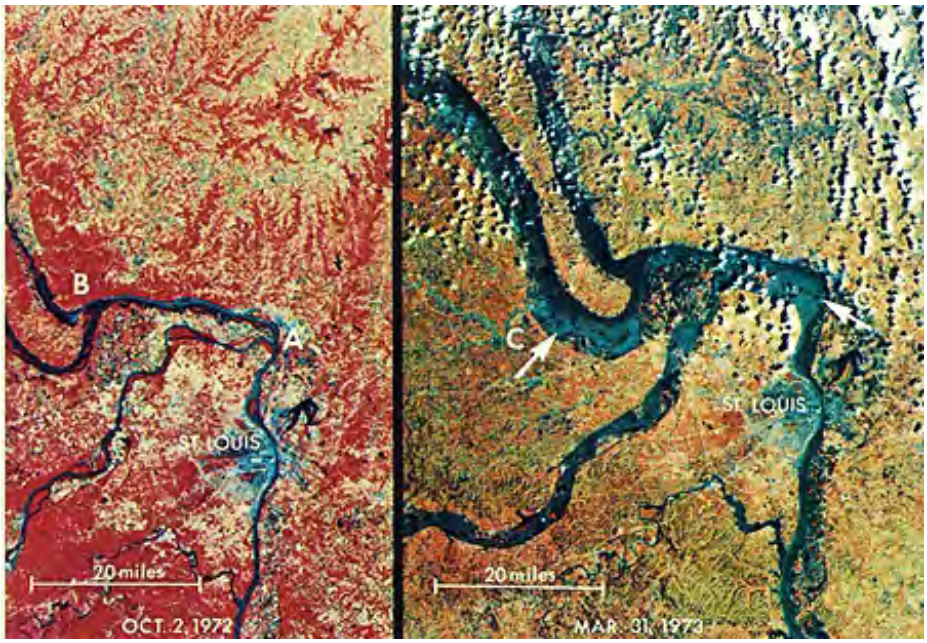
Primary Author: Nicholas M. Short, Sr.

This final page concentrates on various floods that were well monitored by space sensors.

Mapping the Extent of Flooding

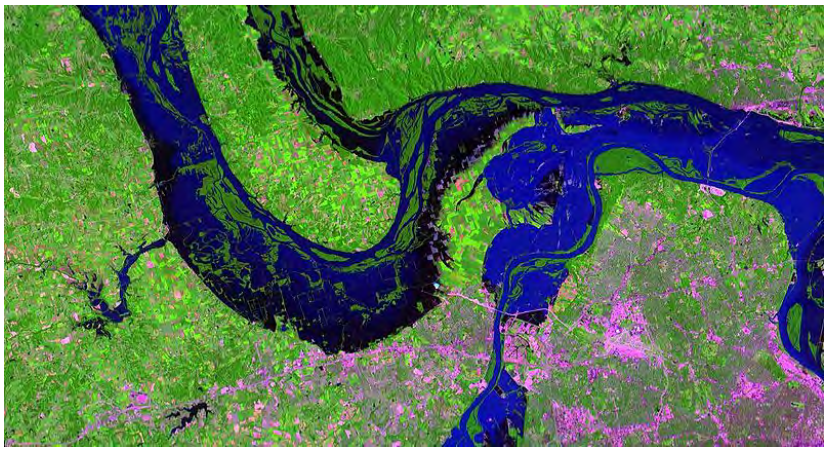
This kind of information was a factor in predicting the major flood in the northern Midwest, late in the Spring of 1997. Satellite imaging gave a critical look at the great floods on the Red River in North Dakota and Manitoba (Canada) that inundated Grand Forks, Fargo, and other towns along the state's border with Minnesota.

Spring flooding is frequent in parts of the Mississippi River basin. A hundred-year flood, i.e., largest expected statistically in a 100-yr span, resulted from snow melt and rain in late March of 1973. This Landsat-1 subimage (with an earlier pre-flood view), captured the extent of flooding on a cloud-free day, showing St. Louis, Missouri (protected from downtown flooding), and the flood plains of the Mississippi, Missouri (joined at A), and Illinois (at B) rivers:

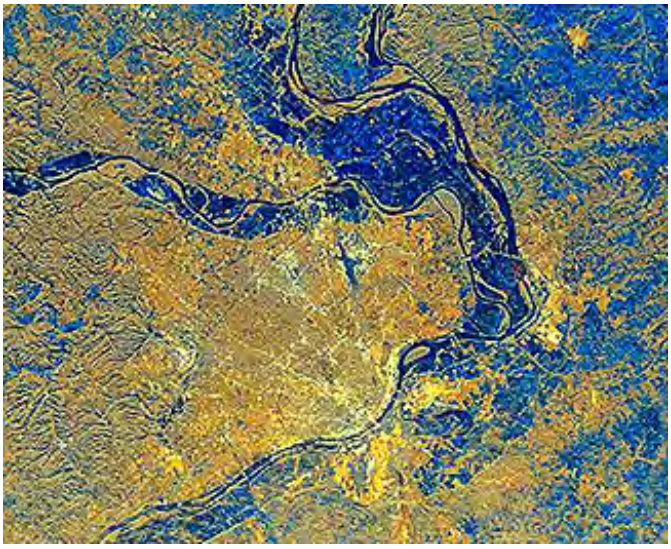


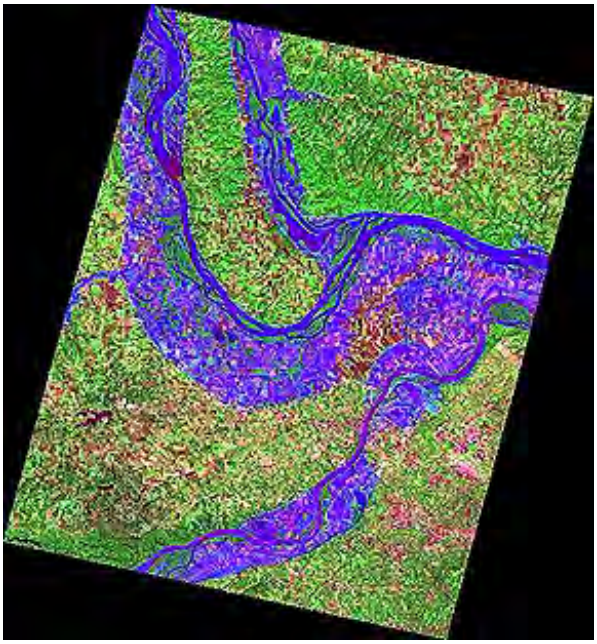
Twenty years later, the Midwest again flooded, worse than before, in fact many places recorded the highest flood levels in their history. After several months of excessive rain that saturated the soil, because of a blocking high pressure system that kept the jet stream relatively stationary, in late July and August of 1993, water levels rose well above flood stage. Areas hardest hit were from Iowa to southern Illinois. Levees broke, inundating tens of thousands of acres. The '93 flood, said to be largest ever on the Mississippi, became the costliest in U.S. history (some estimates approach $15 billion). Satellite imaging played a key role in getting a number of good images of the flooded area.

Once again we examine the lowlands northwest of St. Louis. First is this Landsat image:



The next image was taken by Shuttle astronauts, using SIR-C radar. (Note that the orientation has west near the top.) Below that is an image of merged JERS-1 radar and a SPOT 3-band composite, which offers considerable detail (notice how farmlands show through the water).

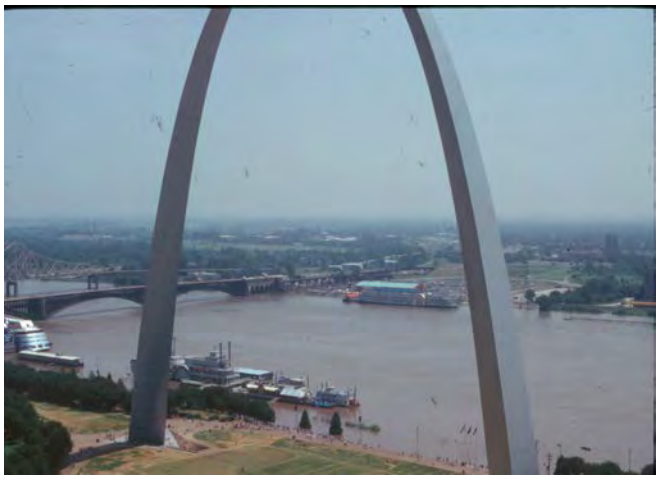




14-41: **Which year does the Mississippi flood seem worse? Why isn't St. Louis flooded?**
ANSWER

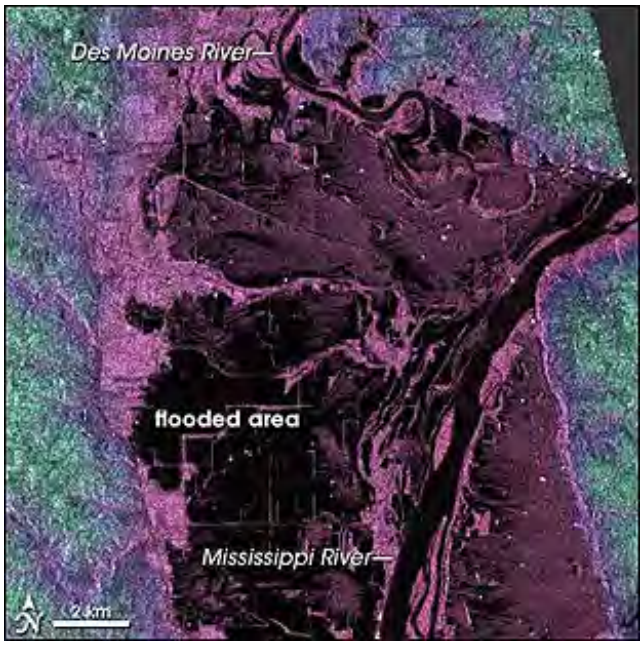
The flooding of 1993 affected many rivers. Floods occur as early as June and as late as October (see this NOAA website for an historical review). The next three photos show the flooding at St. Louis (on August 1 the flood peaked at the waterfront at 49.6 ft, 19 feet above flood stage and

6 feet higher than during the 1973 flood shown above), Jefferson City, MO, and Alton, IL:



Aerial view of the Missouri River flooding on July 30, 1993, in the vicinity of Cedar City and Jefferson City Memorial Airport immediately north of Jefferson City, Missouri, looking south (photography from the Missouri Highway and Transportation Department).

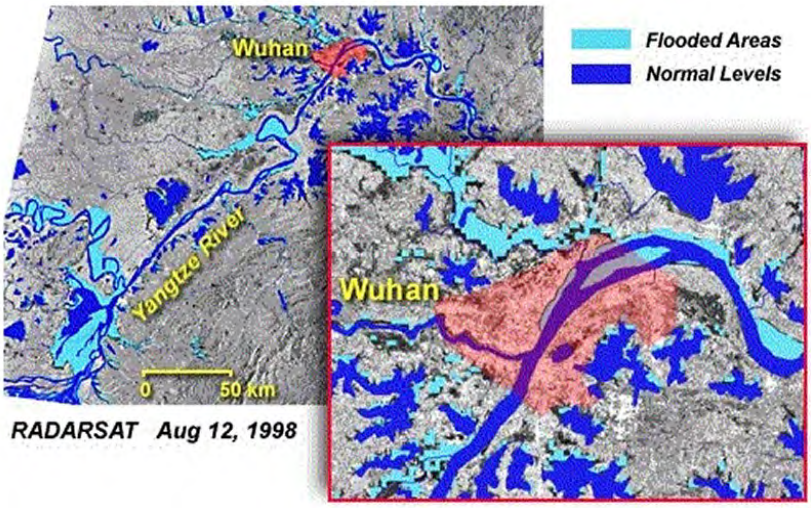
Up river during the same 1993 flood, the SAR radar on ERS-2 rendered the flood in mostly black tones in this scene near Dubuque in southeast Iowa:



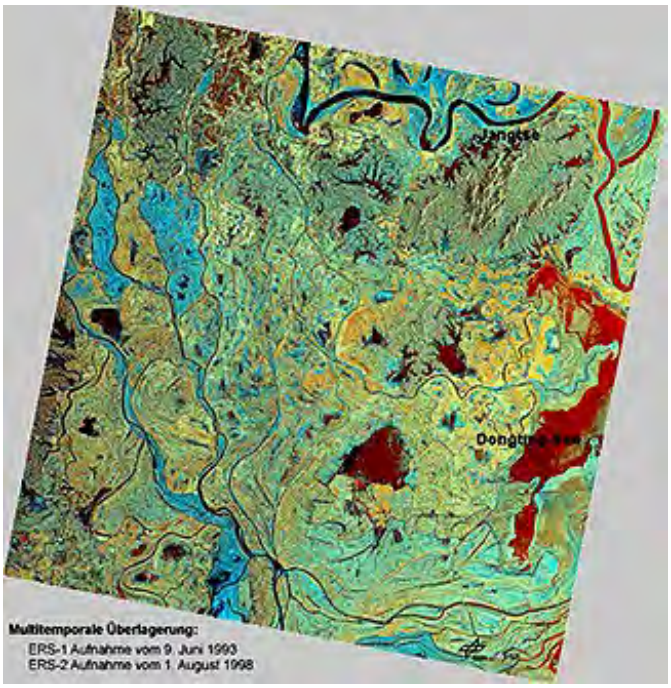
The effects of the 1993 persisted in the lowlands near St. Louis well into 1994, as indicated by this March 1994 photo taken from the Space Shuttle:

Flooding can occur anywhere on all continents except the Antarctic. The next image is a Landsat-1 subscene (February 6, 1974) of the Barcoo River in Queensland/South Australia, flooded by Fall rains. The floodwaters have spread greater than 50 km (31 mi) wide in these low-lying plains, with low rolling hills.

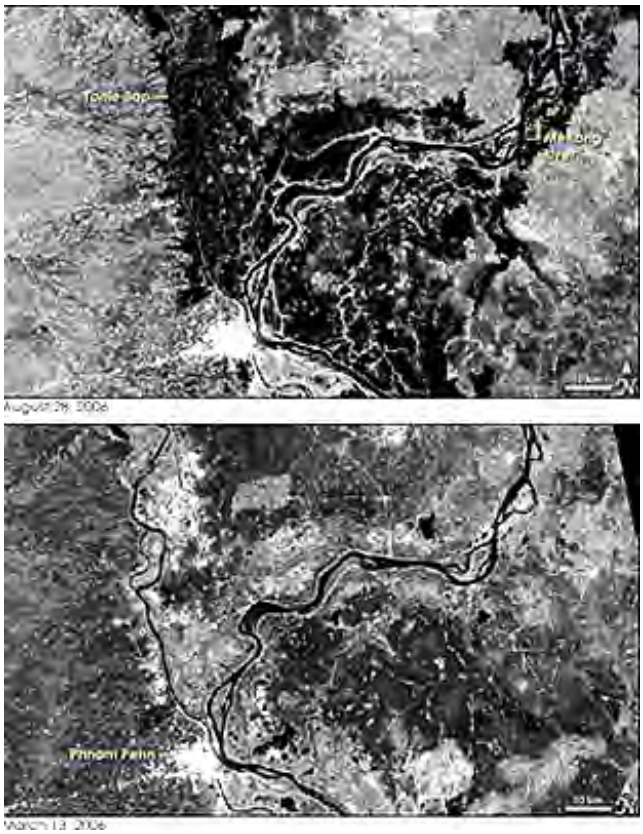
The Yangtse River in China underwent a major flood in August of 1998. Millions were driven from their lowlands homes. This Radarsat image shows the flooded lands near Wuhan:



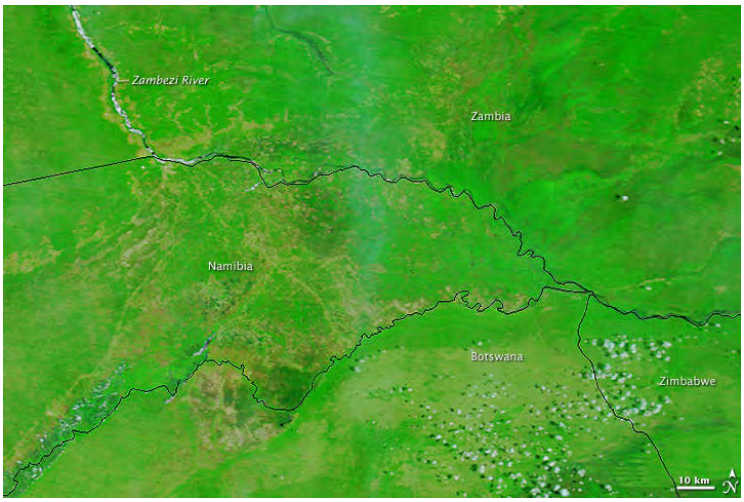
In this unusual image, an ERS-1 radar image taken during June of 1993 is joined with an ERS-2 radar image taken on August 1, 1998, providing a multitemporal or change detection rendition. Both blue and red associate with flood waters.



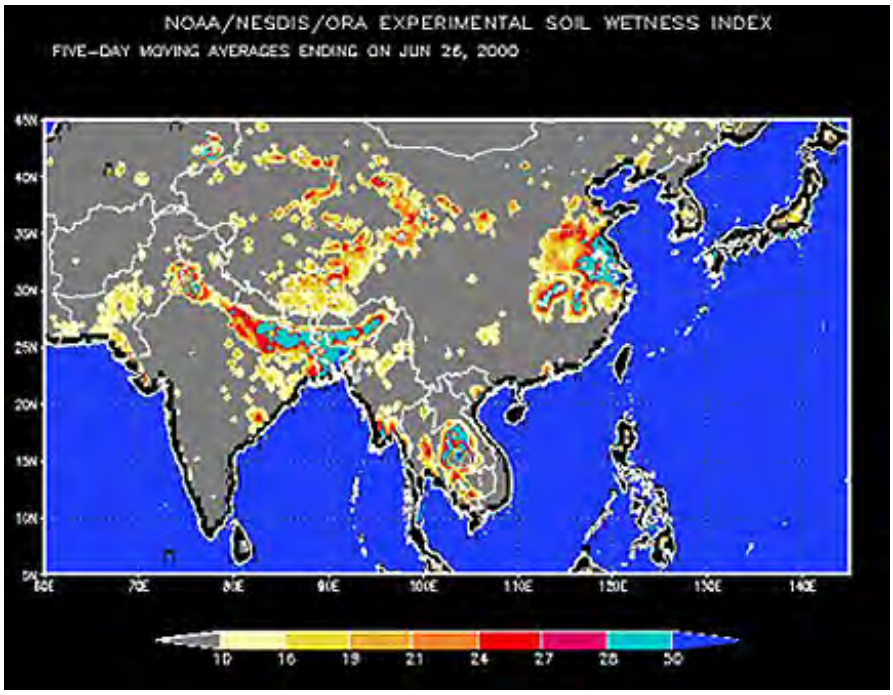
Radar is especially powerful in recognizing floodwater extent since the low reflectivity from water makes it appear dark. Floods can be beneficial as well as harmful. These two Radarsat images show annual flooding in the Mekong River basin in Cambodia. Waters from the monsoon rains collect in the basin and replenish the ubiquitous rice paddies that provide the main food supply for the Vietnamese and Cambodians.



The pair of MODIS images below show parts of Namibia, Botswana, Zambia and Zimbabwe in which the rivers are normal and then in flood:



Finally, check this map showing areas with high soil moisture, a condition that bespeaks of water saturation from earlier heavy rainfall or from previous flooding. This soil wetness map shows much of Asia; maps of all the continents plotting this parameter are made by NOAA/NESDIS.



This lengthy Section 14 purports to convey that the principal use of remote sensing remains surveillance of weather systems and oceans on local-to-global scales. We report this because of the widespread occurrence of water on the Earth's surface (even greater than the 70+% ocean surface area, stated on page 14-1, if we include the Antarctic ice [which stores more than 80% of the world's {frozen} fresh water] and Greenland.

At this point in the Tutorial, we have examined most of the specialized modes of remote sensing (defined by the electromagnetic spectral regions we can use), the spacecraft systems that mount the sensors, and the numerous applications to which these sensors have contributed. In the next Section (15), on Geographic Information Systems (GIS), we look at some systematic ways to integrate remote sensing data into organization, correlation, interpretation, and management of geographically-referenced information. Then, in Section 16, we look ahead to some of the current or recent remote sensing programs, in which individual satellites will simultaneously make meteorological, oceanographic, land surface, and biologic observations to present a unified picture of Earth as a System.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

METEOROLOGY - WEATHER AND CLIMATE: A CONDENSED PRIMER

PREPARED BY DR. NICHOLAS M. SHORT

This primer is dedicated to my first mentor at NASA Goddard Space Flight Center, Dr. William Nordberg (now deceased) who hired me at a critical time in my professional life and thus brought me into the field of remote sensing. Trained and practicing as a meteorologist, Dr. Nordberg (a native of Graz, Austria), took a lead role in the development of ERTS (Landsat) as he expanded his contributions from weather satellites to those doing Earth Observations for a much greater range of applications.Earlier, Bill, as all of his colleagues called him, had played important roles in the Nimbus and TIROS programs. Dr. Nordberg was one of the finest scientists and leaders I've ever served under. I am especially grateful for his tolerance of my demands for more field work - our arguments were vociferous but always controlled by a mutual friendship. By the time I accepted his offer to transfer to NASA (from a National Academy of Sciences Associateship), he had become Director of the Earth and Atmospheric Sciences Directorate. He took a special interest in transforming me from a ground geologist to a space geologist.

I. THE PHYSICS, CHEMISTRY AND ENERGETICS OF THE ATMOSPHERE

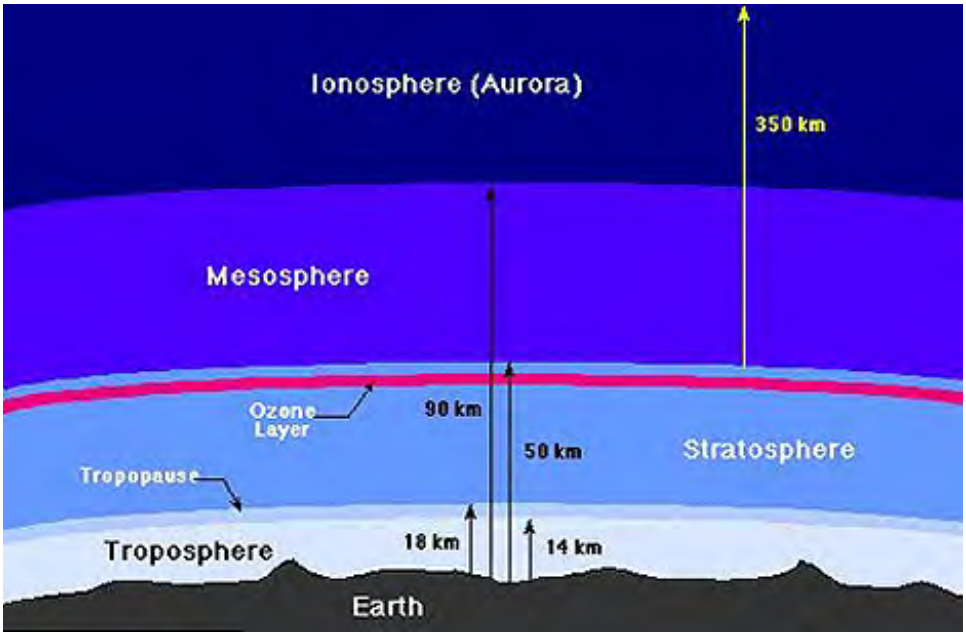
Meteorology (from the Greek *meteõros* meaning "high in the air") is defined as the science of atmospheric phemomena and processes. ***Weather*** denotes the short-termed (hours to a few days up to a week or so) behavior of the atmosphere, generally with the connotation of applying to local or regional parts of the Earth's globe, as it varies in the conditions used to describe weather - fair, rain, warm, windy, etc. Thus, each place is said to have a set of particular conditions, which tend to change over short time spans, that affected people describe as the weather for the day or perhaps as long as the next week. ***Climate*** refers to much longer time frames and describes the common characteristics of weather in broader parts of the Earth's globe. It applies to wider regions and depends on geographic location, physiographic conditions, time of year, and other factors. Climate is thus tied to 1) the larger, longer variations in typical or average weather in a region determined by the seasons - which in turn depend on the location of the Earth, with its tilted axis, as it rotates around the Sun in an annual cycle; and 2) the range of conditions expressed in day to day weather variations and extremes in such properties as temperatures, extent of cloud cover, and duration and types of rainfall/snowfall.

On Earth, the atmosphere is also known as "air" (with the Space Age we now know that some other planets and rarely even their moons have atmospheres; in this Primer we restrict our discussion to the Earth alone). The terrestrial atmosphere is part of the four "spheres": 1) the Lithosphere; 2) the Atmosphere; 3) the Hydrosphere; and 4) the Biosphere. These can, and do, interact and influence one another. The atmosphere itself is a mixture of gases of different compositions, plus water as vapor, liquid, or solid, and suspended particles of various nature. These gases are subject to the rules of ideal gas physics (e.g., Charles and Boyles Laws). Thus, the physics of meteorology becomes the physics of gas behavior as described by temperature, pressure, circulation, density stratification, and energy utilization and transformation. The main

(predominant) source of energy powering atmospheric behavior and change is solar radiation from the Sun. Also contributing to thermal inputs (or exchange) is terrestrial heat flow, in and out movement of heat from water reservoirs (mainly, the ocean), inputs from humans, animals, and plants, and chemical reactions within the atmosphere which can be exothermic or endothermic.

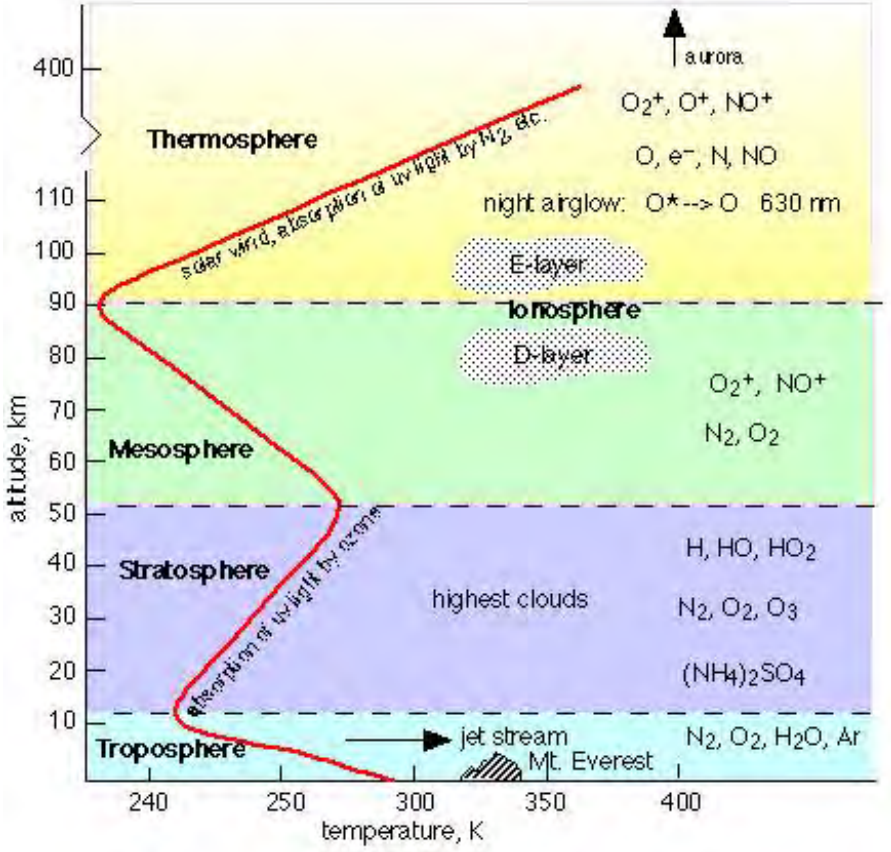
In the brief synoptic treatment of Meteorology that follows on these next pages, the primary sources of information and of illustrations has been taken from the Internet and from ***The Atmosphere: An Introduction to Meteorology****, 4th Ed., by F.K. Lutgens and E.J. Tarbuck, Prentice Hall, 1994 and* ***Physical Geography of the Global Environment****, by H.J. de Bilj and P.O. Muller, John Wiley, 1993*

To begin our survey, lets introduce the terminology that labels the main structural components of the atmosphere as a function of height above the surface:



We live within the troposphere, within which most of both weather and climate activities take place. Jet aircraft are capable of flying in the upper troposphere and lowermost stratosphere (where the air is less dense [less drag on the plane and hence greater speed] and smoother, often cloud-free wind conditions) and experiment rocket planes have penetrated up to 90000 feet into the lower stratosphere. Knowledge of the air above this has come from balloons, sounding rockets, and orbiting satellites, as is considered in this Section.

The layered subdivisions of the atmosphere (page 14-1) are shown again in this diagram which is here primarily to show the elemental species that are distributed in these layers and the general temperature profile from surface to the outer atmospheric fringes.



The increasing thinness of the atmosphere above the Tropopause brings into play different physical phenomena that alter temperature patterns. In the general temperature profile diagram, the pattern is decrease --> increase --> decrease --> increase. In the Stratosphere, the presence of a very small amount of ozone (O_3) has a significant influence on temperatures since it absorbs heat from ultraviolet radiation. This temperature gradually increases until the stratopause (roughly equivalent to the height at which ozone becomes fully depleted) where a temperature inversion (inflection) brings about a resumption of adiabatic temperature drops down to about minus 90°C. This continues through the mesosphere. Another reversal leading to significant rise in temperature begins at the mesopause extending through the thermosphere. Even though its atmosphere is tenuous (very low densities), the presence of tiny amounts of oxygen and nitrogen is sufficient to absorb short wavelength solar radiation, leading to the positive temperature gradient. Levels as high as 1000° C are reached owing to the ability of moving atoms and molecules to attain higher speeds in the very low density atmosphere which means they are more energetic and hence this kinetic energy transforms into increased temperatures.

The composition of the troposphere is dominated by two elements in their gaseous molecular forms, N_2 and O_2. This table lists all notable atmospheric constituents, their relative amounts (in parts per billion), their sources, and their variations over time:

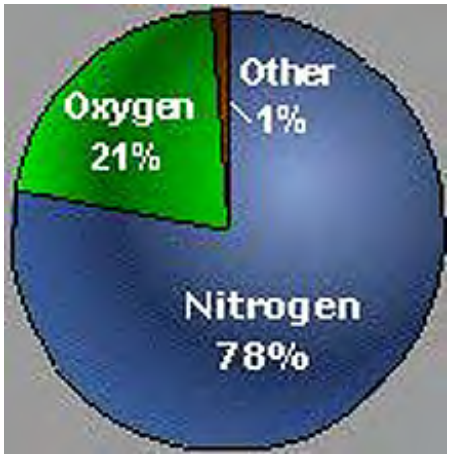
THE COMPOSITION OF THE TROPOSPHERE

species	relative abundance (parts per billion by volume)	source	comment
N_2	7.81×10^8	biologic	long lived
O_2	2.01×10^8	biologic	long lived
H_2O	$10^6 - 10^7$	physical	long lived
Ar	9.34×10^6	radiogenic	permanent
CO_2	3.5×10^5	biologic, industrial	variable, increasing
Ne	1.8×10^4	interior	permanent
He	5.2×10^3	radiogenic	escaping
CH_4	1.6×10^3	biologic	variable, increasing
Kr	1.0×10^3	interior	permanent
H_2	5.0×10^2	biologic, photochemical	variable
N_2O	3.0×10^2	biologic, industrial	increasing
CO	1.0×10^2	photochemical, industrial	variable, increasing
SO_2	$<10^2$	industrial, photochemical	variable
O_3	$<10^2$	photochemical	variable
Xe	9×10^1	interior	permanent
NO, NO_2, NO_x	variable	industrial, biologic	—*
CH_3Cl	6.0×10^{-1}	biologic	short lived
CCl_2F_2	2.9×10^{-1}	industrial	increasing
CCl_3F	1.7×10^{-1}	industrial	increasing
CCl_4	1.2×10^{-1}	industrial	increasing
CH_3CCl_3	9.8×10^{-2}	industrial	increasing
CF_4	7.0×10^{-2}	industrial	increasing
CH_3Br	1.0×10^{-2}	biologic, industrial	possibly increasing

*Relatively short lived, with an average lifetime of roughly one month.

The two most notable compositional numbers are N_2 = 78.1% (7.81×10^8 divided by $10^9 \times 100$) and O_2 = 20.1 %. Some of the nitrogen was derived originally from escape from the Earth's interior but biologic processes have increased its percentage over time. Oxygen has increased its percentage over time largely through photosynthetic processes, starting about 1.5 billion years ago. Argon (derived from radioactive decay of potassium in rocks) is the next most abundant component, about 1%, but can sometimes be matched by water vapor (0.1 to ~1.0%). Carbon dioxide (CO_2) and Ozone (O_3) play small but important roles. Many of the other "trace" gases have either a biogenic origin, a volcanic origin, or result from human activity.

This simple pie chart diagram should help to simplify the generalization of atmospheric composition:

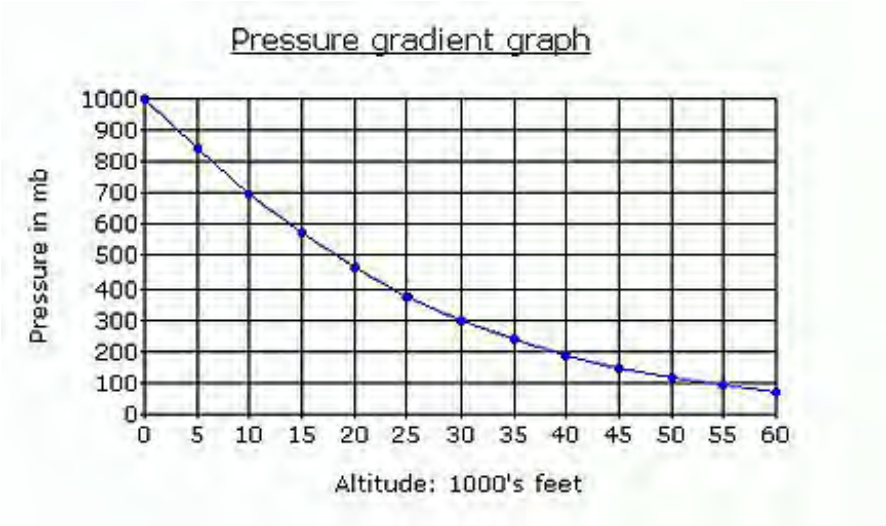


The composition of air up to the outer limits of the stratosphere (out to about 70 km) is almost constant and homogeneous insofar as the three dominant species is concerned. Over time the

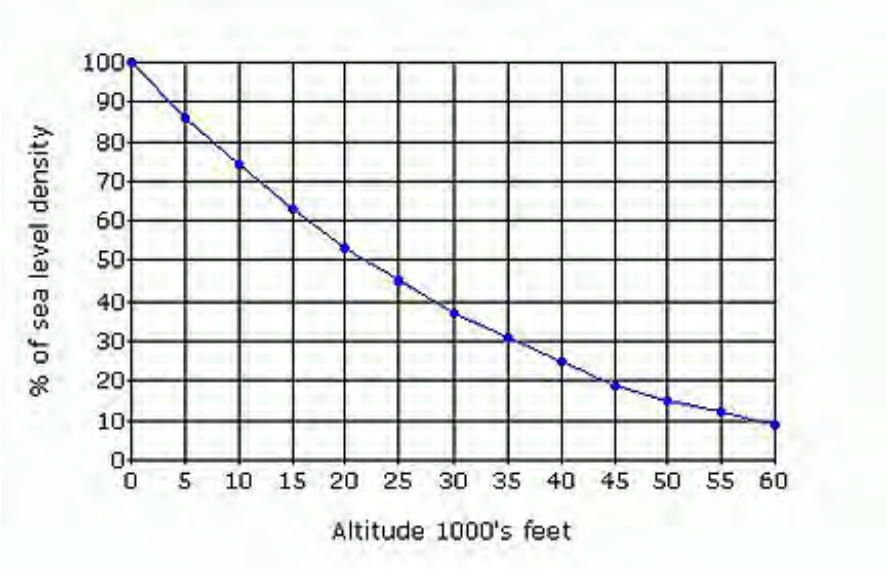
addition of CO_2 and possibly O_2 slightly lowers the percentage of all other constituents (since they must be summed to 100% as a constant). In the upper atmosphere, which has a much less amount of all constituents (leading to lower densities), some of these constituents have enough thermal energy to excape into outer space (hydrogen and helium being light are the principal gases that so leave).

Over the 4 plus billions of years in which Earth has had an atmosphere, its composition has changed significantly. The early atmosphere was largely ammonia and carbon dioxide plus some nitrogen, with little oxygen, but may have had more hydrogen.

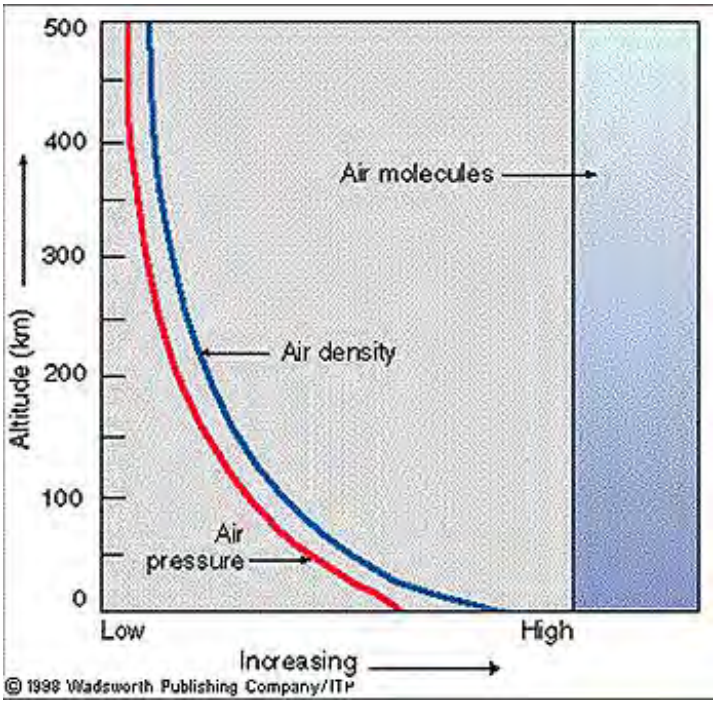
The total mass of the global atmosphere is estimated to be ~5.1 x 10^{18} kilograms. Ninety percent of that mass is in the inner 10 km (6 miles). The average density of air at the Earth's surface is 1.2 kg/m^3. The mean pressure of a column of air at sealevel, coming from gravitational attraction on all gases above it to the outer limits of the exosphere (where gas molecules diminish to the levels of outer space) is set at 1.013 bars, or more commonly expressed as 1013 millibars. A general pressure gradient for the atmosphere out to about 11 miles (18 km) is shown below:



Note that this and the next plot are exponential curves. At about 18000 feet (5.45 km), the pressure is 500 millibars, about half that at se level. This means that most of the mass is concentrated in the troposphere. (For pressures lower than about 400 mb, the amount of oxygen present is insufficient for humans to function for a long period without a supply of oxygen from tanks (climbing Mt Everest is possible without oxygen tanks but dangerous and requires that one stays for only a short time when above the 28000 ft level.) Air becomes progressively thinner with altitude, i.e., its density diminishes as shown in the graph below (which closely follows the pressure curve).



The parallel behavior of diminishing pressures and densities out to 500 km follows this same general exponential pattern:



These pressure and density changes are caused and driven by thermal energy. The atmosphere is almost entirely heated by incoming solar radiation which interacts with air to produce specific temperatures (heat from the subsurface Earth makes a small contribution). Gas expands and contracts in response to variations in temperatures. The differing temperature states in the atmosphere, both horizontally and vertically, give rise to winds, clouds, storms, etc. Movement of the atmosphere is also initiated by pressure changes that result from temperature changes.

In the troposphere, the atmosphere follows the normal behavior of the physics of gases. The change of pressure states just described can be determined by applying the Gas laws. These were formulated between 1660 and 1800. The gas laws were determined experimentally on simple gas systems, in which thermal energy was neither removed nor added. These laws bear a kinship with the first Law of Thermodynamics, which in this form is: A gas's temperature can be changed by gain or loss of heat, or by modifying the pressure (expansion; compression), or a combination thereof. A corollary: If the temperature is increased (decreased) not by heat energy addition (subtraction) but by work done on it (or it does to its surroundings), the change is said to be *adiabatic.*

The first gas law is known as *Boyle*'s Law (for a constant temperature) which is written as PV =

k_1 (k is a constant), or its proportional form:

$$P_1V_1 = P_2V_2$$

Thus, if the volume is increased twofold, the pressure will decrease to 1/2 its previous amount to maintain the equation's equality. That is, if V_2 becomes equal to 2 V_1, then P_2 must decrease by 0.5 (50%) to maintain equal multiplier products on both sides of the equation.

The *Charles* Law, $V/T = k_2$, in which pressure is kept constant, can be written $V_1/V_2 = T_1/T_2$. So, to maintain equality, if a Volume is contracted (V_2--->$0.5V_1$), then the corresponding T_2 must be $0.5T_1$. This results in cooling. Or, again, if a gas is allowed to increase its temperature, its volume will expand. The third relevant law is that of *Gay-Lussac*: For a constant volume, $P/T = k_3$ or $P_1/P_2 = T_1/T_2$. Under those conditions, if the temperature is increased, the pressure will also increase. (In the kinetic theory of gases, increasing the temperature makes the gas molecules move faster and hence their impact on surfaces transfers more energy and increases the force [over area] and therefore the pressure; likewise, increasing the volume increases the pathway distances of the molecules and allows them to gain more momentum to strike harder.)

It can be daunting to understand these two laws in the physical sense. Perhaps these remarks will help: Why does the pressure increase when a gas is compressed to a lower volume? The same number of gas molecules now occupy less three-dimensional space. The number of molecules in motion that are striking any specific area on the now smaller container (lower volume) will increase. Thus, the force exerted by this larger number is greater; hence the pressure on the area is higher. What about the temperature effect? As the volume is decreased (compression), the molecules become more crowded. This results in a lowering of the path length of a moving molecule. Its free motion is thus reduced. Since temperature is a measure of the kinetic energy KE (the Boltzmann equation: KE = 3/2kT), the reduction in motion means less energy and a lower temperature. Conversely, expansion results in a lessening of collisions, greater path lengths, and thus more KE and higher temperatures.

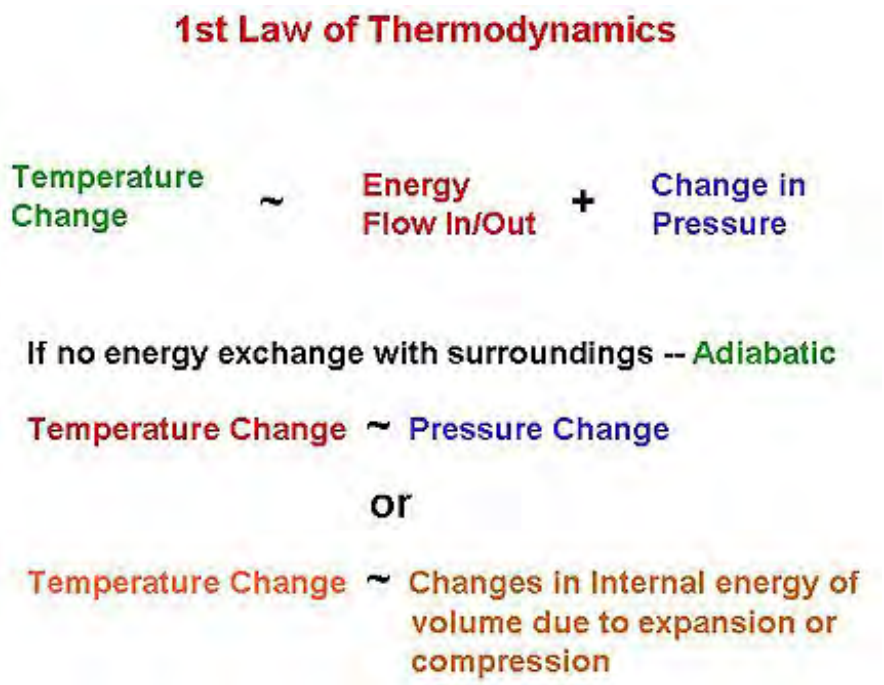
The behavior of gases can also be expressed in the form known as the Ideal Gas Law: $PV = nRT$. where n = the quantity of gas (in moles) and R is the gas constant of proportionality (in the SI system, R = 8.314 J/mole-K). (Note that T in this formula is absolute temperature in Kelvins K = C + 273; P is in Pascals). Using this formula, it is evident that if V increases (gas expansion), temperature likewise increases; if P decreases, the T falls - as indeed it does moving upwards through the troposphere. Also, if n is divided by its volume, the result is ρ, the density. The Ideal Gas Law becomes $P = \rho RT$. For the case in which T is held constant, the pressure will increase with any increase in density (as we shall learn later, denser air masses also are higher pressure systems).

The Gas Laws do not apply in the same way if all three variables, P, V, and T, are allowed to change freely. In particular, the notion of volume increase leading to temperature increase (Charles Law) would seem to be violated. But consider this: If the gas is compressed (from externally applied pressure not the same as its initial internal pressure), its volume will decrease and from Boyles Law its internal pressure will increase. That increase in P will, from Gay-Lussac's Law, lead to an increase in T. The reverse happens when the gas is decompressed, so that the decrease in P that results causes the T to fall (this can take place during evaporation).

Applying these laws can be tricky. This balloon behavior model affords a clue: Begin with a gas inside a thin-walled rubber balloon. Let its internal pressure be equal to that of air exterior to it. If the balloon gas pressure increases, to adjust (equalize pressures again) the balloon will expand in volume. If the exterior gas were to experience an increase in temperature, its volume would increase, its pressure decrease, and the balloon once more would expand to restore pressure equality.

A somewhat different slant on gas physics results when the atmosphere is considered as a thermodynamic system. Thermodynamics considers P and T changes from the standpoint of *work done* and *energy exchanged*.

The atmosphere is made up of a number of species of elements and compounds. Each contributes to the total atmospheric pressure (each has a *partial pressure)*. The behavior of gases from the viewpoint of physics involves principles of thermodynamics. Fundamental to this is the first law of thermodynamics, as expressed here:

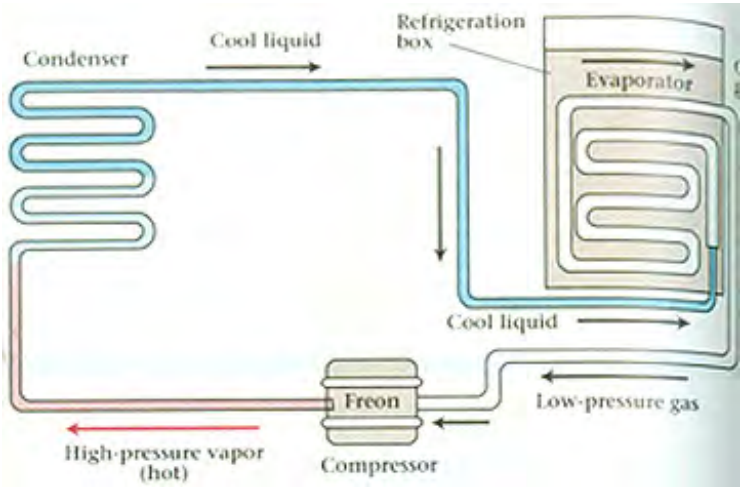


This can be stated in this simple formula: Q = T + W. The internal energy (heat content) Q of an air parcel subject to movement within the atmosphere may be changed, either by a change in temperature or by doing work on its surroundings. When the parcel rises, its density decreases as the parcel expands (does work on its surroundings); conversely, a downward moving parcel is compressed (increases its density) and warmed. (Expansion reduces pressure; contraction increases P.) (One way to look at this: less dense gas has its particles further apart, so the number exerting impacts per unit area decreases [lowering pressure]; contraction pushes particles closer so as to increase impacts (raising pressure].) If the change in density is **adiabatic** (no energy added or removed, i.e., Q = 0; if thereis heat exchange one outcome could be a constant temperature - the isothermal case), then expansion is accompanied by cooling (W is +; T is -) whereas contraction (compression) involves warming (W is -; T is +). This is a fundamental property of the thermodynamics of gases and is a key concept in dealing with air parcels and masses since the general response of rising or falling air is always adiabatic.

From thermodynamics we learn that as an *adiabatic* (no external heat added or removed) gas expands, it cools. This takes place when air moves upwards into the atmosphere and experiences ever lower pressures, thus expanding and cooling. Conversely, compressing (reducing volume through pressure) the gas, which requires an energy input, causes it to heat up. Adiabatic temperature change is a fundamental concept that underlies much of what happens in observing meteorological phenomena in the inner air shell around the Earth. This behavior is consistent with that deduced from the action of the Gas Laws for the case *where P, V, and T all vary*. But, as we shall see, a complicating factor is the role of latent heat released or added during changes of state (condensation and evaporation). We shall return to thermodynamics on the next page.

Further treatment of thermodynamics is beyond the scope of this meteorological mini-Tutorial content. However, some idea of how thermodynamics affects gases can be gleaned from this brief review of how an ordinary house refrigerator works (the principles also apply to air

conditioning and heat pumps). Keep in mind that the system acts as though not subject to the condition in the Boyle and Charles Laws that one variable (H, P, V) be held constant, so that all are free to change. Follow the description of the process, aided by this illustration:

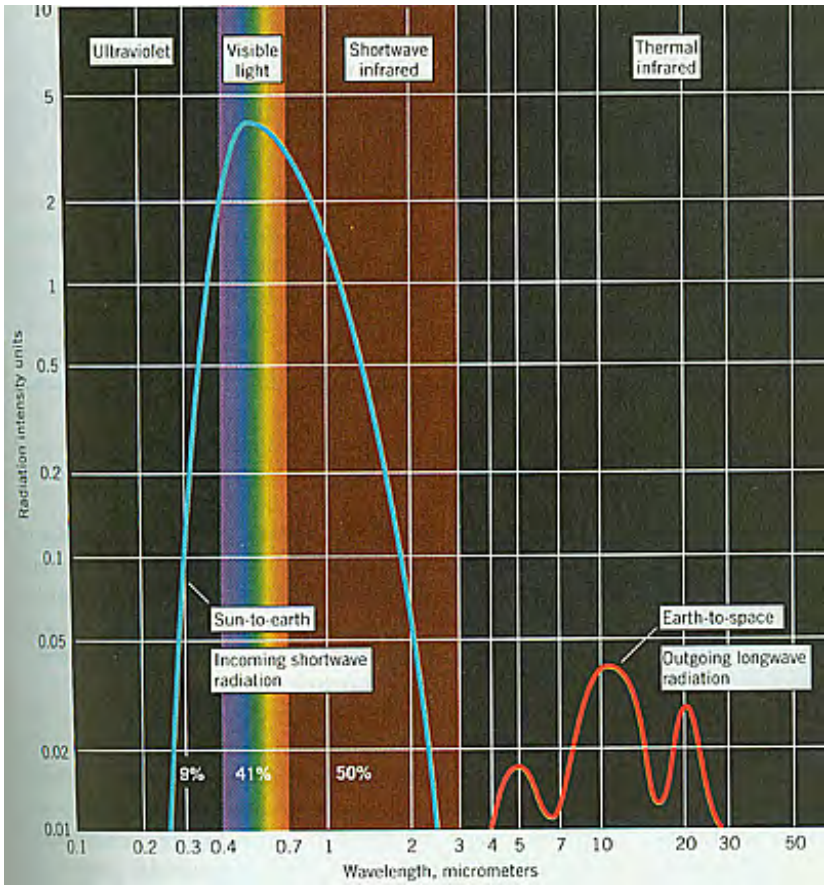


From Contemporary College Physics; Jones & Childer

Start with the refrigerant gas: usually either Freon gas or Ammonia. In the compressor unit, this gas when compressed is heated. It passes then through condenser coils, where the gas is cooled to a liquid (giving off or exchanging heat). This cool liquid is then transferred to a large evaporator coil inside the used part of the refrigerator; the coil assembly is kept at a lower pressure. Evaporation under this reduced pressure requires heat to convert the liquid to a vapor; this heat is supplied by the air (and contents) inside the refrigerator, so that this vaporization process results in heat loss and a corresponding temperature drop. The cooled gas then returns to the compressor to repeat the cycle. So, in this chain of changes of state, the heat inside the refrigerator (that part where one wishes to keep its occupants cool) is removed during evaporation and is transferred to the moving Freon which then releases the heat at the condenser.

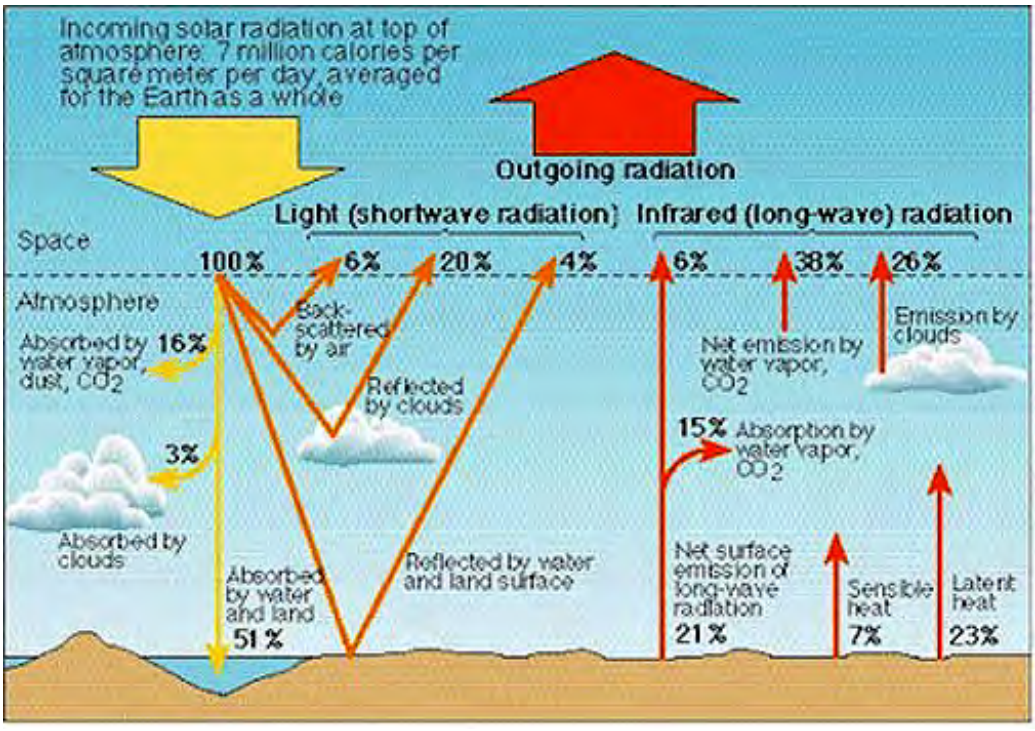
Gas behavior is at the heart of concepts that relate to the movement of atmosphere (air), as will become evident in these three pages. For now, just accept this statement: Temperature changes in the atmosphere bring about both pressure and volume changes; the meteorological system is open (no confinement). The resulting large units of air (parcels or air masses) will undergo movements (winds are a manifestation) brought about by pressure differences (high pressures move air towards low pressure parcels) and also by density differences (decrease in volume produces an increase in density which drives that air towards regions of air having lower density). Gravity, the Earth's rotation, and centipetal forces are also factors.

The atmosphere is kept in motion and develops its weather/climate characteristics primarily through energy input as **heat** (as a term this refers to quantities of thermal energy). Heat actions produce various temperatures (a measure of the degree of atomic/molecular motions through heat inputs and withdrawals). As we saw above, temperature variations give rise to differences in pressure (hence development of pressure gradients that drive gases into motions that include wind) and volumes. The dominant source of heat affecting the atmosphere comes from solar irradiation or insolation (heat added from internal Earth flow adds a small amount to the land/ocean bottom surface). This insolation (= spectral irradiance, given in units of Watts/square meter) is primarily short wave radiation as shown in this diagram (see also page 9-2) which also indicates that some of that radiation is returned to and through the atmosphere as long wave radiation. The atmosphere has windows of transparency in the wavelengths shown.



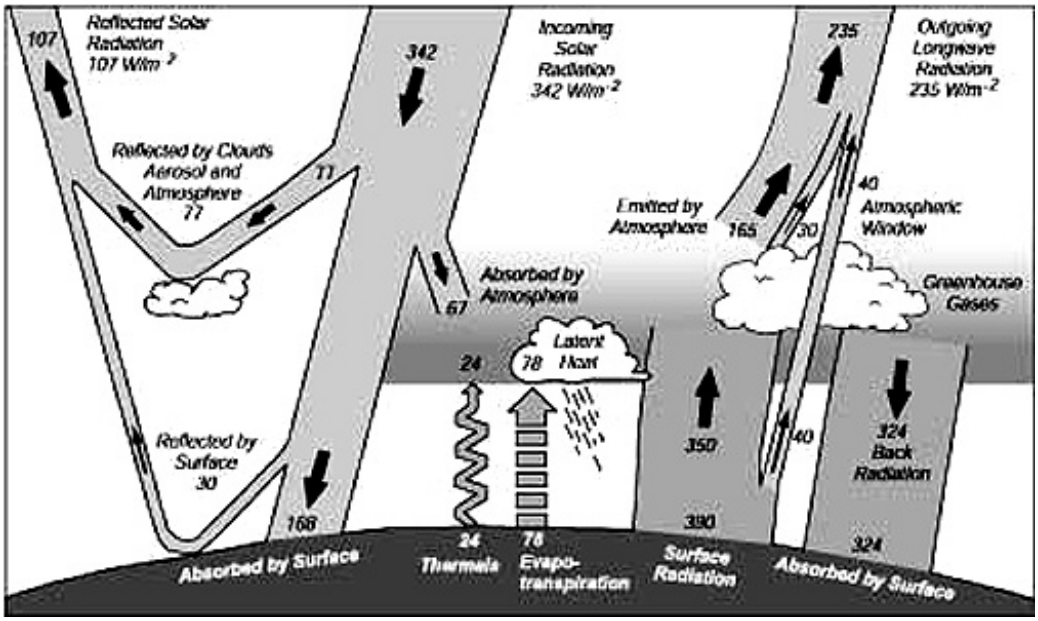
The average amount of solar energy flow as shortwave irradiance reaching the outermost edge of the atmosphere is defined as the *solar constant*. This has a value of 1370 W/m^2 (note: Watts has time in its definition, so it describes a flow rate). The solar constant can also be expressed in terms of heat energy flow and is ~2.0 calories per square centimeters per minute.

The incoming radiation will react with a wide variety of "targets" From this one can calculate a partitioning of heat energy to various recipients such as the ground, water bodies, vegetation, air gases, clouds, etc. This can be shown diagrammatically as an Energy Budget or Heat Budget:



The global energy flow (radiative budget) can also be quantified in terms of irradiance and radiant emittance, in units of Watts/m^2 as shown in this diagram, first published by Kiehl and

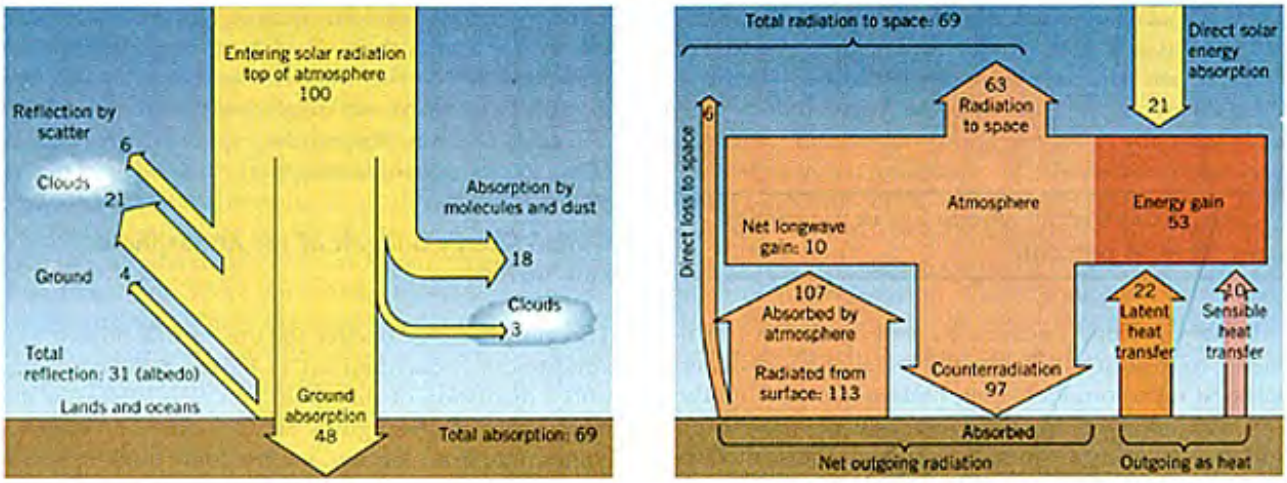
Trenbirth (1997) :



In these two balance charts, which show the same thing but in somewhat different partitioning modes, the heat or energy flows must ultimately balance within the partitioning system in an equilibrium mode. But when the system is treated in terms of net effects leading to accumulation of heat in the atmosphere the outward flow into space will be less than the inward energy received from the Sun, so that warming results (see below).

The system utilizes conservation processes which are designed to attain a balance. The processes consist of inputs/outputs (flows) of *radiant heat*, *sensible heat* (that transferred by convection/conduction), *latent heat* (used or released in change of state processes such as water condensing into rain, freezing to ice or evaporating to cloud condensation), and stored *ground (soil) heat*.

Consider this pair of diagrams which are related to the first balance plot above, but which specifically allot heating modes numerically to show the balance between incoming and outgoing radiation of a thermal nature:



From Strahler and Strahler, Introducing Physical Geography, 1994

Drawing from the Strahler text, and condensing their explanation, let us account for the numbers in the diagram pair. In the left diagram incoming solar radiation is arbitrarily set at 100 energy units (expressed as percent). From the incoming short wave radiation 31% is reflected by the ground, air molecules, clouds, and dust, representing an averaged albedo (reflected radiation). The atmosphere's dust and clouds absorb 21%. The remaining 48% is absorbed by land and ocean surfaces. The right diagram describes the various ways in which outgoing long

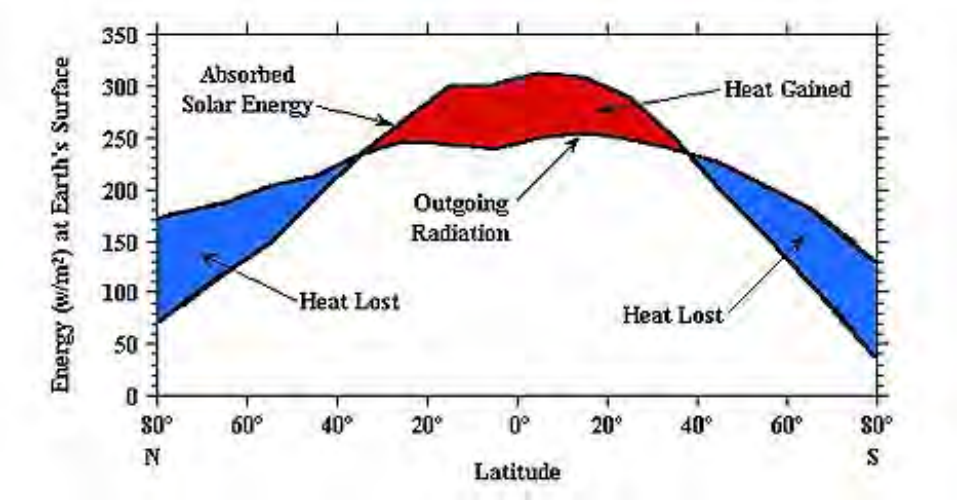
wavelength radiation is ahsorbed by the main components of the system: land, water, air, and outer space. As emitted surface radiation 107 units are absorbed by the atmosphere and 6 units escape into space - this is a total of 113 units.

Where does the extra percentage come from? The atmosphere counterradiates 97 units to the surface which must be added to the 48 units of incoming received and absorbed, giving a total of 145 units introduced to the surface as combined short and long wavelength radiation. On the right side, 10 units of sensible heat and 22 units of latent heat (mostly from ocean evaporation) combine with the 113 units of long wave radiation loss, yielding 145 units that balance the opposing flows. In the atmosphere itself, the partitioning of energy is somewhat different: The 63 units radiated to space are added to the 97 units counterradiated to the surface to give a total of 160 (percentage) units removed from the atmosphere. This is balanced in the long wave ledger by the 107 units released from the surface by radiant emission plus 21 units from outer space, and the 22 and 10 units from latent and sensible heat release respectively. The sum is 160 units as additions to the atmosphere.

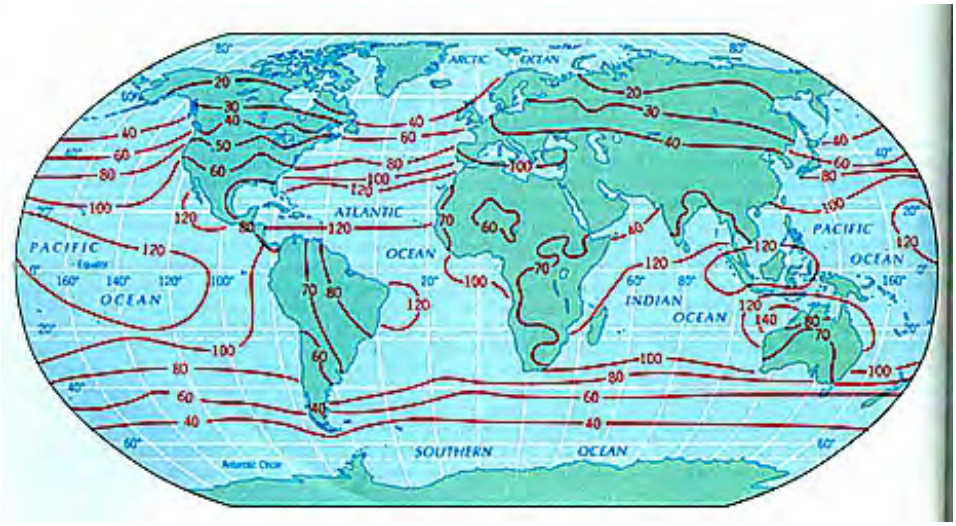
However, atmospheric warming itself has its own numerical input that does not balance: This is based on the 145 units of incoming direct ground radiation and counterradiation (48 + 97 units) but of that number only 113 becomes outgoing radiation. The difference of 32 units accounts for a net gain in heat-producing energy in the atmosphere that yields the temperatures that characterize our planet's temperatures (whose ranges result from interactions as described above). This is analogous to the "Greenhouse Effect" described on the third page of this subsection.

In fact, one must also consider a somewhat different parameter - net radiation balance (not just heat) - in which the outgoing radiation (397, in units of calories per square meter) is less than the incoming radiation (469 cal/m^2) on an averaged global annual basis. This takes into account differences between heat response in the oceans and the land and between latitude locations.

Of great importance is the geographic distribution of net radiation. The excess of radiation at low latitudes results from radiation coming in faster than it goes out; the converse (slower in, faster out) takes over at the higher latitudes. This build-up of heat around the equator and depletion in polar regions is responsible for the poleward flow (transport) of heat energy to equalize the total energy distribution. That is one of the driving forces in the circulation patterns that mark Earth's atmosphere. This diagram shows that the low latitudes in the course of an annual seasonal cycle gain more heat whereas the higher latitudes loose heat.



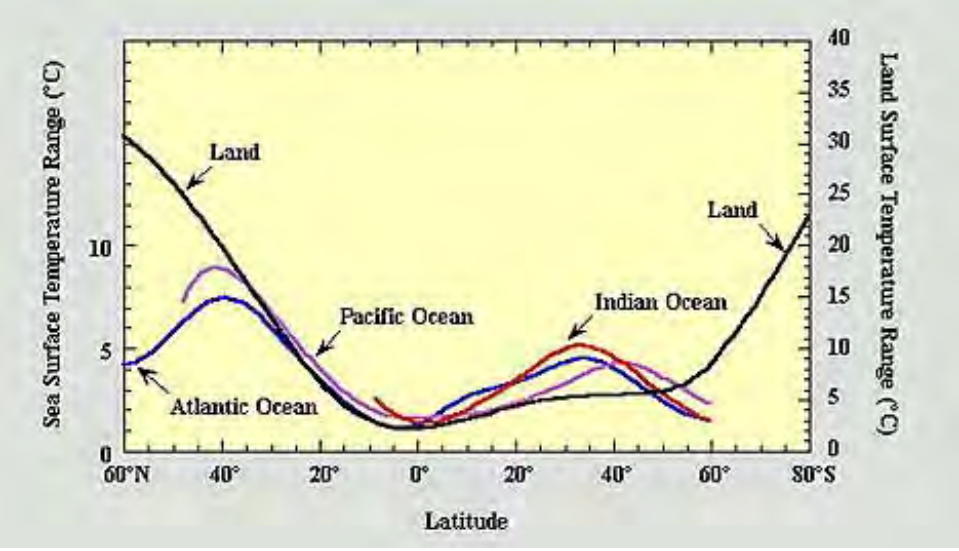
This figure summarizes the global net radiation distribution at the Earth's surface (in units of 1000 calories per square centimeter).



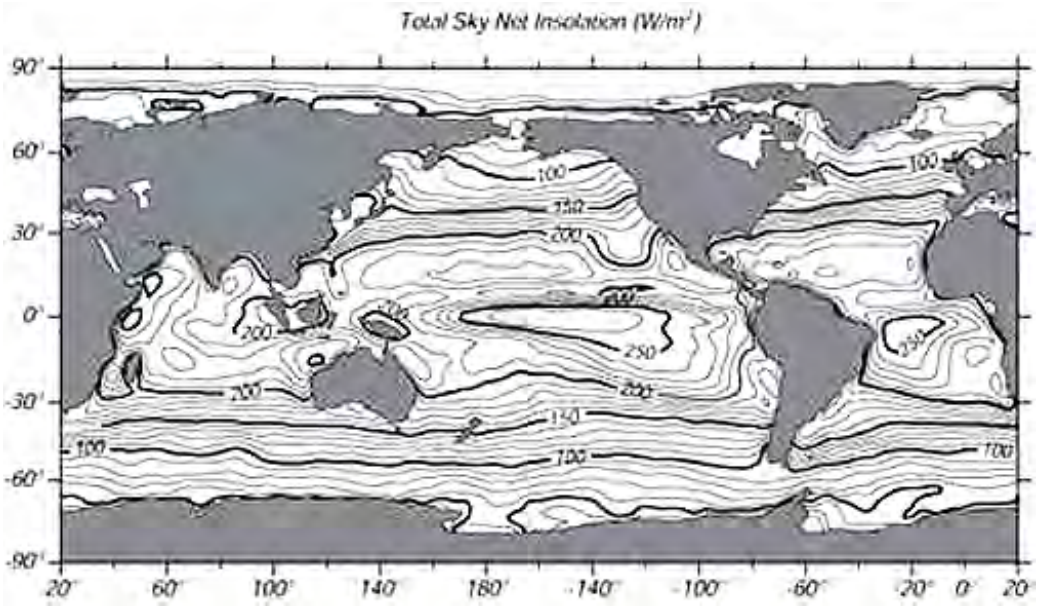
From de Blij & Muller: Physical Geography of the Earth's Environment

The variations in net radiation leading to gains and losses should not be surprising. The above is a somewhat idealized picture that treats the atmosphere as a uniform entity rather than as a system in which inputs and outputs vary over the globe resulting in different conditions from place to place (such as albedo, cloud cover, etc.) that change with time. The same is true for the above heat budgets in the sense that the percentages vary with affecting conditions but should average out over time. It is the differences on a geographic and temporal basis that gives rise to the vicissitudes of weather and climate we observe on our planet.

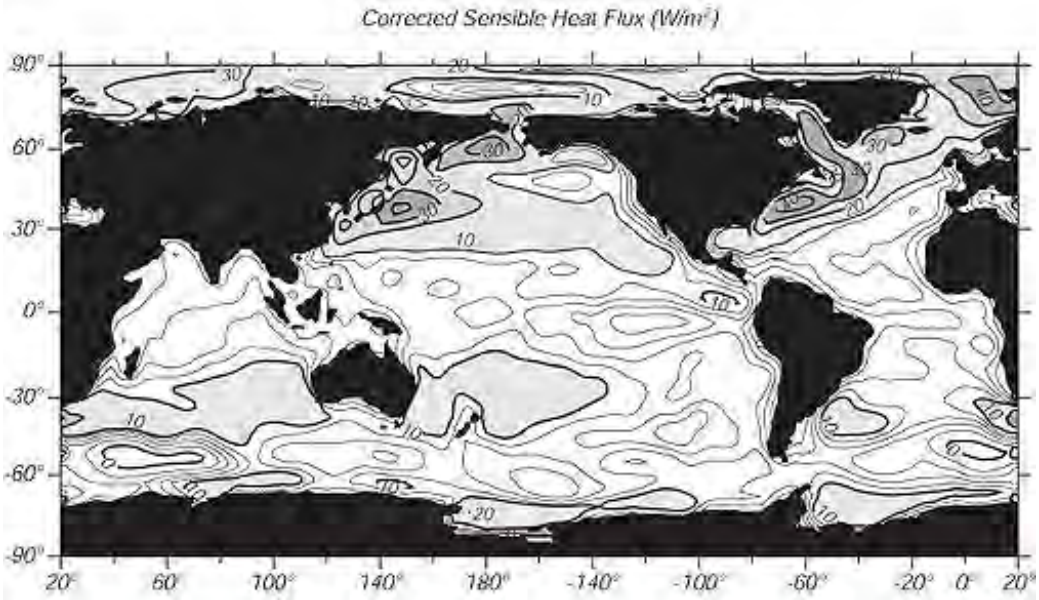
This variation in heat content over the seasons affects the range of temperatures in marine waters, as shown below for several different oceans. In general, surface sea water experiences a minimum of temperature variations in low latitudes and a somewhat higher range of changes as the mid-latitude regions are approached. Note that land and sea temperatures are nearly equal in the tropics but in the polar regions the land is much warmer than seawater.



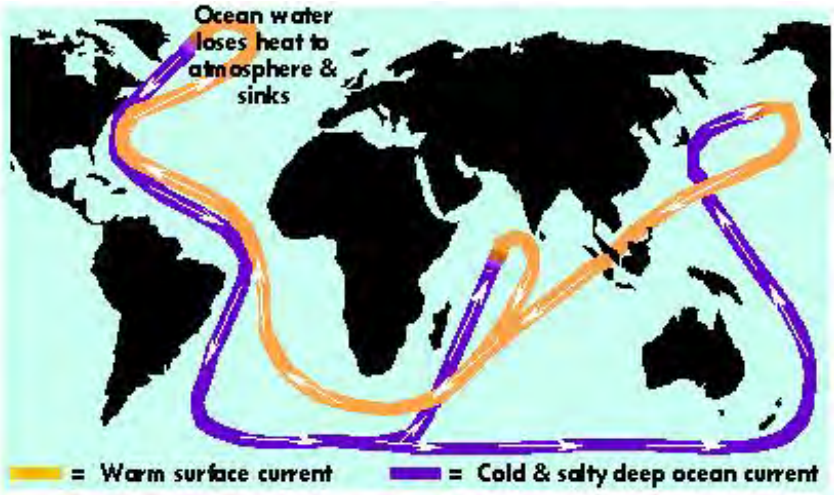
The oceans act as major heat reservoirs that have significant effects on the roles that heat and radiation play in generating meteorological patterns. This generalized diagram shows the global variations in incoming solar radiation (irradiance) over the oceans.



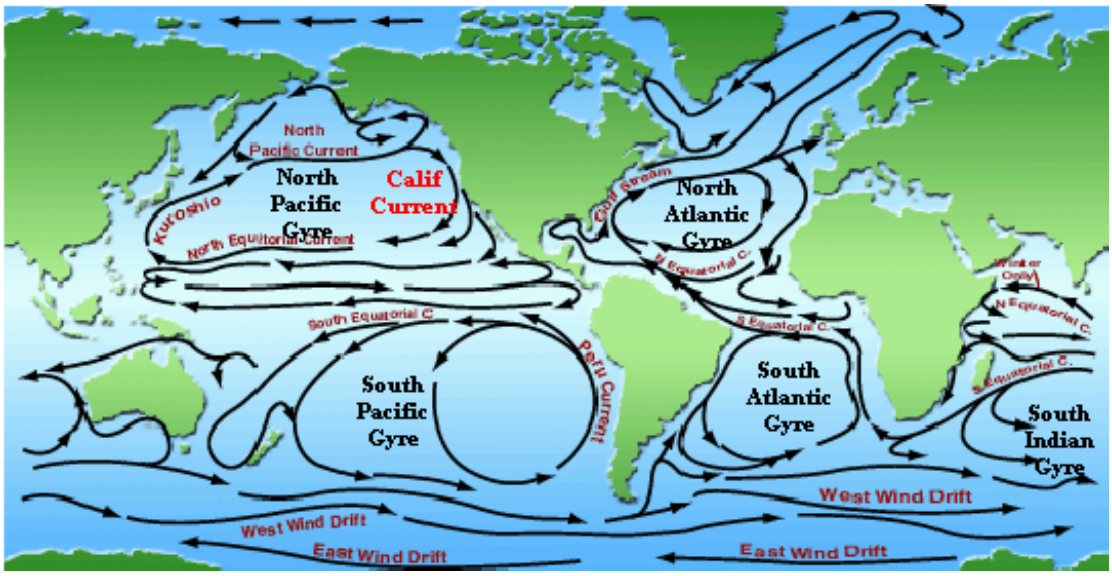
A fraction of that will escape the ocean water reservoir as sensible heat returned to the atmosphere whereas another, larger fraction enters the ocean as stored heat.



The oceans experience differential heating, generally being warmer at the surface. Cooler water, being more dense, will sink. Since there are temperature differences within the oceans at various depths, these give rise to current movements. The general pattern of water transfer is shown here:

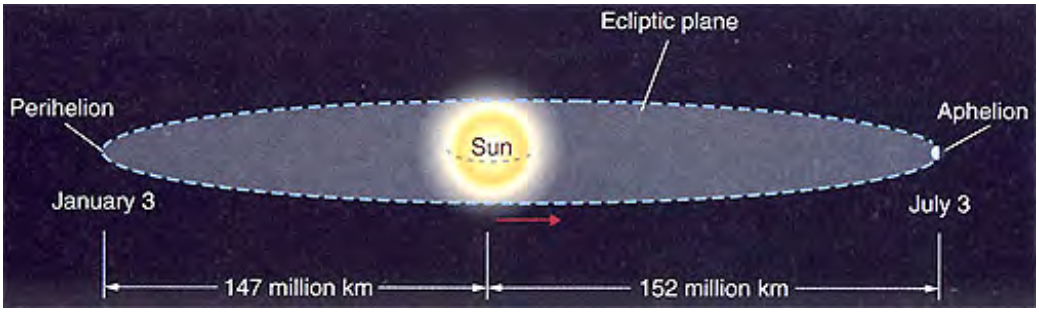


In the above illustration and the one below, the patterns (gyres) are modified by the presence of land masses in "irregular" dispositions (locations) around the globe - the present configuration of continents results from the stage of plate tectonic movements now acting on the earth continental and oceanic lithospheres (crust and the uppermost mantle). Another factor: there is much more land in the northern hemisphere, so that 81% of the southern hemisphere is ocean water compared with 61% in the northern hemisphere; this suggests that marine thermal processes play a larger role south of the equator.



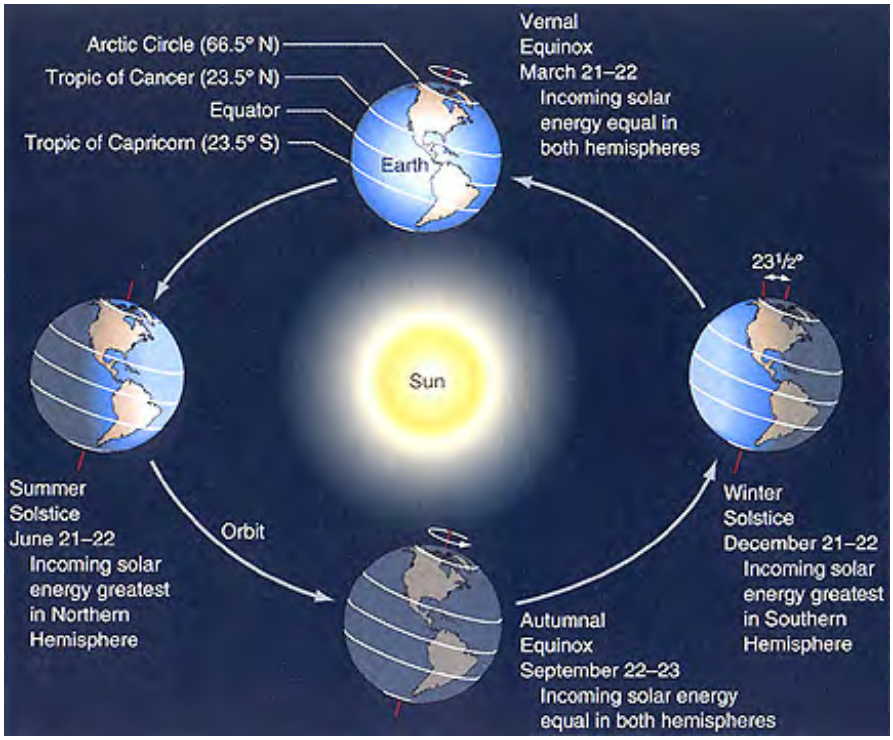
The surface currents result from cold water in higher latitudes moving equatorward, displacing the more heated (warm) waters that move poleward as a thermally driven circulation ensues. Movements are also affected by wind forces. A warm oceanic current transfers some of its heat to adjacent lands whereas the cooler current modifies land temperatures around coastal regions (best example, the West Coast of the U.S.). Best known of these surface currents to U.S. easterners is the Gulf Stream (a warming current). Benjamin Franklin was among the first to recognize its significance in weather control and pointed out that ships sailing from Europe westward took longer to reach North America since they often encountered the eastward flow of the Gulf Stream.

Earth is characterized by its **seasons**, which are extended periods in which weather and climate have distinctive average temperatures, rainfall, and storm frequency and type conditions. In much of the Earth beyond the tropical regions around the equator, the seasons also are marked by extended periods of active vegetation growth alternating with vegetation dormancy. The seasons more than any other factor influence human, animal, and vegetation activities over the course of a year. The reasons for seasons are mainly twofold: 1) the position of the Earth relative to the Sun during its annual 365.4 day revolution around the Sun; 2) the tilt of the Earth's axis of rotation. The first factor, shown here, has a much smaller effect than the second; distances from the Sun vary by less than 10% so that insolation intensity is not great but has some influence.



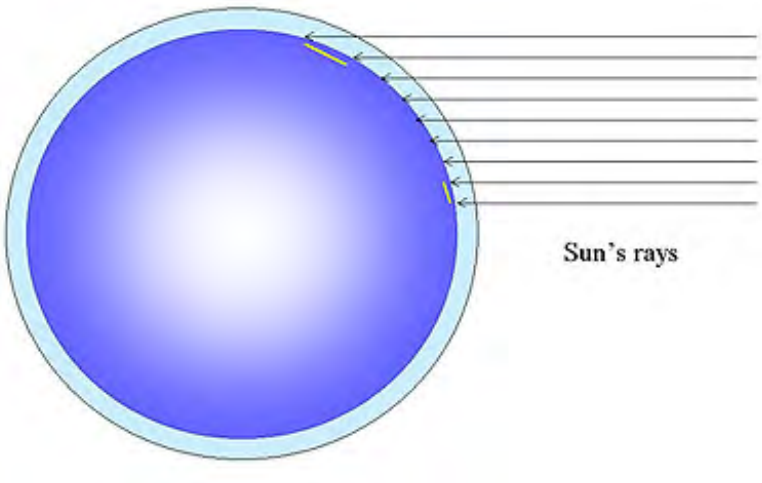
The axis of Earth is tilted 23.5° from the plane of the ecliptic (the plane traced by the Earth's path around the Sun; the other planets also move in paths close to this plane). This tilting as we

shall see shortly is the key to the changing seasons. Lets look first at the total path during that year, noting the direction of tilt.



At the Summer Solstice, the 23.5° tilt is leaning towards the Sun in the northern hemisphere; at the Winter Solstice the southern hemisphere is leaning 23.5° towards the Sun (it is winter in the northern hemisphere but summer in the southern hemisphere). Twice a year at the Vernal (norther Spring, southern Fall) Equinox and the Autumnal Equinox (northern Fall; southern Spring), the position of the Earth relative to the incoming Sun's rays is such that the length of the day is equal in both hemispheres.

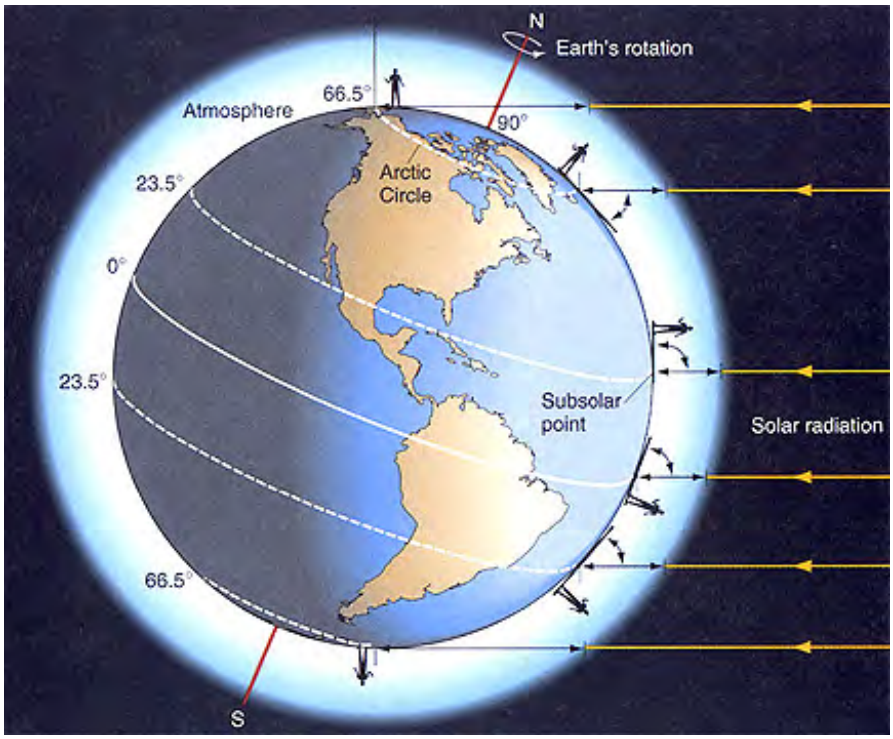
To establish why the Earth's axis tilt is the prime cause of seasonal variation, start first with this diagram:



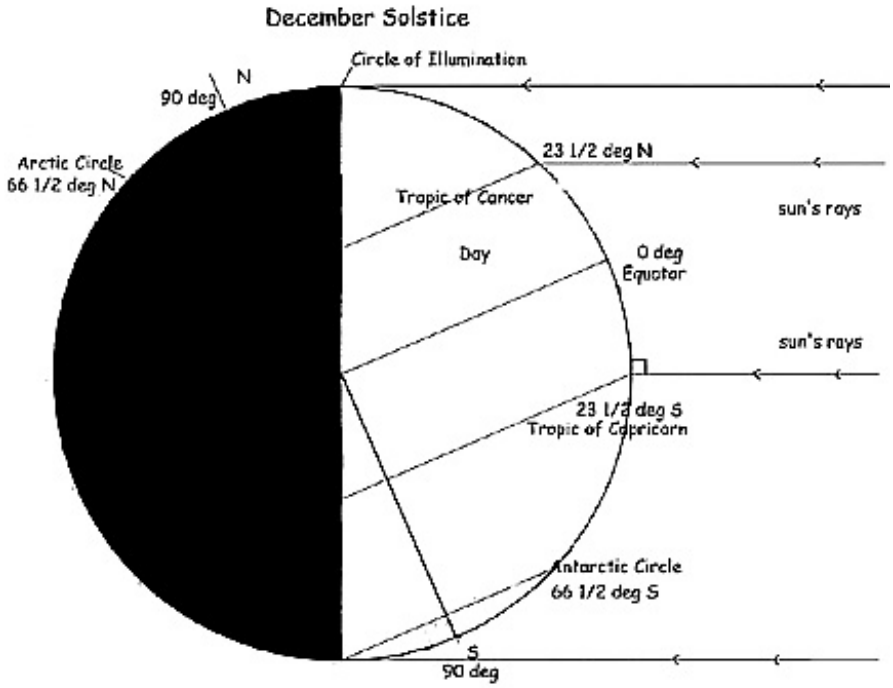
In this simplified diagram, the sphere has a vertical rather than tilted axis (top to bottom). Note that at the equator, the length of the path of sunlight through an atmosphere of uniform thickness and identical density distribution is the shortest of all rays. That path length increases systematically moving to either pole. The shorter the sun ray path, the less is the interaction between the insolation and anything in the atmosphere that deflects, absorbs, or re-radiates the irradiance. Thus a greater proportion of radiation hits the surface or reacts in the troposphere. The result is increased warming (maximum at the equator). As the path lengthens poleward, the amount of radiation reaching the surface and lower atmosphere decreases, thus producing

progressively less warming poleward. The mean (and maximum) temperatures will therefore be lessened going from low to high latitudes.

Now, examine the case where the axis is tilted 23.5° towards the Sun. The effect is to move the North American continent (or Eurasia) clockwise downward towards the ecliptic and hence into regions of shorter pathlengths. Mid-latitudes and even polar regions receive more surficial radiation and thus become warmer. This warming during that position of the pole during times of year close to this orientation leads to summer conditions. Vegetation commences renewed annual growth as the conditions approach as spring progresses; subpolar and polar ice undergoes some degree of melting.



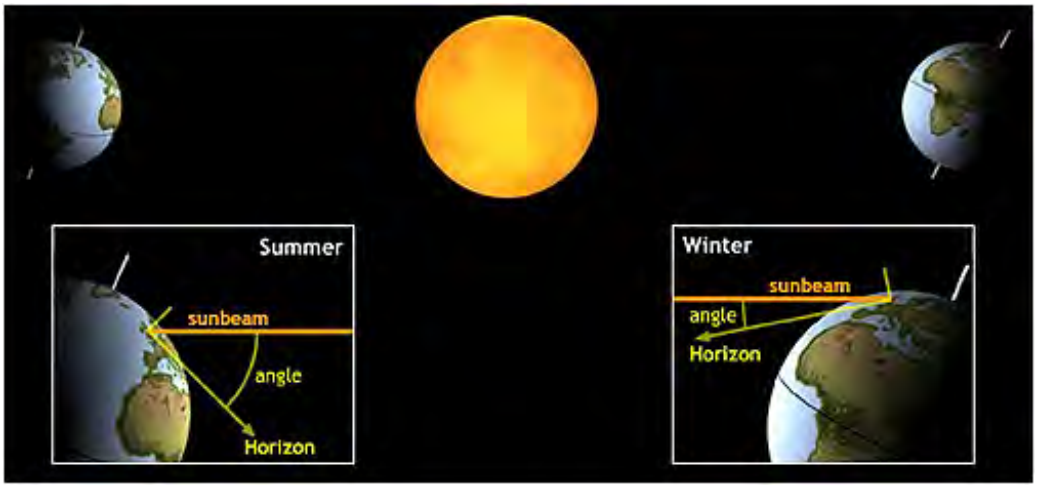
Then consider the reverse case, a 23.5° tilt away from the Sun at the time of winter solstice:



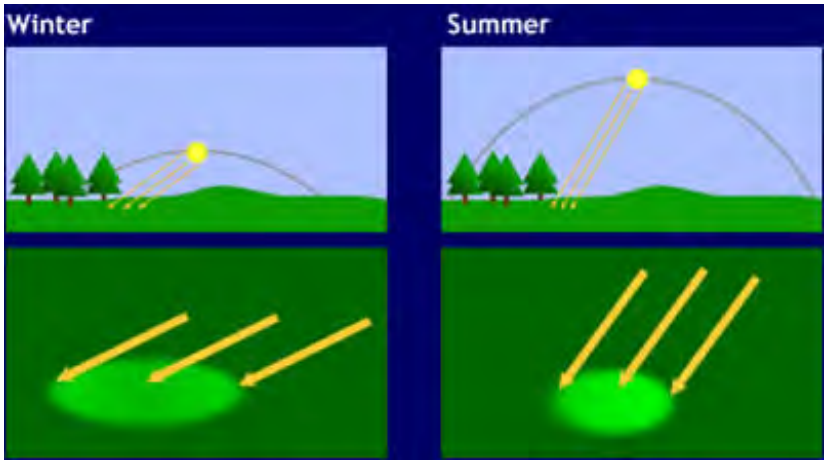
In this situation lands and oceans in the southern hemisphere move upward into the path patterns where the lengths through the atmosphere decrease and more warming radiation reaches these areas. This is the condition for southern hemisphere summer. Northern lands and

seas have also moved upwards into the zones of greater path lengths and hence surface-troposphere temperature decreases. This is northern winter.

At the Equinoxes, the Earth's equator lies directly beneath the *subsolar point*. Between Summer and Winter solstices the equator is shifted a total of 2 times 23.5 or 47° over this six month period (the same occurs in the opposite direction between Winter and Summer. This oscillation accounts for the seasons. At intermediate times the heating patterns relative to the Earth's geography tend, on average, to fall between the extremes. In either hemisphere, this next diagram applies.



In the summer case, the angle between a direct line to the Sun and the nearby horizon (where Earth's surface meets the atmopheric base in the line of sight) is higher than in the winter case. This means that the Sun appears higher in the sky at Noon in Summer than in Winter. This is depicted thusly:

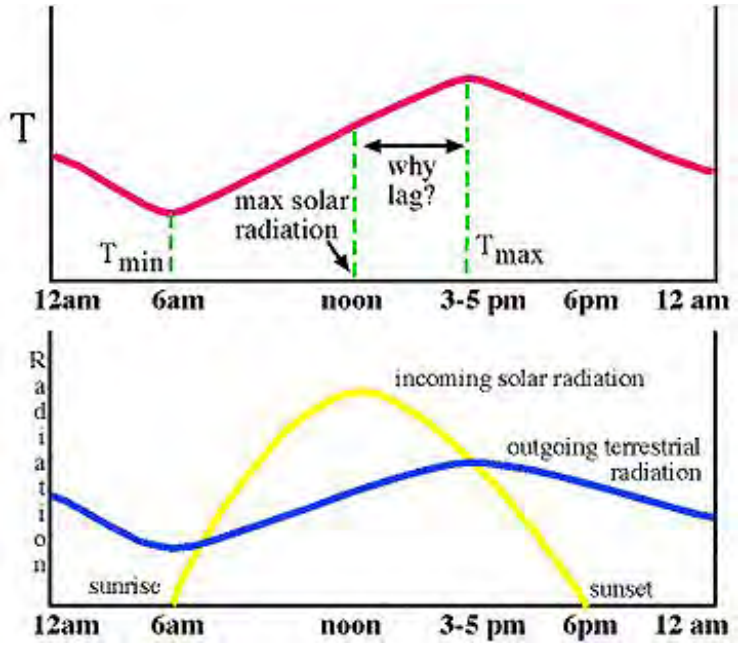


The last diagram indicates two things: 1) a higher sun angle will also lengthen the time in 24 hours in which the Sun shines at a given place; and 2) the more concentrated solar beam leads to greater warming of an area. At the two Equinoxes, the length of day at the Equator is equal to the length of night (neglecting dawn and dusk effects due to light scattering). In the northern hemisphere's progression from Winter to Summer the duration of daylight goes from a minimum to a maximum (as it does for the southern hemisphere cycle 6 months apart). The further north one goes in, say, North America during any given day, the length of daylight (summer) or darkness (winter) increases as its season progresses. In Arctic regions, in mid-summer the Sun is always above the horizon ("perpetual" daylight) but in the winter months, polar regions may experience total diurnal darkness. (The same is true for winter in the southern hemisphere; those doing studies in the Antarctic may leave the continent at such times, especially since then the weather is also at its coldest.) In a hemisphere, the transition between Winter and Summer leads for a given region (at some latitude) to a gradually progressive warming; conversely, the trend from Summer to Winter is an overall cooling. Low latitude regions, such as equatorial

tropics, experience some changes in daily temperature variations (that will occur simply because of non-heating during the night hours), but with usually a minimal range and a higher average temperature spread over the full year.

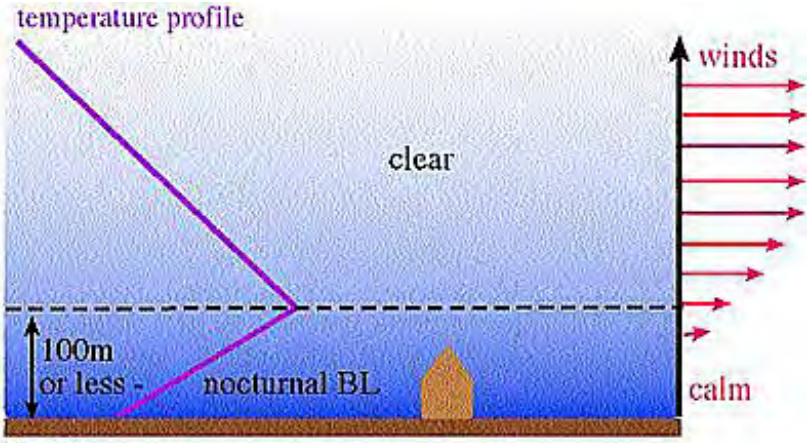
Note too that the maximum and minimum temperatures do not necessarily coincide with the Summer and Winter Solstices *per se*. There is a lag effect that tends to delay the hottest days to the July months and the coolest days to the January months in the northern hemisphere.

As a transition to the second page on regional weather systems, which have both a seasonal and a daily character at individual locations, let's take a quick look at the temperature and pressure changes that typically mark a single day's activities. First, how does temperature vary at some given spot, say, in the eastern United States, during a 24-hr day:

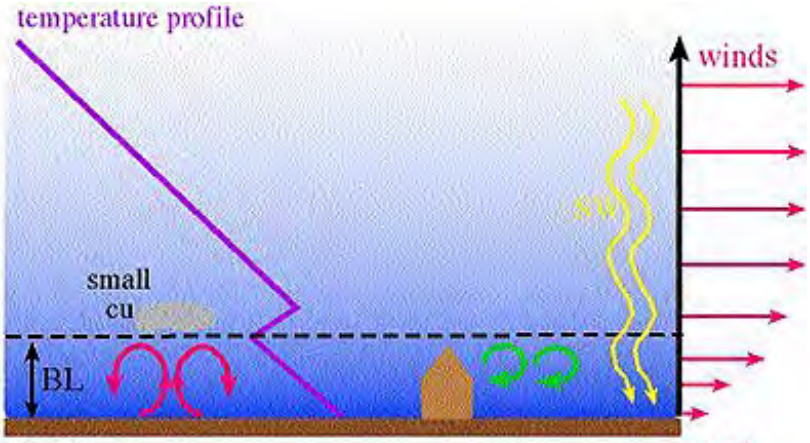


From the lower chart, incoming solar radiation peaks around high noon. Outgoing radiation reaches its minimum around dawn. But, it also attains its maximum about 3 to 4 hours after noon, thus, not coincident with the radiation peak. This lag is the result of several factors, chief of which is the effects of thermal uplifting and winds that carry heat upwards and slow down the surface temperature rise until mixing in mid-afternoon produces the hottest time of the day.

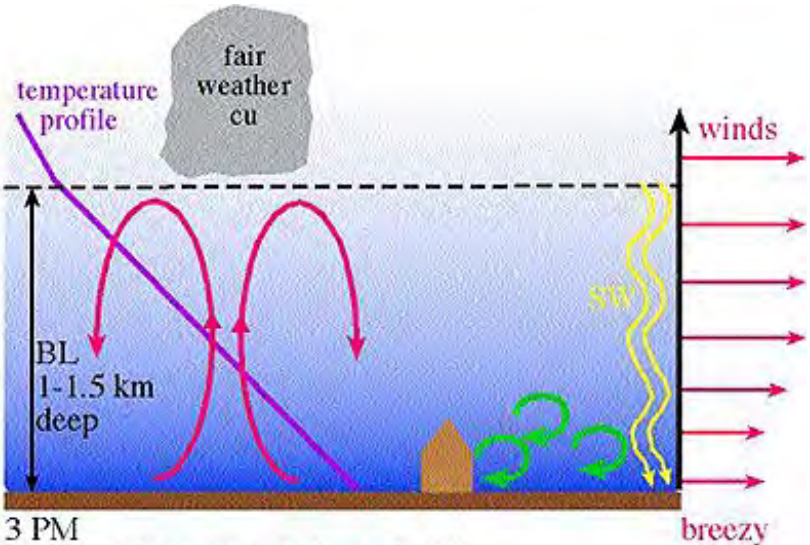
Earthlings experience their weather conditions on the ground within a general layer of air called the boundary layer. This is the bottommost layer that has a thickness of 1 to 1.5 km (up to a mile) in the afternoon but which shrinks to about 100 meters (330 ft) deep in the night. The next, largely self-explanatory, four illustrations show temperature conditions, winds, and air movements at 4 times during daylight hours in a 24-hour thermal cycle. Note that maximum activity in the boundary layer (BL) involves upward flow of air heated at the surface in the afternoon. On a summer day, with local high pressure air and blue skies, clouds of the cumulus type (see next page) are the result.



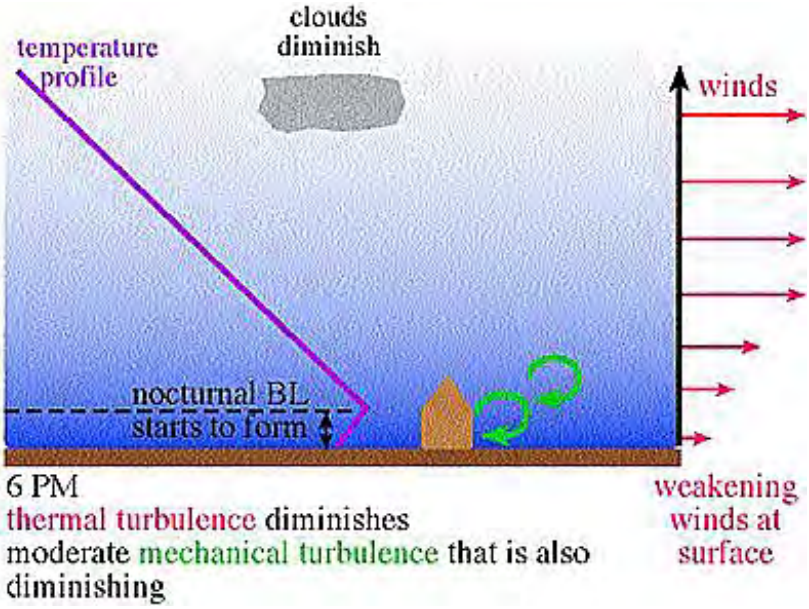
6 AM
No thermal turbulence
Little/no mechanical turbulence



11 AM
weak-moderate thermal turbulence near surface
weak-moderate mechanical turbulence near surface



3 PM
strongest thermal turbulence in BL
strongest mechanical turbulence in BL



We have completed this survey of the general characteristics of the atmosphere. Next up is consideration of the behavior of individual air masses.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The first U.S. meteorological satellite was TIROS-1 launched in April of 1960. Since then NASA, NOAA and other agencies (e.g., DOD) have sent scores of satellites in Earth orbit or into geostationary positions, using increasingly sophisticated sensors to observe various properties of the terrestrial atmosphere, and its climate changes. Other nations have followed suit, so that today TV Weather reports always contain supporting imagery showing regional meteorology ranging from local areas to entire continents. Meteorology itself is a major component of the Hydrologic Cycle which also embraces Oceanography and land Hydrology. This first page considers the general background to "Metsat" observations and cites references on the Net and in books that provide the interested user of this Tutorial with sources that expand knowledge of the principles of Meteorology/Oceanography and other aspects of the "water world". Some quality links to other weather-oriented sites are given.

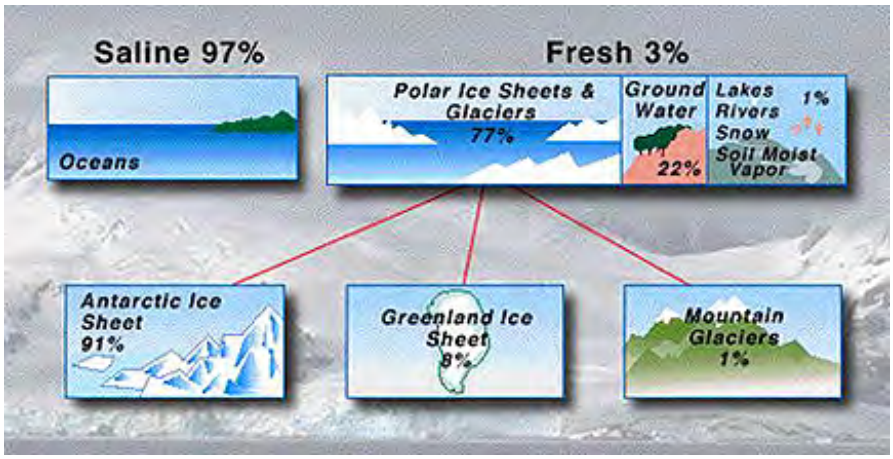
THE WATER PLANET: METEOROLOGY, OCEANOGRAPHY AND HYDROLOGY FROM SPACE

The Hydrologic Cycle; Meteorological Satellites (General)

In the early days of the space program, particle and field physics, communications, and weather satellites were the principal applications of the new technology that grew out of Sputnik. Exploration of the planets, using cameras on space probes, soon followed. Most agree that the highpoint of the first 25 years was the landings on the Moon. Earth-oriented monitoring of the environment and human activities on the surface began in earnest during the 1970s. Astronomical observatories were placed in orbit in the 1990s as launch vehicles became more powerful.

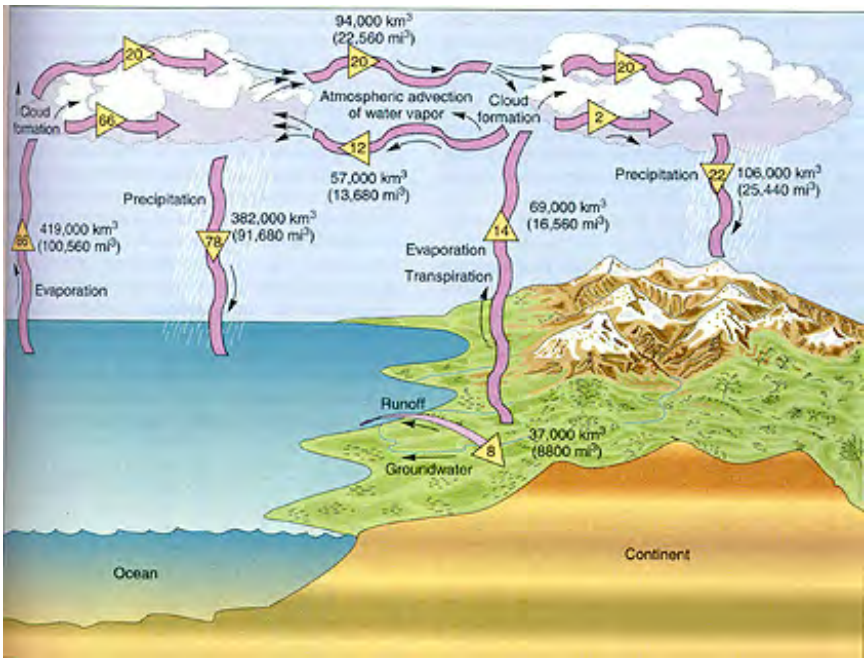
These many uses define the values of space exploration. As far as the U.S. public (the taxpayer) is concerned, however, the first visible payoff from the space program was the images of clouds and weather systems that began to appear on television news in the late '60s. Thus, aspects of the science of meteorology, which experienced a quantum leap of capabilities, were brought home to the proverbial "average man on the streat". Now, with the value of hindsight, we can make this statement: **One of the biggest benefits from the space program has been the observations over the years from meteorological satellites (metsats) of local, regional, continental, and even worldwide patterns of weather and climate behavior that have led to a profound understanding of how atmospheric systems work such that it is now possible to make accurate and timely forecasts for periods of 7 to 10 days ahead so that planners and the public can now depend on reliable predictions in their daily lives.**

We start this Section that deals with meteorological applications by summarizing the distribution of water, both saline and fresh, above, on, and in the outer Earth:



What is particularly striking about this chart's content is the large amount of fresh water locked up in snow and ice, the size of the groundwater reserve, and the fact that rivers and lakes, which are obvious bodies of water familiar to us, actually contain only a tiny fraction of the Earth's supply of fresh water.

Oceans and large freshwater bodies cover more than 70% of the Earth's surface. At any moment, around 50% of that surface, land and sea, is hidden from satellite view by clouds. Over smaller areas, but still significant, rain, descending from these clouds, impacts on the surface to run off and then coalesce into streams and rivers. This great system of interconnected water circulation comprises the hydrologic cycle, as summarized in this diagram:

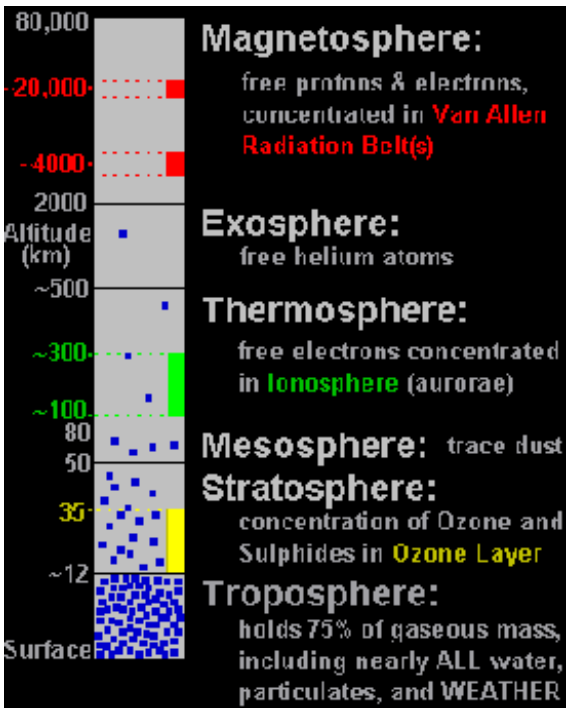


(Christopherson, R.W., GEOSYSTEMS: An Introduction to Physical Geography, 2nd Ed. © 1994. Reproduced by permission of Prentice Hall, Upper Saddle River, New Jersey)

14-1: **What in the diagram accounts for the least amount of water relocation? What category of temporary water storage is not cited in the diagram? ANSWER**

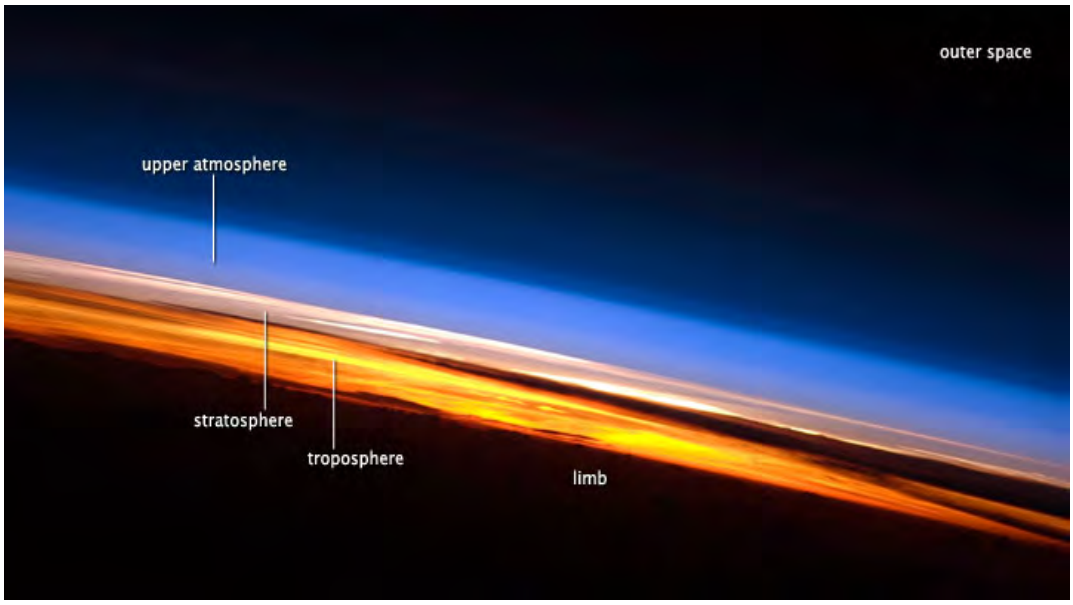
The numbers associated with this chart clearly demonstrate that the oceans not only hold the bulk of the planet's water but are the source of most of the precipitation that constantly recycles water. Water in transit appears as circulating (wind-driven) visible clouds and invisible water vapor or as water mobilized in fluvial systems. In addition to the ocean bulk, which accounts for nearly 98% of the volume of water at or near the Earth's surface, most of the remaining fraction is ice, mostly in the Antarctic and Greenland, snow (much being seasonally ephemeral), and freshwater lakes.

Most remote sensing observations tend to be two-dimensional, that is, the sensors look at the surface or very near surface. Geophysical remote sensing is, in part, 3-dimensional as the instruments used can provide information about the supracrustal rocks (e.g., sedimentary sections) and the crystalline basement crust and some techniques provide information about the mantle and even the Earth's core. Meteorological remote sensing is primarily directed towards making observations of the atmospheric profile - the column of air above the ground which varies in temperature, pressure, and composition. The atmosphere is layered or zoned, with general subdivisions (based on altitude and on composition and physical properties) named as follows (a more detailed version is on page 14-1a):



An astronaut photo shows most of this thick atmosphere, appearing from outer space as a blue band which diminishes in color outward:

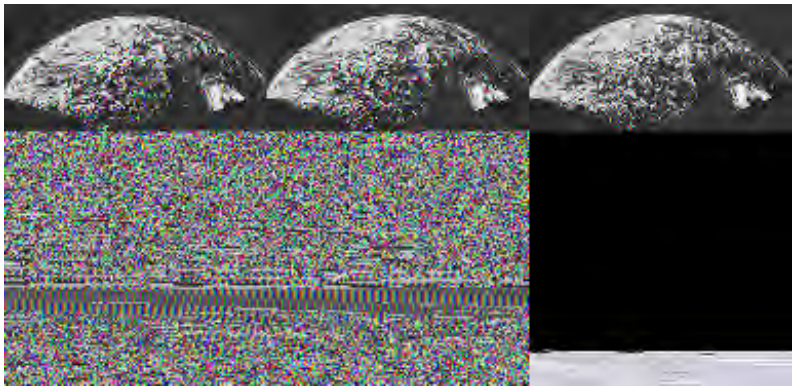
This more recent photo taken near sunset shows the lower layers to better advantage:



The advent of satellites after Sputnik (in 1957) opened large regions in sweeping vistas for directly observing atmospheric properties, weather systems, oceanographic conditions, and water runoff on continents and islands. We could easily combine a series of adjacent scenes, acquired during short time periods, in mosaics to give global coverage on a daily basis. In time, satellites placed in geosynchronous orbit afforded nearly instantaneous coverage of near-hemispheres of the Earth that could rapidly update views of cloud decks and circulation patterns over almost any part of the world. Ironically, the very thing that compromises observations of the land and open ocean, namely clouds, is the prime target of meteorological satellites (Metsats). As more versatile sensors evolved, they quantitatively monitored various other atmospheric or oceanographic properties, such as the stratosphere, tropospheric temperatures, Earth radiation budget, air chemistry (e.g., ozone, CO_2, sulphur compounds, and aerosols), wind and sea current movements, sea-ice, and marine biotic nutrients.

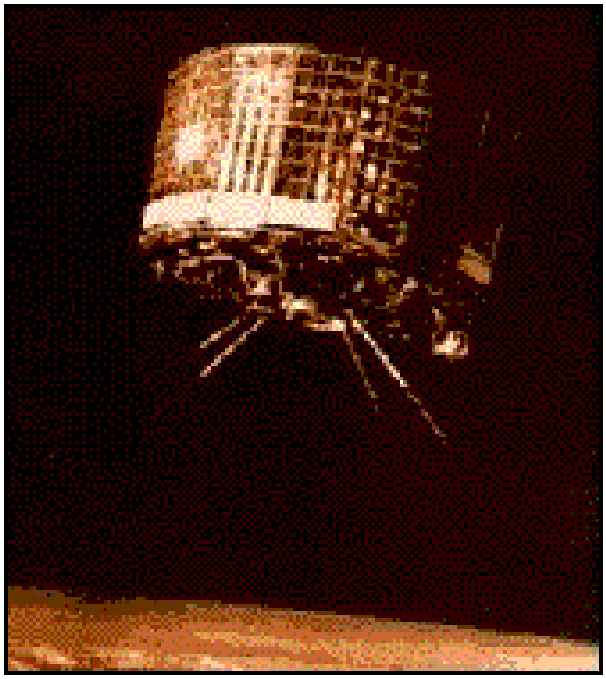
14-2: **What spatial attribute of a sensor system is highly desirable in a meteorological satellite?** **ANSWER**

Common sense tells one that satellites or spacecraft hundreds to thousands of kilometers above the Earth looking down towards its surface will see large areas at a time. Clouds and other weather phenomena scanned over wide vistas would give meteorologists a much better handle on moving weather systems. This became possible after Sputnik; some of the early efforts to utilize space for meteorological studies are briefly described on page 14-4. We show an image that surely kindled excitement among the weather people - indicating the value of the panorama approach - even before orbiting satellites went on line. It is a mosaic of photos taken from an Aerobee rocket that reached great heights during a 1954 mission from its launch site at White Sands Proving Grounds, New Mexico. A significant part of the northern hemisphere, with its various cloud systems, was viewed in this composite.



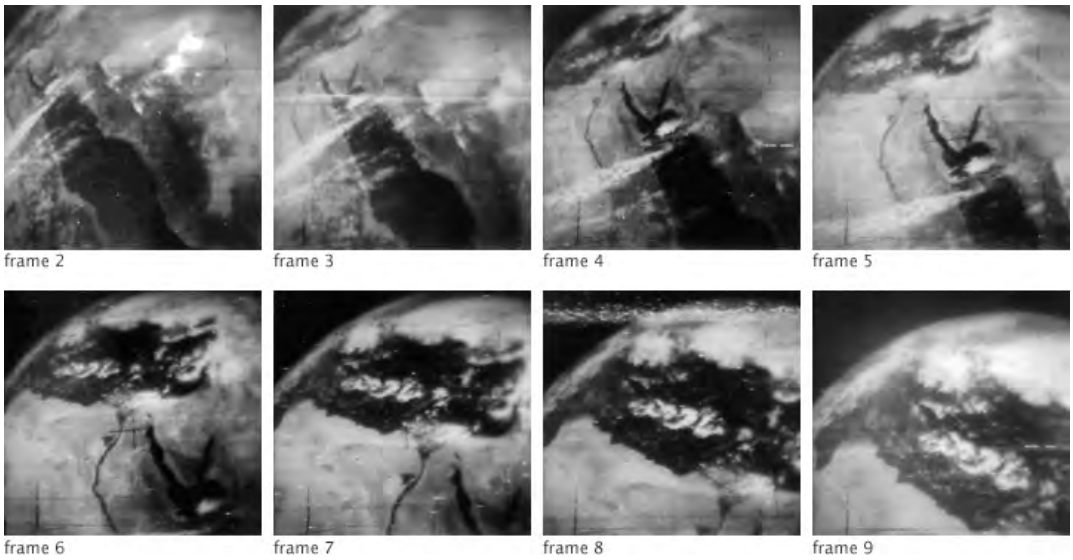
The first ever United States satellite in Earth orbit designed specifically to image and monitor

conditions on and above the surface was the meteorological satellite called TIROS-1, launched on April 1, 1960 soon after NASA came into existence. It looked like this.



It was powered by solar energy collected from the 9260 solar cell plates on its exterior. Being small (42 inches or 106 cm) in diameter, it carried two TV cameras, one with low the other with high resolution. Here is the first image taken by the "Adam" of meteorological satellites.

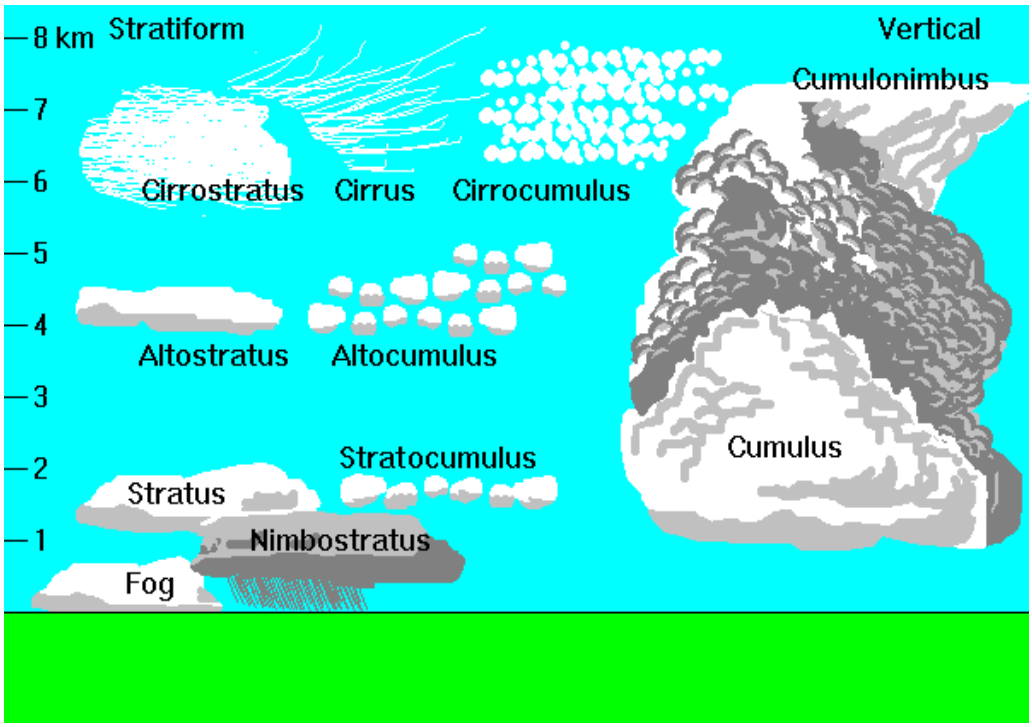
Below is a series of image frames centered on the Mediterranean Sea that constitute the first set of meteorological imagery taken by Tiros, the first metsat:



During its short three-month lifetime, Tiros-1 took 23,000 images of Earth. The TIROS program provided the first accurate weather forecasts based on data from space, demonstrating that it was possible to use satellites to observe weather. The satellite has a long legacy. TIROS-1 led to nine more TIROS satellites, seven Nimbus-series meteorological research satellites, 14 Geostationary Operational Environmental Satellites, 19 NOAA Polar Orbiting Satellites, and many more meteorological satellites maintained by the Department of Defense and other nations.

From 1959 through 2009, many countries launched satellites (approximately 330 - but some failed), such as the United States, the former Soviet Union/Russia, Japan, China, India, Italy, France, and the European Space Agency, primarily to provide current timely data for weather system monitoring and forecasting but also to conduct scientific studies to better understand the atmosphere, the oceans, the Earth's force fields (ionosphere and magnetosphere), solar radiation, and related aspects of the environment. In contrast, so far, they dedicated fewer satellites to land observations. Astronauts conducted meteorological experiments during Mercury, Gemini, Apollo, Shuttle, and Mir flights. The International Space Station is also a good observation platform.

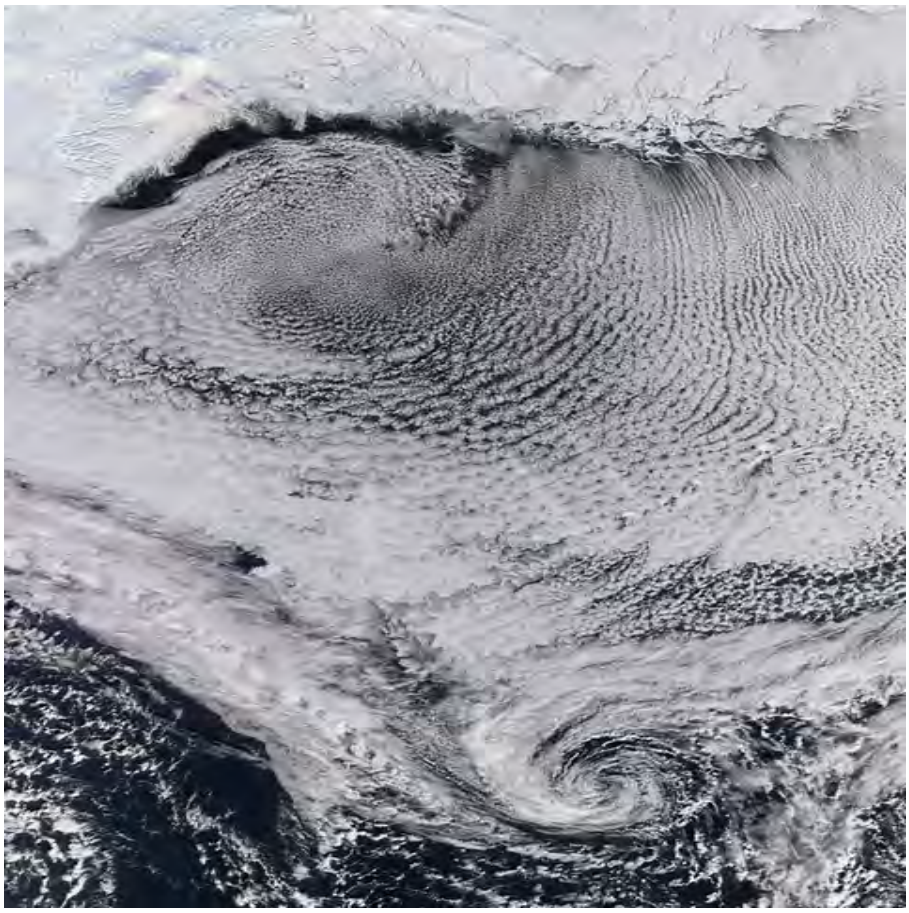
Clearly, the most widespread applications to date of remote sensors operating from space platforms have been to image water - either in oceans, lakes, and streams or in the air as water vapor - in its main functions in the Earth System. But, as the sensors improved, the ability to measure temperatures, and indirectly pressures (highs and lows), in the atmosphere as well as wind speeds and the rates of movements of air masses became possible in a quantitative way. But, the single most obvious thing that meteorological satellites could observe from the beginning are clouds. Most cloud formations (cumulus, stratus, nimbus, and combinations thereof) are straightforward and obvious as to types, but some unusual atmospheric effects expressed by clouds warrant special attention. Two good websites that treat the general characteristics of clouds are at Wikipedia and the University of Illinois. For reference to the various cloud types discussed in the next paragraphs, consider this diagram:



Let us now display over much of the rest of this page images taken by satellites of cloud systems and types. Often entire regional systems can be imaged in a single broad view satellite view. We will begin this coverage of clouds by considering moving air systems at regional scales.

The atmosphere on Earth is in constant motion, being driven by differences in pressure (pressure gradient). The changing pressures are associated with differences in temperature (as described in the tutorial beginning on the next page). Locally, atmospheric movement is recognized as wind. The wind may seem to move in a linear pattern but can change direction and even swirl (shifting or spinning).

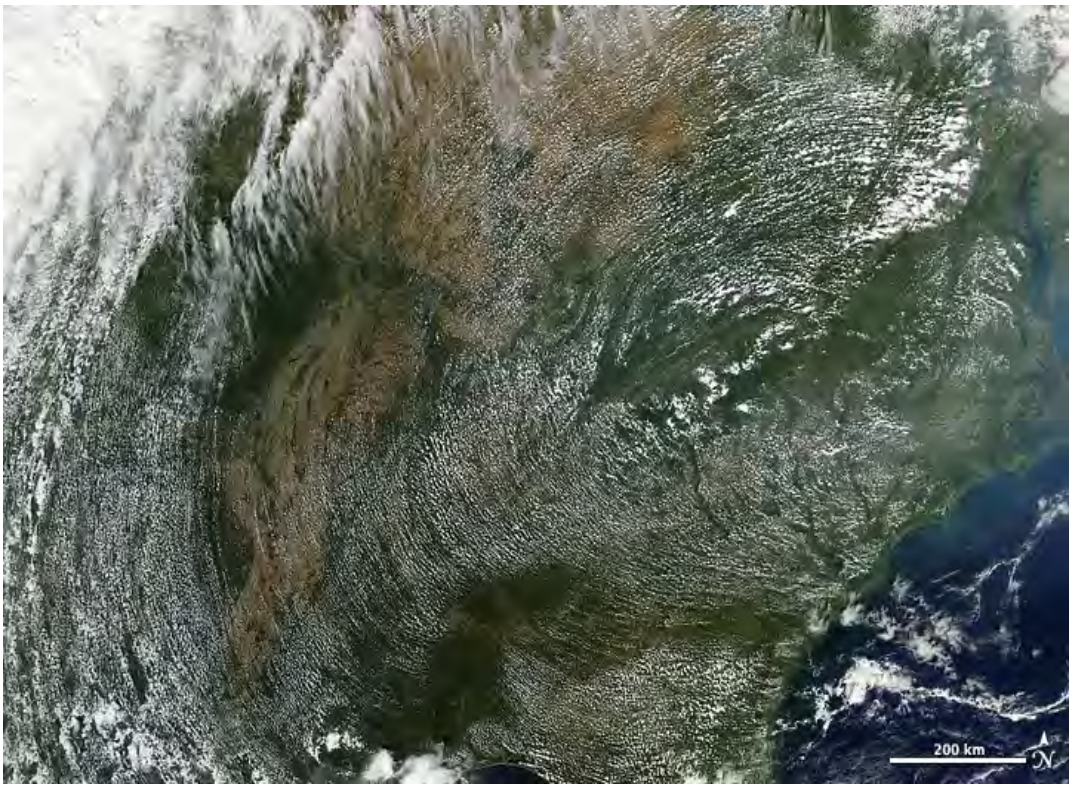
Clouds result when water vapor is cooled enough to condense. Clouds can move in circulating atmospheric patterns. These are distinctive in shape, approaching spiral forms (gyres). These circulation patterns are the result of the Coriolis effect(Wikipedia website). When the air is clear, satellites may not "see" the pattern. Clouds make the pattern visible. The circulating clouds when visualized over time (either from geostationary satellites or multiple satellite passes) seem to be moving like wheels. In the northern hemisphere, those cloud systems that appear to rotate counterclockwise (hands on clock would move "backwards") are termed 'cyclones'; those rotating clockwise are called 'anticyclones. Both modes of circulation are evident in this image.



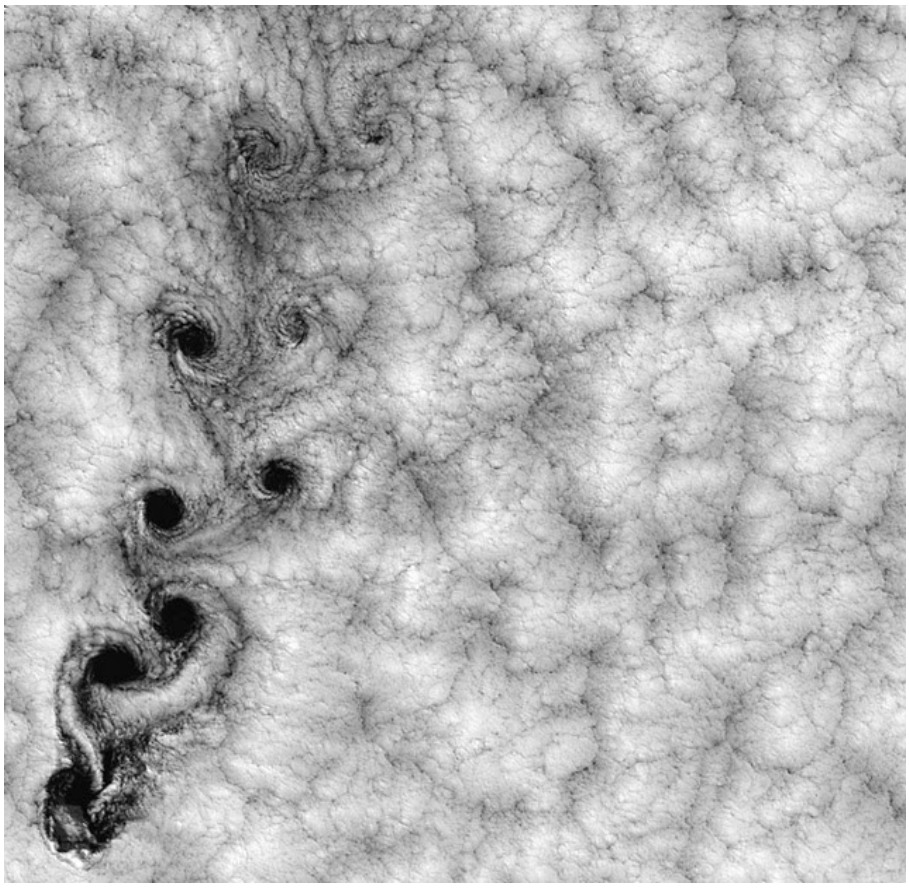
One of the most striking, and common, cloud patterns is the cyclone. In the northern hemisphere, this cloud system has a counterclockwise (ccw) spiral swirl that is often associated with a major low that delivers rain and even stronger storms, such as hurricanes. Wide field views can encompass the full systems of clouds comprising such lows.

Here are two images: the first shows a single cyclonic low and the second two cyclonic cloud banks off the coast of Iceland:

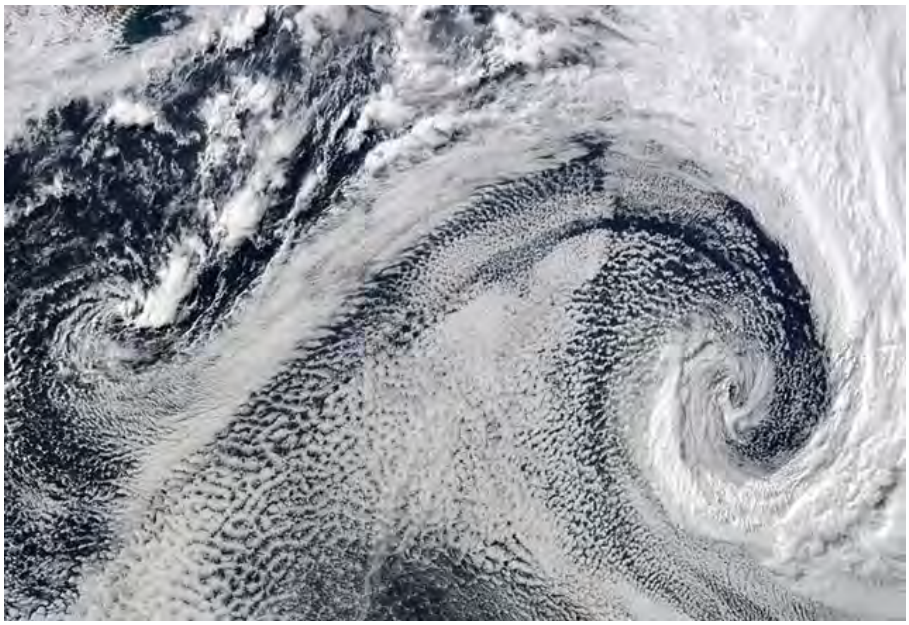
Atmospheric highs are often associated with fair weather systems; hence, the clouds may be few and a high is hard to recognize. This MODIS image shows a high with cumulus clouds.



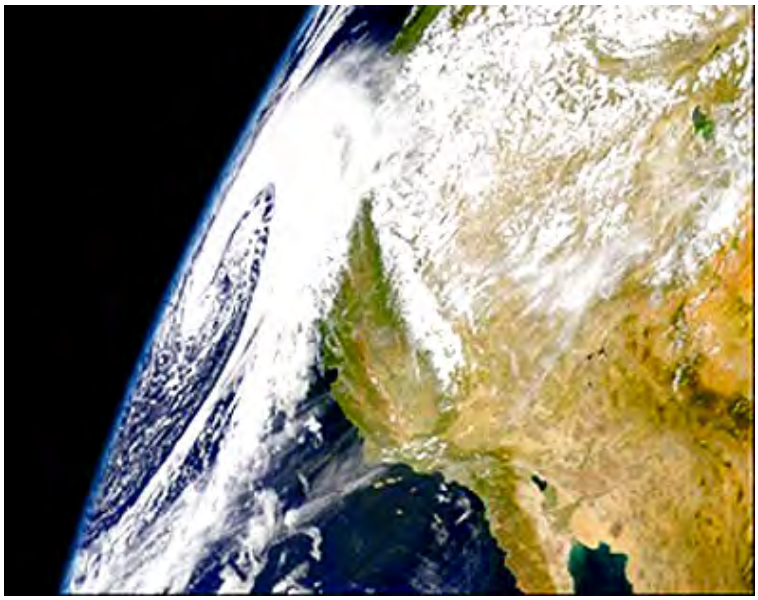
Multiple highs are evident in this next image:



In the southern hemisphere, again owing to the directions imposed by the Coriolis effect, the direction of rotation is reversed, with highs moving counterclockwise and lows clockwise.



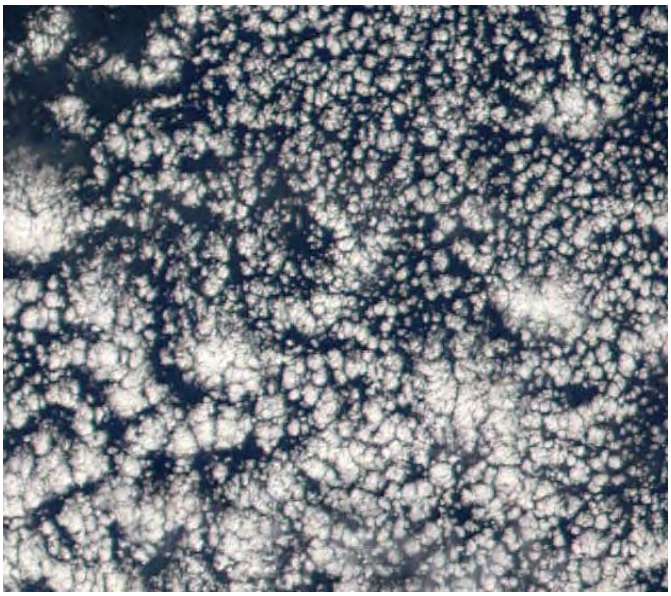
The appearance of highs and lows is affected by the angle at which they are imaged. This view of an incoming storm low approaching the coast of California was made by the SeaWIFS instrument on the OrbView-2 satellite. The low has been compressed into an ellipse because this view actually looks to the horizon, causing the storm clouds to appear distorted owing to the curvature of the Earth.



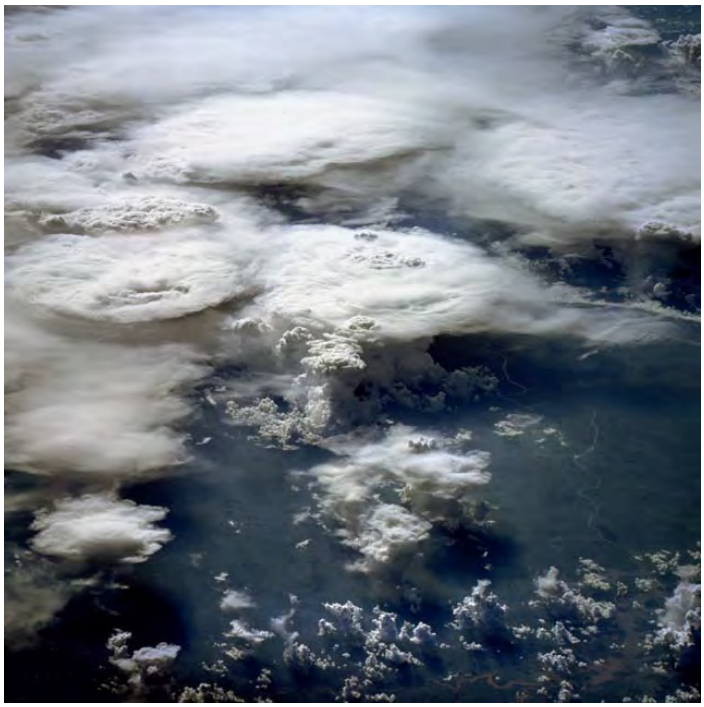
The spiral pattern can develop over land or water bodies less than ocean-sized. A ccw spiral assemblage of clouds has formed over the Black Sea that touches northern Turkey, southern Ukraine and other neighboring countries. The adjacent land has few clouds, mostly independent of the cloud cover over the Black Sea as heat has evaporated water and condensed it into this forming low. This MODIS image covers 700 km (430 miles) on a side:

The cloud type most familiar to most of us is "cumulus", sometimes known as "fair weather clouds". Here are cumulus clouds over the Amazon jungle:

Similar to cumulus clouds are these puffy clouds, often called "cell" clouds to denote their origin by advection from the ocean surface:

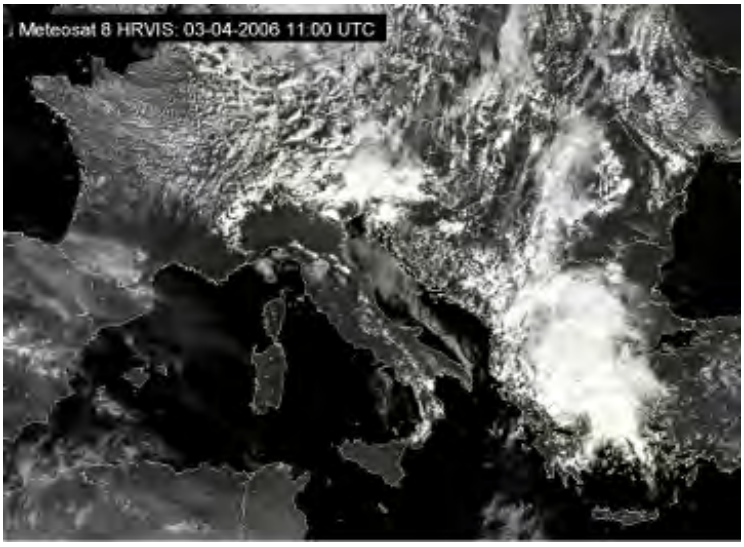


Cumulonimbus clouds can rise well up into the atmosphere and may be quite wide. These thunderhead clouds show as broad tops when seen from above, either from space or from high flying aircraft/



Cumulonimbus clouds can be dispersed laterally by wind currents, producing what has been termed an "anvil cloud". Here is a view from the International Space Station.

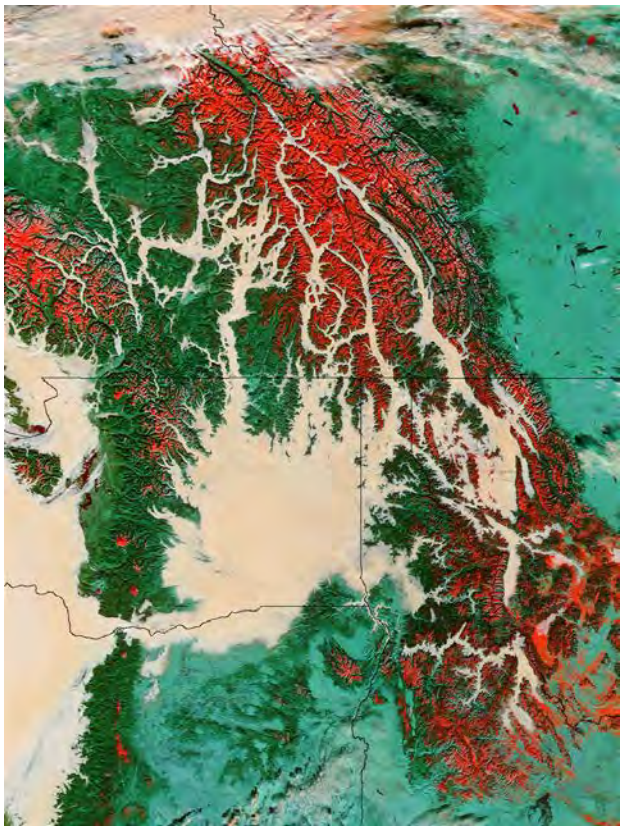
Cumulonimbus clouds can coalesce to form a huge continuous cloud cell, as seen in this meteosat image of the Balkans:



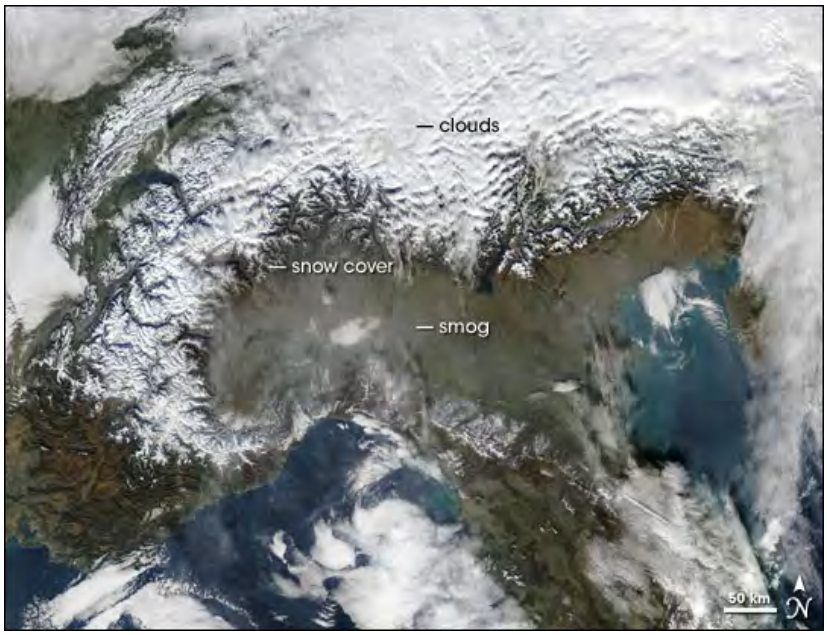
At the other extreme in altitude is fog. Fog is simply a cloud bank so low that it pervades the environment on the ground, often creating conditions of poor visibility. One of the most famed of fogs is that which can blanket the London, England area, producing "spooky" conditions that seem to be favored in movies featuring that city. In mid-December, 2006 several nights of thick fog covered London, making travel difficult as the airports were closed. Here is this fog as seen by MODIS:

We saw an example of valley fog in Section 6, on the page that includes San Francisco. Here is another satellite image that shows this type of fog in the San Joaquim Valley:

Both snow and fog appear in this California scene. How can one tell them apart? Using different bands to produce a color composite, the two can be separated as seen in this MODIS image of the northern Rocky Mountains in which snow is shown in red, fog in peach, and vegetation in green:



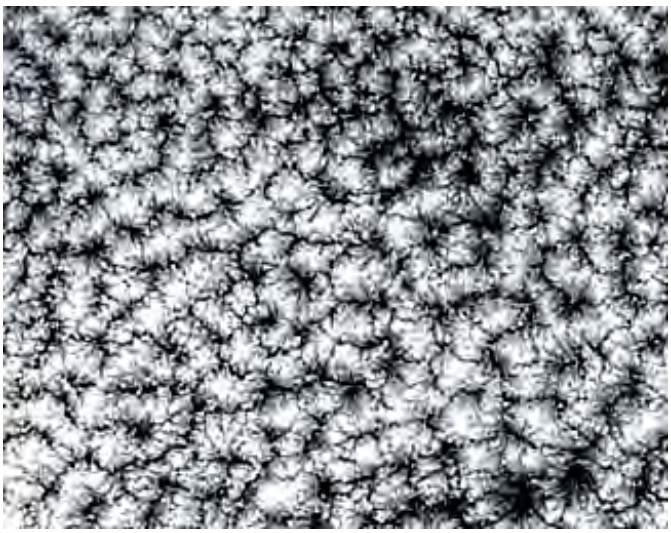
Smog is a type of fog in which harmful chemicals may comprise the bulk of the cloud or may mix with natural fog. Smog is human-induced, resulting from the burning of fossil fuels (in autos) and power and industrial plants. Here is a mix of fog and smog in northern India, and a second image in the Po Valley of Italy:



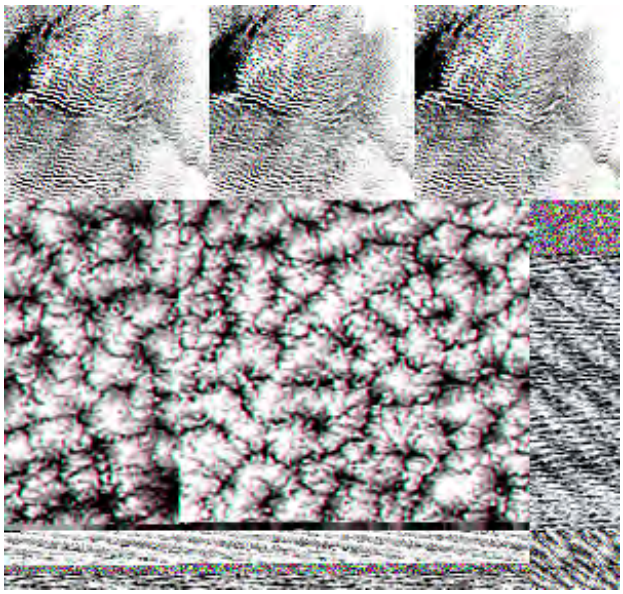
Still another dark cloud maker is volcanic ash and steam, producing what is called "vog". This example shows the blackish vog over the island of Hawaii, with the three sources of volcanic material labeled in the image:

Volcanic eruptions produce distinctive clouds both from the vantage of the ground and from space. These are examples:

The scale of Landsat images, covering 180 km on a side, is especially suited to showing clouds in some detail but over an area in which their context is well displayed. This Landsat image shows a pattern of stratocumulus cloud cells, each about 7-10 miles (10-15 km) over the Pacific Ocean. As warmer moist air rises in convection cells over the ocean and cools, condensing the water vapor into cloud droplets (which usually coalesce to form clouds), cold air then sinks around the sides of the cells.

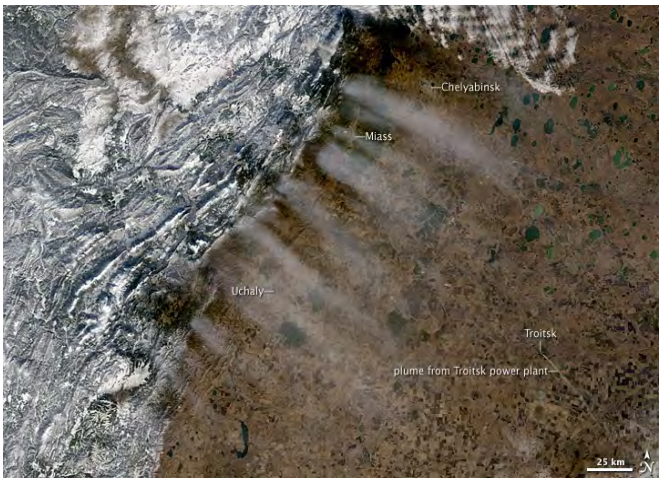


Stratocumulus clouds can often arrange themselves in waves, much like ripple marks on sand dunes, as evidenced in this Landsat image taken over the Barents Sea, near the Kola Peninsula:

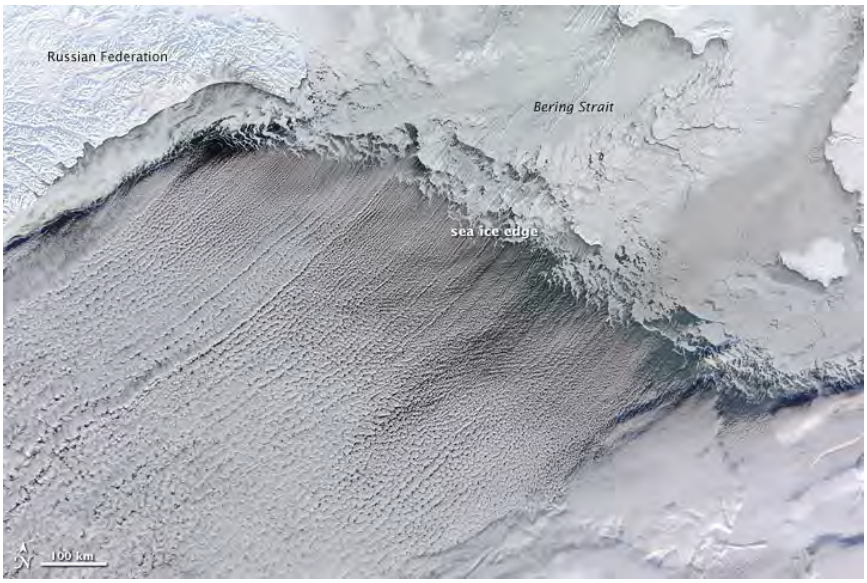


Stratocumulus clouds are common above the oceans, as seen here in this MODIS image of the west coast of the United States. Of special interest is the crosslink between these clouds and low fog along the coast and especially in Puget Sound (state of Washington) and the San Francisco Bay (California).

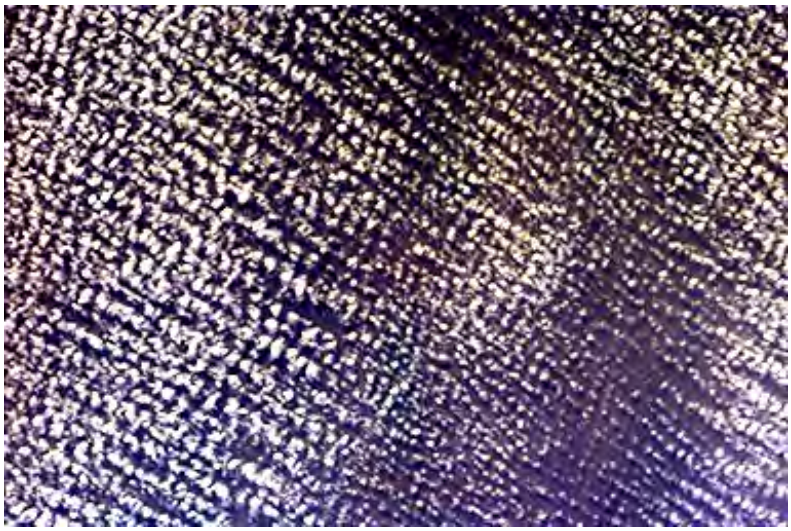
Note in the above image that the large areas east of the Sierra Nevada and Cascade mountains show up as browns in this color version. Nevada, Oregon, and Washington there consist of arid (desertlike) country owing to the *orographic effect*, which results when moisture-laden air is forced upwards by mountains, cooling the air and causing significant precipitation, leaving drier air to proceed past the mountains. This effect is evident in this MODIS image of the eastern Ural Mountains of Russia, with thin linear cloud formations over the plains of western Siberia as air moves eastward and downward off the high terrain.



The next visible image, taken on an afternoon over the Bering Straits off Alaska, shows a series of cumulus clouds aligned in *cloud streets*. The cumulus clouds result from radiative heating over land, which forces buoyant bubbles (thermals) up. The cumuli are the visible tops of these thermals. They are aligned by wind shear.



The next image, taken on an afternoon over Kenya, shows a series of cumulus clouds aligned by wind shear in a pattern similar to cloud streets. The cumuli are the visible tops of these thermals.

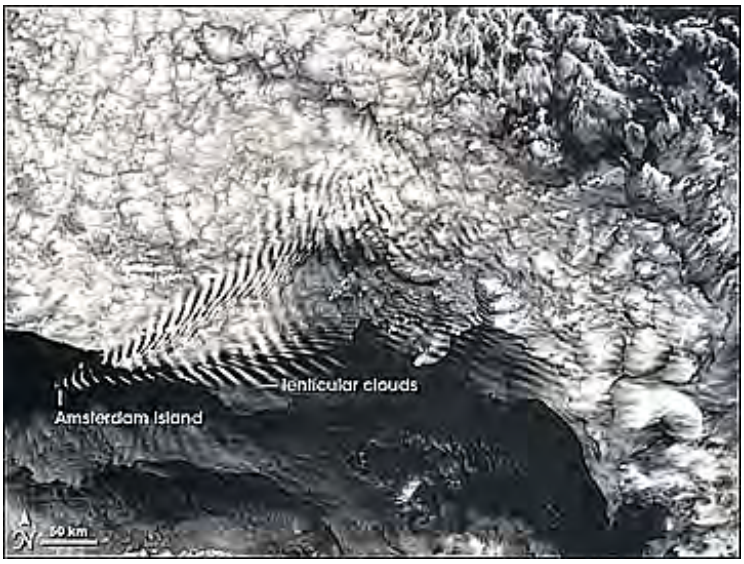


Cloud streets can be seen in a fuller context in this satellite image of western Hudson's Bay in Canada. The winds are blowing eastward off the winter snow and ice on land over the open ocean:

This next image was taken by the MISR sensor on Terra. A chain of swirls, known as von Karman vortices, has formed from stratocumulus clouds on the leeward side of the Beersburg volcano (about 2.2 km high) rising on the Jan Mayen island (Norwegian) 600 km east of Iceland. Winds streaming past this local obstruction induce the rotational perturbations that are expressed downwind as the vortices.

Similar gyres have formed in the stratocumulus cloud field over the ocean off the West African coast, as the northerly prevailing winds blow around the Canary Islands.

A particularly pleasing assemblage of clouds is seen in this image over the Indian Ocean. Most intriguing is the V-shaped waves east of the small Amsterdam Island. That island serves as an obstruction which disturbs the air so as to produce the distinctive lenticular waves beyond it.



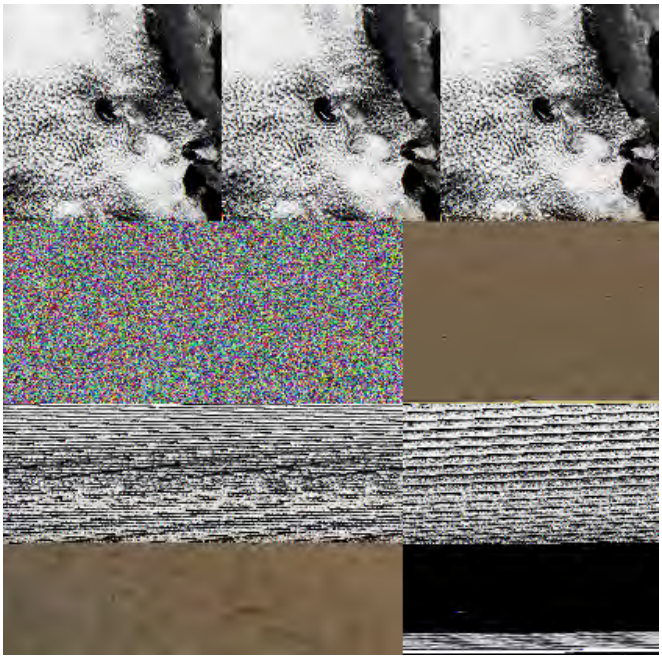
The effect of obstructions, such as mountain peaks on small islands, is even more striking in this image:

Some cloud forms are both odd and rare. An example of this is the "roll cloud", an elongated single cloud that has an elliptical cross-section. It is produced by the passage of a cold front over moist air.

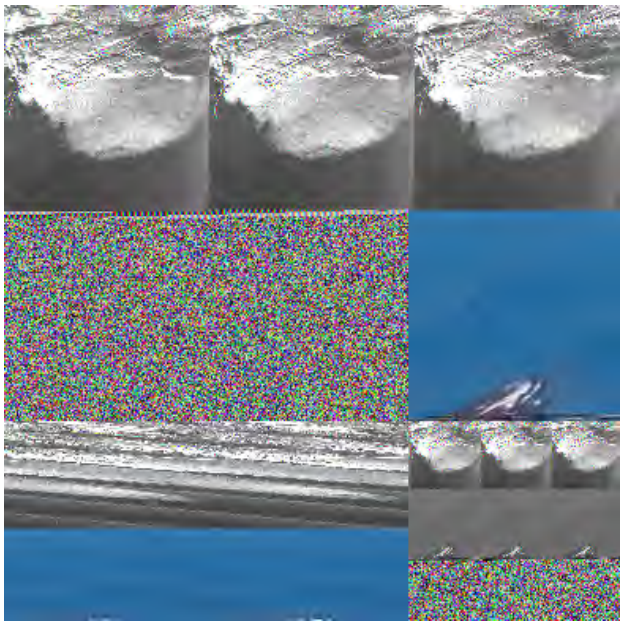
Another unusual cloud type is the "lenticular cloud", of which multiple ones have formed over Mount Rainier in Washington state:

Sometimes such clouds have been mistaken for hovering UFOs (but the cloud eventually dissipates):

As you will see on subsequent pages, most satellites whose prime purpose is to gather meteorological observations and atmospheric properties data produce images covering a wider area of view than the above examples. One non-Metsat satellite, HCMM (the Heat Capacity Mapping Mission), has an FOV yielding a swath width of 716 km (447 miles). It has yielded a number of cloud-rich images in which the details of individual cloud groups are evident. Consider this view which shows von Karman cloud vortices in the Pacific Ocean west of Baja California. They begin to form near the coast as subsiding air caused by local coastal convection sets up the eddying motion that generates these clouds whose bases are perhaps about 300 m (1000 ft) above the sea surface.

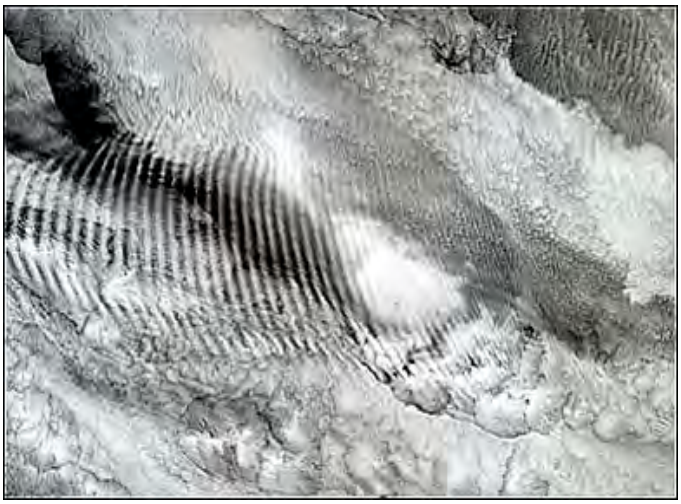


Now consider this next pair of HCMM images taken during the day on December 5, 1978, which reveal some characteristics of air masses. The top image is in the Visible-Near IR (0.5-1.1µm); bottom is a Thermal IR (10.5-12.5 µm) image.



At first glance, the top image looks like a bank of clouds extending over a featureless cold body which could be the ocean somewhere. But the bottom image shows that dark area to be land and the cloud cover to be uniformly cold. The area of clouds is now uniformly dark - hence very cold - compared with the somewhat warmer land mass that lies around the border between north Texas and south Oklahoma. In the top image, the land surface was uniformly cold and hence featureless. In the bottom image, the boundary between visible land and cold clouds indicates the sharp front between the advancing colder air mass on the north and warmer air to the south.

Another special cloud feature is gravity waves formed at the top of stratocumulus clouds. Formation mechanisms are covered on page 14-1d (in the Meteorology tutorial accessed at the bottom of this page). Here is an example of these "ripples" as they formed over the Indian Ocean:



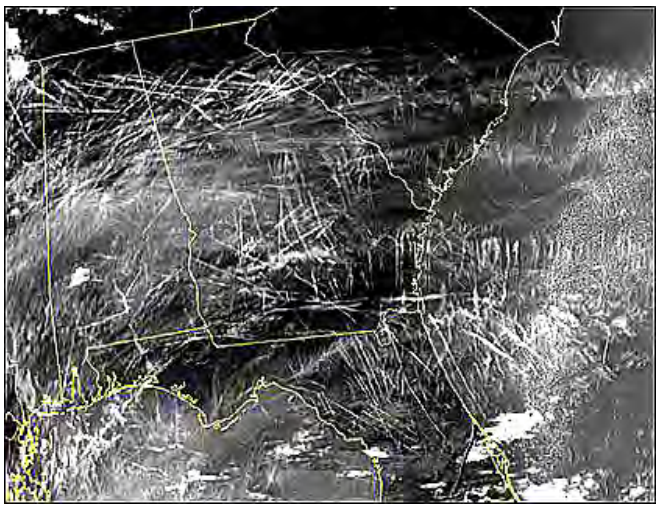
An unusual cloud type that forms in the Mesosphere, around heights of 80 km (50 miles), is the *noctilucent* type. They are visible mainly during deep twilight and result from the Sun's rays reflecting from high altitude ice crystals. Here is a brilliant example:

Noctilucent clouds are encountered by astronauts viewing them from the International Space Station. This is a typical photo showing the clouds in blue:

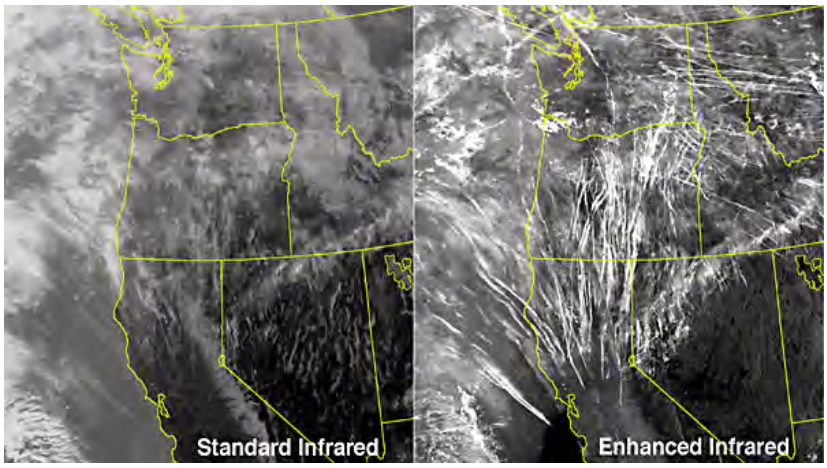
One unique type of cloud is manmade. Contrails occur when water-laden exhaust from jet engines condenses. A narrow line of moisture makes up the contrail; or, the contrail may consist mainly of ice crystals. Winds eventually dissipate it; in some instances conditions permit the contrail to survive for many minutes (their straight lines do distort). Contrails are believed to

affect weather by raising both short and long-term temperatures (one estimate is for about a third of a degree per decade). Here are two views of contrails: the first shows contrails developing behind a six engine bomber; the second shows contrails above Belgium developed as air warfare resumed during the 1944 Battle of the Bulge.

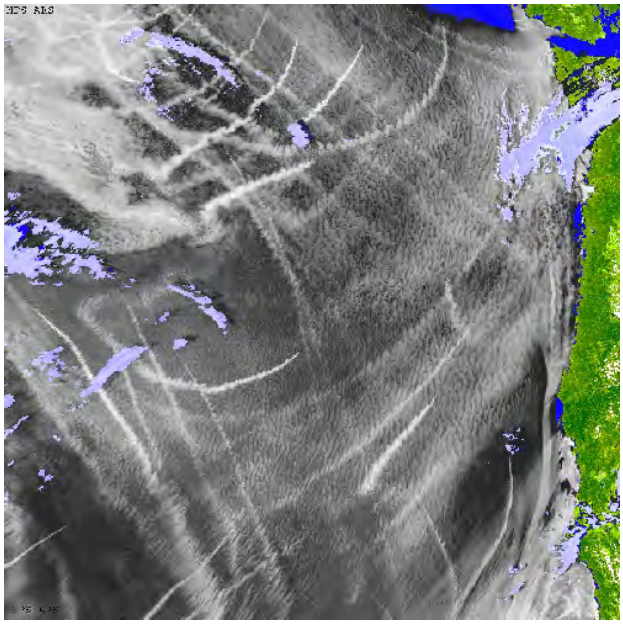
Contrails are readily spotted in all kinds of space imagery. Here is a MODIS image taken over the southeast U.S. on January 29, 2004 showing a large number of contrails (at times more than 2000 planes are over the North American continent at any one time):

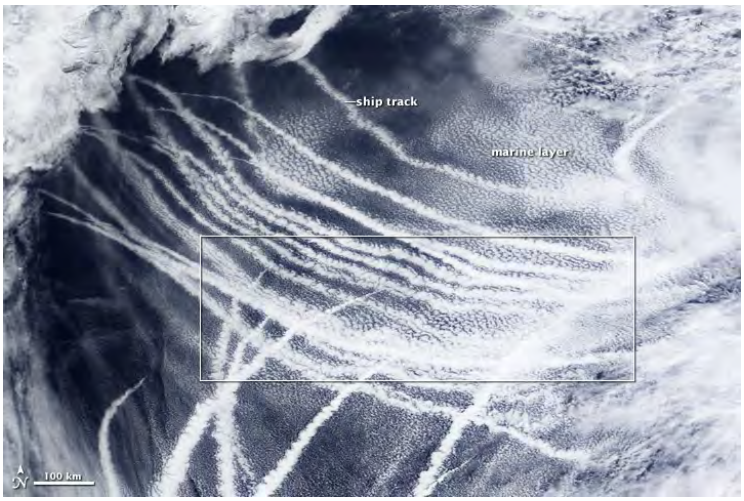


But contrails can show up better in certain wavelength bands, as illustrated here:



Contrails can form at low altitudes as well as high. Condensation of moisture around particles of soot, etc. released from the stacks of large boats form what is called "shiptracks", as shown in these two examples:



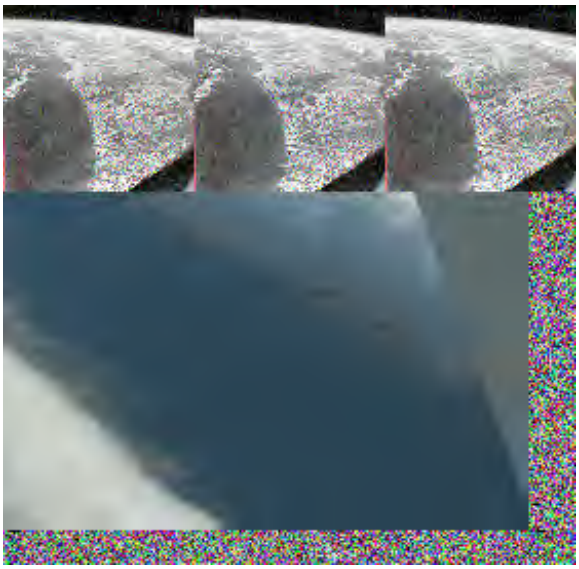


This next cloud type would seem to be related to contrails but it is natural, not manmade. It goes in Australia by the name of "Morning Glory" cloud. It appears to require a special mode of air flow in a streamlike pattern. Here it is as seen from a light aircraft:

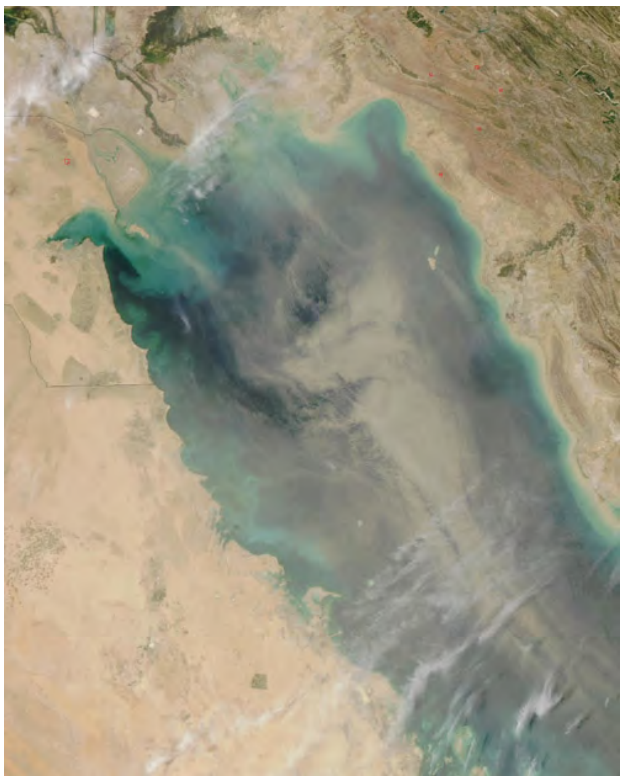
Clouds and vegetation often have a cause-effect relation. Vegetation, mostly trees and grasses, introduces notable amounts of moisture through the evapotranspiration process into the atmosphere where the excess forms clouds. Below are two photos from the EarthKam on STS-76 (see page 12-5) that illustrate this. The first shows the island of Trinidad off the Venezuelan coast, fringed by clouds where marine air become enriched with moisture from the coastal trees. The second is a view of the Amazon River in Brazil; near the river, vegetation is from marshlands that give off much less moisture (hence cloud sparsity) than the thick tree canopy of the surrounding jungle.

Ocean waters often have few clouds whereas neighboring land supports more owing to vegetation abundance. This is the case in this photo from Gemini V showing clouds over the

Florida Peninsula but their absence in the Gulf of Mexico and the Atlantic Ocean:

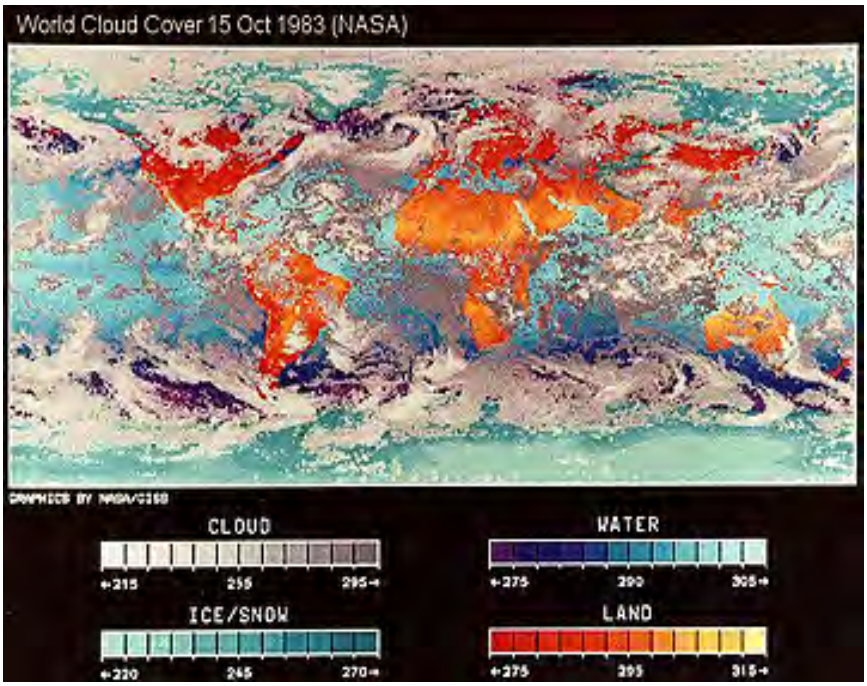


Although technically, sand storms are not clouds (they are nearly devoid of moisture), they produce an effect that can be described as "brown clouds", consisting of dust and sand particles. Here is a sand cloud over the waters of the Persian Gulf:

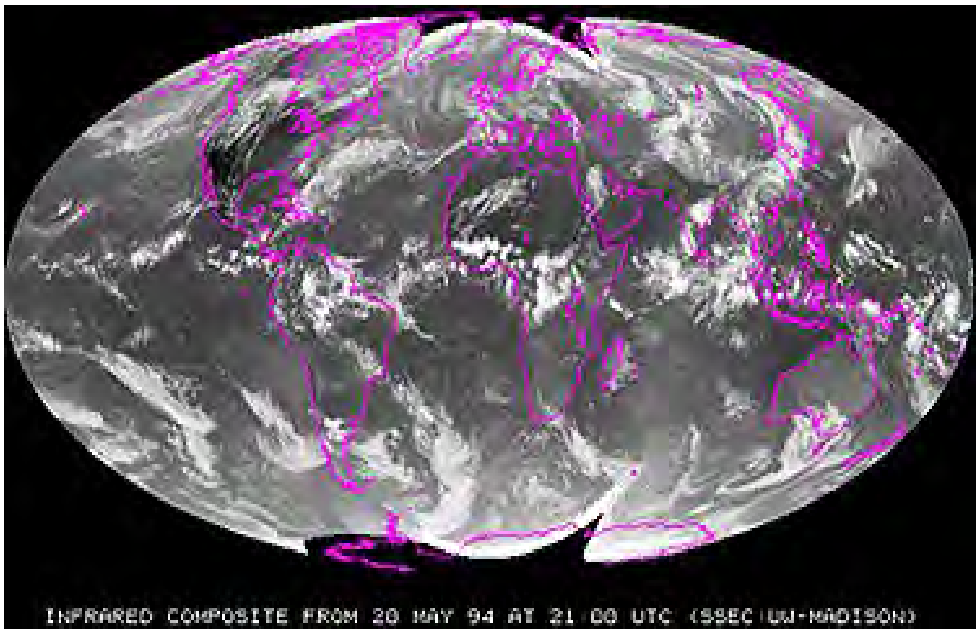


Elsewhere in the Tutorial examples were given of "auroras", cloudlike apparitions in the sky caused when solar cosmic rays hit the atmosphere. Seen at night, the auroras are often colored in blues, greens, and reds, indicating spectrally which atmospheric gas elements were being excited by the rays so as to give off light of specific wavelengths. Here is a picture of an Aurora Borealis (northern hemisphere; in the southern hemisphere the term is "Australis"):

We will close this first page by covering some of the ways in which weather and climate information is delivered to the public. As their ultimate achievement, meteorological satellites can give near real time *global* coverage of the active but transient weather systems of our planet. As we shall see shortly, both polar and geostationary orbiting satellites provide observations that facilitate this. That worldview is well illustrated by this 1983 map of cloud patterns and sea ice:

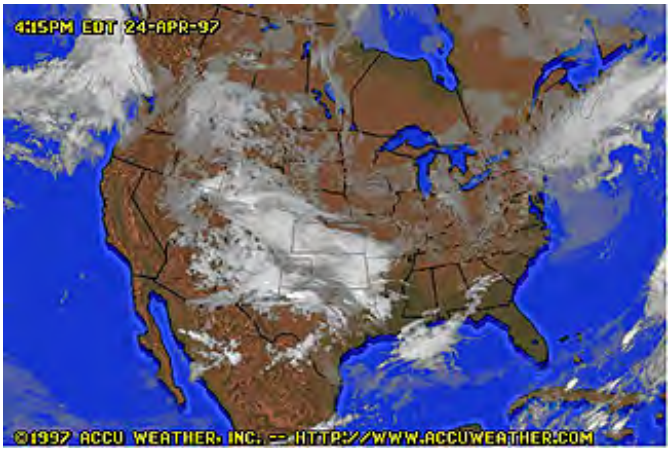


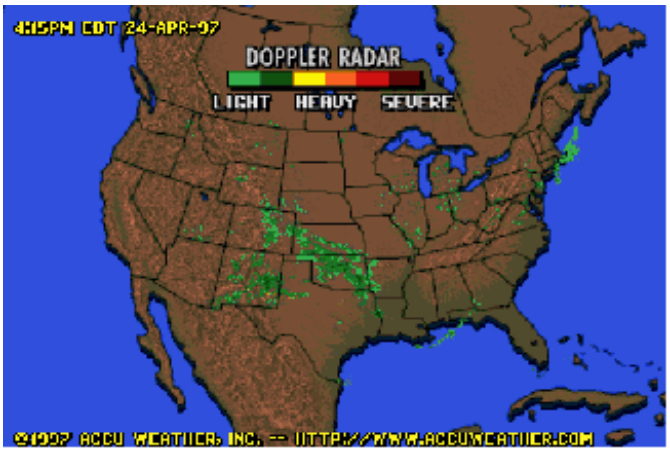
Such global views of cloud cover over a day or less are gathered routinely by making composites (mosaics) from geostationary satellite imagery. Infrared imagery usually shows sharp contrasts. Here is a GOES image that shows the worldwide cover for the 20th of May, 1994:



To most of the general public, including many in the technical fields, the one incursion of Earth-observing satellites into everyday life comes during the Weather segment of the TV news. We are familiar with synoptic views of clouds over our home region, as well as panoramas across the continent in which we live. These weather maps usually come from visible and thermal IR bands on sensors mounted in geostationary satellites. Even more common are images made by ground-based Doppler radar systems that sweep circular pattern. Radio signals bounce off (are scattered by) particulates, such as raindrops or ice, and return to the antenna, yielding estimates of precipitation amounts and wind speeds (using the Doppler principle [see page 8-2]), Doppler radars detect phase shifts in successive pulses, and employ the Doppler effect in which the target produces an increase in frequency as it approaches (or is approached) relative to the radar and a decrease as the distance between it and the radar increases.

Currently, in the U.S., the National Oceanographic and Atmospheric Administration's (NOAA) National Weather Service operates most weather radars. Their Next Generation Radar (NEXRAD) network consists of S-band radars at 164 stations across the country. This system especially detects and warns of severe storms, tornadoes, and flood-generating heavy rains. Here are two examples: cloud patterns on the top and precipitation on the bottom, for the 48 states, downloaded from the Accuweather site (http://www.accuweather.com). on the day this paragraph was written.

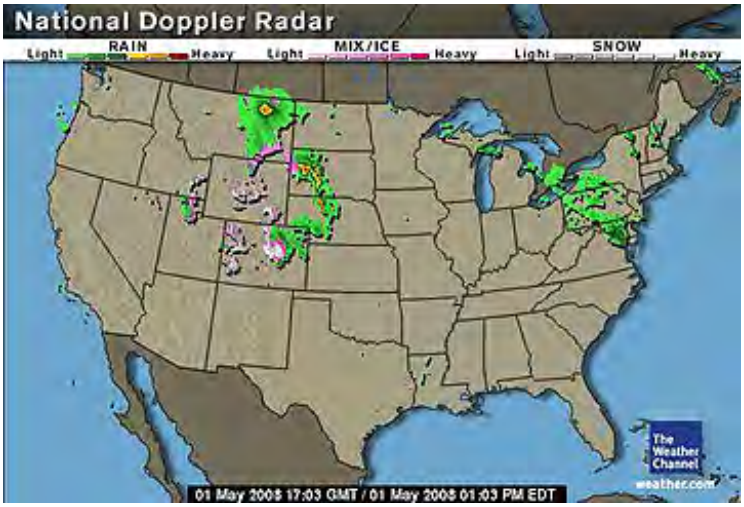




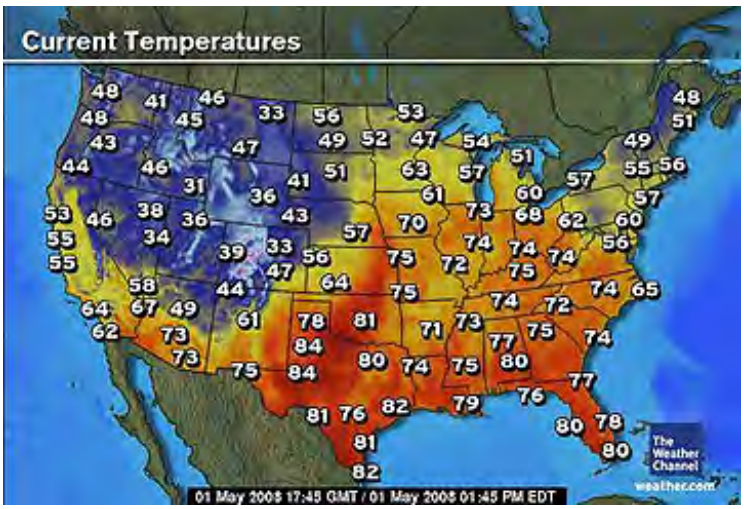
By arranging frequent observations into a time-lapse sequence (usually over an interval of the last 6 to 24 hours), the system creates and displays a " movie" of advancing weather systems from local to continental scales. .

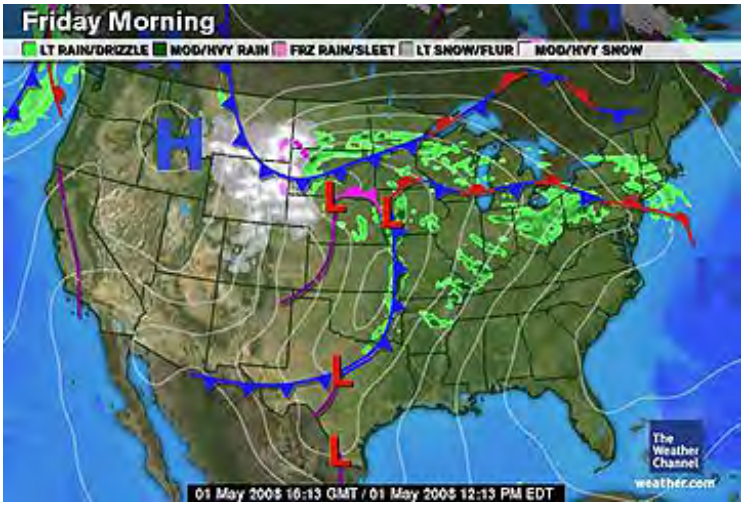
14-3: **What particular types of weather maps or images, in terms of what they show, do you remember seeing on the Weather segments of the local news shows you watch on TV? ANSWER**

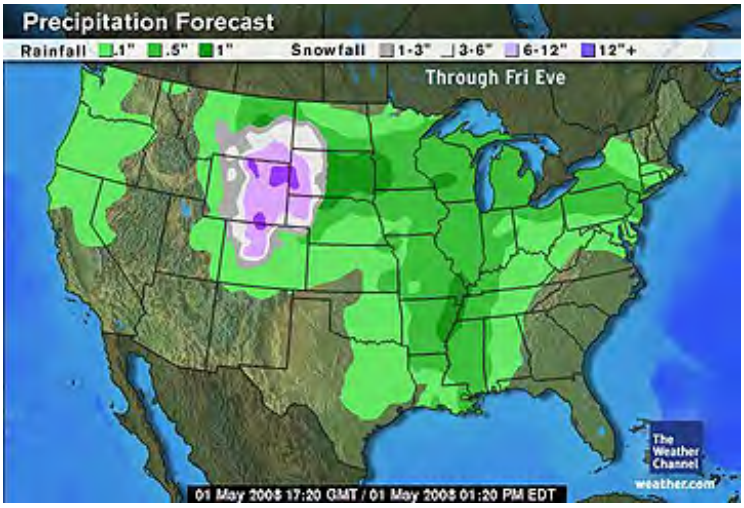
Maps such as the Doppler radar images just shown are the starting point for your local weatherman's daily presentation on the news. Those U.S.-Canadian viewers on TV cable have access to the The Weather Channel where a variety of maps appear throughout the day, and are constantly updated. We downloaded from this site a series of maps for "May Day", May 1, 2008 to show you the type of information available from TV, the Internet, IPOD, etc. Here they are - read the captions for explanations.:



IR Satellite
CLOUD TEMPERATURE
COLD
COLDEST
The Weather Channel
weather.com
01 May 2008 16:35 GMT / 01 May 2008 12:35 PM EDT

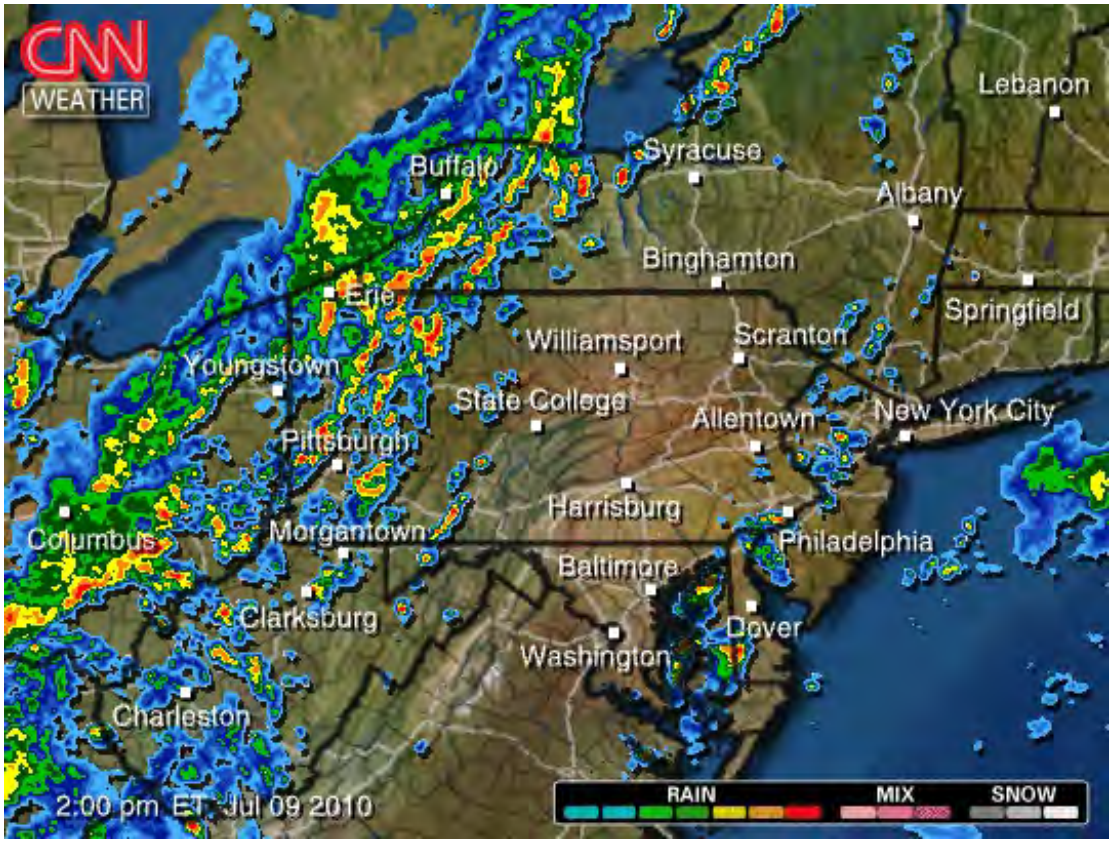

Current Temperatures
The Weather Channel
weather.com
01 May 2008 17:45 GMT / 01 May 2008 01:45 PM EDT


Friday Morning
LT RAIN/DRIZZLE
MOD/HVY RAIN
FRZ RAIN/SLEET
LT SNOW/FLUR
MOD/HVY SNOW
H
L
The Weather Channel
weather.com
01 May 2008 16:13 GMT / 01 May 2008 12:13 PM EDT


Precipitation Forecast
Rainfall .1" .5" 1"
Snowfall 1-3" 3-6" 6-12" 12"+
Through Fri Eve
The Weather Channel
weather.com
01 May 2008 17:20 GMT / 01 May 2008 01:20 PM EDT

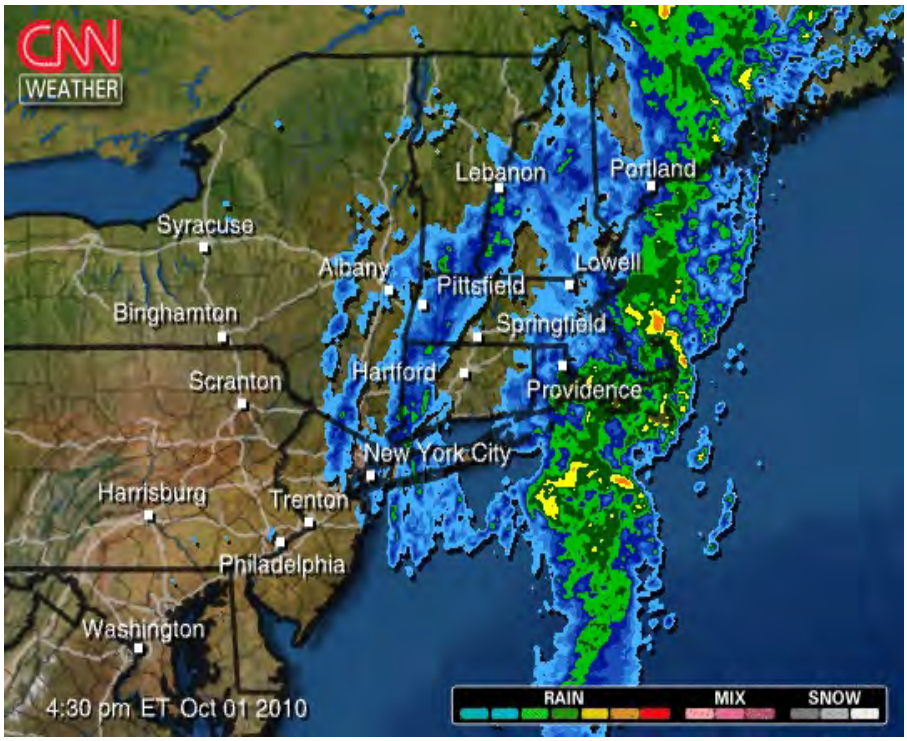
The rapid movements of weather systems make streaming video an effective tool to watch clouds and fronts in motion. NOAA has such a "movie" showing you the GOES infrared images of clouds in North America for the current day back 24 hours, accessed at Clouds. Finally, JPL has a webcast from its von Karman Series that looks to the future of weather and climate studies. Access it through the JPL Video Site, then the pathway Subject-->Von Karman Series 2003 --> Format -->Webcast --> Search to bring up the list that includes "New Weather and Climate Tools for the 21st Century" (with emphasis on AIRS, the Atmospheric Infrared Sounder), February, 2003. To start it, once found, click on the blue RealVideo link.

Internet service are such today (2010) that is easy to get current weather information in real time directly online. Click on this image to see the eastward moving cold front from the west (note that there is a "backdoor" front advancing inland from the Atlantic) on July 9, 2010, the day this image was integrated into the Tutorial.

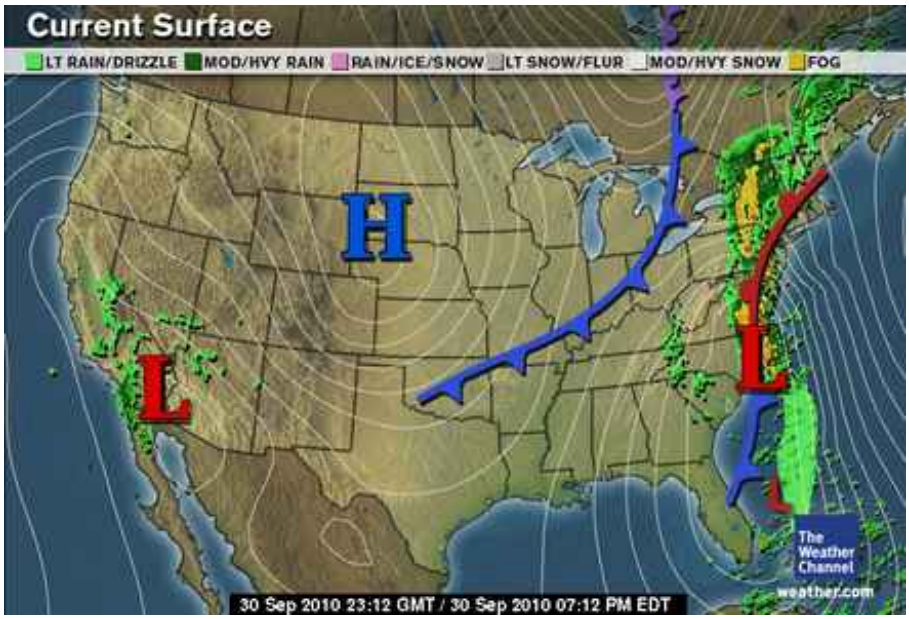


This form of streaming video helps to visualize the changes in a weather system over short periods. Unusual patterns of precipitation can be highlighted. A case in point is the extensive storm over the eastern U.S. between September 28 and October 1, 2010. The weather system

was driven by Tropical Storm Nicole which caused extensive damage in Jamaica before heading up the east coast of North America. Here are the videos for September 30 and October 1; note the distinctive northward flow:

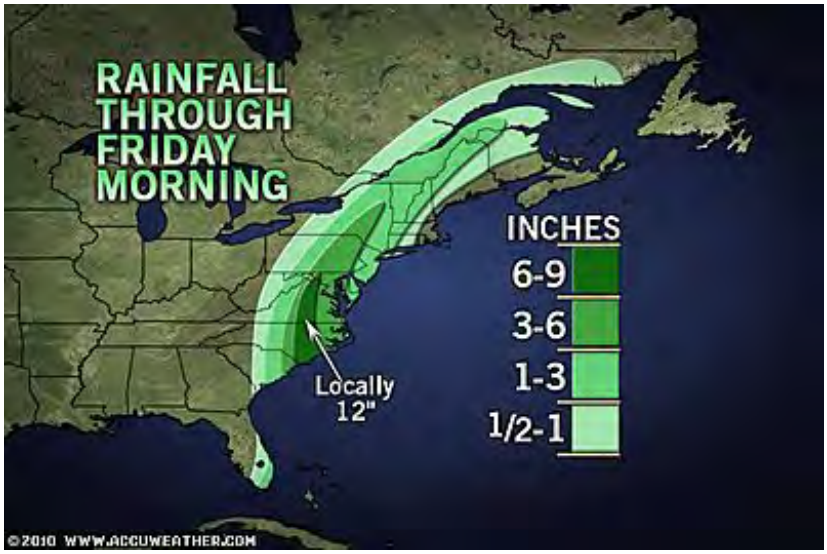


Note the location of pressure high (west) and the low associated with Nicole.

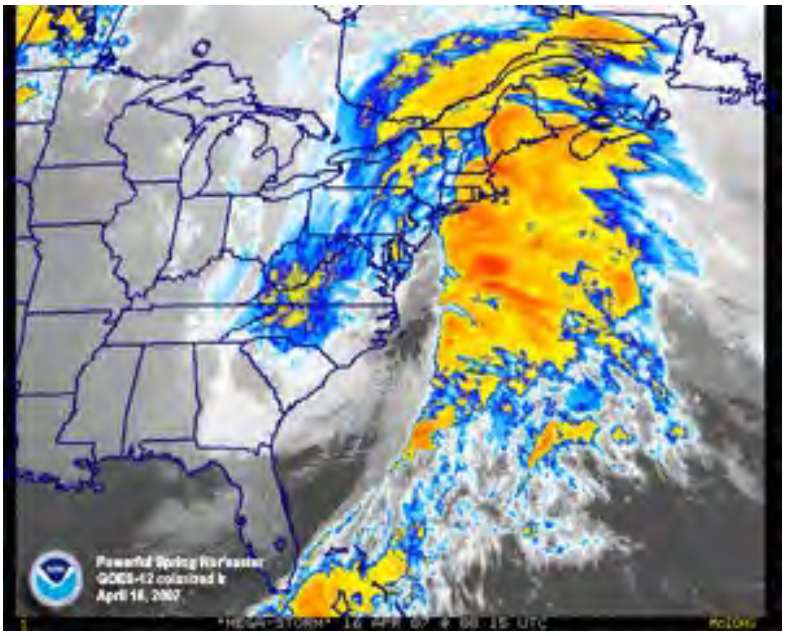


The system had a moderately low temperature distribution, as seen in this IR image.

At the time of this storm, the eastern U.S. was experiencing a summer drought. But in the last week of September many areas saw more than 20 inches of rainfall. The writer's (NMS) home area received a much needed 4 inches in one day. Wilmington, NC received 22 inches from this storm. Flooding ensued:



Here is another example of putting a storm into motion. The "noreaster" type of storm (winds coming from the northeast on the western side of a low in the northern hemisphere) is capable of snowfalls of several feet or more in the eastern U.S. (such thick falls are fairly common in western U.S. mountains):



One would think that this type of time-lapse streaming video would be an ideal way to depict and follow the path or track of hurricanes. The writer spent two full days searching the Internet for one good example of this. A check of more than 600 websites failed to find a single adequate live action case - surprising! Two examples of a composite image showing hurricanes on successive days were the best found:

Most readers of this Tutorial also follow weather reports on their local TV and in their newspaper. One thing people usually check each morning is the weather expectations for the day. In the evening the information sought is probably the 5 or 7 day forecast. CNNs website offers a 10 day forecast. How good are these prognostications? Surprisingly accurate, at least for the next 5 days or so, and not bad for periods of a week or more. What allows this? Four things in particular: 1) better atmospheric instrumentation such as balloon sounders released from many more stations; 2) better computer models that use daily measurements as well as longer term data as inputs; 3) better regional observations in real time; and 4) numerous images and other data acquired by satellites. This fourth "better" may be the most important. And this is where remote sensing from satellites plays a major role. Metsats are one of the prime triumphs of the space program.

Section 14 reviews the history and accomplishments of this use of Metsats to monitor the daily changes of the Earth's weather systems, oceans, rivers, and snow/ice and to conduct long-term research into the interactions between the atmosphere and hydrosphere that control the

meteorological state of the planet. This subject is vast - worthy of its own web site - and many web sites dealing with it now exist, as you can ascertain by doing an Internet search. In this Section, we will introduce only a digest of the types of observations - mainly by presenting images and a few graphs - in this brief, simplistic, and generalized treatment. We emphasize meteorological applications, with an abbreviated summary of selected oceanographic (surface temperatures, seastate, currents, and phytoplankton distribution) and hydrologic (flooding, water storage, and drainage regime) uses. Many people working through this Tutorial may lack knowledge in meteorology, and even more so, in oceanography and hydrology. For those who seek a broad overview, or wish to delve into greater technical and scientific details, we suggest perusal of a good introductory Meteorology or Oceanography text or, quidker yet, the relevant chapters in a Physical Geography text, such as:

- Satellite Meteorology: An Introduction, S.Q. Kidder and T.H. Vonder Haar, 1995, Academic Press, p466.
- Images in Weather Forecasting: A Practical Guide to Interpreting Satellite and Radar Imagery, 1994, Cambridge University Press.
- Meteorology Today: An Introduction to Weather, Climate, and the Environment, 1995, C.D. Ahrens, West Publishing Co.
- The Atmosphere, F.K. Lutgens and E.J. Tarbuck, 1993, Prentice-Hall.
- Introductory Oceanography, H.V. Thurman, 1988, MacMillan Publ. Co.
- Introducing Physical Geography, A. Strahler and A. Strahler,1994, J. Wiley & Sons
- Geosystems: An Introduction to Physical Geography, R. W. Christopherson, 1992, MacMillan Publ. Co.
- Weather for Dummies (a very readable overview)

WRITER'S NOTE: None of the Internet tutorials which I checked out provided satisfactory (to me) overviews of meteorology, weather, and climate. Too condensed and disjointed! Likewise, there is a paucity of material available online that instructs in oceanography or hydrology. So, in 2004 I developed a four page Meteorological Mini-tutorial that follows this page which surveys the basics of Weather and Climate in a continuous, cohesive summary. You can access the first page by clicking here. (If you choose not to engage in this instructive diversion, the NEXT button below will bypass this Tutorial and carry you to page 14-2.)

We close this first page with a parenthetical input, since nowhere else in the Section does the following information seem relevant: 1) the hottest temperature ever recorded on Earth was at Azizia, Libya: 57.8° C or 136° F [Death Valley, California once reached 134° F]; 2) the coldest spot was -89° or -121° F, in Siberia; 3) the wettest area was around Llaro, Columbia which received 323.6 inches of rain in one year; the driest spot is in the Atacama desert of Chile which receives on average less than 1 inch of rainfall per year.

Primary Author: Nicholas M. Short, Sr.

The prime sensor used now for many years on U.S. Metsats is AVHRR, whose characteristics are given. Several other satellites that have proved valuable in weather monitoring are listed with brief narratives.

The remainder of this Section will incorporate four general topics: 1) the historical development of Metsat programs; 2) examples of the kinds of data/information, as rendered in images and maps, supplied by these systems; 3) emphasis on the global nature of many of the observations; and 4) consideration of oceanographic and land hydrologic investigations. We start with an introduction to instrumentation on board meteorological satellites by considering the history of the venerable AVHRR that has flown on many satellites.

Metsat Instrumentation; AVHRR

The chief tool for conducting meteorological observations is still any sensor that images in the Visible spectrum. The target of interest is primarily clouds: their types, distribution, extent, rates and directions of movements, and potential for rainfall. Clouds and other components of weather systems - the atmosphere, and the oceans - are also well-suited to sensing in various regions of the Infrared. Water and gases have distinct absorption bands in these regions as well as characteristic emissions. Generally, when we mention an infrared sensor regarding Metsats, we refer to the thermal-infrared region of the spectrum, particularly in the interval from 10-14 µm but also the 3-5 µm region. In those segments of the EM spectrum clouds are cooler than the land and usually the water as well. If imaged conventionally, clouds would appear dark, contrary to their visible-region appearance. The image is customarily inverted (photographically when printed, or simply by numerical transformation of the digital data to make low values large and high values smaller, prior to printing or displaying) to produce images with clouds as whitist features in the scene.

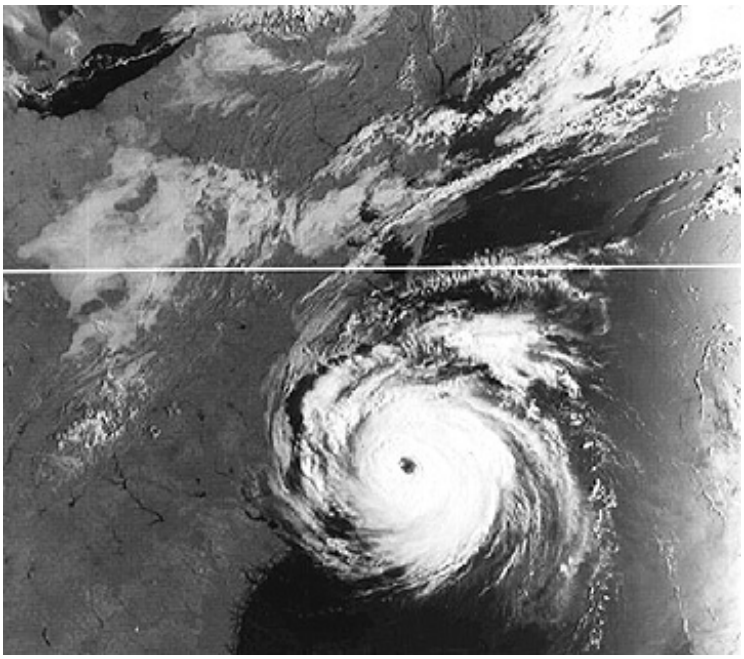
For nearly 30 years now, the workhorse imaging sensor on many Metsats has been the Advanced Very High Resolution Radiometer (AVHRR). The AVHRR has five channels, whose characteristics are:

Band Number	Wavelengths (µm)	Main Uses
1	0.58-0.68	Day clouds; snow; ice
2	0.725-1.10	Surface water; snow; ice
3	3.55-3.93	Fires; clouds at night
4	10.30-11.30	Day/nite cloud & surface temperatures and mapping
5	11.50-12.50	Same as 4; water vapor

14-4: **How many of these bands are infrared; how many thermal?** **ANSWER**

AVHRR has flown on NOAA's polar-orbiting satellites, starting with TIROS-N, and is still very much in use today (GOES series). The ground resolution at nadir for this instrument, for a field of view of 1.4 milliradians (producing a swath width of 2,400 km [1,491 mi]), when flown at 830 km (516 mi) is 1,100 m (3,608 ft).

Without further ado, we preview a typical AVHRR product: a NOAA-6 view in the visible region of Hurricane Diana, off the U.S. East Coast on September 12, 1984. Hurricanes generally are the most photogenic of the large weather systems, as we shall see again.



14-5: **What is the black spot close to the center of the hurricane called? What happens within this feature? ANSWER**

In general, s do not require the higher resolutions of land-imaging Metsats, because their principal targets are large assemblages of clouds. The highest resolution system, at 600 m (1968 ft), are sensors onboard DMSP (Defense Meteorological Satellite Program) spacecraft. The tradeoff is to couple the sensor(s) to an optical telescope with a large field of view (FOV) that creates a wide swathwidth or full disk coverage, depending on orbital parameters. We can join successive swaths to produce timely mosaics that cover large regions or most of the Earth over a 24-hour cycle. Therein lies a big difference between Metsats and Landsats, in that the former seek clouds, while the latter can't see through them, although clouds sometimes happen to be the targets of interest. Landsat, SPOT and other satellites produce excellent subregional cloud pictures suited to mesoscale analysis.

Other imaging sensors in active use on U.S. Metsats (we detail several later in the Section), include the two-channel Operational Line Scanner (OLS) on DMSP satellites, the 6-channel Coastal Zone Color Scanner (CZCS) on Nimbus-7, and the Total Ozone Mapping Spectrometer (TOMS) on several recent satellites. We don't review sensors on foreign satellites except for brief mention of pertinent characteristics when we introduce their images.

The Heat Capacity Mapping Mission (HCMM; page 9-8) provided useful visible and thermal imagery of clouds. The Earth Radiation Budget Satellite (ERBS) measures incident and reflected, and longwave radiation to determine the energy budget related to solar insolation and outgoing re-radiation (with a terrestrial target). Active radar and passive microwave sensors also operate on some Metsats. They can be particularly effective in detecting and tracking several forms of precipitation, such as rain droplets. Radars on Seasat and the SIR series (flown on Shuttles) produce images specifically for oceanographic studies. Nimbus 5 and 6 carried the Electrically Scanning Microwave Radiometer (ESMR) and Nimbus 7 included the Scanning Multichannel Microwave Radiometer (SMMR). The Special Sensor Microwave Imager (SSM/I) operates on DMSP satellites. Japanese, Canadian, and European satellites also mount radar or microwave systems, which look at water and land targets.

We can only scratch the surface (really, the atmosphere) on this subject of Metsat instrumentation. For example, the Nimbus series actually supported 24 different experimental

systems, involving a variety of sensors. The interested user should consult Kidder and Vonder Haar (reference: previous page).

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

This page is divided into two main topics. The first part considers the two main categories of meteorological satellites: those that follow near-polar orbits and image strips along their pathways and those placed at about 37000 km (22000 miles) in geostationary orbits (at the equator, the satellite moves at the same velocity as the Earth's surface beneath it). By placing a small number of geostationary satellites at different points above the globe, the whole Earth can be monitored simultaneously in real time for its moving weather patterns. A generalized system of groups of orbiting satellites is described. The second half deals with specialized instruments onboard some Metsats that are capable of taking readings or "soundings" through the atmosphere at various depths to produce profiles of such variables as temperature, water content, CO_2, etc. From these data, maps of atmospheric properties at various elevations can be constructed.

Classes of Metsats

Meteorological satellites fall into two general classes in terms of their orbital characteristics. We can group most into the Polar Orbiting Environmental Satellite (POES) category. These launch into orbits at high inclinations to the Earth's rotation (at low angles with longitude lines), such that they pass across high latitudes near the poles. Depending on orbital altitudes, angular velocities, and inclinations, these satellites strive to be sun-synchronous, that is, to cross the equator southbound about 11° westward (as Earth rotates underneath) with each trip around the world (about 105 minutes long), so that they pass over some reference position (e.g., the equator) at the same local time. This time is usually between mid-morning and mid-afternoon on the sunlight side of the orbit. Thus they image their swaths at about the same sun time during each pass, so that lighting remains roughly uniform. Of course the clouds change with each orbit, but their broad patterns and positions remain mostly unchanged in the short orbital periods involved.

From this method, we can make a daily mosaic from the swaths, which is a good general summary of global weather patterns for that period. This same orbital configuration applies to Landsat, SPOT, and some of the other land observers. Other POES members have inclinations or other orbital constraints, such that they cross equivalent latitudes at different times of day, allowing observations over various times in the diurnal cycle. Most POES orbits are circular to slightly elliptical at distances ranging from 700 to 1700 km (435 - 1056 mi) from the geoid. At different altitudes they travel at different speeds.

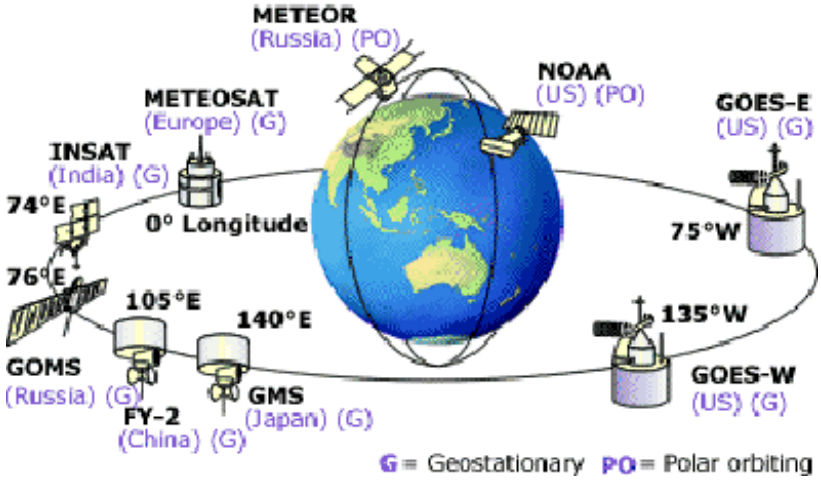
The second type of Metsat is in the Geostationary Operational Environmental Satellite (GOES) class. These satellites are *geosynchronous*, meaning their orbits keep them synchronized with Earth's rotation, i.e., they take 24 hours to complete one orbit. To have a 24 hour orbital period, they must keep an orbital altitude of 35,780 km (22,234 mi, or about 5.61 Earth radii), which sets their speed at 3.07 km/s (6,868 mph).An equatorial point travels underneath at a speed of about 0.465 km/s (1,040 mph).

When these satellites orbit above the equator, with zero inclination, they are also geostationary (fixed) relative to a point on the equator, so that they observe the Earth without any significant relative motion. At this distance, and with a wide FOV, they see the Earth as a full disk, but the area covered is less than a hemisphere, being about 1/4th of the planetary surface.

The Earth and satellite, synchronized, rotate relative to the Sun, so that part of each day, the satellite's sensors face the night side. But, because of this night effect, to view the Earth in daylight continuously requires at least three (preferably four) geostationary satellites located equidistantly around the globe *above the equator*. Under these orbital conditions, the sensors can look at the disk at any moment, gaining a synoptic image, in which it views all meteorological elements in real time. We acquire visible images typically 30 minutes apart and infrared images less often. GOES platforms usually operate in pairs, spacing about 75° apart. We refer to current GOES units above the western hemisphere as GOES-East and GOES-West.

14-7: **Which type of satellite most likely gives you the cloud movement panorama you see in the nightly TV news? ANSWER**

This next figure offers a quick look at most of the major Metsat systems (POES and GOES) now operating to provide worldwide or hemispherical coverage.



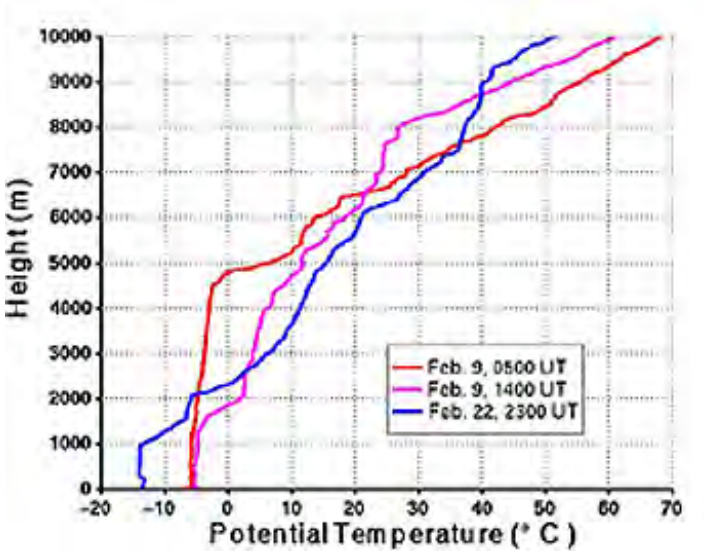
Most of these will be discussed later in the Section as we go through the various Metsats operating in both the past and present. Before taking this sojourn, a diversion is necessary to emphasize the three-dimensional nature of meteorological remote sensing.

The 3-D Atmosphere: Atmospheric Sounders

Throughout this Tutorial so far, the various applications of remote sensing have dealt almost exclusively with obtaining information pertaining to two-dimensional targets - principally, the Earth's surface. (Several aspects of Geophysics, considered in the Introduction, are concerned with subsurface measurements.) But it is obvious that remote sensing of the atmosphere - a thick layer around Earth - must be three-dimensional (3-D) in its efforts. The concept of probing the atmosphere to determine its properties from surface upwards is embodied in the term "atmospheric sounding".

Sounding is a nautical term that applies to the ancient practice of dropping a line with weights overboard to determine the depth of waters being traversed by a ship. Atmospheric sounding is similar but is accomplished by releasing "weather balloons" that rise upwards carrying measuring instruments.

Temperatures, pressues, moisture content, wind speeds, and composition are the principal variables sought by a weather probe. Temperature is the most frequent parameter thus measured, as seen in this plot:



It would take a huge fleet of weather balloons constantly being sent aloft to get a good picture of the atmosphere's physical state over large areas (actually, volumes). Now, satellites can perform this job much better. Obtaining 3-D atmospheric data with satellite sensors is difficult but doable. These sensors are in effect sounders, which can look vertically through atmospheric levels. Limb sounders look off-nadir through longer paths nearly tangent to the Earth's surface. Most operate on this principle: various properties of atmospheric gases are temperature-dependent. Wavelengths of emissions and absorbances (peaks and troughs) will vary with temperature. If those wavelengths can be measured, variations of temperature with depth into the atmosphere (related to altitude above the surface) can be sounded. Since temperature is the prime parameter to be sensed, the thermal infrared and the passive microwave regions are used almost exclusively in atmospheric sounding.

The writer's attempts to surf the Internet to obtain a clear picture of how atmospheric sounders operate has proved difficult and unrewarding. One passage on a website maintained by CEOS (the Committee for Earth Observations by Satellite) is the best found. Three paragraphs from that site are reproduced here (in italics):

Atmospheric sounders generally make passive measurements of the distribution of IR or microwave radiation emitted by the atmosphere, from which vertical profiles of temperature and humidity through the atmosphere may be obtained. Oxygen or carbon dioxide is usually used as

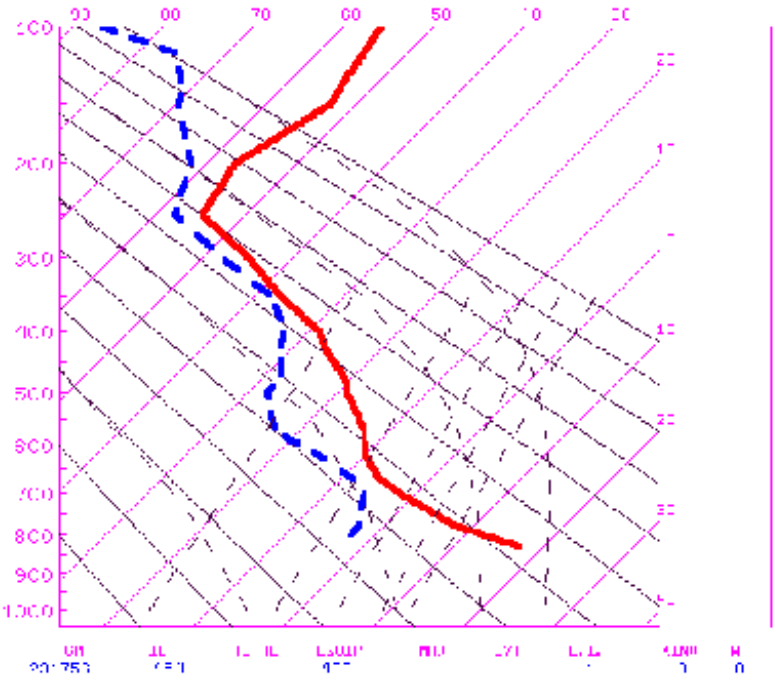
a 'tracer' for the estimation of temperature profiles since they are relatively uniformly distributed throughout the atmosphere, and hence atmospheric temperature sounders often measure radiation at wavelengths emitted by these gases. For humidity profiling, either IR or microwave wavelengths specific to water vapour are used. Most measurements are conducted in nadir viewing mode.

Sounders are able to estimate profiles of temperature and humidity by identifying radiation coming from different levels in the atmosphere. This is achieved by observations of the spectral broadening of an emission line, a phenomenon which is primarily caused by intermolecular collisions with other species, and which decreases with atmospheric pressure (and therefore is a function of altitude).

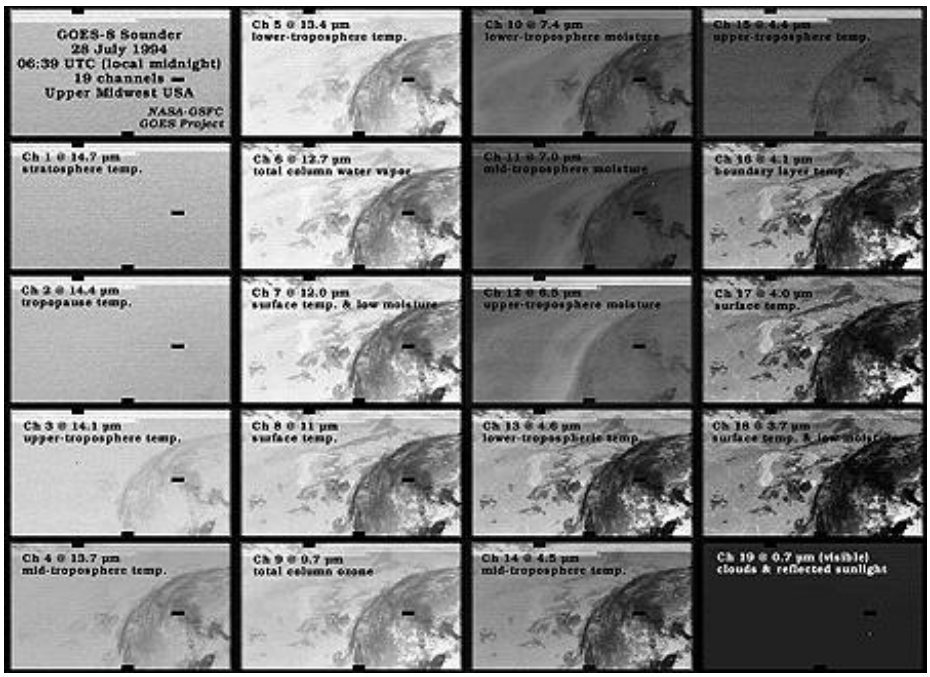
Microwave sounders have the ability to sound through cloud and hence offer nearly all-weather capability; their spatial resolution (both vertical and horizontal) is generally lower than that of the IR instruments. IR sounders are routinely used to provide temperature profiles from a few km altitude to the top of the atmosphere with a temperature accuracy of 2-3K, a vertical resolution of around 10km, and a horizontal resolution of between 10 and 100km.

The way in which the AIRS sounder (see below) determines temperatures/water vapor throughout the atmosphere is described by T. Pagano (pers. comm.) in this way: The individual spectral channels have weighting functions that peak at different altitudes in the atmosphere. A Singular Value Decomposition retrieval (there are other methods too), combined with a radiative transfer program (computes the expected upwelling radiances given a temperature and water vapor profile) is used to solve for the temperature and water vapor profiles. It does this by minimizing the observed spectrum with the calculated spectrum.

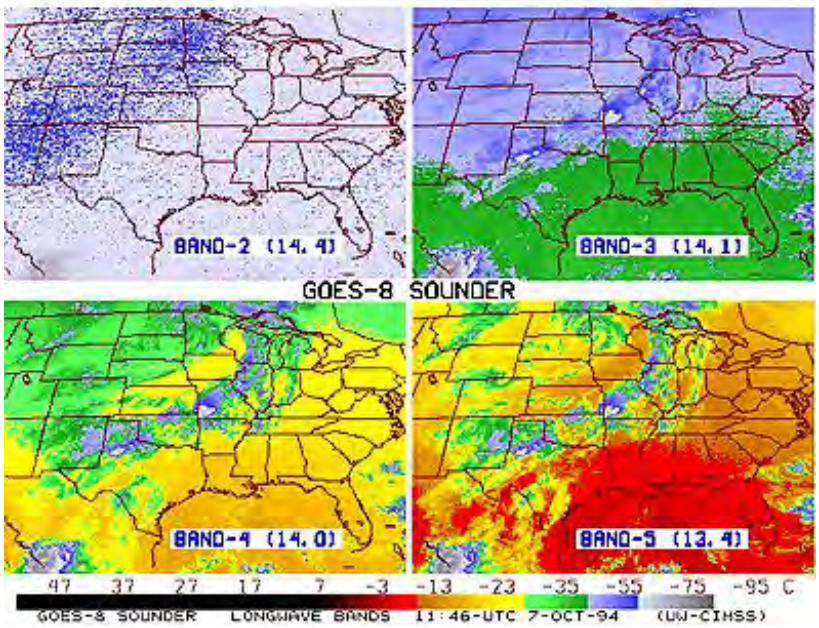
Among early Metsat instruments, we can mention as typical: the High Resolution Infrared Sounder (HRIS - 20 channels that cover from the visible to 15 µm, with a coarse resolution at 42 km); the Stratospheric Sounder Unit (SSU); the Microwave Sounder Unit (MSU); the Visible Infrared Spin-Scan Radiometer (VISSR), a two-channel sounder on GOES, which also may have the GOES 1-M sounder; the Solar Backscatter UltraViolet (SBUV-2) radiometer; the Limb Infrared Monitor of the Stratosphere (LIMS); and the Stratospheric Aerosol and Gas Experiment (SAGE). Sounders are especially suited to obtaining temperature, water vapor, ozone, and other trace-gas data that we can plot in altitude profiles. Here is an example (somewhat degraded) of temperature (red) and water vapor (blue) profiles over Denver, Colorado, on January 13, 1997, as calculated from the GOES-8 Sounder (see page 14-7 for details); the ordinate replaces altitude with pressure (in mbars).



The sounder on GOES-8 uses different channels to sense atmospheric properties. This composite panel show images made by various channels:



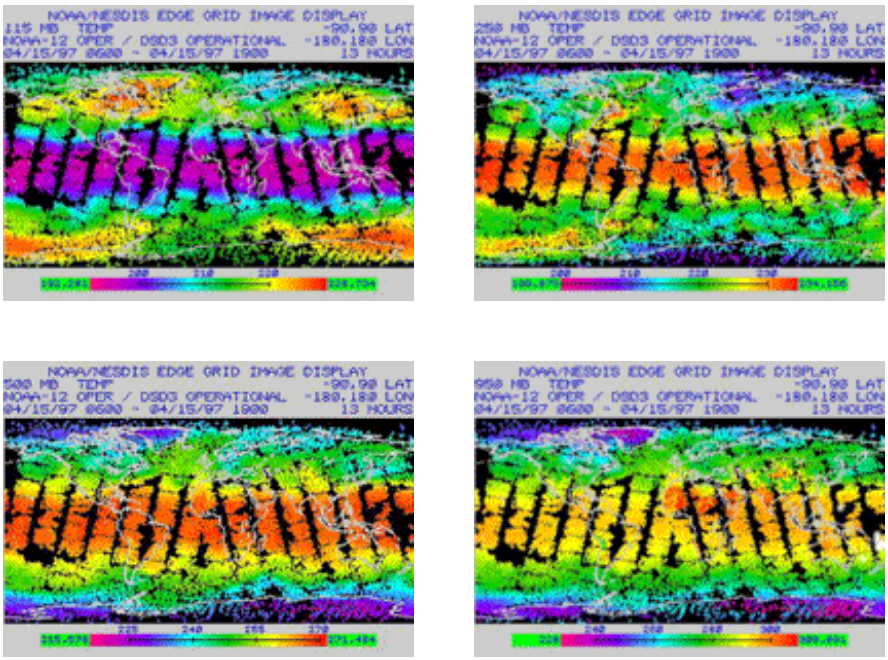
One way in which a sounder can sense levels in the atmosphere is illustrated in the diagram below. Each temperature map was made at a specific wavelength that is capable of measuring the radiant temperature in the atmosphere at some particular altitude (or depth within the air column, which can also be expressed in terms of atmospheric pressures in millibars). The map progression from upper left to lower right follows a trend of decreasing height above the surface. Note especially in the lower pair that the lateral temperature variation is from cooler in the northwest to warmest in the southeast.



Soundings can also be made in the microwave region. At different frequencies (or wavelengths, especially in the millimeter range), the return signals give blackbody-equivalent temperatures at different depths.

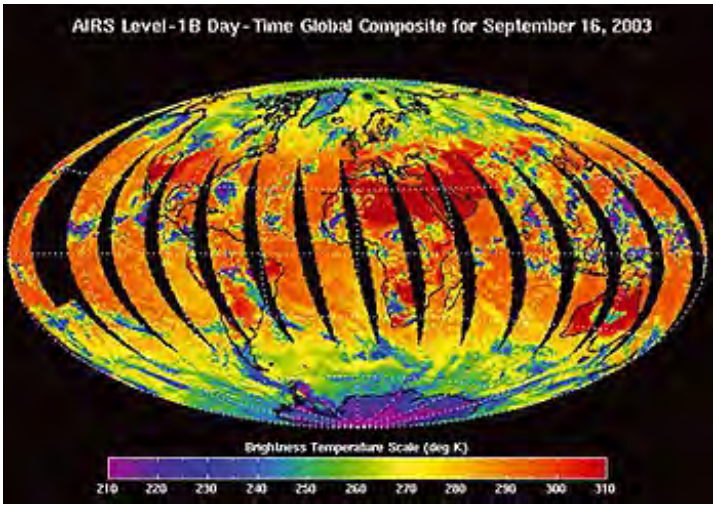
However, because satellites sweep over much of the Earth, particularly those in near-polar orbits (see below), and thus, take data readings continuously as they orbit, we can present the sounding data as global maps. As an example, consider these maps of global temperatures at four different altitudes (specified as pressures in millibars), as obtained by the TIROS Operational Vertical Sounder (TOVS) on Nimbus 12 on April 15, 1997 (115 mbars is

approximately at 15 km and 950 mbars is near the surface):

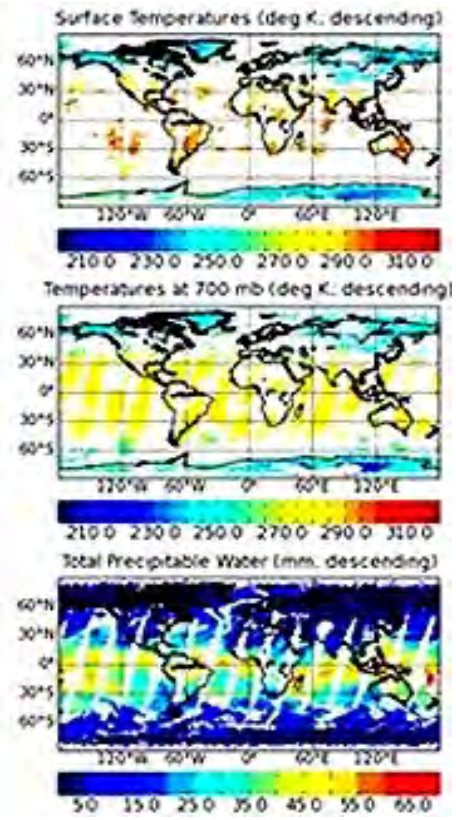


14-6: **For a location on the equator, from the above 4 image panels, estimate the temperature profile on this date (April 15, 1997). ANSWER**

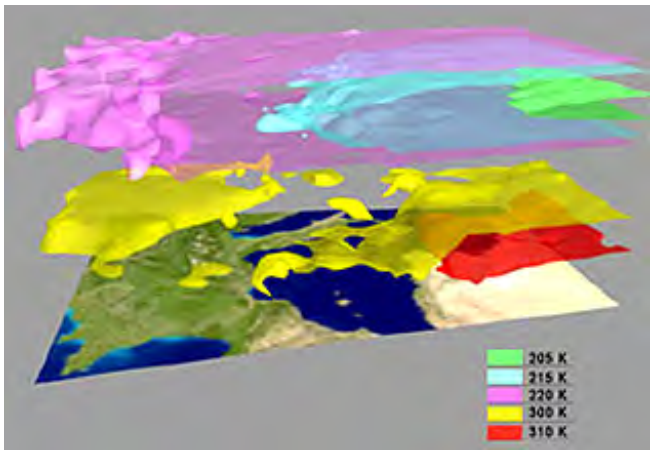
We display a number of similar maps made from atmospheric sounder instruments on Metsats elsewhere in this section. But, for now let us point out that the EOS Aqua satellite has three of the most sophisticated sounders yet flown in space (see Section 16). Of these, we mention here AIRS - the Atmospheric InfraRed Sounder, which is described on this JPL AIRS site. AIRS consists of a hyperspectral sensor that has 2378 individual channels sensing in the thermal spectrum. It can produce both temperature and moisture maps at various atmospheric levels and profiles into the atmosphere - all on a daily basis. (In practice, profiles are not produced routinely, only as needed. Here is a typical temperature map produced by AIRS:



Every day AIRS produces temperature maps for the surface and the 700 mm level and for total precipitable water. Here is a typical set for August 24, 2006:



Researchers working with AIRS data have prepared some novel illustrations. Here's one that shows temperatures above Europe in a quasi-3D mode:



With this digression into the topic of sounders as new background, we will now proceed to review the history of Metsats beginning with TIROS and Nimbus.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

The historical background leading up to the launching of fleets of meteorological satellites is briefly touched upon. The first U.S. group, the TIROS series, is described. Greater discussion of the Nimbus program, including a table listing the various instruments used in the series still flying, expounds upon its accomplishments. Instruments singled out for illustrative examples are: HRIR; IDCS; SCMR; ESMR; SMMR; TOMS; and SBUV; other sensors are mentioned. NOAA satellites are mentioned here if they supported any of these sensors.

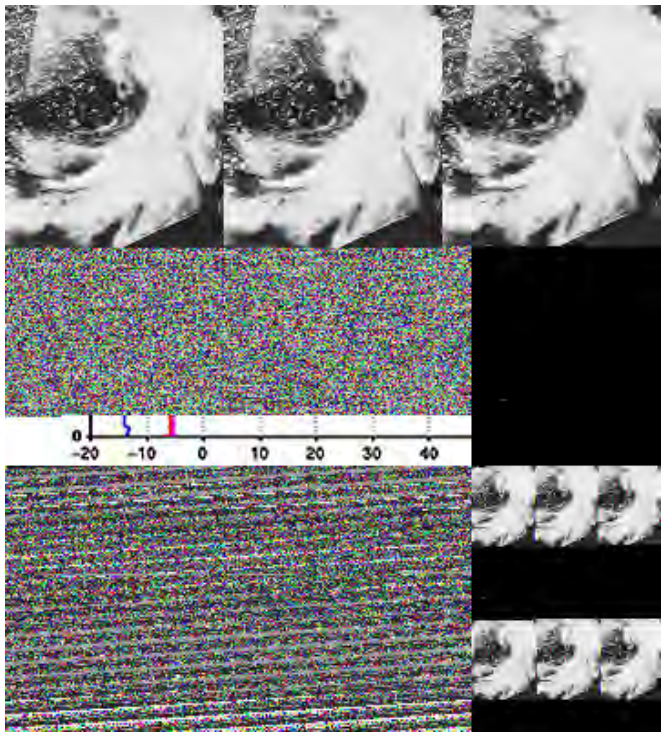
The United States Metsat program traces its formal inception to the launch of Vanguard 2 in February of 1959, just 12 months after our first ever successful satellite, Explorer 1. The spacecraft failed because of a wobbly attitude. An improper orbit on the next attempt, with Explorer 6, led to a few low resolution pictures of Earth. NASA achieved success with Explorer 7, orbited on October 13, 1959, which carried a radiometer designed by Verner Suomi, America's pioneer in satellite meteorology, and his colleagues. Using black and white hemispheres on the sensor to calibrate, this instrument measured solar and terrestrial radiative energy (determining reflective and absorbed components) to estimate radiation balance.

TIROS and Nimbus

The first experimental meteorology series began with TIROS-1, launched on April 1, 1960. TIROS stands for Television Infrared Observational Satellite. The main instrument was a vidicon, which is a modified television camera that scanned through 500 lines, each containing 500 pixels. Although desinged to function primarily to learn what that kind of satellite can contribute to meteorology, TIROS (and Nimbus) both provided useful insights as well into land observations for earth resources applications.

Because of its fame as a first, this satellite (a second copy) has been enshrined at the National Air and Space Museum in Washington, DC:

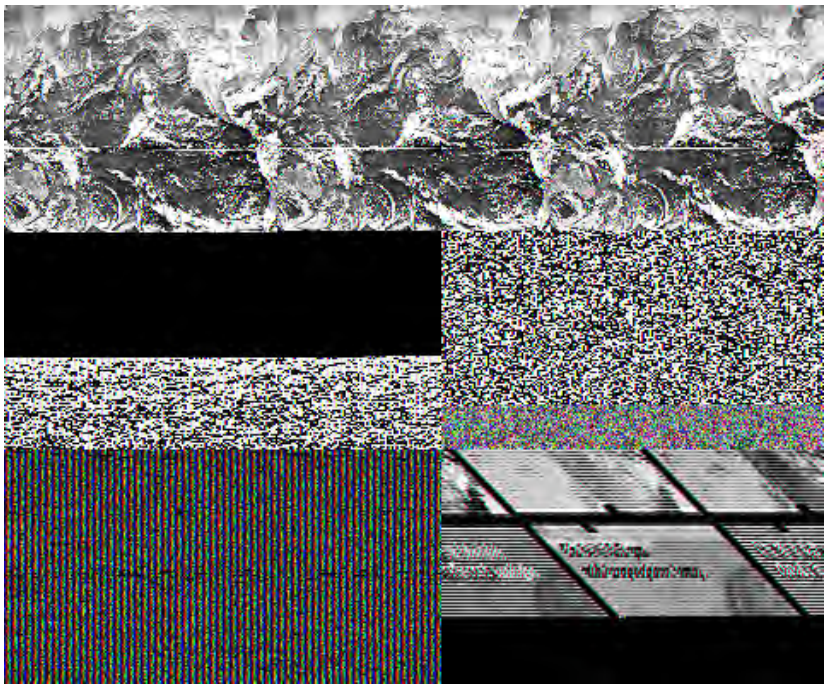
We show here one of the nearly 23,000 images returned from TIROS-1:



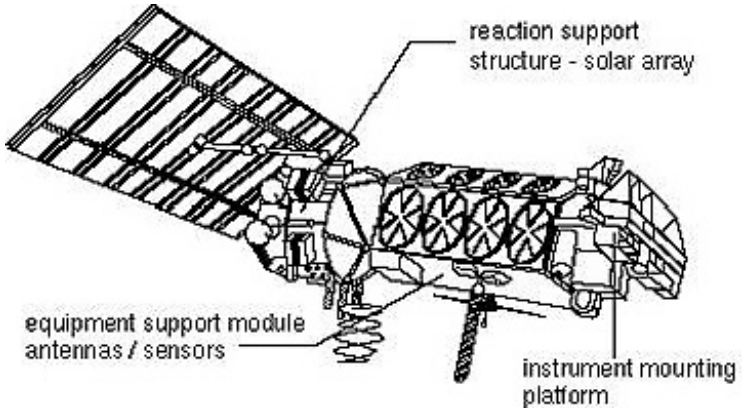
14-8: **Why is there a blocky disjointedness associated with this image?** **ANSWER**

TIROS images are mosaicked to provide global panoramas of cloud cover:

TIROS-2 included the Medium Resolution Infrared Radiometer (MRIR). The TIROS-8 vidicon had an 800-line resolution and was the first to use the Automatic Picture Transmission (APT) technique. The TIROS series (10 in all) was non-polar following orbital inclinations between 48° and 58° . The first eight TIROS maintained orientation by spinning at 12 revolutions/minute, which limited the time during which the camera pointed at Earth. TIROS-9 followed a "cartwheel" spin pattern (spin axis perpendicular to the orbital plane), which facilitated coverage to allow strips of images that users could mosaick to provide global composites:

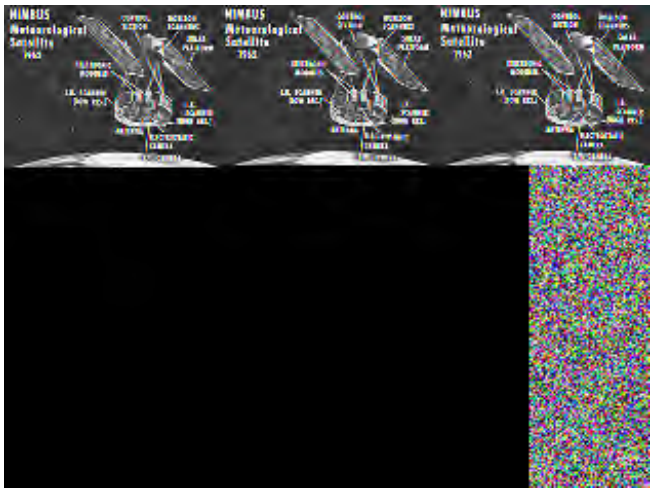


TIROS-N, the last in the series, was much different from TIROS-1, as seen here:

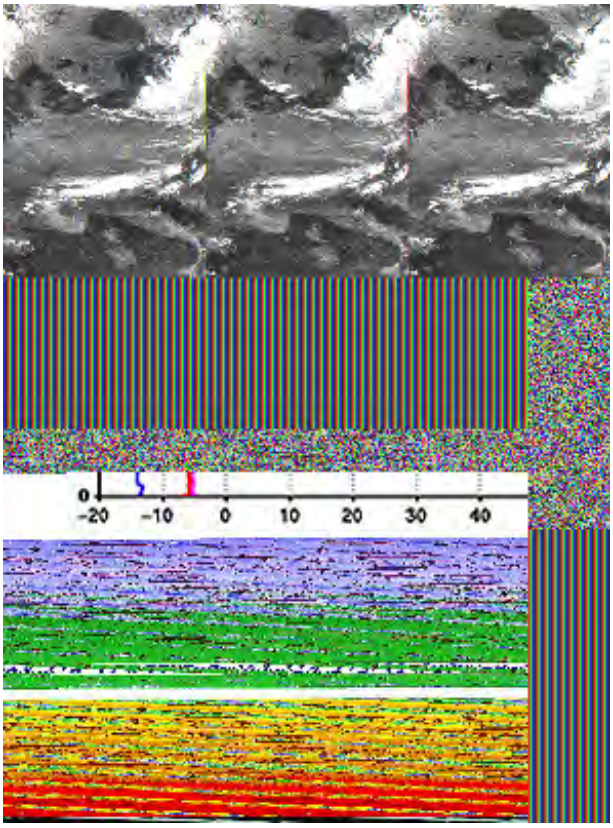


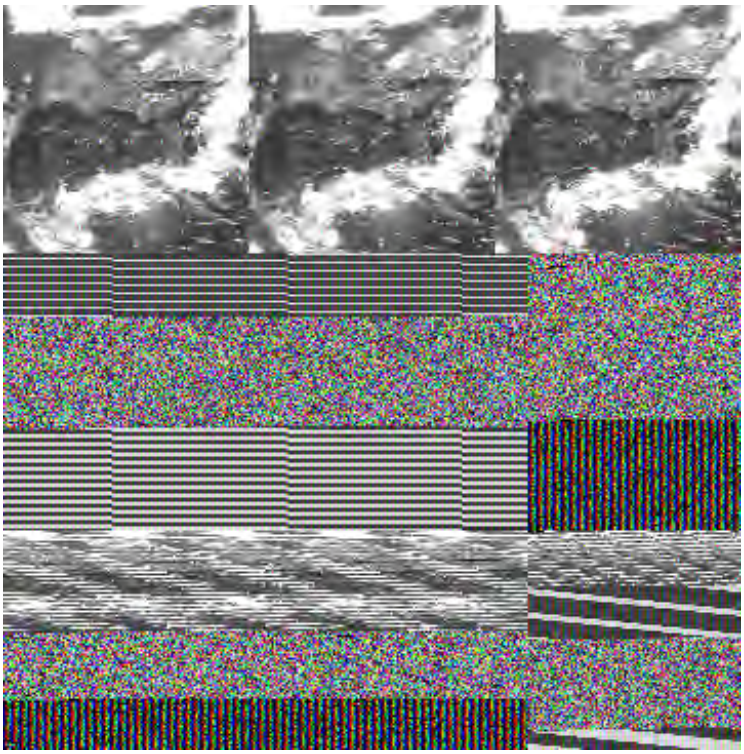
The next fleet of Metsats, the Nimbus series (total of seven), flown from 1964 to 1978, was dedicated largely to research and development experiments. We name the instruments used in this series in this table:

Nimbus-1, launched on August 28, 1964, was the first to be put in a sun-synchronous orbit (now the norm for polar satellites). It used a three-axis stabilization technique (based on flywheels) that kept it pointed constantly at Earth. Nimbus 1, shown below, looks remarkably like ERTS-1 (Landsat). It should - Landsat used the same spacecraft but of course had different sensors.

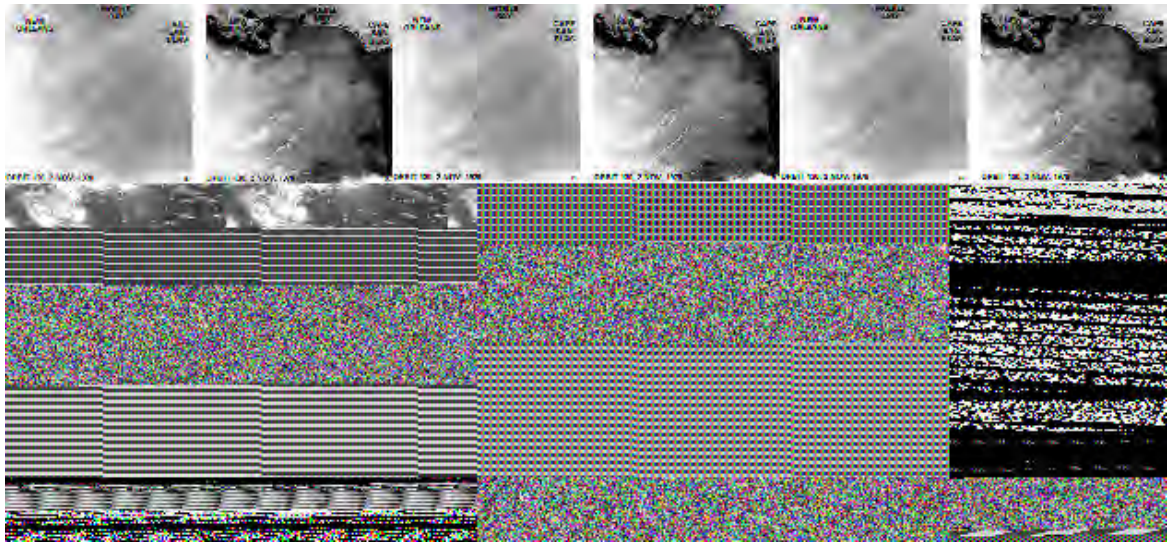


Nimbus carried several instruments including the Advanced Vidicon Camera System (AVCS, an APT) and the High Resolution Infrared Radiometer (HRIR), which operated in the 3.6-4.2 µm interval. An example of a Nimbus 1 HRIR image, taken at night over western Europe, appears next, on the top (note the distortion that enlarges Germany and Sweden relative to southern countries - the Italians may be aggrieved by the shrinking of their "boot"!). On the bottom is a visible Image Dissector Camera System (IDCS) image of the southeast U.S., as seen by Nimbus 3:



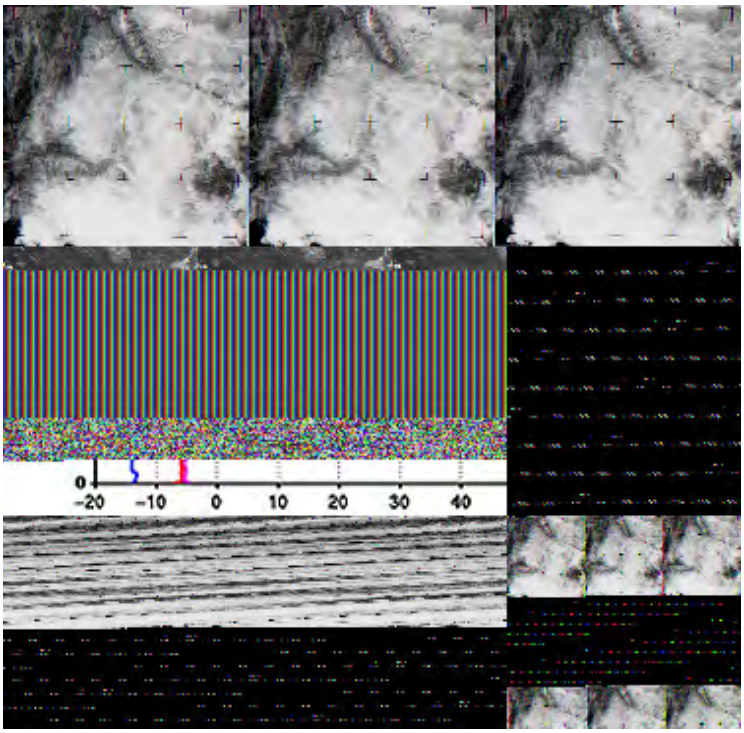


To compare the information obtained in the visible (left) and the thermal infrared (right), look at this image pair of the Gulf of Mexico acquired simultaneously by Nimbus 4.

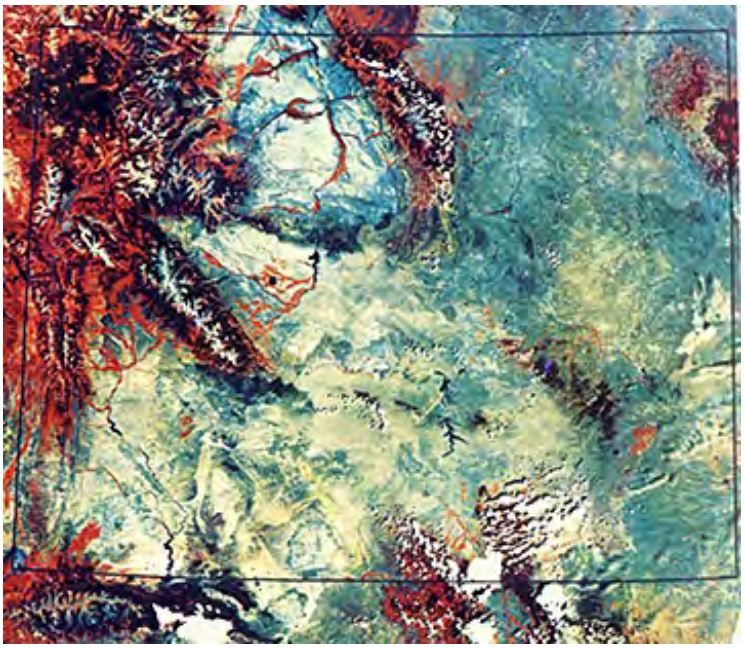


14-9: **The thermal image shows extended groups of swirly white patterns. Are these clouds or thermal patterns in the Gulf of Mexico seawater? Cite a logical clue to your conclusion.** **ANSWER**

In the next image we get a feel for what one can see at a resolution of 1.1 km (0.7 mi), attained by the IDCS visible channel on Nimbus 3. Under optimal viewing conditions (no clouds), parts of southwestern Wyoming and a bit of Utah appear here.



The basins generally are highly reflective, and the mountains tend to stand apart because of dark tones associated with evergreen vegetation. Compare this image with this next Landsat 1 MSS mosaic of the entire State of Wyoming that was made by General Electric for me (NMS) as part of my Wyoming project in which I was Co-Investigator with the Geology Department at the University of Wyoming:

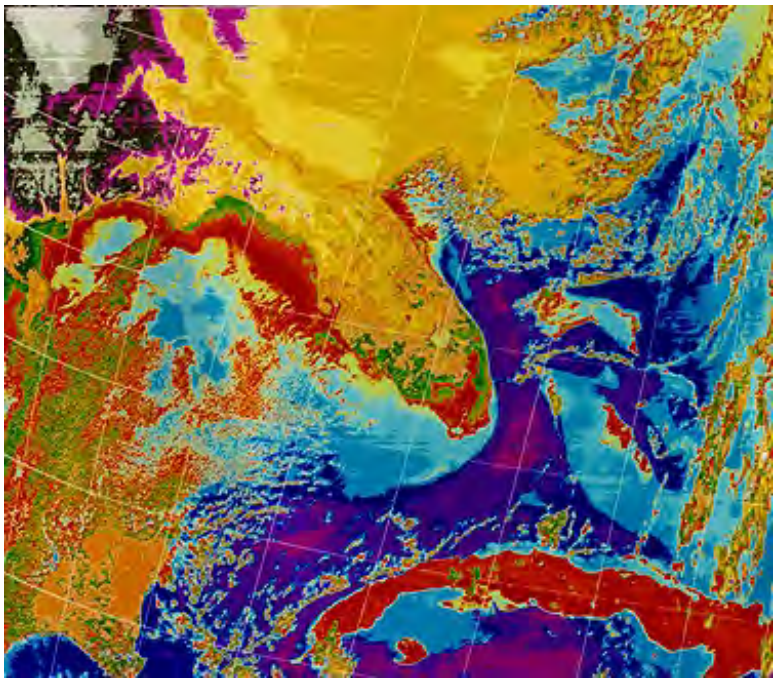


14-10: **The above Nimbus image was the first remote sensing product ever worked on by the writer (NMS) when I transferred in 1970 from the planetary to the remote sensing programs at NASA Goddard Space Flight Center. Just for fun, why don't you fit the Nimbus image into the Landsat mosaic - and check the answer. ANSWER**

Nimbus 3 (launched April 14, 1969) also was the first to use atmospheric sounders extensively, along with its Infrared Interferometer Spectrometer (IRIS), operating between 6 µm, and its 25 µm and Satellite Infrared Spectrometer (SIRS), sensing in the 15 µm region. Nimbus 4 (April 8, 1970) carried the Backscatter Ultraviolet (BUV) radiometer, becoming the first Metsat to measure atmospheric ozone.

Two instruments on Nimbus 5 (December 11, 1972) are of special significance. The Surface

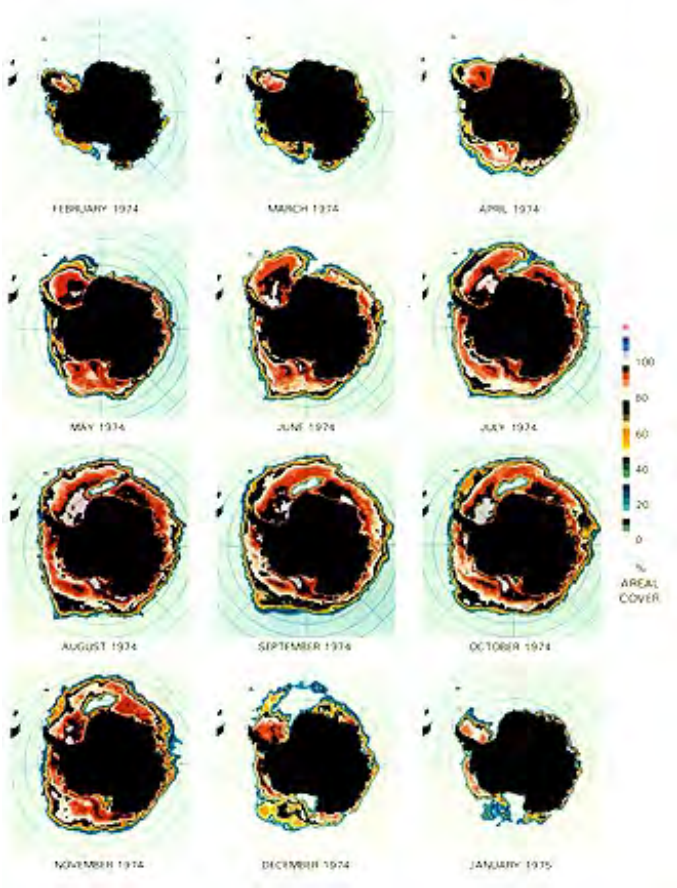
Composition Mapping Radiometer (SCMR) uses two thermal bands, 8.4-9.5 µm and 10.2-11.4 µm, to produce color-coded temperature maps, such as this image of Florida and Cuba and surrounding waters, made from the 8.8 µm channel:



Ratios of radiant temperatures measured by the two bands provide a qualitative estimate of SiO_2 content of rocks and soils. This process uses the concept of "restrahlen", a German term that refers to decreased emissivity because of resonance vibrations associated with silicon-oxygen bonds in silica tetrahedra. As the silica content increases, the emissivity decreases more and also shifts as wavelengths become longer.

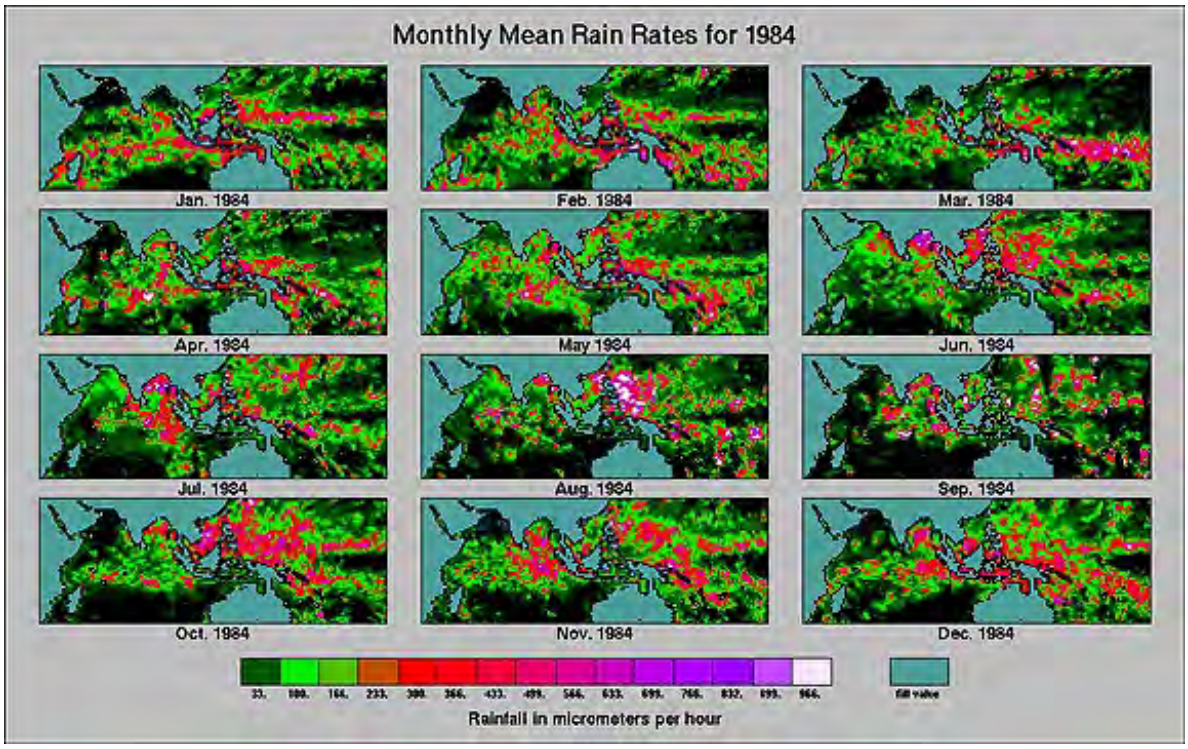
14-11: **How might this silica shift be used practically?** **ANSWER**

The Electronically Scanning Microwave Radiometer (ESMR) on Nimbus 5 operated at a 19.35 GHz frequency (1.55 cm wavelength) to sense brightness temperatures of the surface and atmosphere. This instrument was capable of sensing surface ice temperatures, especially in the polar regions, as shown in this time series of maps that plot the percentage of ice cover around Antarctica on a monthly basis in 1974. The ESMR was also the first to use microwave absorption to estimate precipitation (rain rates) by quantifying increases in optical depth, which correlate to higher brightness temperatures. The ESMR on Nimbus 6 (June 12, 1975) was set at 37 GHz (0.81 cm).

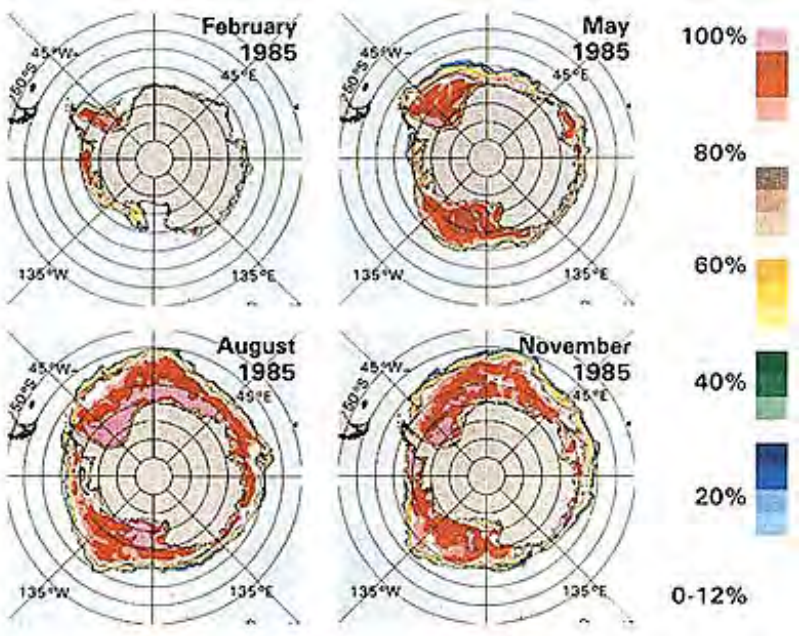


Nimbus 7 carried eight highly complex sensors which were all improved versions of sensors previously flown on Nimbus satellites. They were a Limb Radiance Inversion Radiometer, a High Resolution Infrared Radiation Sounder, an Earth Radiation Budget experiment, a Scanning Multichannel Microwave Radiometer (SMMR), a Pressure Modulated Radiometer, a Solar Backscatter UV/Total Ozone Mapping Spectrophotometer, a Temperature, Humidity Infrared Radiometer and a Tropical Wind, Energy Conversion and Reference Level experiment.

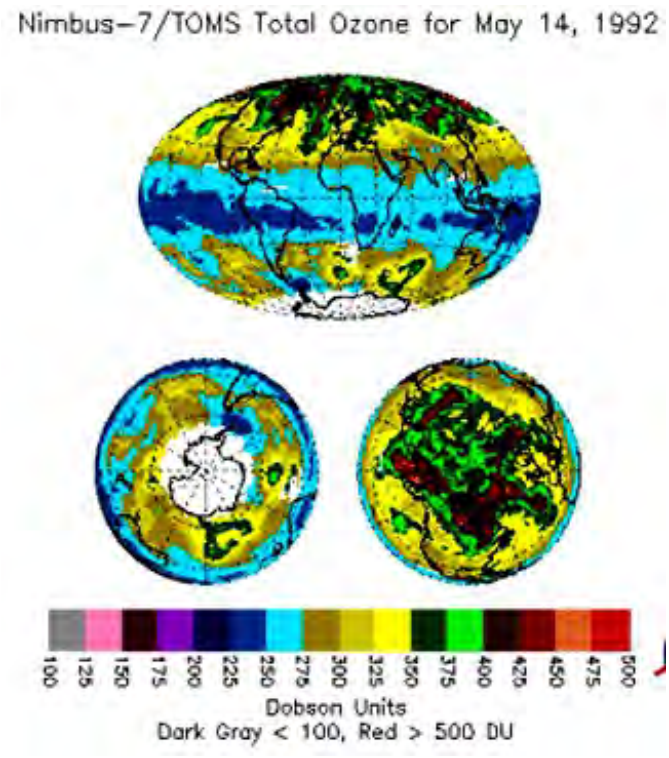
The Scanning Multichannel Microwave Radiometer (SMMR), flown on Nimbus 7 (launched October 24, 1978), and again on Seasat, consists of a 5-channel (0.81, 1.36, 1.66, 2.80, 4.54 cm), dual-polarization instrument that provides data of value to many applications: ocean circulation, low altitude winds, water vapor, cloud liquid water content, sea ice type, extent, and concentration, snow cover, moisture distribution, and rainfall rates. Data on tropical rainfall for the months of 1986 over the Indian Ocean and South Asia is one such product:



The SMMR monitored ice in the Antarctic shelf during 1985, as displayed for four months here:



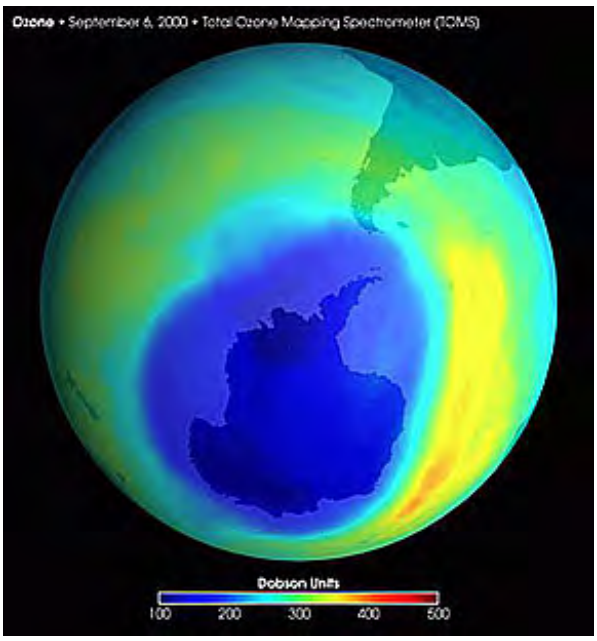
The first Total Ozone Mapping Spectrometer (TOMS) on Nimbus 7, which is still operating, measures UV reflectivities at 0.312, 0.317, 0.331, 0.340, and 0.380 µm. It calculates ozone quantities from the ratio of the returns in the 0.312/0.331 µm wavelengths. As an example of a global ozone map, here is the plot of data obtained on May 14,1992.



(Note: A Dobson unit is the response at 4 wavelengths by a Dobson spectrometer from which the total ozone from ground to the outer atmosphere can be measured and then recalculated as a compressed column of ozone equivalent to a 0.01 mm thick, measured over a fixed area centered on Labrador in eastern Canada and adjusted to a STP of 1 atmosphere and O° C.)

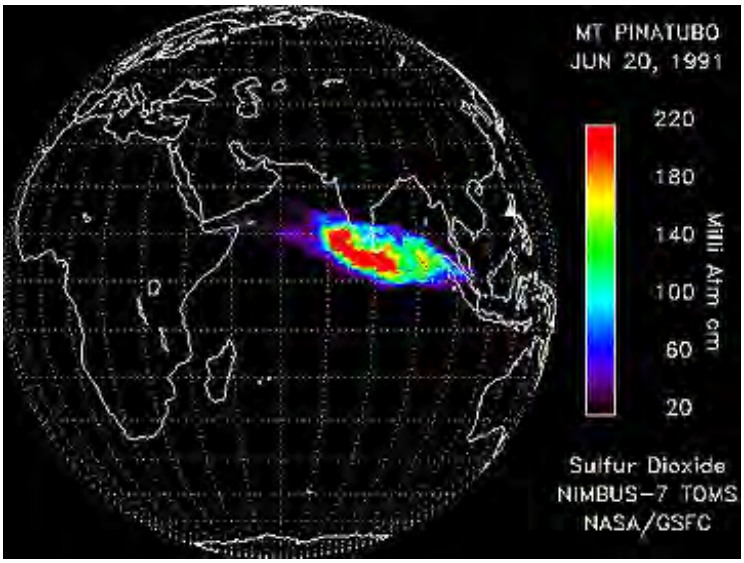
In the above maps, the ozone content is higher in high latitudes in springtime. The lower values closer to the equator are of concern because high ozone content affords greater protection from harmful UV radiation.

The South Polar ozone hole has grown wider since 1992. Here is a TOMS image of much of the southern hemsphere, centered on Antarctica, which in September, 2000 disclosed the largest expanse of high ozone levels yet recorded; the hole since has shrunk a bit.

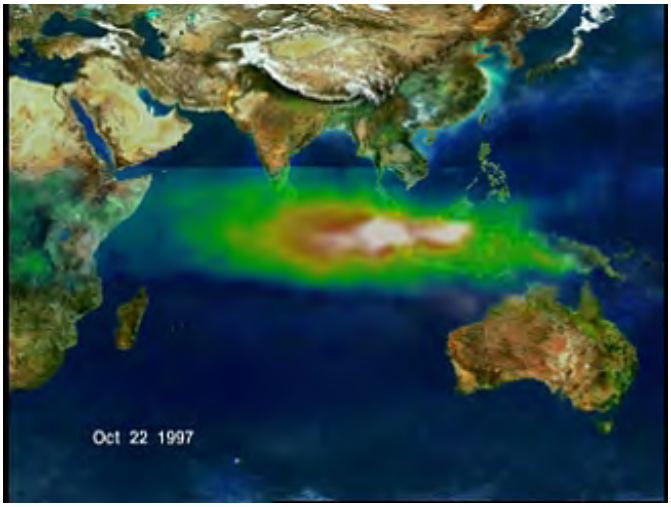


We can also use the TOMS to monitor SO_2 in the atmosphere. After some major volcanic eruptions, it tracked extensive clouds of SO_2-enriched ash and gases injected into the upper atmosphere daily across much of the world, until they dissipated below detection levels. Here is the status on June 20, 1991 of the cloud produced by the Mt. Pinatubo eruption in the

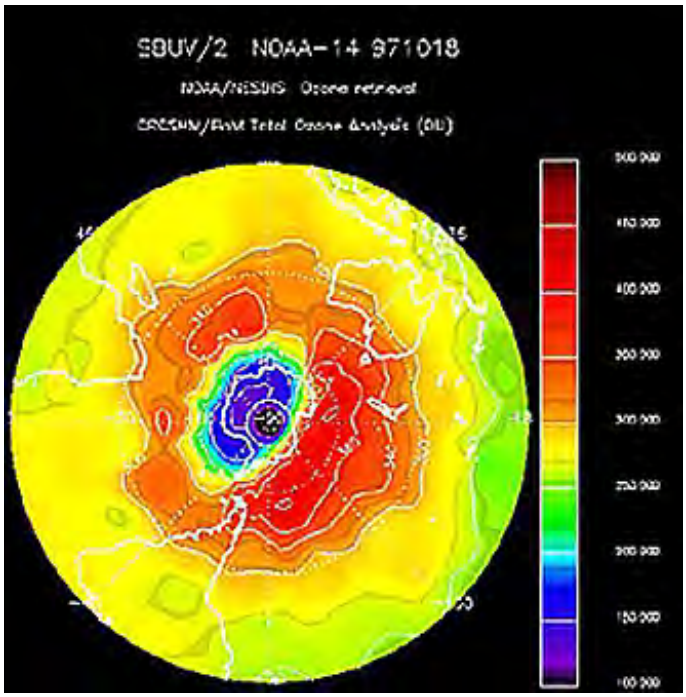
Philippines.

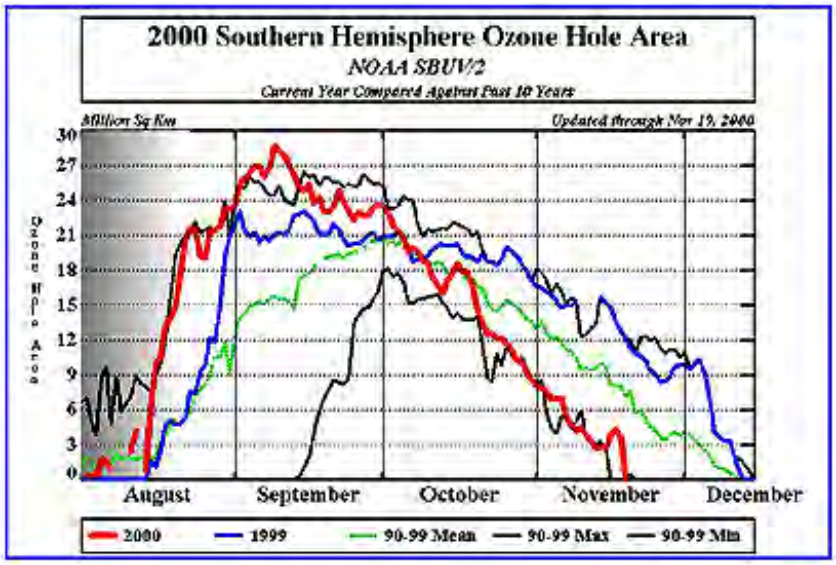


TOMS has proved very useful in monitoring other kinds of atmospheric disturbances such as massive smog buildups that persist over time. Forest fires in Indonesia and elsewhere from September through November of 1999 led to a huge elongate trail of gases that were carried by winds westward to Africa. This TOMS image should the smog and associated clouds near its maximum extent:



The first Solar Backscatter UltraViolet (SBUV) sensor was on Nimbus 7 (also SBUV/2 on NOAA-9, -11, and -14), sharing some of the components of the TOMS. The SBUV had 12 channels in the UV region providing coverage between 160 and 400 nm. Shown below is a SBUV/2 map of ozone distribution (in Dobson Units) for the South Polar (Austral) Spring. Then, beneath that are graphical data sets for areas of coverage in that region for the months of September through December of the years 2000 and 1999, and three curves summarizing grouped data for 1990-99:





Both the TIROS and Nimbus programs were initiated by NASA. Later ones in the two series included NOAA as a major participant and eventual operator. There are other U.S. and foreign Metsat series: For a nearly complete listing of polar orbiting meteorological satellites (the POES group), consult this Colorado State website.

PREVIOUS NEXT

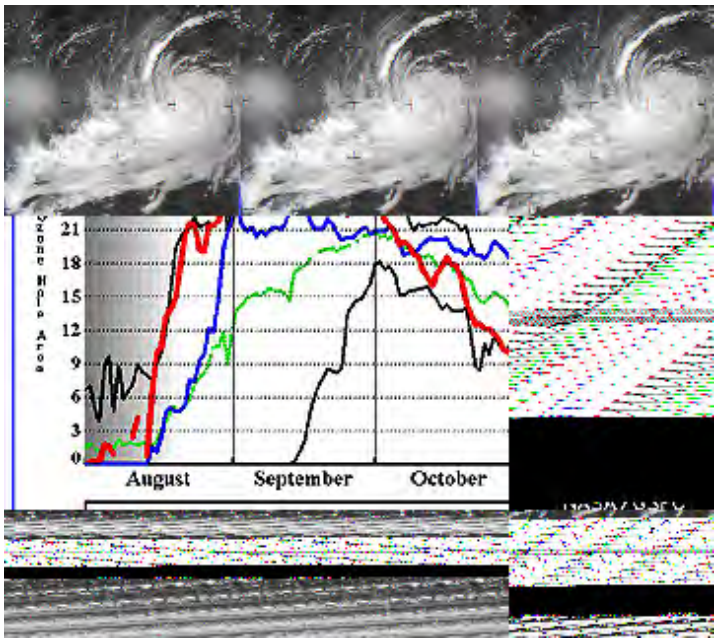
Primary Author: Nicholas M. Short, Sr.

Both military and civilian meteorological satellites have been operating in space for almost four decades. Some are specialized, with particular mission goals or objectives to measure parameters not routinely determined by the main stable of Metsats. ESSA is a series of 9 satellites making environmental observations, including weather phenomena. The military DMSP series looks at Earth both in the Vis-NIR and the thermal regions. It is tailored to provide data that support military operations, with direct data readout available to personnel worldwide. One of its sensors, the SSM/I, senses microwave radiation relevant to both cloud and rainfall conditions. TRMM is a joint U.S./Japan satellite capable of global coverage of rainfall, particularly in the Tropics.

ESSA, DMSP, SSM/I; TRMM

The first U.S. Metsats dedicated to routine operations (rather than, in part, experimental) were the ESSA 1-9 series flown for the Environmental Science Service Administration (formed from the U.S. Coast & Geodetic Survey and the National Weather Service as the predecessor to NOAA which itself was chartered by Congress in 1970 as the umbrella organization that included still other agencies). The ESSA satellite group that began on February 3, 1966, and ended on February 26, 1969 evolved into the current NOAA series. The AVCSs and APTs were the prime sensors.

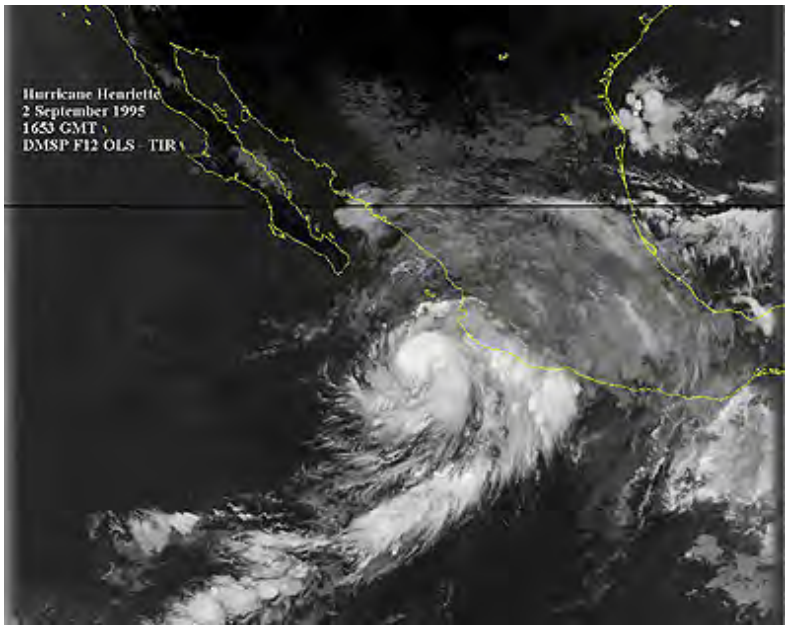
Hurricanes and cyclones were special targets that ESSA routinely monitored from their early to final stages of development and movement. Here is a view taken on September 1, 1966, of Hurricane Faith as it moved towards Cape Hatteras on the North Carolina coast.



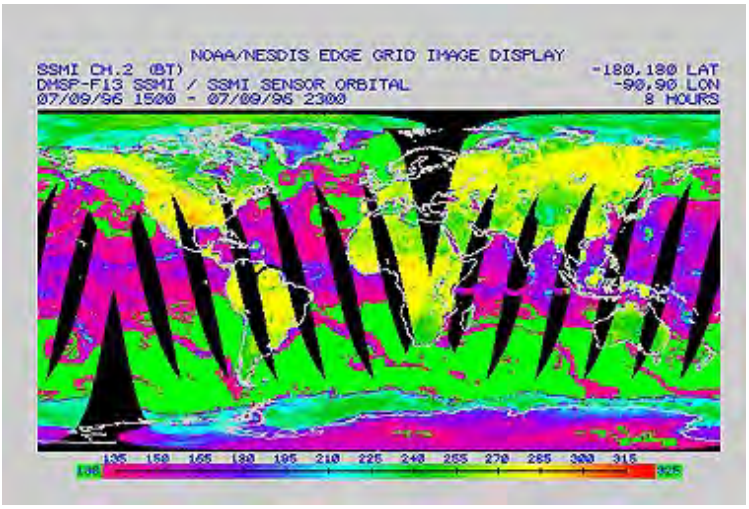
14-12: **What's missing in this image?** **ANSWER**

The U.S. Air Force program in meteorology, essential to its military operations, began with the launch of the first Defense Meteorological Satellite Program (DMSP) satellite on September 16, 1966. The principal instrument was the Operational Linescan System (OLS), which imaged in the visible and thermal IR (TIR) regions. This program, involving many launches in several series, continues today, and much of its data are declassified. Below is a TIR image of Hurricane

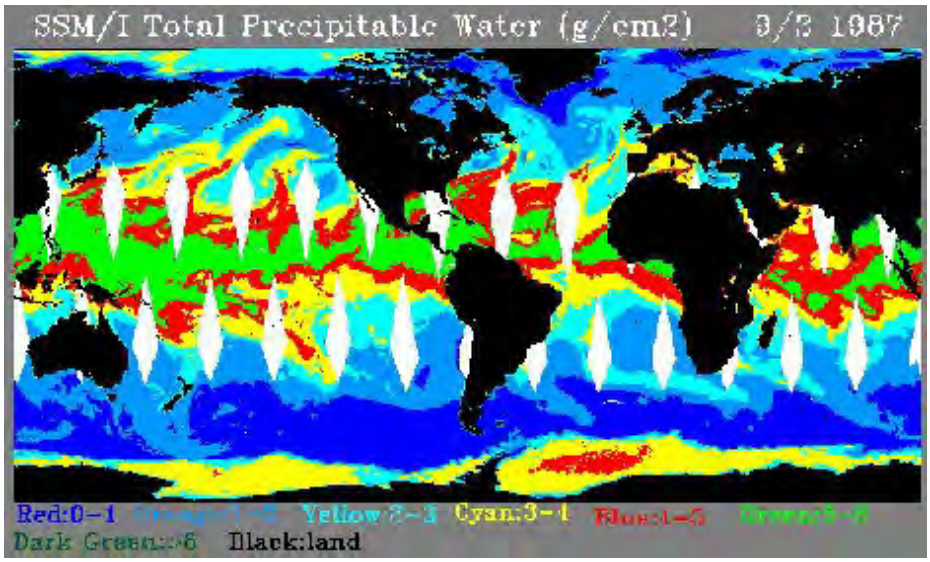
Henrietta, spawned in the East Pacific, which struck the west coast of Mexico on September 5, 1995.



Later DMSPs also carried the Special Sensor Microwave Imager (SSM/I), which measures vertically- and horizontally-polarized radiation at 19.36, 22.23, and 85.5 GHz and horizontally at 37.0 GHz. Here is a global brightness temperature map, derived from Channel 2 data for July 9, 1996:

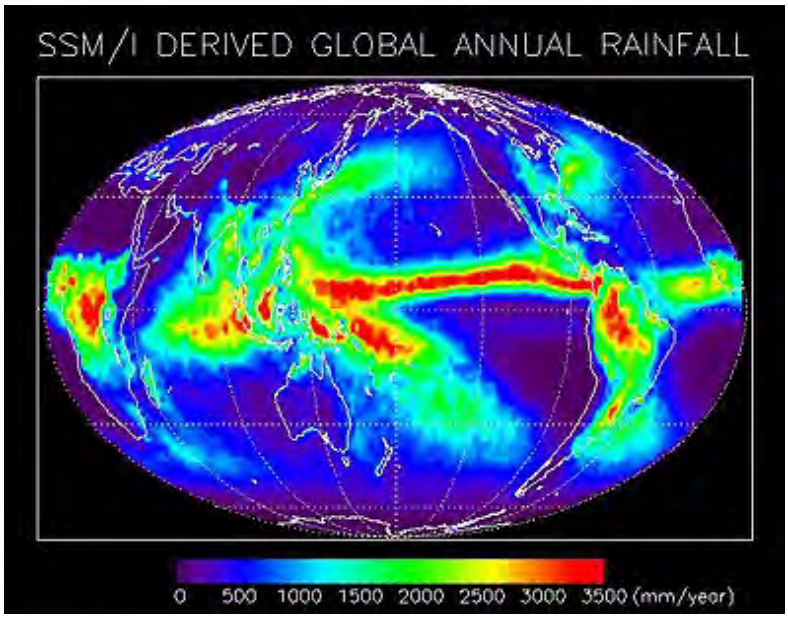


This sensor, a kin to SMMR, detects rainfall and clouds, such as displayed in this world map that shows Total Precipitable Water (in g/cc) as determined from SSM/I data, for September 3, 1987:



14-13: **Why is the green in the above map where it is? ANSWER**

Over the course of a full year, we can average SSM/I readings to provide a global map that summarizes annual rainfall (in millimeters), as presented in this example:



The first satellite whose primary mission is to measure precipitation is the Tropical Rainfall Mapping Mission (TRMM), a joint research project between the U.S. (NASA) and Japan (National Space Development Agency: NASDA). A basic objective of TRMM is to obtain estimates of the vertical profile of the latent heat (heat resulting from a change of state), released through condensation of water vapor in the atmosphere, especially in the Equatorial Intertropical Convergence Zone (ITCZ). The TRMM is a large spacecraft that a joint engineering team built at NASA's Goddard Space Flight Center and successfully launched from Japan on November 27, 1997. The TRMM Observatory has three primary instruments:

1. The Precipitation Radar (PR), built in Japan, measures 3-D rainfall distribution. The PR is an electronically scanning radar that operates at 13.8 GHz using horizontal polarization with a 129-slotted waveguide antenna. It has a horizontal resolution of 4.3 km (2.7 mi) at nadir and a scanning swath width of 220 km (137 mi).

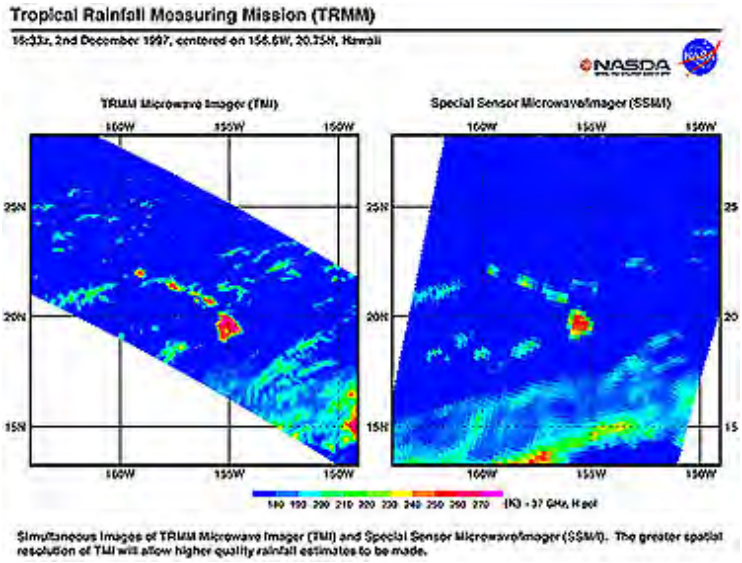
2. The Multi-Channel Microwave Radiometer (TMI), collects data on the integrated-column, precipitation content, its areal distribution, and its intensity. The TMI operates at five frequencies, ranging from 10.65 GHz (45 km [28 mi] spatial resolution) to 85.5 GHz (5 km [3.1 mi] resolution). Dual polarization at four of the frequencies provides nine channels. At a 65° scan angle, the swath width is 760 km (472mi).

3. The Visible Infrared Scanner (VIRS) provides high resolution (2.1 km [1.3 mi] at nadir) information on cloud cover, cloud type, and cloud top temperatures; being a radiometer it operates at 0.63, 1.6, 3.75, 10.8, and 12.0 mm wavelengths.

In addition, the spacecraft's load includes the Lightning Imaging Sensor (LIS) (at 0.777 mm wavelength) capable of picking out lightning flashes associated with active rainfall and the Clouds and Earth's Radiant Energy System (CERES), a broad-band, scanning radiometer, designed to measure reflected and emitted radiative energy at the surface and for the atmosphere and its constituents.

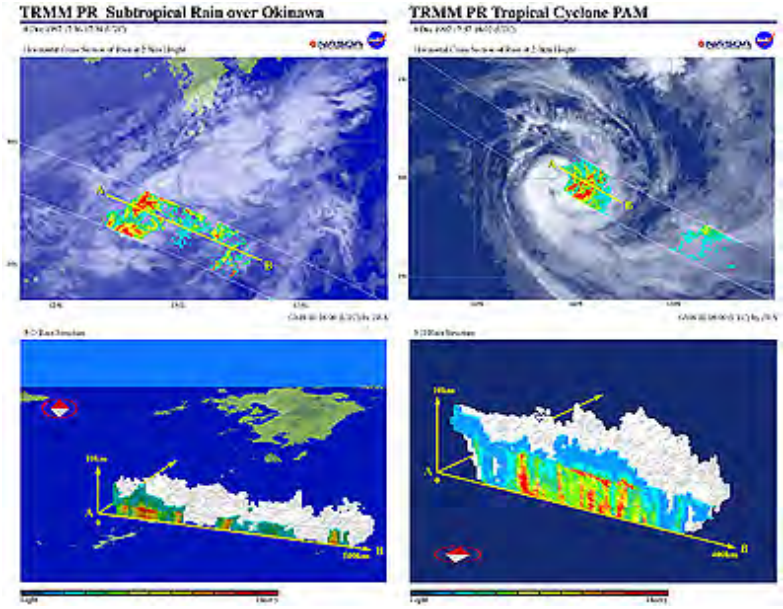
A goal of this low-altitude (350 km [217 mi]), non-sun-synchronous (precessing) satellite is to provide monthly precipitation data over 500 x 500 km grids, especially for the tropical-ocean regions. A network of ground stations to produce corroborating ground truth is under development. You can find more information and updates on this spacecraft and its products at this NASA Goddard TRMM Home Page.

By December, 1997, the TRMM investigators released the first images on the Internet. We reproduce representative ones here, and you can access others via the above URL. The first compares images over the Hawaiian Islands, obtained simultaneously by TRMM's TMI sensor (left) and the SSM/I on the current DMSP satellite. The improved resolution of the TMI is evident. The islands were largely cloud-covered with localized rainfall (reds).

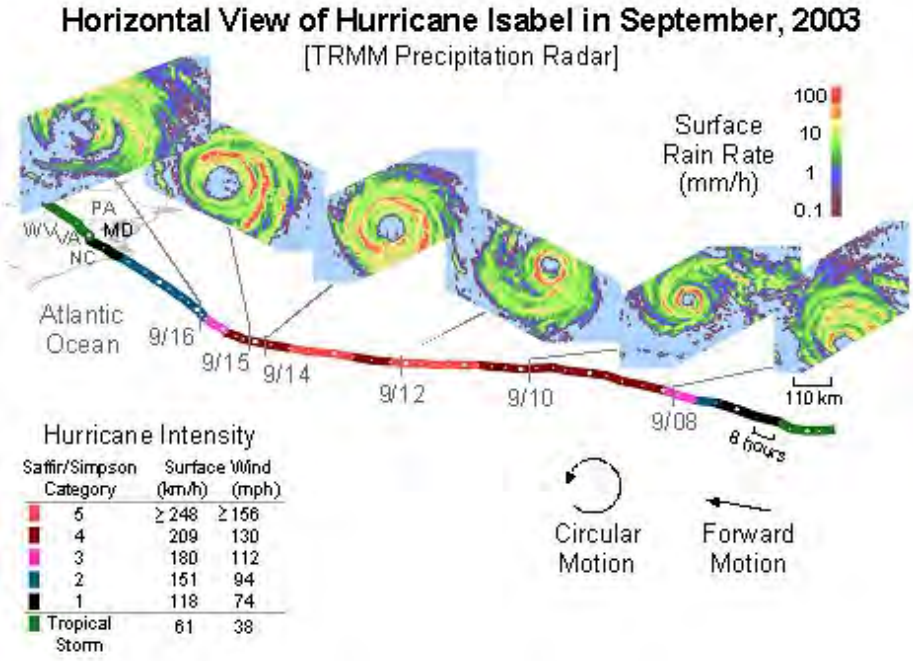


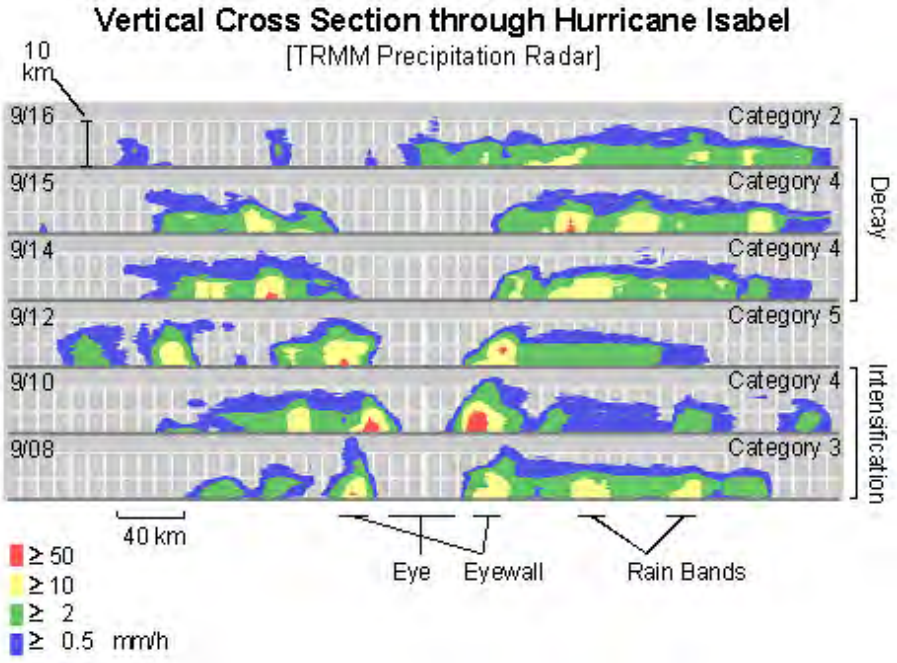
Simultaneous images of TRMM Microwave Imager (TMI) and Special Sensor Microwave/Imager (SSM/I). The greater spatial resolution of TMI will allow higher quality rainfall estimates to be made.

The second images show Typhoon Pam in the western Pacific, south of Okinawa in early December. The imaged cloud tops appear gray, and the colors in the flat projection represent relative rainfall intensities. The side view (cross-section) is an extrapolation of data to indicate rain distribution within the cloud mass.

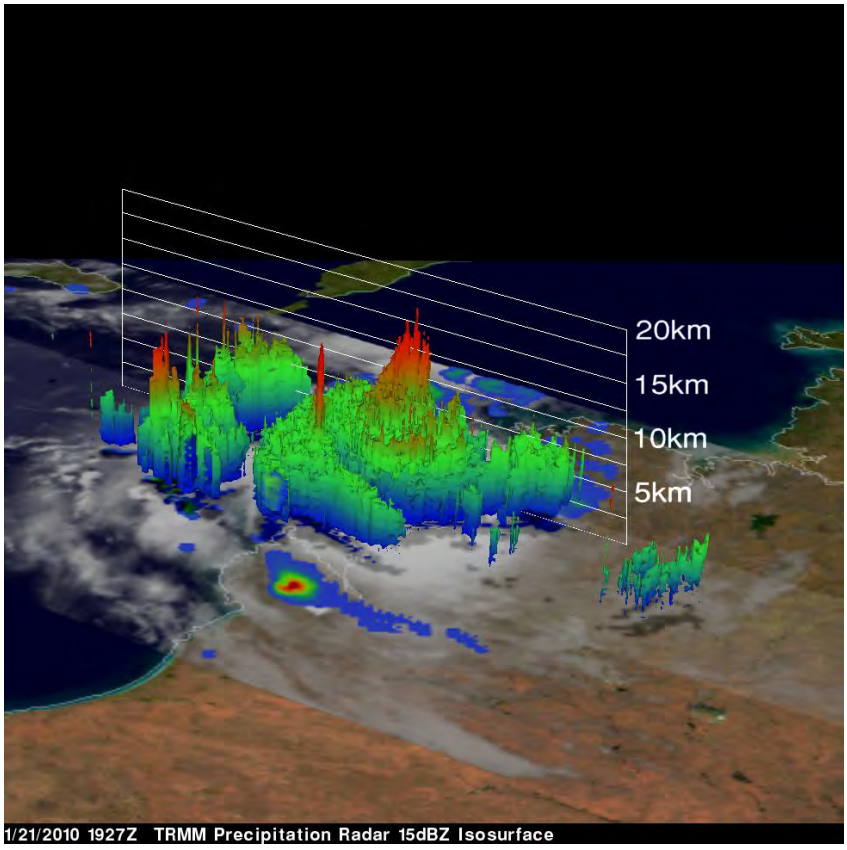


TRMM has proved very adept at monitoring the rainfall coming from a hurricane as it is tracked along its path. That is evident in these two views of Hurricane Isabel as it neared the coastline of the eastern U.S. in September, 2003.

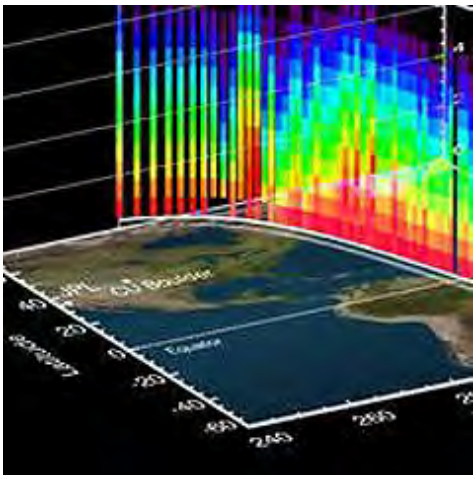




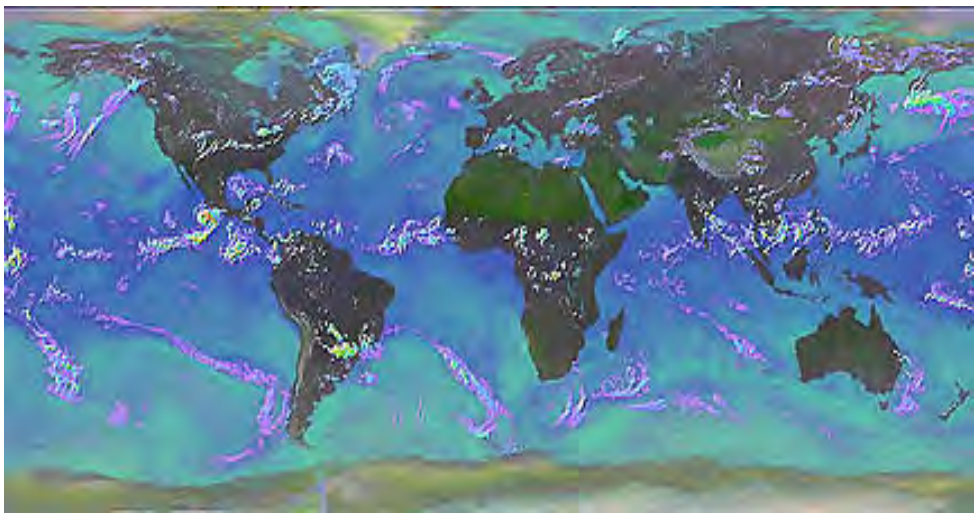
TRMM's radar can produce a three-dimensional image of a hurricane's structure, including an indication of cloud heights. Here is a diagram of hurricane Magda off western Australia in January, 2010:



TRMM can conduct both on-going applied observations that help to monitor rainfall conditions and more scientific studies that are revealing some of the properties of moist atmosphere. Two forms of water exist: Light water (hydrogen has no neutrons) and heavy water (hydrogen has one neutron). Being lighter, the first water type should preferentially migrate further out (up) into the atmosphere. This is observed in this bar graph profile of water in the atmosphere, as measured by TRMM's TES (Thermal Emission Spectrometer). In the rendition, red is heavier isotopic water; blue is lighter. Note the fact that over large water bodies, the heavier water extends to higher elevations in the atmosphere.

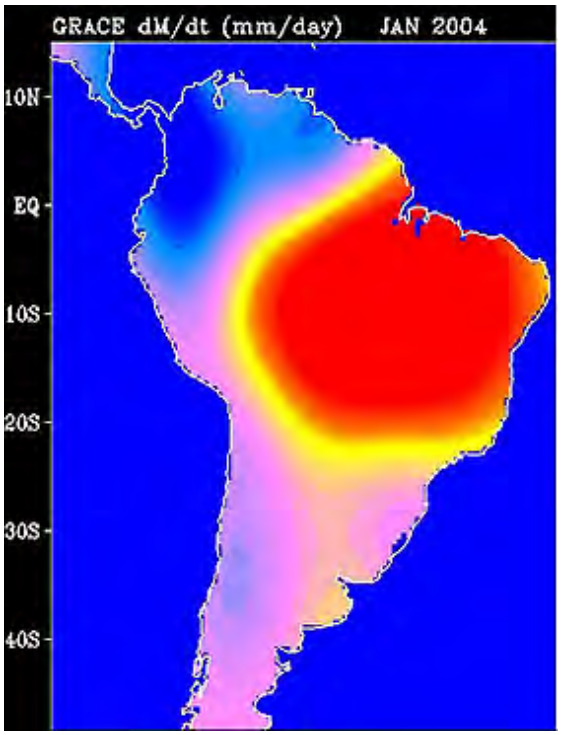


The next image is a composite from several-days coverage showing worldwide rainfall during part of December. Major cloud masses are in white, areas of moderate or intermittent rainfall in purple, and heavier rains in reds, yellows, and greens.

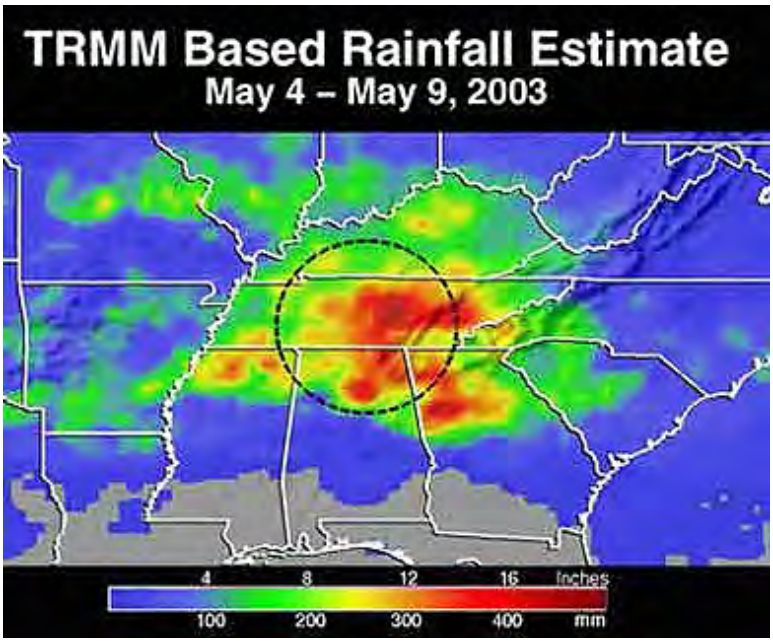


14-14: **In several parts of the world, there are long, linear patterns of purple rainfall. Suggest a possible explanation. Can you locate a small hurricane in this image?**
ANSWER

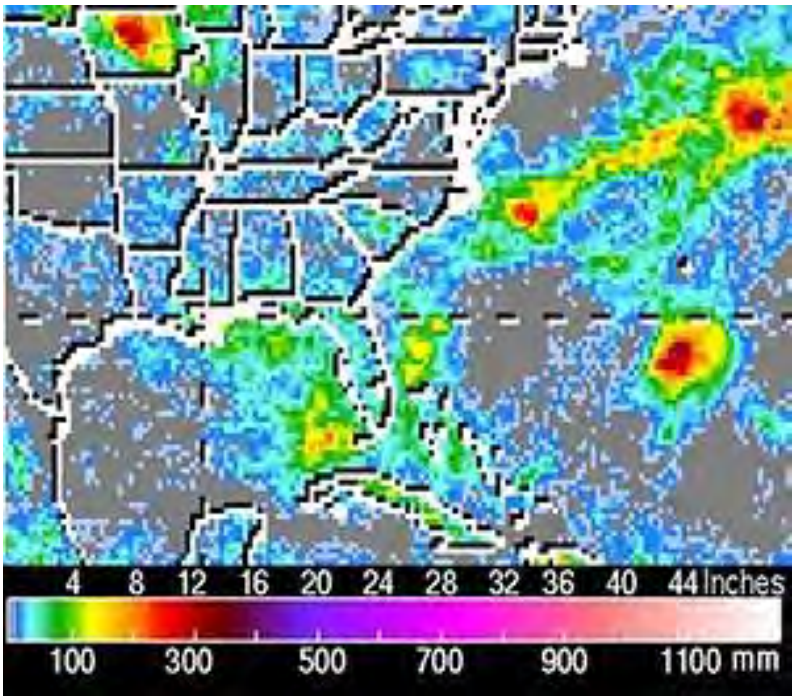
One of TRMM's prime accomplishments is its ability to estimate cumulative rainfall in selected areas over time spans averaged over about a week or longer. In a study of the entire continent of South America for a full year, average rainfalls during that period have been measured using TRMM, Quiksat (a scattering sensor), and GRACE (a geophysical gravity-measuring satellite). The balance between rainfall IN and stream flow OUT was also determined. Here is a smoothed map of rainfall distribution for the month of January, 2004:



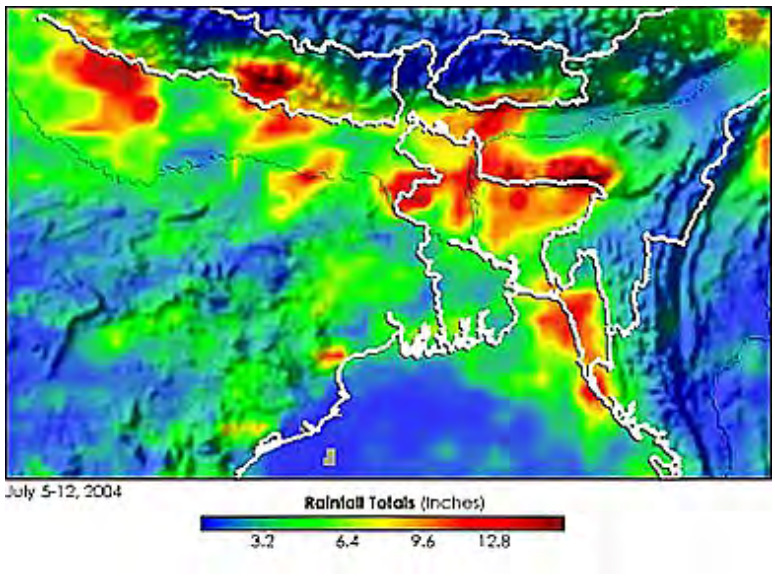
Because of its orbital migration, TRMM cannot revisit the same area in shorter (1-2 days) repeats and thus misses some rainfalls associated with thunderstorms, etc.) The Spring of 2003 was very wet in the U.S. Southeast, with a higher than normal number of tornadoes. In this map of rainfall, TRMM determined that up to 16 inches of rain fell in Tennessee and parts of adjacent Alabama and Georgia over a 6 day stretch in early May.



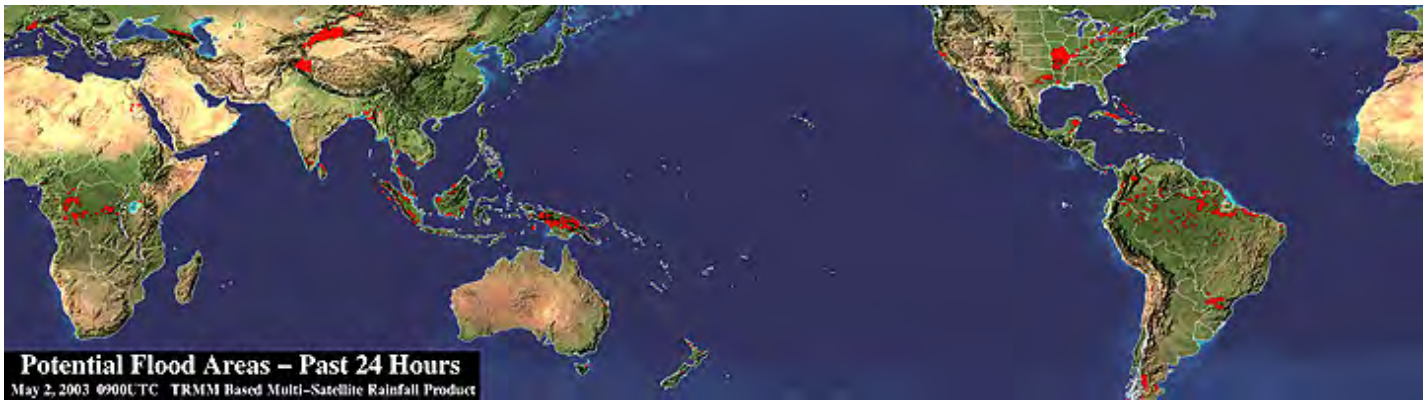
TRMM continues to determine rainfall amounts over various time interval through August of 2010. Here is a map of rainfall in the Gulf of Mexico over a 7-day stretch, important to those responsible for cleanup after the BP oil spill of 2010:



Thus TRMM is well-suited to making measurements that predict parts of the world that are, or will shortly be, experiencing widespread flooding. This composite image shows the magnitude of rainfall within a week in early July 2004 during an exceptionally heavy monsoonal deluge in eastern India, Bangladesh, and the Malayxian Peninsula:



TRMM now has the capability of producing maps that are effective in making flood forecasts. These maps single out (in red below) areas of land that receive 1.5 inches or more in a 24 hour period. They in turn can be used to predict possible flooding conditions at 3 and 7 days after the heavy rainfall. Here is a worldwide rainfall map (>1.5 inches in red) for May 2, 2003:

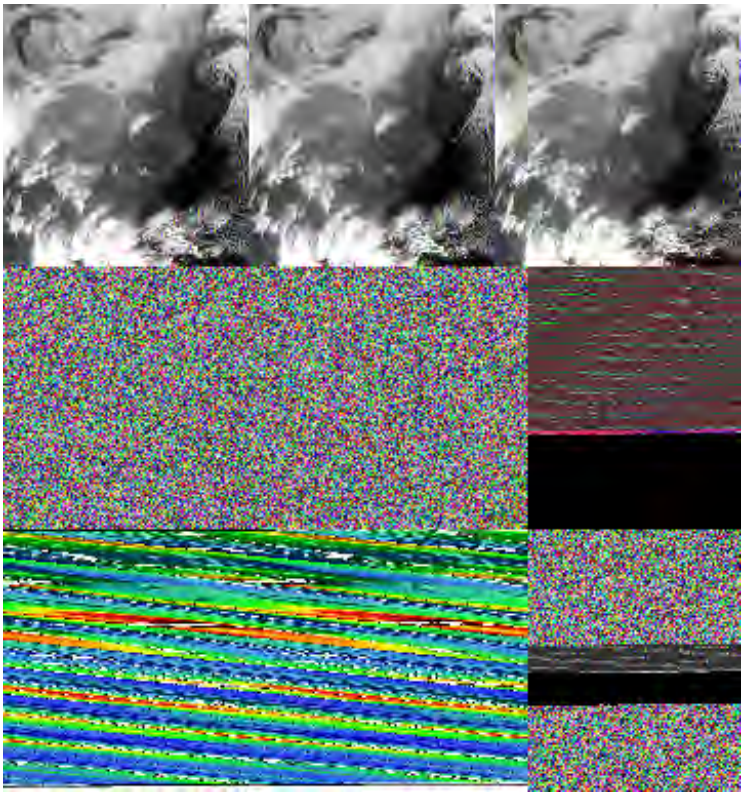


Primary Author: Nicholas M. Short, Sr.

The TIROS and Nimbus series of Metsats were initiated by NASA (mainly at the Goddard Space Flight Center). The National Oceanographic and Atmospheric Administration began developing and flying its own series, NOAA-6 through NOAA-15, as its mission made it the responsible agency for operational civilian meteorological satellites. Several TIROS satellites were dove-tailed with the NOAA program. Most of the NOAA series carry a version of the AVHRR sensor, usually with three bands in the Vis-NIR and two in the thermal region. Several hurricane images typify these NOAA products. European, Russian, and Chinese polar satellites are briefly reviewed.

The NOAA Series; Other Polar Metsats

The present-day series of operational Metsats can trace its inception to the launch of TIROS-M on January 23, 1970. Also called ITOS (for Improved TIROS Operational System), it included several vidicon cameras and an IR radiometer sensing between 3.4-4.2 µm. The first of the NOAA series (TIROS 2 to 5) followed this initial satellite. Here is a typical image, showing the eastern U.S., made by the NOAA-4 Very High Resolution Radiometer (VHRR), operating in the near-IR region.

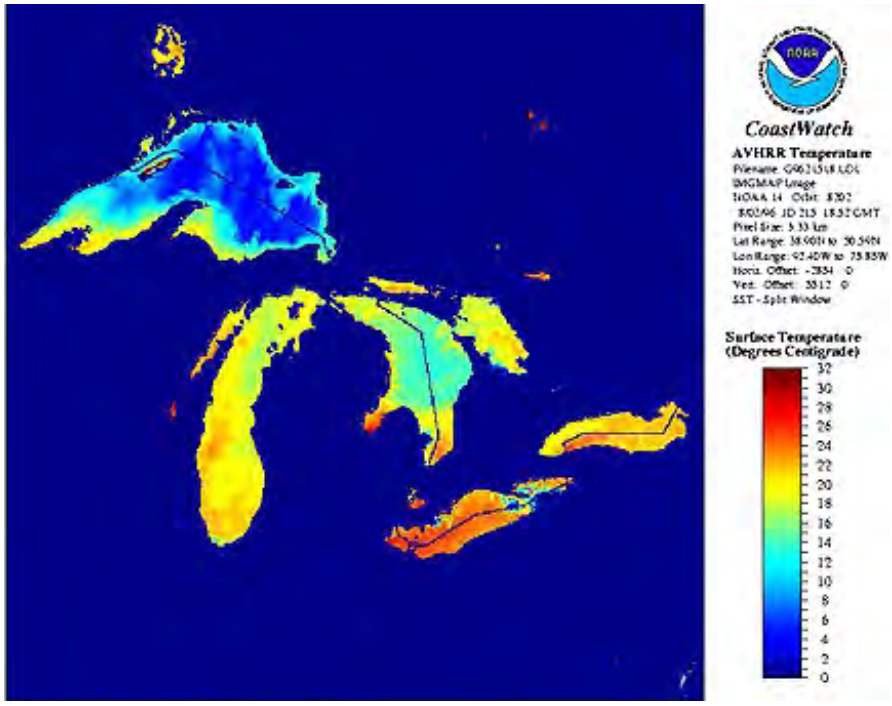


A new series in this generation began with the launch of TIROS-N on October 13, 1978. This satellite was the first in that series to carry the AVHRR (described on page 14-2), along with the first sounder, TIROS Operational Vertical Sounder (TOVS), designed to profile temperature and water vapor. The TOVS is actually a three instrument complex: the High Resolution IR Sounder (HIRS-2), with 20 channels; the Stratospheric Sounding Unit (SSU), with three channels near 15 µm, and the Microwave Sounding Unit (MSU), a passive scanning microwave spectrometer with

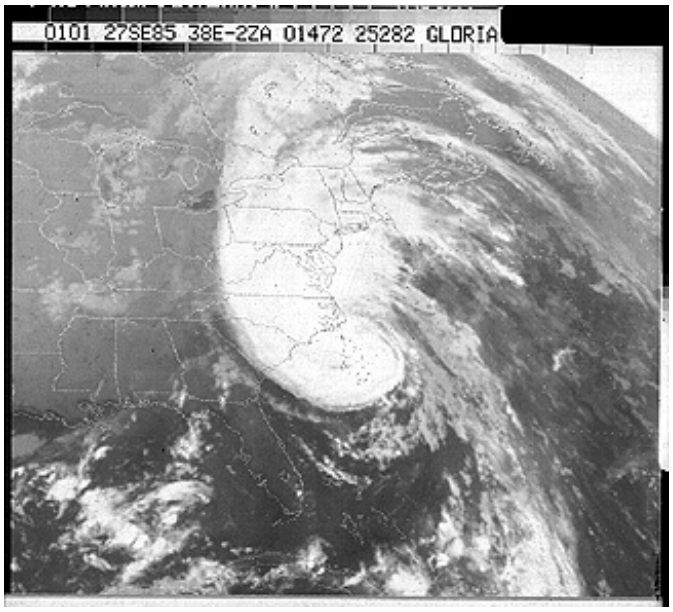
four channels in the 5.5 µm interval.

These instruments are the mainstays of the subsequent NOAA-6 (launched on June 27, 1979) through NOAA-15 (May 1998) Metsats. Even-numbered NOAA Metsats have North to South equatorial crossing times near 7:30 A.M. and have orbital repeat periods (re-occupy approximately the same paths) of four to five days. The odd-numbered ones cross the equator from North to South at night (2:30 A.M.) and have eight-nine day repeat periods. NOAA strives to keep as many as four satellites in this series operational at any one time.

The thermal bands on AVHRR can generate temperature "maps" of water bodies, both oceans and lakes. Here is a thermal image of the temperature in early Fall in the five U.S. Great Lakes:



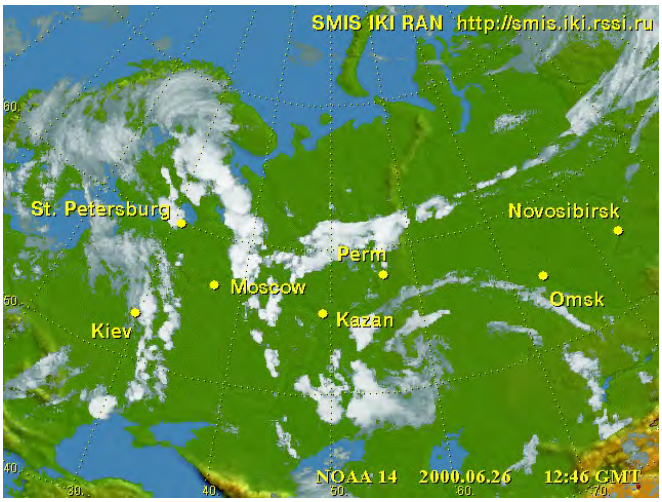
Here is a NOAA-9 AVHRR image in the visible region of Hurricane Gloria, as it neared the East Coast on September 27, 1997:



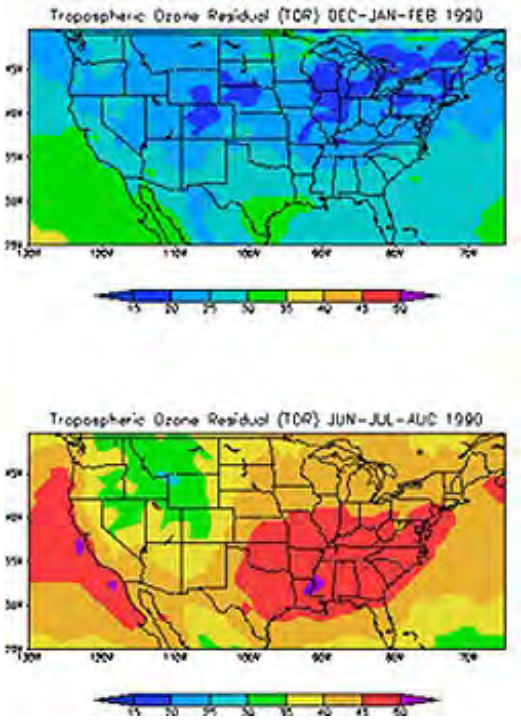
Like most multi-channel systems, we can combine different AVHRR-band images into color composites, as demonstrated by this NOAA-14 rendition of Hurricane Fran on September 4, 1996:

14-15: **This is a classic portrayal of a hurricane's appearance. Deduce from the directions of motion implied by the swirls in the hurricane's spiral pattern which way the circulation (clockwise or counterclockwise) of the winds is taking place.** **ANSWER**

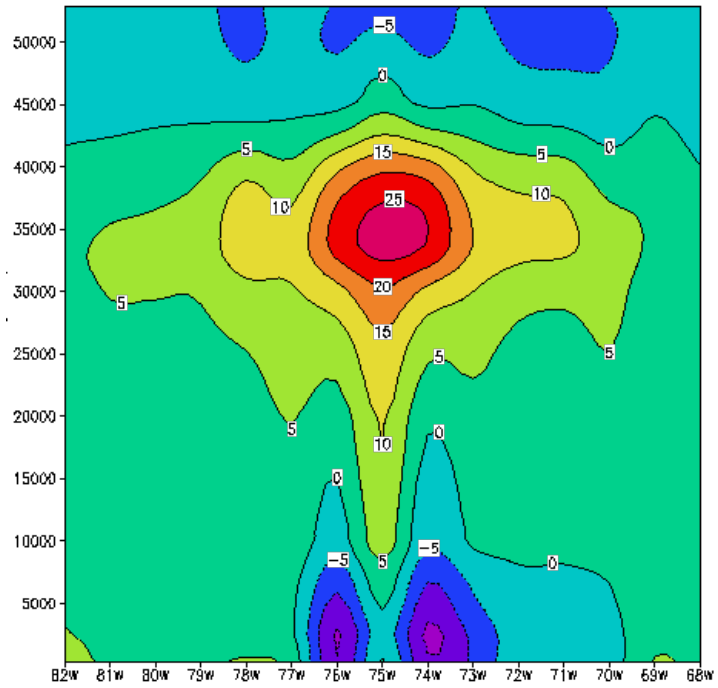
The NOAA daily images are compiled and displayed in different formats. Here is a NOAA-14 image taken on June 26, 2000 that covers European Russian eastward into Siberia.



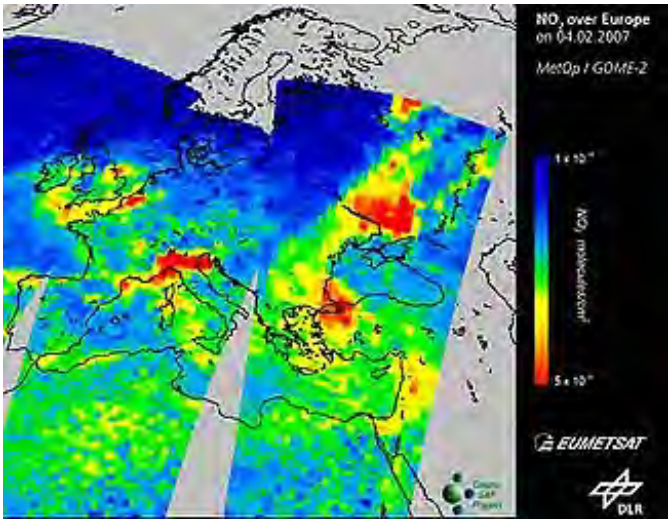
The NOAA series also began the important task of looking at the distribution of ozone in the troposphere. TOMS (described later) and the SBUV (Solar Backscatter UV) are the principal instruments. Here are measurements made for two periods in 1990.



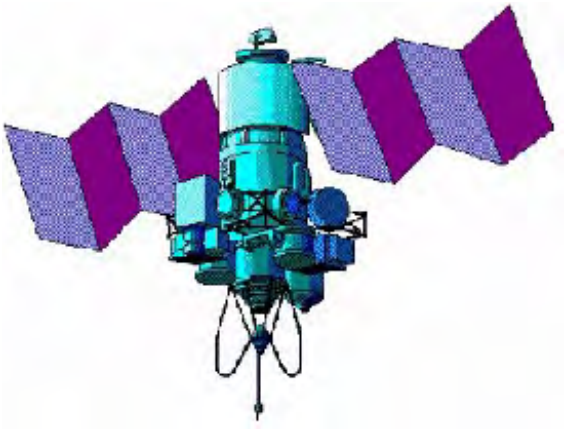
The latest in the NOAA group, NOAA-15, has these instruments: AVHRR/3; AMSU A & B (sounders); HIRS (an infrared sounder); SEN (Space Environmental Monitor); SAR (Search and Rescue); and DCS (Data Collection System. Its AVHRR now has a sixth band, at 1.6 µm. The AMSU-A is capable of producing temperature profiles, as exemplified by this cross-section through Hurricane Bonnie.



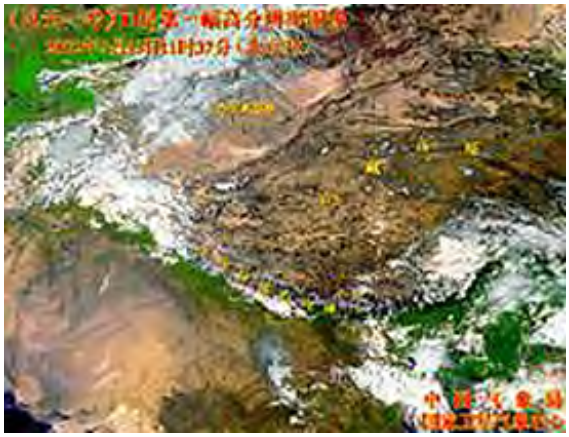
ESA has only recently started to operate polar Metsats. EuMetSat, a consortium of 18 countries, launched MetOp-A on October 19, 2006. It is a big satellite, as shown below, which mounts 13 sensors. As examples of its products, observe the Visible image of clouds and a plot of nitrogen dioxide over Europe:



The former Soviet Union/CIS has launched more than 25 polar-orbiting satellites in its three Meteor Series, starting in March of 1969, and continuing through 1994. Other Metsats were part of their Kosmos series



China, too, is now in the Metsat field. Here is its polar-orbiting Feng Yun-1 (translates as "Wind Cloud") satellite (FY-1C); beneath is one of its regional images of China.



Next, we examine another class of Metsats that examine near full Earth discs from geostationary orbits.

PREVIOUS NEXT

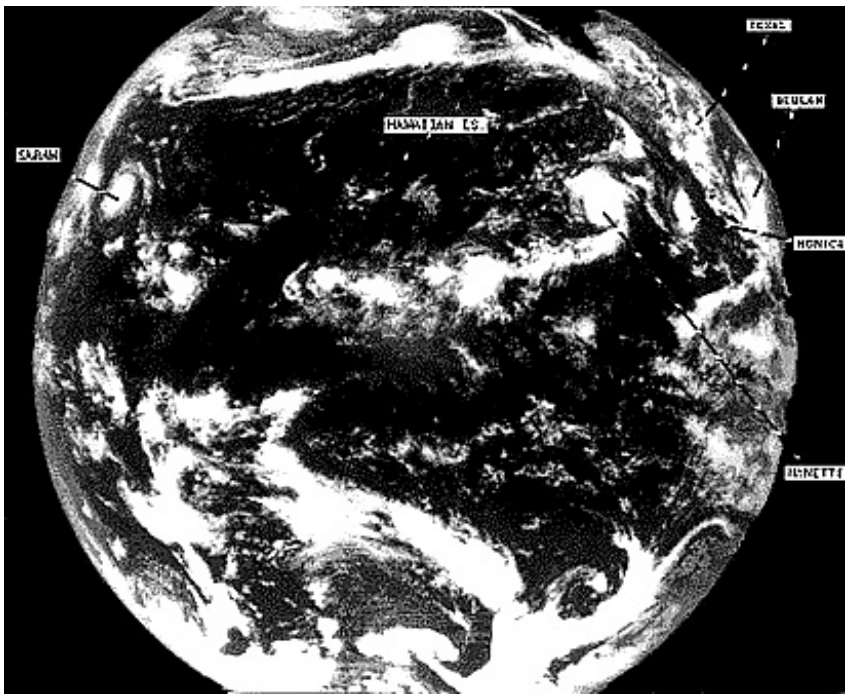
Primary Author: Nicholas M. Short, Sr.

The various geostationary satellites from several countries, either now functioning or having become non-operational, each provides synoptic views of about one-fourth of the Earth's surface. The part (it remains the same) they cover depends on where they are located relative to a point along (close to) the equator. By completing their orbit once every 24 hrs, they stay fixed relative to Earth's equatorial rotation and thus are "stationary" in space. This page concentrates on the U.S.'s GOES series but similar satellites from other nations' space programs are discussed. Thus, in today's world, mankind has an excellent network of satellites monitoring the Earth's weather systems that can be continuously viewed.

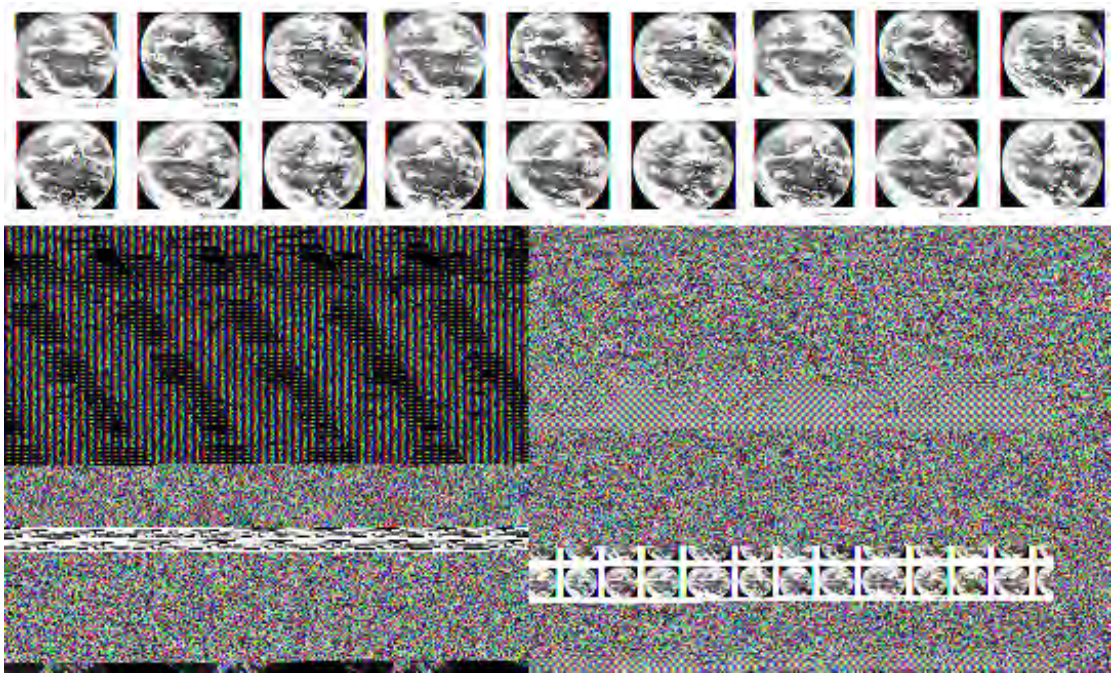
Geostationary Satellites

We switch now to consider the second general class of Metsats: those in geosynchronous orbit. That orbit is achieved when the spacecraft is "parked" above the Earth at 35,800 km (22,300 miles) and is moving along a circular path around the planet at approximately 11052 km/hr (6802 mph). A point on the Equator that remains directly underneath is traveling at ~1667 km/hr or 1042 mph. At these speeds there is no relative motion differences, so that the observing satellite is synchronously locked into a geostationary position above the hemisphere it is intended to view and (unless it drifts) will always view the same scene.

Much of the imagery shown on TV News Weather segments comes from geostationary satellites. Launched on December 7, 1966, the first such satellite was the Applications Technology Satellite-1 (ATS-1), which employed the Spin Scan Cloud Camera (SSCC) to obtain visible images of the western hemisphere. Here is the first image obtained by ATS-1, which was initially centered above Eduador (but drifted west), in which several cyclonic disturbances are present in the scene:



The value of synoptic coverage over short time spans was quickly proven by images such as these:



14-16: **Comment on at least one mega-weather pattern in the northern and in the southern hemispheres as time progresses from January 2 through the 7th.** **ANSWER**

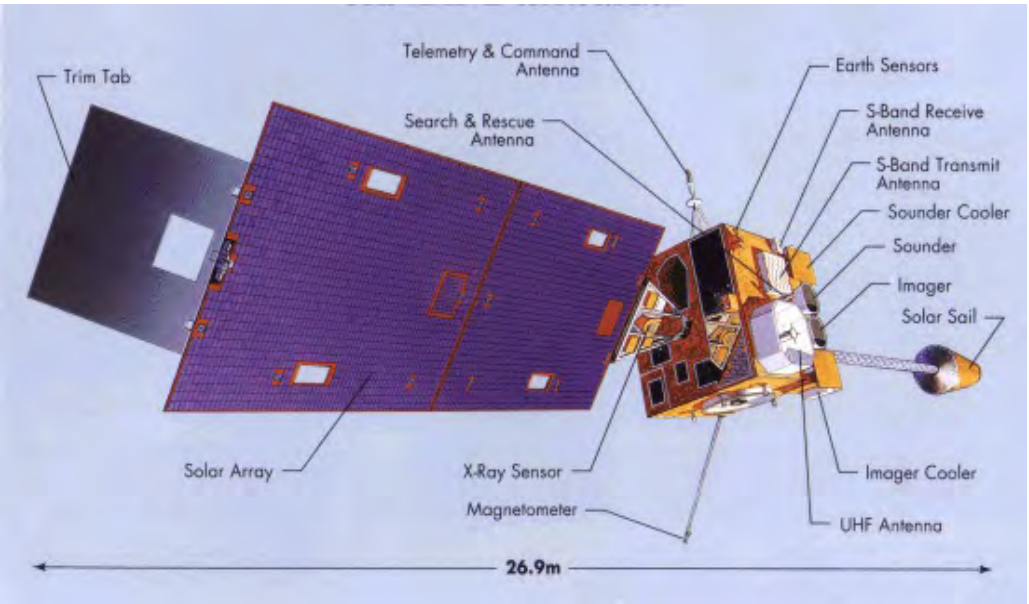
Rapid improvement in vidicon technology led to the first color system, the Multicolor Spin Scan Cloud Camera (MSSCC), on ATS-3 (launched November 5, 1967), producing striking portraits of entire continents such as this view of South America:

14-17: **Is that a hurricane at about 1:00 o'clock in the hemisphere, off the Spanish coast?** **ANSWER**

The beginning of an operational system dates to May 17, 1974, with the launch of the first of two Synchronous Meteorological Satellites (SMS). Each SMS carried, as its principal sensor, the Visible Infrared Spin Scan Radiometer (VISSR). These were predecessors to NOAA's Geostationary Operational Environomental Satellite (GOES) series-the kingpin of present day geosynchronous systems, covering the western hemisphere. Here is a photograph of GOES-1 on the ground awaiting mating to its launch rocket:

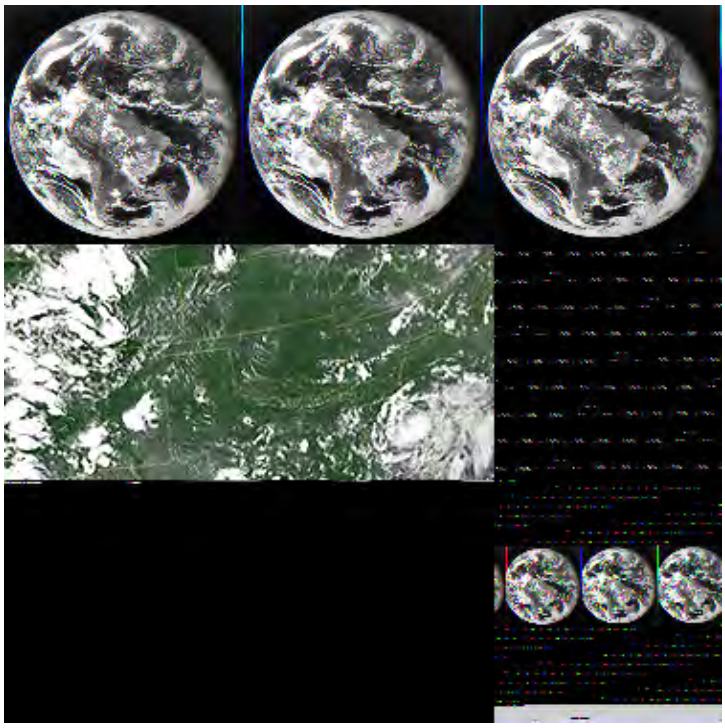
GOES-1 arrived in a geostationary orbit at 135° W, soon after its launch on October 16, 1975. Others have since been launched at two to three year intervals . Meteorologists refer to the satellites covering the Atlantic Ocean and the eastern U.S. as GOES-East (located above the equator at 75°W longitude), and those over the Pacific as GOES-West (at 135°W longitude). Together, they provide coverage of both the Atlantic and Pacific, as shown in this drawing which also illustrates the full disk nature of the view:

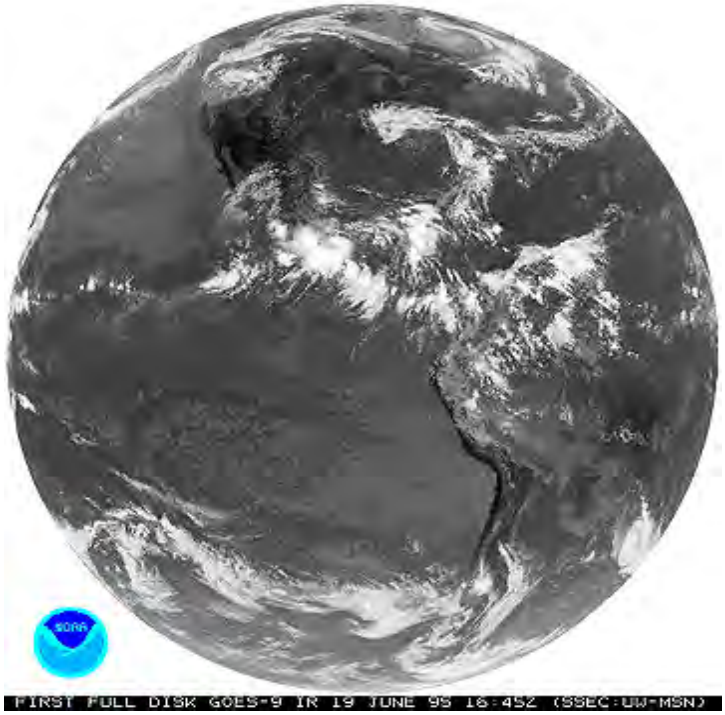
A more advanced series of GOES spacecraft, called the GOES-IM group, began with GOES-8, shown here as an artist's drawing:



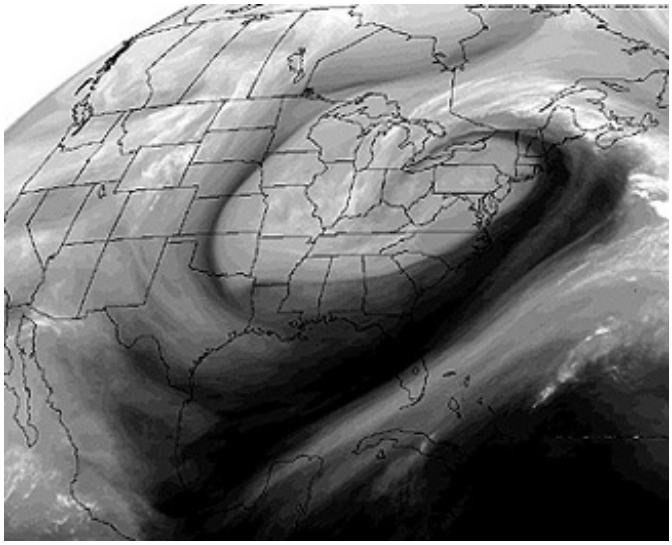
To cover the entire Earth, four GOES would be needed. However, other parts of the world are monitored by other systems (see next page). As of mid-2005, GOES-8 has been decommissioned, GOES-9 is in backup status, GOES-10 (West) and GOES-12 (East) are operating, and GOES-11 also is in orbit but in storage until GOES 10 fails.

Starting with GOES-8 (April 13, 1994), the IMAGER operates all of the time, rather than providing periodic views. A second sensor, the SOUNDER, provides profile data through the atmosphere. To exemplify GOES imagery, we now show the first visible image (top) from GOES-1 on October 26, 1975 and the first test IR image (bottom) from GOES-9 on June 19, 1995 (that satellite failed early in its operational life):



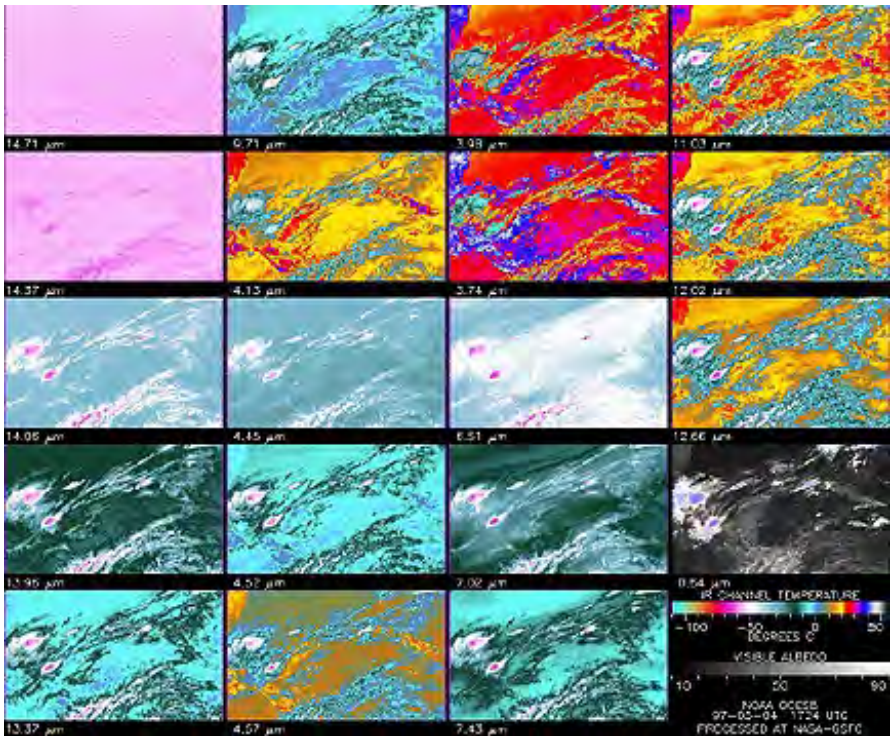


These hemispherical images can be subdivided to concentrate on specific areas. Here are a GOES-8 (East) image that focuses on a large continental storm on March 20, 1994:

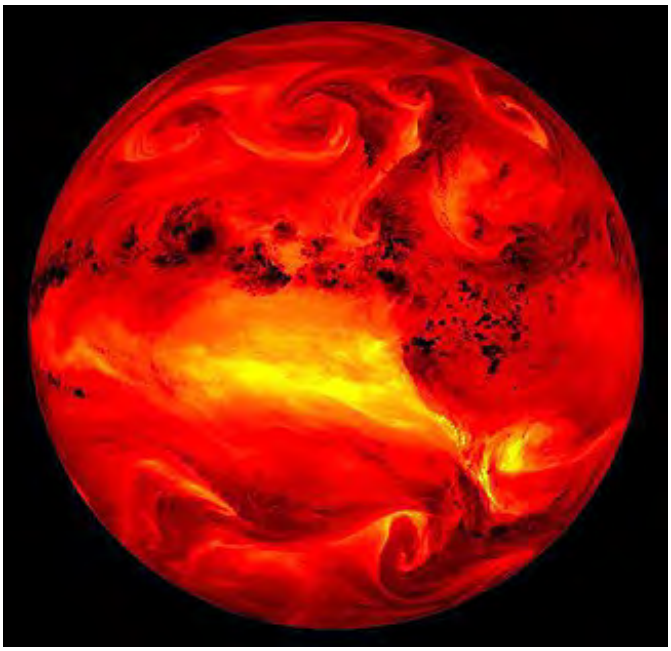


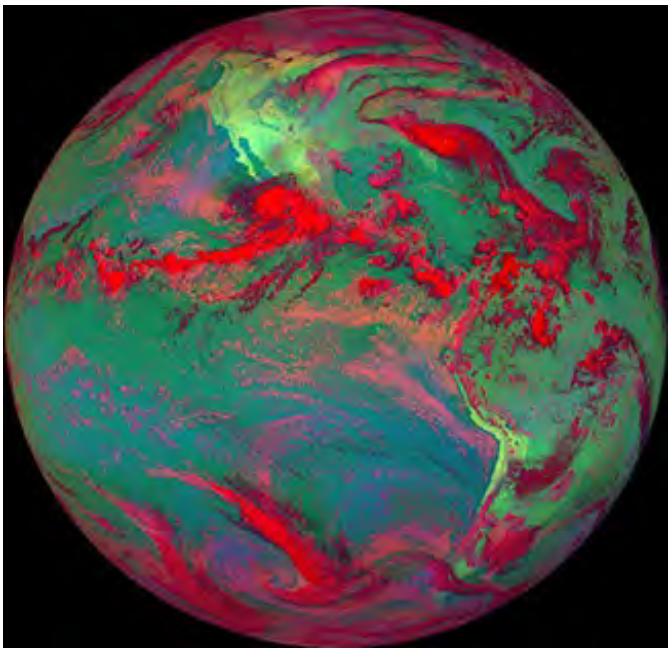
14-18: **The March 20th storm resulted in a huge snowfall over its northern segment. The cloud pattern resembles a hurricane but there are differences. Mention several of these. ANSWER**

The GOES-8 sounder has a visible band and 18 thermal bands, which are sensitive to temperature variations related to CO_2, ozone, and water vapor at different atmospheric levels. Each band is made into an image, to which colors are assigned, to identify thermal differences, as demonstrated in this panel of images taken on May 5, 1997.



Unusual color composites can be made from different channel images. Below on the top is a colorized rendition of the 6.7 µm channel image from GOES-8, which is sensitive to water vapor distribution, highlighting a big U.S. storm on March 20, 1994. Below it is a false color image made from Channel 6.7 µm = blue; 11 µm = green; Visible = red. This GOES-8 image displays Hurricane Hugo, as seen on September 21, 1989.





The GOES-NP group is scheduled to begin their launch sequence in 2006. An even more advanced satellite, GOES-S, may fly by 2012.

Various nations have now launched geostationary satellites. A complete listing is found online at this Colorado State website.

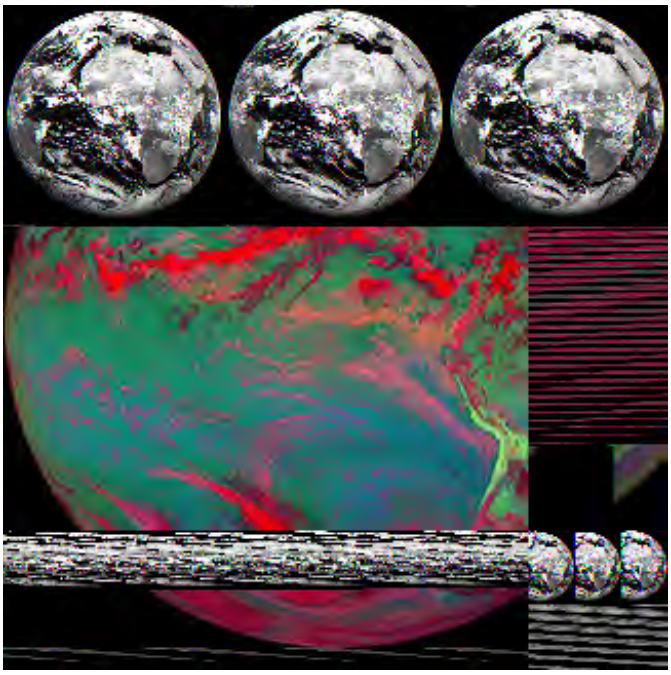
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

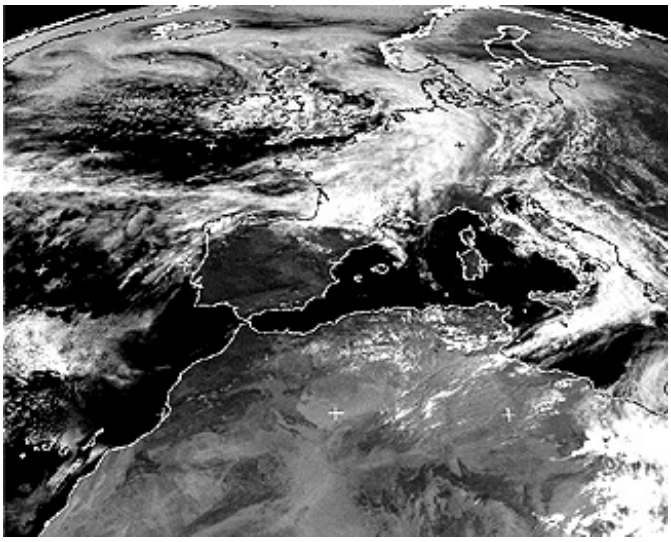
We continue coverage of geostationary satellite imagery by showing examples of images acquired at different times and subglobal locations by non-U.S. satellite systems including Eurmetsat's Meteosat, Russia's GOMS, India's Bhaskara, centered on the Indian Ocean, Japan's Himawari series, and China's Feng Yun series..

International Geostationary Meteorological Satellites

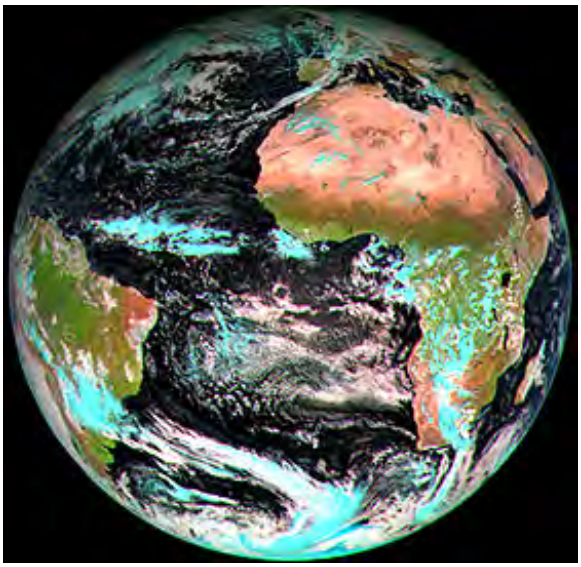
The European Space Agency (ESA) began a geosynchronous satellite program (EurMetSat) in 1977 with Meteosat-1. All told, they have placed six Meteosats in orbit. These satellites sense in three spectral bands: 0.4 - 1.1 µm, 5.7 - 7.1 µm, and 10.5 - 12.5 µm. A recent Meteosat image, showing all of Africa and most of Europe, across the South Atlantic to Brazil, was taken during the day of June 26, 2000:



This is a Meteosat-6 visible band image, a subset from the full scene, looking over western Europe, made on April 28, 1997:

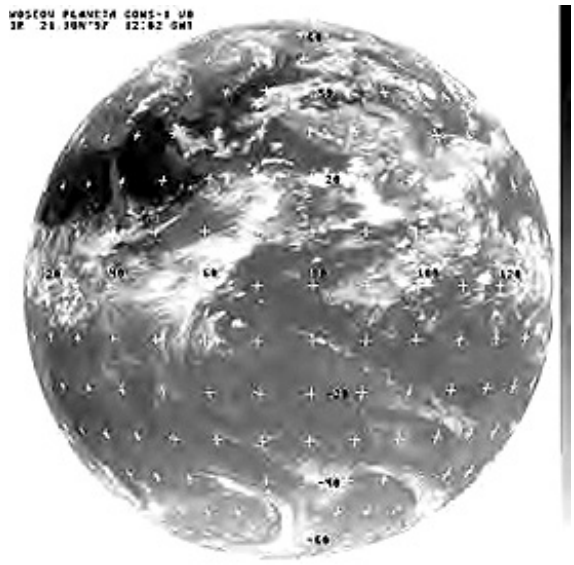


Recently, second generation Meteosats have been locked in orbit. Here is a scene centered on the Atlantic:

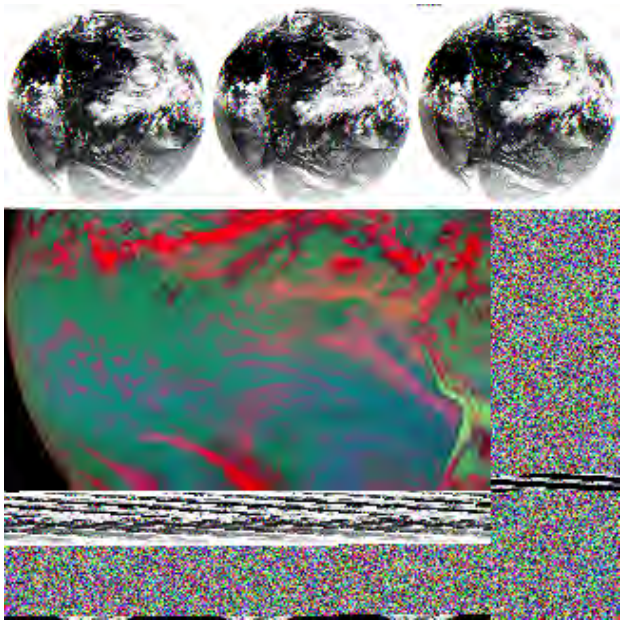


14-19: **From an airplane, should you be able to see across the Strait of Gibralter on that February day?** **ANSWER**

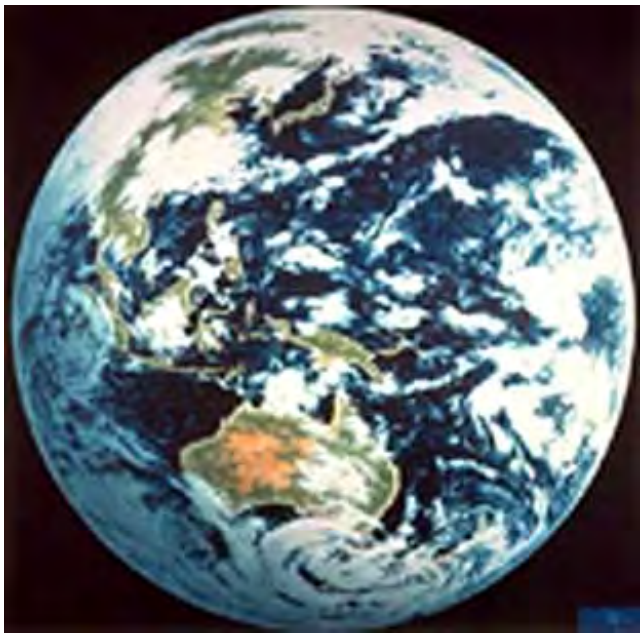
Russia started a new geosynchronous weather satellite program (GOMS; also known as Elektra) in 1995. Here is a GOMS-1 thermal band image over the Indian Ocean, made on February 28, 1995.



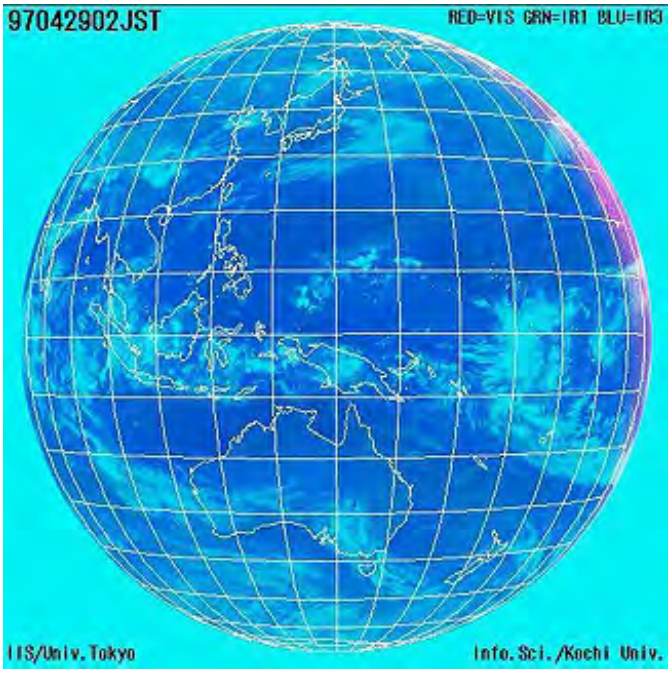
India and Japan have been active in designing, building, and launching Metsats since the late 1970s. India's efforts began with the launch of Bhaskara 1 on June 7, 1979. There has since been two series (four each) of Insats (Insat-1A on April 10, 1988; Insat-2D in 1995), each with a visible and a thermal IR band. These are multi-purpose satellites also dedicated to communications and broadcasting. We show a representative Insat image (thermal IR; night) covering eastern Africa and southern Asia in the top image.



Japan commenced its Geostationary Meteorological Satellite (GMS) program (Himawari series) on July 14, 1977, sponsored by its National Space Development Agency (NASDA). Their two goals were to create an operational system and to contribute to the worldwide Global Atmospheric Research Program (GARP) effort. This is a typical GMS color image of the western Pacific:

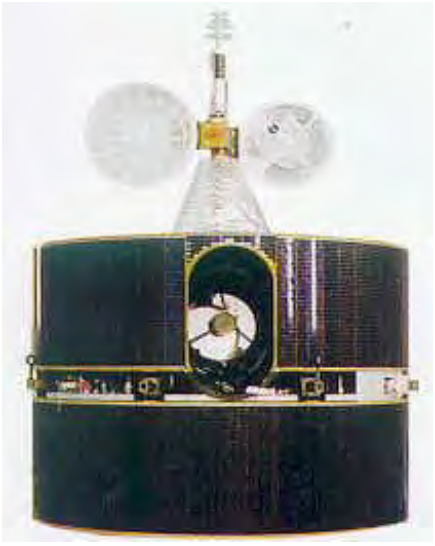


Positioned at 140° E, the current unit, GMS-5, with a VISSR in the visible, thermal IR, and water vapor (6 μm) bands, produces black and white, individual, full-Earth views and color composites, such as that shown on below over the western Pacific.



14-20: **Two large land areas, both of which we think of as commonly with minimal cloud cover are in the upper image above heavily overcast. What areas? ANSWER**

China has initiated its own geostationary satellite program. There have been several Feng Yun-2 satellites in this series. This is a view of the first.



This is a typical image from Feng Yung-2:

14-21: **For global synoptic coverage, what is the distinct advantage that the geostationary image composite has over a comparable effort using polar orbiting Metsats? ANSWER**

Same day (near real time) coverage from all the above satellites is available at the Dundee University (UK) web site (you must establish a password).

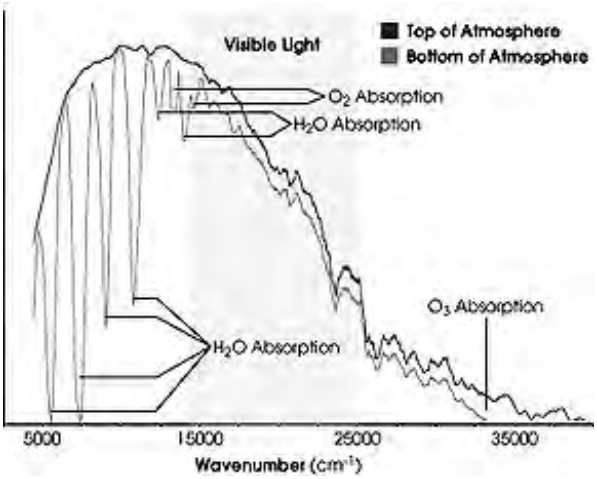
PREVIOUS RST NEXT

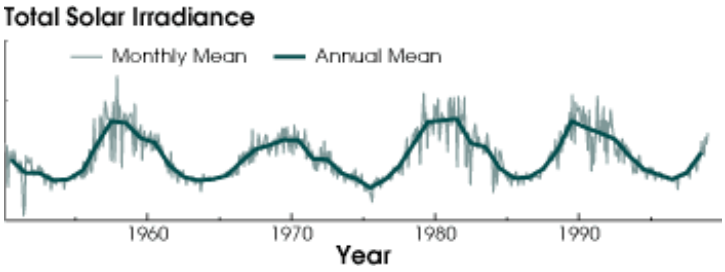
Primary Author: Nicholas M. Short, Sr.

NASA has always played a key role in conducting basic studies in meteorology from space. It is usually joined in these efforts by NOAA and by university researchers. One-time only special missions, using sensor systems developed for specialized measurements, have been flown. Three of these dedicated programs are described on this page: ERBS; UARS; ADEOS (a Japanese satellite with sensors from NASA, NOAA, and other contributors); Earth Probe is mentioned.

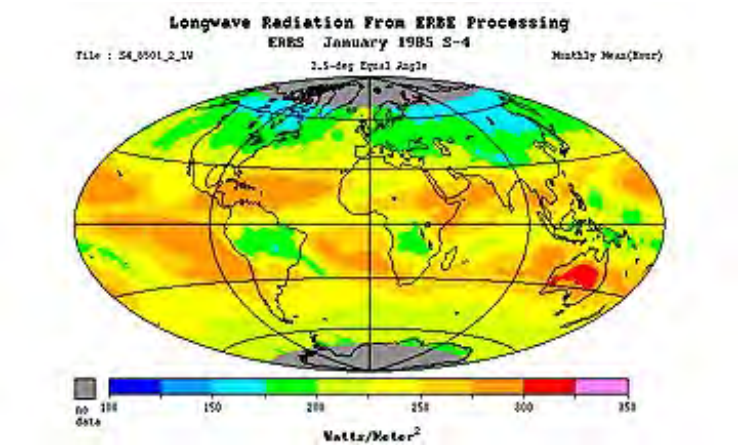
ERBS; UARS; ADEOS

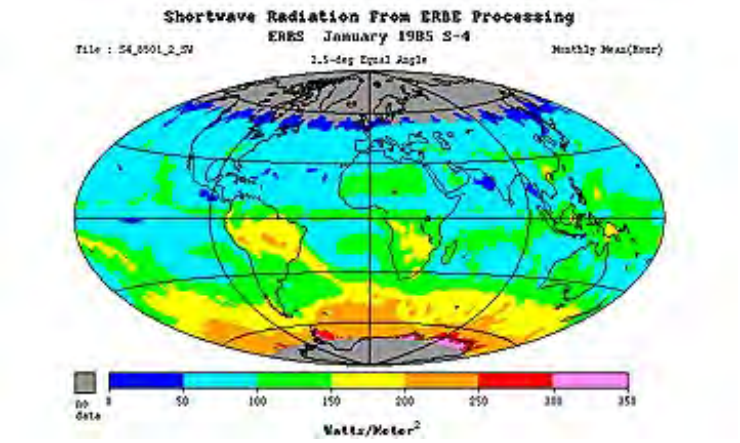
Similar to some of the first Metsats, several U.S. missions in the past 18 years have been flown to primarily perform atmospheric/climate research. Of particular importance in these studies is daily and seasonal knowledge of the amount of sunlight - total solar irradiance - reaching the Earth. Satellites above the main atmospheric column are especially adept at measuring this. Over the years, the satellites have led to these two general plots:



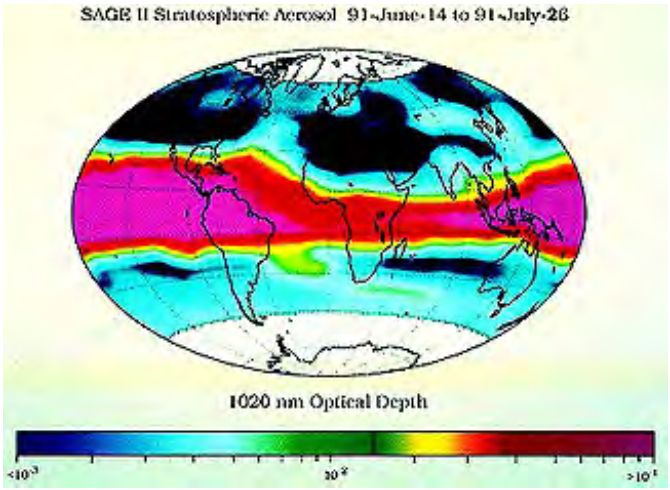


On October 5, 1984, the Earth Radiation Budget Satellite (ERBS) went into a non-synchronous orbit that permitted sampling at various times of day. The ERBS consisted of two instruments: a scanner and a non-scanner, designed to measure diurnal variations of incoming solar radiation and its differential absorption by the atmosphere. The latter had five broadband channels, one looking sunward, and the others at Earth, over wavelength intervals within the 0.2-5.0 µm range. The scanner's three channels undertook similar functions. The ERBS also carried the Stratospheric Aerosol and Gas Experiment II (SAGE-II), a limb sounder, which used four bands (0.385, 0.45, 0.60, and 1.0 µm) to measure aerosol and Rayleigh (fine-particle) scattering, whenever the satellite is in position in its orbit to view sunrise and sunset. The next images show ERBS-determined, long-wavelength and short-wavelength, radiant-flux densities during January, 1985.



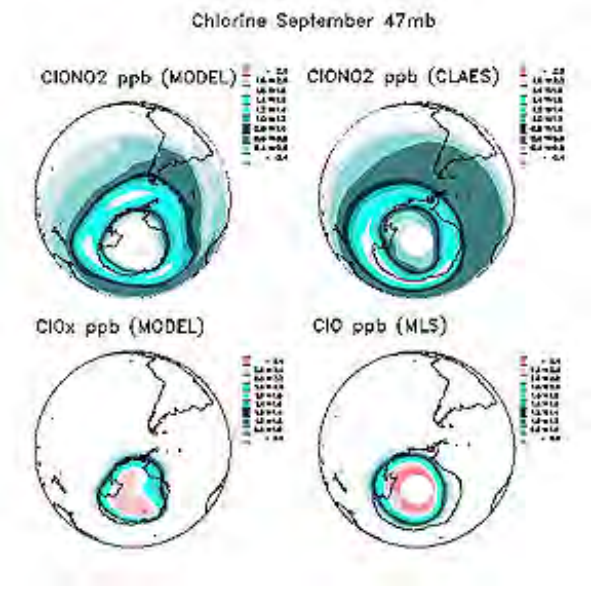


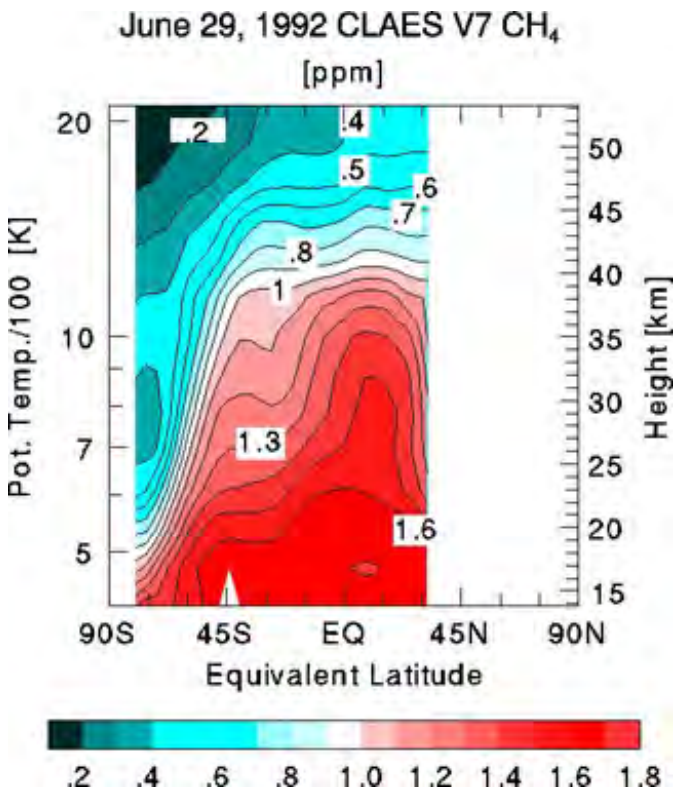
Here we show aerosol concentrations over a period of six weeks, during summer 1991, as measured by SAGE-II. Note: these instruments flew on other satellites, including two in the NOAA series.



14-22: **What were ERBS and SAGE actually measuring? ANSWER**

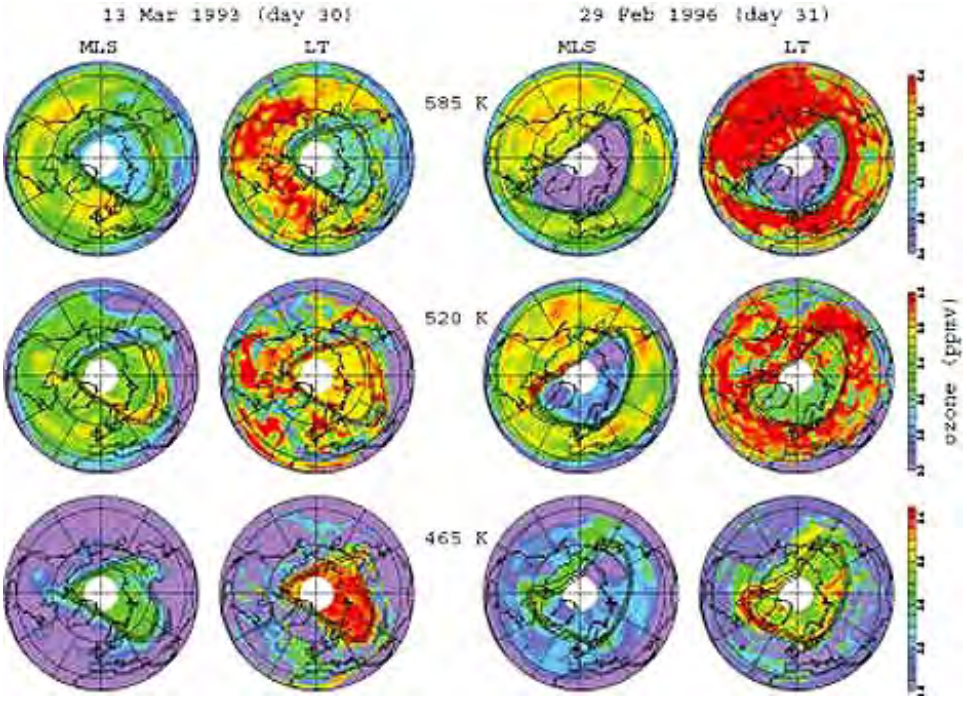
The Upper Atmosphere Research Satellite (UARS) deployed from a Shuttle (STS-48) on September 12, 1991, with ten instruments including the Microwave Limb Sounder (MLS), the High Resolution Doppler Imager (HRDI), and the Improved Stratospheric and Mesospheric Sounder (ISAMS). In the next images, on the top is a plot of the distribution around the South Pole of chlorine monoxide (ClO) and nitrogen dioxide (NO_2), in terms of predictive models and actual results from the MLS and Cryogenic Limb Array Etalon Spectrometer (CLAES). On the bottom is a CLAES plot of methane (CH_4) distribution through the atmosphere along a specific orbit.





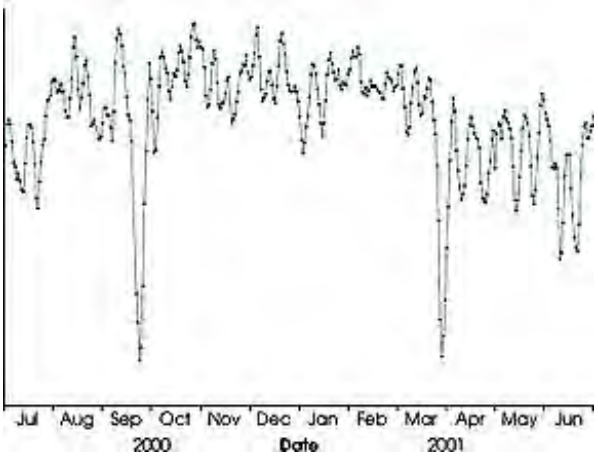
14-23: **Where does the methane (CH_4) concentration appear to be maximum? Speculate on why.** **ANSWER**

Among the many gases studied by UARS is ozone (O_3), of which much attention is being paid to the so-called Ozone Hole at each Pole because the depletion of ozone can have serious consequences to health (loss of ozone increases the risk of damage to life by solar UV rays). This set of maps shows variations in ozone in the Arctic region as detected by UARS' MLS instrument in wintertime during two dates three years apart; the distribution is at different elevations in the upper atmosphere as indicated by temperatures (degrees Kelvin):



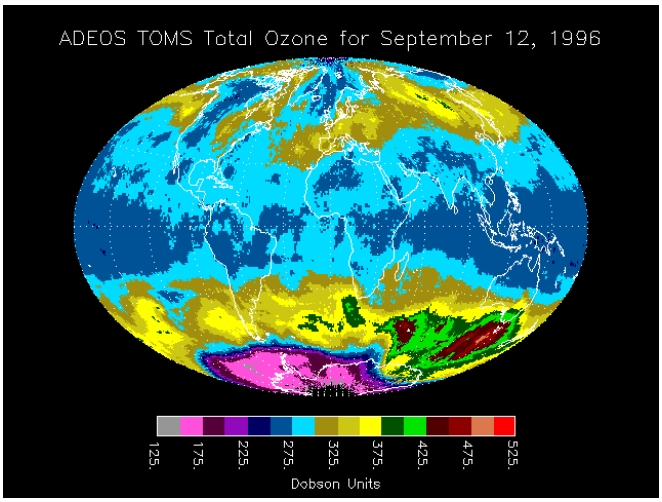
UARS is a good example of the use of multiple sensors to study the properties of a single target medium, in this case the atmosphere. As such, it is a precursor model of the use of multisensors to monitor the Earth's diverse environments, as is being done now in the EOS Terra and Aqua programs (Section 16). A worthwhile review of UARS is found at its Home Page.

Between 1980 and 1991, three solar irradiance satellites, the ACRIM (Advanced Cavity Radiometer Irradiance Monitor) series, were placed in operation. Here is a typical plot of measured solar irradiation:

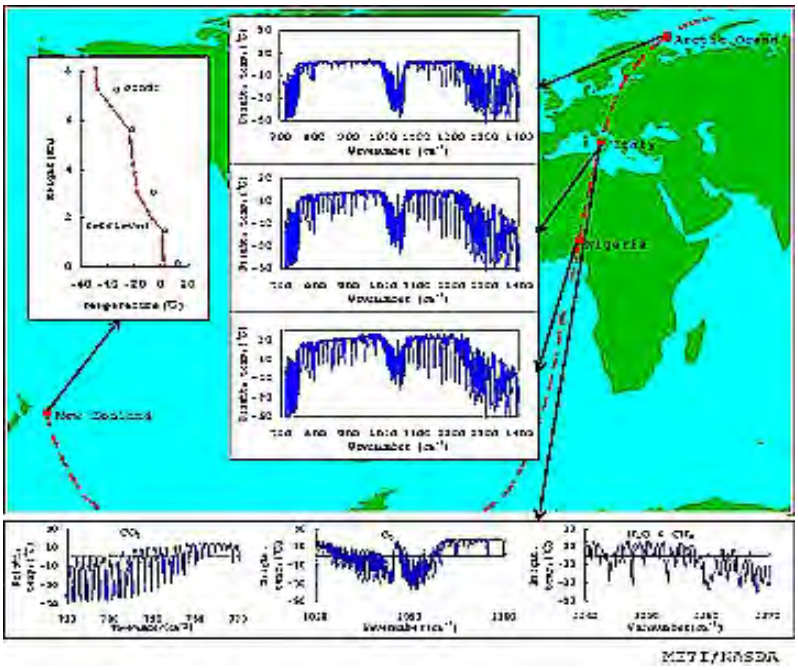


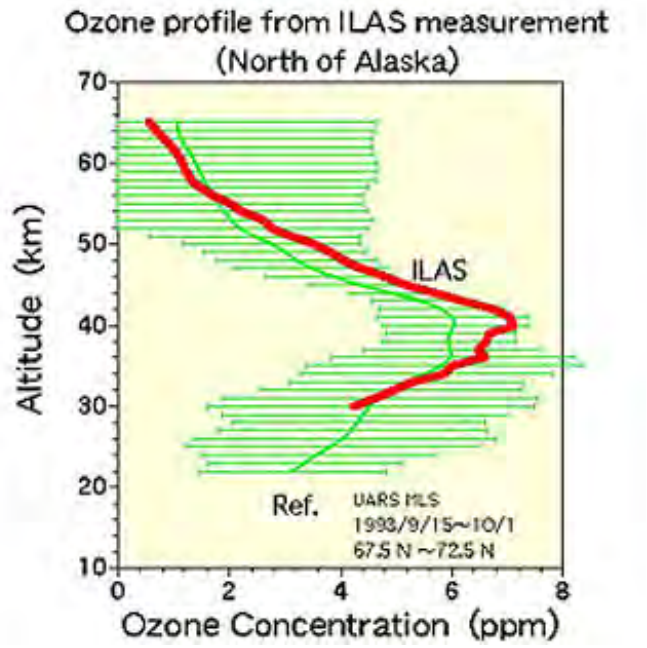
The Advanced Earth Observation Satellite (ADEOS) is a Japanese spacecraft, launched on August 17, 1996 and renamed by them as MIDORI, to conduct land, sea, and atmospheric studies using ten instruments supplied by NASA, NOAA, NASDA (Japan) and CNES (France). Sadly, this versatile satellite failed on September 12, 1996 but not before sending back some excellent, proof-of-concept data.

It is now becoming common to piggy-back sensors from several nations on a given satellite. Among ADEOS sensors are a NASA Scatterometer (NSCAT), discussed on page 14-12. Another NASA instrument is TOMS (Total Ozone Mapping Spectrometer), developed at Goddard Space Flight Center. Besides ADEOS, TOMS has flown earlier on Nimbus 7, Meteor-3 (Russian), and since ADEOS on Earth Probe (see below). Here is one of the few TOMS maps from ADEOS.



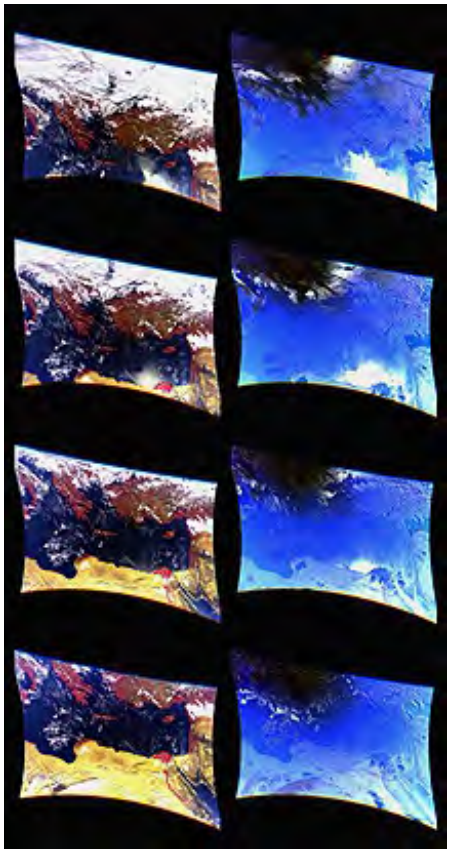
Next, we show examples of ADEOS sounder plots using instruments supplied by two Japanese agencies: the Interferometer for Monitoring of Greenhouse gases (IMG), which looks at absorption bands for CO_2, CH_4, H_2O, O_3, and NO_2 in a 0.715-2.0 µm channel, CO in a 2.0-2.5 µm channel, and CH_4 in a 2.32-3.05 µm channel; and the Infrared Limb Atmospheric Spectrometer (ILAS), which examines ozone distribution.



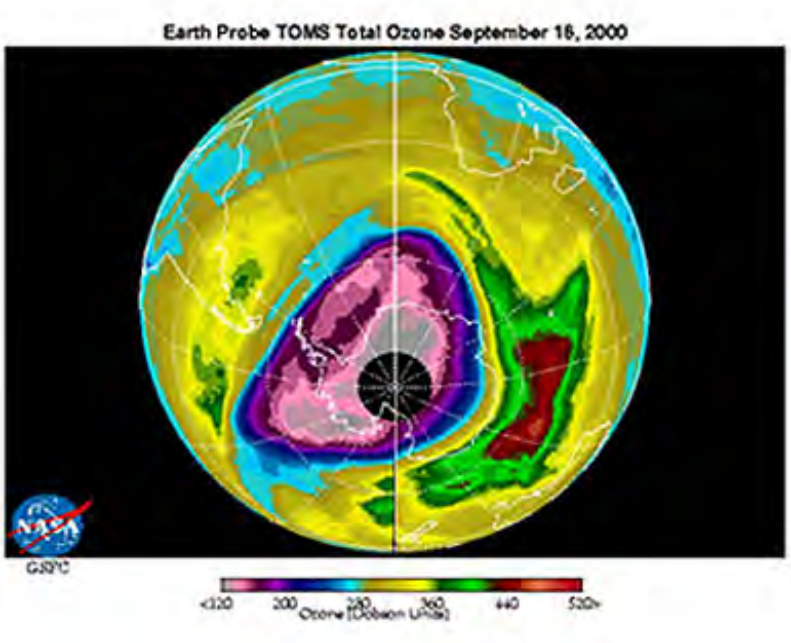


14-24: **Is it reasonably safe to fly in a jet with respect to ozone as a contaminant? What are the horizontal green bars in the right diagram? ANSWER**

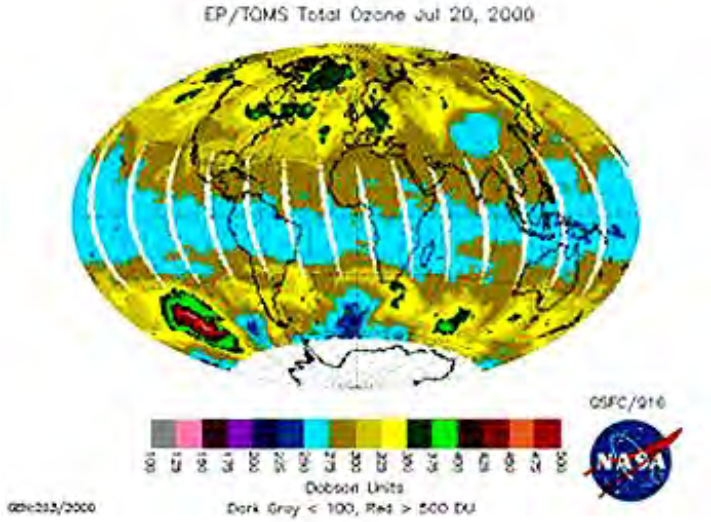
Another ADEOS instrument, developed by the French, is POLDER (*Pol*arization and *D*irectionality of the *E*arth's *R*eflectance). It includes 8 channels in the Visible and Near IR and can take images from different directions. It has a polarimeter that allows polarized light properties to be assessed. The image set below shows part of the Mediterranean in 4 successive views on the left and polarized light components (dominated by blue from scattering) on the right.



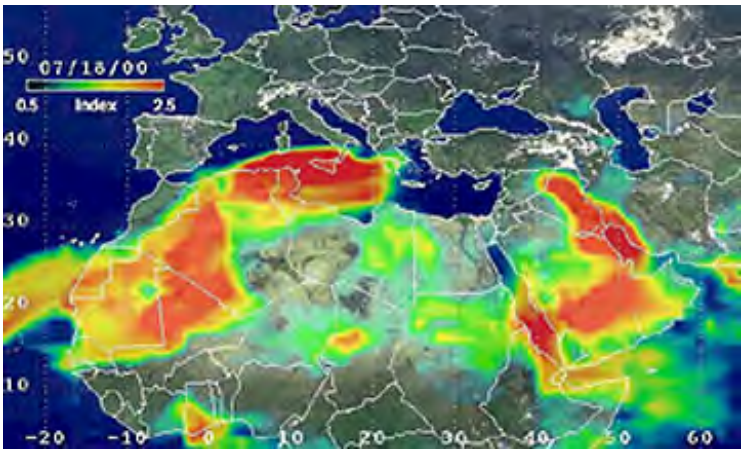
The early loss of ADEOS meant that one satellite monitoring ozone was no longer able to supply this critical data set, first collected in 1978 by the TOMS instrument on Nimbus-7. Fortunately, continuity was preserved by the launch in 1996 and continuing operation of Earth Probe (EP) which included a TOMS. This is a typical result for the South Polar region:



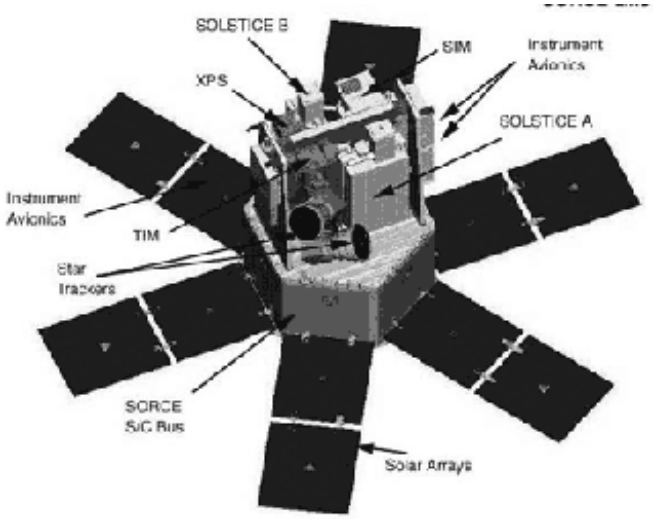
The next image was taken over a short period around July 20, 2000:



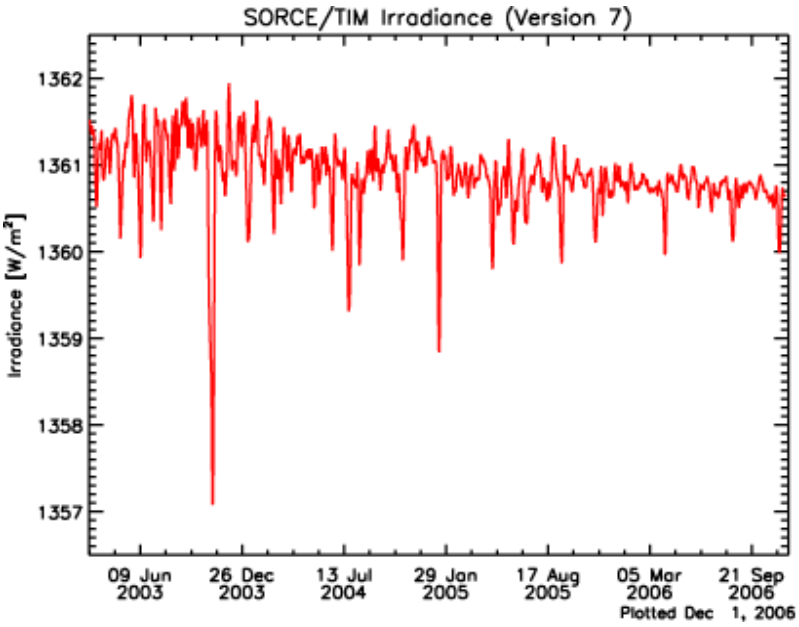
The previous week, the EP TOMS produced a neat image of dust storms raging in North Africa and the Near East. Here, TOMS was measuring aerosols associated with the dust storm:



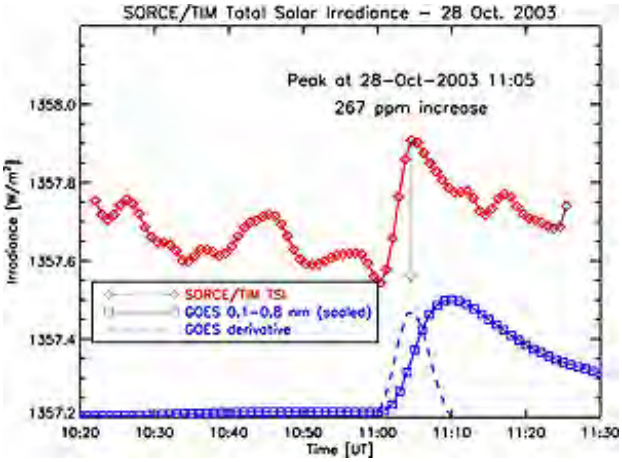
A recent earth radiation budget satellite is SORCE (Solar Radiation and Climate Experiment, launched on January 25, 2003 and managed jointly by NASA Goddard and the University of Colorado. Here is the spacecraft:



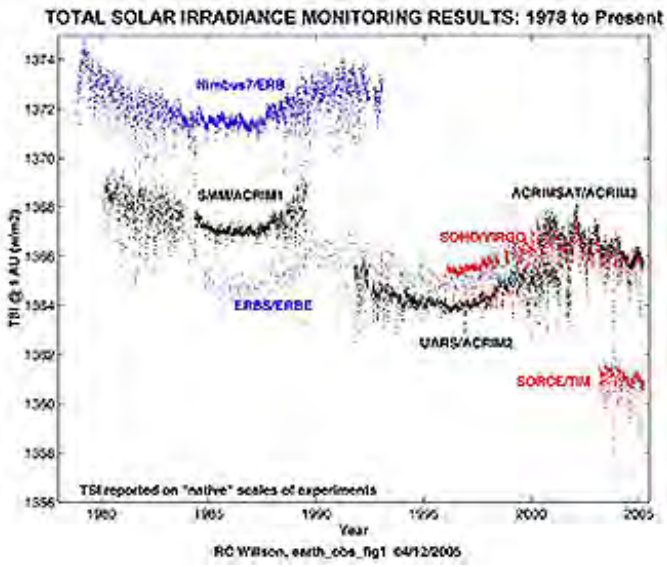
This is a typical data set for its main instrument, TIM (Total Irradiance Monitor):



SORCE has proved invaluable for monitoring both the occurrence and the effect of solar flares on incoming radiation:



The instruments on the several irradiance monitoring satellites are carefully calibrated to correspond to "absolute" values. Nevertheless, each has returned slightly different levels of irradiant energy, as indicated by this diagram (note the small range of values on the ordinate):



PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

On this page, we are pleased to add a study from a reader of this Tutorial who offered to prepare a synopsis of his Internet website that utilized GIS. Dr. Scott Madry and his team of associates have for many years been conducting a research project into locating archaeological ruins in a part of the Province of Burgundy in France. In recent years they have used a combination of SPOT and radar imagery together with multi-layered GIS analysis to help in finding ancient ruins such as hill forts in the rolling terrain around the Arroux River near Autun, south of Dijon, France. Part of their approach is to derive line-of-sight maps that pinpoint favorable higher locations where these forts may once have been built but are now eroded down or dismantled so as to escape ready visual detection. Remote sensing helps to disclose patterns that may be structure foundations.

AN INTEGRATED GIS/REMOTE SENSING CASE STUDY IN FRANCE

Guest Writer: Dr. Scott Madry

Note to CD-ROM Users: You must go online to access some of the links cited below.

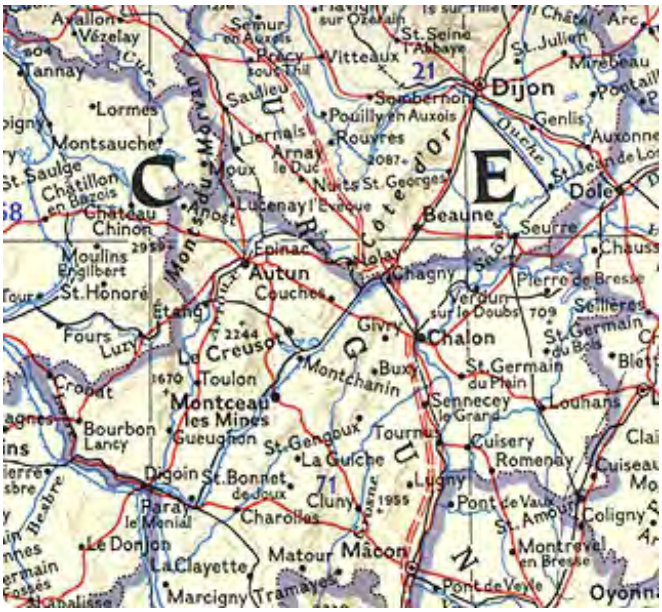
Background for this Study of Archaeological Sites in Burgundy

Another case study can show the integration of GIS and various types of remote sensing data for a rather unusual application, archaeology. This research project is a long-term analysis of the application of Geomatics to study the interaction between different cultures and the landscape over time in Burgundy, France. An American interdisciplinary team has conducted research in the Arroux River Valley region of Burgundy for over two decades.

The team uses an integrative approach termed historical ecology: the multiscalar analysis of the interaction of culture and the environment over time. A period of over 2,000 years (from the Celtic Iron Age, Gallo-Roman, Medieval, to the present) is being analyzed. The overall goal of the research has been to understand long-term interaction between these very different cultures and the physical environment. Various remote sensing, GIS, and GPS technologies have been used to support this work, which has been conducted by Dr. Scott Madry.

This map of France shows the location of the study area circled in red.

The area of Burgundy, whose chief city is Dijon, is shown in an extract from an Atlas map; note the location of Autun.



A feel for the landscape of parts of Burgundy is provided by this ground photo.

Study Area

The research area consists of the valley of the Arroux River, a tributary of the Loire River, in the Burgundy region of east-central France.

The study area is shown from space with this SPOT image that has been merged with DEM data to give a three-dimensional sense of relief (somewhat exaggerated); Autun is in the dark area in the upper right:

The valley is located in the modern Department of Saone-et-Loire, with the river flowing from its headwaters in theMorvan Mountains to the Loire along a northeast-southwest trending line.

Along the heights above the river a series of Celtic hillforts are located. These have figured significantly in the history of the region, as well in the research activities. Foremost of these is Mt. Beuvray, in the northwest corner of the study area. This impressive mountain was the site of the great Celtic city of Bibracte, capital city of the Aedui, the powerful Iron Age Celtic polity whose territory was centered on the Arroux. Bibracte reportedly had a population of over 30,000 in 52 B.C., when Julius Caesar quartered his legions there. Other, smaller, hillforts follow the river to the south, while others follow the opposite side of the river.

The major modern city in the valley is Autun (see photo of the ancient Roman city walls as they look today). It was founded by Julius Caesar as Augustodunum Aedorum, home of the relocated Celtic inhabitants of Bibracte after the Roman conquest. Caesar's conquest of Gaul, including his time in this area, was documented in his Gallic war commentaries "De bello gallico commentarii".

While the region retains a mostly rural character, there are several modern threats to cultural

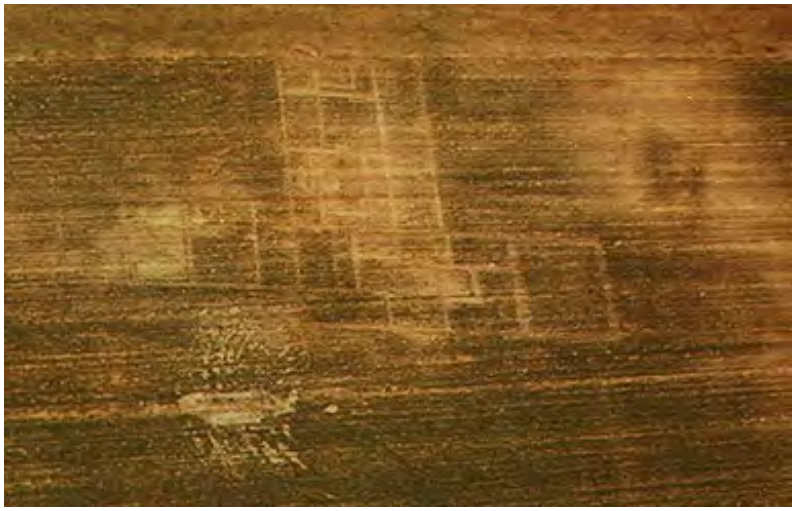
resources in the region. The most serious of these are a series of gravel mines (sabliers) that run along the banks of the river. These mines are taking much of the land immediately adjacent to the river. The mines are destroying a large number of archaeological resources in the area, and one aspect of the project is to locate and document these sites before they are destroyed.

Aerial Photography

Archaeologists have been using aerial prospecting and photography since just after World War One to locate buried structure, roads, and other features of ancient landscapes. Several archaeologists who served as pilots in that war noticed strange circles and square patterns from the air that were not visible on the ground. After the war they returned to these locations to find that these were actually archaeological sites, ancient roads, etc. Early pioneers of 'aerial archaeology' conducted the first such surveys in Europe and the Middle East in the 1920's. Charles Lindbergh also conducted similar surveys in Central America in the 1930's. How is it possible to see ancient landscapes and the remains of structures that can be a thousand years old?

Faint lines and color changes visible from the air are often invisible on the ground, and can be caused by buried cultural remains. Aerial archaeologists refer to these as crop, soil, and shadow marks. Crop marks form because there can be noticeable variations in crop vigor, color, or height when crops or natural vegetation grows over buried walls or other cultural remains. These are called 'negative' crop marks, as the crop is less vigorous due to the lack of moisture or root vigor caused by the buried walls. The opposite are 'positive' crop marks, where the crops are taller or more vigorous when growing over pits or post holes. There is more moisture and better root growth, so the crops grow better. These differences are heightened in times of crop stress, such as a drought. In the major French drought of 1976 thousands of new archaeological sites were discovered from the air. Soil differences also can be visible, such as where a road or ditch was filled in with soil from a different place. These are called 'soil' marks. Small variations in topography causing shadows that are visible early or late in the day, these are called 'shadow' marks.

15-20: The aerial photo below shown a Roman Villa discovered in 1979 during a low level aerial survey for this project. This is a 'negative' crop mark. What kind of detail of the structure can you see? Look in the upper right portion of the photo. There is a green square, a 'positive' crop mark. What do you think that this might be?ANSWER

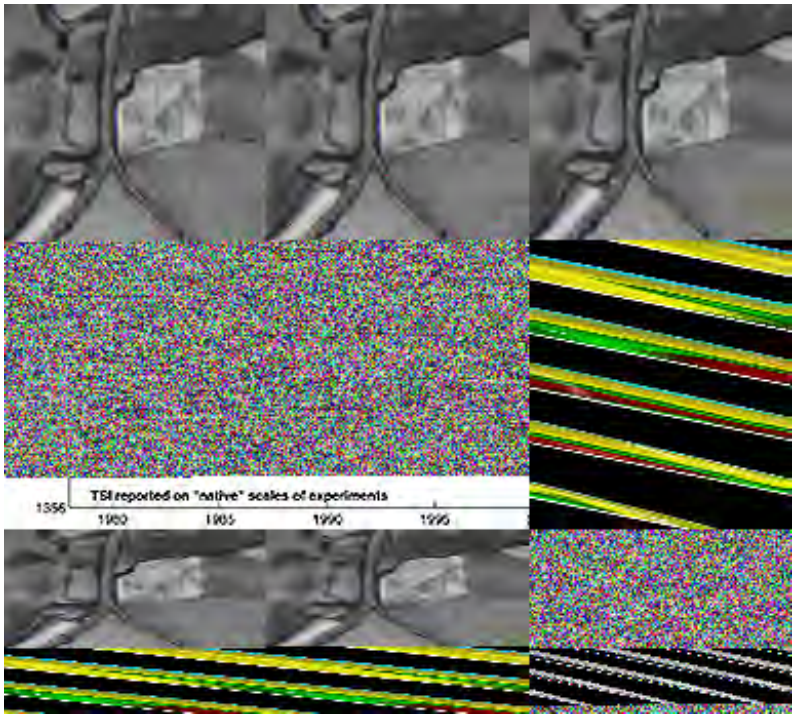


15-21: The ground photo below was taken the same day (see the person walking in the background for scale). Can you see the outline of the structure? If you were walking along, would you be able to tell there is a buried structure here?ANSWER

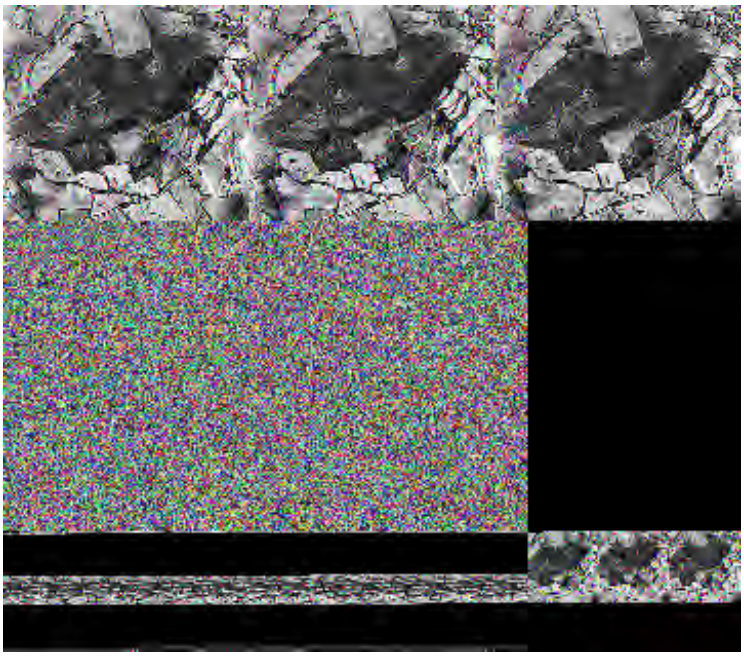
Project Historic Aerial Photography

An excellent source of remote sensing information for archaeologists is archival aerial photography, the older the better. U.S. Army Air Corps aerial recconnaissance photos of the region dating from September 1944 were acquired from the U.S. Defense Intelligence Agency using a Freedom of Information Act Request (FOIA). These photos were acquired during the Allied push into the region near the end of the Second World War. Over 200 black and white vertical aerial photos of a scale of approximately 1:40,000 were acquired and have been manually and digitally analyzed to search for archaeological sites, roads, etc. Numerous features have been located and mapped.

15-22: Look at the photo below. What kind of marks do you see? What do you think they are? ANSWER



15-23: Below are a 1944 aerial photo of Mt. Dardon and the 1983 1:25,000 IGN map. What differences do you see?ANSWER



Modern Remote Sensing Data

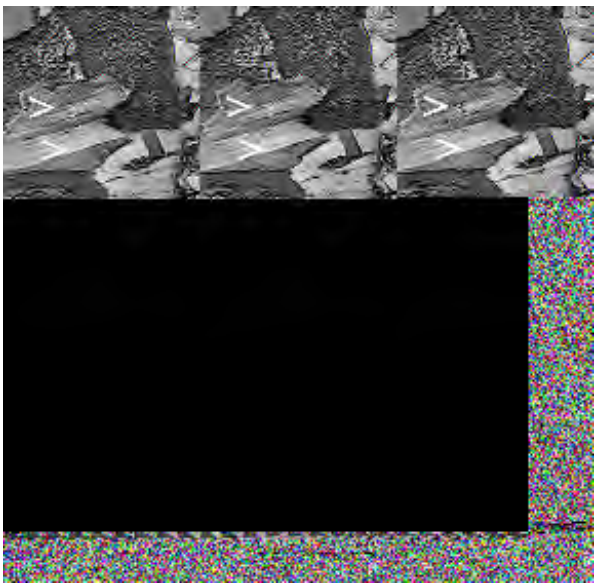
This project has used a variety of modern remote sensing data. Since archaeological remains are very small, airborne data are very useful in conjunction with satellite data.

The Aries scanner system is a French airborne digital radiometer system that was used on this project. It was constructed by the Laboratoire de Météorologie Dynamique (CNRS, France). It has two channels, normally set with one in the area of the visible and near infrared, and the other in the thermal portion of the spectrum (Perisset and Tabbagh, 1981:185). Internal calibration of the thermal scanner can provide apparent temperature recording capability. The spatial resolution of the data is dependent on aircraft elevation, but 1-2 meter data are typical for missions such as ours. For this project, the ARIES scanner system was mounted in a single engine Pilatus aircraft, and three corridors within the research area were flown in 1987.

This airborne thermal scanner image shows the location of the Roman villa structure which has been destroyed by a gravel mining operation (shown by a star on the image). The field is now a lake. This is a fairly common situation in Europe, and part of the project activities is to locate archaeological resources so that they can be studied before they are destroyed by modern activites.

15-24: In the airborne thermal image above can you see any patterns in any of the fields?ANSWER

A somewhat more ambiguous example appears in the image below. This is the area in the upper left of the above infrared image. The "arrow(>)" marks point to a faint but perhaps meaningful curved (arc) pattern which may also be an ancient wall or some other buried structural feature.



15-25: How can thermal IR imagery help in finding buried archaeological remains?ANSWER

SPOT Imagery

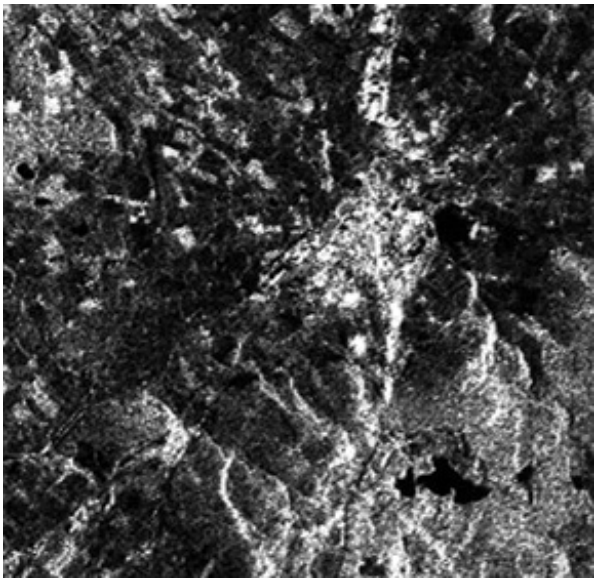
The Autun project has used a variety of satellite imagery over the years, including Landsat MSS

(80 meter) and French SPOT (20 and 10 meter) data. The SPOT data has a spatial resolution of 20 meters for multi-spectral data, which records information in three bands of the spectrum, and a 10 meter spatial resolution for a panchromatic band. The resolution of these images available from space can provide significant improvements in the utility of these data for regional archaeological and environmental applications, especially (as in France) where the field size is very small. Here is a false color SPOT image that includes part of the research area. In this image, Autun is at the upper right and Mt. Beauvrag at the upper left.

Accurate modern landcover maps were produced using SPOT satellite imagery. The upper part of this next image/classification corresponds to the above SPOT scene; look for the blue horseshoe pattern to fit it in.

Radarsat Imagery

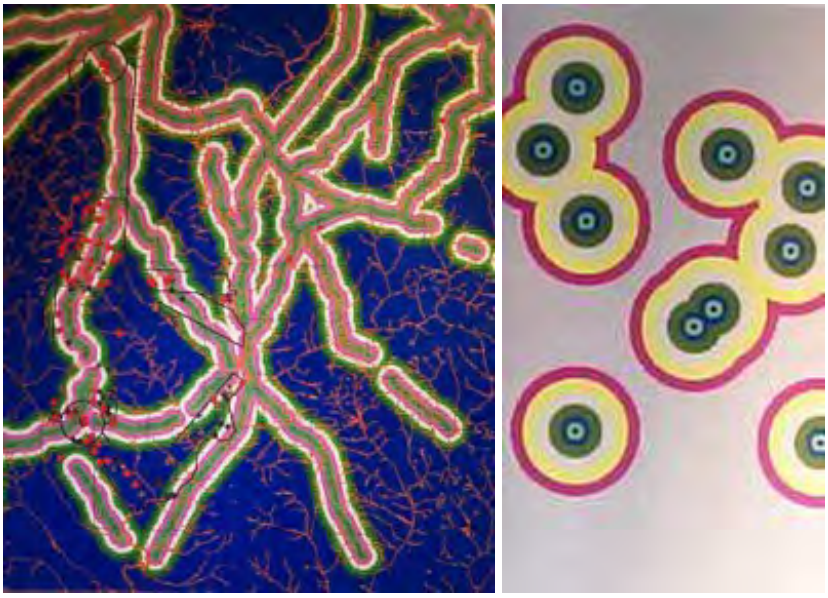
Canadian RADARSAT-1 satellite imagery has been acquired for the region on 4 November, 1998, shown below. With a spatial resolution of 8 meters, this system is different from SPOT or Landsat in that it is an active radar system that sends its own burst of electromagnetic radiation down to the ground which bounces off the surface and recorded by the satellite. This system can operate day or night and through cloud cover. It provides a different and new way of visualizing the area. The first scene below shows the area around Autun. Below it is another Radarsat image that illustrates the hilly topography in parts of the study area.



GIS Analysis: The Project GIS Data Base

The GIS data base covers an area of about 30 by 60 km, covering the majority of the Arroux River Valley and its immediate environs. The current basic raster layers of the GIS data base include: *elevation* (generated from the French digital elevation data), *aspect* (derived from the digital elevation data), *slope* (derived from the digital elevation data), SPOT images (20 meter false color infrared), SPOT images (10 meter panchromatic), *land use/land cover maps* (derived from Spot image data), *geology* (generated from 1:80,000 geology map of the upper 2/3 part of the region), *faults* (from the same 1:80,000 geology map), *hydrology* (from the three 1:50,000 topo maps), *modern roads* (from the three 1:50,000 topo maps), *ancient roads* (from project information and old maps), *known Celtic hillforts* (from project information and old maps) *data layers showing different distance categories*, or buffer zones, from roads, streams, faults, archaeological sites, hillforts, and ancient roads (also generated from the data above). Additional data have recently been added that were derived from the 1:25,000 maps.

Two of the thematic maps listed above are of special utility in the GIS analysis described below. On the left is a map of the Celtic road network developed in or after Roman times. On the right is a map of known hill forts of the time. These are helpful in calculating distance from any vantage point to specific features.

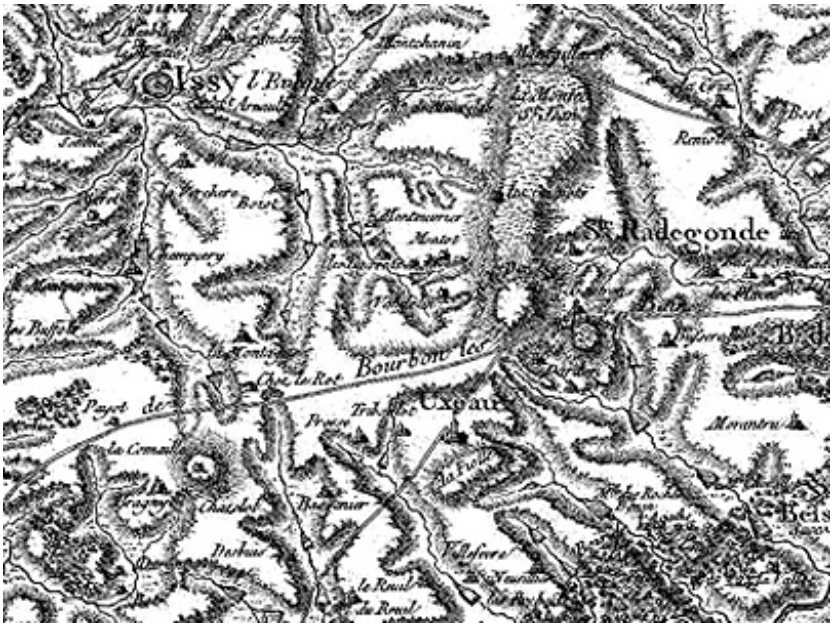


Historic Map Digitization

An original year 1659 map of the region and two 1759 maps of the research area were digitized using 4,000 by 4,000 digital CCD array device that converts maps or photographs into digital format. The 1659 map which was scanned from the original, entered into the GIS, georeferenced,

and entered into the GIS system as a data layer like any other. This map is quite general in its spatial accuracy, but it is the oldest map yet found that covers the research area. Here is part of this map:

Two 1759 maps were also scanned into the system and patched together to make a single data layer. Here is the Cassini map centered around Mt. Dardon. It clearly shows the roads, rivers, topography, and every individual house and structure.



These 1759 maps were produced for the famous Cassini triangulation survey of France conducted in the mid 1700's. This massive work took three generations of the Cassini family to complete, and was the first accurate survey of an entire nation using modern surveying and mapping techniques. Reproductions of the maps are available for all of France from the Institut Geographique National in Paris. These historic maps are extremely detailed and accurate, much more so than the older map we possess created 100 years before it.

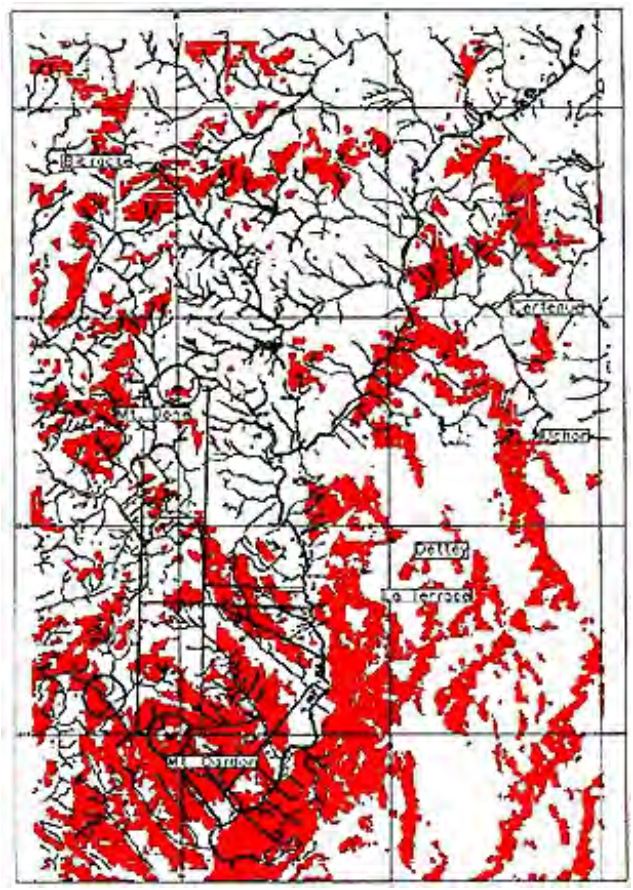
The Cassini maps provide us with the oldest, reasonably accurate, record of the location of roads, bridges, towns, and villages of the area. It also shows information about the vegetation and land-cover of the area at that time. The maps contain an astounding level of detail, and have been digitized to represent a series of GIS layers representing the cultural and environmental makeup of the area at that time.

GIS Analysis: Line-Of-Sight Determination

One of the more interesting results from the GIS analysis was the line-of-site analysis (Madry and Rakos 1996). This GIS technique allows one to determine what parts of the landscape are visible from any given location. The research demonstrated that the old Celtic road network connecting the hillforts of the area tended to follow within the line-of-sight of the hillforts, rather than take more direct paths (as originally proposed in Madry and Crumley 1990). We have run the same line-of-site analysis from four locations on each of the known Celtic hillforts in the research area . These four line-of-site maps for each hillfort were then combined to generate a map of complete inter-visibility from each entire hillfort, assuming that the forts were manned by watchers from each of the four "corners" of the ramparts. An eye height of 5 meters above the terrain was used, assuming that towers of just over 3 meters height were strategically located around the ramparts (and that the eye level was just under 2 meters above that height).

These individual line-of-site maps are combined to produce the map below which shows (in red)

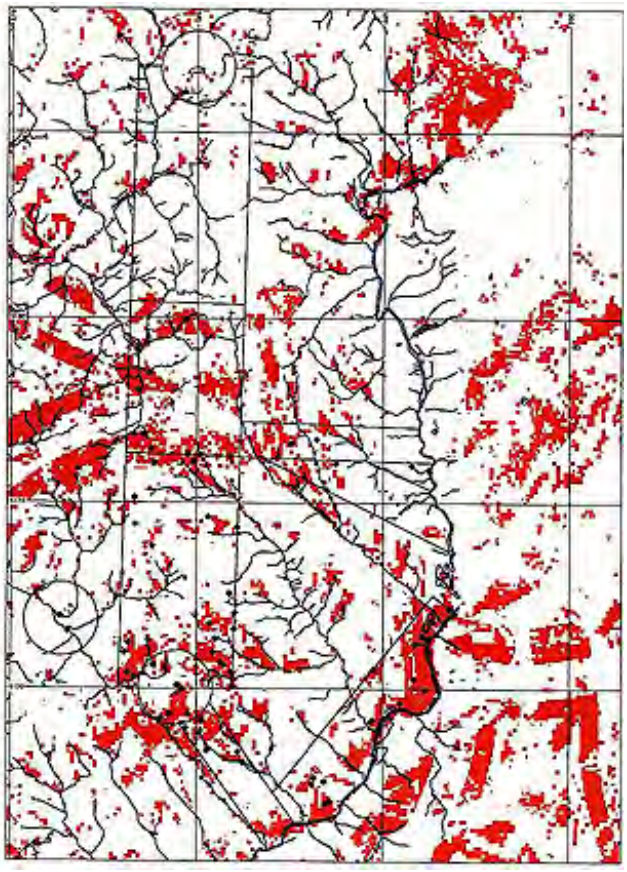
the total portion of the research area and each ground transect that is within line-of-site of each hillfort, and for the total network of hillforts in the region. This analysis shows that the Celtic roads definitely tend to follow paths that stay within the view of the hillforts, even if they are less direct or require a steeper climb. Additional work was done to model the location of Celtic and Gallo-Roman roads where the exact location of segments is not known.

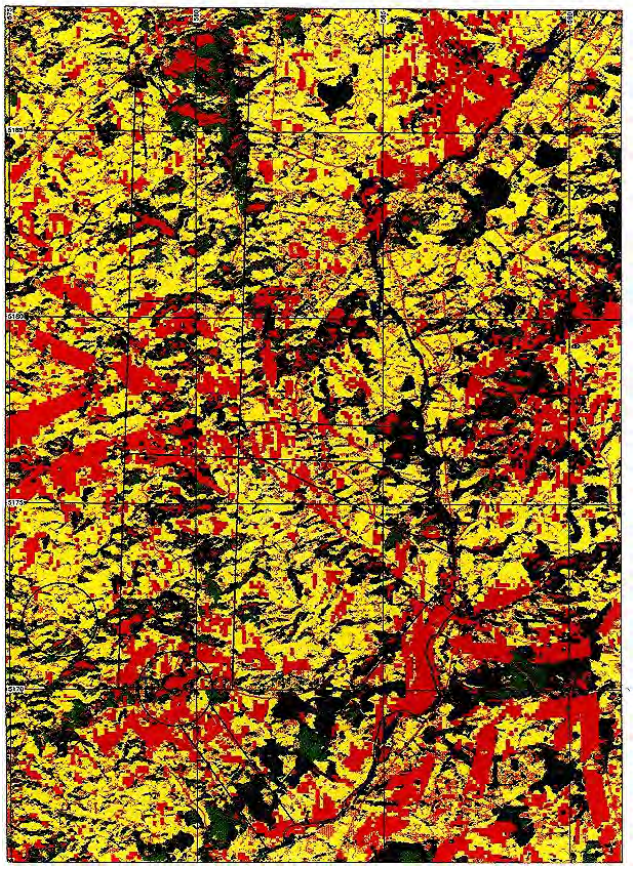


Predictive Modeling

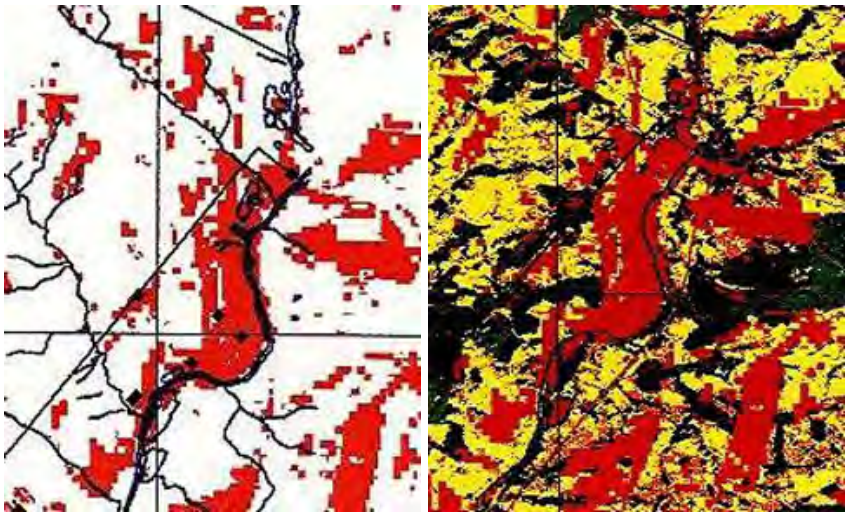
Site location modeling is a useful GIS product for archaeologists who wish to locate and protect unknown cultural resources. It allows us to model where archaeological sites of a given period ***may*** be located, based on the known site locations and various environmental and cultural data in the GIS. Predictive models were developed on this project using site data generated from field surveys conducted in 1978 and 1979. A model based on environmental and cultural data was created that accounts for 78.9% (45 of 57) of all Gallo-Roman sites in only 29.2% of the total area that was field surveyed. The same model also includes 69.2% (36 of 52) of the Iron Age sites and 80.3% (49 of 61) of the Medieval period sites located in the field survey. This model was then generalized to include a much larger area surrounding the transects (4.7 times as large). New layers in the GIS containing these locations were produced along with new maps showing the areas with the highest probability of site locations. These areas of higher probability of archaeological sites have a high correlation with areas that are threatened by current gravel mining activities in the area.

Current research involves analysis of aerial photography and site surveys using GPS in the areas predicted to have higher site potential by the predictive models. Coordinates of high probabililty areas are entered into GPS units, which are used in the field to search for the exact location of the archaeological sites. GPS receivers are also used in the aircraft while conducting aerial prospecting and photography. The GPS allows us to locate our position and the location of the photographs much easier than looking on maps for individual fields while the aircraft is banking and turning.



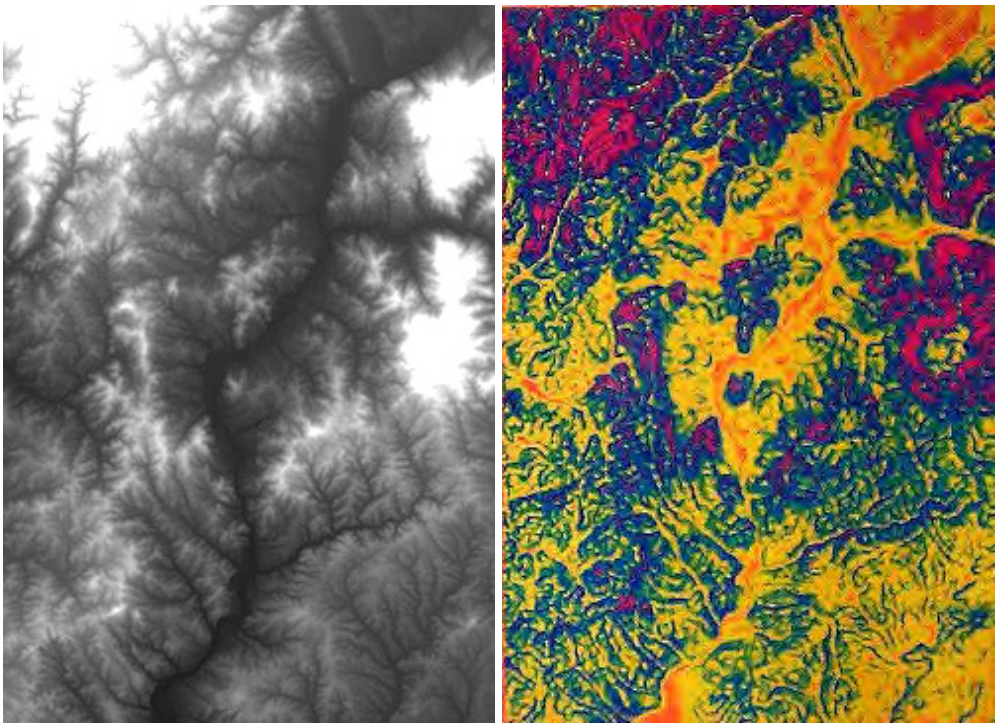


These Roman sites tend to cluster along the rivers in flat bottomlands, and also along the Roman roads. Celtic living sites tend to be more in uplands and near the Celtic hillforts on the hilltops. The largest single area of high probability is the area shown below, next to the river, where the Roman villa was located and later destroyed. The gravel operations continue to work in the area, and researchers are trying to locate additional sites using these techniques before they are destroyed.



Visualization and Simulation

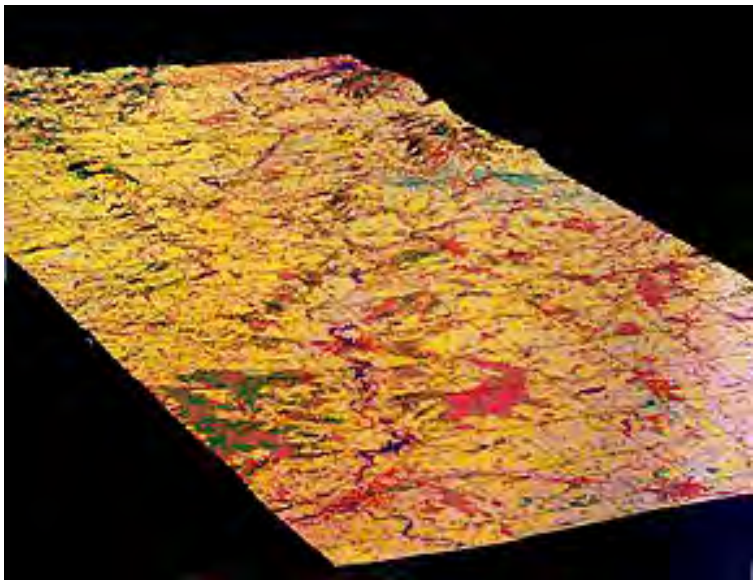
This project has used a variety of visualization tools to assist in better understanding both the data and the region. New technologies allow us to view our data in interesting ways from the laboratory when it is not possible to be in the field. For example, we can "fly" through the study area and look at different GIS layers draped over the terrain. First we need a digital elevation model (DEM) showing the topography of the region. For this project this was generated from the French 1:25,000 scale topographic maps.



Shown on the left above is the gray-scale DEM image of the Arroux river valley. Lower elevations along the river are shown in black, shading to white which represents the highest terrain in the region, including the site of the hillforts. From this a slope steepness map, seen on the right, can be derived using colors to indicate small intervals of slope values in degrees. These data can be displayed in several ways: one is the SPOT image of the study area with topographic effects superimposed, as we saw near the top of this page.

Drawing upon an appropriate topographic rendition, we can create perspective views that ultimately can be combined to produce a 3-D fly-through of the valley (described below), which shows the region in three dimensions as we fly up the river. We may choose to overlay any individual or combination of GIS data layers in these views.

As seen from an oblique aerial perspective, the GIS scene for an instant might look like this:



To generate the fly-by, we have applied a program that uses input data such as we have created to generate a movie-like run through the scene. Quicktime Virtual Reality (QTVR) lets us pan and zoom interactively on our computer (or over the web) to see the actual view from given locations in the study area. This is very useful in determining the accuracy of Line of Sight GIS computations, and also in allowing people to experience a virtual visit to the research area on the Web. To view the Quicktime VR movie you will need to download a free plug-in available from Apple Computer (the software runs in both Macintosh and Windows). For futher information, click on the Visualization word found on the Home Page of the Web site set up by this research team. That site also has background that supplements the overview on this Tutorial page and has the .mov file. It is your choice as to trying to download that file and run the movie. Unless you already have Quicktime, that program takes up to an hour to download.

The visualization is outward from the perspective centered around the summit of the Celtic hillfort of Mt. Dardon you have seen in the images. You can see the commanding view this hillfort had of the surrounding terrain, and portions of the Celtic ramparts which have survived intact.

Aerial imagery, and even more, GIS data are somewhat difficult for many people to relate to. They are abstract and it is hard for many people to relate what they see on the computer screen with what is actually on the ground. Visualization tools like QTVR allow us to bridge the gap between the imagery and reality, providing the ability for us to 'visit' the research area and compare the actual landscape with what we see on the various images and GIS layers.

Conclusions

This project has used a variety of tools to study the regional archaeological patterns of a large area in France. Particular emphasis has been placed on finding archaeological sites and then predicting the location of unknown sites before they can be destroyed by modern land use practices. It is the combination of new technologies and techniques, such as aerial photography, GIS, GPS, remote sensing, and visualization, that provides the greatest value to the researchers. Work on this project is continuing.

For more information off the Internet:

Time's patterns pieced An article for the UNC Endeavors magazine about this project

Satellite remote sensing for archaeology

Aerial archaeology An excellent overview of aerial archaeology by one of the best sources

GRASS visualization Includes links to other visualization programs

USGS GIS Provides an overview of GIS

Aerial archive links (U. Wien) An rich site dealing with archaeological topics, including use of aerial and satellite imagery

Bibliography for Burgundy Archaeology Study:
Papers by Scott Madry (co-authors indicated)

Line-of-sight and Cost Surface Techniques for Regional Archaeological Research in the Arroux River Valley, (with L. Rakos), 1996, in: New Methods, Old Problems: Geographic Information Systems in Modern Archaeological Research, edited by H. D.G. Maschner. Center for Archaeological Investigations, Occasional Paper no. 23, pp. 104-126.

Applications of Digital Terrain Data, Archival Maps, and Historical Aerial Photographs for Regional Archaeological Analysis, (with R. Wincek), April 2, 1995, paper presented at the 23rd annual meeting of the Computer Applications in Archaeology conference, Leiden, the Netherlands,

Un Emploe des Systemes de Teledetection et d'information Geographique dans une Analyse Regionale de Repartition Spatiales Archaeologiques, 1991 In: Archeologues et Ordinateurs, Lettres d'Information du Centre de Recherches Archeologiques. 17:1-11. Paris: Centre Nationale de la Recherche Scientifique.

An Application of Remote Sensing and GIS in a Regional Archaeological Survey (with Carole L. Crumley), 1990. in Interpreting Space: GIS and Archaeology, K. Allen, S. Green, and E. Zubrow, eds. Taylor & Francis, London.

Application des Systemes d'Information Geographique Informatiseé et de Teledetection en Releve Archeologique Regionale Bourgogne, France, et Caroline du Nord, USA., 1990 in Journees d'Etude Internationales: Teledetection et Cartographie Thematique en Archeologie, Valbonne, France CNRS/CRA.

Remote Sensing and GIS Analysis in Large Scale Survey Design in North Mississippi, Jay Johnson, S. Madry, and T. Sever, Vol. 7, Number 2, Winter 1988, pp. 124-131 Southeastern Archaeology.

Remote Sensing in a Temperate Regional Archaeological Survey in France, 1987 in Regional Dynamics: Southern Burgundy from the Iron Age to the Present, edited by C. L. Crumley and W. H. Marquardt-published by Academic Press.

Téledetection,(Remote Sensing), 1986, in Vol. No. 56 pp. 2-6. in Echos Du Passé, Revue Periodique de l'Association Les Amis Du Dardon, Gueugnon, France.

La Grenouilliers Perdue, 1985, Vol l. No. 54. in Echos Du Passé, Revue Periodique de l'Association Les Amis Du Dardon, Gueugnon, France with Dr. Carole L. Crumley.

Remote Sensing in Archaeology, in Archaeology, May-June, l983 (l8-l9). Reprinted in French in Echos du Passé, Bulletin du Les Amis du Dardon, Gueugnon, France, l985.

Primary Author: Dr. Scott Madry email: madry@informatics.org
Informatics International, Inc. and Dept. of Anthropology, UNC-CH

PREVIOUS NEXT

In 2005, another comprehensive example of the use of GIS, this time with aerial photography rather than satellite imagery, was found on the Internet. In this study, a Brown University student's undergraduate thesis, conditions favoring surival of an increasingly rare wildflower on Block Island, RI are defined and mapped using GIS.

A GIS APPROACH TO PRESERVING A RARE FLOWER HABITAT IN BLOCK ISLAND, RHODE ISLAND

This page will try to reconstruct an innovative application of GIS to a very specific topic. The work reported on was part of a Senior-Honors thesis done by Matthew A. Vadeboncoeur at Brown University who graduated with a Sc.B in Environmental Science in 2003 . His thesis title is *Using GIS to prioritize Land for Management in the Conservation of a Rare Species: A Landscape-based Metapopulation Method for Northern Blazing Star on Block Island, RI*. His goal was to identify areas on the island of optimal suitability for preservation of the wildflower as an aid to decision making by the citizenry and the Nature Conservancy regarding future land development. (Mr. Vadeboncoeur was contacted by email to invite him to prepare this page and to give permission to use his work, but no reply was forthcoming; therefore the write-up comprising this case study is solely the effort of the present writer [NMS].)

Liatris scariosa var. novae-angliae is a beautiful wildflower whose English name is Northern Blazing Star found over much of New England but which is disappearing from Rhode Island. Block Island, RI is still a favored habitat. Two views of this lovely flower are seen below:

The flower favors grasslands and meadows, and at times near swampy growth. Here is a typical habitat on Block Island and a picture showing its occurrence in grassweed growth.

Northern Blazing Star on a Block Island footpath

The entire state of Rhode Island is shown in this Landsat image. Block Island lies within Long Island Sound about 20 km (12 miles) south of Naragansett Bay.

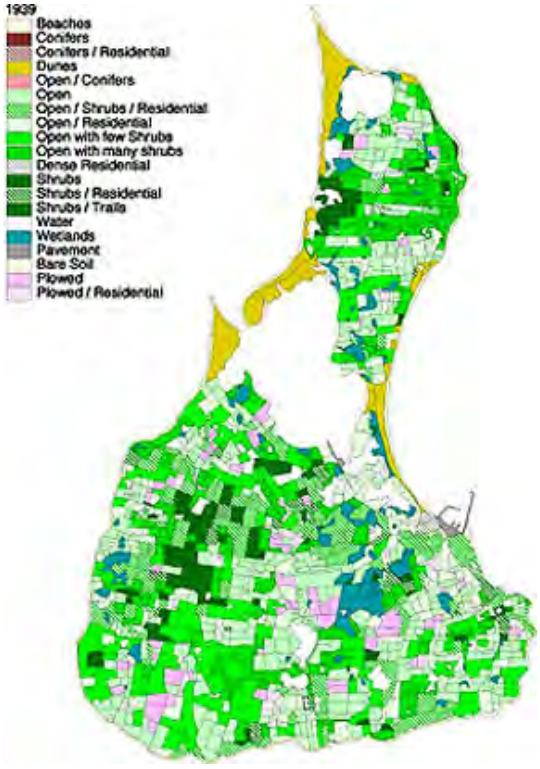
An aerial oblique photo shows the entire island. Beneath that is a map of the island, which is about 11 km (7 miles) long

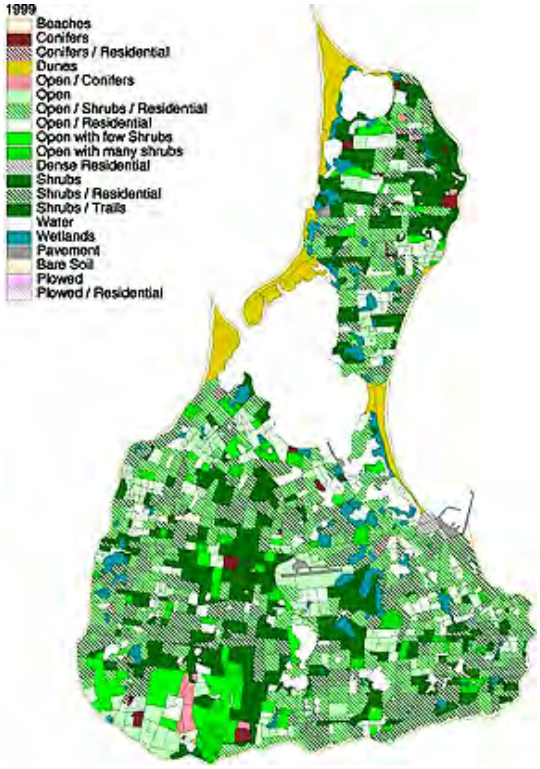
Block Island is a smaller version of Martha's Vineyard in Massachusetts in that it has permanent residents, residents in summer homes, and many tourists. This view shows much of the island (in winter) including the Great Salt Pond with most homes near shore but some inland.

The next three pictures help to give a flavor to the setting and scenery of Block Island. Read their captions

The material exposed at the cliff in the second picture is stratified to unstratified glacial drift. Block Island is the remnant of a glacial moraine deposited in the Pleistocene. The island's soils are thus derived from transported drift.

As is characteristic of most of southern New England, there has been significant, sometimes drastic, conversion of land use as more people flock to seashore environments for part to full time living. Compare the land use maps for Block Island for 1939 and 1999, a 60 year span.



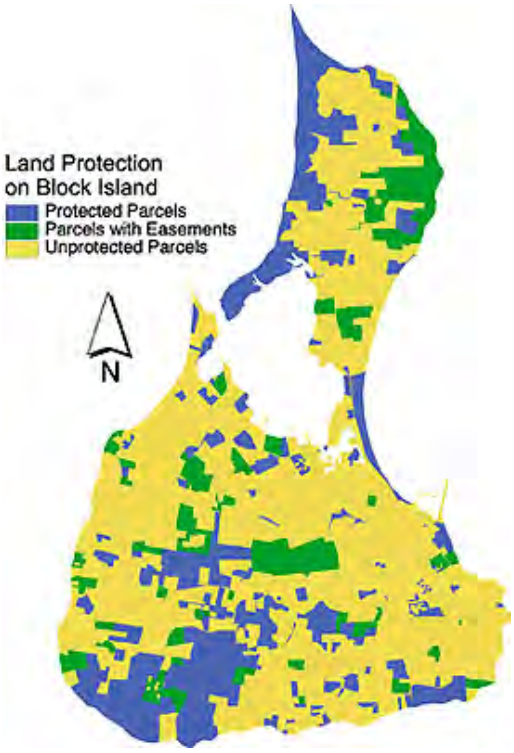


As expected, the largest change has been the conversion of natural cover to residential land and associated landscaping. This of course is encroaching on the habitat of the Northern Blazing Star:

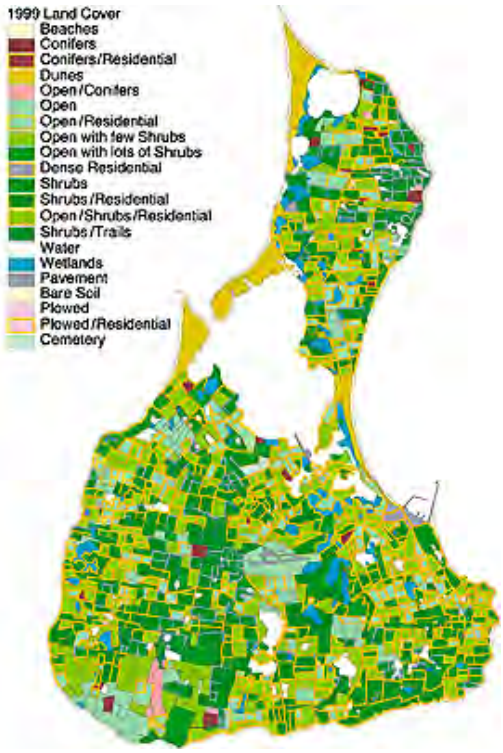
Northern Blazing Star habitat and encroaching development on Block Island

Matt Vadeboncouer approached the problem of selecting and recommending parcels of land for limitations on residential use by using a combination of false color aerial photography, maps made by Soil Conservationists, and field work.

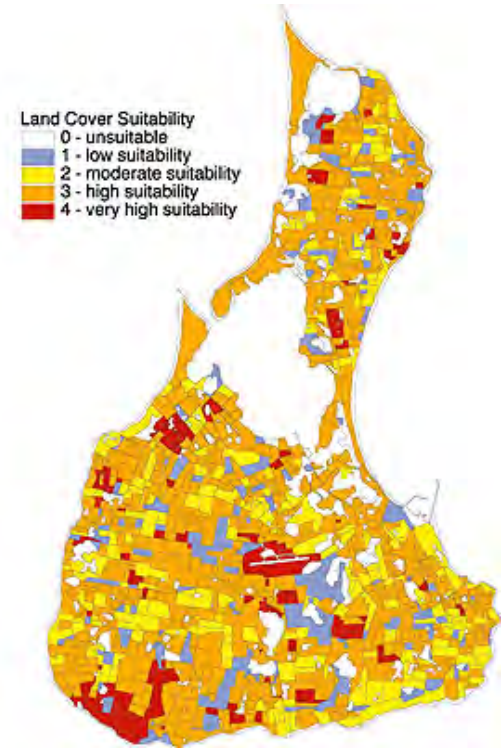
Primary input maps were made using the ESRI ArcInfo GIS system. The first of these digitized a general rating of the Land Protection status as of the start of the project.



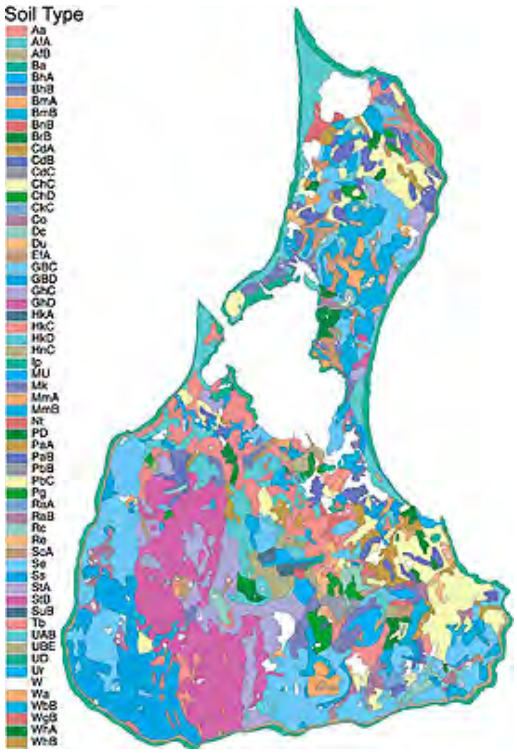
The second map is a detailed reclassification of land use in the early 2000s made by Vadeboncouer (using 1999 data).



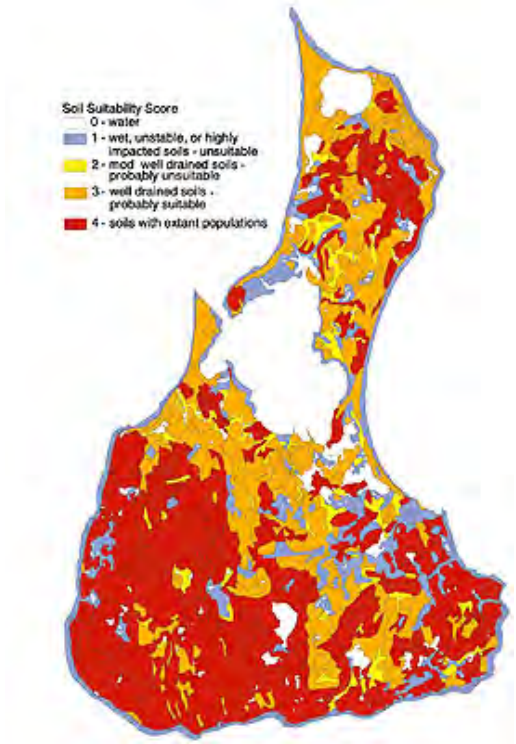
Next, a land cover suitability map was generated, in which the highest suitability pertained to favored Blazing Star environments.



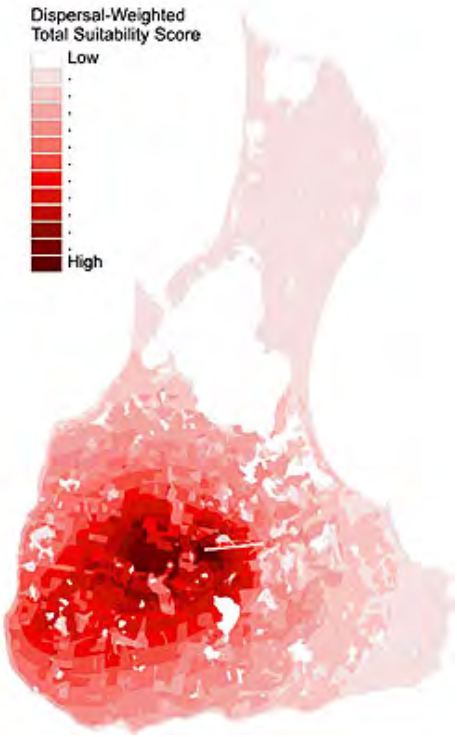
Next, a soil classification using earlier work was put into the GIS theme map base.



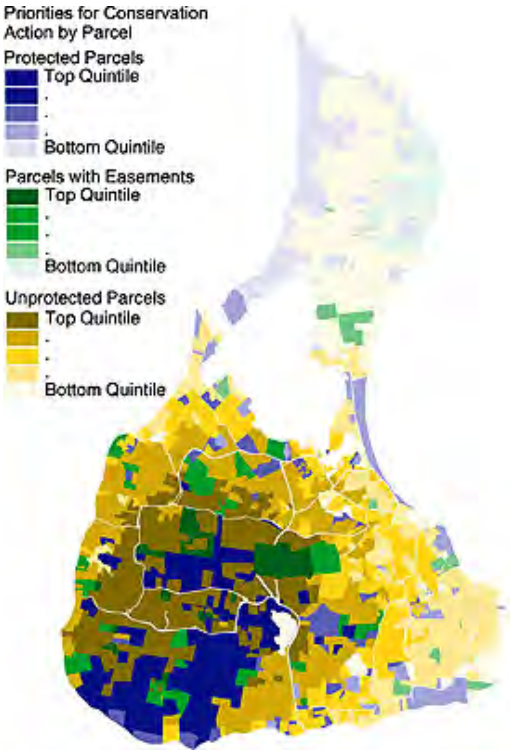
From this and other data, a soil suitability map was derived in which category 4 indicated soils best for supporting Liatris scariosa; category 3 was also within an acceptable level.



Using appropriate probability functions, a Total Suitability Score was assigned to the cells containing the land parcels.



From these and other map inputs the final output was a map denoting Conservation Action by Parcels. Those in blue identified land where the island decision makers could choose to limit development in order to better preserve and manage natural conditions including those that protect continued presence of the cherished Blazing Star plants.



The map below is not from Vadeboncouer's study but is a recent one referred to by the developers for managed land use and open space which could eventually be reallocated to human use.

This study is both a tribute to the capabilities of GIS and to how a college student's work can actually be turned to benefits for one's fellow citizens while maintaining aspects of the natural environment.

15-26: Which input map benefited the most from remote sensing; do you think the Northern Blazing Star will be saved?ANSWER

PREVIOUS NEXT

Applied remote sensing is (or can be) a key component in various types of decision making. As an example, consider how one goes about the process of converting now vacant natural landscape into an extensive living community: this includes housing developments and shopping and service centers plus sites for industry and transportation and utilities infrastructure. Much of the information needed is in the form of thematic maps (soils stability; topography; proximity to roads, etc.) and various data sets (land ownership; sources of water; local economy, etc.). Over the past two decades, a new systematic approach to gathering, storing, and manipulating the different information types, with the objective of analyzing the tasks and making and implementing decisions, has become highly popular and widespread in use: this is the methodology intrinsic to what is known as GIS or Geographic Information Systems. GIS can be described as an approach suited to extracting data and information from maps, aerial and space imagery, commonly coupled with locational ground data, that can be used to make decisions about applications of the recovered information to a wide range of practical utilizations.

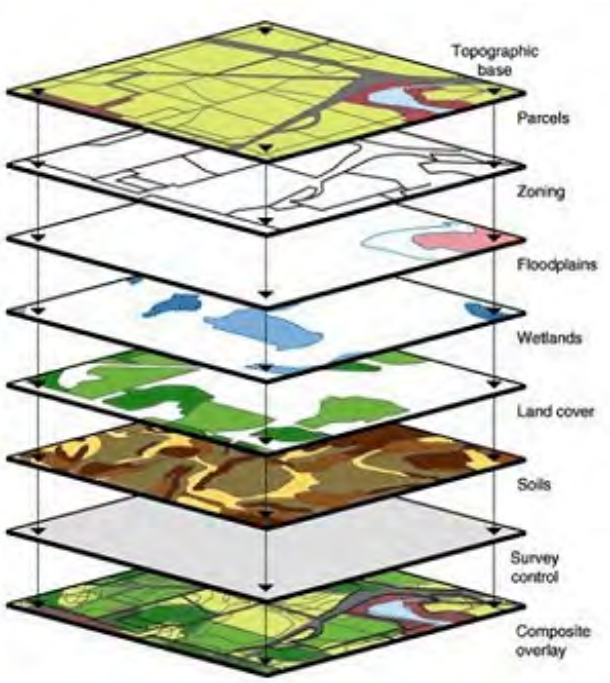
Remote sensing plays several roles - an obvious one is to upgrade certain maps that are outdated by classifying the most recent scene content into classes needed in any GIS analysis. This Section will explain in some detail the functions of a GIS analysis, will show how imagery like Landsat can be a vital part of the process, and will use two applications to illustrate the general end results sought. The first page focuses on some types of thematic maps in common use.

GEOGRAPHIC INFORMATION SYSTEMS (GIS)

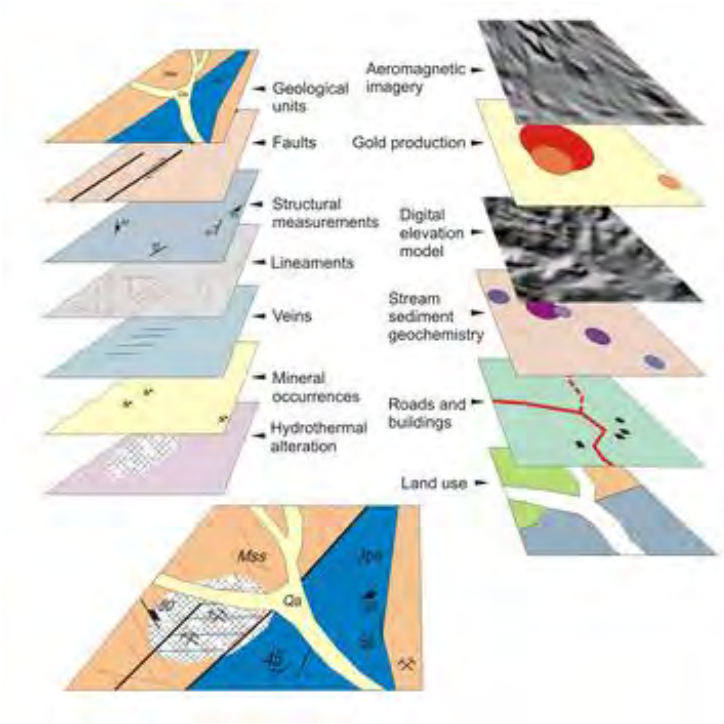
Opening Remarks

As a straightforward and concise definition, a **Geographic Information System (GIS) is a methodical approach (largely computer-based) to problem solving (decision making) by capturing, organizing, merging, analyzing, and applying diverse kinds of data that can be displayed on maps covering a geographical area of interest which therefore shares the same set of specified locational coordinates.**

Before we begin the process of learning what a GIS really is and does, this simple example of using maps to **make decisions** should give you valuable insight. Suppose you are given the task of determining where, in a particular region, a new town should be developed from scratch in the wilderness (but not too far from civilization) that will be used to house people with various connections to opening a major new mine extracting gold ore. After pondering the problem, you conclude that the best way to start would be to acquire some maps showing different types of relevant information that include the areas near the eventual mine. These maps display a variety of **themes**:



Some of these maps are likely to be suited to details concerning one theme, such as the regional geology:



You also deduce that you will need other kinds of information, such as on file in government offices. This ancillary info could be land ownership, available sources of power, sewage, etc, highway maintenance, and the like. When all this is collected, what do you do? You lay out the maps, become familiar with their contents, and start looking for potential sites proximate to the future mine. You shift back and forth from map to map trying to find appropriate clues as to where the best site(s) would be. This requires a sort of mental overlay process, which is difficult. After narrowing possibilities, you then feed in pertinent ancillary information. Finally, you arrive at a decision, but this is still subjective and based largely on a qualitative assessment. The whole process is cumbersome, tedious, and generally inefficient.

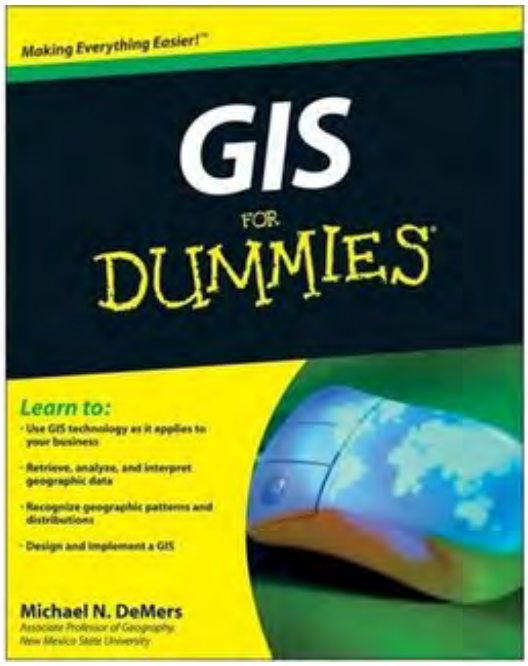
But what if there is some kind of "automated decision making" that allows you to systematically combine the information in individual maps into an interrelated composite that handles diverse data, embellishes that with relevant ancillary data, and then uses a logical program that quantifies the process of final choice as to the most optimal location. That is precisely what GIS

is all about. GIS is the outgrowth of the computer age, and the ease with which data can be manipulated by scanning and other input methodology. GIS centers on maps and other kinds of spatial displays - the starting point in our discussion of this new technology.

As this Section unfolds, it will become obvious that remote sensing data and visual products, being largely geographical in context, can, and often do, play a large role in GIS, primarily as a source of new or updated information about relevant scene content. GIS, in turn, provides a systematic way of utilizing remote sensing imagery in practical applications in the real world.

Before you proceed through this Section, you could benefit from the Overview of the nature of GIS, especially as used by decision makers, by reading through this Wikipedia website. Another overview website has been put together by the USGS. A video describing what GIS is all about has been place on the Internet by ESRI, a leader in the field.

As this Section unfolds, you should become aware of the ever growing importance of GIS in government, business, and academic activities. At universities now, there are degree programs built around GIS. Perhaps the best proof that GIS is a major field of its own is implied in this next illustration:



Maps and Attributes

Before we examine GIS itself, you need to be reminded of the role that maps play in describing the Earth's outer layer(s). Throughout this Tutorial, we have presented photos and images of the Earth's surface and clouds, some accompanied by maps that locate, classify and describe objects thereon. All these objects share a common characteristic: they are visual, miniaturized representatives, or surrogates, of places and classes of features located geospatially on or above the surface, be it land or water. They thus depict aspects of the local, regional, or global geography, which we can locate geometrically in an x, y (horizontal), and z (vertical) plot. We can reference the plot, then, to some form of coordinate and projection system. A map projection is a specific way of transferring points or locations from a spherical globe onto a scaled-down flat surface (the map) according to a systematic, orderly realignment, using a latitude/longitude grid network.

Photos (images) and maps are inherently two-dimensional or planimetric, although we can use techniques, such as contours or shading, to present a quasi-three-dimensional appearance or to extract information about relative elevations above generalized datum planes. The most common type of three-dimensional map is the topographic map, which we already examined in

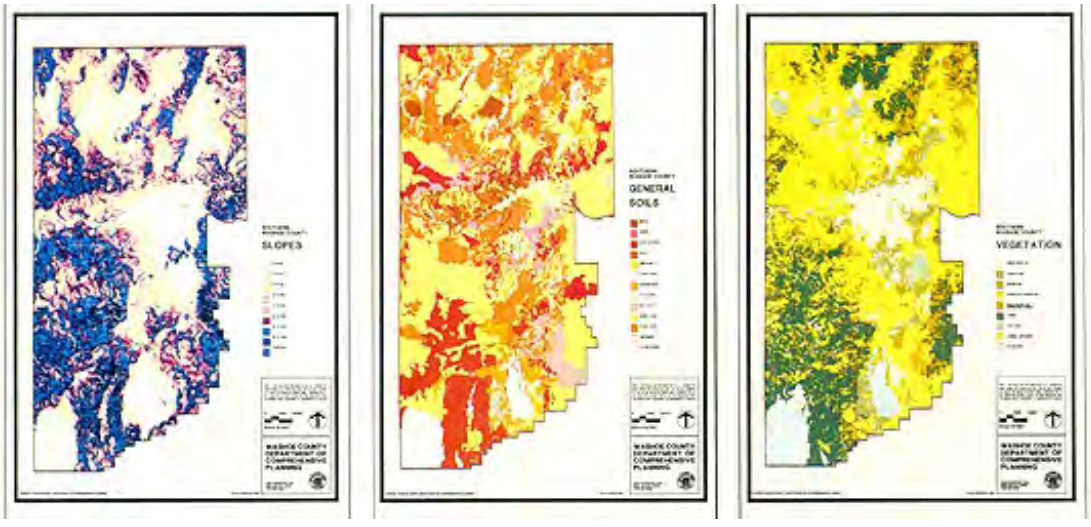
Section 11 (page 11-1). As we noted in Sections 10 and 11, maps in general nearly always have the following, essential information: graphic distance measures, scale, orientation and direction, projection type, and some geographic coordinate system, along with other descriptors and symbols in an accompanying legend. Topographic maps also have contour intervals.

15-1: What type(s) of map is/are the average citizen likely to encounter in everyday life?
ANSWER

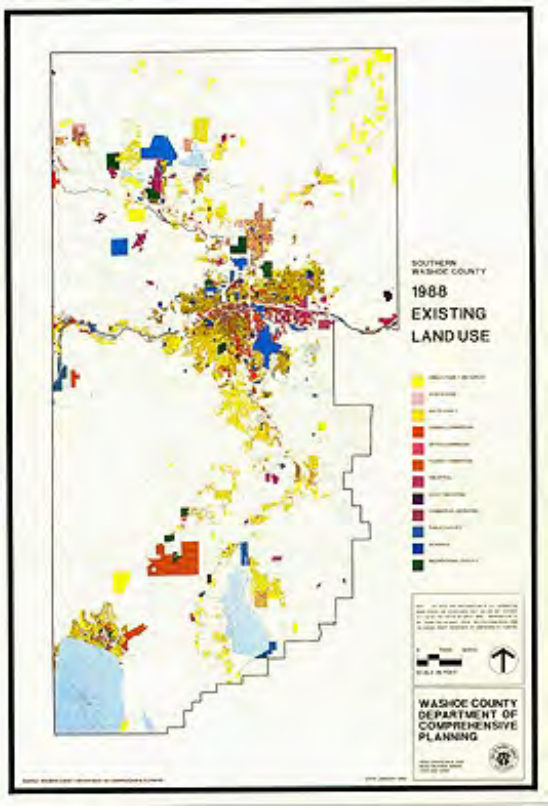
Maps are manmade derivatives made from aerial photos and space images or by surveying their features on the ground, designed to record information of various kinds about the spatial distribution of the specific categories of features present on the surface. A category can also be a theme. We can refer to other categories called attributes, which are general descriptors that are inherently non-spatial (they depend on characteristics rather than the location). A parcel of land, regardless of size, likely contains a diverse mix of features or characteristics that we can assign to different categories; in other words, it has many attributes. We may need a wide variety of thematic maps and associated attributes to fully describe the contents of a surface. Thus, given an area of, say, ten square kilometers, one map may display road networks, another vegetation cover, a third dwellings or functional buildings, a fourth engineering properties suited for excavation, and so forth. We may combine several themes on a multipurpose map, such as showing together roads, buildings, recreational areas, etc. One common type of map portrays land cover, which identifies all appropriate classes or categories that we wish to display within the map scale limits at selected points or locations. A variant of this map is the land use map that differs by detailing aspects of the cover involved in, or of interest to, human activities.

15-2: Think of (and list) at least four other types of specialized thematic maps.
ANSWER

Examples of land cover or condition maps, each devoted to a single purpose or theme, appear here for slopes, soils, and vegetation, distributed over the southern part of Washoe County in western Nevada (the city of Reno lies near the center):

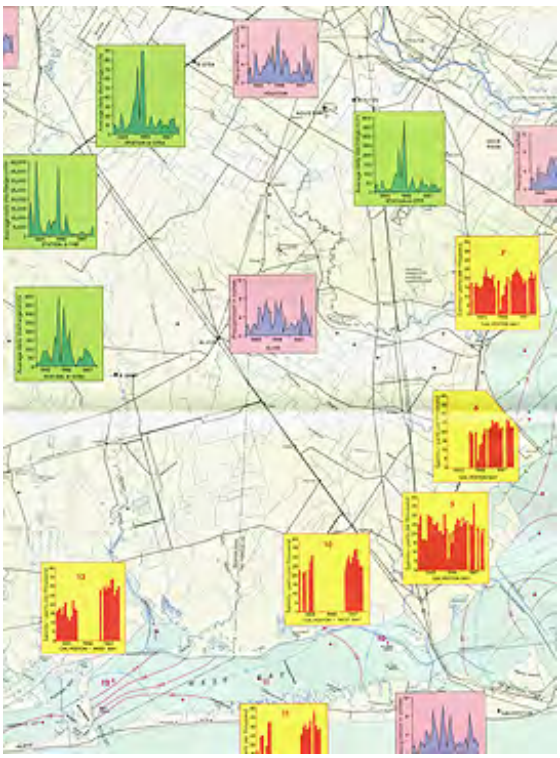


These maps describe natural states of either the surface or materials underlying it. (Note: as we reduce these maps to fit the web page, information in their legends becomes too small to read. However, the importance of the illustration is in the differing patterns rather than the specific identifiers.) Much of the surface is undeveloped, but areas where people constructed buildings, in which to live, work, manufacture, or play, comprise one form of land use, as shown in this map.



15-3: Suppose you decided you wanted to live near Reno, but you are now residing in the eastern U.S. Your great Uncle has died and left you a piece of land. Explain how you could use the above 4 maps, even though "in absentia", to make an educated guess as to problems you might encounter in building a house, so you can start planning even before you get there. ANSWER

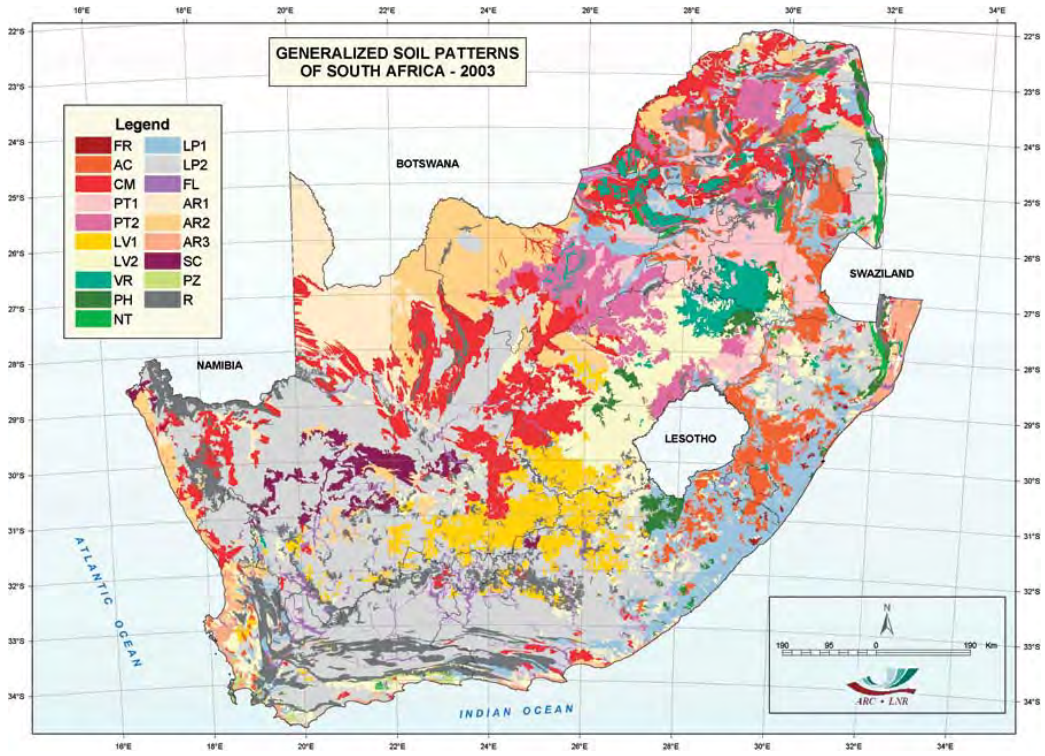
Some maps are composites of conventional, spatial distributions and graphs or other modes of data representation. Here is an example showing an area between Houston and Galveston, Texas, in which average monthly rainfall for the years 1965-67 at selected localities appears as bar graphs:

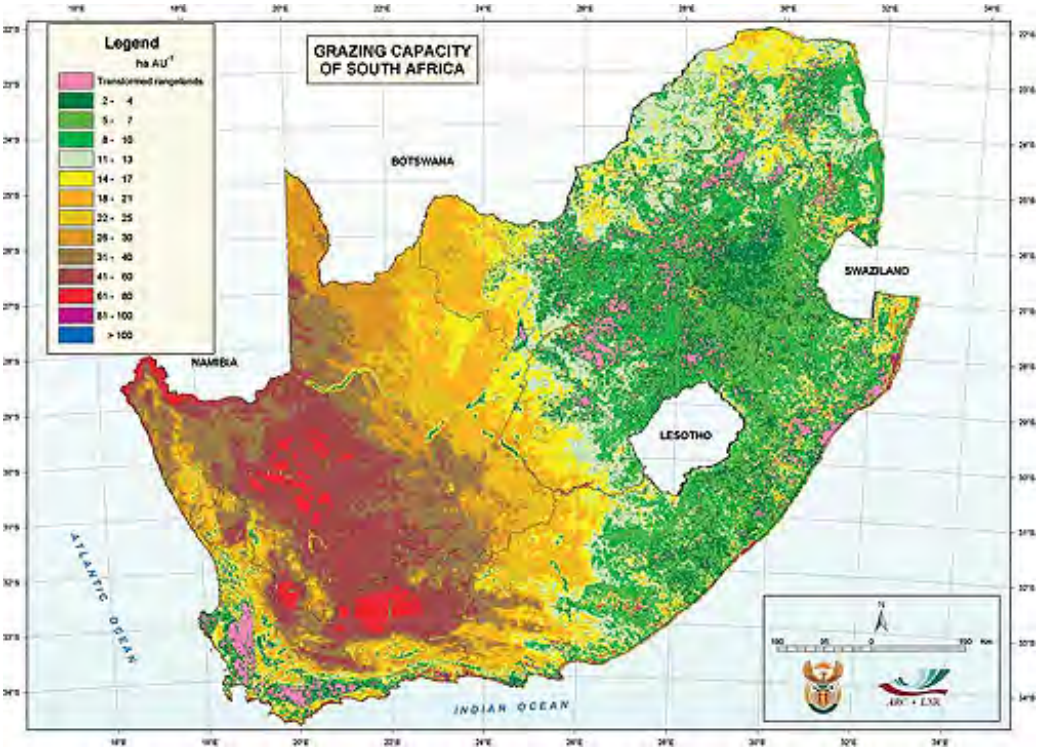


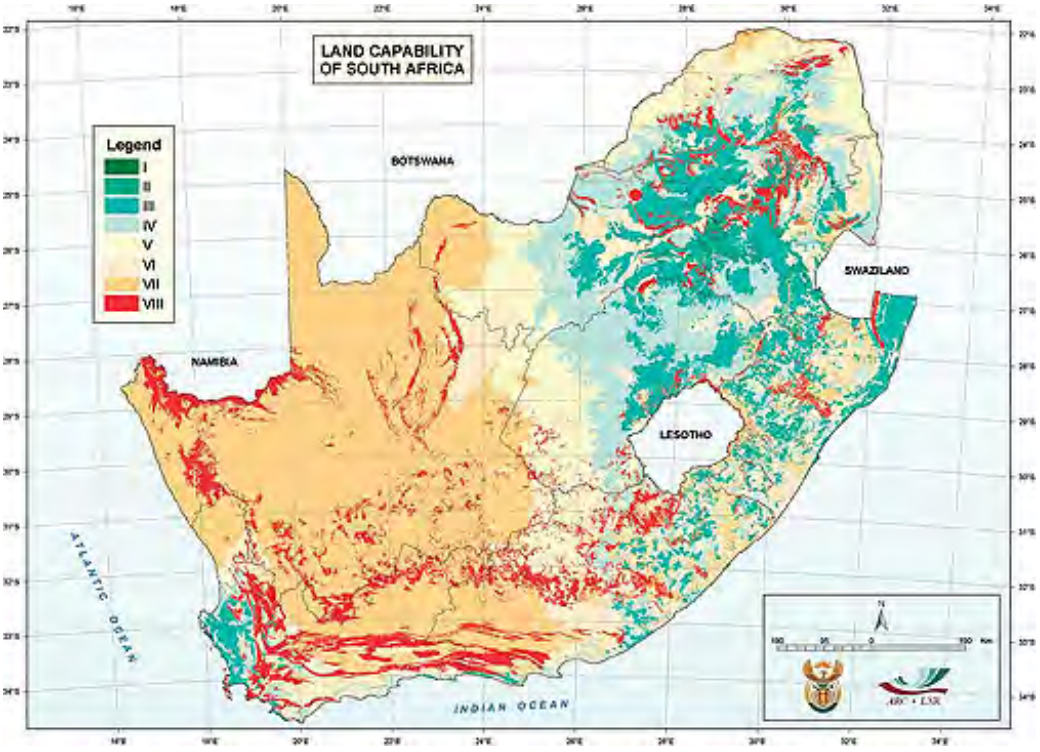
Thematic maps can cover small areas such as a town or township or county or they can apply to

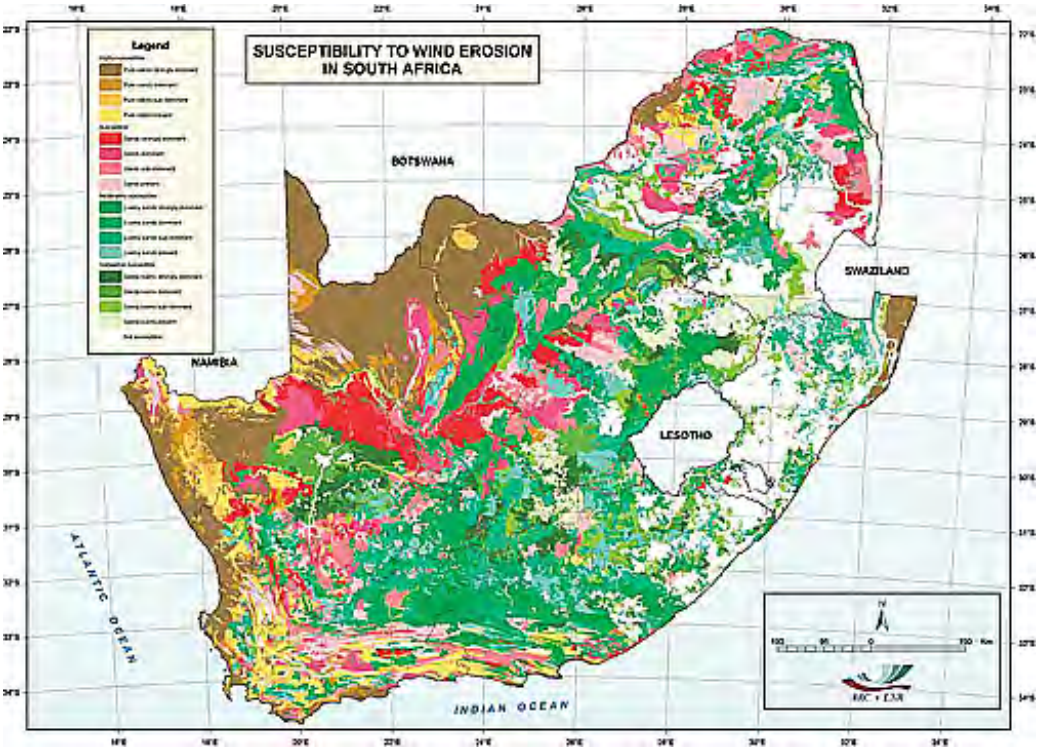
an entire country. Here is a Landsat TM image mosaic that includes South Africa and some neighboring countries.

The next four examples of thematic maps for South Africa are self-explanatory. Taken together, they could be used to decide where to set up farming centers devoted to animal husbandry.

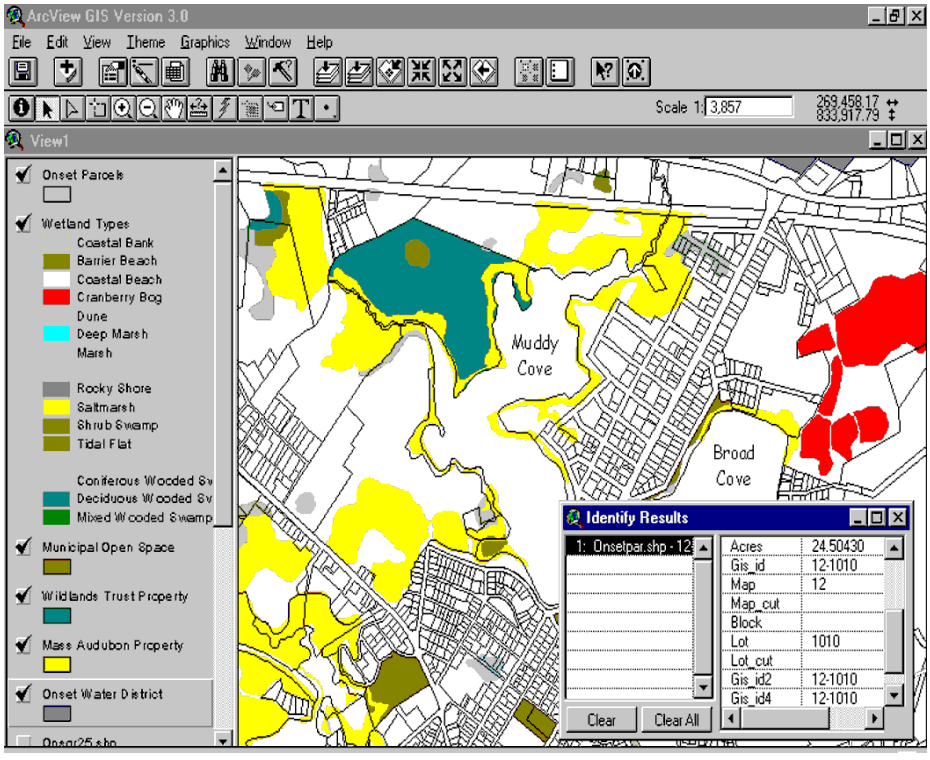



GRAZING CAPACITY
OF SOUTH AFRICA
Legend
ha AU-1
Transformed rangelands
2 - 4
5 - 7
8 - 10
11 - 13
14 - 17
18 - 21
22 - 25
26 - 30
31 - 40
41 - 60
61 - 80
81 - 100
> 100
BOTSWANA
NAMIBIA
SWAZILAND
LESOTHO
ATLANTIC OCEAN
INDIAN OCEAN
100 50 0 100 Km
ARC + LNR

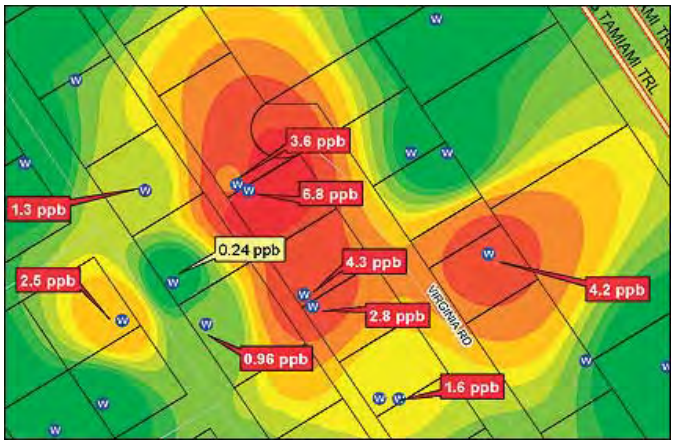

LAND CAPABILITY
OF SOUTH AFRICA
Legend
I
II
III
IV
V
VI
VII
VIII
BOTSWANA
NAMIBIA
SWAZILAND
LESOTHO
ATLANTIC OCEAN
INDIAN OCEAN
100 50 0 100 Km
ARC + LNR

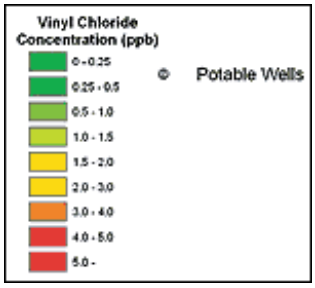


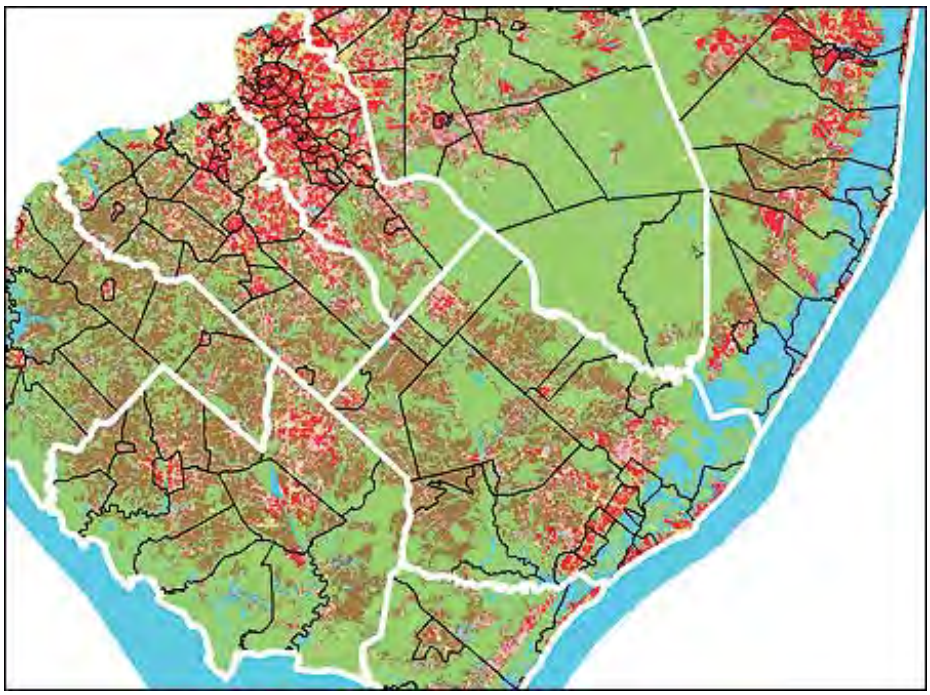
These maps were produced by the Agricultural Research Council - Institute for Soil, Climate, and Water, as displayed in Volume 19 of the ESRI ArcInfo Map Book Gallery . ESRI is the leader in developing and marketing software used to conduct a GIS analysis of interrelated thematic maps. A typical ESRI product is a map itself with accompanying information. Here is an example which is characteristic of an ESRI display:



ESRI's Map Book Gallery is worth your time to visit. It contains a wide variety of illustrations of different map types and uses. GIS is a large subset of the general way in which maps are used to store and present data that have geospatial significance. As will be shown later in this Section, a major objective of GIS is to use a wide range of geographically-based input to solve problems and make decisions. But often the maps produced can almost be stand-alone presentations of information that need not be tied to some practical outcome. To illustrate this idea, here are some typical maps from the Gallery which have worthwhile information *per se*; check their captions for specific descriptions:





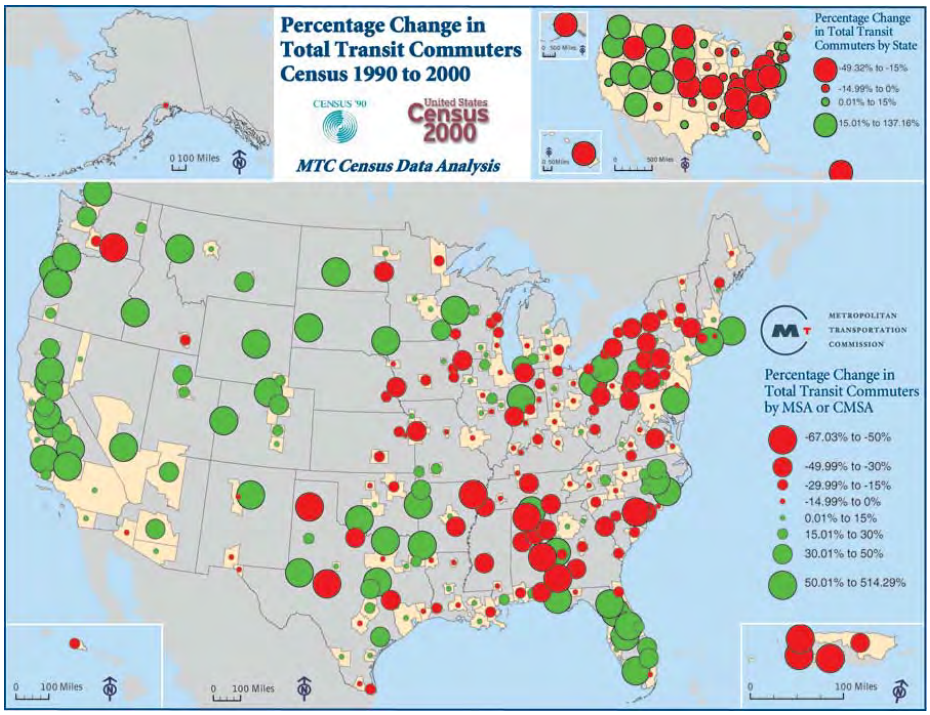


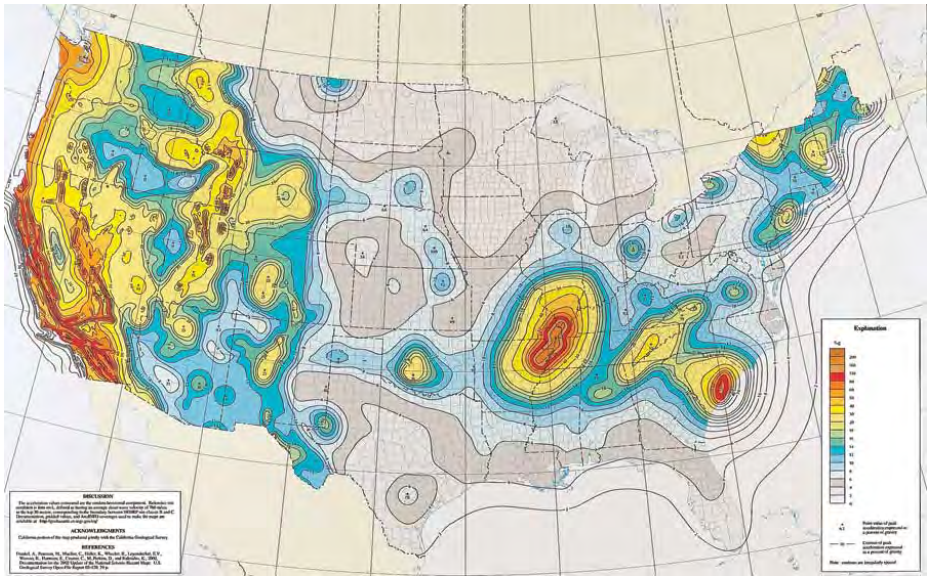
LEGEND

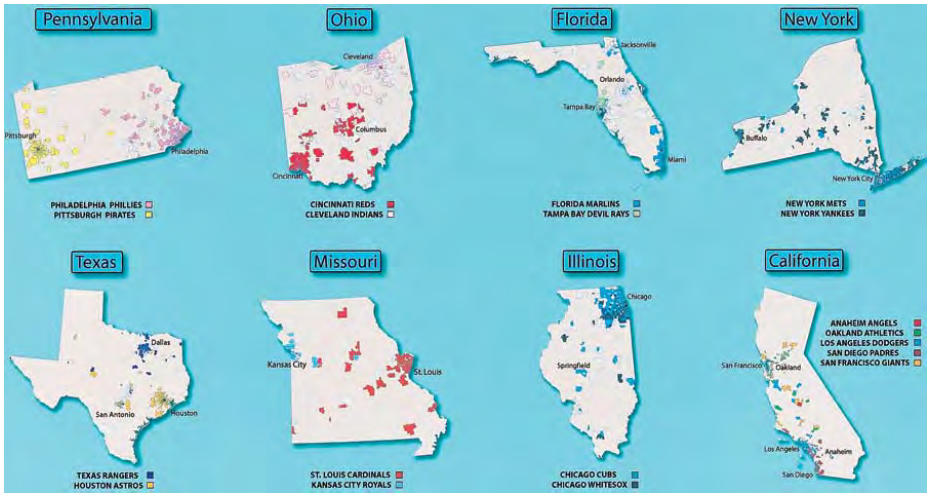
- County Boundaries
- Municipal Boundaries
- Water

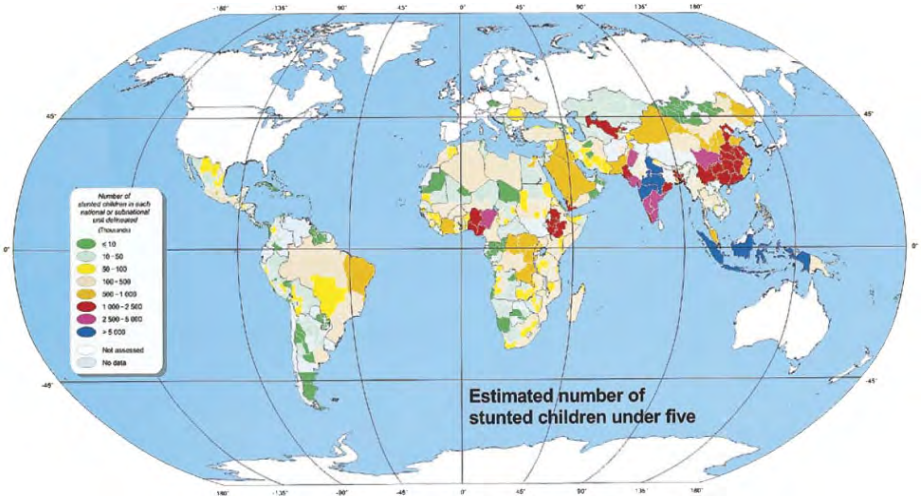
LAND LABEL

- Existing Residential High-Density Development
- Existing Residential Medium-Density Development
- Existing Residential Low-Density Development
- Existing Military Development
- Existing Commercial/Industrial Development
- Other Existing Development
- Environmentally Constrained
- Developable
- Possible Urban Developable
- NJMC Redevelopable


Percentage Change in Total Transit Commuters Census 1990 to 2000
CENSUS '90
United States Census 2000
MTC Census Data Analysis
Percentage Change in Total Transit Commuters by State
-49.32% to -15%
-14.99% to 0%
0.01% to 15%
15.01% to 137.16%
METROPOLITAN TRANSPORTATION COMMISSION
Percentage Change in Total Transit Commuters by MSA or CMSA
-67.03% to -50%
-49.99% to -30%
-29.99% to -15%
-14.99% to 0%
0.01% to 15%
15.01% to 30%
30.01% to 50%
50.01% to 514.29%
0 100 Miles




Pennsylvania
Philadelphia
Pittsburgh
PHILADELPHIA PHILLIES
PITTSBURGH PIRATES
Ohio
Cleveland
Columbus
Cincinnati
CINCINNATI REDS
CLEVELAND INDIANS
Florida
Jacksonville
Orlando
Tampa Bay
Miami
FLORIDA MARLINS
TAMPA BAY DEVIL RAYS
New York
Buffalo
New York City
NEW YORK METS
NEW YORK YANKEES
Texas
Dallas
San Antonio
Houston
TEXAS RANGERS
HOUSTON ASTROS
Missouri
Kansas City
St. Louis
ST. LOUIS CARDINALS
KANSAS CITY ROYALS
Illinois
Chicago
Springfield
CHICAGO CUBS
CHICAGO WHITESOX
California
San Francisco
Oakland
Los Angeles
Anaheim
San Diego
ANAHEIM ANGELS
OAKLAND ATHLETICS
LOS ANGELES DODGERS
SAN DIEGO PADRES
SAN FRANCISCO GIANTS



** Parts of this Section have been freely adapted from Activity (Chapter) 7 of the Landsat Tutorial Workbook; that activity was prepared in 1981 by Mr. William J. Campbell, at that time the Branch Chief of Code 935, the initial sponsor of the present Tutorial.*

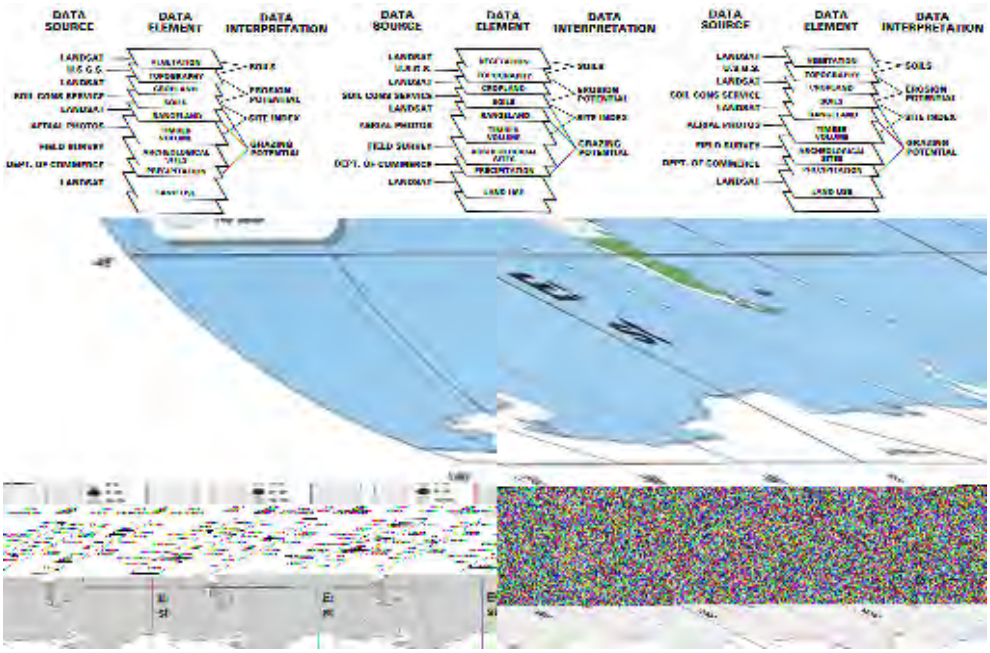
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The concepts of data elements, data interpretation, and data sources are explained. Although today multiple data element maps are commonly superposed (combined) by computer processing, analog methods can be used instead in a more limited way. The meaning of geospatial data types is considered.

Data Elements and Models

The interrelations among variables can be incredibly complex. To control the production process, we formulate a model for crop growth that accounts for the dynamic interactions and cause-effect responses of the appropriate factors. A model in this context establishes the functions, sequences, and feedback effects of the determining variables in the vigorous operation of a system.



As evident from the above diagram, we take information from various conventional sources, along with data from Landsat or similar Earth-observing satellites, and often combine everything to produce data elements, which we further convolve to produce interpretation maps and other output types. Then these outputs become viable parts of environmental resource planning, site selection, and other outcomes that benefit from modeling.

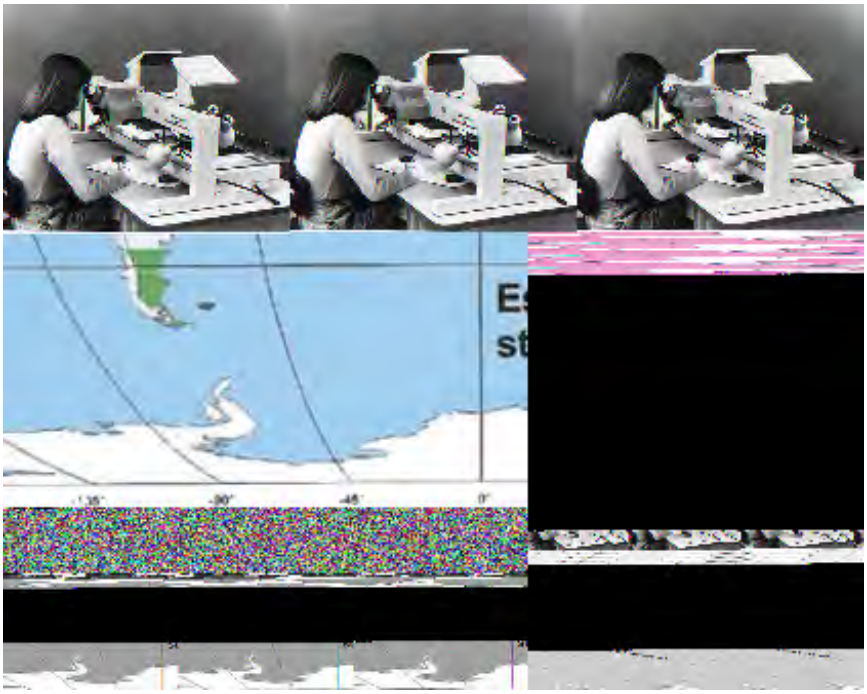
Interpretative maps are derivative, i.e., they result from decisions to produce new object categories, stemming from combinations of several others. For example, a map may plot the likelihood of runoff erosion, which we can deduce from maps showing soil properties, topography, rainfall, stream patterns, and other related factors. Variants of this map type are allocative or suitability maps (e.g., transportation routing), evaluative maps (e.g., earthquake damage potential), and predictive maps (e.g., 100-year flood coverage).

15-5: Let's say that you have, during one of your vacations in the western desert, found some enticing but still elusive evidence that an area you traversed during a hike had some unusual surface expressions of possible minable silver ore in a body near,

but below, the surface. State some appropriate steps you would take to end up with a Data Interpretation map that you could use to persuade some investors to help you "stake a claim" and then develop the ore body. ANSWER

Until the 1970s, the traditional way of using geospatial data, making decisions based on combining information sources, and applying models by laying maps on a table was unwieldly and vexatious. Interpreters inspected these maps individually, and mentally compared data sets - visual and tabular - while seeking to narrow parts of an area to locations that they pinpointed as suitable on several defining maps. In principle, they could physically overlay two or three maps, treating several relevant themes by placing them on a light table (particularly if they were rephotographed as transparencies) to check on pre-selected, favorable conditions, where patterns on the maps superposed in a positive way. This is, of course, inherently inefficient, and often the interpreter would simply glance back and forth between maps. One can improve the interrelationship process by laying a grid over each map and extracting information according to some relative merit or weighting, organized by location within the grid. Managing these data, usually in tables, proved labor-intensive, slow, and often ineffective.

Devices were developed to allow more systematic comparisons between maps. These comparisons occurred on a one-to-one basis by superimposing maps and photographs through an optical projector that allowed for differences in scale and even projection. One such device, still in use today, is the Zoom Transfer Scope (ZTS) manufactured by Bausch & Lomb.



The operator mounts a photo or a map at one level of the ZTS and manipulates mirrors and lenses so that its image, projected through prisms, appears to register (align) in the view scope over a second map or photo at the bottom. The operator then draws tracings from the first map/photo on a thin sheet over the second one or takes a photograph using a polaroid camera within the optical train, in which the two views merge. We show here an illustration of this, in which an operator initially rephotographed a standard road map (at 1:500,000), so that all black or colored lines appear white, and then projected it through the ZTS onto a 1:1,000,000 print of a Landsat false color composite, showing Philadelphia, PA, and Trenton, NJ:

Obviously, this form of image combination is severely limited to two images. However, one or more additional layers are feasible, if we display map and photo reproductions in other colors, but multiple registration becomes difficult.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The history and growth of GIS is examined and its principal phrases elaborated. The planning process by which GIS serves as the framework in problem solving and decision making is diagramed. The interactive components of Data Management within the GIS framework are also shown in a flow diagram.

The Development and Evolution of GIS

The maturing of the computer age has greatly changed the manner in which we can merge, compare, and manipulate multiple maps and other data sets. Computers and their software have significantly enhanced data handling capacity (capture, integration, flexibility). A powerful new tool, known as the Geographic Information System (GIS) emerged in the 1970s. Many of those who developed GIS were inspired by the 1969 publication of the classic ***Design with Nature*** by Ian McHarg (Doubleday/Natural History Press), a leading landscape architect then at the University of Pennsylvania. This document pointed the way to planning and decision-making through comparative, integrated maps and related data types.

McHarg's book inspired a young geographer, Jack Dangermond, to become one of the pioneers in GIS. He founded a company, ESRI, that has since become a world leader in this field. Dangermond came to NASA Goddard in the early 1980s to demonstrate this newly emerging technology to staff of ERRSAC, including the writer (NMS) at the time. He started with the basics. His "pitch" was remarkable - I was converted into an "instant fan". The case study he used - getting an activist group of citizens in northwest Colorado to agree on a master plan to develop the region - was totally convincing. My one regret is that my work mission never brought me into a full use of GIS, other than peripherally during the nuclear power plant project described later in this Section.

ESRI has grown into the industry leader in GIS technology and applications. A 2007 review of how ESRI's main product is applied is given by Jack Dangermond on an ESRI website.

Since its inception, GIS has become a major growth industry, now conducted worldwide at the multi-billion-dollar level. It has blossomed into the main way for using maps (novel and practical) in most endeavors that focus and rely on geographically-based data of many kinds. Because remote sensing has routinely provided new images of the Earth's surface, it too has become intertwined with GIS as a means to constantly and inexpensively update some of the GIS data (such as land use and cover). The Association for Geographic Information defines GIS as:

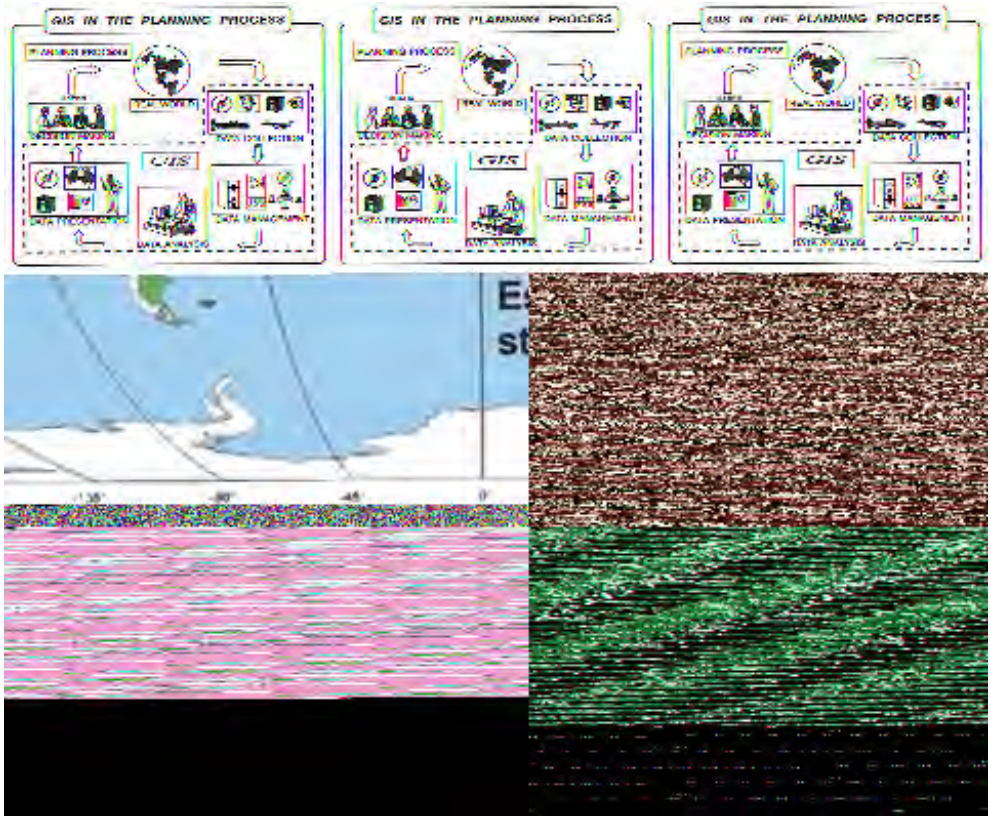
A system for capturing, storing, checking, integrating, manipulating, analyzing, and displaying data which are spatially referenced to the Earth' (usually) land surface.

A simpler working definition is: *A computer-based approach to interpreting maps and images and applying them to problem-solving.* The inclusion of computers to store, process, manipulate, interpret, and display GIS information is the critical ingredient that separate modern GIS from the more conventional (traditional) methods of using maps and correlative data prior to the 1970s.

15-6: In their book on GIS, J. Star and J. Estes (see references at the end of this section) mention the four essential "M's" involved in the use of GIS by planners, resource managers, scientists, and others. These are operations that are commonly carried out

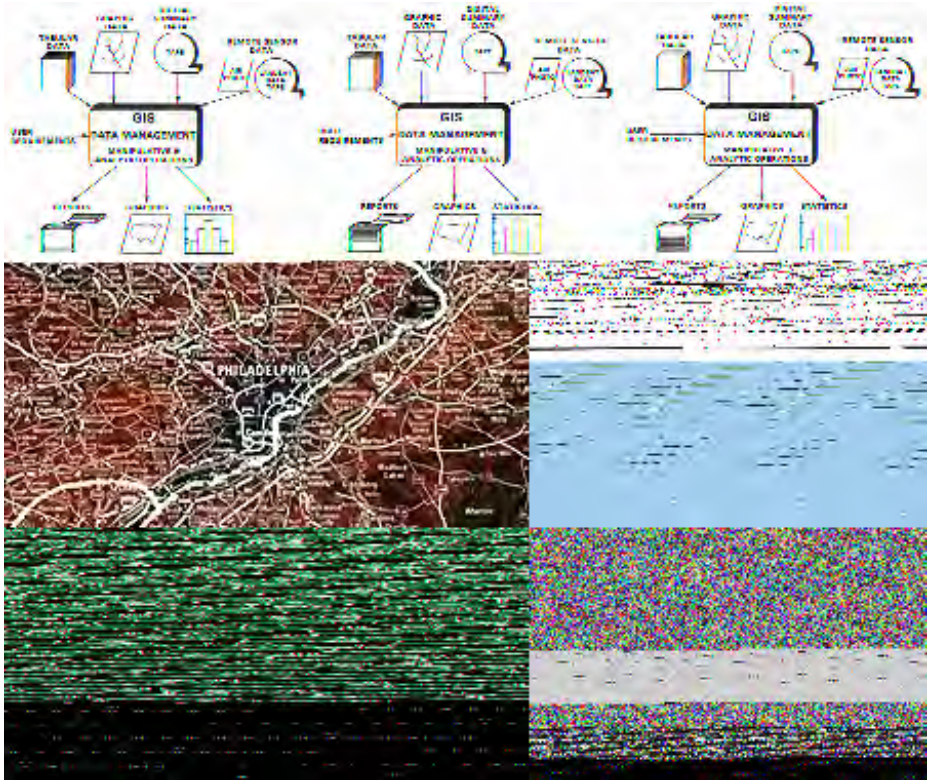
in succession. Try to guess what these M's are. ANSWER

We synopsize the role of GIS in the general planning process for site selection, environmental management, and other geographically-dependent applications in this diagram:



From B. Davis, GIS: A Visual Approach, ©1996. Reproduced by permission of Onword Press, Santa Fe, NM.

The driver for this closed-loop operation is the constant need for timely information about human activities and expectations concerning life in the real world. The specifics underlying those needs define the types and amounts of data/information required. Once GIS users stipulate the specifics, they collect the data from multiple sources, such as already published maps and tabulations, current field observations, surveys, and aerial/satellite imagery. In the next step they convert the varied data into computer-compatible formats. The heart of the GIS operation lies within various techniques for analysis that users have devised as GIS evolved. They then present their reports, displays, new maps, statistics, and other kinds of computer-based, information-oriented products to decision-makers (often, themselves). The test of value then happens by applying the results in the same real world that dictated the initial requirements. Data management through a GIS involves all of these facets:



15-7: What is the "driver" or key determinant in the above Data Management diagram?
ANSWER

PREVIOUS NEXT

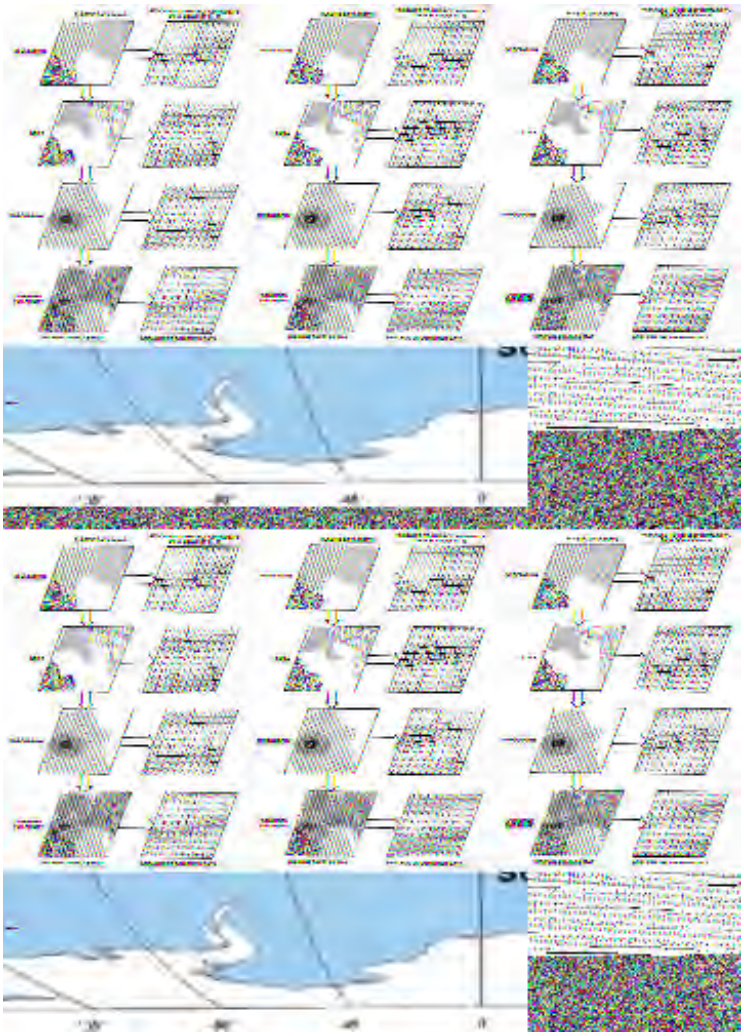
Primary Author: Nicholas M. Short, Sr.

This page begins with a quick look at some of the major Internet GIS sites that offer you much more insights into how GIS works, along with links to many organizations that deal with services and products or present case studies. Then, the main topic turns to one aspect of the decision-making process, using as an example the numerical evaluation of several thematic maps in terms of optimal conditions related to critical factors involved in choosing a site for a hypothetical project. These are combined into a suitability map in which the highest scores designates the best potential sites.

Decision Making: Suitability Determination

We embark now on further exploring what GIS does, how it works, what software systems are available, and what typical end products and applications look like. Before moving on, you may wish to gain further insights into GIS by visiting sites on the Internet. We have picked three whose links are given here. A useful summary is found at the GIS Development site. The leading U.S. software providers are Leica Geosystems and ESRI, which also offers training courses.

Perhaps the best way to appreciate the power of GIS, even before examining the design and function of the data handling system, is to introduce a typical case study and the thinking behind the steps in a site-suitability analysis. The following diagram summarizes the rationale behind such an analysis:



Suppose that three factors or variables, among the attributes that describe a geographic area under consideration, are essential in determining best sites for, say, a land development venture: Vegetation, Topography, and Soils. A data element on a map represents each variable. In this case each map shows the characteristics and distribution of the members or classes within the element theme, e.g., different types of soils and their properties.

The most obvious manipulation performed on each map was to assign numbers to parts of the map that are the weighted estimates of suitability, ranging from 0 (least suitable) to N (highest number for most suitable).

We usually designate one map, often a cartographic map, as the base, over which we lay the others, (each then constitutes a data layer) either manually or digitally. Now, some soils, vegetation cover types, and elevations are more favorable than others in specifying their role in the site-selection process. Thus, for a certain intended use, we prefer high areas over low areas. So, we assign relative heights numerical ratings, say from 1 to 5. We can mark soils with optimum drainage by higher numbers in a scale of 1 to 8. We then subdivide a data element map into cells in a grid. We assign each cell a value based on its thematic rating. We can incorporate other kinds of data, e.g., tables representing some condition in the cells, provided there is some spatial connection. When we combine the maps sharing the same cells, each comprising a data layer, the values sum for each cell (ranging from lowest numbers = worst suited to highest = best suited). In a modern GIS, we do this digitally. The outcome is a map, in which we judge areas with the highest resultant scores the most favorably suited.

15-8: The numerical data and their values need to be assigned to each cell. This is called encoding. For the soils map, as an example, suggest 5 ways in which this encoding can be done. ANSWER

Primary Author: Nicholas M. Short, Sr.

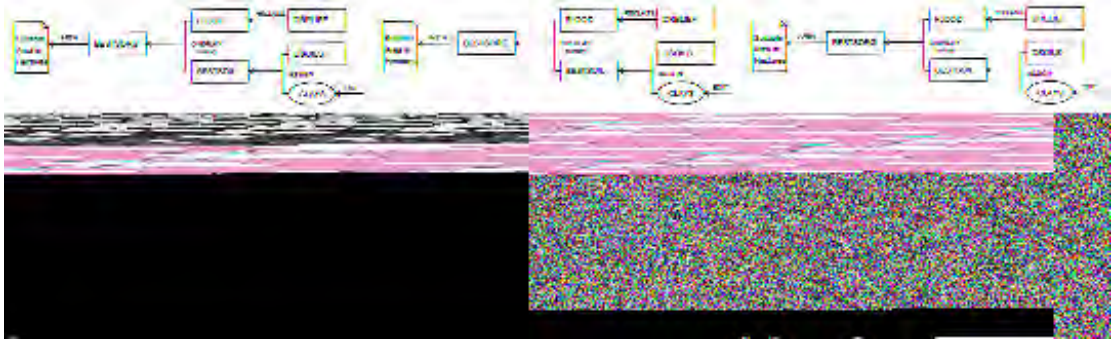
The time has come to run through an actual example of how the analysis and integration of thematic data, shown in the form of simple maps, will by appropriate combinations and numerical evaluation produce one of the most common outputs in a GIS study - a site suitability map. The IDRISI image processing and GIS analysis program are used to create this end result. The actual task is to pick the best location for growing sorghum (a crop) along the Senegal River in Africa. This is displayed in a map called BESTSORG.

A GIS Case Study in Africa

We extracted (with permission) the following from an example in the Database Query section of the Tutorial Exercises in the IDRISI for Windows Users Guide. This simplified example contains only a few of the many data elements that we normally consider using in the model. It does illustrate the workings and rationale underlying a GIS in operation. To manipulate data easily, it also relies on a different ranking mode within each element, which reduces to acceptable (1) versus not-acceptable (0).

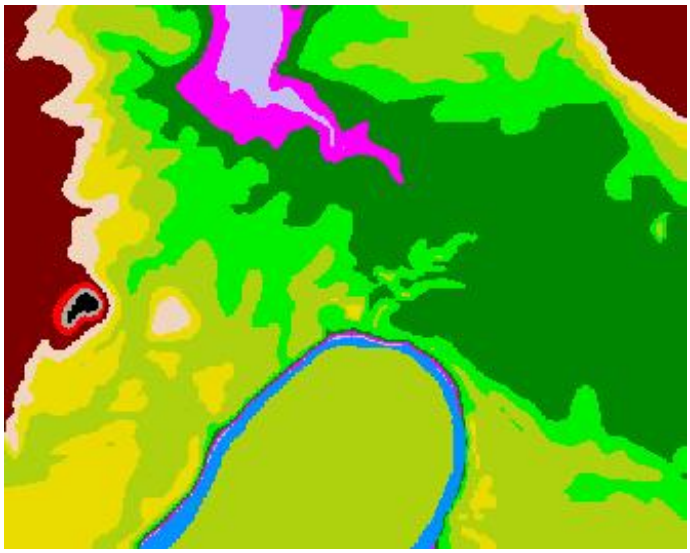
The problem is identifying the best areas for planting a specific crop within a valley along the north side of the Senegal River at the border between Mauritania and Senegal, two West African nations along the Atlantic Ocean. The cereal crop is sorghum, a tall grass cultivated in tropical climates that yields grains suited as silage and fodder or, in some varieties, a cane sugar-like juice called sorgo. Many of the factors in the list on the second page of this section, pertaining to agricultural siting, apply in this case as well, but here, we consider only key ones - soil types, relief, and periodic flooding (which promotes sorghum growth).

This following flow chart outlines the sequence of steps for using GIS to automate the decision about suitable areas:



(NOTE: In the graphics that illustrate these steps, we leave off the legend created by the IDRISI program, because there is no provision to copy it when the we convert the image to a ".tif" format. We identify the category color(s) for each graphic in the text.)

The first factor is the elevation variations or relief. Field surveys have led to a digital elevation model (DEM) data set, from which we produced this general topographic map image (three-meter contour interval) (DRELIEF in the flow chart is just the file name, as are the others):



(The sequence of increasing elevations, as color-coded is: blue (lowest, associated with the river), medium gray, purple-red, dark green, medium green, olive, yellow, tan, brown, red, darker gray, and black.)

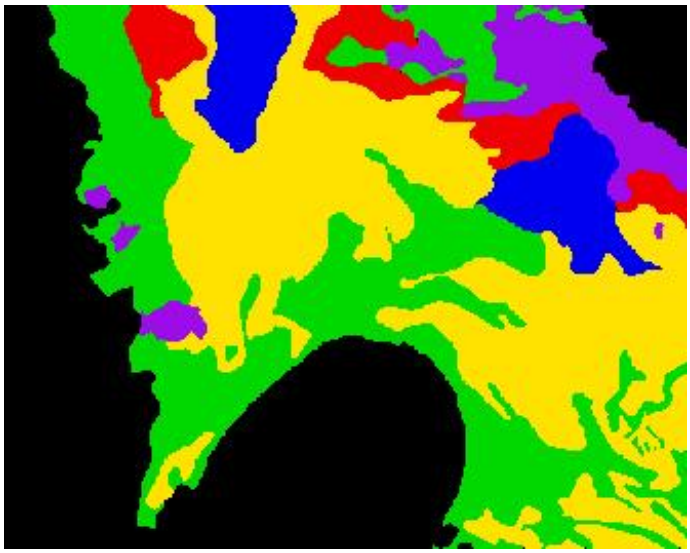
15-9: There seems to be something strange about this elevation map. The river is the lowest part of the scene, yet it appears to back up against the medium green pattern. What gives? ANSWER

The relief or range of elevations in this scene is 33 m (about 108 ft), stamping the local region as relatively low. Gentle hills occur in the west and northeast sections. A lowlands extends across the river, with an unusual lower area (a gully?) shown in the north center of the image. This depression controls the location and retention of water during floods that take place routinely during the rainy season. Flood stages can reach 9 m (30 ft) above the normal stream level. Using this elevation range as a constraint, we used an IDRISI routine, called RECLASS, that gives boolean algebra values of 0 and 1 to other ranges of values, to produce an image (named FLOOD) that defines the area of flooding (in blue). Here, we assigned 1 to all elevation values (in blue) less than 9 m, i.e., those prone to flooding, and 0 to all values above this level (rendered in black in the image).

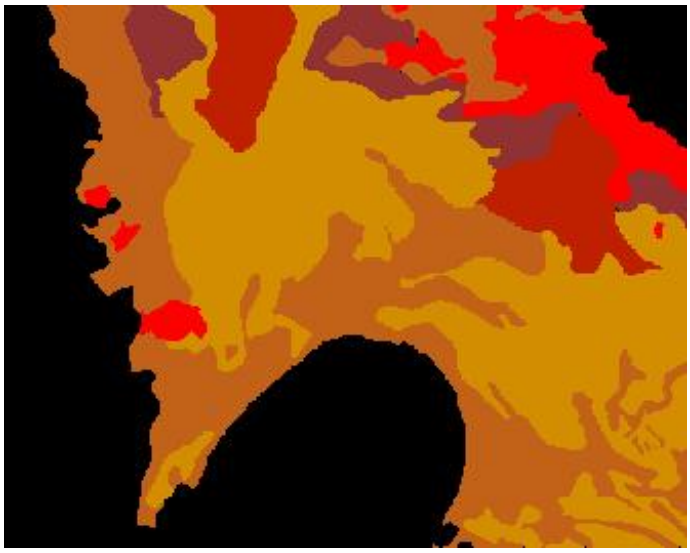
15-10: Why is the area beyond the river flooded where it is in the pattern map? ANSWER

During this stage, the soil types in the region differentially absorb floodwaters, such that, some soils retain enough water over the long term to provide a continuing supply of moisture during the growing season. Thus, soil type is another critical factor. Here is a map (DSOILS) of the

distribution of five soils within the region:



In this map, blue = heavy soils (1); yellow = clays (2); red = sandy clays (3); green = levee soils (4); purple = stony soils (5). The IDRISI program allows us to give each soil a suitability ranking, as follows: best (5) = clays, then in decreasing quality, levee soils (4), sandy clays (3), heavy clays (2) and worst, stony soils (1). The black areas are not considered further, for other reasons. Now, the trick becomes that of producing a map (SORGSUIT), in which we rate the soils by suitability, using the IDRISI *ASSIGN* program:



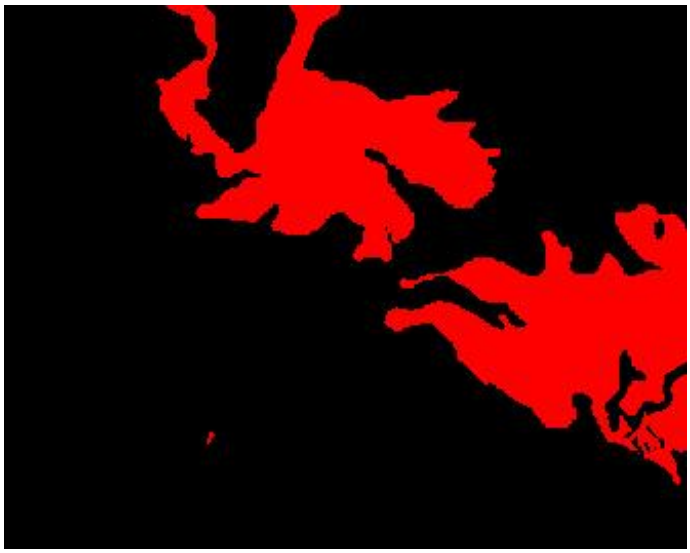
In this map, red = 1 (least suitable); brown = 2; dark brown = 3; yellow brown (most suitable) = 4; black = 5 (not considered). Clearly, this map has the same distribution pattern as the soils map above, but now, we instruct the computer to give each soil a numerical value (an attribute that represents its productiveness for growing this crop), instead of a soil-type name. We now simply generate a map that shows the location and sizes of patches of clay soil (in green) that constitute the best soil (named BESTSOIL) for growth:

As the flow chart indicates, the final step in the suitability process is to combine the information in the FLOOD and BESTSOIL maps as an *OVERLAY* to generate a decision end product map, called BESTSORG. Again, we require binary logic using Boolean algebra. On each of the intermediate maps, there are just two themes or patterns, one in color and the other in black, representing the idea of "good" or "appropriate" (flooded, optimum soil) and "not suitable" respectively. Let "acceptable" = 1, and "not suitable" = 0.

Each data cell (the grid is implicit in the above maps but is not shown) for each map then has either a 1 or a 0 within it. Since the two intermediate maps are at the same scale and projection, they fit or register in the overlay process. We could use an analog process, i.e., make transparencies of the two maps and visually overlay them on a light table. But, we can produce the same result digitally, by combining data in the grid cells for each map, following this map-algebra operation:

FLOOD		BESTSOIL		BESTSORG
0	X	0	=	0
0	X	1	=	0
1	X	0	=	0
1	X	1	=	1

Thus, only those cells that are colored (1) in both maps make a product score of 1. These form continuous patches, colored red, in the BESTSORG map produced by *OVERLAY* shown here:



Since we know the area of each cell, the total area associated with suitable conditions is just the product of the number of cells times the unit area. In IDRISI, the two programs GROUP and AREA do this.

15-11: Examine this BESTSORG map carefully. What is controlling the distribution pattern of the red that represents the optimum crop growth conditions? ANSWER

Of course, we would include many other attributes and thematic factors in a real suitability case, as for example: Access (roads, etc.), Market Requirements, Fertilizer Needs, Manpower Availability, etc. We express some of these factors as maps and others as data that we may organize into value tables. We can integrate these data in GIS models that add to the scores in each data cell. And, as we showed in the Site Suitability illustration above, we could apply a different system of ranking, in which we associate a range of rankings, say, from 1 to 10 (rather than just 0s and 1s), with each cell, and then sum them for all the data elements.

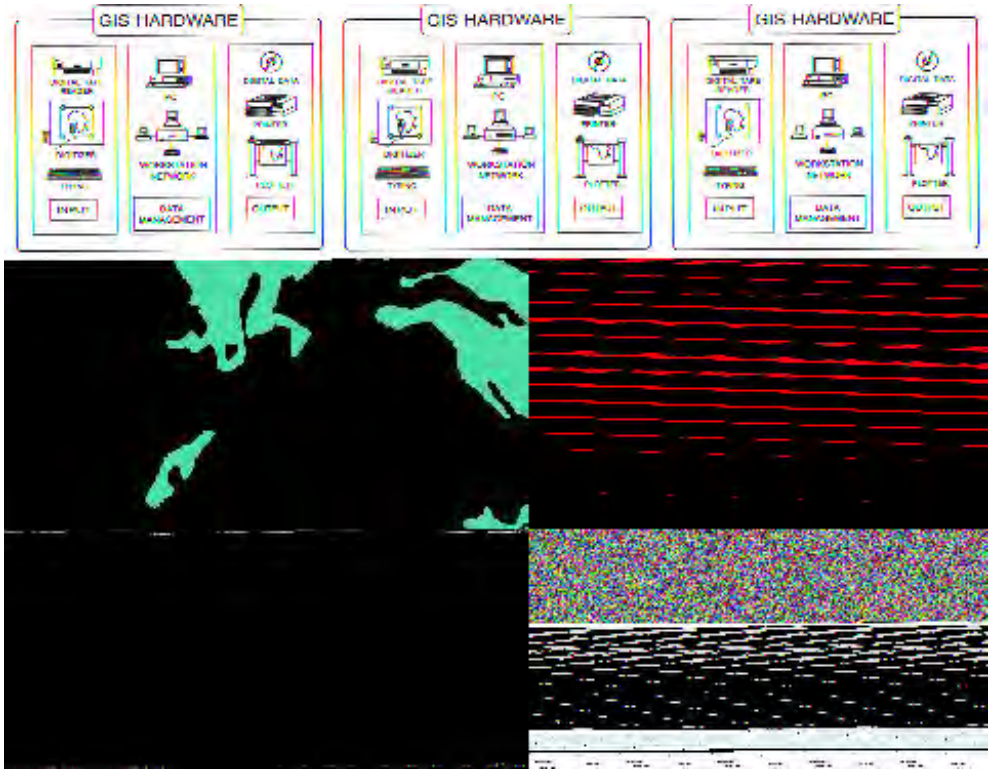
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Conducting a GIS analysis is best done by following some appropriate set procedure. The logic and strategy of this are described on this page. Some numerical data can be entered into a database consisting of tables or other non-image formats. But maps normally will have to be digitized or scanned. The usual steps in carrying out a GIS study for practical applications are set forth in this page.

Conducting a GIS Analysis

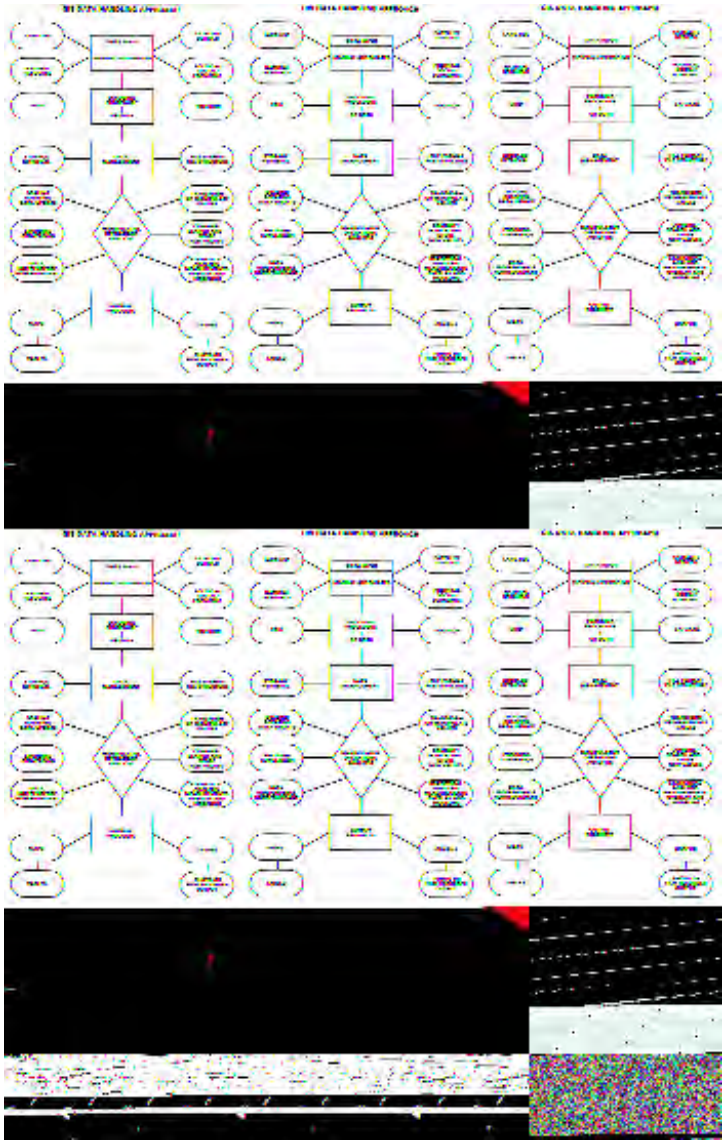
Now that we've shown what a GIS can do, let's consider next, some of the mechanics of conducting a GIS analysis. Today, we do this analysis almost exclusively on computers. We show some typical hardware in the following diagram (missing is a scanner):



From B. Davis, GIS: A Visual Approach, ©1996. Reproduced by permission of Onword Press, Santa Fe, NM.

Many organizations have developed software packages for GIS analysis. The most widely used are ARC/INFO (Unix and Windows NT platforms) and ArcView (a smaller PC desktop version) marketed by the Environmental Systems Research Institute (ESRI) of Redlands, California, a company founded in 1969 by Jack and Laura Dangermond. Another program, called GRASS, was developed by the Army Corps of Engineers and, although they no longer support it, you can learn about this tool at Baylor University's GRASSLINKS.

The following flow chart outlines a general system design for procedures and steps in a typical GIS data handling routine:



A similar analysis flow sequence is present in the USGS GIS tutorial referenced on page 15-5. We reproduce the sequence as the title heading in the subsections of the tutorial: Data Capture (scanning and digitizing) --> Data Integration --> Projection and Registration --> Data Structuring --> Modeling --> Overlay --> Data Output.

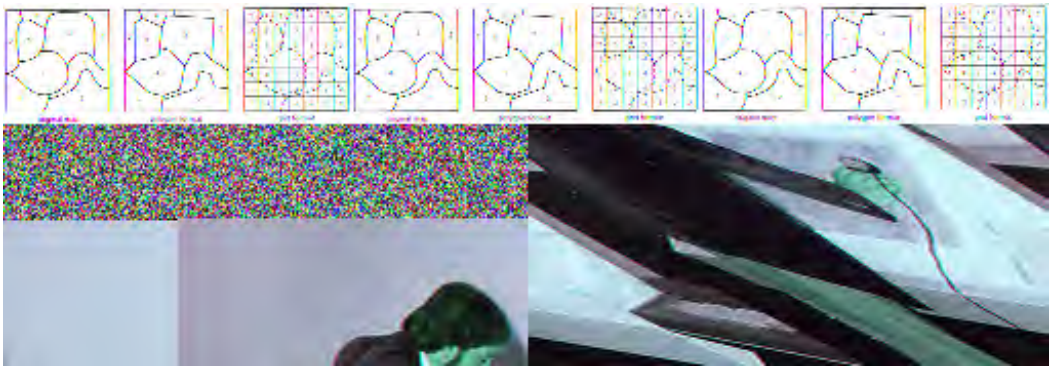
The objective in any GIS operation is to assemble a data base that contains all the ingredients to manipulate, through models and other decision-making procedures, into a series of outputs for the problem-solving effort.

15-12: In the diagram above, one of the entries in the data input end is satellite imagery. Make a list of some of satellite-derived observations that lead to information relevant to GIS analysis. ANSWER

Some input data may already exist in digital form, but we must convert most of it from maps and tables or other sources. We can scan some maps or satellite/aerial imagery and handle the resulting products as digital inputs to individual pixels that are data cell equivalents. Note that the digital system directly records pixels as colors. However, we must recode these values (geocoding), if we want them associated with specific attributes. Although recent technology has automated this conversion, in many instances we still must manually digitize maps, using a digitizing board or tablet, such as we show here:

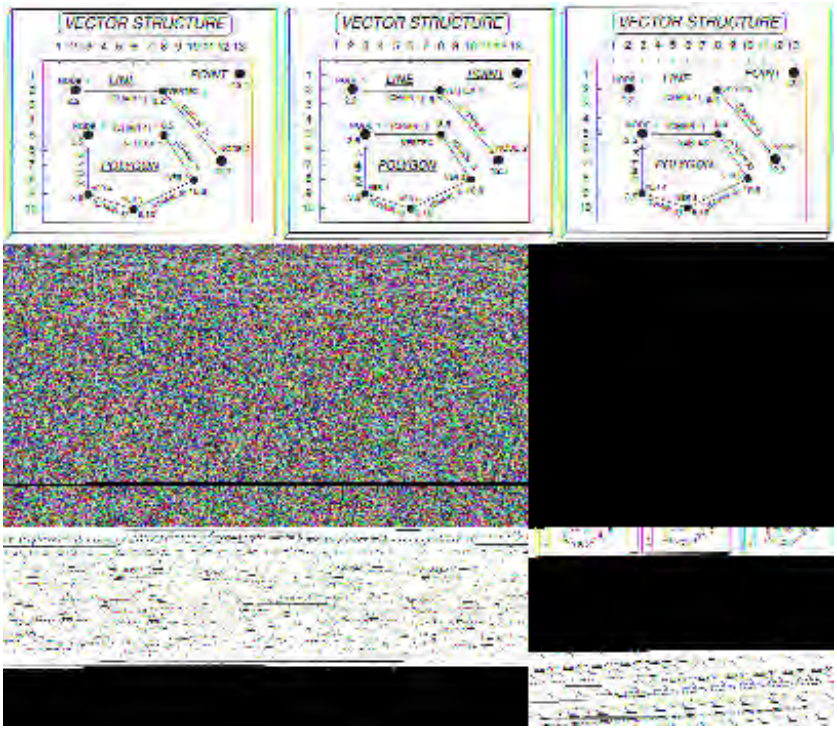
Behind this table, on which a map is mounted, are closely spaced electrical wires arranged in a grid that we can reference as x-y coordinates. The operator (here, Bill Campbell, Chief, NASA GSFC Code 935, who authored the chapter on GIS in the Landsat Tutorial Workbook) places a mobile puck with centered crosshairs over each point on the map and clicks a button to enter its position, along with a numerical code that records its attribute(s) into a computer database. He then moves to the next point, repeats the process, and enters a tie command.

Any map consists of points, lines, and polygons that locate spots or enclose patterns (information fields) that describe particular attributes or theme categories. Consider this situation that refers to several fields (e.g., different types of vegetation cover) separated by linear boundaries:



We can capture the information contained in each field by either of two methods of geocoding: *Vector* or *Raster*. In the center panel, the approach creates polygons that approximate the curvature of the field boundary. If that field has irregular boundaries, we may need many lines. Each line consists of two end points (a vector), whose positions we mark by coordinates during digitizing. We show that process in this diagram:

15-13: In the above diagram, by visual inspection, decide which of the two geocoding methods seems to be more accurate. ANSWER



From B.Davis, GIS: A Visual Approach, ©1996. Reproduced by permission of Onword Press, Santa Fe, NM.

Each point has a unique coordinate value. Two ends or node points define a line. We specify a polygon by a series of connecting nodes (any two adjacent lines share a node), which must close (the main pitfall is that some polygons don't close, so one must repeat or repair the process). We then identify each polygon by a proper code label (numerical or alphabetical). A look-up table associates the code characters with the attributes they represent. In this way, we can enter all the map fields that are large enough to be conveniently circumscribed, and their category values, into a digital database.

In the raster approach, we manually overlay or scan onto the map a grid array of cells, having some specific size. As shown in the right panel (grid format, above), an irregular polygon then includes a number of cells, completely contained therein. The system records these cells' locations within the grid and a relevant code number for each data element assigned to them. But some cells straddle field boundaries. A preset rule assigns the cell to one or the other of adjacent fields, usually related to a relevant proportion of either field. The array of cells that comprise a field only approximate the field shape, but for most purposes the inaccuracy is tolerable for making calculations. While the raster approach can be less precise, it has the advantage of creating numbers that are more easily handled.

Generally, grid cells are larger than the enclosed pixels in pictorial map displays, but the cluster of pixels within a polygon approximates the shape of the field. The relation of cells to pixels makes this raster format well adapted to digital manipulation. The size of a cell depends partly on the internal variability of the represented feature or property. Smaller cells increase accuracy but also require more data storage. Note that multiple data layers referenced to the same grid cell share this spatial dimensionality but have different coded values for the various attributes associated with any given cell.

15-14: From a data handling viewpoint, particularly involving computer manipulations, which method of geocoding, vector or raster, should be easier to process? ANSWER

Data management is sensitive to storage retrieval methods and to file structures. A good management software package should be able to:

I. Scale and rotate coordinate values for "best fit" projection overlays and changes.

II. Convert (interchange) between polygon and grid formats.
III. Permit rapid updating, allowing data changes with relative ease.
IV. Allow for multiple users and multiple interactions between compatible data bases.
V. Retrieve, transform, and combine data elements efficiently.
VI. Search, identify, and route a variety of different data items and score these values with assigned weighted values, to facilitate proximity and routing analysis.
VII. Perform statistical analysis, such as multivariate regression, correlations, etc.
VIII. Overlay one file variable onto another, i.e., map superpositioning.
IX. Measure area, distance, and association between points and fields.
X. Model and simulate, and formulate predictive scenarios, in a fashion that allows for direct interactions between the user group and the computer program.

Developing a GIS can be a costly, complex, and somewhat frustrating experience for the novitiate. We stress that data base design and encoding are major tasks that demand time, skilled personnel, and adequate funds. However, once developed, the information possibilities are exciting, and the intrinsic worth of the output more than compensates for the marginal costs of handling the various kinds of data. In plain language, GIS is a systematic, versatile, and comprehensive way to present, interpret, and recast spatial (geographic) data into more intelligible output.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

In the next two pages, you will see how the search for the best site for building a new electric-generating power plant near Harrisburg, PA, using both GIS and other information, is conducted step-by-step. This was part of a demo organized by NASA Goddard's ERRSAC both to the client (PP&L) and to the general user community. On this page are shown some of the data elements considered as input, a Landsat-based classification, and a series of thematic maps derived from data files at PP&L, state agencies at Harrisburg, and other sources.

The Pennsylvania Power and Light New Plant Siting Problem

Perhaps GIS analysis will come through more convincingly with a second case study. This case is an applications demonstration, developed at NASA Goddard during the early 1980s, when its Eastern Regional Remote Sensing Applications Center (ERRSAC) still operated. The "client" was the Pennsylvania Power and Light (PP&L) Company, a public electric utility, whose service areas include parts of central and eastern PA. One service area is the state capital district at Harrisburg on the Susquehanna River that had been receiving much of its power from the infamous Three Mile Island, site of a near-catastrophic nuclear power-plant accident on March 28, 1979. This experience forced PP&L to have special concerns regarding any of its future sitings, for its own power facilities and for its large customers. Specifically, long range planning called for a possible second generating plant to be constructed in the Harrisburg area. Could both GIS and remote sensing inputs help in the decisions involved in site selection?

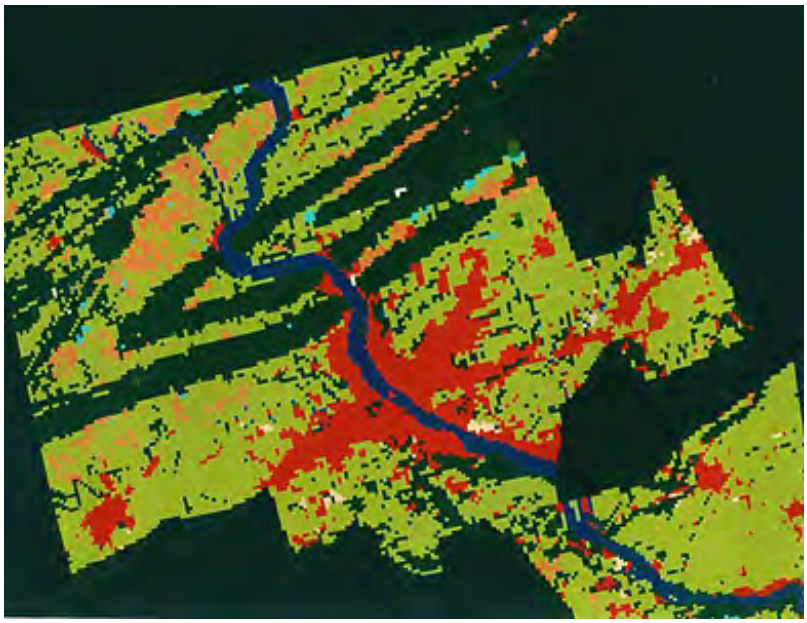
To refresh your memory of what Harrisburg looks like (you would have seen a picture looking northeast from the Interstate 83 bridge in the Exam at the conclusion of Section 1), we show a very similar photo here:

By the late '70s, PP&L had developed a computerized GIS database primarily for land use analysis, environmental impact analysis, energy facility siting, and other technical assistance. Its multivariate data base consisted of 43 data elements distributed in ten general categories, the two largest being terrain units and land use. These are shown in this table:

The standard grid cell size in this data base is 9.2 hectares (22.9 acres). The question prompting PP&L to approach ERRSAC in setting up a cooperative study: Can Landsat images help keep the GIS outputs up-to-date?

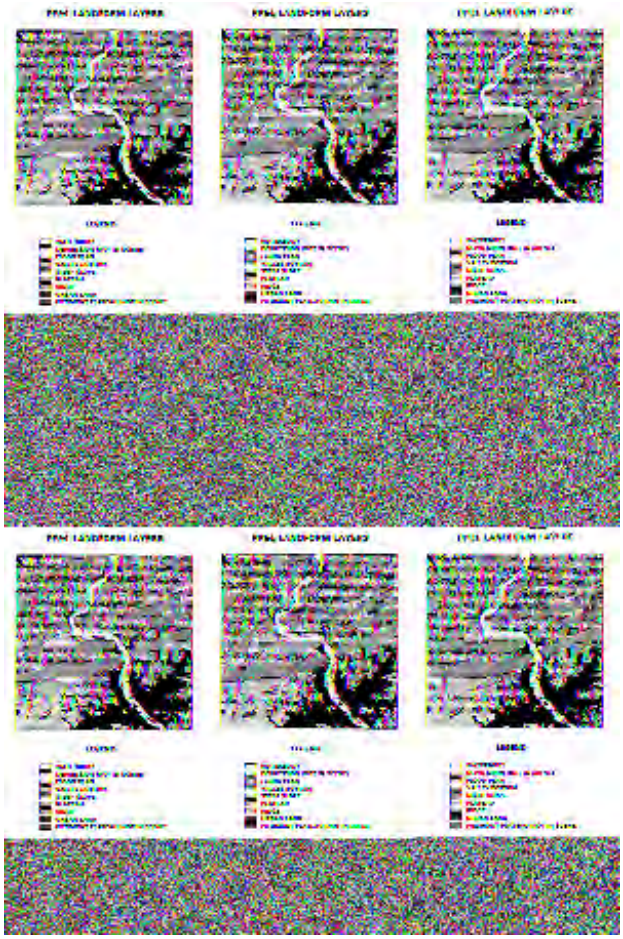
15-15: Landsat MSS resolution is equivalent to 1.1 acres. How did this fact influence PP&L in its thinking about utilizing space imagery in its geocoded data base? ANSWER

From these files, ERRSAC investigators used an ESRI software package to generate a 12-class map of the Harrisburg service area:

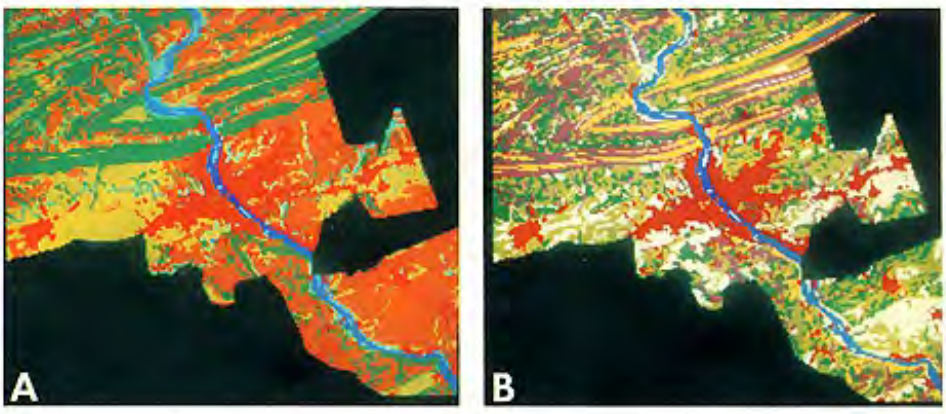


The proof-of-concept task presented by PP&L for a Goddard-directed GIS analysis was to select optimal site(s) for a heavy industrial complex. Two primary limitations were imposed: 1) the site must be within 5 km (3 mi) of a major source of water (the Susquehanna River met this condition), and 2) should exclude areas in the immediate Harrisburg district.

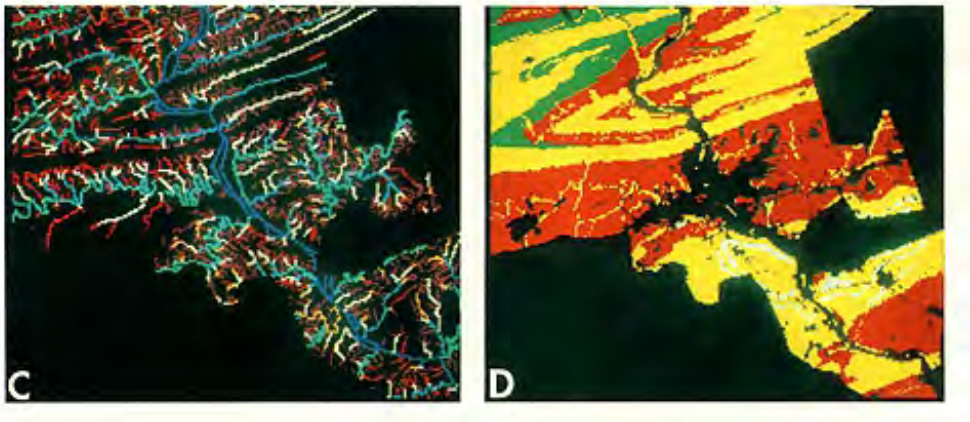
The first step in the analysis was to generate a series of data element maps from the PP&L data base. Initially, these were formatted as printouts using a black and white Versatec Electrostatic Plotter. Here is the output for landform types, as data layers (this figure suffers from the inability of the scanner to differentiate the symbol patterns, some of which are too similar to be resolved):



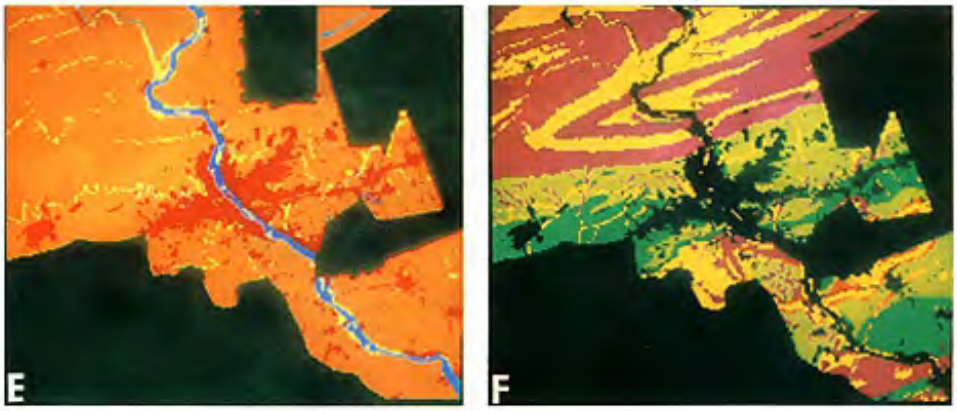
After completing these printout maps, many were converted to color-coded renditions. Below is a set of six (of the 23 total produced) thematic maps using PP&L data digitized at ERRSAC to become part of the multilayered inputs to the algorithms that were used to select acceptable site(s) for the proposed power plant. See description underneath each map as identified by letter for theme represented and values associated with colors used.



- A. Landforms: Urban = Red; Water = Dark Blue; Dissected Plateau = Brown; Steep Slope = Dark Green; Valley Bottom = Greenish-yellow; Depression = White; Floodplain = Aqua
- B. Slope: 0 - 3% = White; 3 - 8% = Light Green; 8 - 15% = Dark Green; 15 - 25% = Purple; > 25% = Yellow; Urban = Red; Water = Dark Blue



- C. Stream Order: 1st = Dark Blue; 2nd = Orange; 3rd = Aqua; 4th = Light Blue; 5th = Yellow; 6th = White; 7th = Purple; 8th = Red
- D. Soil Permeability: Very High = Dark Green; High = Red; Good = Yellow; Moderate = Orange; Poor = White



- E. Flood Prone Areas: Urban = Red; Water = Dark Blue; Flood Prone = Yellow; Not Flood Prone = Orange
- F. Agricultural Potential: No Limit = Dark Green; Few Limits = Light Green; Moderate Limits = Yellow; Severe Limits = Brown; Extreme Limits = Purple

These, and other, maps were helpful to the GIS interpreters and the planners in visualizing the distribution of variations within a given attribute. In the analysis, they stored the different categories within a theme and manipulated them by ordinal rankings in a numeric code.

15-16: Which, if any, of these maps could Landsat have made some direct contribution to, perhaps aided by supplemental information? ANSWER

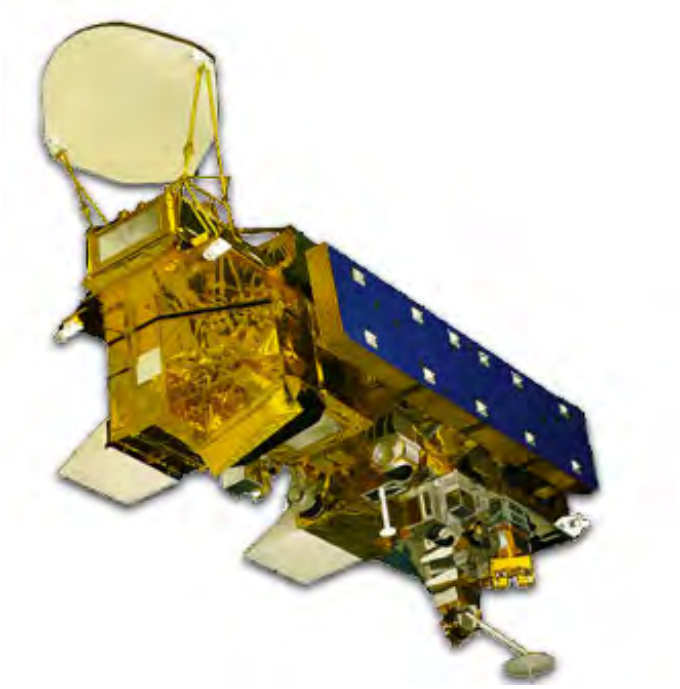
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

As of July, 2002. Aqua has been launched successfully but as of this date only limited imagery or other data have been released. The present page will be fleshed out as the materials become available. Meanwhile, this is a brief synopsis of the program, plus a further look at the European Space Agency satellite Envisat, which is a companion to the Terra and Aqua pair but is also equipped with sensors that supplement the data needed for Earth Observations. Launched in July, 2004 is the third in this Series: Aura.

AQUA; ENVISAT; AURA

The afternoon segment (PM) of the EOS flagships, Aqua, was successfully launched by a Delta II rocket from Vandenberg Air Force base at 2:55 AM PDT on May 4, 2002. Here is an artist's painting of the spacecraft:

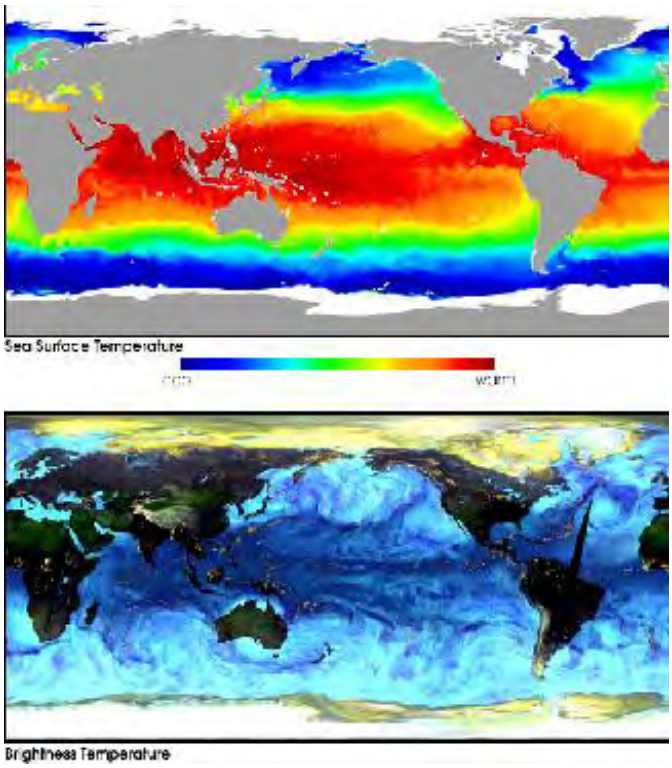


Aqua's initial orbit of 705 km (438 miles) leads to an average equatorial crossing time (going north) of 1:30 PM. Its mission is to complement Terra observations by providing data later in the day; however, the new sensors on Aqua are designed to obtain information of interest primarily to *meteorologists* and *oceanographers*. A better understanding of the planet's water cycle is the goal, as these sensors will lead to improved data on global precipitation and evaporation, radiative balances, humidity and temperature profiles through the atmosphere, soil moisture, etc. The overall objective is to integrate the water studies into a better general understanding of the world's environments. The three host participants in Aqua are the United States, Japan, and Brazil, although other nations are close associates.

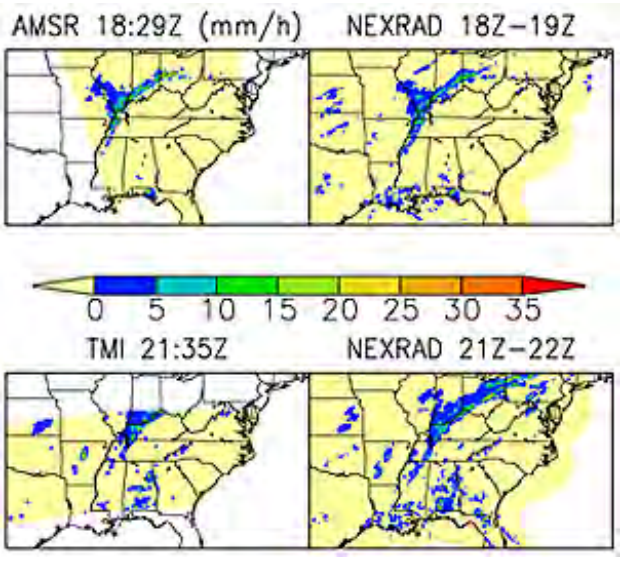
Aqua carries six state-of-the-art instruments in a near-polar low-Earth orbit. The six instruments are the Atmospheric Infrared Sounder (AIRS), managed by JPL, the Advanced Microwave Sounding Unit (AMSU-A), the Humidity Sounder for Brazil (HSB), the Advanced Microwave Scanning Radiometer for EOS (AMSR-E), the Moderate-Resolution Imaging Spectroradiometer (MODIS), and Clouds and the Earth's Radiant Energy System (CERES). Each has unique characteristics and capabilities, and all six serve together to form a powerful package for Earth

observations. Information about each of these sensor systems is rather involved and will not be summarized here, but the reader is encouraged to consult the section on Instruments at the Aqua Home Page.

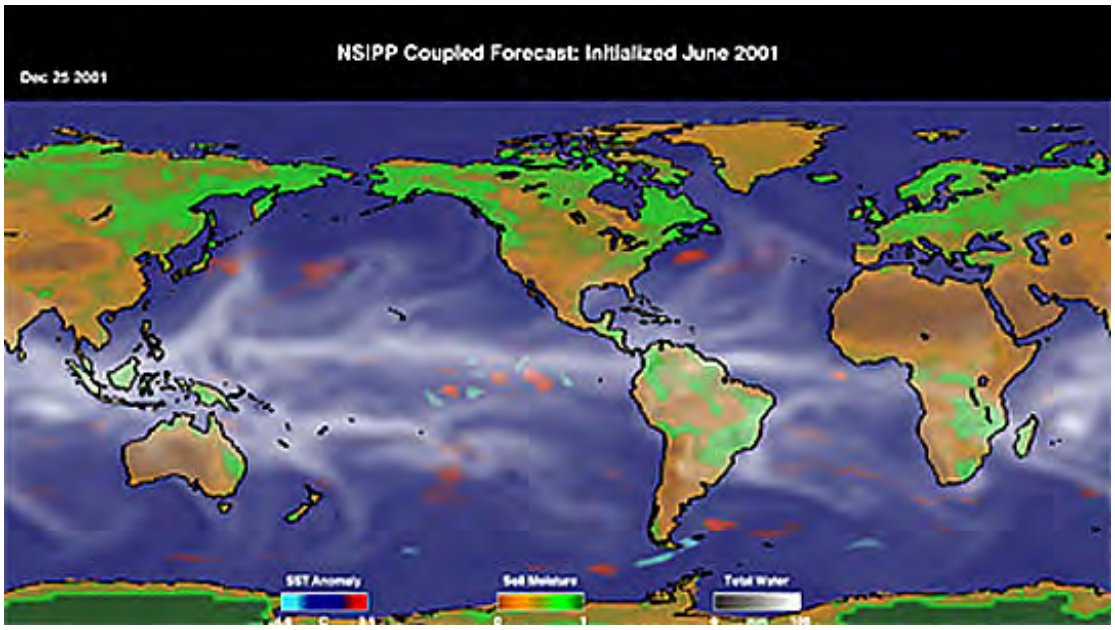
Several months were involved in completing the instrument test phase, after which Aqua went operational. (There were a few minor glitches during this period.) The first imagery released by the Aqua team is this AMSR-E pair that give improved sea surface and brightness temperatures on June 2-4 (three days data integrated):



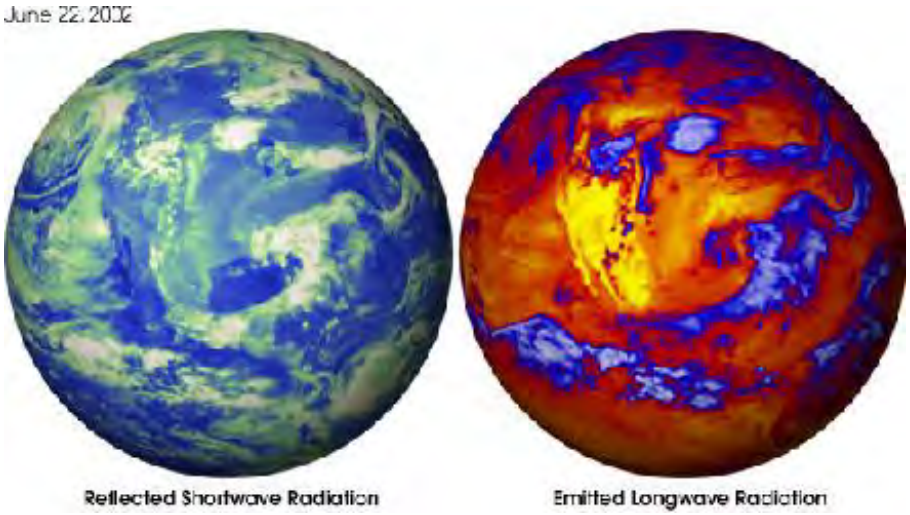
AMSR-E has been used also to locate and measure rainfall on the continents. This set of data maps shows measurements made simultaneously within the same hour on June 5, 2002 by AMSR-E from space and by the U.S. Weather Bureau's ground-based NEXRAD Doppler radar and by TRMM (see page 14-5) three hours later, with corresponding NEXRAD readings. The space and ground measurements were 0.31 and 0.29 mm/hr respectively.



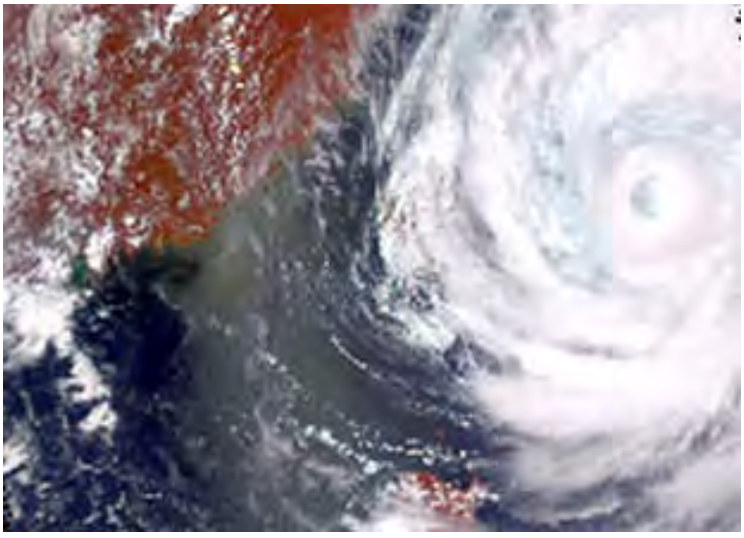
A very important task in applying microwaves to the Earth's surface is the detection of soil moisture and estimation of its amount. Here is a map of the generalized variations in soil moisture on a global basis as determined by AMSR in June 2001.



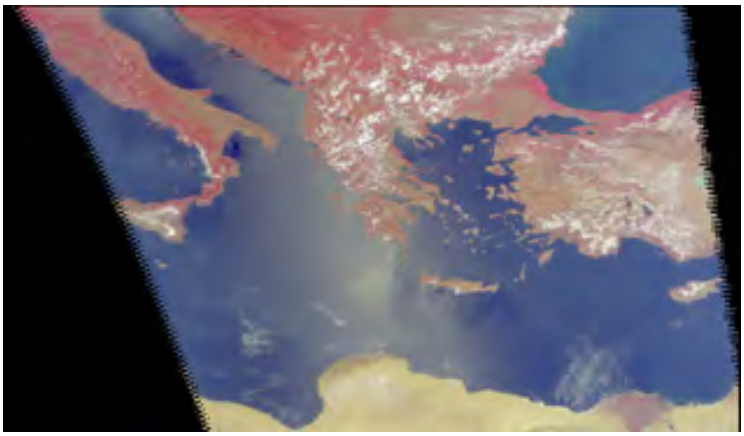
The second sensor to "go public" with data is CERES. We have described this sensor on the previous page. Recall that CERES measures reflected thermal radiation from the Sun and heat from the Earth's surface warmed by the Sun. Here is another pair of images centered on North America and acquired on June 22, 2002:



Then, on July 6, images from the AIRS (Atmospheric Infrared Sounder) group were released. The sensors making up this instrument are designed to measure temperatures along vertical profiles so as to derive a three-dimensional map of the atmosphere. The first image is a Visible view of Tropical Cyclone Ramasun in Asia:

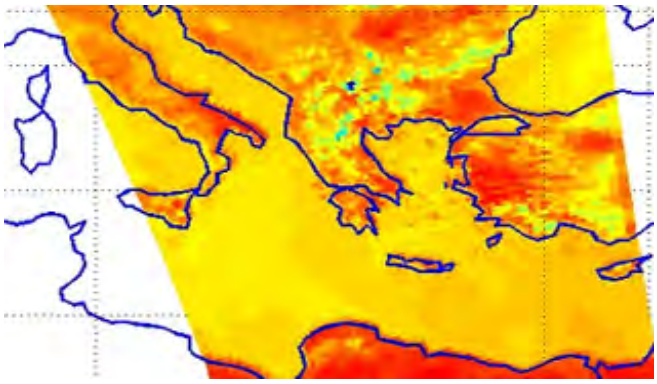


Released simultaneously were images from the AIRS group of sensors for a region from Italy east to Turkey and south to the North African Coast. This is the Visible image obtained from AIRS itself:

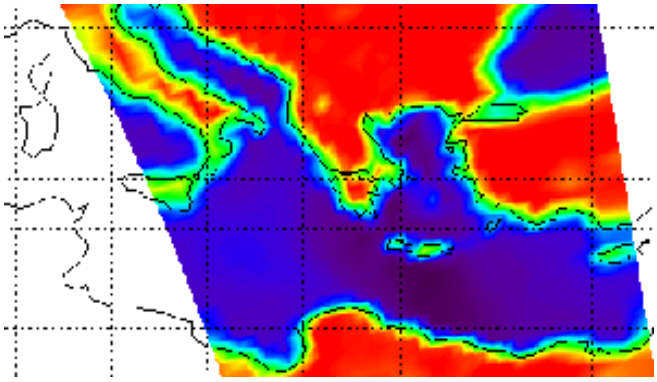


Beneath this visible view as a geographic reference in succession are:

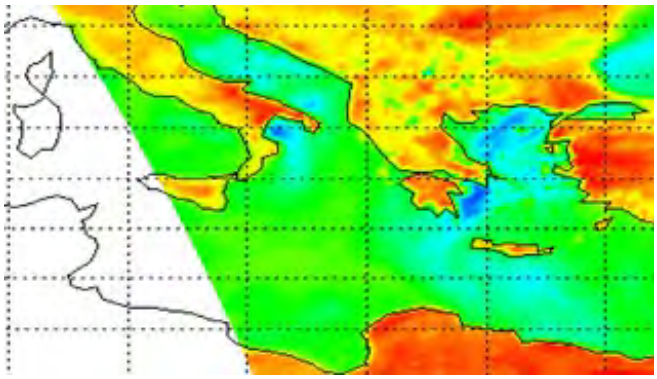
1) An AIRS infrared image at 11 µm, which measures surface and cloud temperatures:



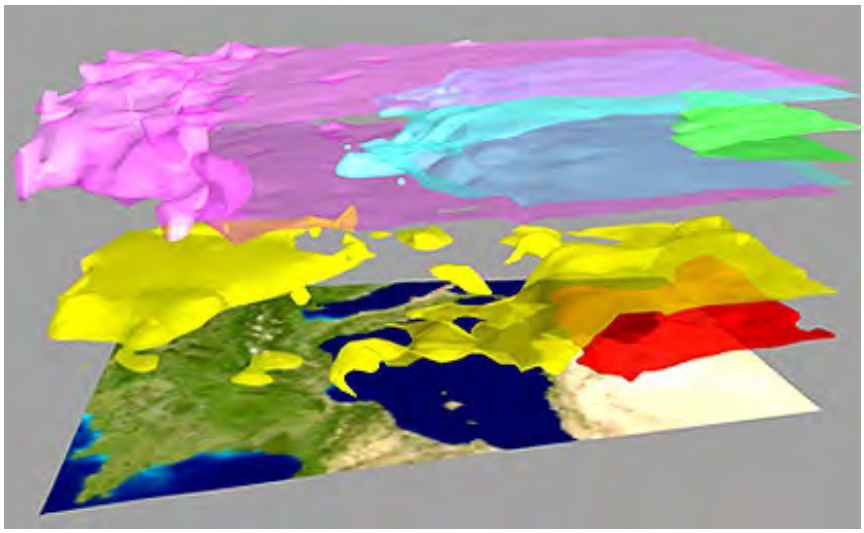
2) An AMSU (Advanced Microwave Sounding Unit) temperature plot, sensed in the microwave at 31.4 GHz:



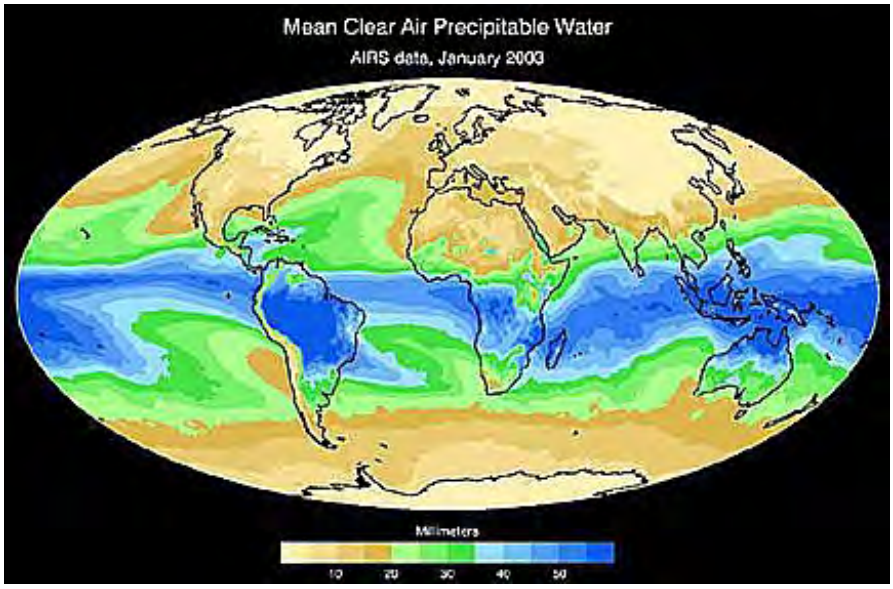
3) An HSB (Humidity Sounder, developed and operated by Brazil) map, made from a sensor operating at 150 GHz, designed to obtain mid-troposphere temperatures, and sensistive to moisture, precipitation, and ice crystals; its dry land temperatures are close to those of AIRS:



Three-dimensional atmospheric maps developed from multisensor observations such as the above three data sets have now been released. Here is the western Mediterranean (France on left and North Africa on right) with the bottom shown as a landscape image, and atmospheric layers at 8 and 12 km shown in colors with red the warmest and purples coolest.

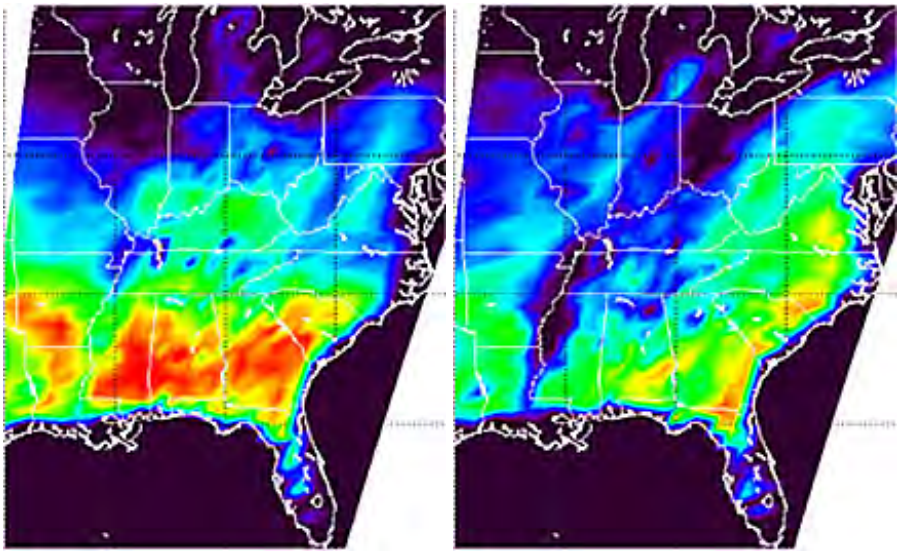


AIRS also measures the total water vapor content of the atmosphere from its base to the outer limits of H_2O occurrence. This can be done globally for an extended period of time, as shown in this map which uses daily averages summed up for January 2004 to arrive at estimates of rainfall potential expressed in millimeters. The tropical zones can thus produce more than a half meter of precipitation over that period.

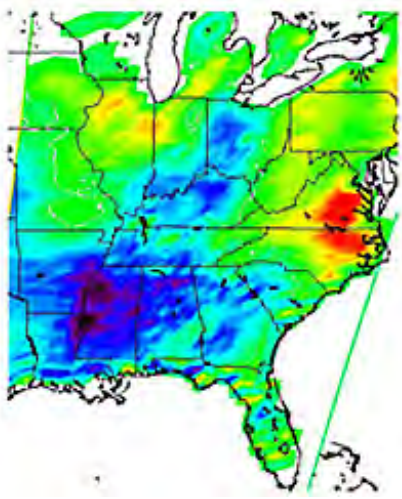


AMSU measures surface emissivities in the microwave region as well as atmospheric emissivities. These in turn are used to calculate temperatures. An excellent illustration of end products from this sensor is given by these observations of the eastern half of the U.S. before

and after passage of Hurricane Isidore in late September of 2002 (these dates are determined by the 16-day orbital repeat cycle of Aqua). The left image was obtained on September 12; the right on the 28th, one day after the hurricane made landfall.

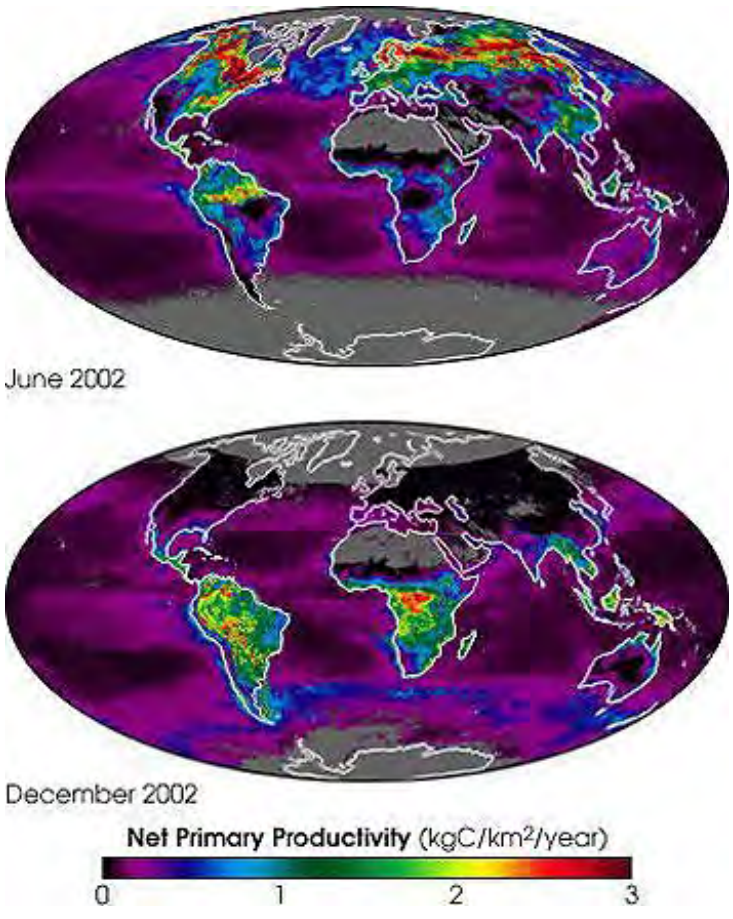


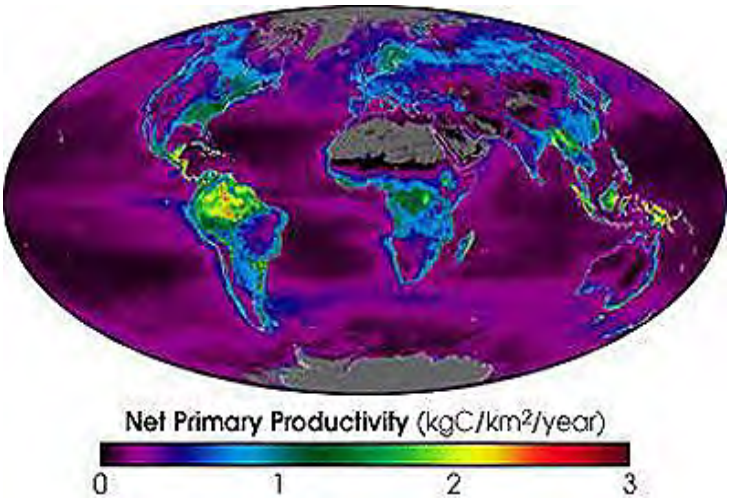
Increasing temperatures range from the colder blues to the warmest reds. In the right image, the blues around the Mississippi River drainage basin result mainly from surface soils that have absorbed water from the hurricane (it appears further north as a blue area) such that the cooling effect of the rainfall into these soils has lowered emissivity. In the image below, differences in temperature determined by subtracting the second from the first AMSU image show a new map in which the hurricane cooling effects stand out in blue and red pinpoints an area of warmer temperatures *in the air* evident in the Carolinas:



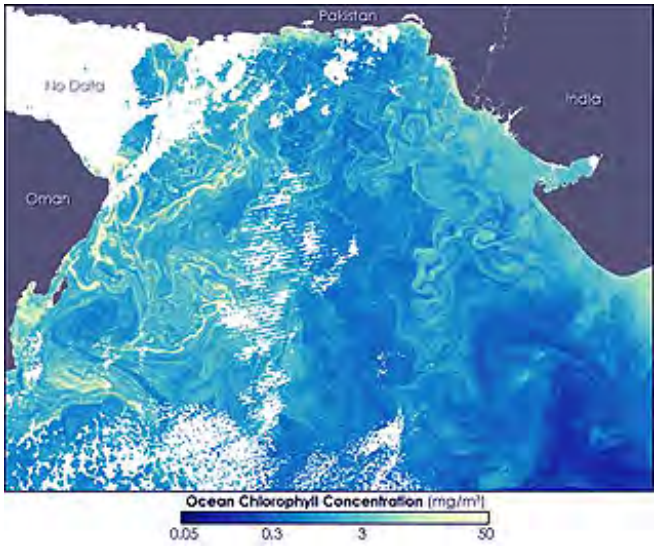
The first MODIS images were released near the end of June, 2002. An example of its large area coverage is this natural color MODIS image, taken on June 24, that shows the U.S. West Coast from northern Oregon to the Mexican border. The red spots are active wildfires in the Klamath Mountains of SW Oregon, and elsewhere in that state. Most of the clouds off northern California are actually a thick fog bank.

MODIS on both Terra and Aqua took measurements of CO_2 throughout 2002. From knowledge of the Carbon Cycle (page 16-4) relations between CO_2 and the amount fixed in vegetation (land and ocean), a measure of productivity could be made at different times of the year. The first pair of plots below show the distribution of global productivity in June and December of that year. The next map indicates the net productivity for the year.





MODIS has also proved its value at detecting plankton and thus complements SeaWiFS and other ocean-dedicated satellites. This image shows chlorophyll distribution in the Arabian Sea between Pakistan and Oman. The higher than normal concentrations of planktonic life are due to a period of significant rainfall in the region.



Other satellites in the EOS program are now operational. ICESat (Ice, Clouds, and land Experiment Satellite; see page 14-14), whose sole instrument is GLAS (Geoscience Laser Altimeter System) and SORCE (Solar Radiation and Climate Experiment) were both successfully launched in 2003 but large quantities of data are still forthcoming.

Aura, discussed below, will concentrate on atmospheric chemistry and climate conditions, looking with its instruments also at the dynamics involved in climate change.

ENVISAT

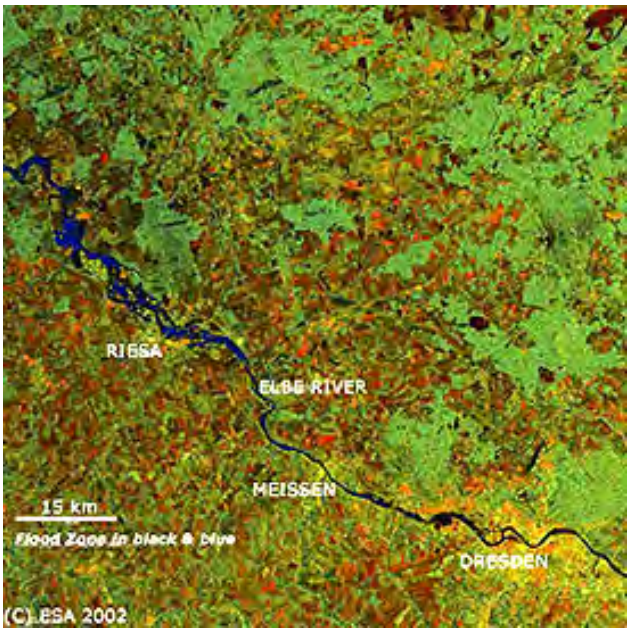
We previously described the ESA (European Space Agency) environmentally-oriented program with its Envisat as the lead satellite, near the second page of the Overview.Check this for details, and note that Envisat was also launched in May of 2002. ESA claims their satellite is the biggest of the EOS family and has the most sensors (10). Count these in this artist's drawing - you should find 10. These are described in the Envisat component of ESA's Web site.

Several images from Envisat appear elsewhere in the Tutorial. Two excellent MERIS (for Medium Resolution Imaging Spectrometer) images are found in the Overview. Here is another MERIS wide swath image showing France, the Alps, and a cloud bank over Germany:

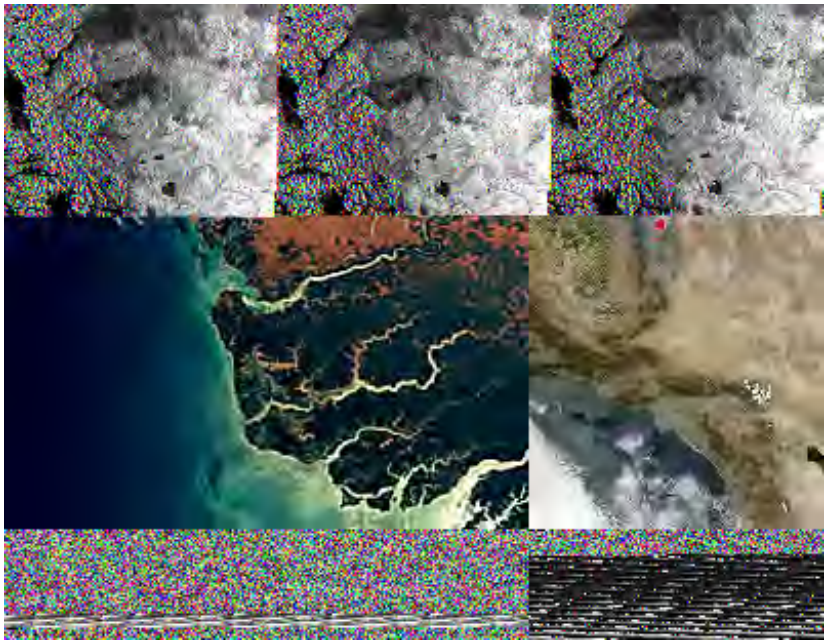
An example of a scene that has strong environmental information is this MERIS image of the West Coast of Africa, including part of Senegal. The Kaffrine River delta is densely vegetated with brackish water coastal vegetation. Dakar is at the end of the westernmost peninsula. Note the sediment from the river.

Data taken by AATRS (Advanced Along Track Scanning Radiometer) are used primarily to determine sea surface temperatures. But several bands can be processed in images that resemble false color composites. These tend to look "funny" because of their extended elongation (along the orbital track) but they have the advantage of being a continuous "picture" taken at one time. Here is one that begins in Kenya and ends in the Sinai Peninsula across the Gulf of Suez; to fit better on the page it has been turned 90° so that north is to the right.

We have seen several ASAR (Advanced SAR) images in Section 8, page 8-7. Here is another, showing the Elbe River south of Berlin, in color because several band modes were used.

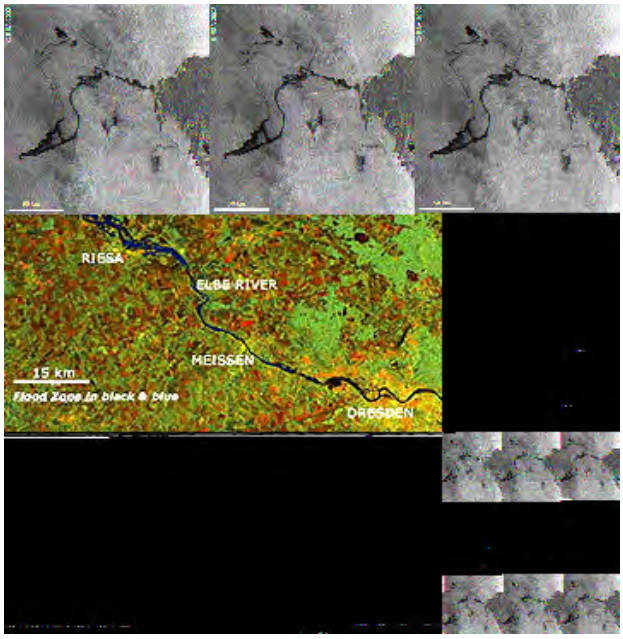


Many of the ASAR images are in black and white. Here is one showing Patagonia in Argentina, with the Andes to the left (west):

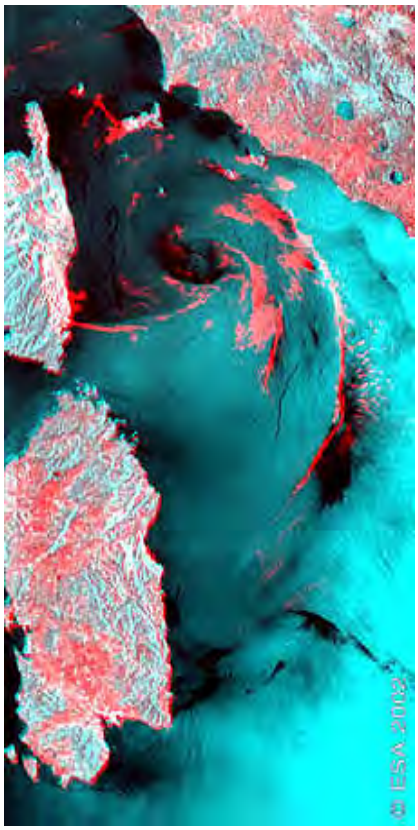


As an example of the environmental utility of Envisat, we look again at the oil spill in the eastern Atlantic of Galicia in northwest Spain, which occurred in November of 2002 when the tanker

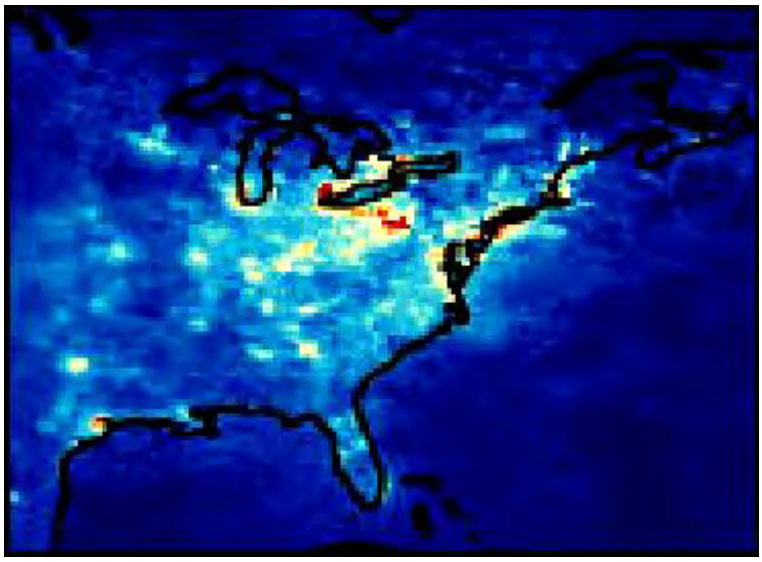
Prestige sank with most of its cargo of 25 million barrels of oil. About 1.5 million barrels did escape, some reaching coastal beaches, as seen here in this ASAR image.



As with several other satellites that mount both visible-NIR sensors and radar, these images can be combined. The three panels below show parts of Corsica and Sardinia in the Tyrranean Sea and the west coast of Italy just north of Rome. The left panel is an ASAR radar image; the right panel was made from MERIS images. The bottom panel is a registered combination of both - red areas over the water are rain clouds; red on land is NOT vegetation, but in the band combinations used this color coincides with vegetation-poor land areas.



One of Envisat's tasks is to monitor various chemicals (including pollutants) in the global atmosphere. Its Scanning Imaging Absorption Spectrometer for Atmospheric Cartography is capable of determining the amount and distributon of the noxious gas NO_2, released largely in auto emissions. Here is a map of the distribution of this gas over the eastern U.S. in June, 2003; yellows and browns are higher concentrations.

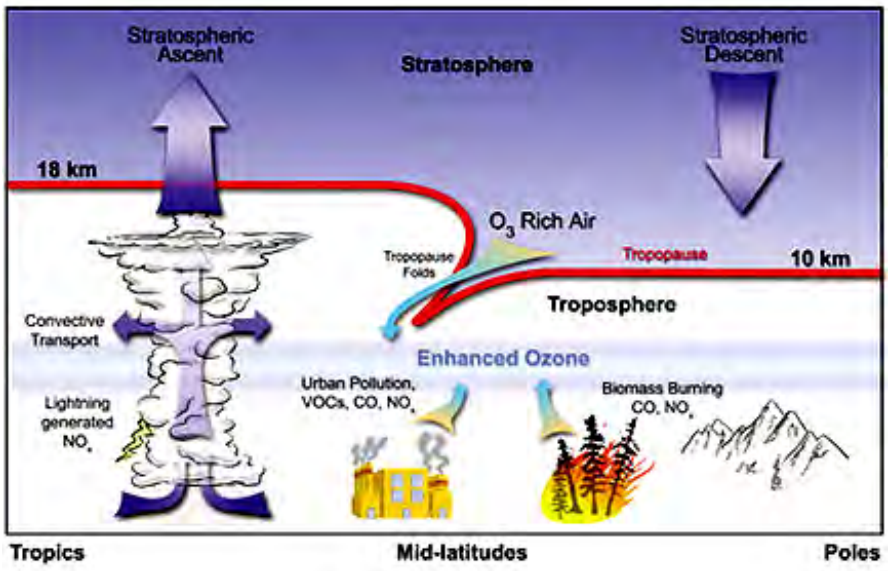


Aura will extend this compositional capability for air analysis and will dovetail results with Envisat.

AURA

The natural chemistry of the Earth's atmosphere is dominated by three elemental gases: O_2, N_2, and Argon; much of the CO_2 in air is introduced by natural processes such as transpiration. Some sulphur compound gases also are brought into the atmosphere by volcanic processes. Ozone, O_3, is likewise in large part present naturally. But as we have seen earlier in this Section, man's activities have altered natural balances of the above chemicals and added still others, mostly in burning fuels. A general diagram illustrating chemical transfer in the atmosphere is as

follows:

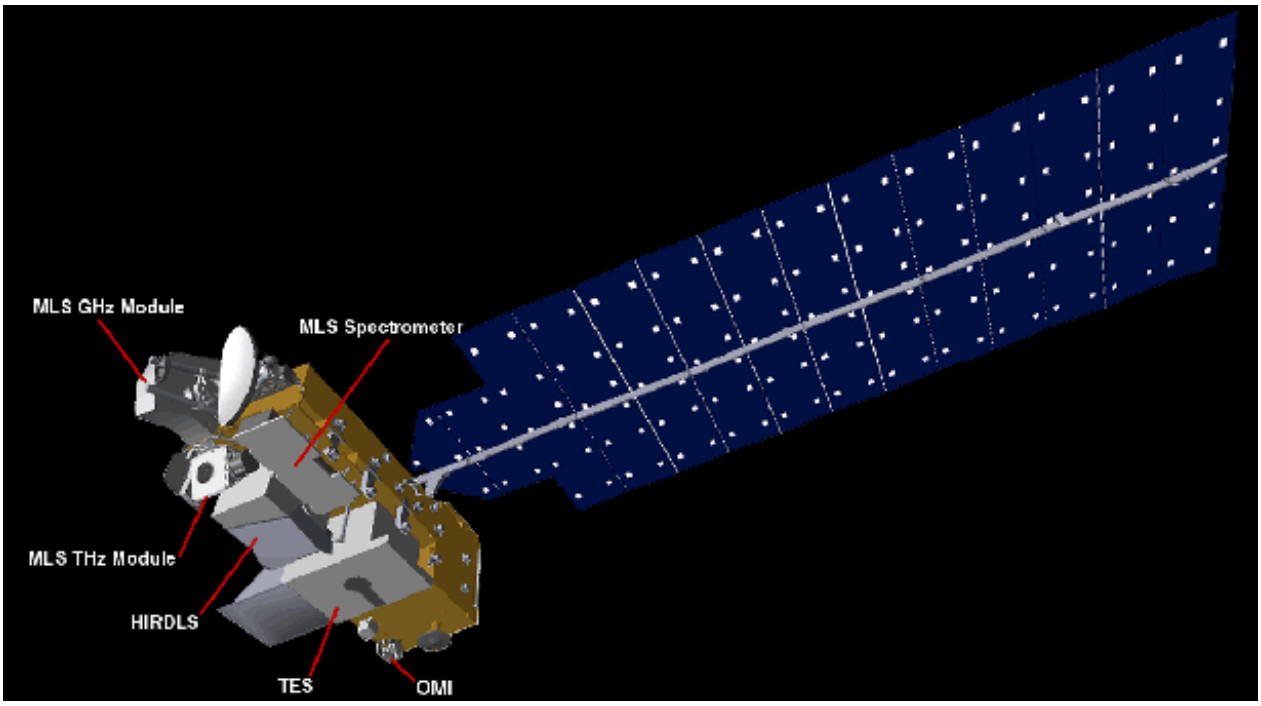


The EOS program has included a spacecraft designed to primarily monitor the Earth's atmosphere. The third flagship in NASA's Earth Observatory spacecraft, AURA, was successfully launched in the early morning of July 15 2004 from the Vandenberg Air Force Base facility near Lompoc, California. Being a night launch, the visual effect of its takeoff, after 6 delays, was a "glorious" sight for the many scientists, engineers, and technicians who had patiently waited until all systems were finally set at "GO".

Aura, like the other EOS spacecraft, has its own website.

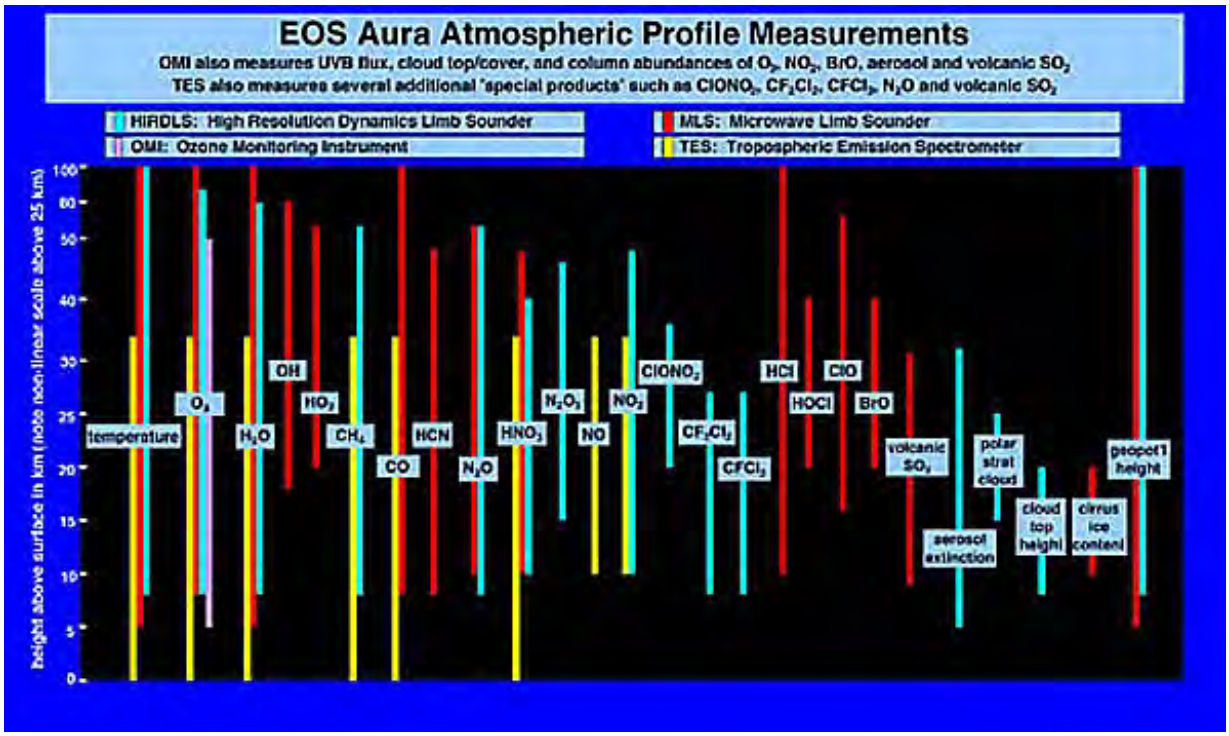
Aura was inserted along an orbital path and altitude that allows its to follow just 8 minutes behind Aqua, so that here is an example of formation flying, discussed on the next page. With two assemblages of different instruments, the Aqua-Aura pair are acquiring complementary data sets.

Its primary mission is to determine the "health" of the world's air envelope. This includes an inventory of global three-dimensional distribution of various chemicals (most as pollutants) in the atmosphere, as well as more detailed measurements of ozone, and evidence of short- and long-term climate change. We show here the $736 million dollar spacecraft (as big as a small bus and weighing over 3 tons) :



Below is the spacecraft "innards" in the assembly facility before its outer cover was emplaced.

Here is a chart indicating the chemicals that can be detected and quantified



The tasks of its four principal sensors are:

The joint U.K./U.S. **High Resolution Dynamics Limb Sounder (HIRDLS)** instrument will measure trace gases, temperature and aerosols in the upper troposphere, stratosphere, and mesosphere.

The **Microwave Limb Sounder (MLS)** will measure important ozone-destroying chemical species in the upper troposphere and stratosphere and measure trace gases in the presence of ice clouds and volcanic aerosols. MLS was built by the Netherlands and Finland in conjunction with NASA.

Ozone Monitoring Instrument (OMI) is Aura's primary sensor for tracking global ozone change and will continue the record of space-based ozone monitoring that began in 1970. NASA's Jet Propulsion Laboratory built OMI.

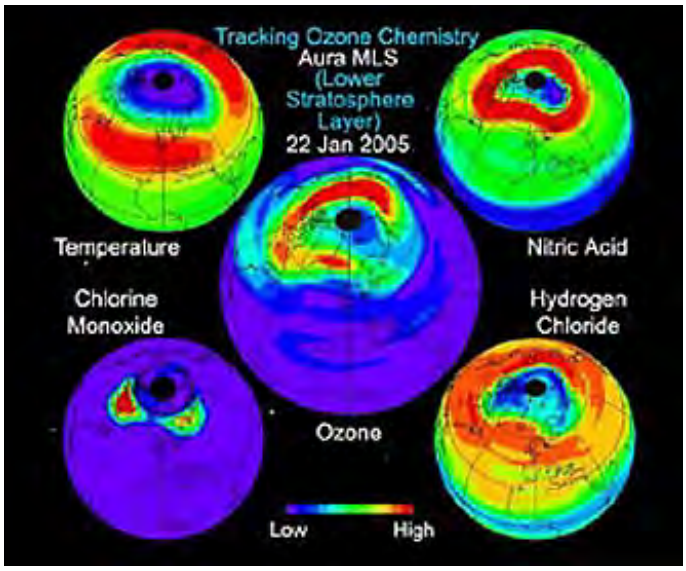
Tropospheric Emission Spectrometer (TES) will measure tropospheric ozone directly and other gases important to tropospheric pollution. JPL also made TES.

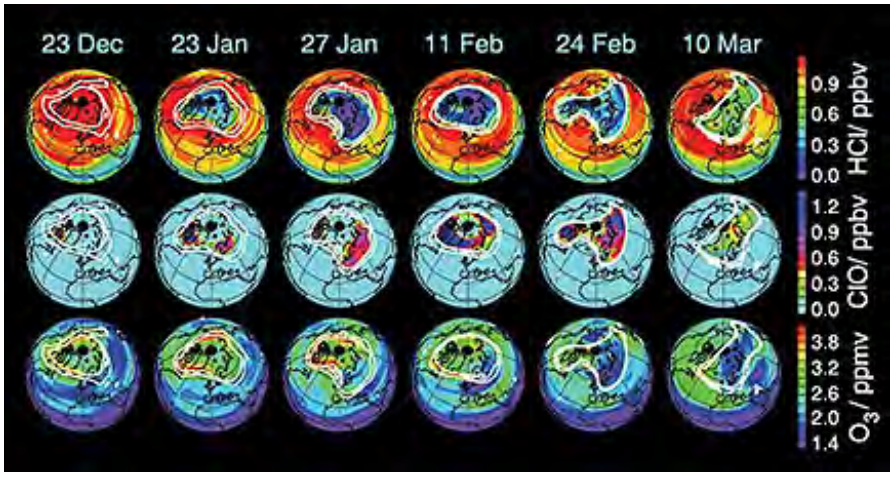
All four images are capable of making ozone measurements.

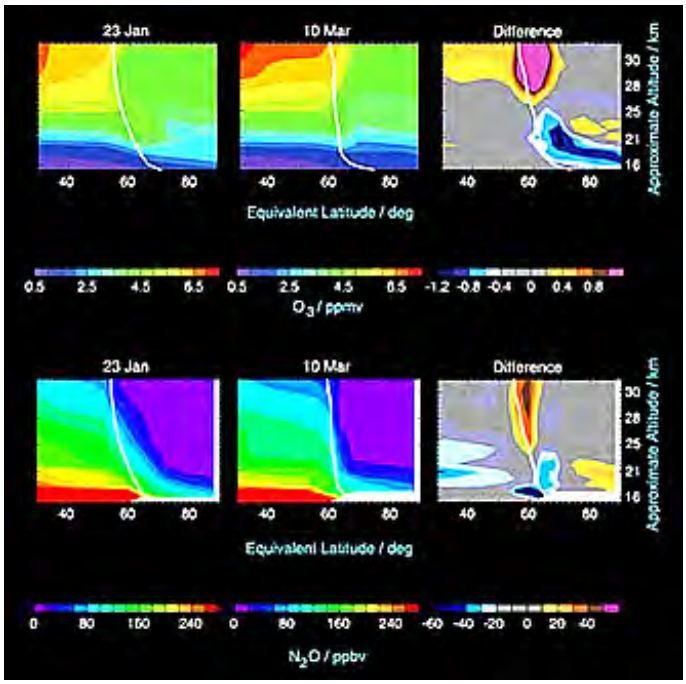
As with most satellites of this size and complexity, it took almost 90 days to purge the system of unwanted gases, check out instruments, adjust orbit, and do other appropriate housekeeping.

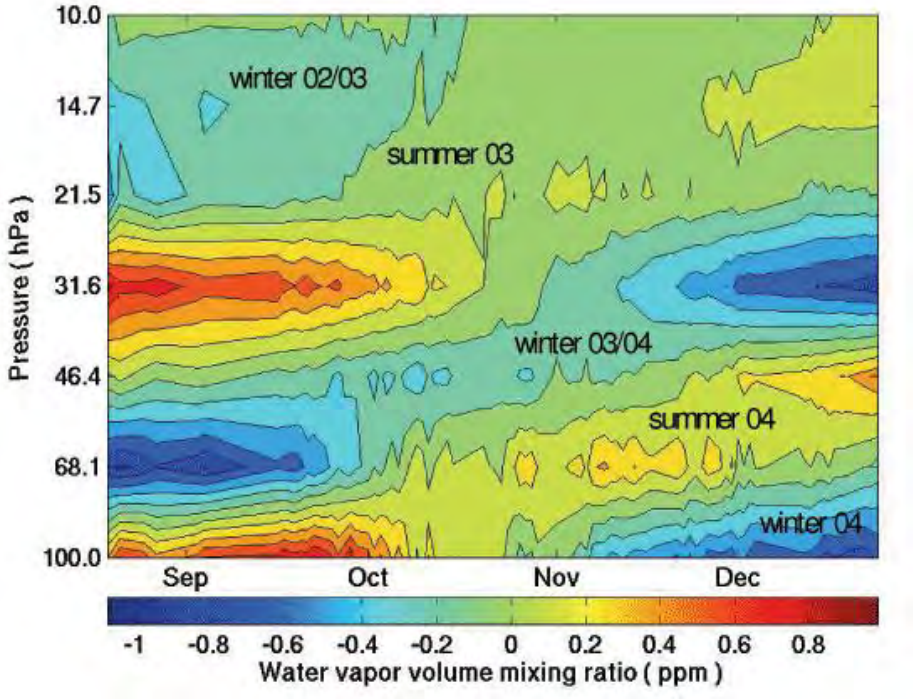
One major disappointment soon was evident. The Sun Shield on HIRDLS had upon opening become clogged with some debris in this instrument shaken loose during launch. The Shield had deployed only 20% of its full open condition. This severely compromises data acquisition because of the limited field of view. Hence, images from this sensor have not been released.

Far better luck befell the other three instruments. We look first at some MLS images; read their captions for brief descriptions.

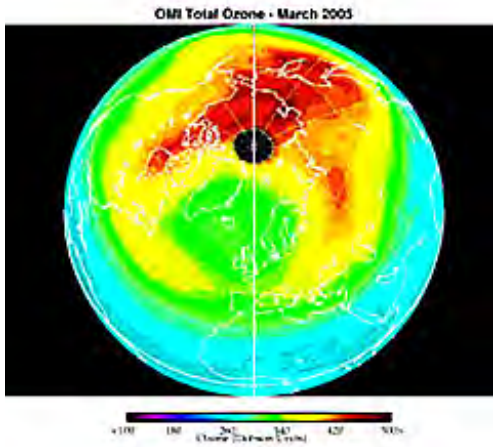

Tracking Ozone Chemistry
Aura MLS
(Lower
Stratosphere
Layer)
22 Jan 2005
Temperature
Nitric Acid
Chlorine
Monoxide
Hydrogen
Chloride
Ozone
Low
High

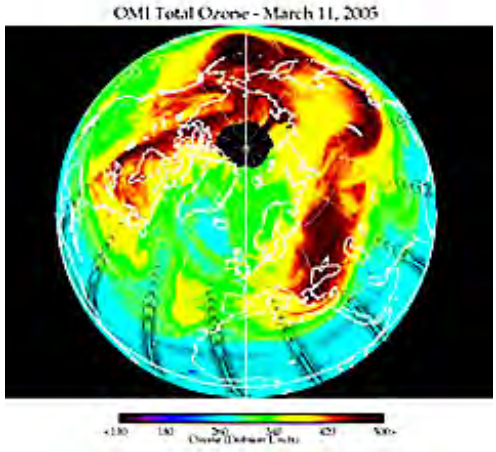

23 Dec
23 Jan
27 Jan
11 Feb
24 Feb
10 Mar
0.9
0.6
0.3
0.0
HCl/ ppbv
1.2
0.9
0.6
0.3
0.0
ClO/ ppbv
3.8
3.2
2.6
2.0
1.4
O3/ ppmv


23 Jan
10 Mar
Difference
Approximate Altitude / km
30
28
25
21
18
40
60
80
Equivalent Latitude / deg
0.5
2.5
4.5
6.5
-1.2
-0.8
-0.4
0
0.4
0.8
O3 / ppmv
0
80
160
240
-60
-40
-20
20
40
N2O / ppbv

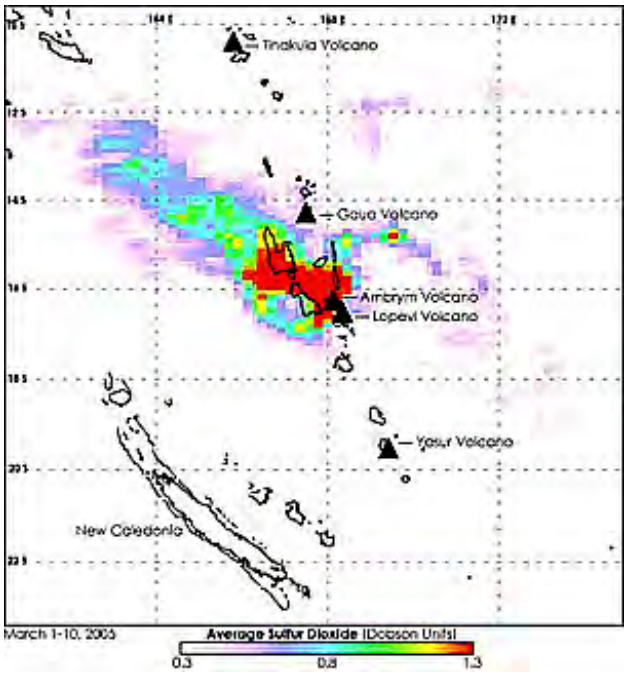


Here are two ozone measurements made by the OMI:

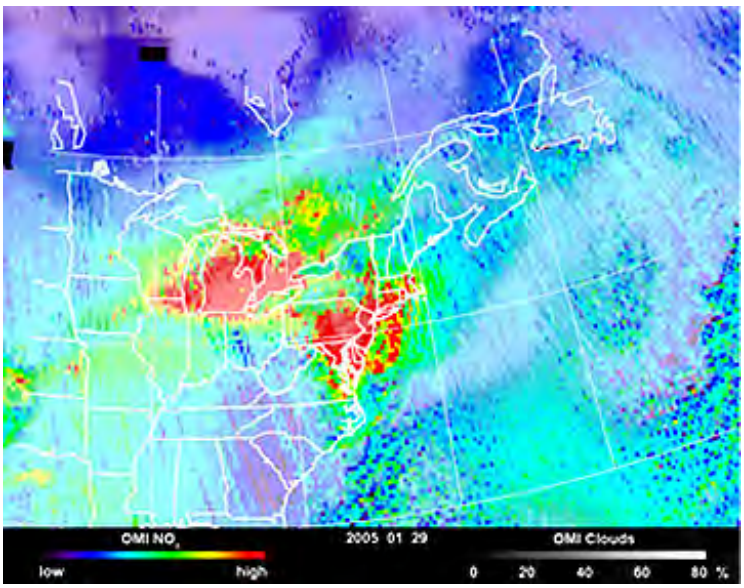




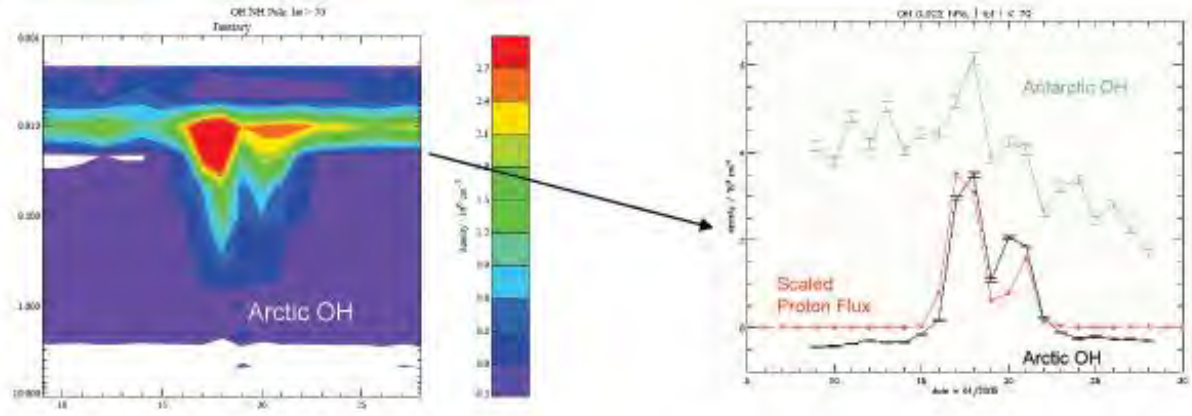
Another example of the OMI data is this map of the a Pacific island group in a small nation called Vanuatu. Here the Ambrym volcano is exuding $S0_2$ gases through vents filled with lake water in large enough amounts to be dangerous to livestock.

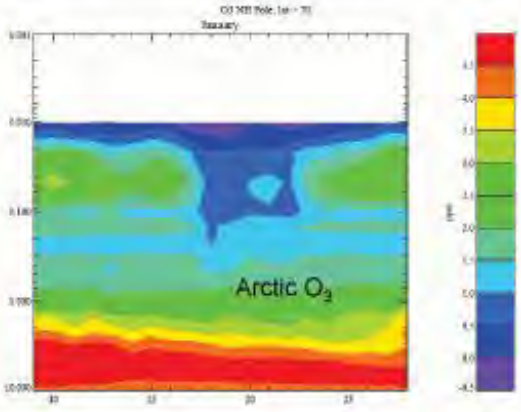


NO_2 concentrations in the eastern United States measured by OMI on January 29, 2005 are plotted in this map:



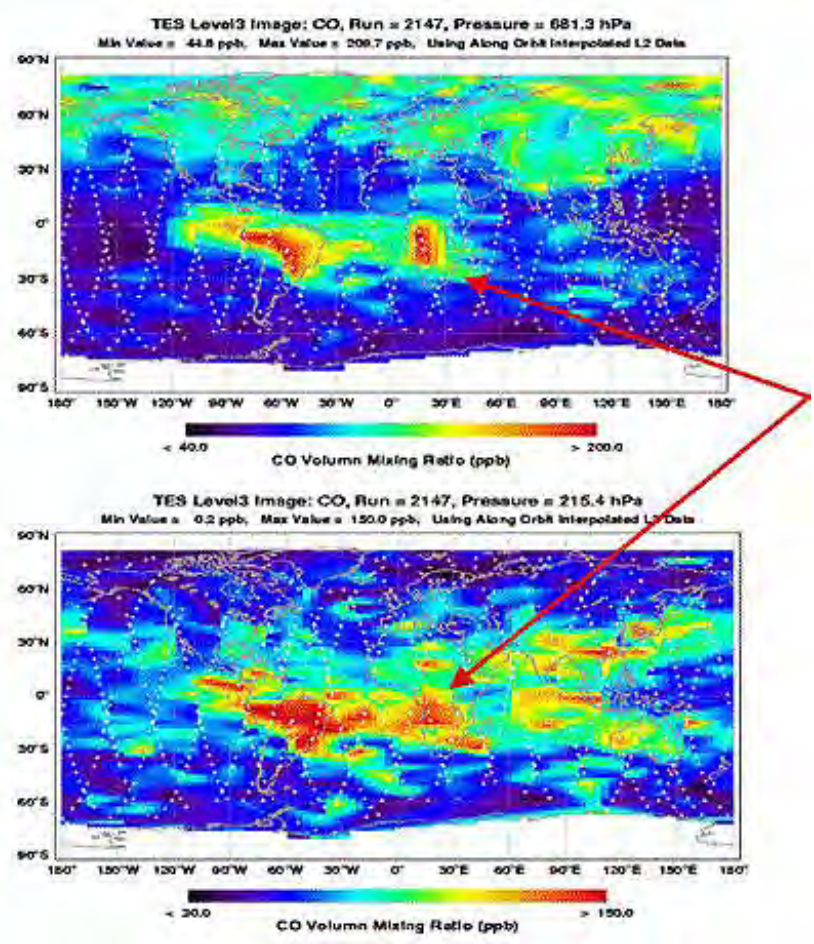
A solar flare on January 17, 2005 apparently had a distinct effect on O_3 and OH in the Antarctic atmosphere which the OMI was able to detect and map; read the description in the diagram for details.



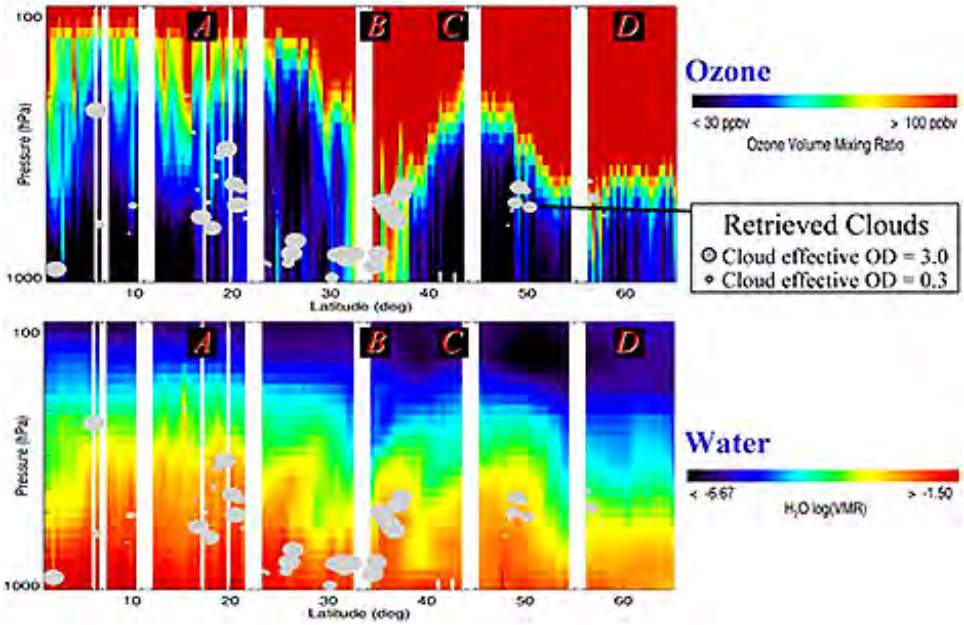


- A major Solar Proton Event occurred on 01/17/2005
- MLS detected an increase in mesospheric OH and a decrease of mesospheric O_3 in the Arctic polar night
  - Arctic results for OH and O_3 are shown to the right
  - Proton flux and OH response for the Arctic and Antarctic are shown above
- MLS also detected an increase in mesospheric OH in the Antarctic polar day
- No corresponding decrease in Antarctic O_3 was found because the background O_3 in the daylight Antarctic mesosphere is already very small

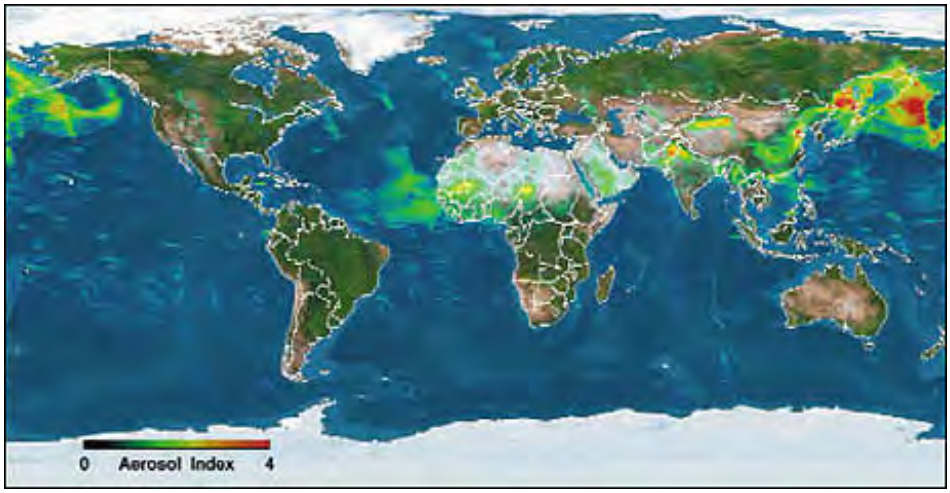
The TES has measured CO levels worldwide at two different altitudes (in atmospheric pressure terms):



The TES obtained vertical variation data for ozone and water along an orbital path from northern Canada to the Yucatan in Mexico.

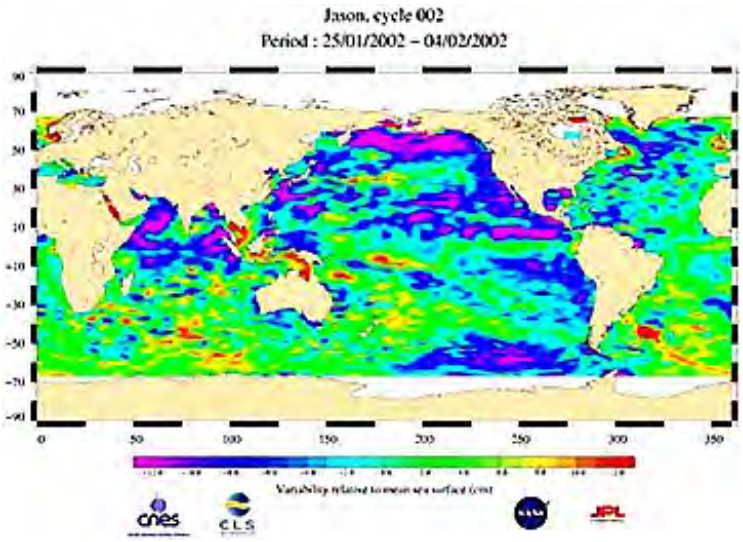


The next is a short-term aerosol map made at a global scale.



The three Earth Observing satellites are all functioning well as of mid-2005. Imagery covering new types of data will be added from time to time.

As the EOS program moves on, some of the companion satellites shown in the diagram on Page 16-7 will be launched. One of these, JASON-1, was placed in orbit on Dec. 7, 2001. A joint endeavor between NASA JPL and the French Space Agency, it will provide improved wind speed measurements over the oceans and will yield data giving more detailed ocean surface topography. Here is one of the first products made public in 2002.



When these EOS satellites are all up and flying, they will represent one type of *Satellite Formation Flying* configuration. Another is the NPOESS program which will be developed in the first decade of th 21st Century. Page 16-11 discusses these innovations.

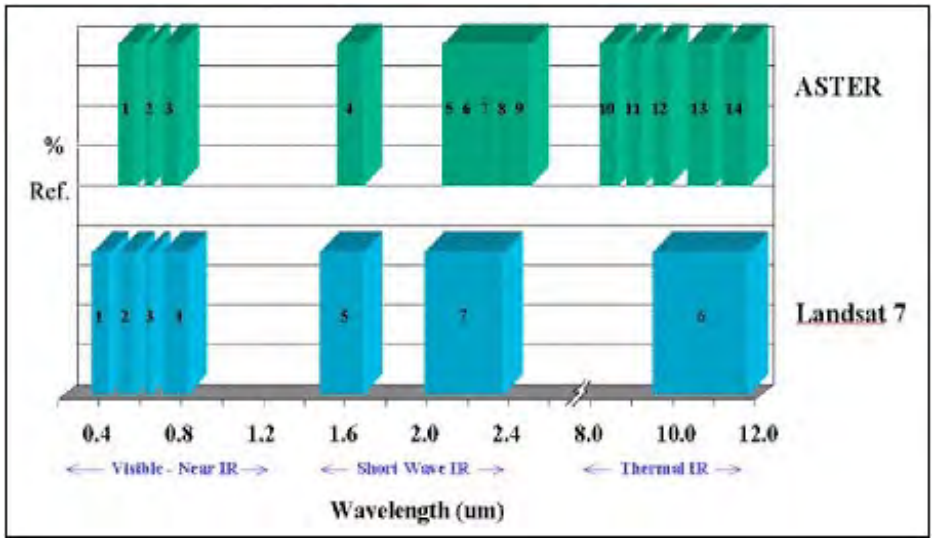
PREVIOUS NEXT

Primary Contact: Nicholas M. Short, Sr.

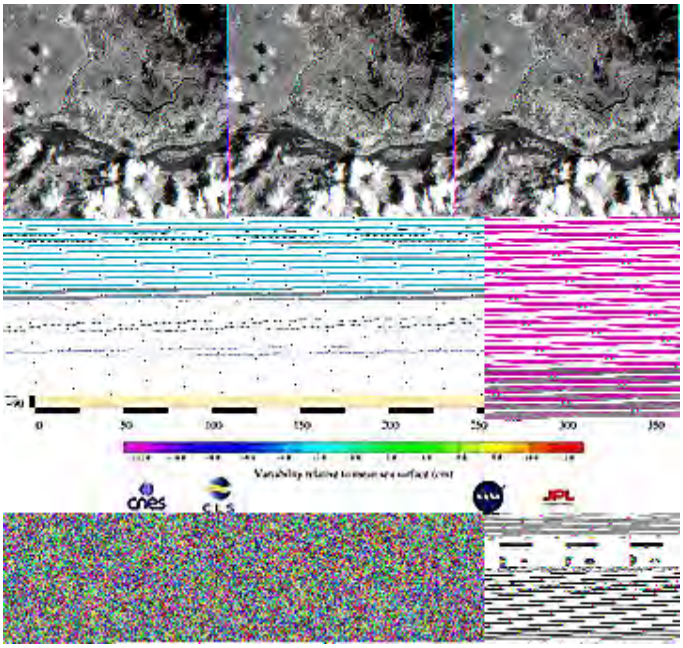
We continue our familiarization with Terra instruments, looking at ASTER, MOPITT, and CERES.

ASTER, MOPITT, and CERES

ASTER produces images over a range of wavelengths. This plot compares the ASTER bands with those on the Landsat Thematic Mapper; the two sensors are similar but ASTER has multispectral capability by having 5 thermal bands:



The first image we look at is a black and white view covering part of the Visible-Near IR spectrum which depicts a segment of the San Francisco River in Brazil. Note its similarity to Landsat images but at 15-m resolution



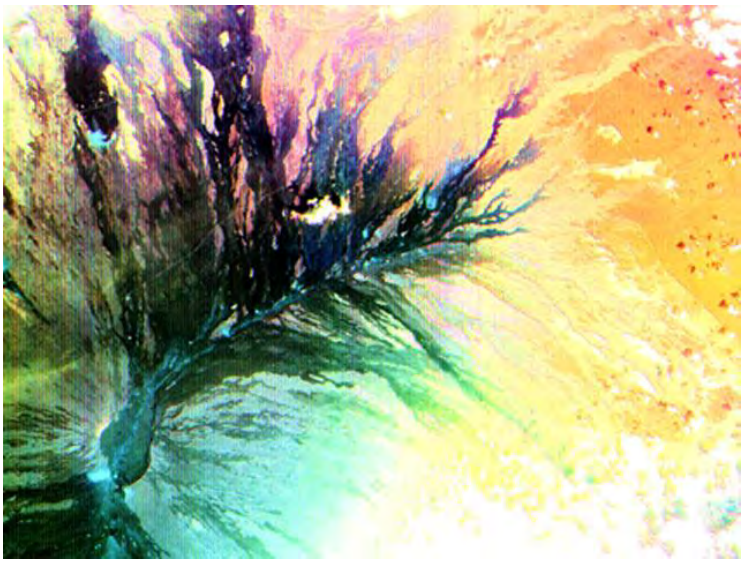
Look next at an ASTER thermal IR image taken at night, showing (in light tones signifying warmer temperatures) the Red Sea and a small land area (dark; cooler) in Eritrea in eastern

Africa. This is similar to HCMM images examined on page 9-8 but the image area is just 60 x 60 km and the ground resolution is 90 meters.

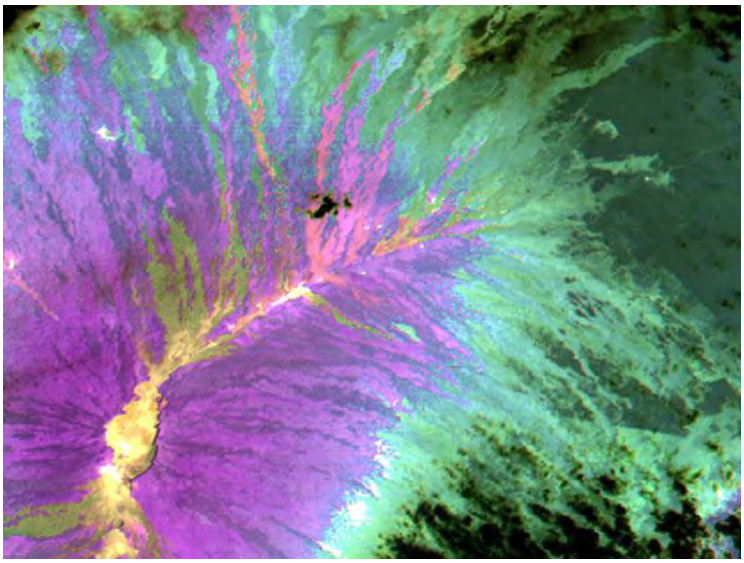
ASTER contains enough wavebands to choose several to make false color composites that appear almost identical to those characteristc of both early and later Landsats. Here is Calcutta, India produced in this mode:

The fourth image is also a false color composite made from three of the Visible-Near IR bands merged wth the higher resolution panchromatic camera image.. The 15-meter resolution is capable of picking out many of the larger buildings in Reno, Nevada.

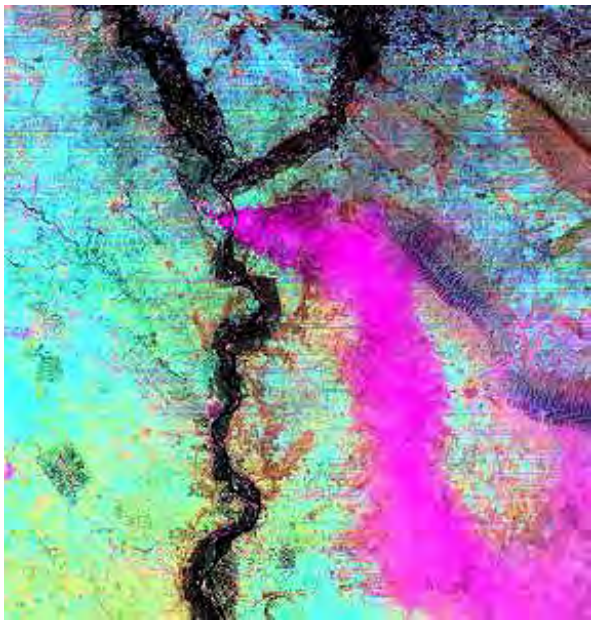
Since ASTER has channels (bands) in both the Visible and Near Infrared (also designated SWIR for Short Wave Infrared), it can produce color composites from both reflected and thermal multispectral radiation. Look first at this image showing the summit and eastern flank of the huge Mauna Loa volcano on the Big Island of Hawaii. This false color composite singles out a dark black (recent) lava flow emanating from a fissure that sends tongues out against an older, more weathered lava surface.



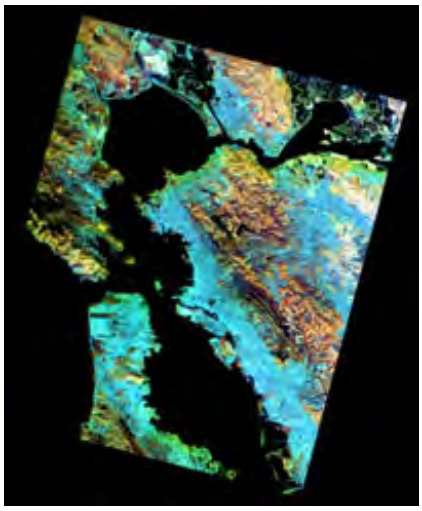
The companion ASTER scene is a multispectral thermal image made from bands at 10, 12, and 14 µm (rendered in blue, green, and red respectively). Note the resemblance to images produced by the TIMS airborne system (page 9-7). The sharp contrast of the black flow, seen above, with neighboring lava is now obscured as the two areas are more uniformly purple, indicating a similarity of composition. The aqua areas beyond are previous (older) flows.

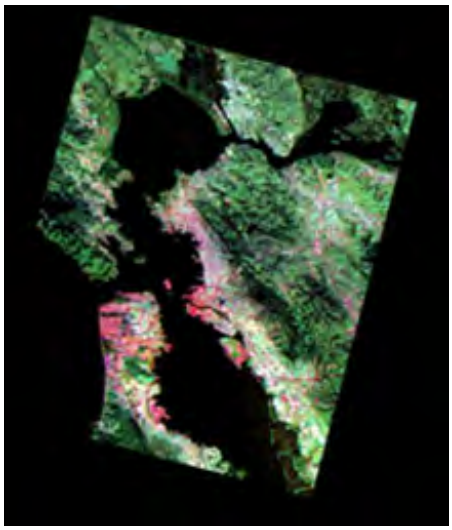


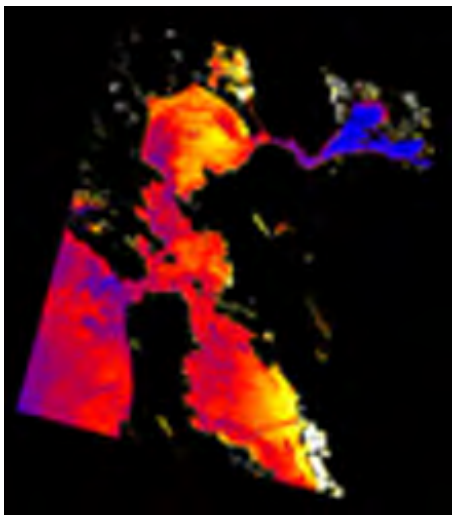
The ASTER thermal bands can actually indicate something of the chemical composition of materials, including gases. A fire near Mosul in the postwar period in Iraq produced smoke that was enriched in noxious SO_2. This takes on a distinctive color (pinkish-purple) in the false color composite shown here:



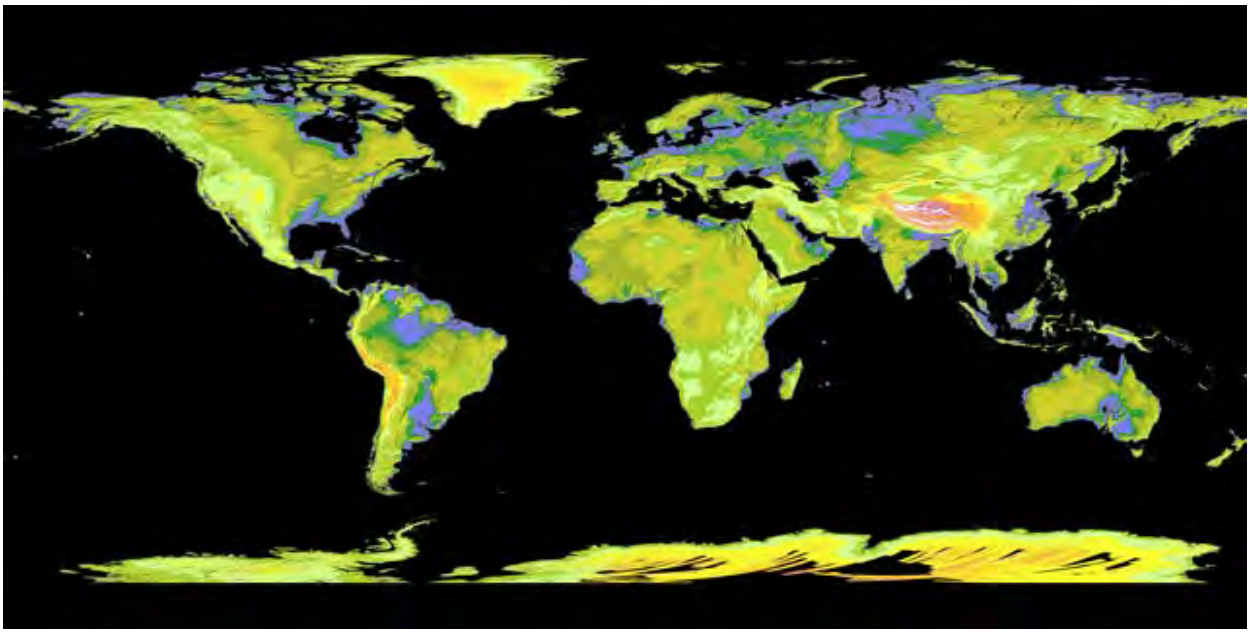
The next three ASTER images display the now familiar San Francisco Bay area. The top image is made from SWIR bands. The middle image is a composite using thermal data with visible bands: the warmest temperatures (red and straw yellow) are found in San Francisco and across the Bay in the Oakland group of cities (more industry). The bottom image blacks out the land and uses thermal data to depict temperature variations in the water, with the usual convention of reds being warmest and blue coolest.



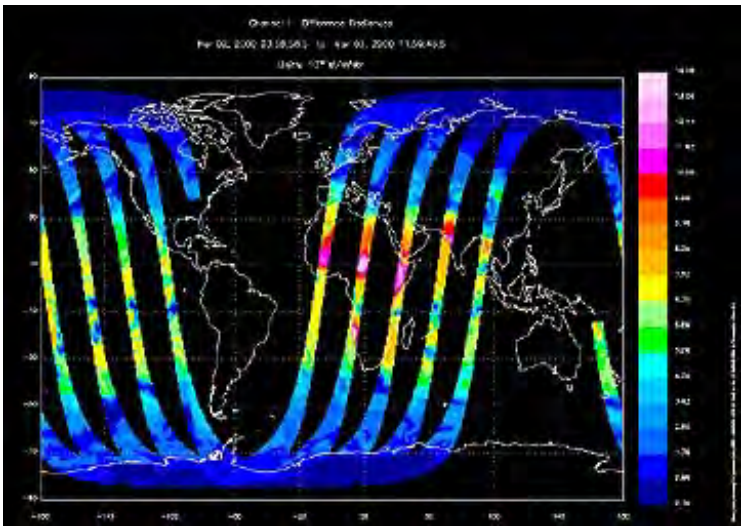




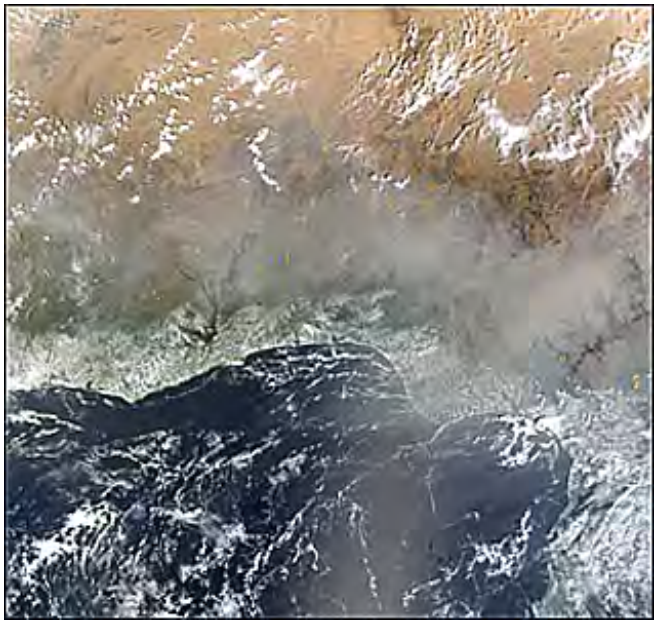
ASTER data are acquired in stereo mode. The Japanese space program has had the task of producing a low resolution global map of the Earth's land topography. Results were released in June, 2009 in two versions, both of which are shown below (Greenland and the Antarctic are shown as high because the thickness of the ice is considered part of the topography):

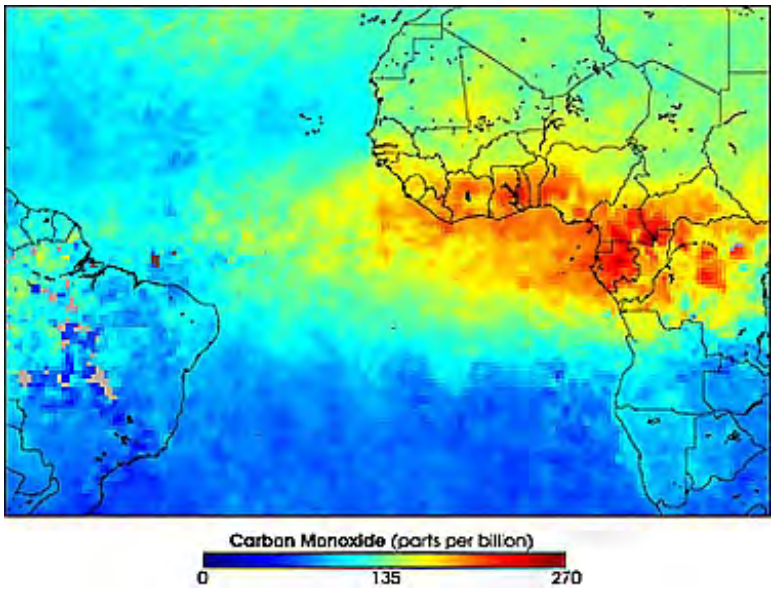


MOPITT's prime task is to measure concentrations of CO (carbon monoxide) (using the 4.3 µm channel) and methane (CH_4) (using the 2,4 µm channel) in a vertical column through the atmosphere. The image below shows radiance variations measured by Channel 1 within a series of strips covering part of the Earth's globe but the black areas represent regions still to be traversed. Reds and yellows are high and blues low quantities of radiance.

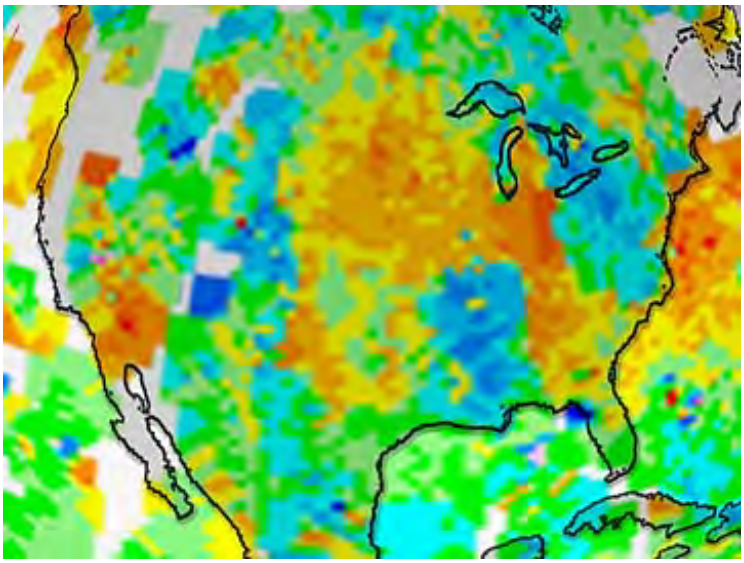


Generally, higher concentrations of CO are associated with either smoke or pollution, or a combination of both. When agricultural burning (a practice common in Africa and South America) on a grand scale occurs, the smoke plumes combine and extend over large areas. Below is a MODIS image of such a plume on February 17, 2004 and a map of CO concentration in parts per billion (ppb) derived from MOPITT data. The levels shown can have adverse health effects:

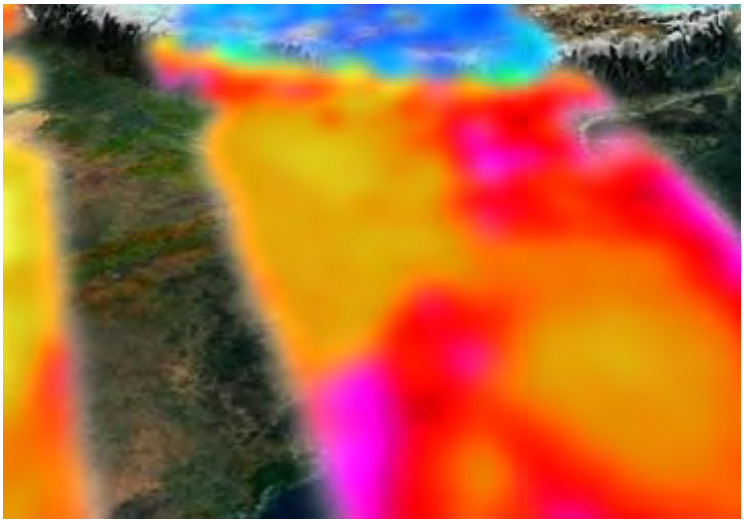




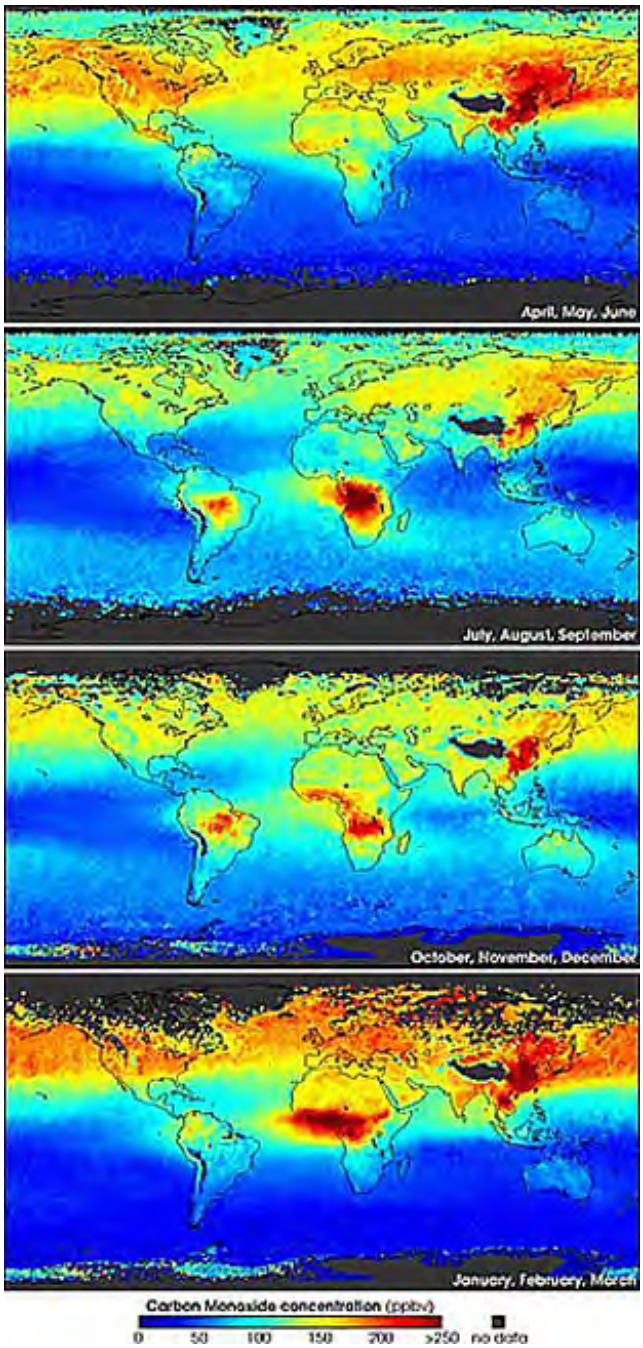
Carbon Monoxide (CO) concentration values are derived from radiance differences with measurements taken at different times by the several channels in this instrument; other ancillary data are then added to determine amounts. The amount of variations can then be converted into maps by assigning colors to the determined ranges. In the next image, the data show CO distribution over large regions of North America; yellows and browns are higher and greens and blues lower CO readings.



A different rendition indicates variations in CO along strips that pass over southern Asia and India:

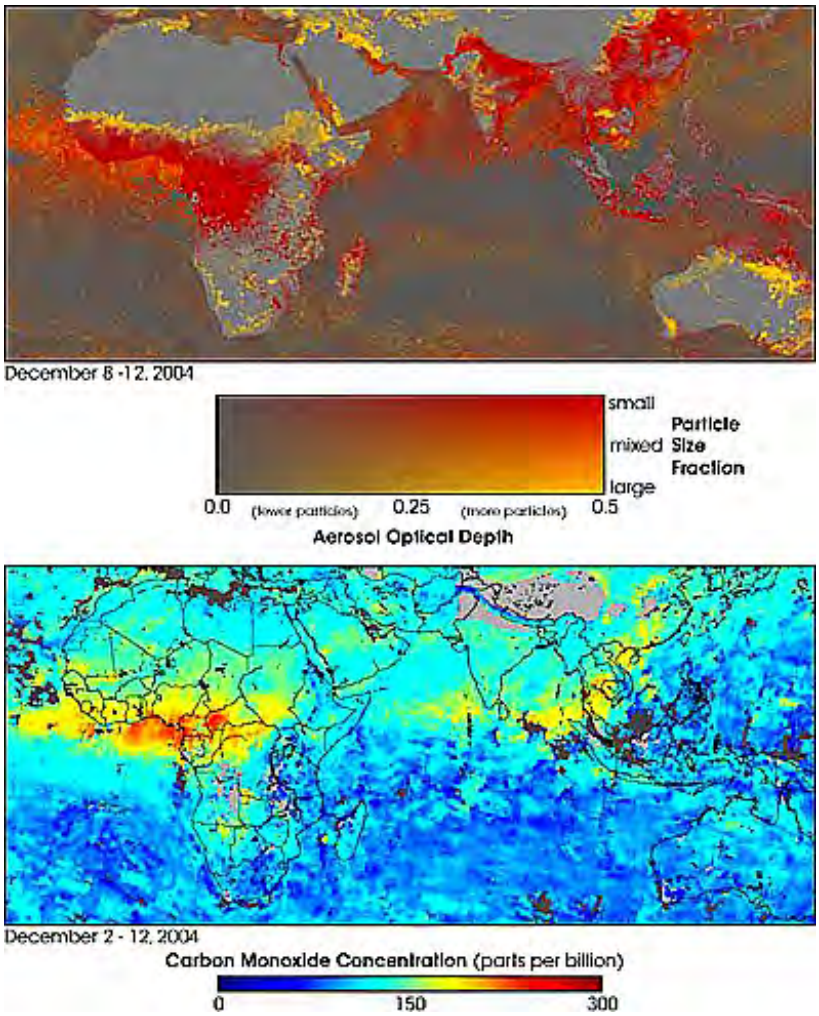


And, over time the CO measurements allow plots of variations over the entire globe, as indicated in this series of 4 seasonal (3 months) maps. Note that the highest concentrations in the temperate latitudes of the northern hemisphere are in the winter months

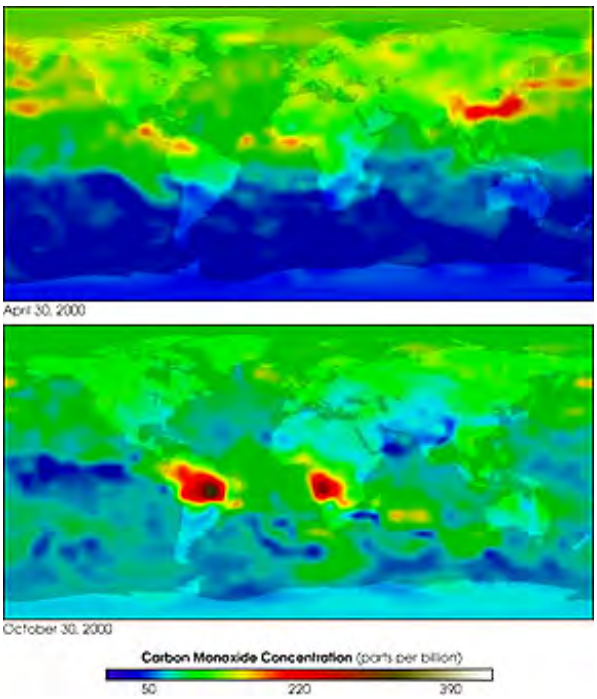


Source: Cathy Clerbaux; NCAR

There is a close correlation between aerosol distribution and increases in CO, as seen in this pair of Terra images:

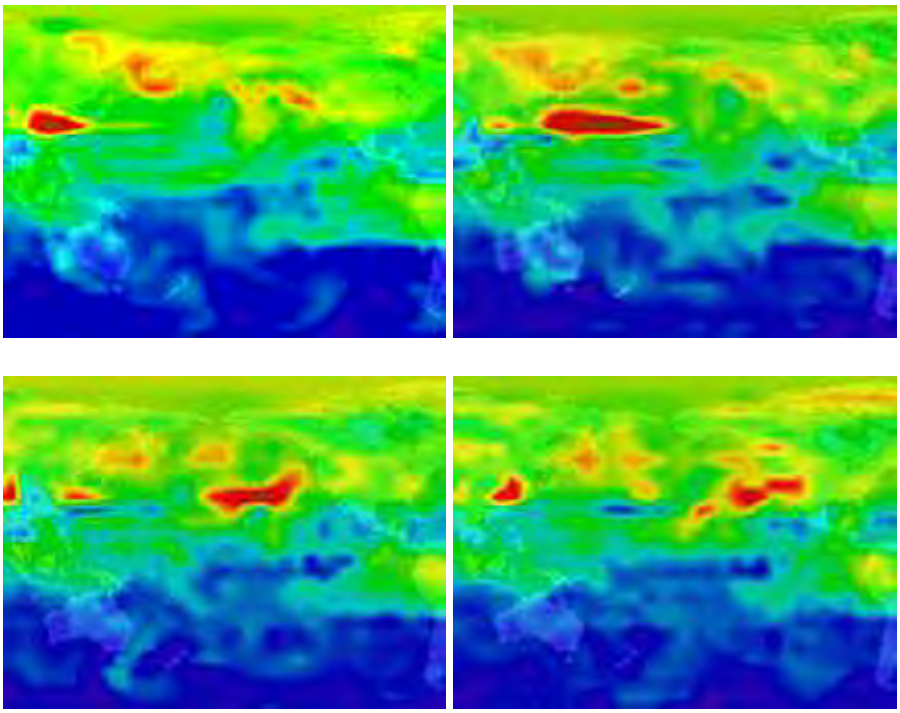


Global studies of CO distribution are leading to a better understanding of why certain areas show higher concentrations. The next two maps taken at different times in the year 2000 show the influence of burning forests and grasslands (especially in the lower image where the orange colors are found mainly in the Amazon and the tropical forests of west central Africa).

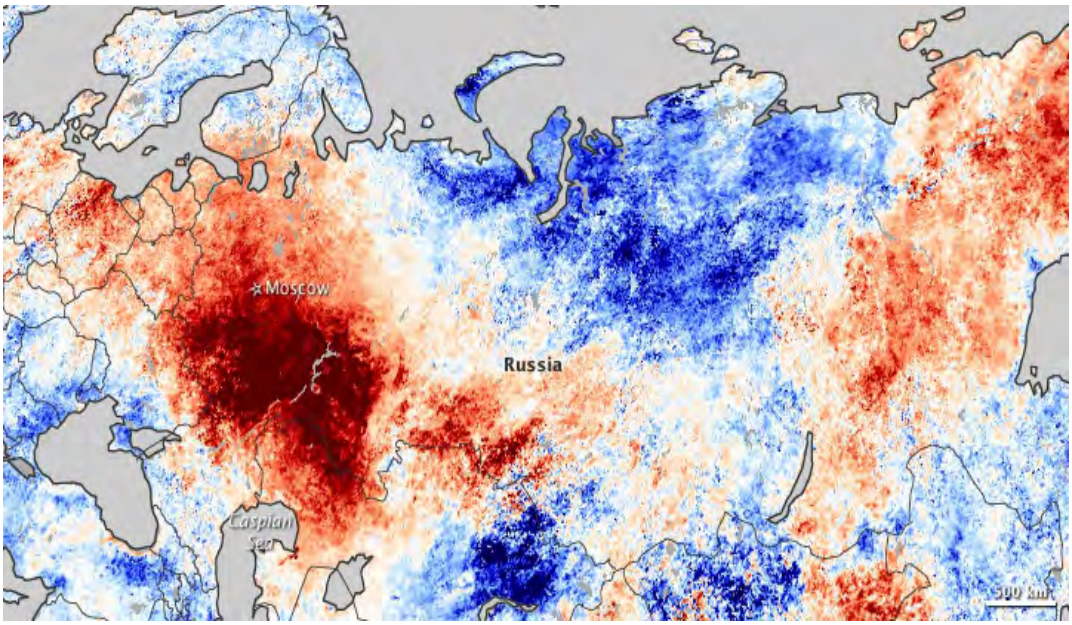


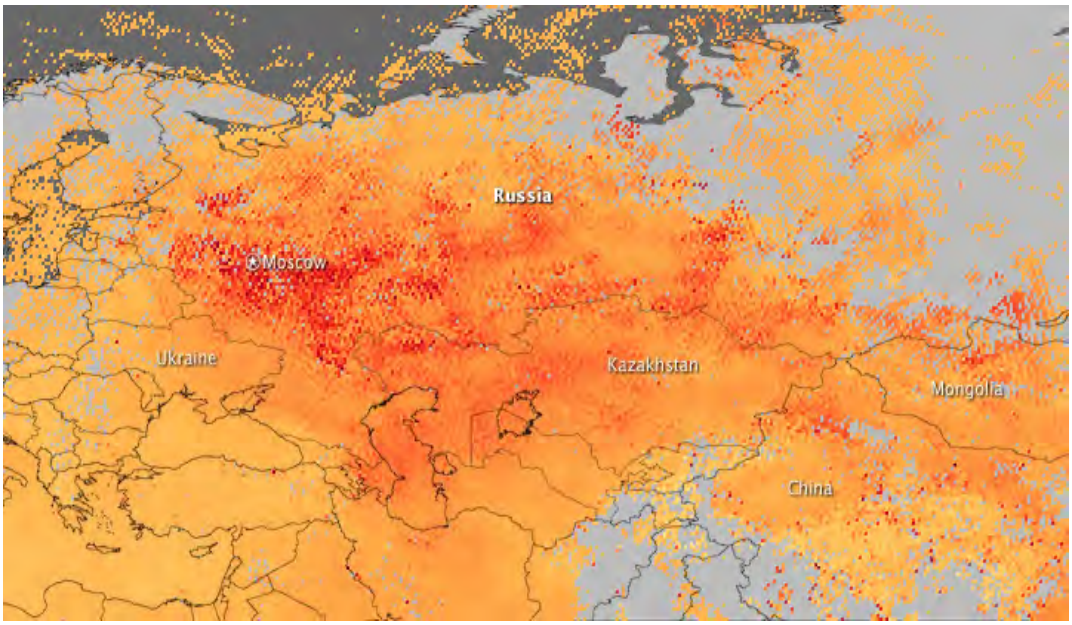
MOPITT can follow over short time spans specific plumes of CO gas that exhibits notably high concentrations. In the sequence below, an elongated plume begins off the coast of Asia and moves across the Pacific, reaching the west coast of North America. The dates of observation are: Upper left = March 10, 2000; Upper right = March 12, 2000; Lower left = March 13, 2000;

Lower right = March 15, 2000.



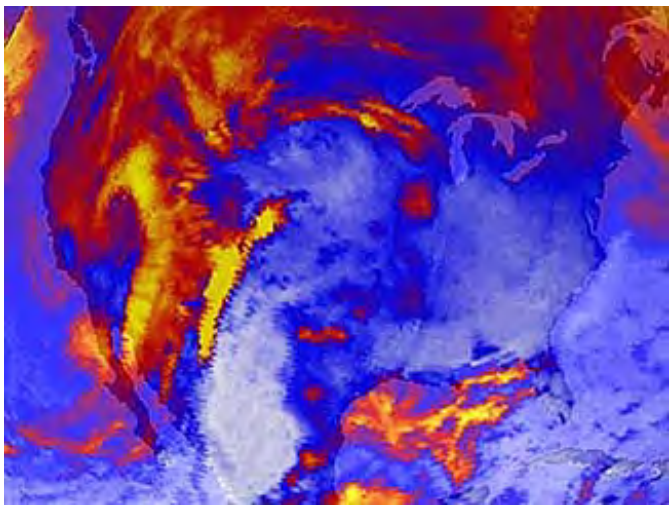
Monitoring of CO from space took on a sense of urgency in July and August of 2010. Large areas in both western and eastern Russia experience, in places, the hottest summers on record. Temperatures in Moscow reached 100 F. and above on several days - almost unheard of historically. The heat led to drought which in turn dried out forest trees and grasslands. Hundreds of individual brush and forest fires caused smoke to spread over wide expanses. As a result CO levels rose to dangerous levels, so that many citizens used masks when outdoors. The next suite of photos summarizes the story of these Russian fires; read each caption for description and details.



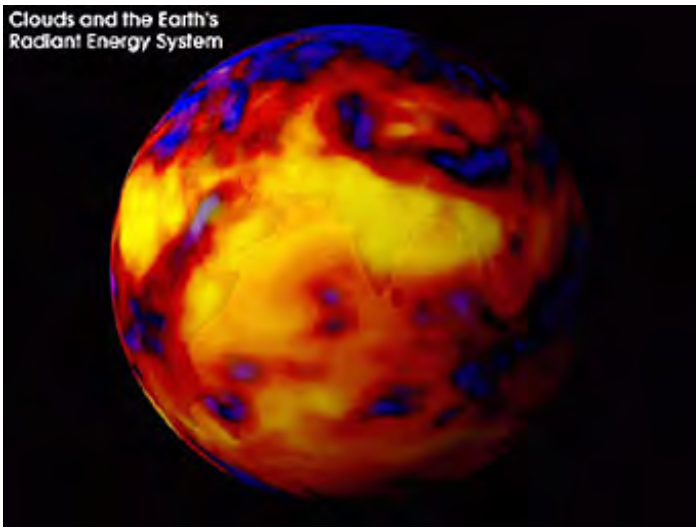

Russia
Moscow
Ukraine
Kazakhstan
Mongolia
China

Moscow
100 km
N

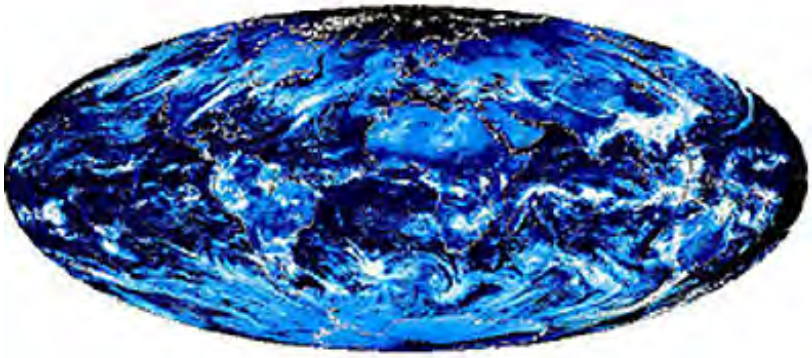
CERES measures both reflected and emitted radiation attributed to the Sun's irradiance. The image below is a map of emitted radiation from the surface and atmosphere, caused by solar heating, for the United States in early April. Reds and yellows are warm; blues and grays cool.

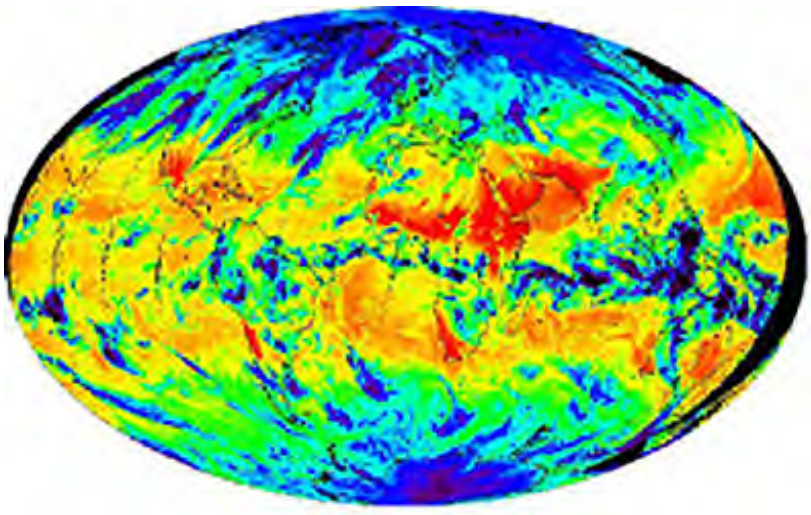


Part of the eastern hemisphere of Earth is shown in this later summer CERES map. The blue areas are primarily caused by large cloud banks, whose temperatures register as cool relative to the much warmer land surfaces.



The final pair of images in this initial data set from Terra's first days of measurements shows, on top, a global map of reflected solar radiation at the top of the atmosphere as sensed by the CERES instrument for a 24-hour period and on the bottom emitted thermal radiation from both the atmosphere and land surfaces (with color assignments of red/yellow to warm and green/blue to cooler).





The next logical step in the EOS series would be to fly a satellite with similar sensors to Terra that passes over the globe in the afternoon. This is the mission of EOS-PM, or Aqua, described next.

PREVIOUS NEXT

Primary Contact: Nicholas M. Short, Sr.

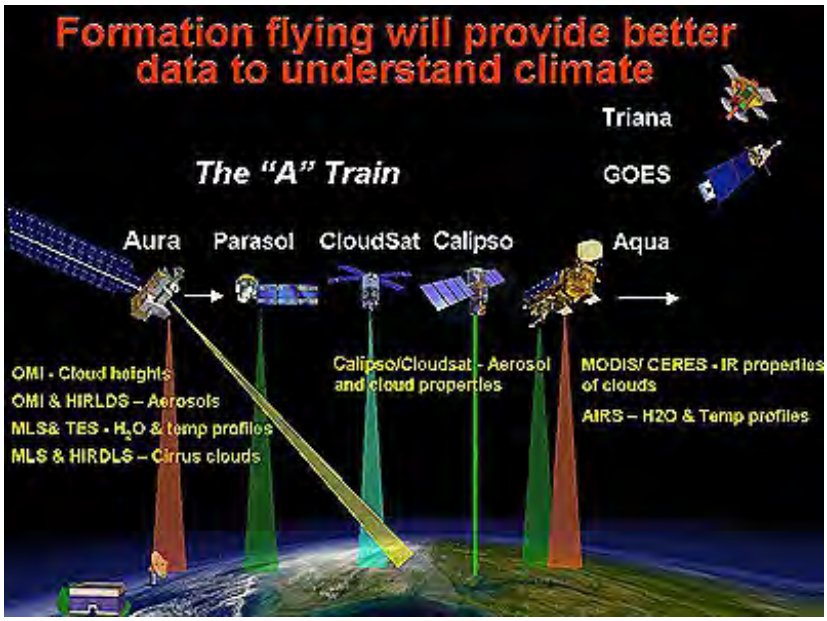
At the beginning of the 21st Century, literally thousands of artificial, unmanned satellites have been placed in orbit around the Earth. Many are still operational. To date, most have been stand alone single satellites dedicated to specific missions. Towards the end of the last century, NASA and other government and private space organizations have begun to group satellites together in various orbital configurations or by meeting common objectives using related sensors. They are also linked by intersatellite communications or through ground stations that direct them as a group. This mode of operation is called ***Formation Flying****. These synchronized programs tend to yield symbiotic results. One important program to be actuated in the first decade of the current century is NPOESS.*

Satellite Formation Flying; NPOESS

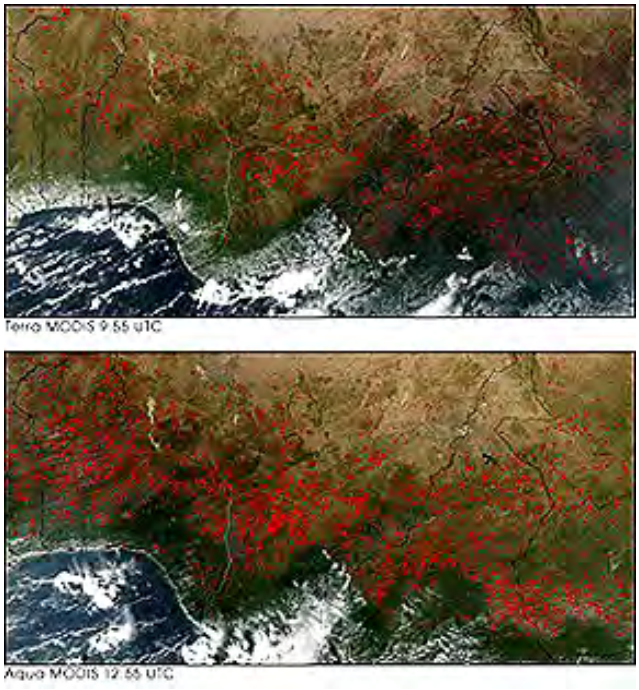
During the preceding Sections we talked about individual satellites, each in its own orbit and making measurements usually independent of other co-operational satellites. But the notion of multi-sensor and multi-platform satellites has already been introduced. Thus, when more than one Landsat or SPOT or IRS, etc., satellite is in orbit and functioning at any given time, the possibility of multitemporal data gathering is activated. Weather satellites are a prime example of having different satellites all looking at the Earth during a given period of time. Data from them can be correlated and integrated to provide information on weather during any specified day. Another variation of such integration occurs when data from simultaneous observations or successive observations or from different platforms are combined by some mode of registration, usually done through computer processing.

In the past 10 years or so, and especially in recent years, the concept of satellite formation flying has crept into the thinking of those seeking different approaches to new programs. Some have described their awareness of birds, such as geese, flying in structured flocks (or bombers in the Second World War) as the inspiration for this concept. Formation flying can be defined as "groupings of duplicate or similar satellites, having sensors in common or are complementary (related), that talk to each other and share data processing (onboard and/or by means of utilizing comparable ground stations and facilities), payloads, and mission functions." The multiple satellites are, in one or more ways, said to be synchronized. Two sites that briefly overview satellite formation flying are found at these SpaceFlight Now and EO-1 web pages.

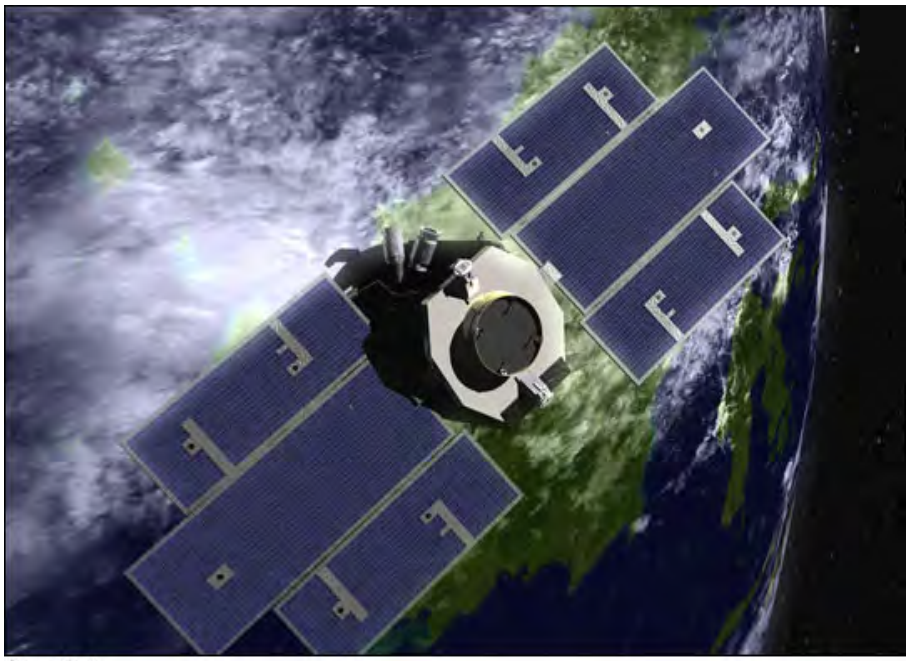
The Terra/Aqua pair and other EOS satellites are part of an integrated series of satellites that are, in effect, prime examples of formation flying or synchronization of satellites dedicated to gathering data/information on climate topics. Terra is in formation with Landsat-7, EO-1, and SAC-C. Aqua, which follows Terra orbital pathways about 3 hours later, has now been joined by CALIPSO, CloudSat, and Aura, and later by PARASOL. This chart summarizes the formation scheme for these satellites:



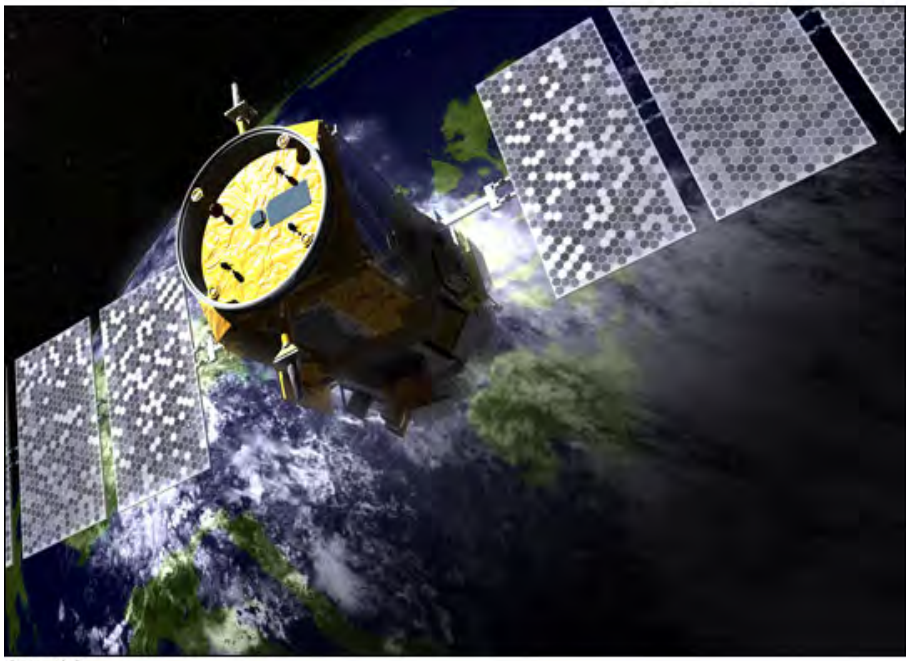
This next image shows the value of flying the Terra-Aqua pair in close formation in which each satellite covers the same area on the same day but separated by several hours when passing over a scene. At 9:55 AM Terra imaged the Central Africa coastal part of western Nigeria. At that time, a large number of small fires deliberately set by local farmers (the "slash and burn" custom) to clear the land for a new crop planting are clearly evident. When Aqua passed this scene 3 hours later, the number of fires had increased significantly.



Another EOS satellite pair - CloudSat and Calipso - were launched together on April 28, 2006. Their orbital positions are adjusted so that they are close-spaced in formation. Here are the two:



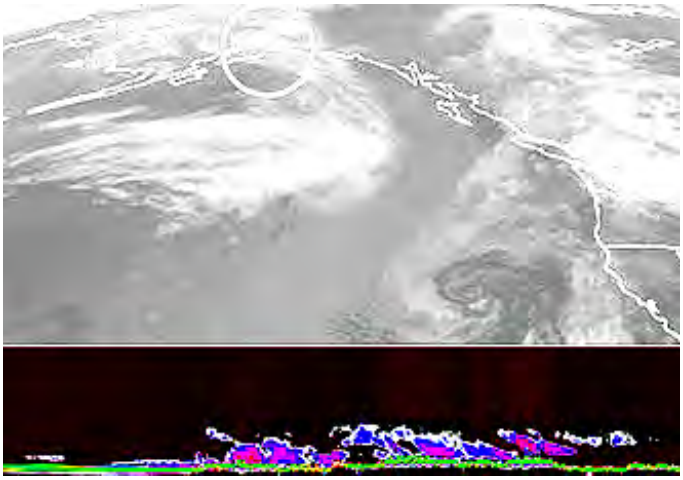
CloudSat

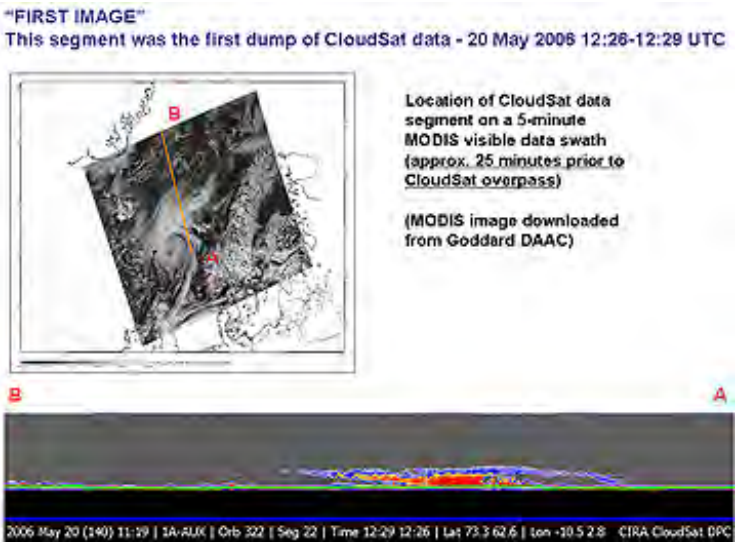


CALIPSO

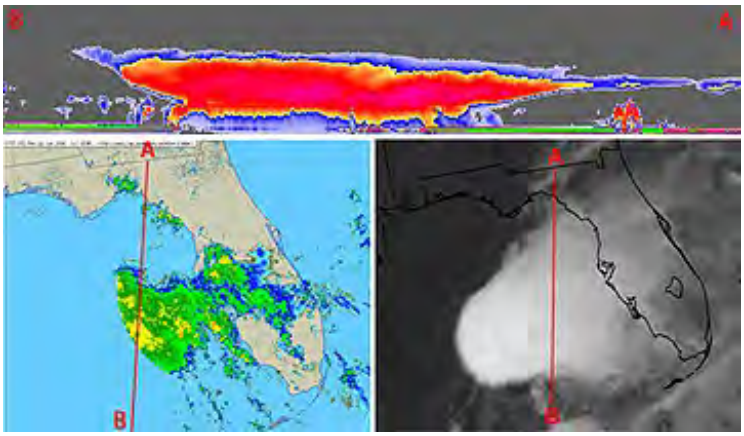
Both examine cloud structure and properties in 3-D. CloudSat uses radar to determine layer stratification; Calipso uses lidar to sense particulate distribution.

The first vertical CloudSat radar images were released in June, 2006. Here are two examples:

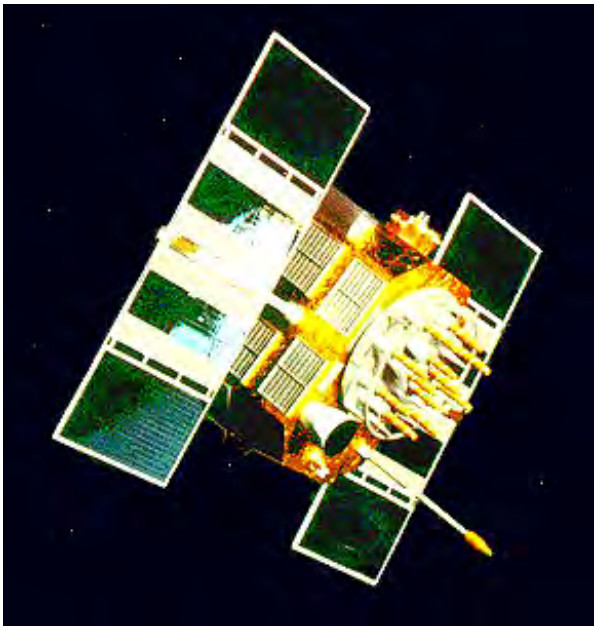




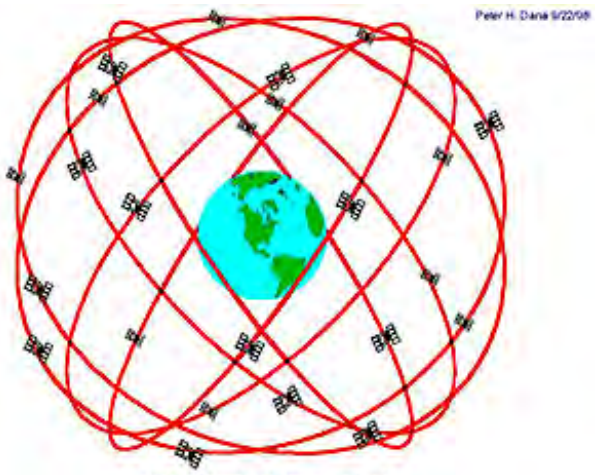
One of the early payoffs from CloudSat's capability was to obtain a vertical cloud pattern slice through the first Tropical Storm, Alberto, as it approached the west coast of Florida on June 12, 2006:



A classic example of formation flying is given by the Global Positioning System (GPS), which we considered earlier on Page 11-6. There we learned that the U.S. is operating a program called NAVSTAR which facilitates location of any point on Earth to a high degree of accuracy (read that page for principles; you are probably familiar with some of its many applications, such as getting directions while driving your expensive car, locating yourself while hunting, or sending your position when shipwrecked at sea). A NAVSTAR satellite looks like this:

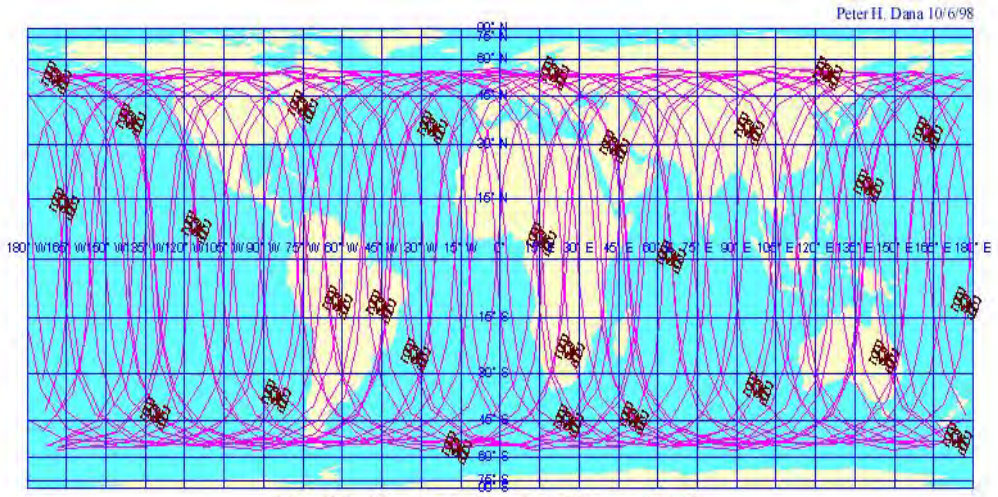


At present there are 27 operational NAVSTAR satellites (24 in the basic array; three others as "spares" or "reserves" to be emplaced as individuals fail). They all support the same sensor system, essentially a radio transponder which communicates with receivers distributed worldwide and linked to parent ground stations. Their orbits are carefully specified and maintained to have mutual regularity in spacing. Thus:



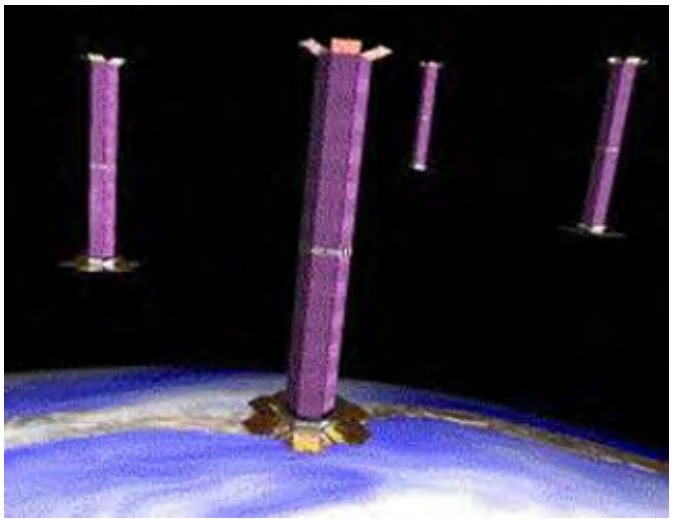
GPS Nominal Constellation
24 Satellites in 6 Orbital Planes
4 Satellites in each Plane
20,200 km Altitudes, 55 Degree Inclination

When plotted on a flattened global projection, the regularity of the orbital paths becomes apparent:

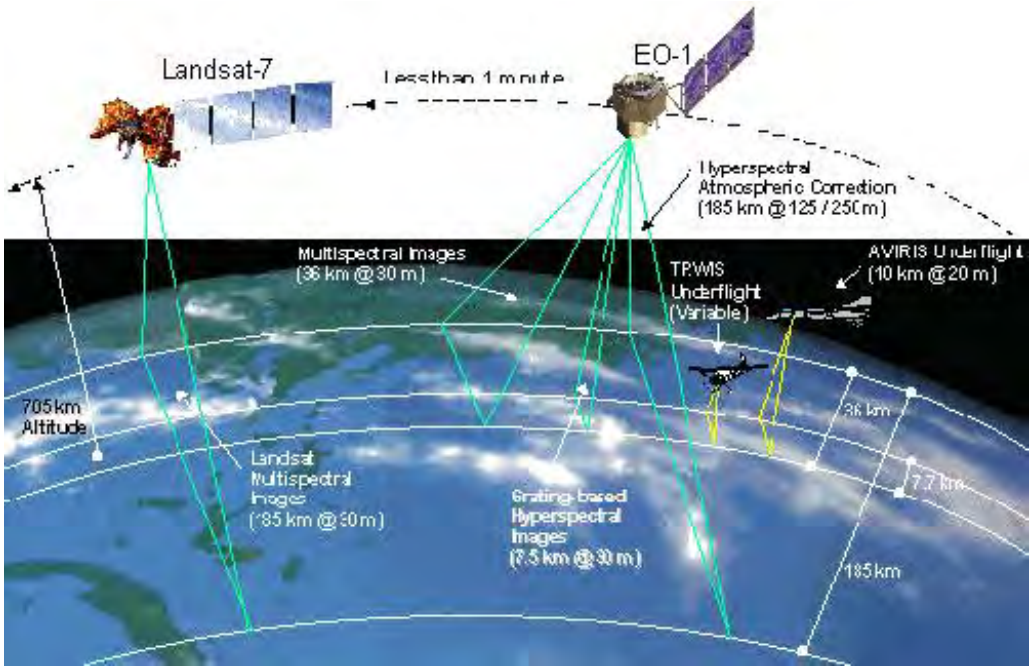


Global Positioning System Satellites and Orbits
for 27 Operational Satellites on September 29, 1998
Satellite Positions at 00:00:00 9/29/98 with 24 hours (2 orbits) of Ground Tracks to 00:00:00 9/30/98

This type of formation flying is known as the "Constellation" configuration. The satellites are spaced at considerable distance from one another but maintain communication with one another as well as with ground stations, through relays. Another configuration of similar satellites is called "Cluster", in which a small group of satellites are located relatively close to each other as they follow orbits that permit them to remain the same distances apart. Here is a suggested arrangement of a group that would comprise the TechMap 21 system:



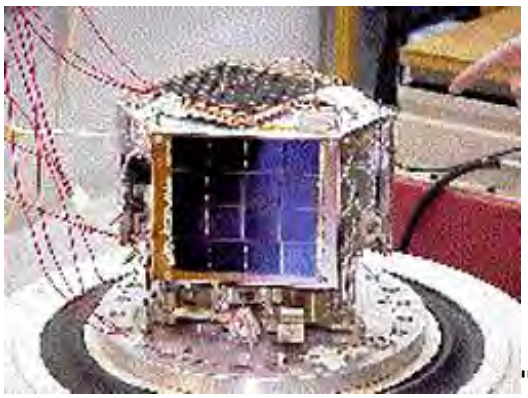
Still another formational configuration can be termed the "Trailing" mode. It is illustrated by the present pairing of Landsat 5 and EO-1 which follows the first satellite in the same orbit, so that EO-1 crosses the same "real estate" at a slightly later time (about 1 minute). This diagram illustrates the two satellites in their common orbit and includes some of the operational parameters.:



An obvious advantage to this is that the trailing satellite not only sees slightly different temporal conditions, but can be equipped with one or more different sensors that provided additional sensing capabilities. In other words, if new sensors with improvements or operating in different parts of the spectrum are developed after the earlier satellite is placed in orbit, it would be next to impossible to add these to that satellite but the second satellite would be so close to the first in time and space, so that the effect is almost as if the sensor(s) has been emplaced on the first one.

One added benefit in grouping satellites in a co-ordinated program is that, under appropriate circumstances, each can be built and launched at lower costs. And, if one of a group fails after orbiting, the cost of replacing it would normally be less that replacing or Shuttle-servicing a large satellite operating as a stand-alone.

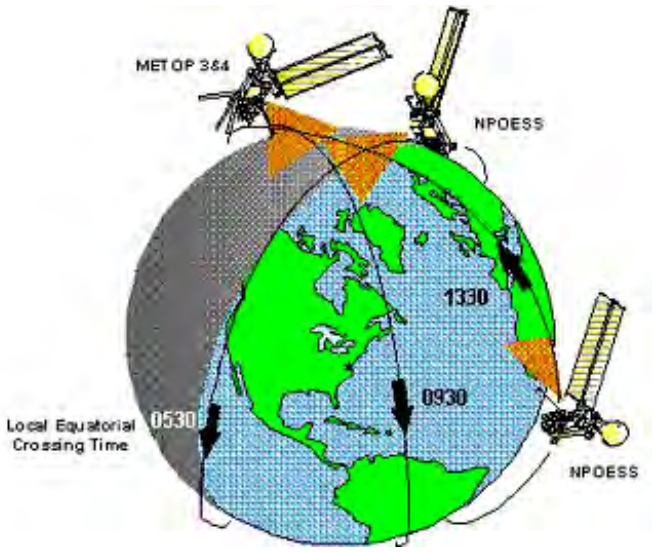
Among candidates for formation flying are very small and lightweight satellites called nanosatellites (under 10 kilograms [22 pounds]), or larger cousins called microsatellites (under 100 kilograms). An example of the first is the experimental communications satellite called SNAP-1 (shades of the first Sputnik!) developed by the University of Surrey, and shown here:


">

Another example of formation flying of small objects will be the Magnetospheric Constellation, consisting of 100 microsatellites whose orbits are set to provide simultaneous coverage on a global scale.

Application categories especially suited to formation flying include astronomy, communications, meteorological, and environmental. In astronomy, multiple satellites at specific spacings open up the opportunity to use them as components in an array that permits the principles of interferometry to increase effective resolution. JPL is planning an astronomical project called SIM-Lite that will use multiple spacecraft to study stars by interferometry. A launch date has not yet been set

A major effort is underway to combine and coordinate programs to gather environmental and meteorological data now acquired independently by several agencies. Thus, NASA, NOAA, and the Dept. of Defense are joining forces to set up NPOESS (National Polar Orbiting Environmental Satellite System), scheduled to be fully operational around 2008. Three satellites, each with up to 10 sensors (most but not all duplicated on each satellite), will be placed in orbits that cross the Equator at times four hours apart. This is the current planned configuration:



The new spacecraft will be modeled after the present day POES program run by NOAA. More about this ambitious scheme can be learned by visiting these online sites: Air Force participation and NOAA, which describes some of the candidate sensors. Tasks considered for this grouping include improved 3 to 5 day forecasts, storm tracking, and crop management. Prior to launch, the proof of concept will be tested by the OSSE (Observing System Simulation Experiment).

To close, this last page highlights a trend illustrated by the EOS series of using pairs to multiples of satellites that fly in regular patterns, spacings, or time separtions, such that, coordinated together, the result is a symbiotic increase in information.

The next page gives some useful references and further reading pertaining to the EOS program.

Primary Contact and Author of this page: Nicholas M. Short, Sr.

Suggestions for Further Reading

Earth System Science: A Closer View - Report of the Earth System Sciences Committee (ESSC), NASA Advisory Council, Washington,D.C., 1988

Managing Planet Earth - A Special Issue (September 1989) in the Scientific American magazine

Earth Observations from Space: History, Promise, and Reality - Committee on Earth Studies, Space Studies Board, Commission on Physical Sciences, Mathematics, and Applications, National Research Council, Washington, D.C., 1995

1995 MTPE/EOS Reference Handbook - EOS Project Science Office, National Aeronautics and Space Administration/Goddard Space Flight Center, Greenbelt, MD, 1995

The State of Earth Science from Space: Past Progress, Future Prospects - American Institute of Physics, New York, 1995

Earth Observing System: Science Contributions to National Goals and Interests - Earth System Science Center, The Pennsylvania State University, University Park, PA, 1995

The Blue Planet: Introduction to Earth System Science - B.J. Skinner and S.C. Porter, John Wiley and Sons, New York, 1995

Distributed Active Archive Centers: Supporting Earth Observing Science in 1995 - National Snow and Ice Data Center (NSIDC) Communication Group, NSIDC, Boulder, CO 1996

Written by: Mitchell K. Hobish email: ***mkh@sciential.com***

This Section surveys one general but vital idea and then addresses how the Earth Sciences community is responding to implementing the outgrowths of this idea. A new field seems to have emerged in the last 20 years. It goes by the imposing name: Earth System(s) Science. But, its ingredients have been around for a long time (some of the components make up the contents of Physical Geography textbooks). The basic notion is this: instead of treating such fields as meteorology, oceanography, botany, ecology, living creatures, aspects of geology, etc. as discrete, self-contained knowledge bases, recognize that they are all part of the same system that controls the Earth, and hence all pointed towards a central common theme, and then cross-link (integrate) the natural shared aspects into a single holistic approach to understanding the Earth. The practioners of this new emphasis on the "Earth System" are now united by participating in joint study efforts such as the International Geosphere and Biosphere Program and by having laid out specifications for highly efficient data gathering, both in/on/above the Earth from the field and through a new family of satellites (the EOS program). This first page explores some of the background to the fundamentals of a coordinated Earth Observation effort and cites several of the important information links to the programs.

EARTH SYSTEM SCIENCE, MISSION TO PLANET EARTH, AND THE EARTH OBSERVING SYSTEM

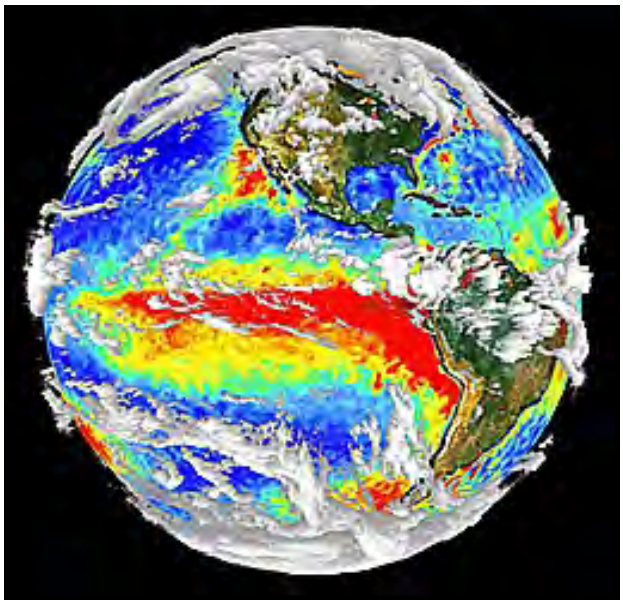
Guest Writer for parts of this Section: Dr. Mitchell K. Hobish *, Consultant

Overview of ESE (MPTE) and EOS; Global Changes

It was probably when Apollo astronauts, on the way to the Moon, first saw our home planet in its entirety from their unique vantage point in space that humans began to view Earth as a single entity, rather than a conglomeration of diverse political entities . To this day, astronauts on their first flights all note the absence of "painted" national boundaries, as they look down upon Earth from orbit. Instead, land, oceans, and clouds dominate the view.

16-1: Look closely at this rather dark view of some part of Earth. What continent(s) are visible? ANSWER

As far back as the Apollo program, even before Landsat, scientists and others connected with the space program realized that earth-looking satellites, as their sensors improved, could provide a wealth of data and derived information about the Earth as a whole and about selected regions by operating in a coordinated way. Different spectral intervals, look angles, resolutions, etc. would provide complementary data sets useful in many applications. Thus, early on, the value of acquiring data from fleets of satellites was so compelling that planners began to devise programs that would supply such inputs. One variant would be to have different sensors on the same spacecraft. This next image was made as a composite using data supplied by land, meteorological, and oceanographic satellites that show clouds, land surfaces, and ocean water temperatures. It serves to introduce you to one of the main ideas in this Section: the Terra space platforms that bears five sensor systems which can gather data integrated spatially and in time which are invaluable in monitoring the Earth's natural systems and the impact of humans on such systems.



'Humans' is the keyword in the above paragraph. To the best of our knowledge, Earth is the only planet in the Solar System that supports life (although organic molecules may exist on one or more satellites of Jupiter and Saturn, and perhaps once on Mars [page 19-13]). Life, in all its myriad forms, virtually covers Earth. No matter where we look for it, we find it. Much of our

planet's geophysical and biological phenomena take place in a relatively thin shell of fluid (the atmosphere and oceans) that is about as thin in proportion to the Earth as a sheet of paper wrapped around a basketball. Most other relevant life activities stay on the land surface,within an even thinner zone of the uppermost layers of soil and rock. And yet, the complex interactions between the biosphere and the geosphere all take place within that thin shell. Just about everything that concerns us as living beings depends upon the integrity of these shells of land, sea, and air.

16-2: Can you think of any place (or condition) at the Earth's surface where life is not to be found? ANSWER

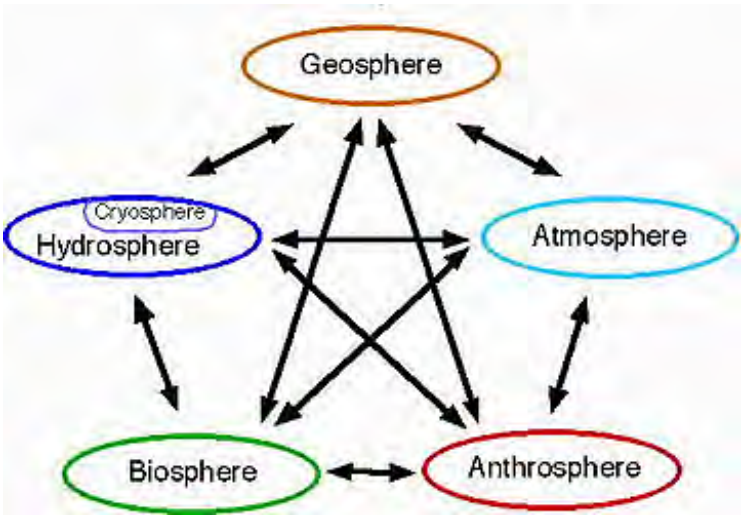
The shells result from eons of dynamic processes that began as the Earth formed. These processes, taken together, constitute global change. Without global change, we humans and much of the rest of the biosphere would not exist, because global change generated an oxygen-containing atmosphere, our protective stratospheric ozone layer, and global temperatures that support life (due to the greenhouse effect) as we know it. Until the last few thousand years, global change has been dominantly a "natural" process.

Recent observations have led scientists to conclude that human activities contribute to global change, that our industrial and land-management practices increase the rate of change of several geophysical phenomena, and that some changes may be deleterious to the biosphere. Nations around the world have banded together in a wide range of scientific and policy-based activities to determine the nature of human contributions to global change and to determine the effect such changes can have on our lives. These are gathered under the umbrella known as the International Geosphere-Biosphere Program (IGBP), which is, in essence, a massive effort to understand and learn to manage the world's environments.

16-3: Think of at least three such vital phenomena whose change can be directly traced to Man's activities. ANSWER

As one of several U.S. government agencies involved in the U.S. Global Change Research Program, NASA instituted the Mission to Planet Earth (MTPE) initiative which it later renamed Earth Science Enterprise (on the Net at ESE), based on the space-based constellation of satellites and sensors known as the Earth Observing System (EOS). This section deals with some of the observations obtained in recent years that have led to the formation of MTPE/EOS, why NASA established it, what NASA designed it to accomplish, how NASA approaches those goals, and what anyone can do with the data. . For some general information about this and related programs, click on NASA's Earth Observatory Home Page. Also, the EOS Program Office publishes a bi-monthly News booklet which can be accessed on the Internet at The Earth Observer or can be obtained through postal mail by emailing to the address found on that page.

We have already used the term "Earth System Science" and will fully explain its meaning on page 16-3. For now we will simply state that ESS is an amalgam of several major components, as depicted in this next diagram. These are well-established fields of study of the Earth. You may notice that 'System' is sometimes replaced by 'Systems'. The plural is favored by some practioners to connote the role of these components each as an interrelated system which act in consort as part of the overall System (this is similar to James Lovelock's idea of "Gaia"). Most use the singular which implies that the components are subsystems. In this sense, the Earth operates as a unified system in which various fields of scientific study concentrate on one of its different parts.



---

* Mitchell K. Hobish, Consulting Synthesist, prepared this unit , with some later additions by N.M. Short

PREVIOUS NEXT

Section 16 Author: Dr. Mitchell K. Hobish, Consultant (mkh@sciential.com)

*At first glance, the themes considered under the purview of Earth System Science (ESS) might seem to be a hodgepodge of unrelated subjects. The common thread, however, is **global warming**. While the actual interrelationships may at first glance seem unconnected on this page, it does give several examples of some of the frequently cited topics in different subjects under study (e.g., volcanic eruptions, ocean parameters (particularly sealevel rise), the Greenhouse effect, polar temperatures; El Niño; the Ozone Hole), while showing how observations from space contribute to monitoring these phenomena. Of special interest are the predicted changes in temperatures, sugar maple distribution, and grain yield, regionally or globally, if the current mean carbon dioxide content in the atmosphere is doubled. Satellite observations and measurements play a key role in monitoring these changes; the ability to monitor large areas at once with short repeat times means that these observers are essential tools in controlling the behavior of Earth's atmosphere to minimize the dangers of negative feedback to the temperature status.*

Evidence for Global Warming:

Degradation of Earth's Atmosphere; Temperature Rise; Glacial Melting and Sealevel Rise; Ocean Acidity; Ozone Holes; Vegetation Response

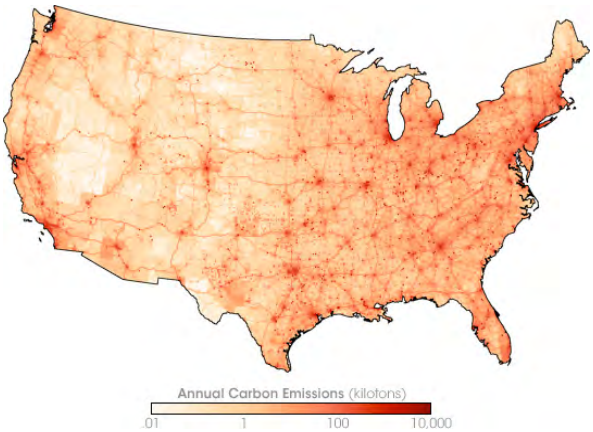
Introduction

"Global change". "Greenhouse effect". "Global warming". The media are full of statements, concerns, guesses, and speculation about these phenomena, as scientists and policy-makers around the world struggle to address recent scientific observations that indicate human activities impact our environment. And yet, each of these is a "natural" phenomenon, as are many others. Hurricanes, droughts, and monsoons all occur without any control by humans to initiate, forestall or moderate them.

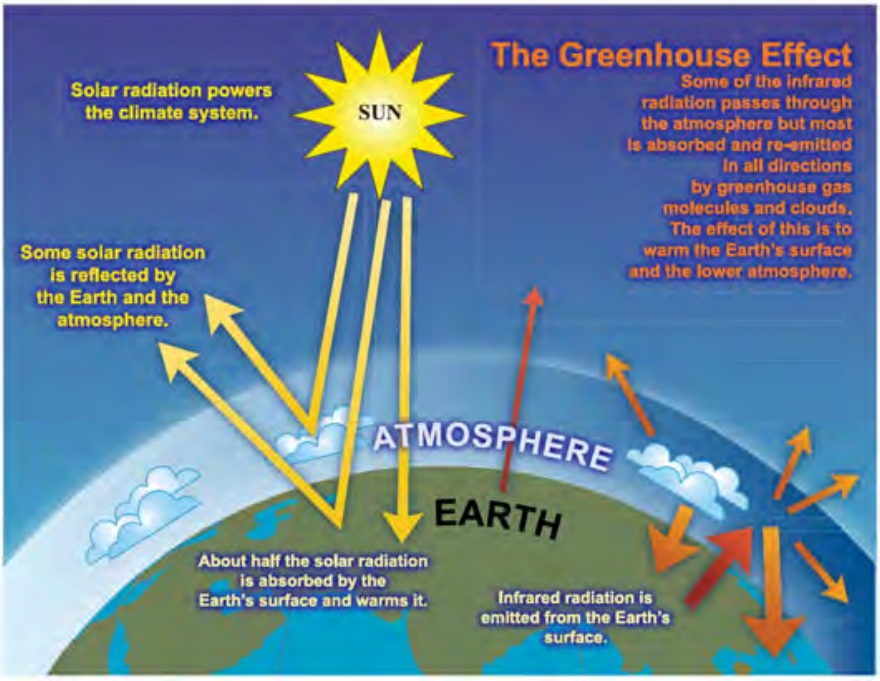
Most readers of this Tutorial already know that global warming is literally a "hot" topic. For those still with relative unfamiliarity, we recommend this Wikpedia review: effects of global warming.

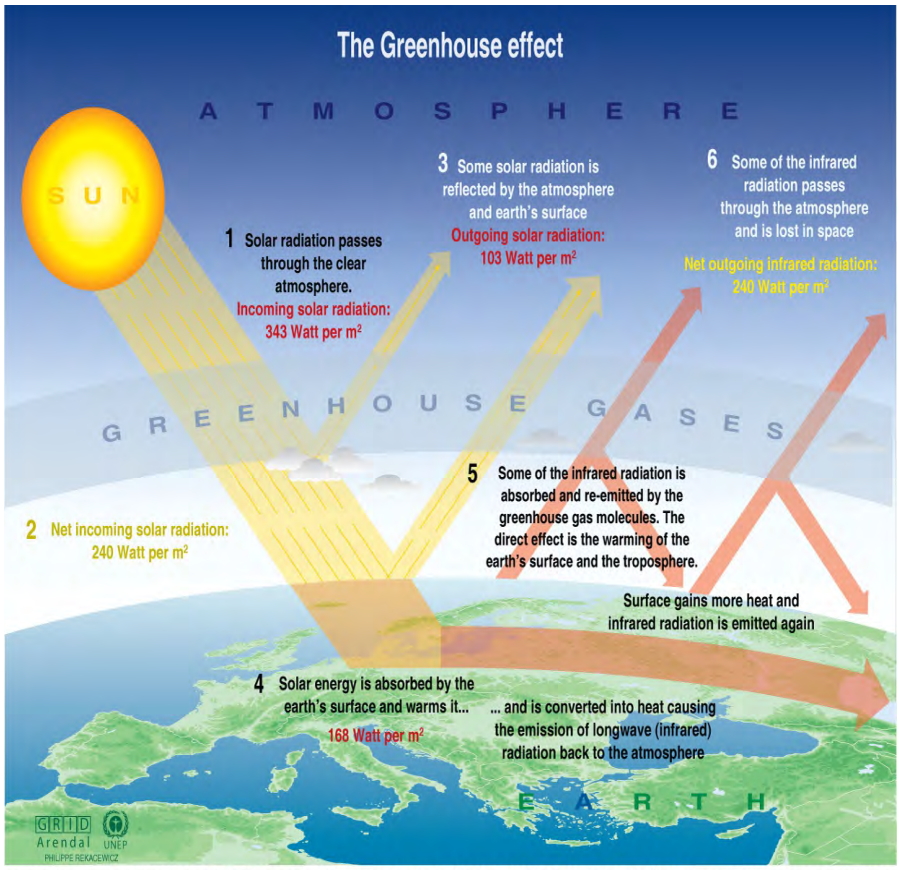
This page considers the nature of and evidence for global warming. There seems almost no doubt that the Earth's climate has been steadily warming overall in the last hundred years: global warming is a fact - a reality. The nagging question - still unanswered to everyone's satisfaction - is: How much of the quantified increase of about 1° C in that time interval is just due to a natural trend probably related to glacial cycles (which themselves are temperature-dependent) and how much is additional warming accelerated by human activity as the industrial revolution reaches rapidly increasing levels? The present page provides an overview of some of the answers to this question. Since it was first written, the writer (NMS) has encountered new lines of evidence - pro and con. Rather than try to splice these into the text, I have elected to add a new page (16-2a), accessed at the bottom of this page.

Most people who follow the news accounts of global warming have come to realize that the primary culprit is human contributions of heat-affecting gases into the atmosphere as these are released by everyday processes such as burning coal to produce electricity and consuming gasoline to run automobiles. The most blamed gas is carbon dioxide (CO_2), but carbon monoxide, methane, sulphur dioxide, and other gases also contribute. The United States is often cited as the principal polluting nation, although now China and India - each with more than a billion people, many of whom are now driving cars - are becoming major players. In the U.S., cities and surrounding metropolitan areas are the chief sources of CO_2, as evident in this map made from ground and aerial sounding monitors but also containing some input from satellite measurements:



These gases are the key factor in the Greenhouse Effect that is warming our atmosphere to the levels we now record. The Greenhouse Effect results from the trapping of solar radiation that reflects from the Earth's surface by these (and other) gases. This is illustrated below in two complimentary versions.





Sources: Okanagan university college in Canada, Department of geography, University of Oxford, school of geography; United States Environmental Protection Agency (EPA), Washington; Climate change 1995, The science of climate change, contribution of working group 1 to the second assessment report of the intergovernmental panel on climate change, UNEP and WMO, Cambridge university press, 1996.

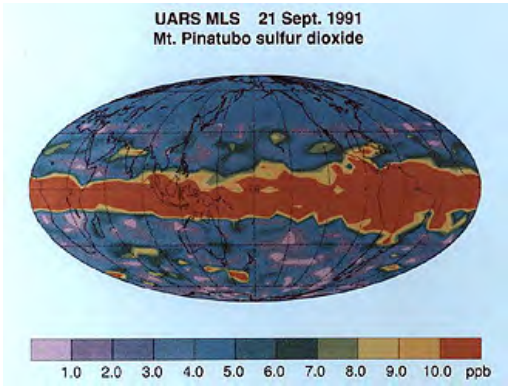
The atmosphere is essentially transparent to incoming solar radiation. After striking the Earth's surface, the wavelength of this radiation increases as it loses energy. The gases that are involved are opaque to this lower energy radiation, and thus trap it as heat, thereby increasing the atmospheric temperature. As these gases increase, due to natural causes and human activity, they enhance the Greenhouse Effect, and may raise temperatures even more. If the climate warms, the vegetation belts will tend to move northward, changing global ecological and biome patterns. Other effects may be discerned in precipitation patterns, sea level changes, and more.

The Role of Natural Phenomena in Warming

To fully understand global warming, one must first understand the operations of the Earth System (described on later pages), since natural sources of temperature-affecting gases play an important role. We can learn about our planet's interacting physical systems by observing the results of such natural phenomena, and use our knowledge to explore human-induced changes. Thus, one factor that may in itself induce and account for some of the regional and global temperature changes is volcanism. Remote sensing is effective at monitoring such events. Consider, for example, the eruption of a volcano, such as Mount Pinatubo in the Philippine Islands in 1991, that happened without human intervention. This volcano had been dormant for more than 600 years.

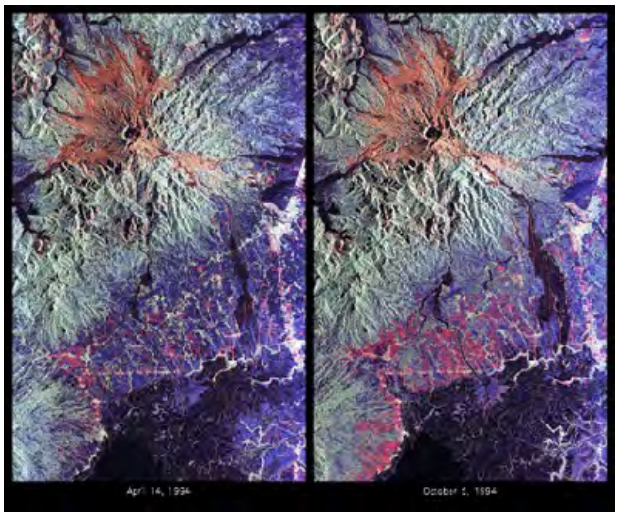
When a volcano erupts it spews millions of tons of ash, debris, and gases into the atmosphere, not to mention the lava flows from some volcanoes. Because of the presence of instruments - on the ground, at the ocean's surface, and in space - meteorologists/environmentalists observed

a cloud of sulfur dioxide (SO_2), emitted by Pinatubo, make its way westward, extending well past India within twelve days of the original eruption. Monitoring has been done by the UARS (Upper Atmosphere Research Satellite). By three months, that cloud had completely encircled the Earth, as shown from space (below), and inside of a year SO_2 particles in the atmosphere were providing gloriously-colored sunsets all over the globe and lowering global temperatures, as well. Clearly, an erupting volcano impacts more people and places than just within its immediate vicinity.



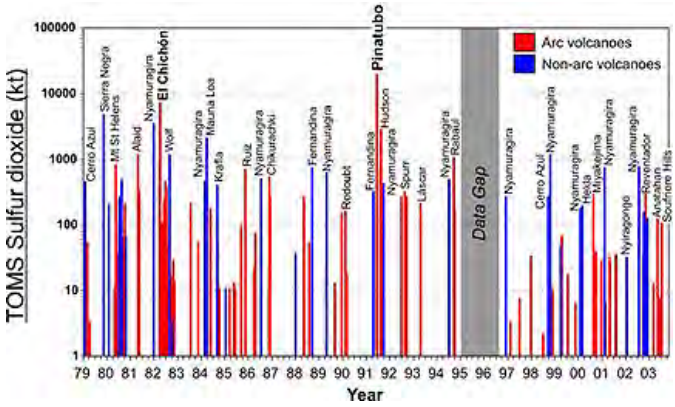
16-4: Why does the Pinatubo ash tend to stay confined to a wide equatorial zone?
ANSWER

Mt. Pinatubo offers a stellar example of how monitoring from space can continually update the status of transient ground events (that, in unpopulated areas [not the case in Luzon] may go unnoticed). Below are two SIR-C radar color composites (L-band HH = red; L-band HV = green; C-band = blue) taken in May (left) and September (right) of 1995, both showing the effects of the 1991 eruption. Ash from an earlier phase of the 1991 event appears in red just above the summit of the volcano.

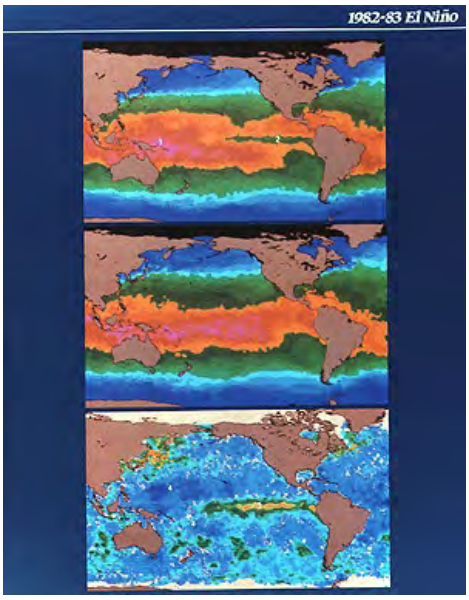


16-5: There is one noticeable change in the right radar image. Find it. Make a guess as to its cause (clue: the volcano didn't erupt between May and September of 1995).
ANSWER

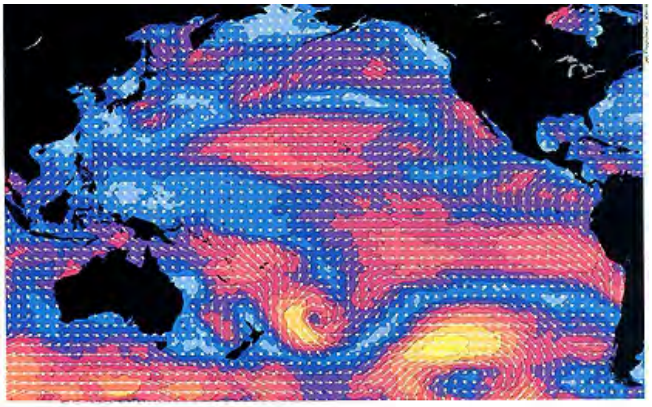
Volcanoes are a main source of SO_2 rises in atmospheric chemistry. The TOMS instrument (see page 14-9) has been making measurments of SO_2 in the air for more than 30 years now. The plot below shows the amounts detected for major volcanic eruptions from those on island arcs and others not on a direct plate boundary for the period from 1979 to 2003:

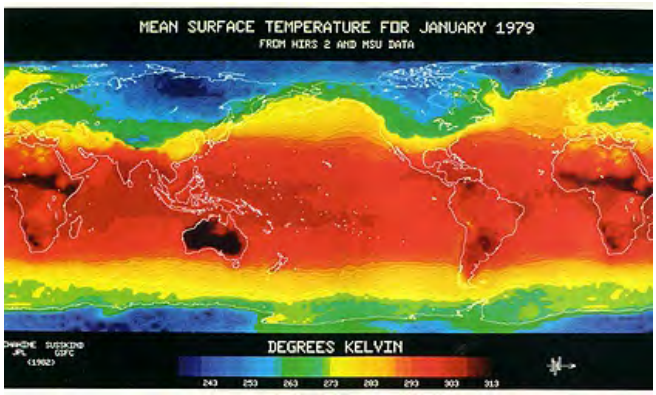


Another fascinating example of a natural phenomenon we know as El Niño, because it occurs with some regularity (although not complete predictability) around Christmas time. El Niño refers to the baby Jesus, whose birthday we celebrate at the end of December. When an El Niño occurs, a pool of warm water from the western Pacific Ocean moves eastward to the western coast of South America. In the process, weather patterns around the world changes often to the detriment of human populations–as do South American fish populations. In non-El Niño years, fish are abundant in this region, because of the cold, nutrient-filled waters. When an El Niño occurs, that cold water flows deep into the Pacific, and fish populations decline dramatically, with concomitant effects on humans whose livelihood depends on those fish.



Other such phenomena abound. Some are readily observable by space sensors, particularly meteorological and oceanographic measurements (see Section 14). As examples, consider these near-global plots. The top from comes from the Seasat Radar Scatterometer and shows prevailing wind patterns over the oceans at supercontinental scales. The bottom one depicts the mean day, night, and day-nite temperatures of the Earth's land and sea surfaces, averaged for January of 1979, from Nimbus 6's High-resolution Infra-red Sensor (HIRS) 3.4 and 4.0 µm channels (page 14-4), integrated with MSU microwave and infrared data. Suffice to say that even without human contributions Earth is a dynamic system, one that changes routinely and often drastically.

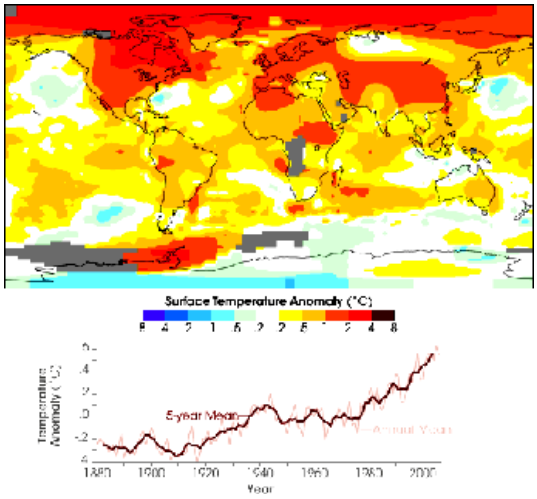


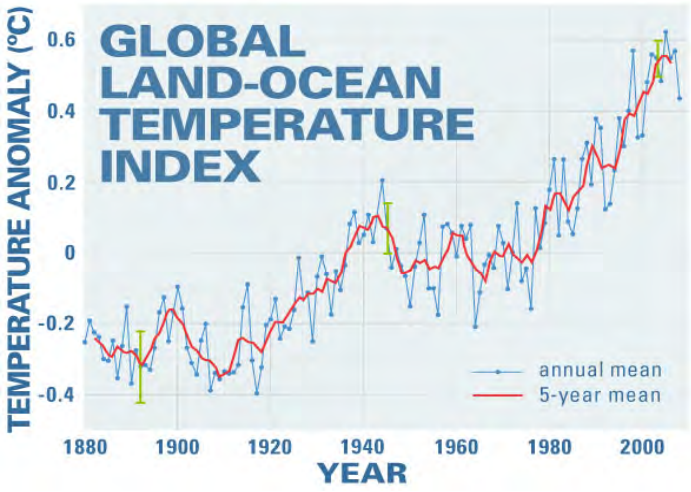


Trends and Patterns in Temperature Changes

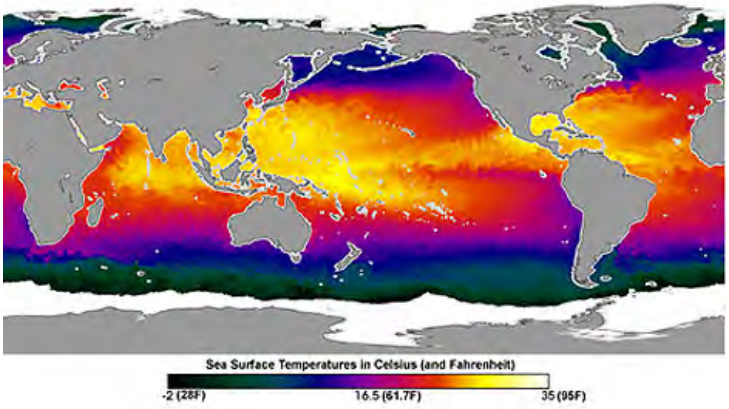
Clearly, global mean annual temperatures are rising, and we continue to monitor this condition with space observations to help settle the question: how much is just a natural trend (e.g., inevitable interglacial warming) and how much is due to man's activities? Calculations show that the burning of fossil fuels (mainly coal, petroleum derivatives, and natural gas) add about 6 billion metric tons of carbon (as the element) to the air annually; each year also, deforestation permits an extra 1-2 billion metric tons of carbon to reach the atmosphere.

As an indication of how much worldwide temperatures have risen in the last few years, this next map of the globe shows the geographic distribution of temperature anomalies, with measurements from various sources as compiled by NASA GISS (Goddard Institute for Space Science), for the year 2006 - the fifth warmest on record. The plot of temperature increase since 1880 is shown but is hard to read. So, beneath it is a recent (Oct., 2009) version that clearly presents details:

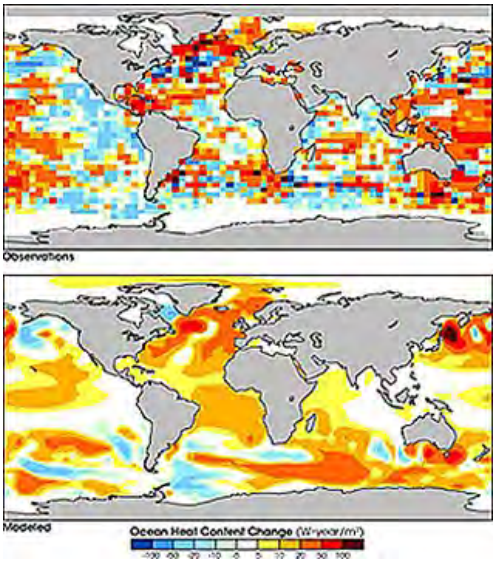




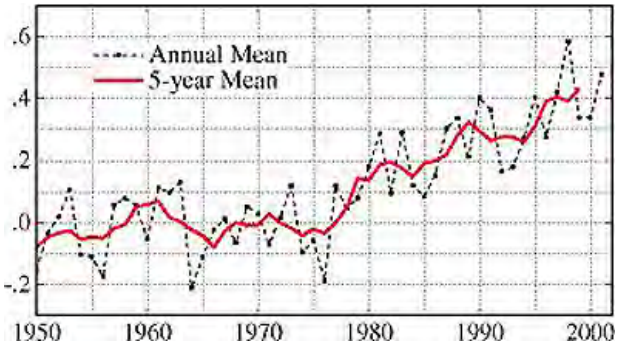
GISS is continually refining its models. Daily inputs improve the credibility of the predictions. Now that Aqua (page 16-11) is operational, its AMSR instrument can provide global data covering very short time spans. This map is of worldwide surface temperatures on August 27, 2003:



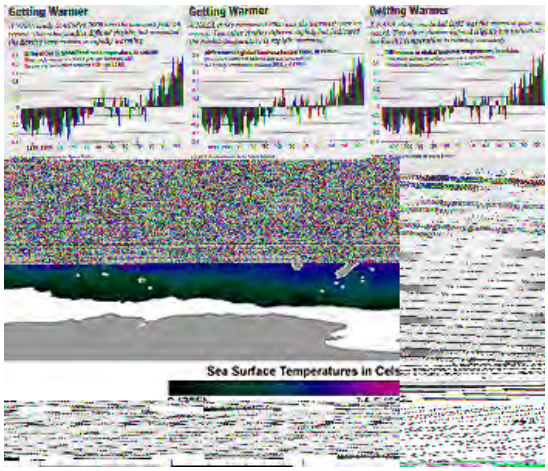
Using temperature data, GISS has now published a map of sea surface heat content anomalies (measured to a water depth of 70 m) integrated over a ten year period from 1993 to 2003. From typical data (top illustration) they have derived a model distribution of temperature variations. Since the upper ocean waters serve as efficient reservoirs for heat storage, the patterns of distribution of excess, normal, and deficient heat (in Watts-year/m^2) thus determined are proving helpful in weather and climate forecasting since the temperatures have strong effects on atmospheric heating.



The trend of temperature changes on annual and 5 year-running averages proves that in the second half of the 20th Century the global atmosphere has been slowly warming - cumulatively totaling about 0.5 ° C, as shown here:

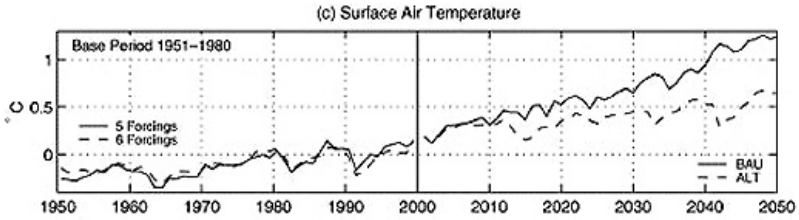


A more recent graph released by GISS covers departures of global average annual temperatures from a 30 year average as baseline over the period from 1870 to 2005 (the latter being the warmest year on record):



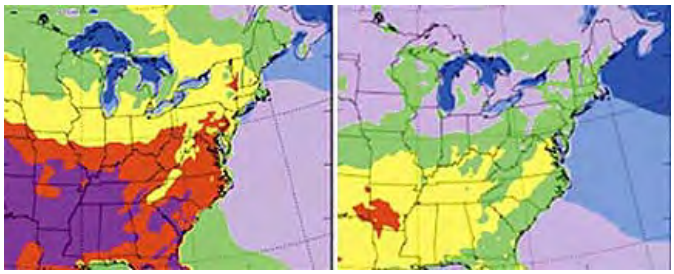
From the Washington Post

GISS has designed its model to predict both sea surface and surface air temperatures up to 50 years in the future. Here is a recently published forecast, which suggests that under current trends the temperature rise will be somewhat greater (more than 1 °C) than the last 50 years:



Other models predict an even more extreme temperature rise - if little or nothing is done to alleviate the causes. One forecast holds the average temperature will increase by 3 to 5 degrees Fahrenheit in the next 50 years.

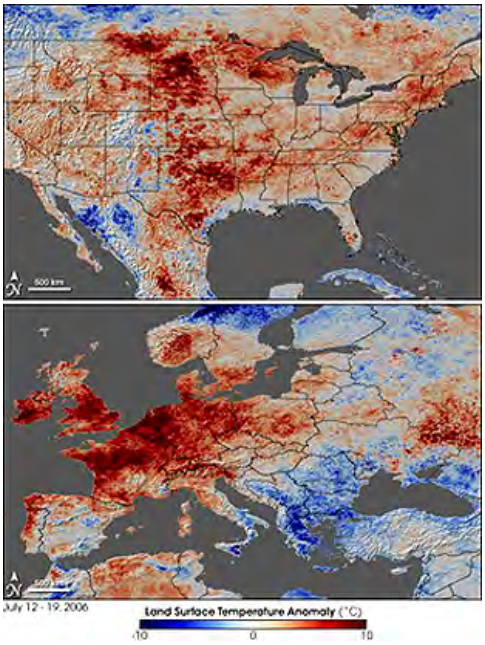
Perhaps most alarming of all is the result of an extended study by NASA's GISS. Using a sophisticated model, their researchers projected temperature rises by the year 2085 for the eastern U.S. For comparison, using this model they produced an estimate for the year 1993 (right map) which was closely matched by reality:



The colors denote peak July temperatures as follows: Violet = 79-85° F; Green = 86-92° F; Yellow = 93-99° F; Red = 100 - 107° F; Dark Purple = 108-114°F. Under the 2085 conditions, rainfall in this region will diminish and vegetation will become much like that around Phoenix, AZ today. Available water will become scarce, so that desalination of the Atlantic ocean would have to provide much that is needed.

That this warming trend is already happening elsewhere in North America is convincingly demonstrated by temperature data for the state of California. In the illustration, temperatures (in Fahrenheit) across the state and in Los Angeles have risen more than 2.5° F in places in the last four decades and more generally almost 4° F statewide in the last 120 years:

So, is there any sign this is happening everywhere? Much debate has started in the middle of 2006 because of very intense heat waves during July-August in both the U.S. and Europe (the latter seldom experiences the extremes that occurred then) as shown in these plots obtain from metsats:



These excess temperatures point up the problem that both advocates and disbelievers of global warming face in interpreting the abnormalities. Thus: Does this period of high temperatures (St. Louis, MO had almost a week of temperatures at or above 100° F) reflect the onset of the predicted hotter summers or is this just another of the drastic warm spells that have happened infrequently in the past? The "jury is still out", although the unusual European situation seems to favor the warming hypothesis.

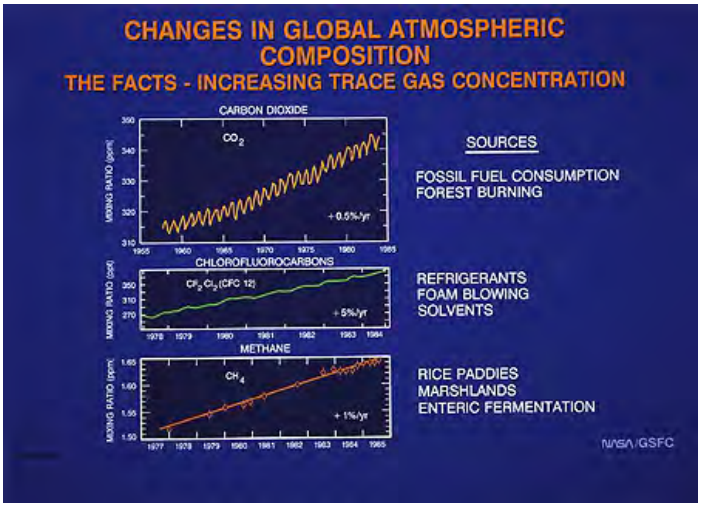
Changes in Atmospheric Composition

In addition to the observed phenomena described above, scientists have accumulated other relevant environmental recordings. For example, levels of several trace gases in our atmosphere have been rising and continue to rise. This change in atmospheric gas composition may be the most serious threat to the global well-being of Earth's environment. Information about this has been nicely summarized in an October, 1997 pamphlet entitled *Climate Change: State of Knowledge*, prepared and distributed by the President's Office of Science and Technology (OST), from which information and selected illustrations on this page have been extracted.

One of these gases, carbon dioxide (CO_2) has been increasing at an accelerating rate since the middle of the last century, after many, many years of essentially stable levels. Why is this? A simple explanation is that the Industrial Revolution began at about the time these increases started. With that social upheaval came the use of biomass and coal for fuel to support these new industries. Burning such material generates CO_2.

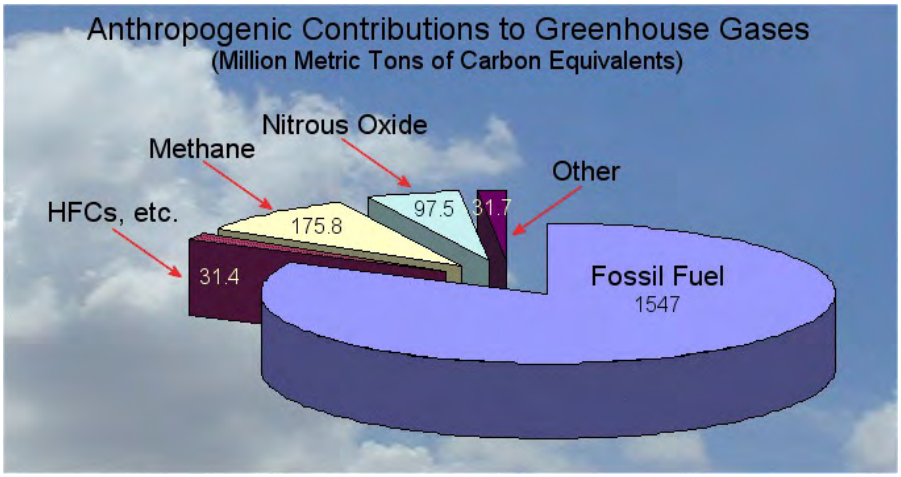
Other trace gases have been rising, as well. Carbon Monoxide (CO) is particularly deletorious. Methane (CH_4), from rice paddy production and enteric fermentation, is increasing, as are

chlorofluorocarbons (CFCs) that have been used for many years as a refrigerant and to produce foam. Methane is a much more potent greenhouse gas that CO_2.



16-6: The carbon dioxide curve shows regular oscillations (much like a sine curve). Why? With what do these three curves best correlate? ANSWER

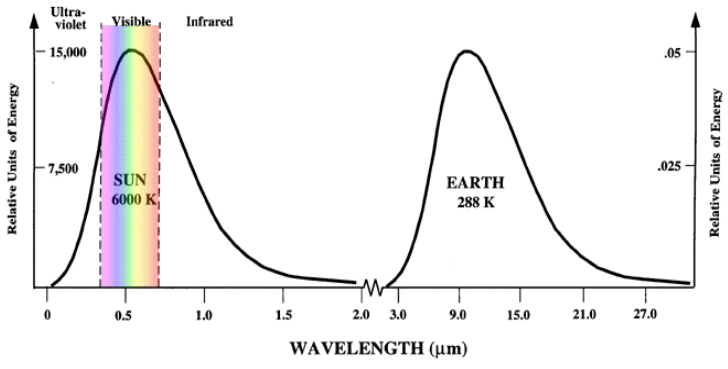
This pie chart shows the relative contributions of the different gases involved in atmospheric contamination:



For those unfamiliar with carbon dioxide and its properties, consult this Wikipedia website

As you will learn shortly, CO_2 is generally considered to be the prime culprit in any supposed global warming. For the next few paragraphs, we will divert from this theme to review some basic facts that provide a background for discussing CO_2's role in moderating the atmosphere:

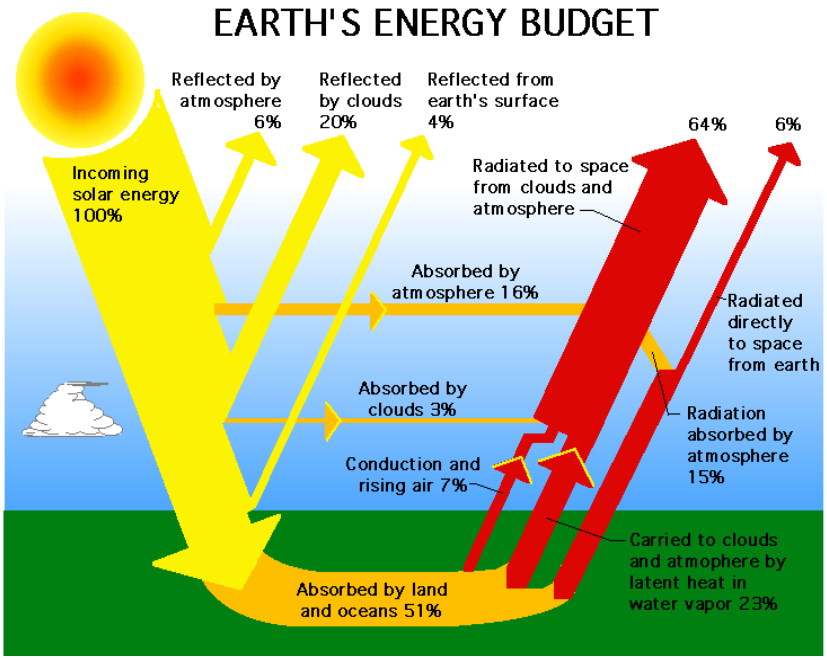
As we saw in the Section 14 subsection that presented a "short course" in meteorology, the Earth's atmosphere obtains almost all its thermal energy from the Sun. This diagram is based on the general blackbody radiation curve (see Section 9). It shows the peak wavelength for radiation input from the Sun and the peak wavelength of the Earth itself as a thermal body of average temperature of 288° Kelvin. (The two curves have the same height in the plots; in fact the height for Earth is greatly subdued, indicating its energy output is much lower.)

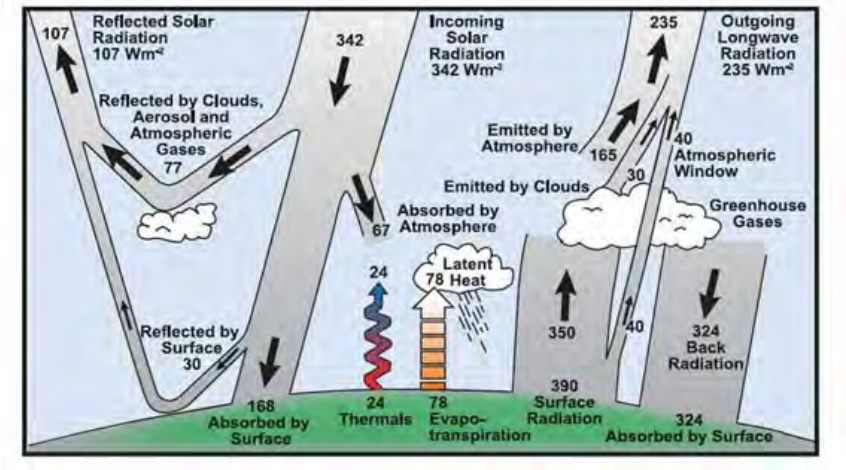


Comparison of the emission spectra of the sun and the earth. Note the huge disparity in the amount of energy emitted by the sun (left-hand scale) and the earth (right-hand scale).

The Earth's Energy Budget

The next two illustrations are plots of the so-called radiation budget that balances incoming with outgoing radiation, first in percents and then in power terms.





The reason for showing the above two illustrations is just to call attention to the complexities involved in the physics of the atmosphere. Various processes and interactions are involved in heating up the atmosphere. Changes in both short-term and long-term temperatures depend not only on adding a single heat-absorbing gas but on other factors that either exacerbate or mitigate the thermal state of the Earth's air.

The Effects of Carbon Dioxide on Warming

This next diagram shows the absorbing bands associated with water and carbon dioxide, the two princpal "contaminants" in the atmosphere besides the primary constituents nitrogen and oxygen. The total amount of water vapor in the atmosphere remains essentially constant although the distribution (locations) at any one time do vary. The amount of carbon dioxide added or subtracted through natural processes (controlled mainly by vegetation) is seasonal but the overall level as annualized is fairly constant.

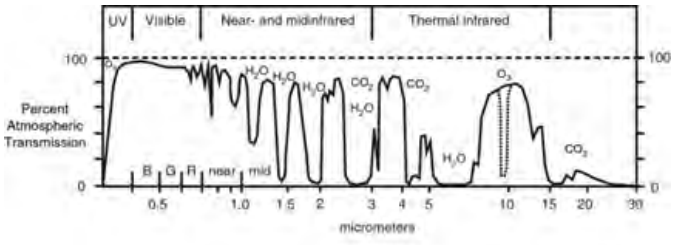
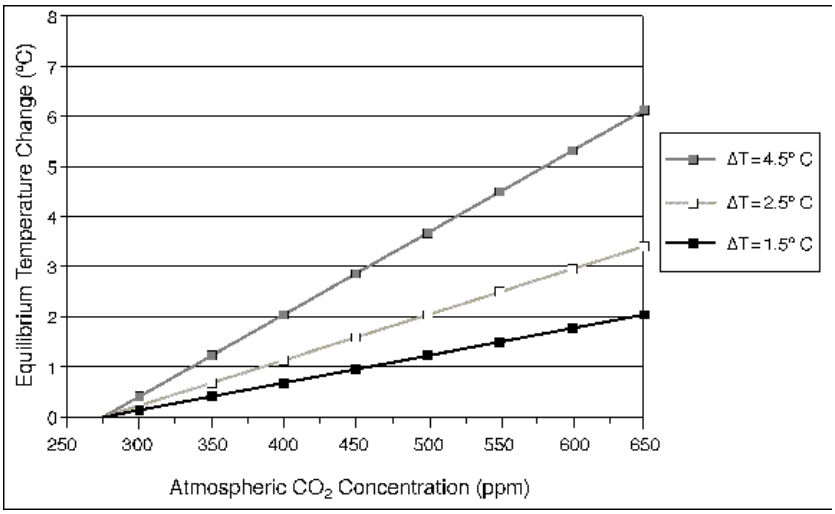


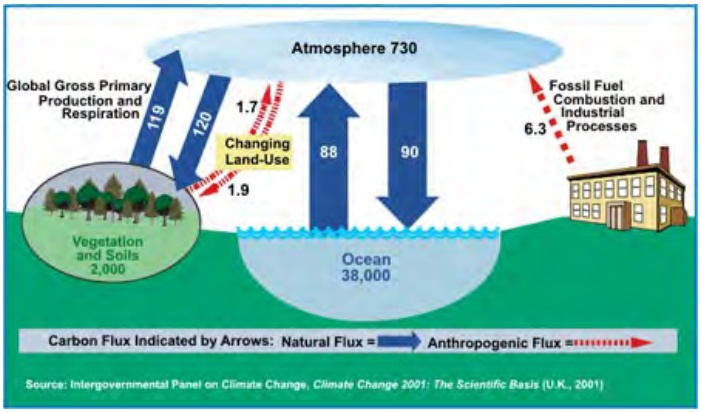
Figure 2.03
Atmospheric windows and absorption bands

In almost all discussions of global warming caused by CO_2, the plots used to demonstrate this effect show temperature versus time (years). This results in two kinds of plots: 1) one that shows variations (ups and downs) over long periods; and 2) one that shows a steady increase in recent years. This is pertinent information but a third plot would be equally informative: one that simply shows how temperature increases with concentration (the amount of CO_2 in a fixed volume). This is a straightforward type of measurement in Physics. Yet the writer had great difficulty in finding such a plot after searching through hundreds of images on Google. The one shown below consists of measurements made on a mixture of Nitrogen and Oxygen (in atmospheric proportions) containing also variable amounts of CO_2-gas in a chamber that was subjected to thermal radiation; the three plots refer to different rates of allowed temperature increases:



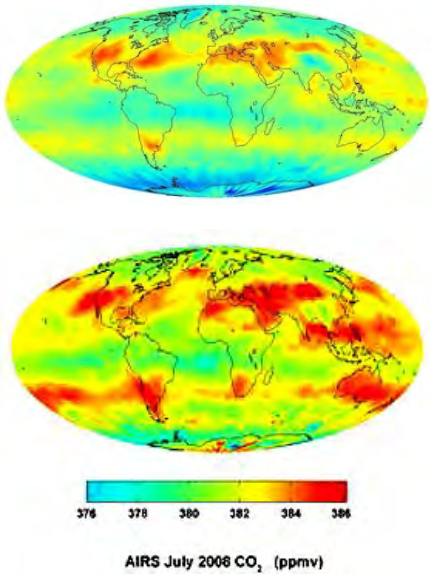
The irrefutable message from this plot: experimentally, it is clearly demonstrated that the addition of increasing amounts of CO_2 to an otherwise constant atmosphere causes a systematic rise in gas temperature. The ranges shown in the above plot are pertinent to levels that might be attained in the future if carbon dioxide emissions are not curbed.

This is a good moment in our discussion to place the CO_2 variations into the context of their distribution in the atmosphere, land vegetation, and the oceans *in terms of* the total amounts of carbon itself (but excluding the carbon found in limestone and dolomite rocks on the continents); a variant of this is shown on page 16-4. Note that the bulk of the carbon is found in the oceans, mainly in the form of carbonate-based shells of plants and animals, but also as dissolved carbonate and bicarbonate ions (see below):

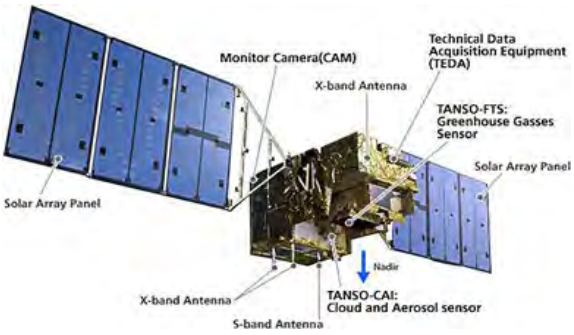


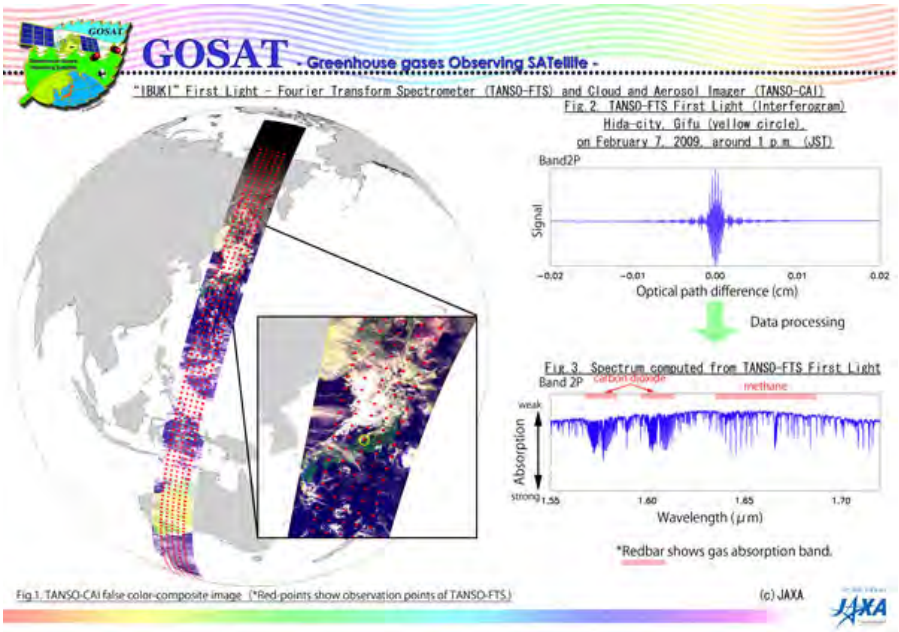
Satellite Observations of Carbon Dioxide Distribution

Now, back to other aspects of information relevant to global warming. Later in this Section we will familiarize you with dedicated satellites that gather data on atmospheric gases and other contributors to global warming. This bears on a main thesis in this Tutorial, that satellites are proving to be a powerful, almost unique, tool for monitoring the whole Earth at the global scale. For now, we will mention two such observers. The first satellite is Aqua, discussed on page 16-11. Here are the results obtained by the AIRS sensor on the Aqua satellite which led to a global map of carbon dioxide at a height of 8 kilometers (5 miles) above the Earth's surface; the upper image applies to July of 2003 whereas the lower is five years later, showing the spread of CO_2 in the mid-troposphere:

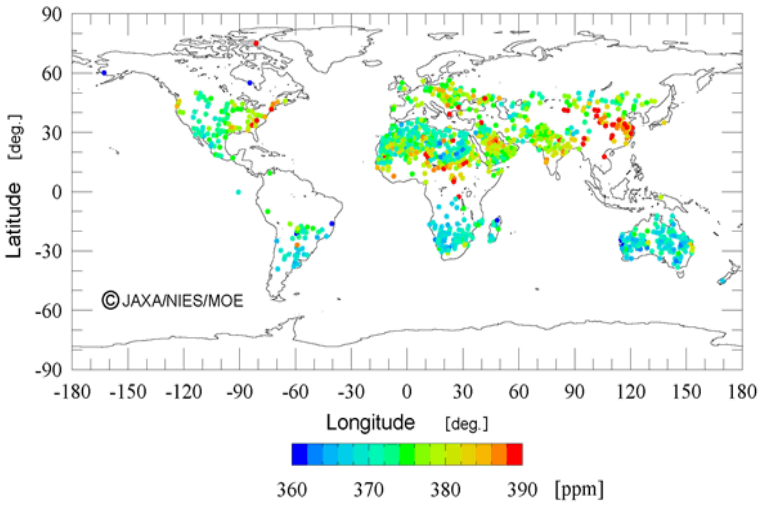


NASA built a sophisticated satellite, the Orbiting Carbon Observer, which would have provided valuable data on CO_2 distribution in the lower atmosphere. Unhappily, the launch on February 24, 2009 failed when during the ascent the fairing that covered the satellite did not jettison properly. But by good fortune the Japanese space program JAXA had just shortly before successfully launched GOSAT (also called Ibuki) which is designed to make similar CO_2 measurements. Here is a drawing of the satellite and below it the first data set released (the satellite is still undergoing adjustments):

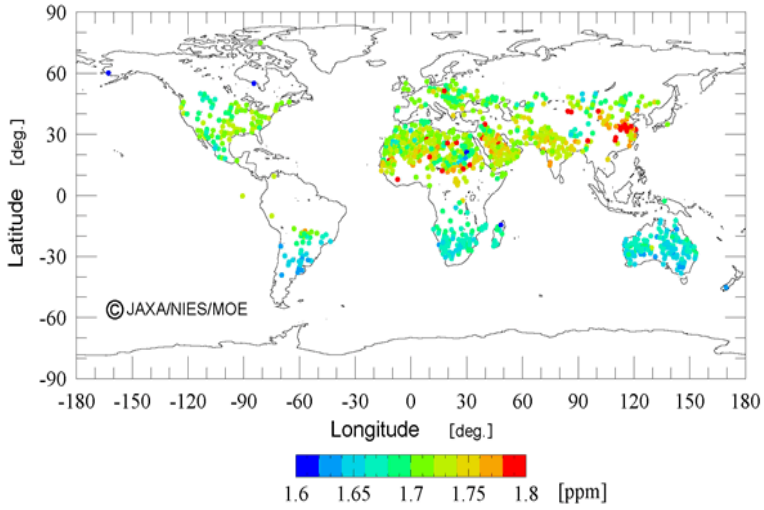




In June 2009, the first maps of CO_2 and CH_4 distributions worldwide were released, as shown here (more comprehensive plots will follow):



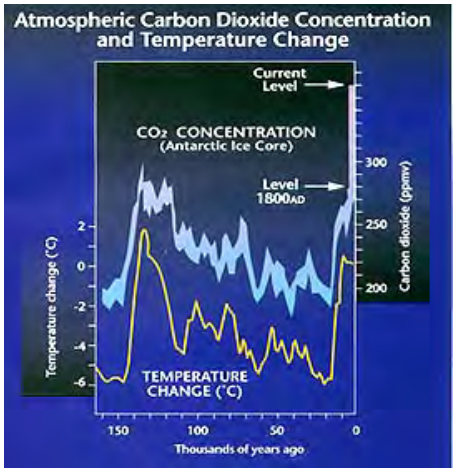
XCO_2 (Column averaged dry air mole fraction)
(uncalibrated data)



XCH_4 (Column averaged dry air mole fraction)
(uncalibrated data)

Historical Records of Carbon Dioxide in Glacial Ice

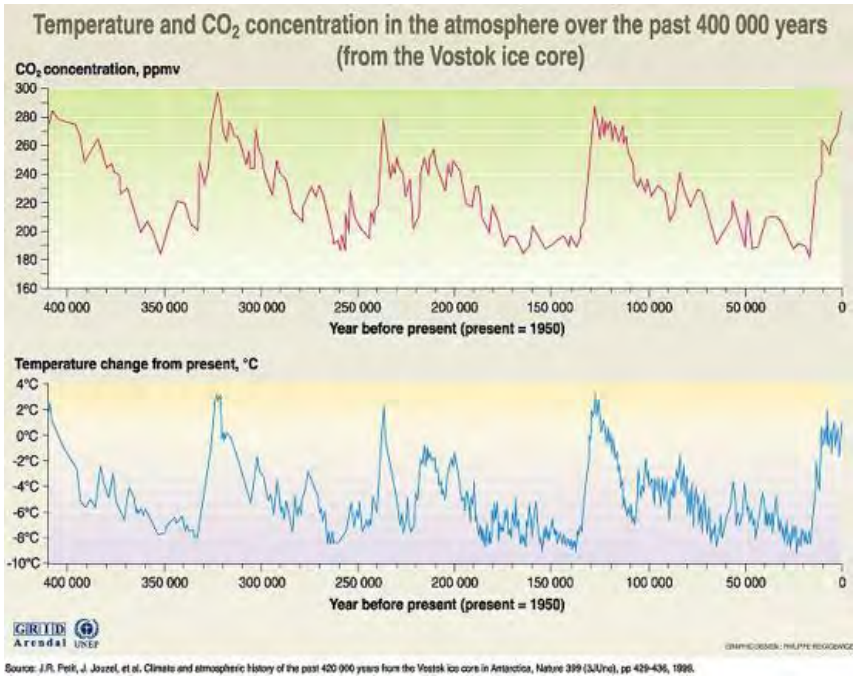
These new data sets will be invaluable. Scientists need to establish a long-term data base for global warming, which includes temperatures from previous centuries. One way to estimate past temperatures is from CO_2 measurements of air trapped in glacial ice that has been accumulating for thousands of years. The plot below shows the general trend in rising CO_2 during the last 160,000 years (includes recent ice ages), with estimates of temperatures (derived from a CO_2 model) shown for comparison.



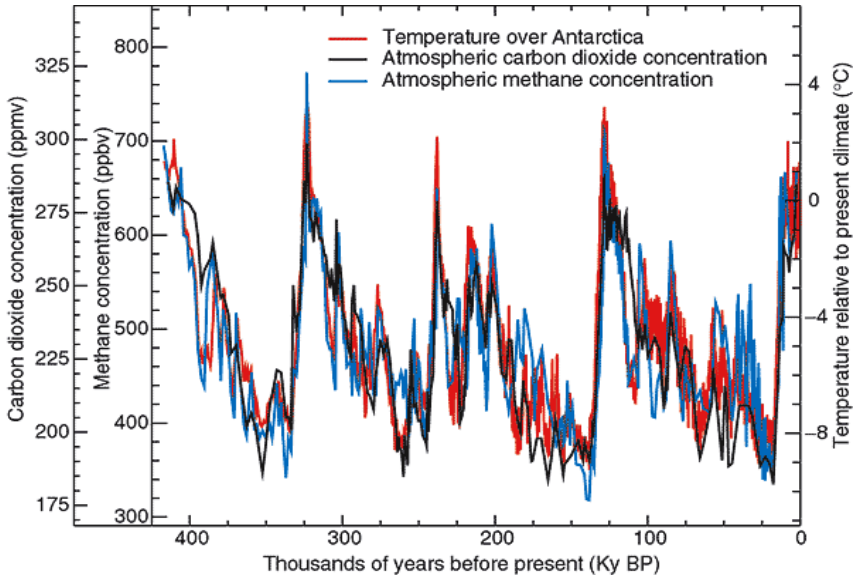
This may be a very important graph bearing on the arguments for and against global warming. The graph shows a temperature maximum about 130,000 years ago, with a steady decline thereafter until a rapid rise beginning about 15,000 years ago. This rise resembles the same

increase leading to the 130,000 peak. The question that must be asked: Is this just a repeat in the normal cycle associated with Pleistocene glaciation? The follow-up question: Is there any strong evidence that the more rapid temperature rise of the last 50 years (or, going back to the Industrial Revolution) is abnormal? The closest to the truth may be that there is now a temperature spike added to a trend that has natural causes related to glacial processes.

Even more impressive is the correlation between CO_2 content in cored ice and derived temperatures during the last 400,000 years, as shown in these plots, both using data from the Vostok ice core (Antarctic):

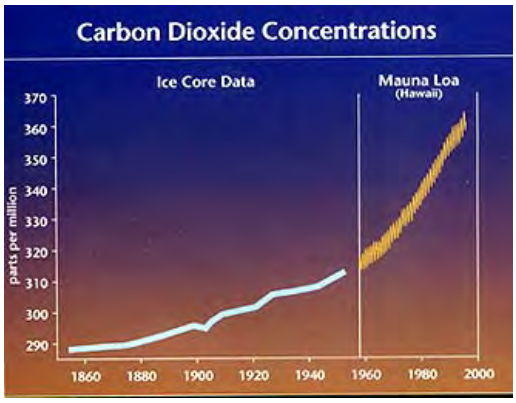


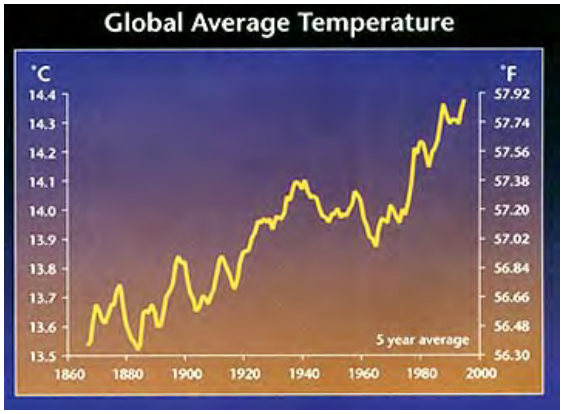
Source: J.R. Petit, J. Jouzel, et al. Climate and atmospheric history of the past 420 000 years from the Vostok ice core in Antarctica, Nature 399 (3/Una), pp 429-436, 1999.



Note: These high and low range values of temperature correlate well with the periods of continental glaciation in the northern hemisphere and the warm interglacial intervals over this extended time period.

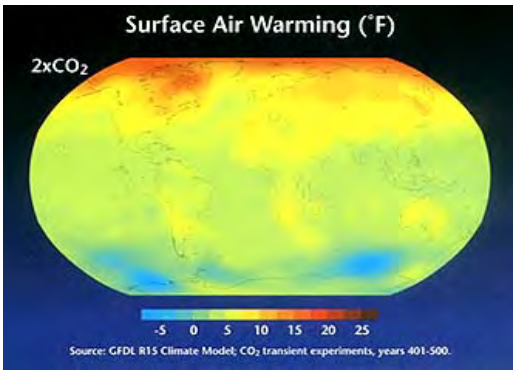
The next graphs provide more details on the increasing levels of CO_2 during the last 140 years. The top shows carbon dioxide concentration increases, based on ice core measurement until 1960 and Mauna Loa Observatory measurements thereafter. Below it is the measured temperature changes averaged for the entire world; the trend upwards, amounting to about 1.5° F, shows some irregularities (not smoothly cyclic) which result from other climatic factors. The Mauna Loa level in 2006 was 384 ppm.

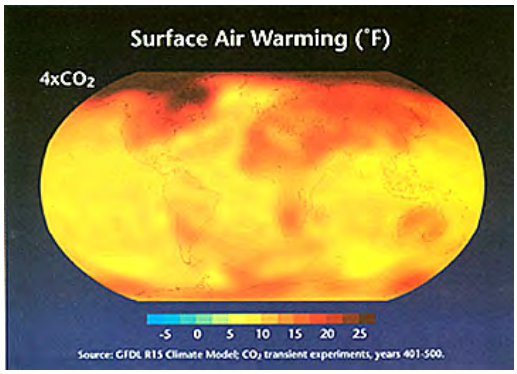




Global and Polar Distribution of Temperature Effects

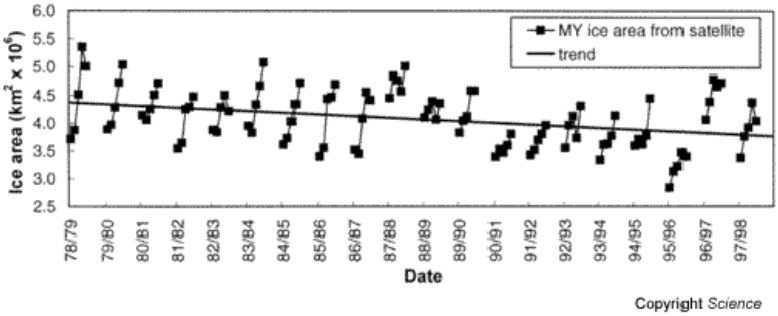
Armed with the data that relate CO_2 increase to temperature rise, models allow prediction of the global distribution of changing temperatures if the total amount of carbon dioxide in the air were to double or quadruple from today's values (arbitrarily set at 275° F). The 4x case could be truly catastrophic, leading to worldwide deserts unless the added heat in the atmosphere also notably increase vegetation growth (to tropical forms). Thus:





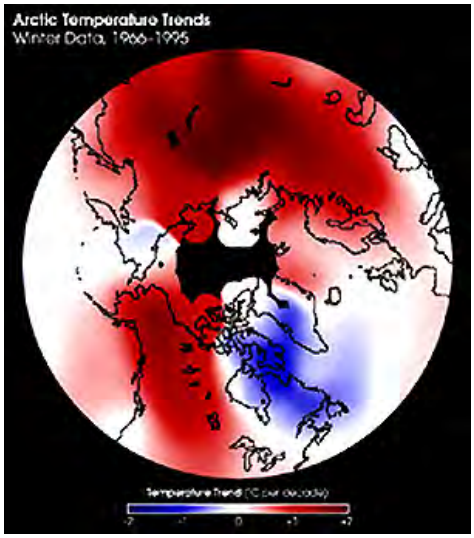
One obvious consequence of the significant rise in CO_2 in the northern polar latitudes would be melting of Arctic Ocean and Greenland Ice Cap ices, releasing huge quantities of stored water that would have an extremely serious impact on global sea levels. A rise to 550 ppm CO_2 by the end of year 2100 would bring about rises from 8 to 37 inches, enough to greatly modify present beach and shoreline configurations and even submerge such cities and New Orleans and Venice (which is also presently sinking because of groundwater withdrawal).

Evidence has been building, through satellite and ground measurements, that the Arctic region of the northern hemisphere has been steadily warming. The main consequence has been the shrinking of the area capped by ice in the Arctic, as shown in this plot based on ground data:

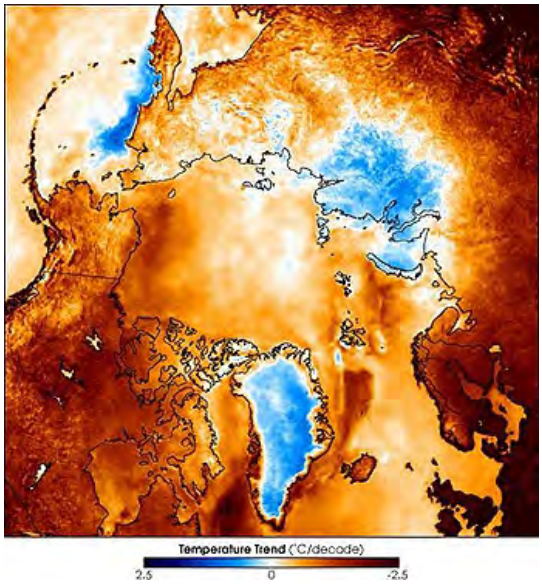


Copyright *Science*

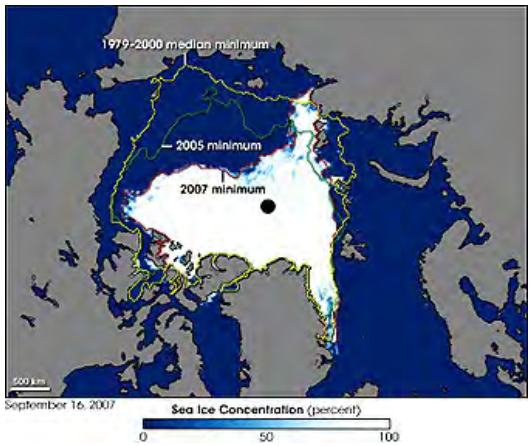
The general change pattern of Arctic temperatures is summarized in this map of the upper northern hemisphere using data from the last 30 years. The Arctic ice shelf has responded both by thinning and loss of ice at the margins. Further south, the western U.S. and Siberia are both in a warming trend. Note that the only region that has cooled during this time is in eastern Canada-Greenland.



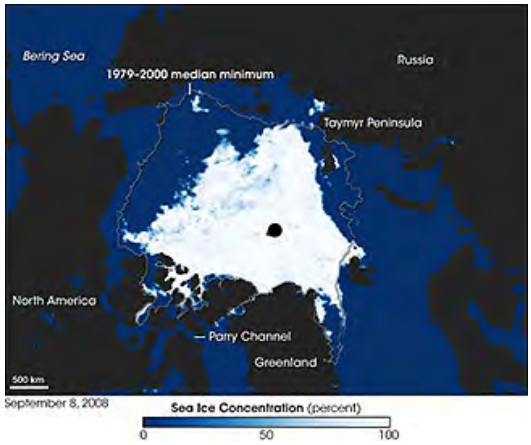
A somewhat more precise map of temperature changes in the Arctic covering the two decades between 1972 and 1992 has been constructed from meteorological satellite data as shown here:



The net effect has been an overall shrinkage of the outer extent of the main sea ice cover within the Arctic Circle:

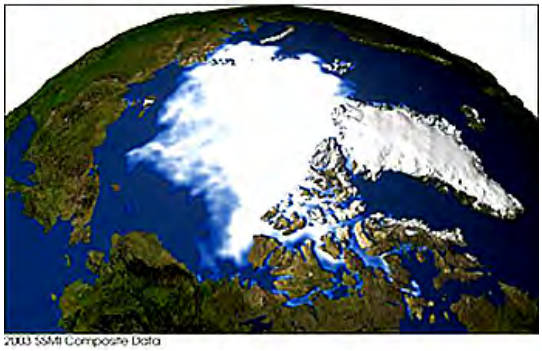


Owing to greater precipitation, the sea ice around the North Pole increased somewhat in 2008. But surrounding waters were more open. This meant that the famed Northwest Passage shipping lanes were even better suited to navigation by cargo ships from Europe to the Orient for the summer months, obviating the need to transit through the Panama Canal.

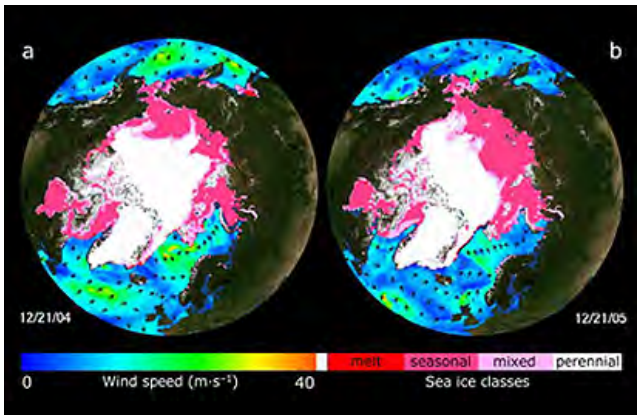


This ice fluctuation can be displayed visually using data from the SSM/I (Special Sensor Microwave Imager) instrument (on the U.S. Air Force Defense Meteorological Satellite series), taken in 1979 and again in 2003. The Arctic Ocean appears to be opening up off eastern Siberia and northern Alaska.

1979 SSM/I Composite Data

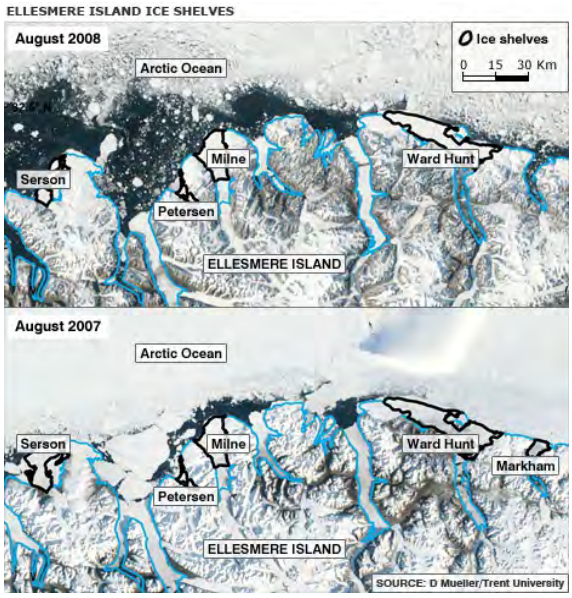

2003 SSM/I Composite Data

More recent observations have increased the alarm. NASA's Quikscat satellite has produced a map of the Arctic's perennial sea ice (ice that remains through the summer) for the years 2004 and 2005:

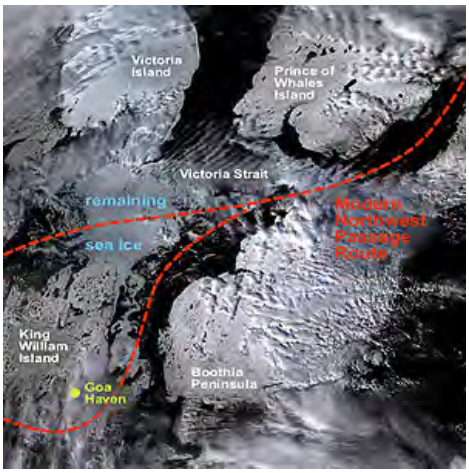


In 2007, the reduction of the most stable ice - perennial, in white - amounted to about 14% in one year - the largest yet observed. Overall, in the last 30 years ice has thinned from an average of 3.5 meters to less than 2 m. The areal extent is likewise shrinking, as this graph shows:

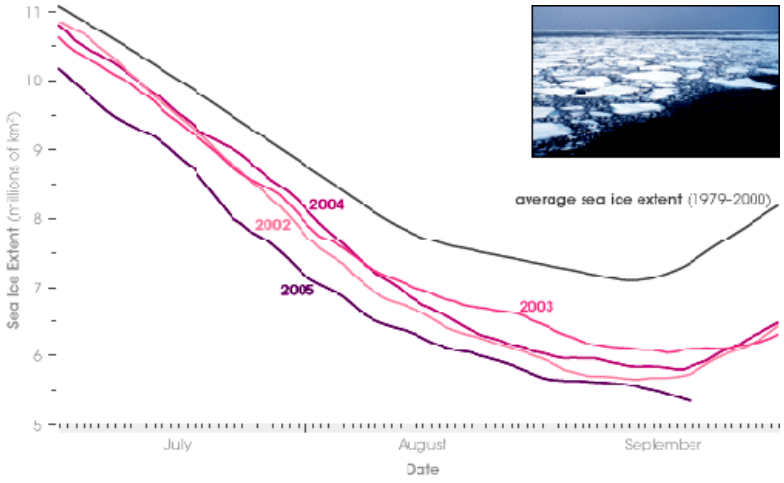
The ice is breaking away from Arctic islands. Here is a one year change next to Ellesmere Island in northern Canada:



Early explorers had hoped for at least a short season when the Arctic ice cap had melted enough to provide open water through which ships could pass - the famed Northwest Passage. In recent years this is now happening on an annual basis, as seen in this image in northern Canada:

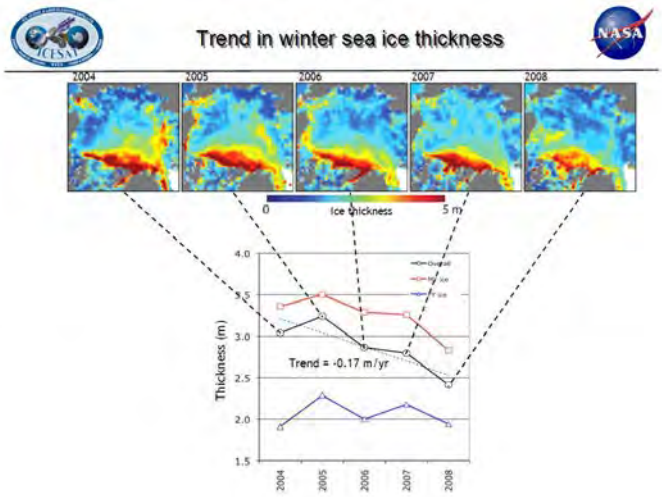


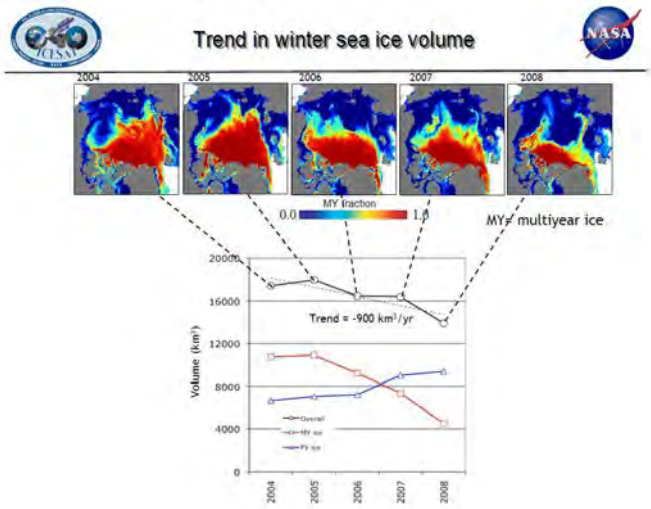
This progressive diminishing of ice cover has been increasing steadily in the last few years, with 2005 now showing the least amount of frozen ice. Data acquired by satellite and ground truth lead to this plot:



Evidence is growing that ice in the Arctic Ocean is accelerating its loss. Near Ellesmere Island in northern Canada, the Ayles Ice Shelf began to break up in 2006. Here are a series of space observations that show the breakaway of a 64 square kilometer (41 sq miles) section of the Shelf over a period of just weeks:

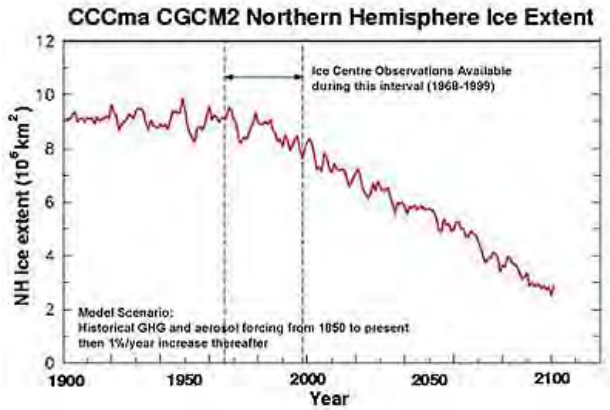
NASA's IceSat has been monitoring the changes in Arctic sea ice thickness and volume since 2004. Here are results through 2008:





The decrease in sea ice has a potentially profound effect on the climate in the northern hemisphere. Ice reflects solar irradiation, thus keeping temperatures colder. As its areal coverage decreases, the reflectance diminishes and the open waters produced absorb more solar heat, causing the regional temperatures to rise. This in turn accelerates the general melting. If the Arctic ocean becomes notably warming, the lands to its south will experience warmer temperatures and the climates of North America, Europe, and Asia will change significantly. Some of the effects of this could be beneficial but others would be deletorius. If this warming trend is just part of the general interglacial cycle, then little can be done by humans to moderate or even reverse it. But if greenhouse gases are the culprit, then it becomes urgent to place these under much tighter control.

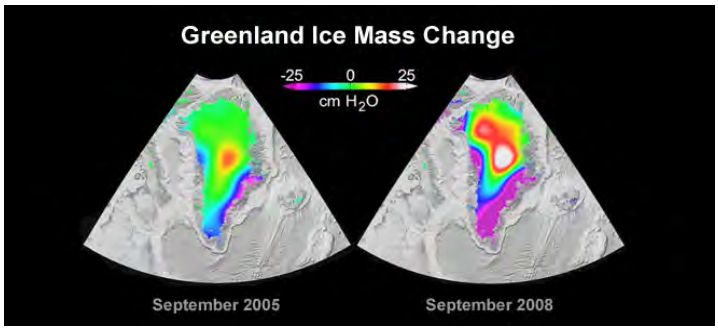
Experts studying the ice shrinkage in the Northern Hemisphere have, using projected temperature rises, produced a plot of area decrease during the full 20th Century and a prediction for loss in the 21st Century. The loss trend started around 1970:



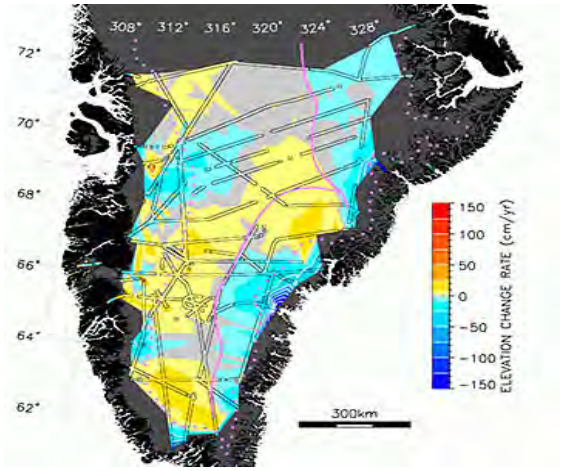
Changes in Continental Glacial Ice as Indicators of Warming

The subcontinental Greenland land mass, mostly covered with ice, is also serving as a proving grounds for detection of widespread melting. Despite the interior cooling of Greenland (suggesting there should be enough snow to increase the ice cap's thickness), the edge of this ice cap is currently shrinking as confirmed by several separate aerial/ground/satellite studies in

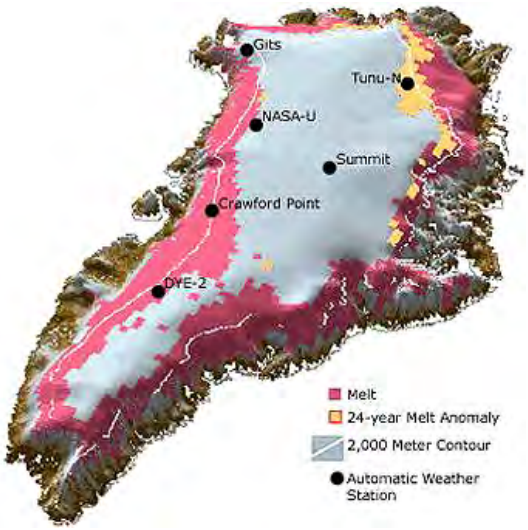
which remote sensing played a part. Most recent is the map produced by the GRACE satellite that uses gravity data to calculate the mass of ice that has melted, or increased, on Greenland:



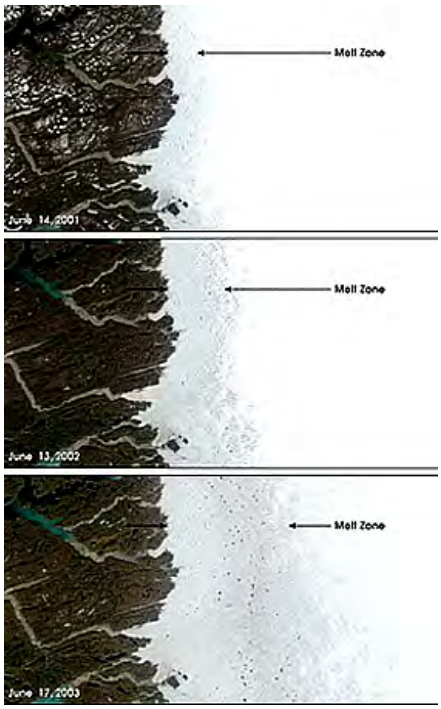
A NASA team made survey passes across Greenland by air using a laser altimeter to measure elevations. A second group used the Global Positioning System (GPS) (page 11-6) to track movements of ice. When both results were analyzed, the map shown below was constructed. It shows increase in ice in the central part of this huge island (possibly related to the recent winter cooling) but significant shrinking near the edges (increases in summer temperature). The net loss of ice is estimated to be 51 cubic kilometers a year, enough to fill a lake 30 miles long, 30 miles wide, and 70 feet deep (spread globally to the oceans this would cause a rise of 0.005 inches a year; this contributes 7% of the total rise now being measured).



This map shows similar results from a different perspective:



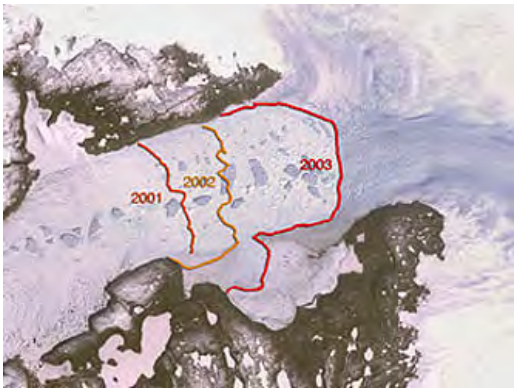
Scientists visiting Greenland in the last few years have remarked about significant increases in the onset of melting along the edge of the western fringes. This is visible in MODIS images taken from 2001 into 2003, with the zone of active melting showing up as a darker gray (presence of water):



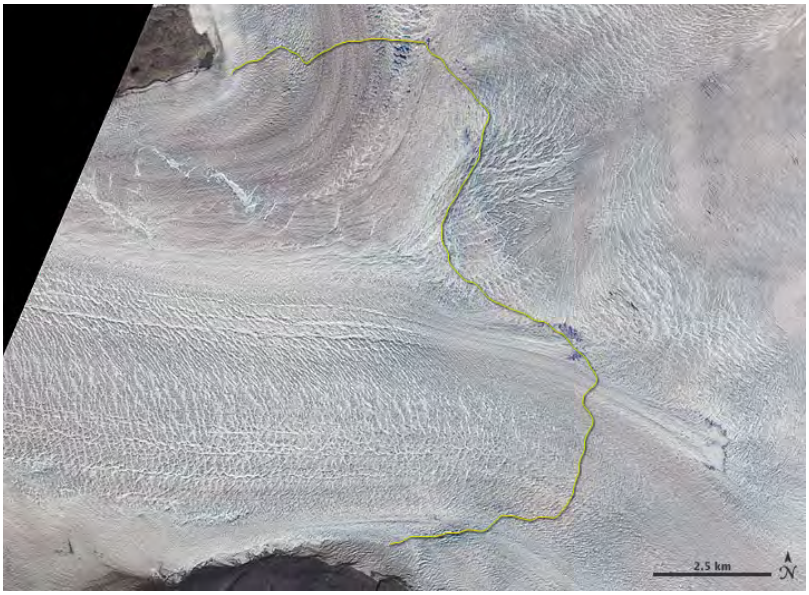
Individual glaciers along the edges of the Greenland ice cap, especially on its western side, have been undergoing retreat and flow rate increases (caused by increased water within the glacier that lubricates its movement) as the ice melts. Three big glaciers on both sides of Greenland are shown in this map:

This Terra ASTER image illustrates the loss through melting at the Jakobshavn Glacier:

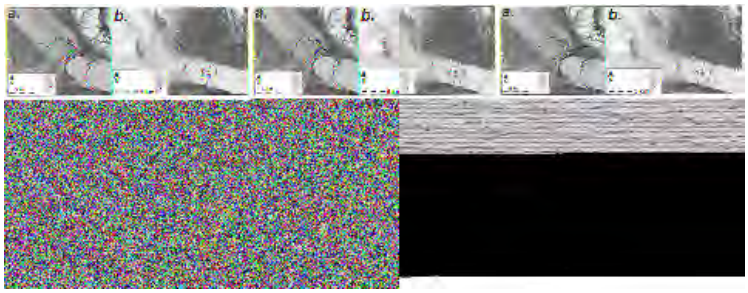
This aerial photo shows the positions of the Jakobshavn glacier's toe or frontal boundary in the 21st Century. This front is now rapidly moving at a rate of 15 m/yr (49 ft/yr), as it calves off ice that enters the ocean along Greenland's western shore.



This next pair of images, made by ALI on EO-1, show in detail changes within the Jacobshavn glacier that have occurred between 2001 and 2010. The degradation of the glacier is obvious.



The two glaciers flowing into the sea along the southeast coast (see above map), Kangerdugssquaq and Helheim, from the latest ground measurements are now flowing (even as their fronts retreat) at 40 m (131 ft) and 25 m (82 ft) per year respectively for the period 2004-05 - the fastest movement rates ever observed for this pair:

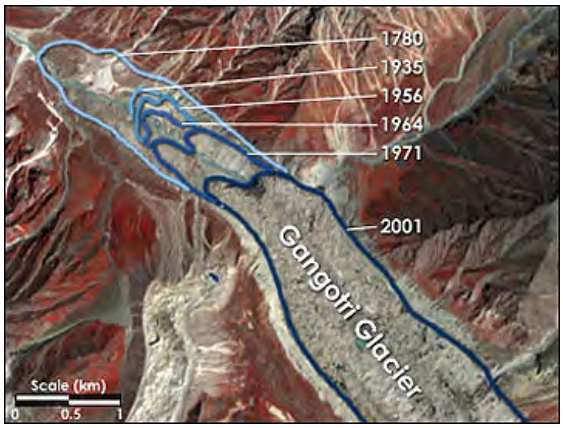


Recent reports indicate that the amount of ice released from the Greenland sheet to the sea has nearly doubled in the last 10 years - the highest rate yet observed. This may be another sign that overall Arctic ice is wasting away at an accelerated pace. In the 21st Century, if this increase continues, worldwide sealevel can rise - from arctic melting alone - by about a meter (3 feet). This space image shows the coast around Helheim with a large raft of icebergs beyond the shoreline:

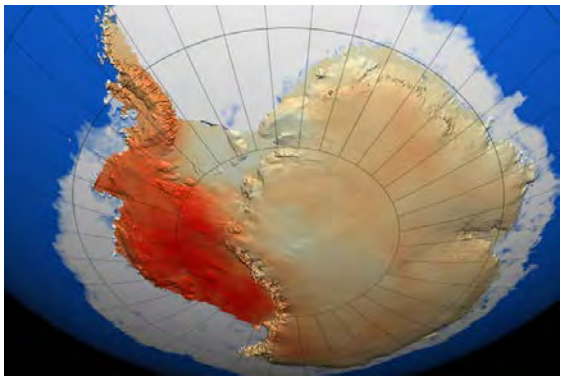
The concern is one of "volume", not just "area". While the areal coverage in the Arctic has decreased significantly - and looks dramatic - the volume loss in Greenland exceeds that of the sea ice. This is the kind of loss, if it accelerates, that will have the most impact on sealevel rise.

Ice melting is worldwide. Thus, melting is being observed in the Southern Hemisphere, mainly in South America and the Antarctic. Here is a ground photo pair showing the Qon Kalis glacier in the Peruvian Andes:

Glaciers in Asia also show notable retreats, as shown here for the Ganglotri glacier:

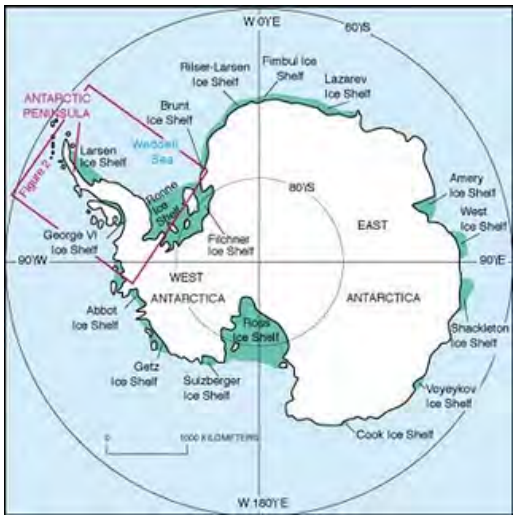


For a time, it was believed that the south polar region (the Antarctic) was not undergoing the degree of warming being observed over much of the rest of the world. But, a study of the last 50 years of satellite and ground station data has disclosed a continent-wide temperature rise of about 0.5 degrees centigrade. This next map shows the distribution of temperature rise in the last decade, with the reds denoting an increase of about 0.25° C grading into the browns that mark an increase of 0.10° C.

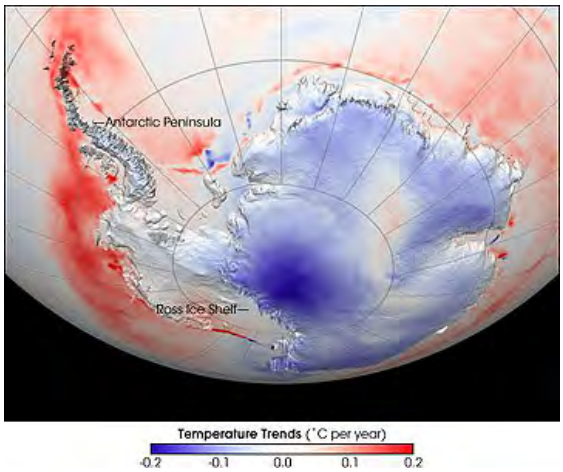


Ice builds up each year in shelves around the Antarctic. This diminishes in the Antarctic summers (Dec.-Feb.). Here is a composite view of Antarctica showing swirls of ice in the seas during an Antarctic summer.

There are many named ice shelves around Antarctica. This map shows the principal ones:

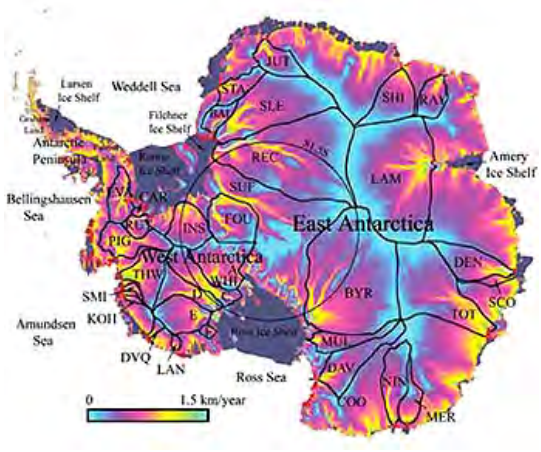


On page 14-14 evidence of major changes - mostly as thinning of ice along the fringes - in Antarctica was discussed. This is in response to notable rises in temperature in sea water along its coastal shelves; in the Antarctic Peninsula, a cumulative rise between 2 to 3°C (3.6 to 5.4°F) has occurred in recent years. This map, made from NOAA AVHRR data, shows that the exterior parts of the Antarctic and surrounding waters has warmed in the last 22 years, yet the interior has cooled. Perhaps the increased evaporation in the seas, where ice calving has exposed more open ocean, has produced more inland precipitation, accounting for the cooling.



Ice along the Antarctic Peninsula (west-northwest of the South Pole) is largely melting and thus retreating, although some measurements show ice advance

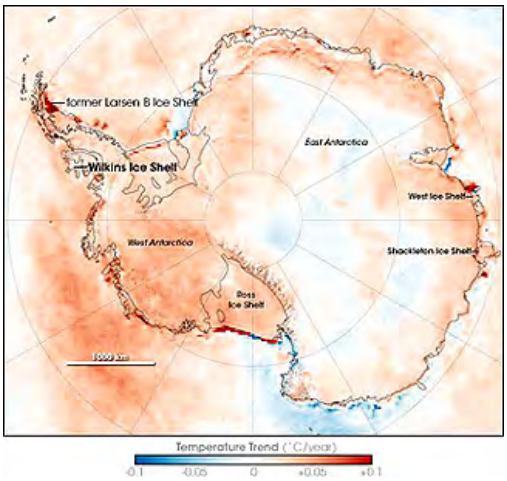
In the last few years parts of the fringing Antarctic ice shelves have broken loose as massive ice bergs floating northward. Use of laser altimetry and interferometric measurements of data obtained by InSAR (Interferometric SAR) are showing up as annual velocities of ice movement both within the continent and along its edges, as seen in this composite image:



The Antarctic ice shelves appear in gray in this map. Letters indicate 33 named glaciers. As the ice sheets thin and break off, the glaciers respond by increased (accelerating) outward flow.

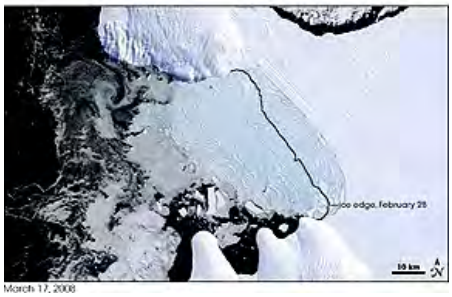
The Ross ice shelf was one of the first in the western Antarctic to show dramatic changes. It has greatly diminished by 2006. Here are a series of MISR (Terra) images of parts of the Ross shelf as these broke off and entered the sea as ice(berg) sheets:

In February of 2008, most of the Wilkins ice shelf disintegrated, as ice cover the size of Delaware broke loose. The changes were imaged by the Terra satellite.



ebruary 28, 2008

February 29, 2008

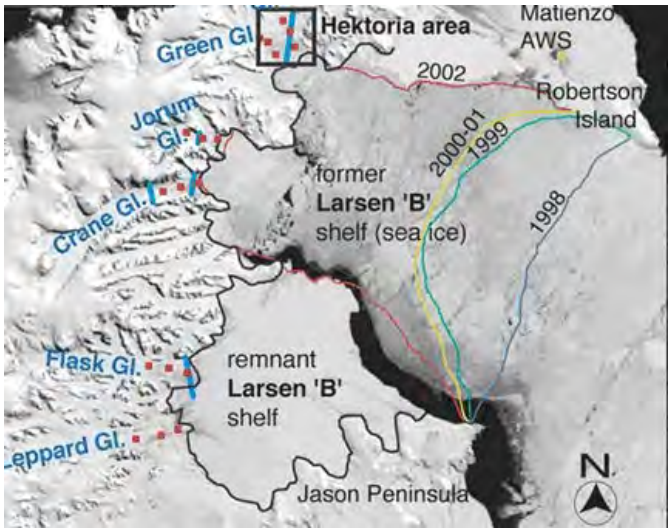


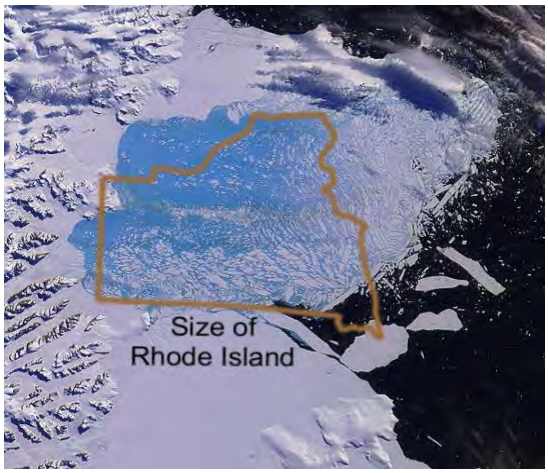
March 17, 2008

Formosasat, operated by the Taiwanese, obtained this detailed view of the disrupted Wilkins ice shelf:

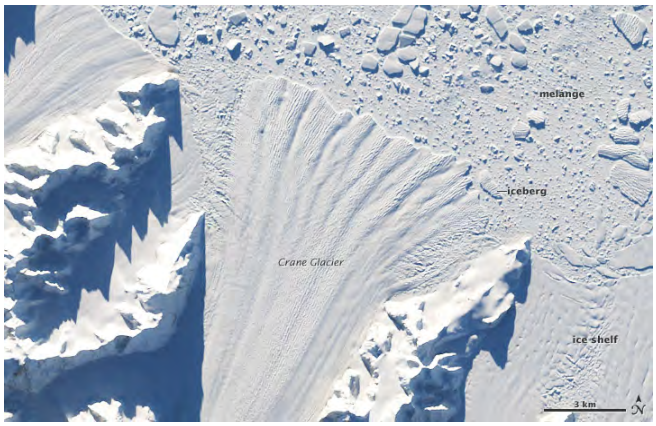
Here is an aerial photo showing broken ice and separated slabs within the now detached Wilkins ice shelf:

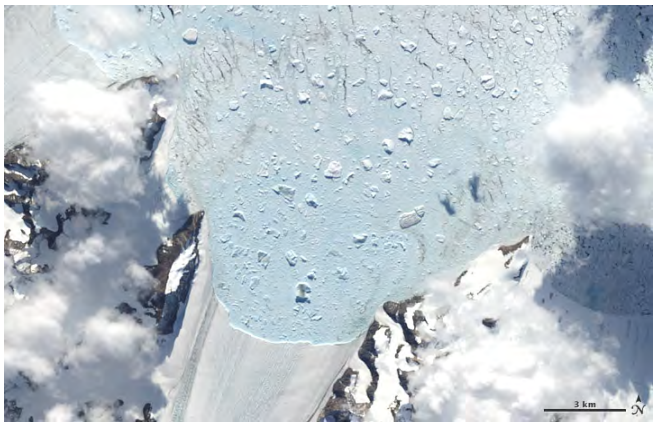
The Larsen B ice shelf, located in the Weddell Sea off the Antarctic Peninsula, experienced almost disintegration in a few years just prior to 2002, amounting to a loss of continuity over an area comparable to the state of Rhode Island, as shown in these two satellite images:





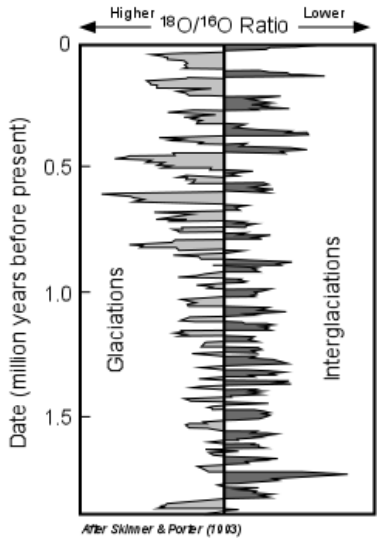
Some of the changes along the Antarctic coastline can be dramatically rapid. Owing primarily to a sequence of warmer than usual Antarctic summers, the Crane glacier, feeding into the Larsen B ice shelf, underwent a collapse that moved its front back more than 6 kilometers from April of 2002 to March of 2003. This was due both to recession of the glacier through wasting and rapid breakoff of the ice as icebergs.





The West Antarctic ice shelves (primarily the Ross Shelf) are losing ~65 cubic kilometers per year, enough to raise ocean levels by 0.16 mm/yr; worldwide sealevel is rising by 1.8 mm/yr (0.7 inches/yr), primarily from Arctic, Antarctic, and Greenland sources. Scientists have calculated that if all of the ice in Greenland and the Antarctic were to completely melt through significant warming the worldwide sealevel would rise 70 meters (230 feet), which would be catastrophic to the large fraction of global populations living near the coasts.

Like the other evidence, a shrinking ice cover in each hemisphere does not confirm a direct response to global warming. Critics of the global warming hypothesis have challenged the conclusion reached from the now verified evidence of the linkage between warming and ice melting. Their argument is simple: Studies of global ice distribution during the last few million years have revealed a cyclic trend of advances and retreats. Debate still continues as to the cause(s) of these cycles. The periodicity is revealed by measuring the ratio of O^{18}/O^{16} in ice cores retrieved from deep drilling into ice caps. O^{18} will increase during times of warming. This diagram displays the trends of warming and cooling determined by this method (individual core segments are dated by C^{14} and other radioactive isotopes):



Thus, natural cycles of freezing/thawing as continental glaciation waxes and wanes may indeed explain the current lines of evidence disclosed by temperature and ice retreat indicators. The crux of the argument between global warming advocates and those conservative thinkers who still believe the recent observations are not alarming and to be expected is this: **The Earth may just be undergoing its latest natural warming in the cyclic sequence**. But taken with other signs of temperature rises, and their effects, one can certainly postulate that global warming is a credible factor in these changes. Some argue that both cyclic warming and human-induced global warming are occurring concurrently. The present situation then may be made worse by the added effects of CO_2 and other gases that are driving the atmosphere into a warmer than "normal" temperature rise.

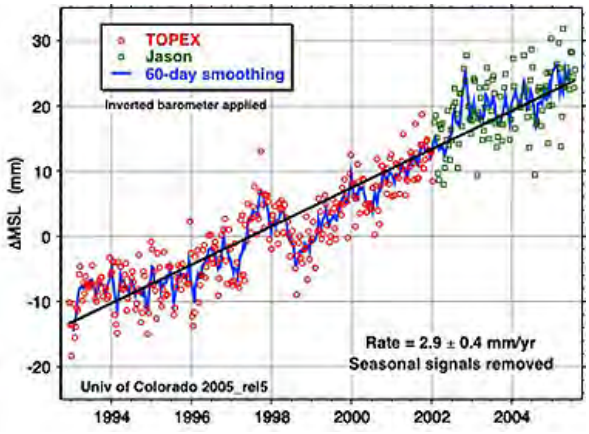
Sealevel Rise owing to Warming-induced Melting

As stated before, ice melting will lead to sealevel rises, both regionally and worldwide. A good overview of this subject is found at Wikpedia's review of sealevel change.

As an illustrative example of how a specific region would be affected, this predictive map of southern Florida shows the extent of inundation if sealevel were to rise just 5 meters (about 16 ft):

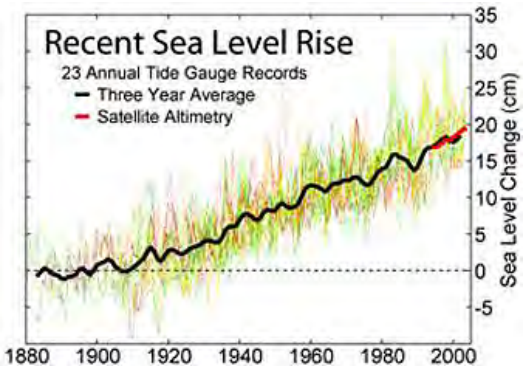
But is sealevel actually rising? Data from the last decade, using satellite imagery, points to a

strong affirmation of this change:



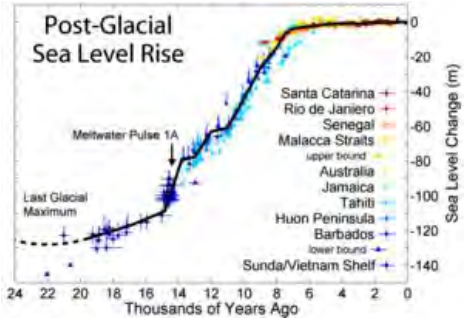
Some other lines of evidence are shown in these figures, taken from the aforementioned Wipedia website:

The first shows sealevel rise trends for the last 125 years as measured at 23 marine gauges:

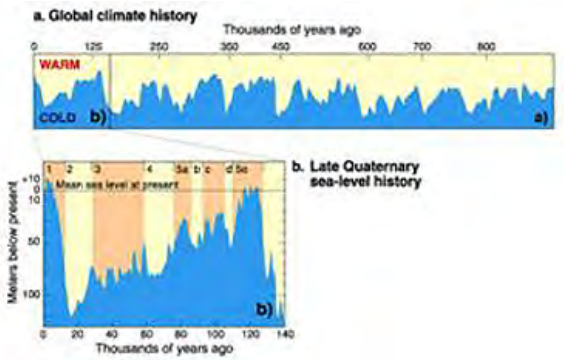


But this question can be raised: Have comparable rises been observed in the past?

From observations made at various islands and continental seaports, sealevel has steadily risen since the end of the last major continental glacial advance over the northern hemisphere; note that sealevel rise has approximately leveled off in the last few thousand years.

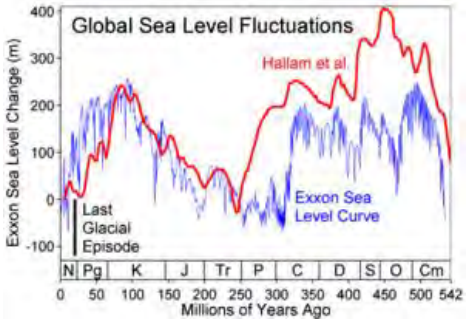


The trends in temperature and sealevel fluctuations over the last 900000 years is shown in this next diagram, which has an enlargement showing the changes during the last 140000 years. The impression is that there has been a cyclic pattern that makes it difficult to say that the current observations are unusual:



Over the last 500 million years or so, sealevel has undergone some broad patterns of change (as

determined in a study supported by Exxon and in a separate study by Halam et al. that has used worldwide data and attempts to compensate for shifts of continents driven by plate tectonics). Higher general levels have occurred at the beginning of the Paleozoic and the end of the Mesozoic, with a minimum during the Pangaea breakup at the end of the Paleozoic. Note that the rise/fall trend during the glacial activity in the Neogene (N) is much less in magnitude than during earlier maxima.



The main conclusion one reaches after looking at the sealevel change data for a long interval of geologic time (assuming the patterns shown are generally accurate) is that short-term changes are highly variable but over millions of years trends of rise or fall proceed more gradually. The current rise may indeed be due to man's activities but alternate explanations (natural fluctuations, etc.) remain competitive. But it is clear that global warming is now taking place, owing probably to several causes, and as a result coastal areas are likely to be inundated (requiring either relocation of cities or emplacement of elaborate dike systems).

Effect of Greenhouse gases on Ocean Acidification

There is another aspect of the relationship between CO_2 and ocean water that was brought to the writer's attention by an article in the July 3, 2010 issue of the *Economist* magazine. Entitled "The Other Carbon Dioxide Problem", it deals with ocean acidification. The gist of the information summarized in the article is recast in the next several paragraphs. A common theme throughout is the influence of feedback mechanisms on marine ecology. The concept of "feedback" is an important one, worthy of a quick synopsis taken directly from this Wikepedia website. Several sentences from that source are extracted here: "Feedback describes the situation when output from (or information about the result of) an event or phenomenon in the past will influence an occurrence or occurrences of the same (i.e. same defined) event/phenomenon (or the continuation/development of the original phenomenon) in the present or future. When an event is part of a chain of cause-and-effect that forms a circuit or loop, then the event is said to "feed back" into itself. Feedback is a mechanism, process or signal that is looped back to control a system within itself. Such a loop is called a feedback loop. In systems containing an input and output, feeding back part of the output so as to increase the input is positive feedback; feeding back part of the output in such a way as to partially oppose the input is negative feedback." An example of positive feedback is the role of water vapor in global warming: increased temperature produce more evaporation, hence more water vapor which, as a major greenhouse gas, itself contributes to further temperature rise. However, the same evaporation process can induce negative feedback if the vapor condenses to form clouds which shield the Earth's surface from heating by solar radiation, thus lowering temperature. Both modes of feedback will be highlighted at the end of the following paragraphs that consider ocean acidification.

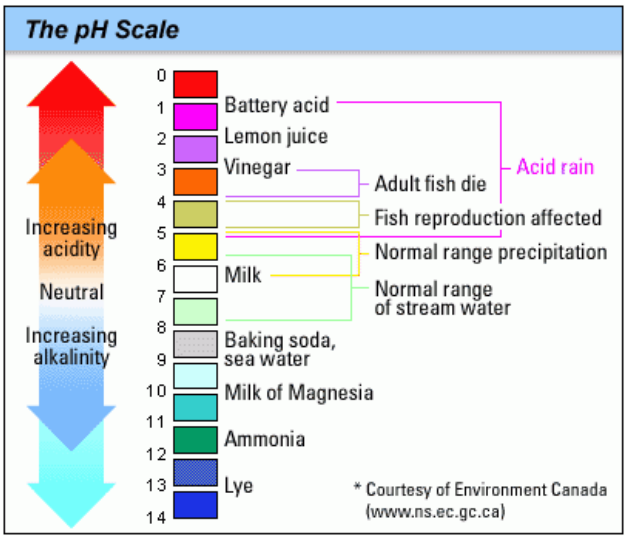
The starting point in examining ocean acidification is simply the scientific fact that as the amount of CO_2 in the atmosphere increases, more of that gas will, in an equilibrium system, react with H_2O to form the weak carbonic acid H_2CO_3. More of that acid makes the ocean water more acidic. Increased acidity will result in the calcium carbonate of coral reefs being more readily attacked and dissolved.

Acidity is a variable that measures the concentration (commonly, in units of moles) of H^+ ions in a aqueous solution. It is usually expressed in terms of pH (p is not a quantitative unit itself, and has no specific name). We will not discuss at length the meaning of pH but will refer you to the Wikipedia entry that treats it in detail. For now, as a "quickie" overview, just consider these points:

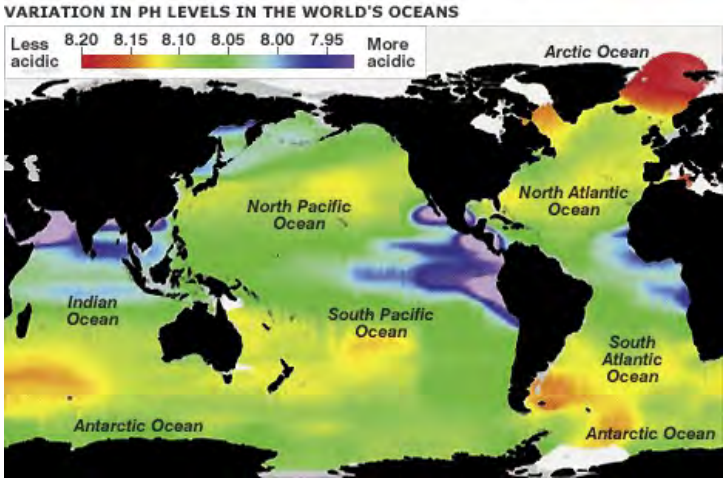
Molecules of water are for the most part stable but a small fraction will dissociate (break apart) in H+ (positive ions of Hydrogen) and OH- (negative ions of the Hydroxyl radical). In the equilibrium state, pure water thus follows this pattern for the fraction dissociated: H_2O <--> H+ + OH-. The concentration of each of these ions in pure water, expressed in terms of pX is defined as pH = - log [H+] and pOH = - log [OH-]. Pure water has a pH of 7, in which there are equal amounts of H+ and OH-. For a more acidic solution, in which H+ exceeds any -OH, take the case where H+ = 0.0001M = 10^{-4}; the log of 10^{-4} = -4; therefore pH = - log [H+] = - log (10^{-4}) = - (-4) = +4 = pH of 4. M is the concentration in moles given in units of molarity. (For any substance, the number of atoms or molecules in a mole is Avogadro's number (6.02 x 10^{23}) of particles. Defined exactly, it is the amount of pure substance containing the same number of chemical units that

there are in exactly 12 g of carbon-12. For each substance, a mole is its atomic weight, molecular weight, or formula weight in grams. The number of moles of a solute in a litre of solution is its *molarity* (M); the number of moles of solute in 1,000 g of solvent is its *molality*.)

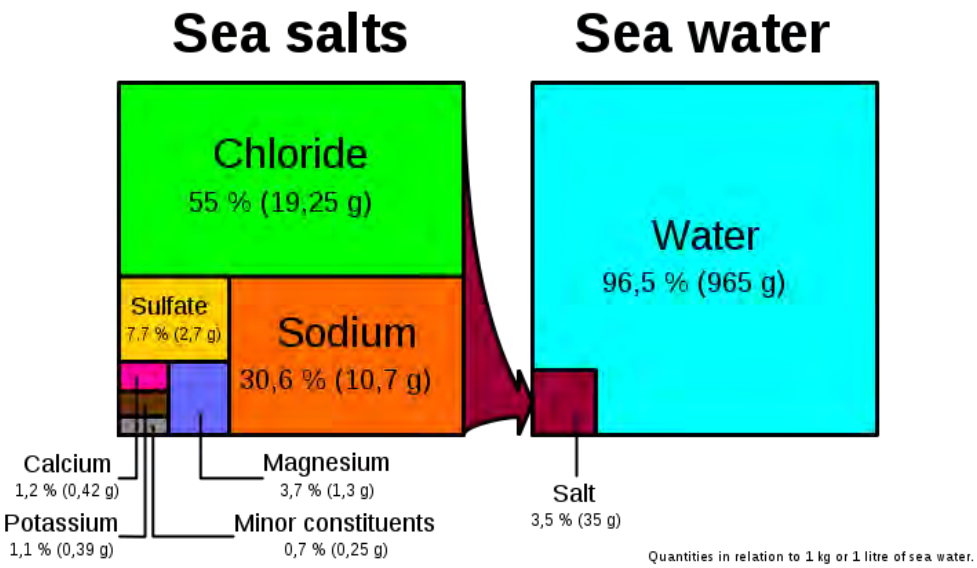
The pH of a variety of substances ranges from just above 0 to 14. This diagram shows some common substances.



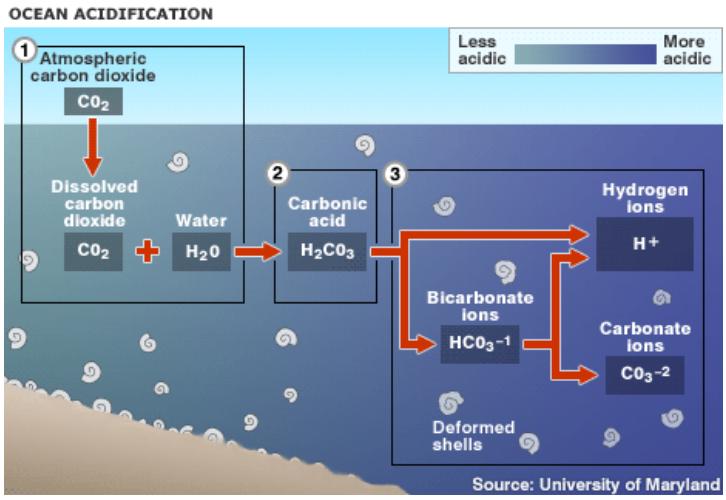
The pH of seawater is on average around 8.1 to 8.2 and ranges between 7.5 and 8.5. The range is controlled mainly by seawater chemistry but is influenced by other variables such as biotic content and atmospheric factors. This is a recent map generalizing oceanic pH variability; note that there is also variability with depth:



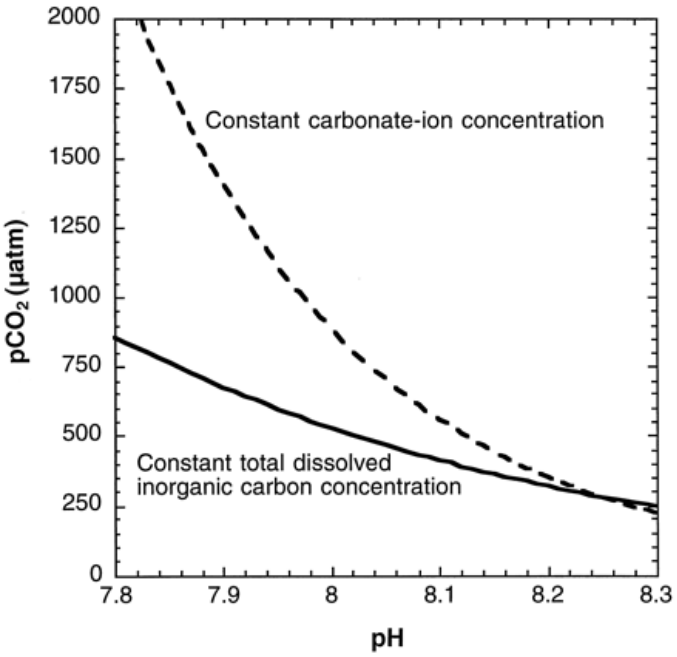
The oceans consist not only of water but of dissolved salts. By far the must abundant are sodium and chloride ions that can precipitate out as halite or common salt. The oceans contain on average 35 parts per thousand (the rest is water) of dissolved salts (or, 3.5%). This diagram summarizes this chemistry:



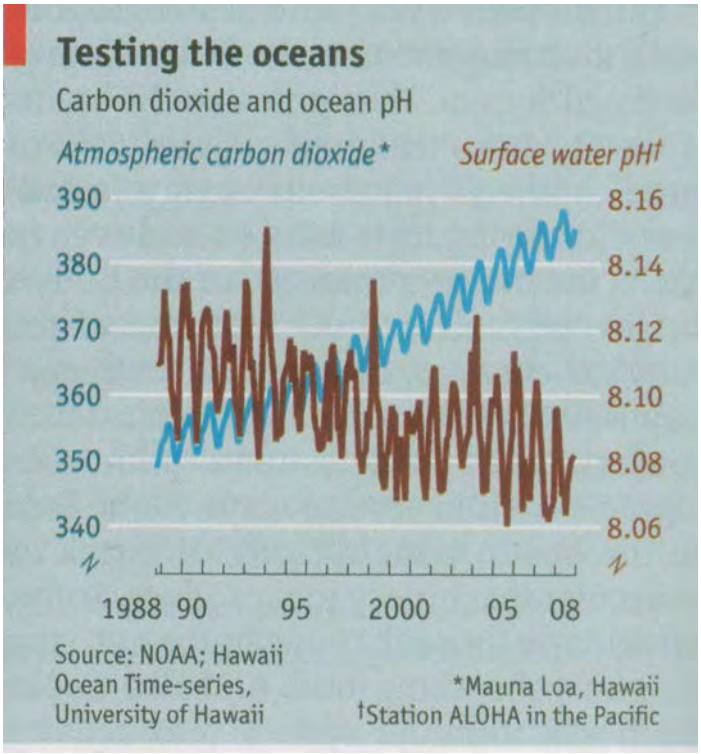
The chemistry of the carbonates of calcium and magnesium is of special significance, as these make up the shells of most marine organisms. Carbonates are sensitive to the amount of CO_2 in the water as this controls both the CO- available to the organisms and the amount of carbonic acid that can form and in turn can dissolve shells, etc. Consult this [Seawater System chemistry] website for a review of this topic. This next set of chemical equations summarizes the chemistry of CO_2 as it interacts with water. The important things to note are 1) an increase in CO_2 leads to increasing carbonic acid, 2) part of the carbonic acid dissociates in H+ and the bicarbonate ion HCO_3^{-1}, of which part dissociates into CO_3^{-1} and another H+. Thus does the increase in carbon dioxide result in adding new H+'s to the oceans.



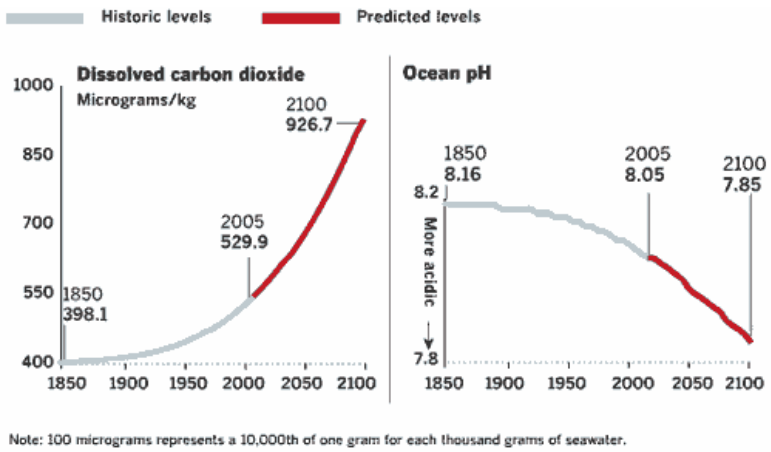
Through experimentation, the lowering of pH (increasing acidity) with progressive rise of CO_2 levels is expressed in this plot:



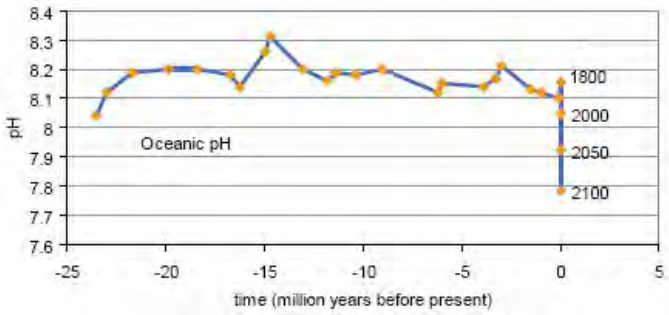
That pH is actually being lowered is demonstrated by this plot (from The Economist article) which shows measurements of CO_2 changes at Mauna Loa in Hawaii and pH off a seaport on the island of Hawaii; the cyclical 'wiggles' are the effect of summer-winter variations.



Using these recent trends as a guide, and taking into account calculations of expected increases in CO_2 emissions from global warming models, this pair of diagrams shows projected effects to the year 2100:

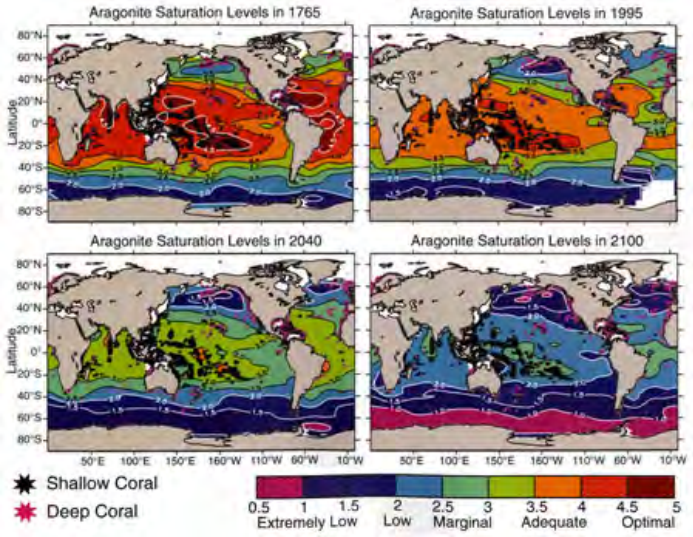


A variant of this (from Turley et al, 2006) shows estimated average ocean pH's from 25 million years ago extrapolated into the future. If this projection is valid, the oceans will become more acidic than any time in that past.



What are the consequences of this change in CO_2 and therefore ocean acidity? First off, corals in reefs are likely to be affected. This is a typical scene showing part of a coral reef:

The polymorph (same composition; different crystal structure) of Calcite (calcium carbonate) is Aragonite. That mineral is found in many marine organisms including corals and free-floating calciferous phytoplankton. This quadruplet of panels shows calculated and projected aragonite levels in the world's oceans for the time span between 1650 and 2100.



This information can be recast in a different way. The six panels below show Aragonite levels for different carbon dioxide amounts in the atmosphere, covering both past, present (350 ppm), and future values.

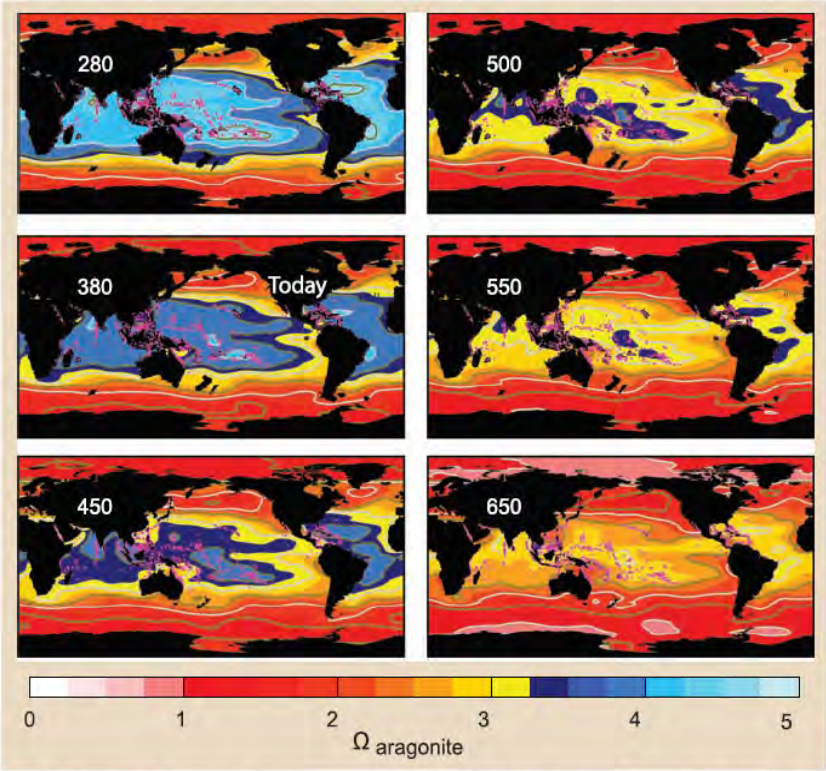


Fig. 4. Changes in aragonite saturation $\{\Omega_{aragonite} = ([Ca^{2+}].[CO_3^{2-}])/K_{sp\ aragonite}\}$ predicted to occur as atmospheric CO_2 concentrations (ppm) increase (number at top left of each panel) plotted over shallow-water coral reef locations shown as pink dots (for details of calculations, see the SOM). Before the Industrial Revolution (280 ppm), nearly all shallow-water coral reefs had $\Omega_{aragonite} > 3.25$ (blue regions in the figure), which is the minimum $\Omega_{aragonite}$ that coral reefs are associated with today; the number of existing coral reefs with this minimum aragonite saturation decreases rapidly as $[CO_2]_{atm}$ increases. Noticeably, some regions (such as the Great Barrier Reef) attain low and risky levels of $\Omega_{aragonite}$ much more rapidly than others (e.g., Central Pacific).

Lets now discuss the role of acidification in terms of feedback mechanisms. A rise in CO_2 leads to a rise in acidity. If this is, in itself, a good thing, then this outcome is positive. Some organisms do flourish more when the seawater is more acidic than usual. Other organisms, however, are adversely affected if they make their protective shells out of carbonate of Ca and/or Mg; this is a negative feedback condition if the shells are more likely to dissolve if the acidity increase is too much. The mechanism: increased acidity increases H+ ions, and more bicarbonate ions, at the expense of decreasing carbonate ions (which can impede shell production in corals, oysters, foraminifera, etc.)

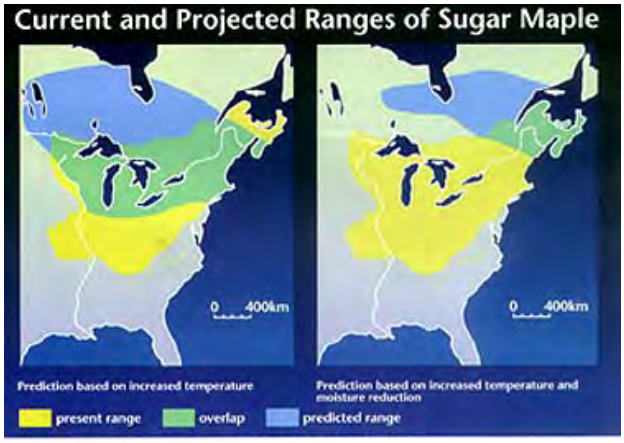
Add to the picture this situation: On land particularly, increased photosynthesis leads to more CO_2 being extracted from the atmosphere, which should impede ocean acidification. However, many land animals release CO_2 through respiration, partially restoring the balance; land plants also release CO_2 when they decay. In the oceans some organisms at depth "respire" as a metabolic process making deeper ocean waters more acidic; these do not depend on

photosynthesis for energy or bodily production. But most shallow water plant organisms (algae, diatoms, etc.) use penetrating sunlight to photosynthesize the CO_2 into organic molecules (starting from sugars). These organisms are the "primary producers" in the oceans and are the central base that establishes and controls the food chain.

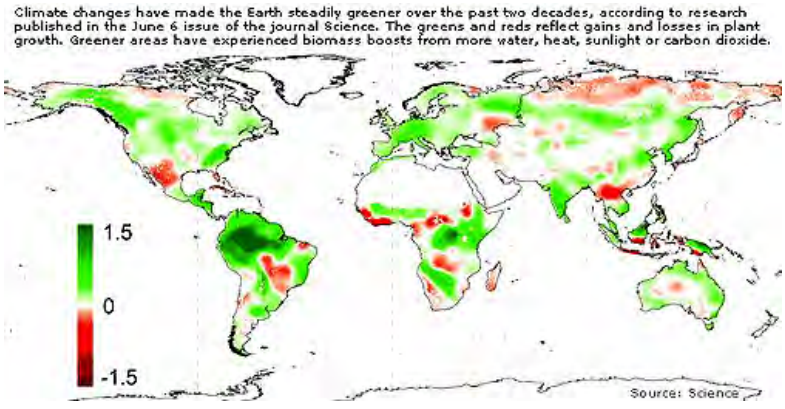
The main point alluded to in the above paragraphs is just that changes in CO_2 introduce a variety of positive and negative feedback mechanisms. Some consequences involve loss of hard parts in organisms - this is negative. Other consequences can lead to certain organisms increasing in numbers - this is positive. But the bottom line is this: While certain plants and animals may thrive if atmospheric carbon dioxide levels continue to increase, corals and other ocean dwellers are more at risk for maintenance or even survival if/when the CO_2 levels lead to dangerous pH levels.

Effects of Global Warming on Vegetation and Ecology

Returning to the theme of consequences of global warming on land surfaces: Another possible sign of global warming would be shifts in the distribution of both natural vegetation and food crops. Consider this next diagram which is a map pair showing, on the left, the current range of the common sugar maple with the shift that could occur from just the predicted temperature rise associated with a CO_2 increase to 700 ppm, and, on the right, a more drastic withdrawal northward if soil moisture reduction is included. This happens simply because the climate zone that favors this maple is sensitive to specific temperature and moisture ranges.



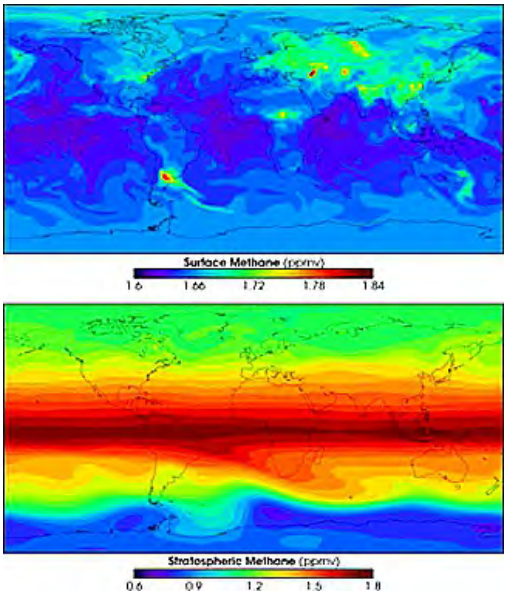
This shift in vegetation is still being demonstrated. Measurements of green leaf distribution and other measures of vegetation (various Veg. Indices) over the last 20 years using satellite data have now revealed a substantial increase in the "greening effect". A study reported in the journal Science includes a map that shows the regions of the world that have experienced greater vegetation development and areas losing vegetation, largely as a result of climate changes:



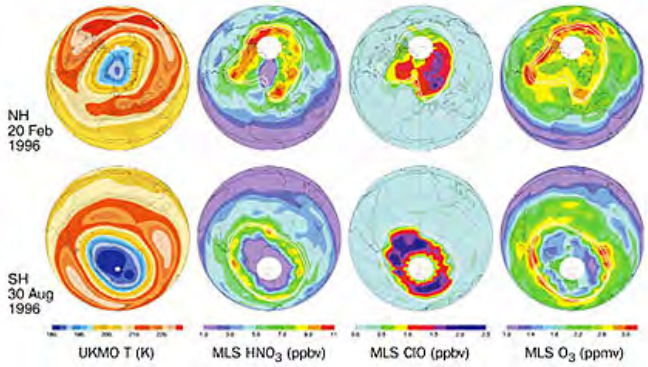
In parts of the Northern Hemisphere, the greatest increase (red is highest) in vegetative cover seen in this summary diagram shows a distinct concentration in the northern parts of Europe, Asia, and North America. To the south, the first Spring leafing is now about 1 week earlier and Fall loss of leaves is almost a week later. The most likely explanation is the warming effect of greenhouse gases and the greater availability of CO_2.

Even more affecting would be the changing conditions suited to supporting certain staple crops. A CO_2 rise to 550 pm would redistribute crop yields worldwide for common grains, as shown below. Note that warming in the higher latitudes of the Northern Hemisphere would favor increases in crop production in Canada/Alaska, Scandinavia, and most of Europe and Russia but changes in South America and most of Africa would move towards drops in yield.

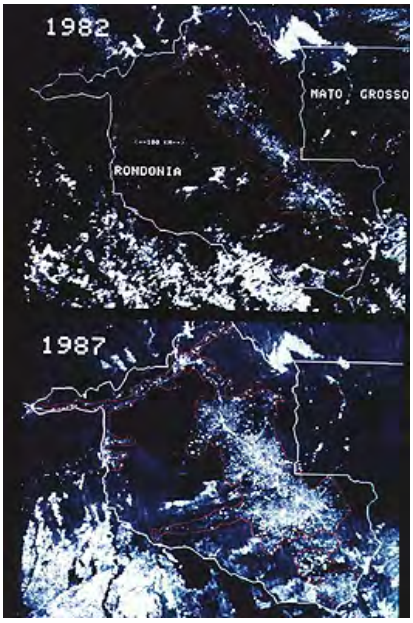
Other gases, including those that contain sulphur compounds, contaminate the atmosphere. One of the products of vegetation decomposition, fossil fuel burning, and even animal flatulation is methane (CH_4). Although present only as traces near the surface and in the stratosphere, this gas rarely is concentrated locally at a toxic level. Using various sources of data, a NASA program has now produced a global map of surface and upper atmosphere distribution of methane:



UARS has measured summer and winter concentrations of several gas contaminants in both earth hemispheres. The first pair of plots on the left are temperature distributions. Note particularly the changes around the polar regions.



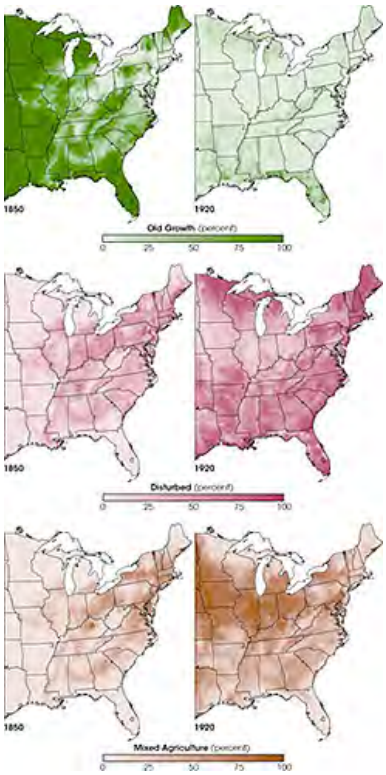
Other human activities may increase the rate of global change. One activity now grabbing attention is deforestation, whereby humans slash and burn, or just clear-cut, huge tracts of trees to use the land for agriculture or the wood for building shelters. As developers denude these large regions, biodiversity decreases, and land-use, water run-off patterns, and local weather phenomena change. Satellite remote sensing has produced dramatic images of progressive deforestation, as witnessed in these two scenes taken five years apart over the State of Rondonia in the Brazilian Amazon Basin by NOAA's AVHRR.



16-7: Estimate the percentage increase in deforestation in the middle of the image pair (eastern Rondonia).ANSWER

In recent years, awareness of the effect on global warming of changing landuse has come to the fore. The argument centers around this: Vegetation serves as a sink for carbon, which in turn influences the amount of carbon dioxide that re-enters the atmosphere. Although a complicated mix of positive and negative feedbacks, this changing cycle of carbon redistribution is strongly tied to the vegetation land cover that is affected by populations determining how the land must be used to support and feed the inhabitants.

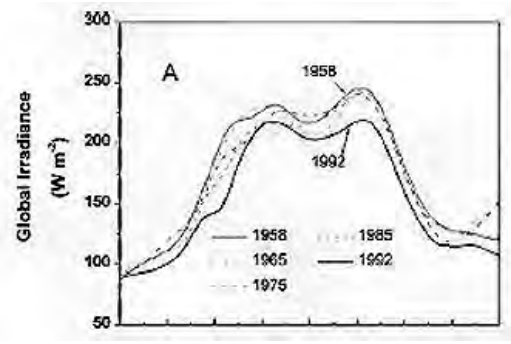
This is neatly illustrated by a historical reconstruction of the clearing of the eastern half of the United States over an extended time frame: from 1850 to 1920 - approximating the period of greatest expansion of that country's people as the railroads spread westward from the Eastern Seaboard. Consider these diagrams:



The principal change was removal of natural growth, disturbing the pristine land cover, and converting much of the land to farming. Although the replacement of woodland vegetation with crops kept some of the capacity for carbon storage, the balance that applied to carbon dioxide release was modified. Similar results have occurred both in the U.S. and worldwide as Earth's human population has notably increased over the last 100 years. Worldwide vegetation cover is decreasing, reducing the capacity to intake carbon from the carbon gases being added to the atmosphere from burning of fossil fuels and other "smokestack" activities.

The Concept of Global Dimming

After all the "hoopla" over global warming as synopsized above, it came as a surprise to NMS, and to the scientific community in general, to learn recently (as I did from a PBS Nova telecast) that there is another major atmospheric phenomenon known as *global dimming*. This refers to the now demonstrated fact that there has been an average decrease in solar irradiation reaching the Earth's surface of 10-15% since the 1950s. In some regions the number is less; in a few places the drop has been as high as 36%. This plot indicates the change (the abscissa, not identified is probably latitudes north and south of the equator):



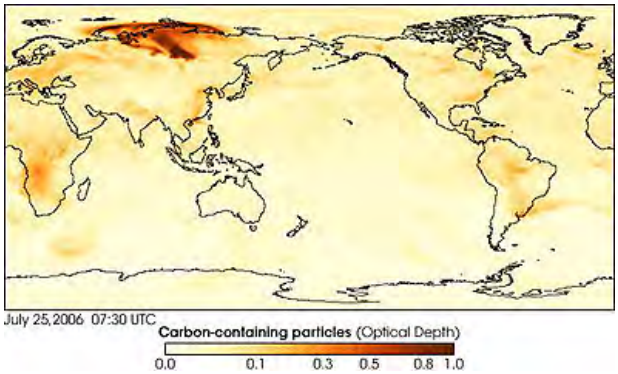
Solar irradiation at the surface is measured quantitatively by pyranometers, such as shown in this cluster;

The decrease in irradiation varies, being generally less over the open oceans and more on land.

The average drop in heavily populated regions is 0.41 W/m^2/year compared with rural regions at 0.16 W/m^2/year. The global dimming effect was first noticed by the English/Israeli climatologist Gerald Stanbill whose published reports were largely discounted. More recent studies by Atsumu Ohmura confirmed the postulate by Stanbill of dimming.

The overall effect of this reduction in irradiation has been to cool the atmosphere by 1 to 1.5 °C. This in turn causes a decrease in water evaporation, and - although not yet verified - shifts in precipitation including less rainfall. Droughts in Ethiopia and West Africa have been attributed, in part at least, to disturbances of the mechanism by which monsoon rains are developed and sent northward from equatorial spawning grounds. A drop in irradiation can mean that plant life will experience less irradiation and hence photosynthesis, now a subject of real concern.

The cause of global dimming is largely due to increased soot, ash, and sulphur particles released to the atmosphere by industrial activity and automobile use. The particles serve as nuclei for water droplets that make up clouds. The clouds in turn become more reflective, sending a greater percentage of incoming irradiation back into space. Much of the particulate matter is carbon-rich, being soot caused in part by regional burning of forests and grasslands (both controlled crop-management burning and wildfires). MODIS on Terra and Aqua produced this global map of carbon soot in the atmosphere; wildfires in northwest Siberia account for the high readings there.



Another factor which seems important in moderating temperatures is the large number of contrails from airplanes, as seen here:

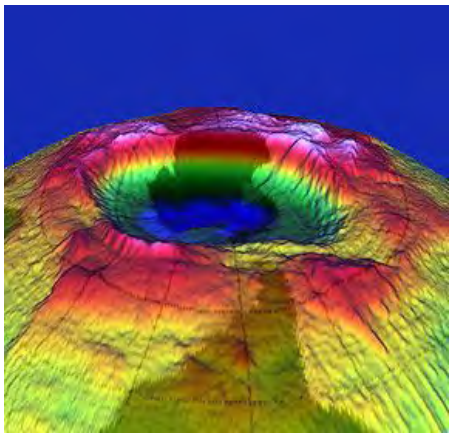
While the contrails may seem insignificant, their role was indirectly verified as a side effect of 9/11 (the Sept. 11, 2001 terrorist attacks in the U.S.). For almost 4 days after the government grounded all commercial flights in the U.S., the amount of measured irradiation increased enough to raise average temperatures by a degree Celsius. Upon resuming flights, this increase was reversed.

What is the main import of this recognition that global dimming is real? The answer is that temperature increases due to global warming are partially offset by dimming (dimming counters about half the increase due to warming). Thus, it follows that warming has been slowed by dimming which if the latter were not acting would mean that the rate of warming (by itself) would have been higher. One might suppose that dimming is therefore a good thing. In some respects, yes. But the consequences of dimming as influences rainfall and crop production may be detrimental and even a major threat to global ecology and various regions which could become both hotter and drier. **The most obvious conclusion: Reduction in both particulates and greenhouse gases should overall be a mandatory necessity for mankind to survive in safety and comfort even as population grows**. The corollary: emissions from all major sources MUST be reduced, preferably more than halved present and projected amounts. Since warming and dimming are global, it follows that this is an international threat and should therefore be met head-on by all nations. It also stands to reason that an efficient monitoring system for atmosphered "health" is vital; using satellites and remote sensing is an essential component.

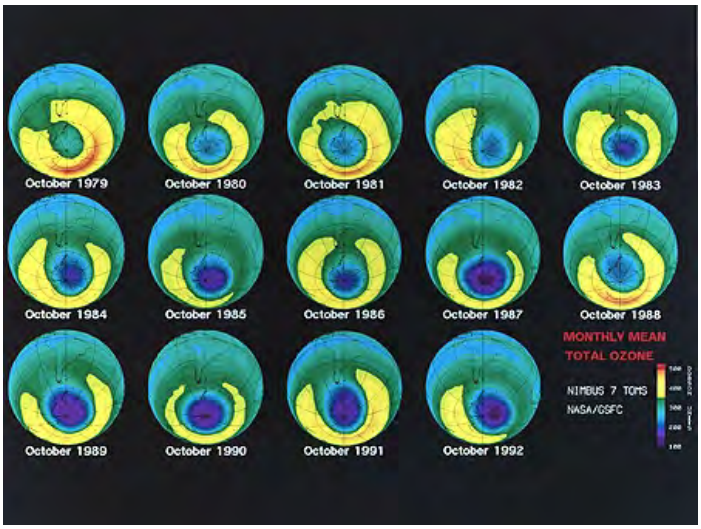
The Ozone Hole

Another observable global change, caused by certain trace gases in the atmosphere, including the CFCs, is the depletion of stratospheric ozone. Ozone in the stratosphere absorbs incoming solar ultraviolet radiation (UVR) that is dangerous to living systems. This UVR causes damage to the genetic material in living systems. Ozone prevents the UVR from reaching the Earth's surface , and so protects us from its harmful effects. Spacecraft sensors have observed ozone depletion during the Antarctic winter, an observation that helped determine the chemistry underlying this process. Over the South Pole, the right combination of cold stratospheric temperatures, ice crystals (or other solids with surfaces upon which the destruction chemistry occurs), and the global wind patterns intensify the process. Space sensors have observed similar (although smaller) depletions in the Arctic, and some of the chemical agents are increasing over mid-latitude regions, i.e., where most humans live.

These depletions have come to be known as "ozone holes". That this is an apt description is evident in this image of the Antarctic variations in ozone level that have been depicted in 3-D by "contouring" the different values:



The size and extent of the "hole" varies with time as is obvious in this series of October Antarctic ozone maps taken between the years 1979 and 1992.



16-8: Comment on the patterns of ozone change you decipher from the above October sequence. ANSWER

Summary Statements:

The Past, Present, and Future Status of Global Temperature Changes

As a consequence: All in all, the evidence seems to be mounting that there is a definite increase in regional and global temperatures. One sign, not commonly cited, is the progressive migration toward the poles of various animals and birds. In North America, rattlesnakes (very sensitive to temperatures) have moved into New England and the Canadian Great Plains. Tropical and desert birds previously almost exclusive to Mexico are now spilling over into Texas and Arizona. Populations of the Snowy Owl have been decreasing in the Arctic (ecologists surmise that this is due to a large drop in the lemning population, the owls chief food source). Some of this temperature rise may by natural (warming trends during interglacial intervals are the norm); some is now recognized by experts as the consequence of manmade perturbations to the atmosphere.

Thus, many observations and other data seem to point to humans as a major causative source, having at least the potential for modifying global phenomena. However, we are not always sure

of this. Scientists continue to wonder if some of the observed changes, such as an apparent increase in atmospheric temperature, are really due to human activities? Or are they part of a natural cycle that we are only now observing in detail, because of the presence of instruments and sensors that were hitherto not available? (Most likely, they are some combination of both a natural trend [as observed in previous interglacial climate histories] and Man's deletorius contributions of wastes to the atmosphere and hydrosphere.) Concerned individuals also must ask whether the current trends will continue and how detrimental they may be.

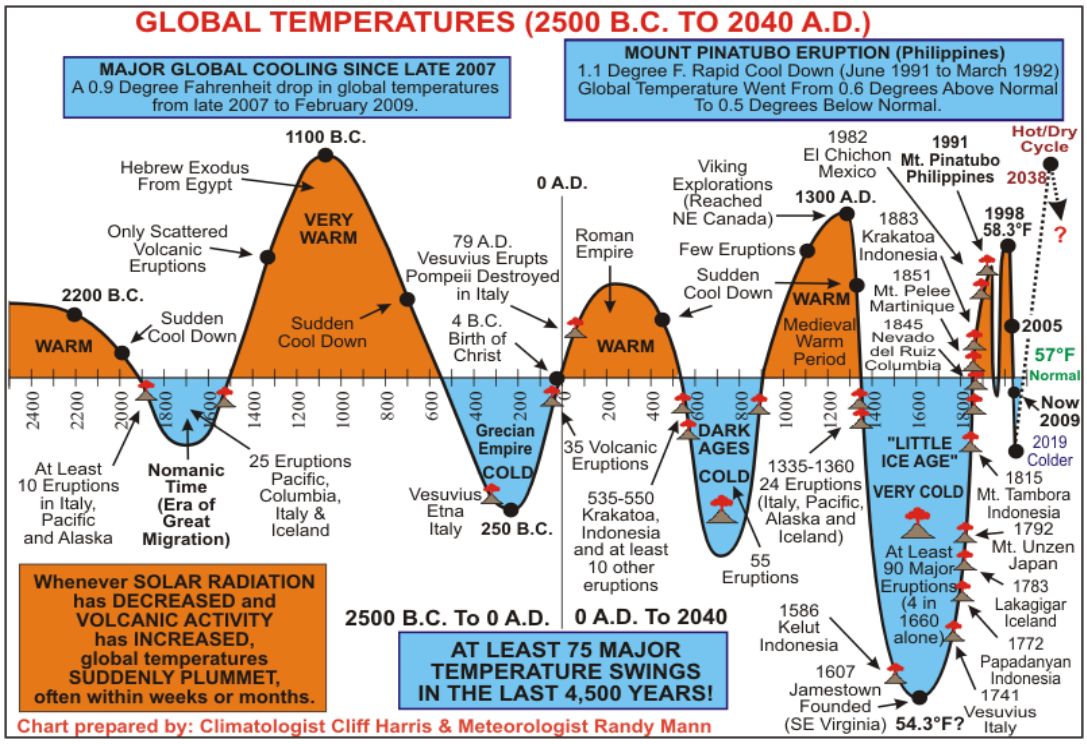
Comment by NMS: The evidence presented on this page may seem "alarmist". But what if the general conclusions and predictions being made by more and more scientists have significant elements of TRUTH. The potential is for *Calamity*. We still are somewhat in the dark as to what may occur if certain ill-defined thresholds of change when temperatures exceed some value(s) - will threatening conditions of climate and responses to its change actually cascade (the above-mentioned accelerating rate of glacial flow over just a few years in Greenland glaciers may be an example). There are many who feel great concern that there is trouble ahead if politicians, globalists, and the public turn their backs to the dangers and languish in the status quo. Both writers of this Section **strongly** recommend that YOU read Al Gore's new book: "An Inconvenient Truth", keeping an open mind and perhaps some skepticism. But if you come away with some conviction that he's largely right, then resolve to take action, raise your protest, and demand the needed rectifications.

As you might gather from the above discussion, global warming is one of the "hottest" topics going on today in both U.S. and International societies. The American people are divided right down the middle on whether manmade global warming is a reality. One recent poll among the general population has 48% of those questioned choosing to be skeptical - they do not believe that human contributions to global warming are a major threat; instead they suspect the trend is due primarily to normal interglacial warming. Most of those polled are not scientists and hence lack the special expertise a scientific background provides, but some scientists themselves have challenged the conclusions cited above and a few even the evidence itself. The subject of global warming is certainly not trivial - it may be vital to survival of the way of life as modern man now lives by. And called for solutions are troubling: they may involve trillions of dollars in expenses and lost revenue, and perhaps even fundamental changes in lifestyles in both developed and emerging nations.

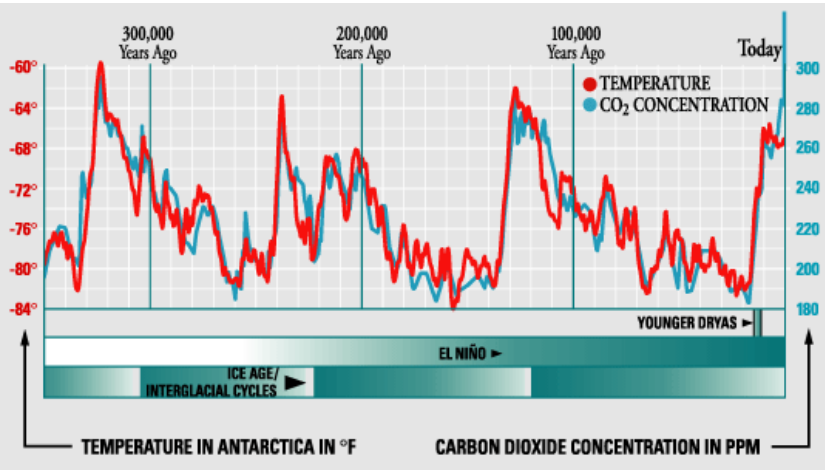
Part of the problem is that the data supposedly corroborating global warming can be contradictory and confusing. If one goes onto the Internet and types in various phrases associated with "global warming", a huge number of 'hits' in both text and illustrations comes up. Some are similar to one another; some are disparate. Some seem to support the 'human factor'; others support the idea that warming is just a part of the natural variability. Obviously, NO CONSENSUS.

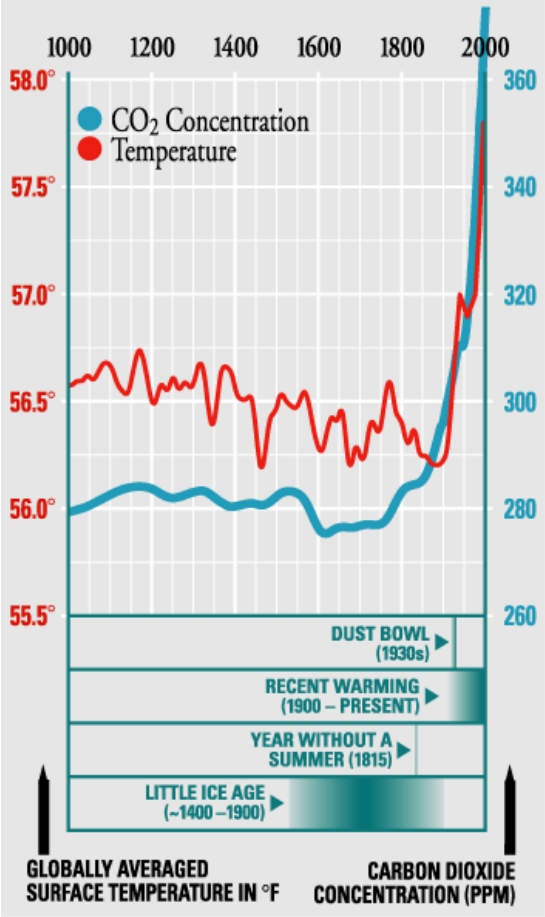
But, the dire warnings that many scientists and others are putting forth today indicate that the "worse case scenario" is realistic enough to warrant immediate and in-depth attention, since its projected outcome would be catastrophic. During the week of the International Climate Conference in Copenhagen (December 2009), the writer (NMS) has decided to expand this page with what may seem redundant relative to the coverage earlier on the page. However, I wish to provide a synoptic overview of what I consider the gist of the argument. I have extracted some key illustrations from the Internet to serve this purpose. Read through the next several paragraphs with an open mind.

The first figure puts the warming history of the last 4400 years in perspective with respect to human history. It seems to favor the natural cycle hypothesis pertinent to climate behavior. It gives little detail on the temperature history of the last 100 years or so.

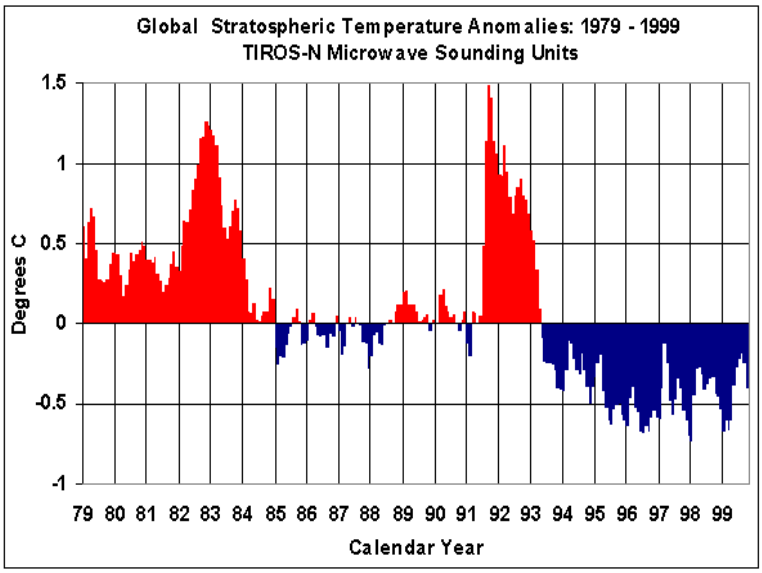


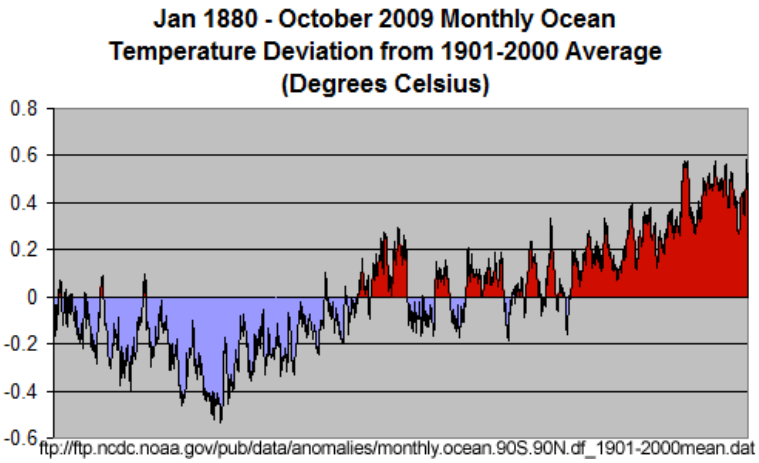
The next two figures cover the climate history of the last 350000 years (top) and the last 1000 years (bottom). The top illustration once more re-enforces the cyclical interpretation of global warming. The bottom illustration is dramatic - it is very similar to one used by Al Gore in making his supportive argument for possible runaway warming - in that temperature is rising in the last 100 years at a rate that is far greater than any previous time in the second millenium.



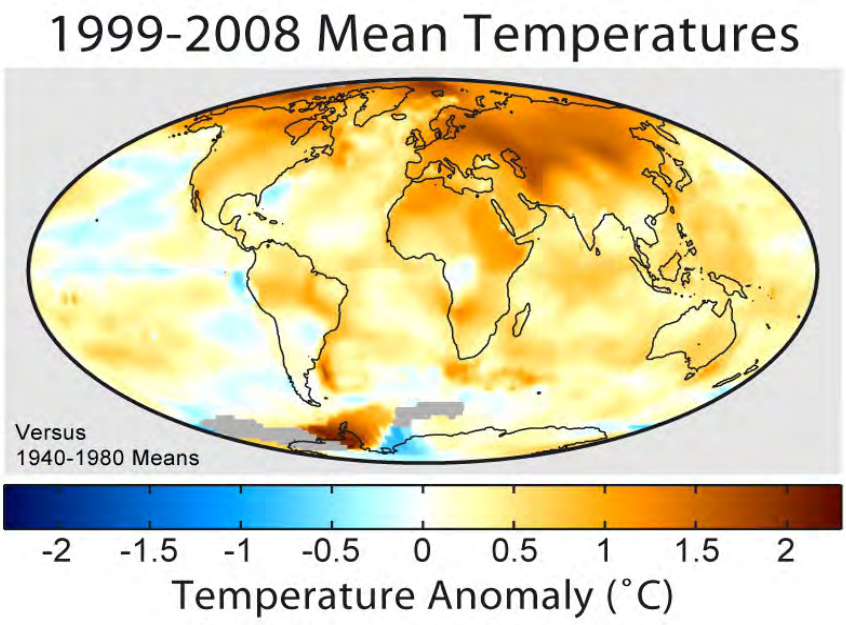


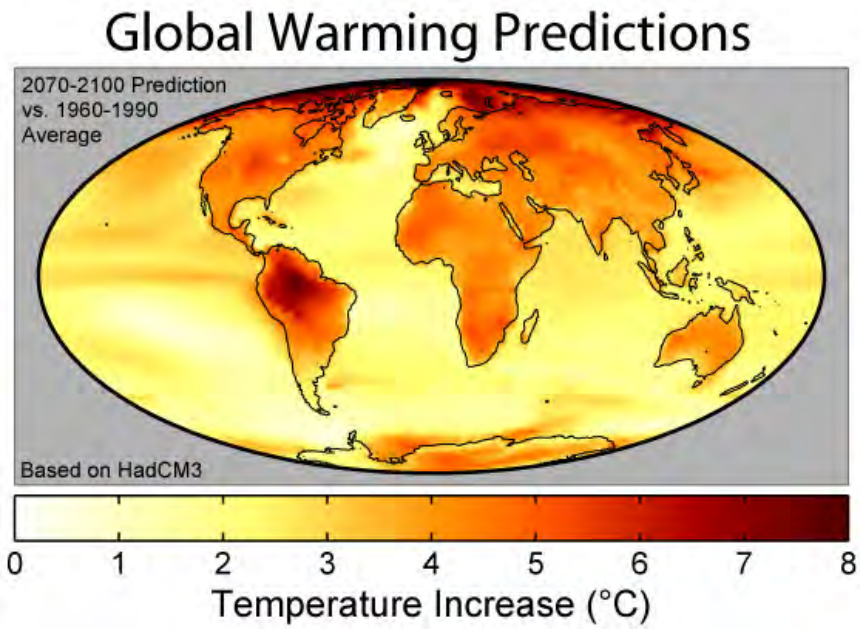
The next two illustrations are examples of data conflict. Both are plots of temperature variations over the last 30 years. The upper one seems to show warming in the early '80s whereas the lower plot seems to indicate cooling. This conflict is unexplained but it could mean different behavior in different (not corresponding) places on Earth. What they do tell the viewer: the data can seem inconsistent and therefore misleading.





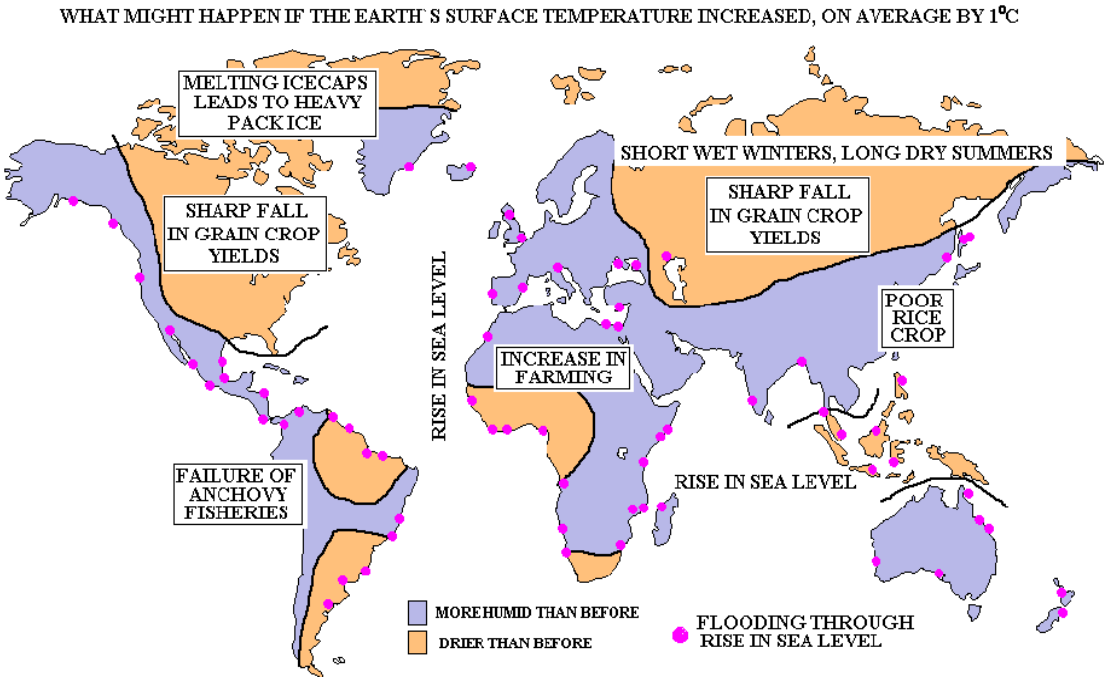
Two global maps are frequently cited on the Internet as indicative of the recent and the future temperature changes worldwide. Examine these:





TO THE WRITER, THE BEST EXPLANATION OF THESE OBSERVATIONS (APPLYING OCCAM'S RAZOR [the simplest explanation is usually the most likely one]) IS THAT THE ATMOSPHERE IS ON THE UPSWING OF A NATURAL INTERGLACIAL WARMING CYCLE BUT HUMANS ARE (THROUGH MISMANAGEMENT OF THE ENVIRONMENT) RESPONSIBLE FOR CAUSING NOTABLY HIGHER TEMPERATURE RISE RATES THAT WILL EVENTUALY POSE SERIOUS THREATS TO SIGNIFICANT NUMBERS OF THE WORLD'S POPULATION.

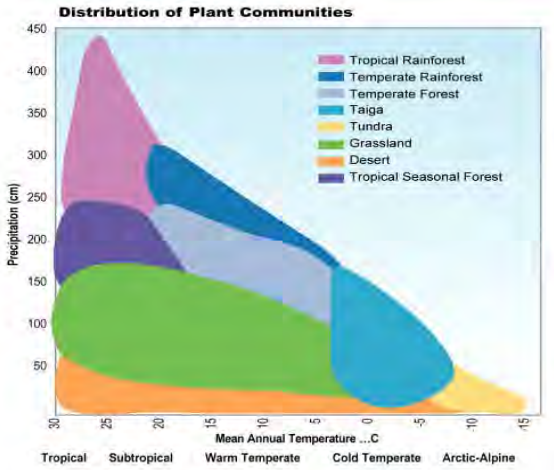
So, what are the likely consequences of all this: Here is a conjectural map of the World in which global warming has induced many major changes (those coming from coastal submergence are omitted). (The map as shown is large but reducing it using the offloaded image file causes the lettering to become blurred.)



Such maps are disturbing in themselves but do not necessarily cause outright alarm. They are rather benign. What does get the "worry beads" in hand are illustrations such as this:

This next "photo" is obviously a fake since New York City isn't now underwater. But it does bring home dramatically the effect of the sea level rise predicted for the worse-case scenario. If in fact the oceans receive water from Greenland and the Antarctic to the extent that some forecast, then many coastal cities would be inundated. Probably New York City would escape the fate shown here because dikes would be built to protect it. Nevertheless, this, and other cities along coasts, would experience radical changes.

While damage resulting from slow but inexorable inundation would occur in the low areas along the coasts, the entire land mass of each continent will be affected. Temperature increases and amounts of seasonal rainfall will modify and displace geographically the major biomes on the Earth's land surfaces. We repeat here a diagram shown on page 3-1. It shows the dependence of any biome on temperature and precipitation. These biomes should survive global warming but they will shift to different regions of the continents. In all likelihood, people will gradually respond to the shifts and relocate but animals may be more unable to adjust and some species may verge on extinction.



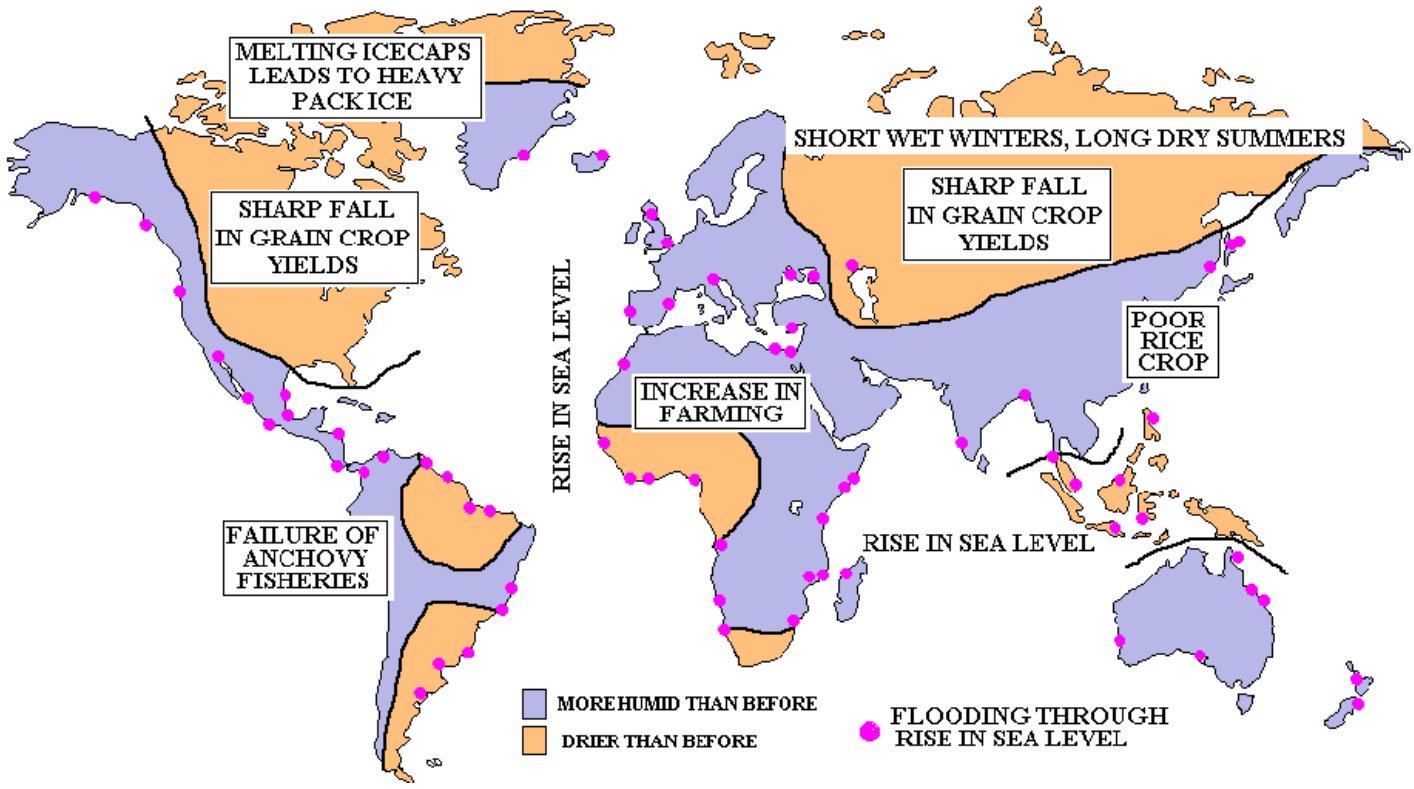
Both temperature and precipitation limit the distribution of plant communities. The climate (temperature and precipitation) zones of some of the major plant communities (such as temperate forests, grasslands, and deserts) in the US are shown in this figure. Note that grasslands' zone encloses a wide range of environments. This zone can include a mixture of woody plants with the grasses. The shrublands and woodlands of the West are examples of grass/woody vegetation mixes that occur in the zone designated as grasslands.

With climate change, the areas occupied by these zones will shift relative to their current distribution. Plant species are expected to shift with their climate zones. The new plant communities that result from these shifts are likely to be different from current plant communities because individual species will very likely migrate at different rates and have different degrees of success in establishing themselves in new places.

Critics of global warming, i.e., those who claim it is not really happening or that the changes are just an interglacial warming trend, cite various lines of evidence they consider as proof that manmade warming is not dominating climate changes. The winter of 2009-2010 is one such instance. In parts of the eastern U.S., record snowfalls (since such records began in the 1800s) in cities like Washington, Baltimore, and Philadelphia are cited as indications that the climate is really not warming. On February 10, 2010 a snowfall of 30 to 30+ inches fell on these cities, as displayed in this MODIS image:

So, is this a valid claim? The brief answer is that the hallmark signature of a "cold climate", namely snow and ice, is not proof in itself that global warming is not underway or that we are entering a period of cooling or perhaps a mini-glaciation. The several heavy snows in the East are the direct consequence of a more active and strong El Niño in the eastern Pacific ocean. This has led to greater evaporation of seawater and hence more moisture-laden storms that travel across North America. In winter, these storms encounter cold Canadian or Arctic air (which is in place as normal) and therefore are resulting in the snow events. Thus, this is a temporal and locational anomaly. Other parts of the world have had a warmer winter than usual. This simply goes to show that proving global warming requires observations over decades and must be evaluated in terms of evidence that can be clearly related to human activities.

We close this page with a conjectural map of the World in which global warming has induced many major changes (those coming from coastal submergence are omitted). (The map as shown is large but reducing it using the offloaded image file causes the lettering to become blurred.)



Mankind can probably live with these changes as we adjust and find ways to accommodate them. But the Gore book describes some climate-influenced events that will be devastating and considers certain possible effects that may be catastrophic. The many varied observations described on this page seem to favor the global warming hypothesis. But we must reiterate one statement made above: **This warming may just be (largely) a natural effect of an interglacial temperature process** (but nevertheless there clearly seems to be a component of the observed temperature increase that correlates well with humans' adverse contributions). If so, not to worry excessively (but some protective response is in order); if not, then the Gore conclusion - that very serious damage to global environments is about to occur - should be a warning that seems to cry out for a quick and drastic response by the planet's polluting nations.

In early 2007, a panel of more than a thousand scientists released a pair of reports summarizing what Science can state with high reliability about global warming. To paraphrase some of their conclusions:

1. Over the past few million years, natural climate changes, driven mainly in response to factors that cause widespread glaciation and alternate interglacial periods, are marked by intervals of slowly rising temperatures (interglacials) and then falling temperatures (continental glaciation). Rates of change are usually less than 1 degree per thousand years, and the total range is probably less than 20° F.

2. During interglacials, some animal or plant species adapted to colder weather have died off (the same may happen during warm weather; new species can develop during either cold or warm climes).

3. During interglacials, sealevel tends to rise and vegetation distribution varies; many biomes migrate poleward; productivity of edible plants may increase or decrease.

4. It is likely that mankind has blossomed since the end of the last glaciation and has become both numerous and dispersed in the last ten thousand years. Now humans are in such abundance as to begin to affect (impact) natural conditions that control the weather, moderate the environment, and otherwise maintain healthy living conditions.

5. In all probability, some of the temperature rise whose rate is increasing is just a consequence of the natural interglacial warming. But a proportion of that rise (amount still guesswork) is almost certainly the influence of detrimental atmospheric gases that are being added from fossil fuel burning and other sources. The rise may be abnormal - it is the sum of both natural and manmade contributions.

6. Since the human factor was not present in earlier interglacial warming trends, its current role in causing global warming cannot be evaluated in terms of outcomes. Many scientists believe that the rate and extent of temperature rise can be excessive enough to cause global equilibrium (adjustments of life to the changes) to be upset, perhaps to a stage that will lead to catastrophe. Work must be done to predict what might happen.

7. Because of the uncertainties, the scientists strongly urge that mankind plays it safe and reduces or eliminates the known causes of atmospheric warming.

SCARED??!!

If so, continue onto the next ADD-ON page which delves into the evidence for current warming owing to both natural and human causes.

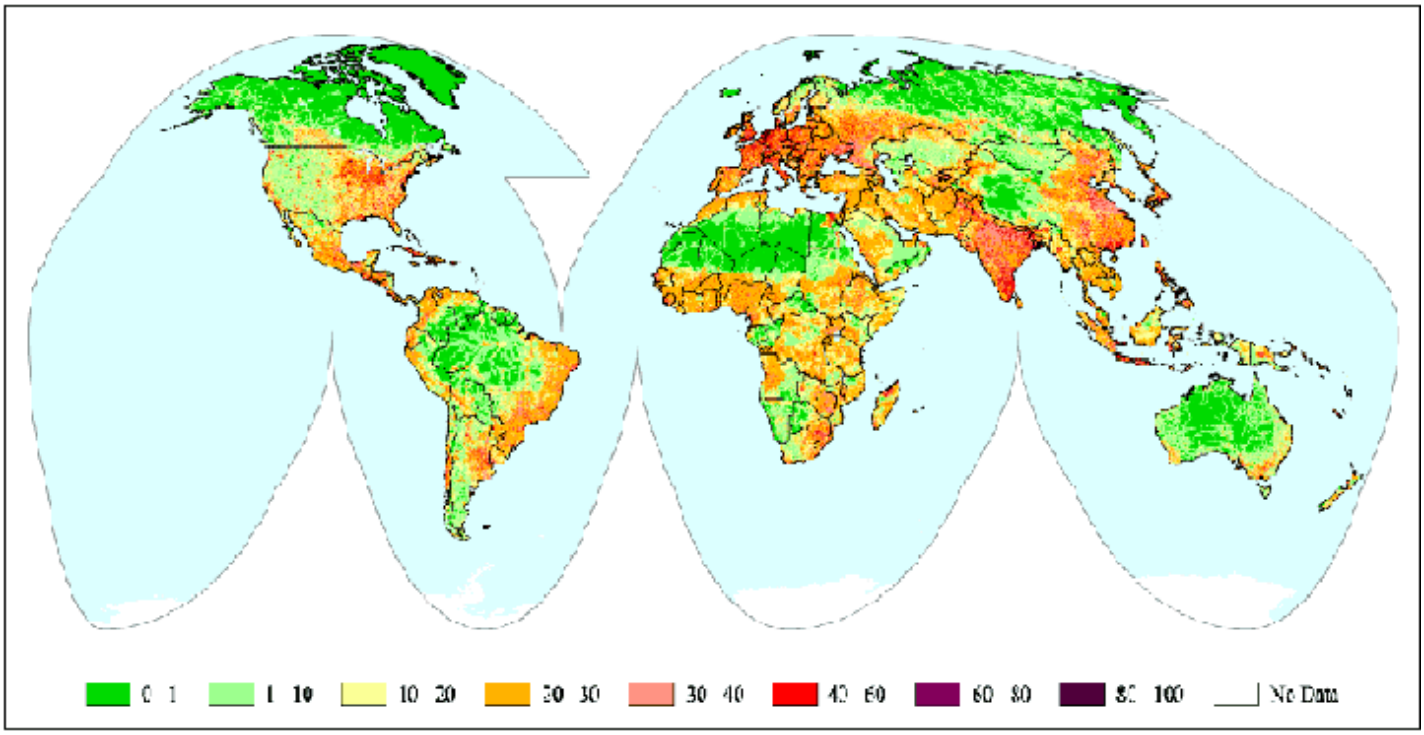
PREVIOUS NEXT

Primary Contact: Nicholas M. Short, Sr.

The growth is the Earth System Science (ESS) approach has accelerated in the last 20 years. A diagram indicates how different disciplines and specialists are involved in its activities. This diagram also lists the fields of science and technology that play roles in ESS. The various natural phenomena characterizing the Earth system tend to occur over different lengths of time (time-scale effect). These may be cyclic, as illustrated by the water cycle.

Earth System Science

The evidence continues to pile up regarding Man's impact on the natural world and its environment. Recently, an international group of scientists prepared and published a global map which they titled "The Human Footprint" that plots biome disturbance at the present time. They based this map (GIS-like in construction) on four prime variables: 1) population density; 2) land transformation; 3) human access; 4) power infrastructure. Satellite data were included in these assessments. Maximum impact is to the right.



As you might expect, Europe, India, eastern China, eastern United States and the coastal areas of South America show the largest biome modifications (in the 40 to 60 range). The minimal damage (greens; low values) in the Australian continent reflects the continuing tendency for its population to reside close to the coasts. Nevertheless, the fact that much of the world's land surfaces have already been damaged (whether this is reversible is still an unsettled question) places the focus on what we must consistently monitor if we conclude that it is necessary to control future human activities to moderate or even restore the natural biomes that are beneficial to mankind.

Clearly, we need some answers, if we are to understand our planet's physical and biological phenomena and how they interact, and to determine human contributions to global change phenomena. We need this information so we can make decisions to reduce any detrimental changes we cause, or to adapt to circumstances we cannot change. We must collect data that allow us to understand the basic forces and rhythms of our planetary system, and to distinguish

global changes due to natural effects from those due to human activities. This global analysis, in turn, requires us to acquire data about virtually every aspect of the Earth system over a long enough period that we can see contributions from long-term phenomena (e.g., an 11-year Solar cycle) and trends. Then, we must use the data to generate information that helps planners and policy-makers to ensure a sustainable environment for all of us.

The evolution of the new field of Earth System Science (ESS) over the last twenty years is largely the logical consequence of these factors:

(1) The growing interest in and concern for the environment

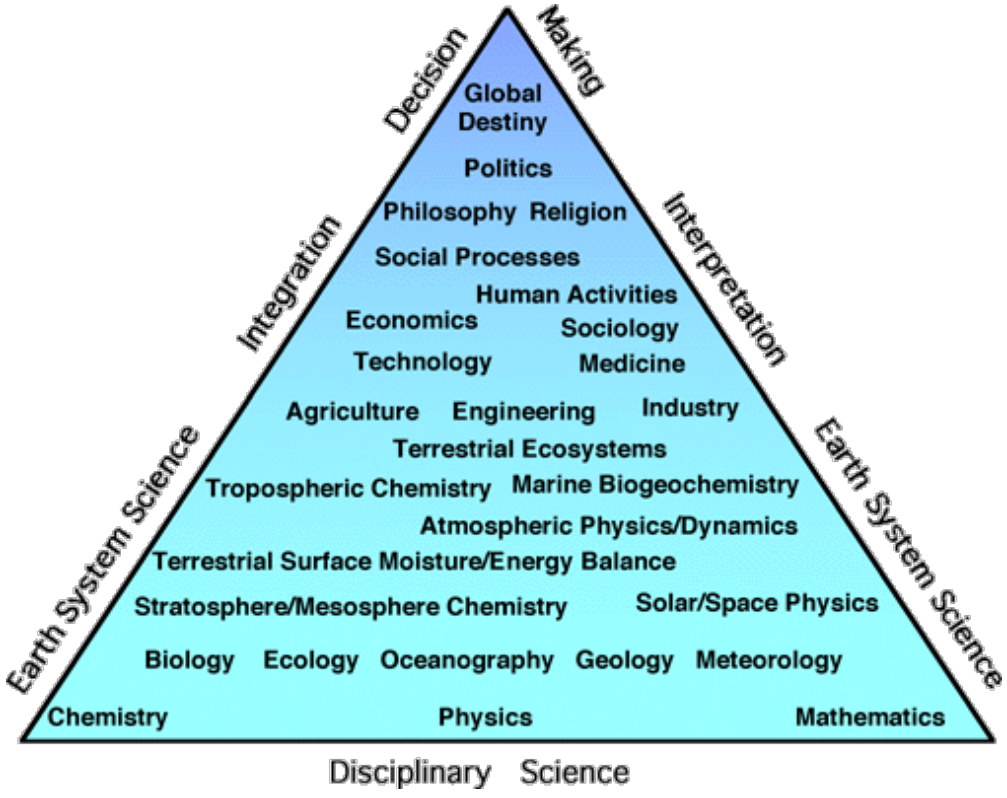
(2) The increasing awareness that biological and physicochemical Earth processes and activities often are mutually interlocked or influential, so that each controls the other to some extent

(3) The recognition that many key processes act on a regional, and often global, scale and that we must examine and interpret them at those levels

(4) The realization that many scientists work in fields and subfields in the Earth sciences that are interrelated and share similar study topics and methods, and hence they would benefit by working together in teams

(5) Technology advancing to allow global monitoring (from space).

ESS's hallmark, then, is that it is an interdisciplinary (multidisciplinary) scientific endeavor made by specialists in aspects of Earth Science - geologists, oceanographers, meteorologists, ecologists - working together with biologists-botanists-agronomists, chemists, and physicists to investigate a wide range of physical and biogeochemical activities that affect the environment (including resources). They do this by treating Earth as a complex but integral system of entities. They concentrate on its functions from a total, or full-Earth, vantage point, i.e., primarily from the global perspective. As now developed, Earth System Science works with the knowledge associated with three major, natural, dynamic, operations: the hydrologic cycle, the physical climate system, and biogeochemical cycles. To the above list, comprised of scientists and technologists, can, in fact, be added such contributing professions as sociologists, economists, and legal specialists.



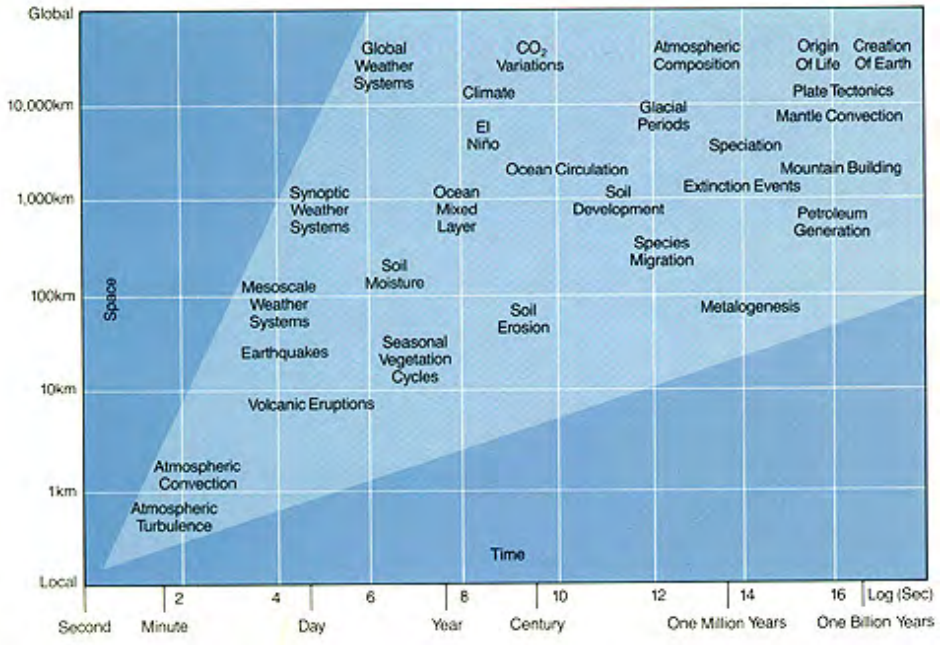
A prime goal for integrating the contributions by the various disciplines in the above diagram is to

determine how the earth system interacts when viewed from differing perspectives. Put another way, Earth System Science studies relevant components, interactions and exchanges, and ecological ramifications among the following arbitrary "spheres," each defined by its own sets of characteristics and internally-related activities and subsystems: Geosphere; Ocean (sphere); Hydrosphere; Atmosphere; Cryosphere; and Biosphere.

16-9: Many universities still don't offer courses called "Earth System Science". What two courses taught for decades in the past come closest to ESS in content? ANSWER

However, ESS is gradually appearing in the curricula of institutitions ranging from colleges to the lower grade levels of Elementary Schools. It is now possible for pre-college teachers to take one or more courses in Earth System Science online at various institutions. For more information, click on ESSEA to access the website that describes these opportunities.

In keeping with the Huttonian Doctrine ("The present is the key to the past" - and almost certainly the future as well), many of the processes have operated continuously at time scales ranging from minutes to vast millenia. Most of these processes also perform at different spatial scales, ranging from local to subcontinental or over vast oceans. For the atmospheric system, the range is at a global horizon. We synopsize this idea in this generalized space-time diagram:



This diagram suggests that we must study some processes over relatively small regions and short time frames, whereas others require worldwide monitoring over extended periods (often beyond a human life span). Nevertheless, even for the latter, we can glean useful information from determining the magnitude and frequencies of changes within the processes, as they proceed through time and shift in space. Changes affecting large segments of the planet, or even the entire globe, will likely have profound effects within the Earth's natural systems.

16-10: What parts of the above time-size plot are not especially favorable to satellite observations? ANSWER

A synopsis of Earth System Science from an educational viewpoint is given on this esse21 website that is sponsored by USRA - the Univesity Space Research Association - a consortium.

PREVIOUS NEXT

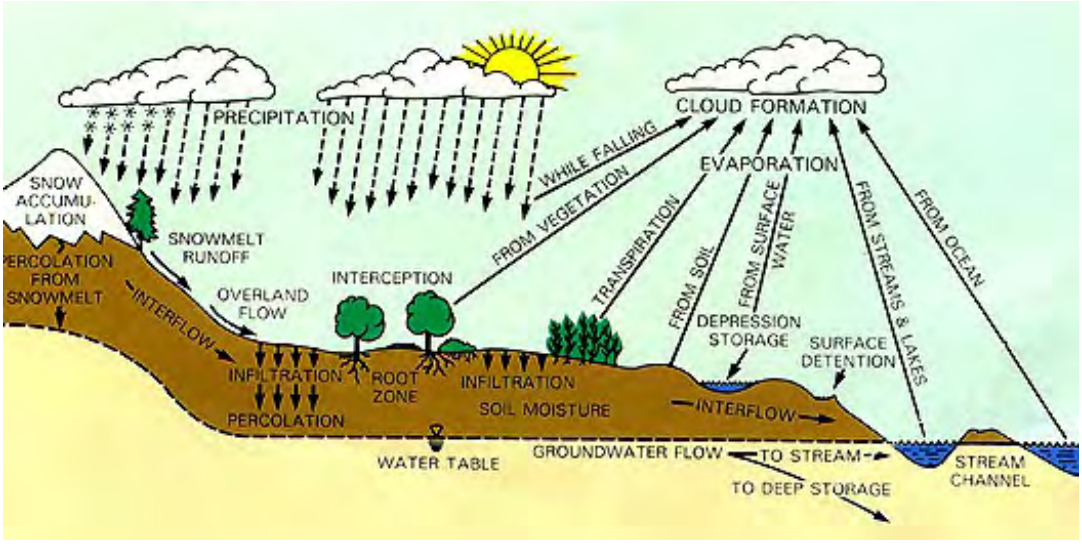
Primary Contact: Nicholas M. Short, Sr.

As stated earlier, many of the Earth's natural systems tend to operate as cycles. The Water Cycle is of prime importance. At the heart of Earth Systems Science (ESS) is the Bretherton cycle. It is briefly explained on this page. Examples of biochemical and carbon cycles further emphasize how various factors and components interact.

Earth System Cycles

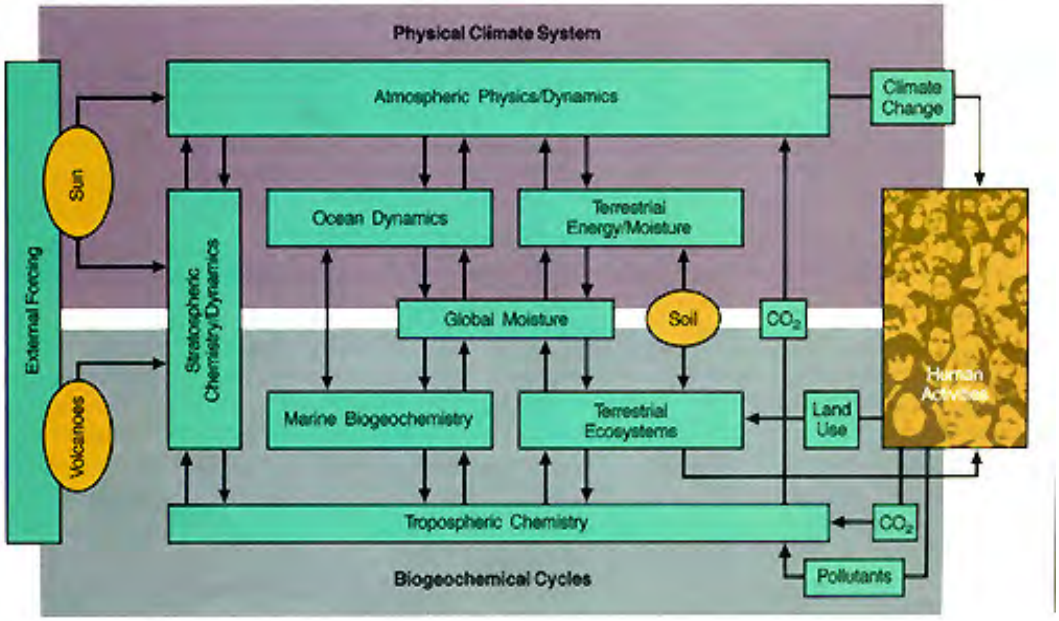
Any treatment of the Earth System needs to recognize the importance of water in developing and maintaining life. Refer to the diagram at the top of page 14-1 for a reminder of how water is partitioned among its environments in the oceans and on land.

On the rest of this page, we show several schematic or flow diagrams that give some indication of the intricately convolved chemical and biochemical components and areas (topics) of import to the major Earth Systems - singly or combined - as they persistently modify the processes acting within them. First to look at is the vital Water Cycle.



This deceptively simple diagram is similar to those found in nearly all textbooks on Environmental Science, Physical Geology, and Meteorology. It shows the pattern of water leaving and returning to the Earth's surface. The process is cyclic in that, eventually, a water molecule leaving the ocean to enter the atmosphere ultimately returns to that source. The rates of movement and the quantities involved (often shown in more detailed diagrams) vary among the specific processes (and within a process this variation changes in place and time of year). But, in the long run, the cycle continues, such that the various amounts maintain an overall mass balance, neglecting losses to outer space and gains from meteorites.

One of the classic diagrams supporting Earth System Science relates fluid and biological Earth processes that operate between the Physical Climate system and the Biogeochemical system, receiving external energy inputs from the Sun and Earth's internal heat and being affected by human activities. We present a simplified version, known as the Bretherton Diagram, below.

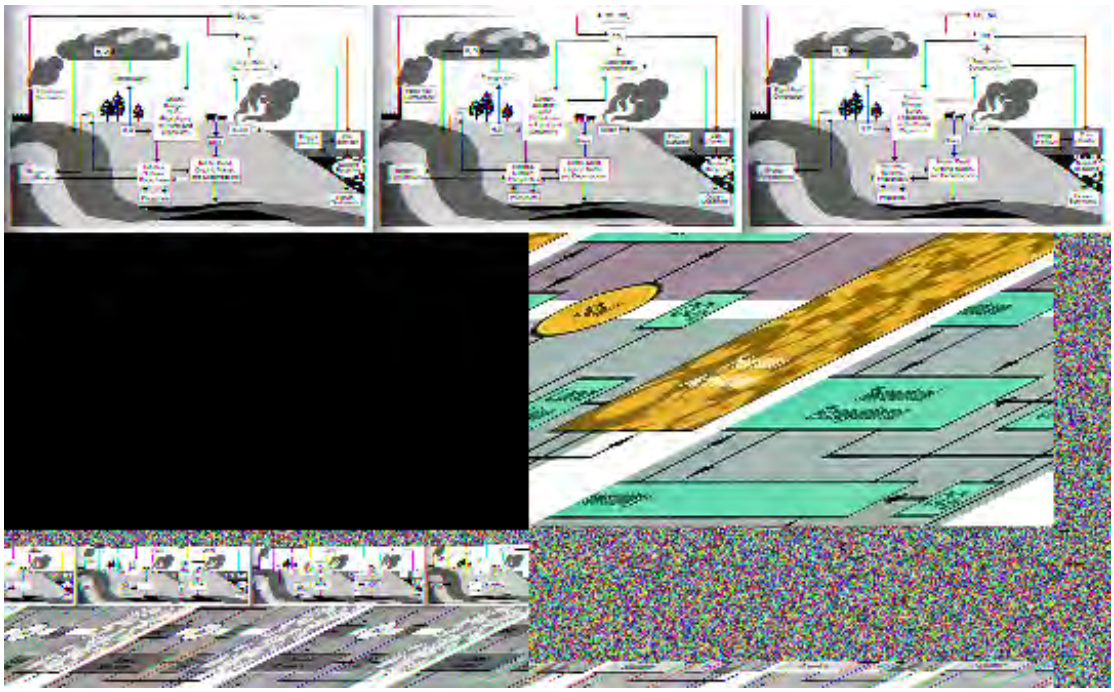


We invite you to peruse this diagram to gain a feel for how these processes, driven by energy, climate, biogeochemistry, and human activities, interrelate.

16-11: What are the main drivers or controllers of these processes? ANSWER

The ultimate driver is solar energy (orders of magnitude greater than volcanic energy or the standard heat flow from the interior [mantle and crust; radioactive decay is the prime source). Estimates of energy in calories from the Sun during a full year (averaged daily over the rotational cycle) are around 13 x 10^{23} cal. per annum (or at any given second, 40 million billion calories/sec). Much of this energy is partioned into heating of the atmosphere, the oceans, and the land surface; most of the remainder is involved in plant photosynthesis.

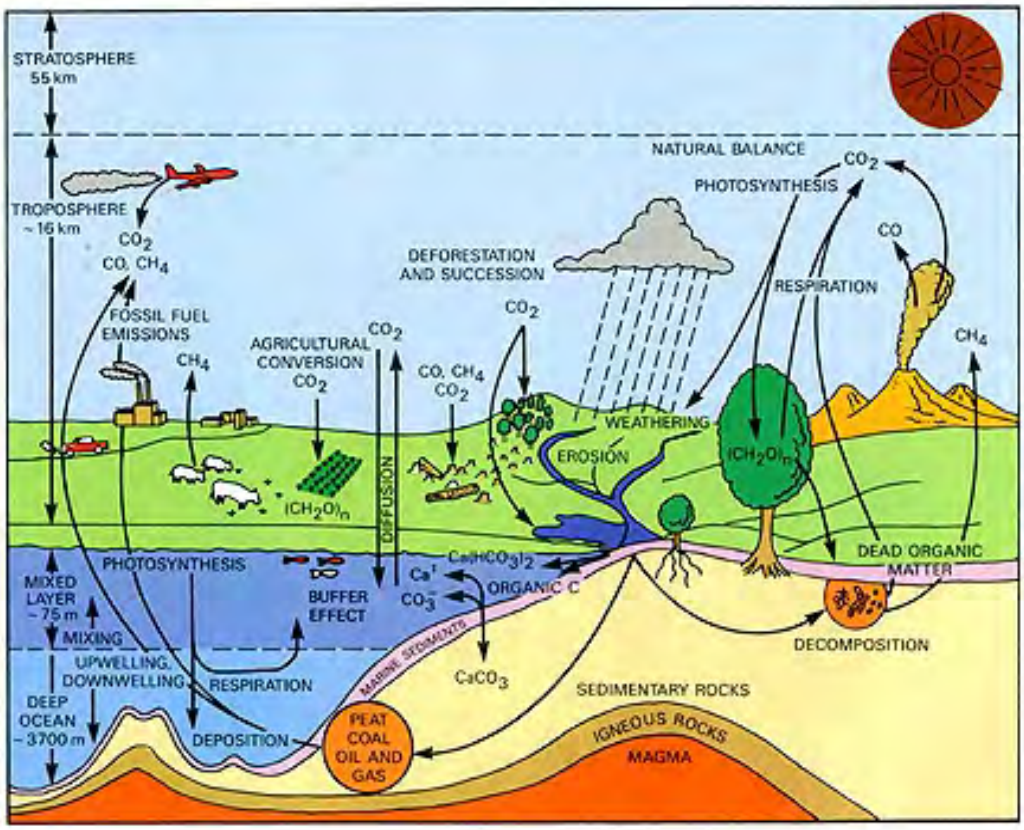
Below, we summarize biogeochemical cycles, which depict the flow of important elements and compounds within the ecosystem as conveyed by biological and (geo)physical processes:



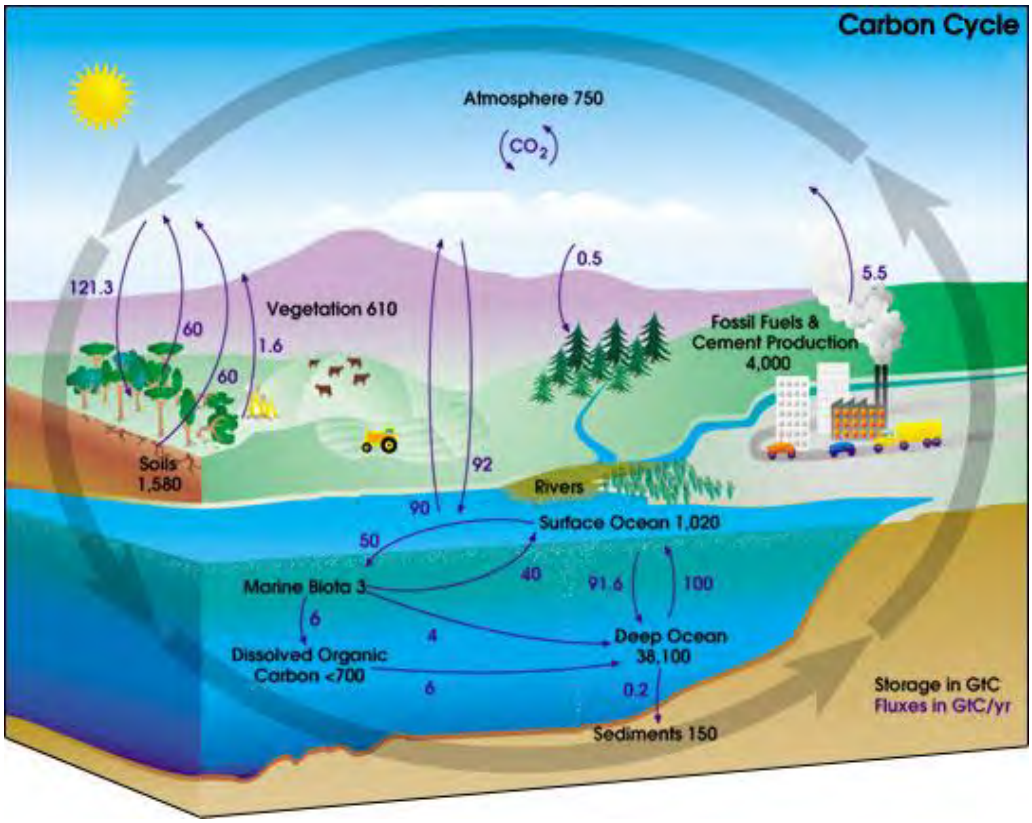
16-12: Do you believe everything you see in the above diagram is correct (sufficient connections)? ANSWER

One of the key cycles is that of the migrating and locating of carbon. Other geochemical

element-based cycles include oxygen, nitrogen, phosphorus, sulphur, and iron. The next diagram outlines the main constituents (carbon and its compounds), and where they reside at any time within the four spheres. We find carbon in the solid Earth mainly in limestones and in petroleum and gas deposits.



Information on the mass balances within the carbon cycle has been obtained with fair reliability. We depict here the exchange of carbon, in units of 10^{15} g C yr^{-1}, among the major reservoirs.



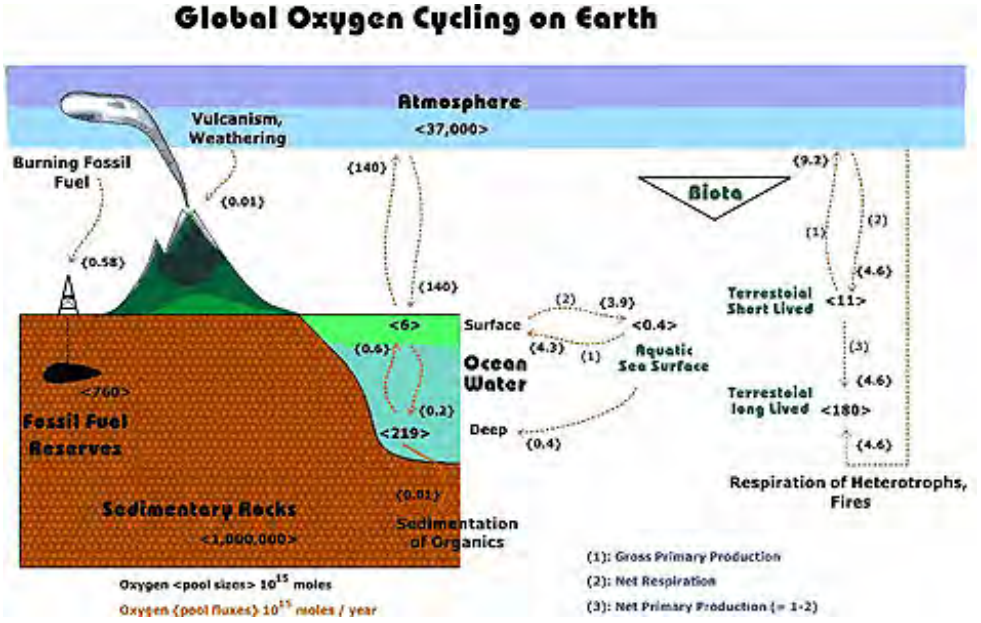
Carbon dioxide is a critical component in this cycle, being free in the atmosphere, dissolved in the

oceans, integral in plant cells, and locked into sediments (limestones).

16-13 Checking the above diagram, what is the main "source" (returning CO_2 to the active environment) and the main "sink" (receiving and storing the CO_2) in the carbon geochemical cycle in today's world? ANSWER

Because of the importance of this cycle, we recommend you visit this Carbon cycle site which reviews the fundamentals that affect this cyclical process.

Almost every important element within the Earth's spheres, and some of the more significant compounds (e.g., CO_2, NO_3, SO_2) have been traced as they move through the systems. There are many geochemical cycles besides those just examined. Those of iron, calcium, phosphorus, sulphur and silicon are well-known. We show one more example: Oxygen.



Cycles, such as those shown above, have interconnected links and passages which influence their components. The process is called *feedback*. A positive feedback increases or enhances the process or activity. A negative feedback reduces or reverses the process. A good review of the feedback concept is found on this Wikipedia website.

Here is an example of feedback that is pertinent to global climate change. In Canada, parts of Europe, and Siberia, and elsewhere, there are large areas that contain peat bogs. These result from partial or inhibited decomposition of accumulations of vegetation, mainly in lakes and swamps. In the geologic past, complete vegetation decay could result in coal deposits. Peat bogs represent incomplete decay, usually under conditions in which colder temperatures are the norm. Peat bogs formed during glacial periods over the last few million years. But now, with the rapid increase in CO_2 owing to human-related activities, such as auto exhaust or smoke from coal-burning plants, the average temperatures and temperature maxima are rising in part because of the greenhouse effect described on page 16-2. This is producing a positive feedback that affects the peat bogs by causing them to warm up and accelerate their decay. As a consequence, this will add signficant amounts of CO_2 into the atmosphere, further exacerbating the rise in temperatures. One negative feedback from this will be the decrease in ice accumulation that means reduction or disappearance of glaciers and ice caps.

From the above, we sense that Earth System Science deals primarily with matters of climatology and atmospheric physics/chemistry, as influenced by essential interactions with the ocean reservoir, water partitioning on the land, biological intakes and effluents, and particles and gases released from volcanoes. Unlike Landsat and similar programs, the emphasis of Earth System Science and its supporting EOS programs has shifted from solid Earth and land use to the

environment.

Accelerated global change is, by definition, a global problem. In addition, as we already noted with the example of the SO_2 aerosol generated by Mount Pinatubo, what happens in any region of our planet may affect many other regions. How, then, do we address and keep checking on these global concerns?

As we shall see later in this Section, the entire field of Earth Systems Science is at the heart of a huge, multinational effort to use space platforms to monitor the Earth's atmosphere, oceans, biosphere, and active geosystems on the land. As part of that complex program, NASA has inaugurated a series of low cost, quick development satellite-based missions that go under the name Earth System Science Pathfinder (ESSP). Beginning in 2002, the hoped-for one to two launches per year will each be highly focused on a single problem. Each mission will be run by the organization that "wins" in the semi-annual Announcement of Opportunity. That organization (be it public or private) will be funded to design, manage, and interpret the mission program, with help from NASA. The first three missions selected are code-named GRACE, Calypso, and Cloudsat. These are described as part of the ESSP website.

PREVIOUS RST NEXT

Primary Contact: Nicholas M. Short, Sr.

Under the "umbrella" of the Global Change Research Program, and other similar programs, many of the world's nations have joined together to expand their study of the Earth System and to find ways to control and manage the environment. In the United States many federal agencies are participants. Hundreds of scientists and applications specialists work together within committees and teams to coordinate these activities; many of these groups include foreign members.

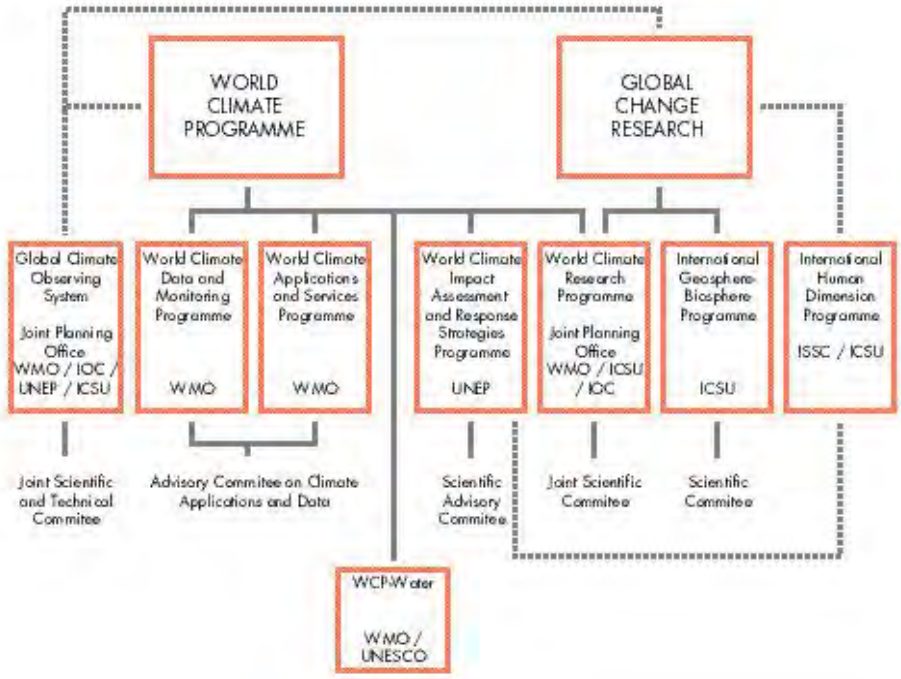
National and International ESS Programs

In the first joint activity of such scope, many nations of the world have established the Global Change Research Program (GCRP). Scientists and policy-makers around the world have met regularly for more than three decades to establish priorities for data acquisition that will help answer many of the questions we face. Members of the Intergovernmental Panel on Climate Change (IPCC) composed a list of seven major geophysical and biological phenomena to monitor. All signatories to the GCRP are addressing these phenomena in one way or another. These seven prioritized areas are:

- Water and Energy Cycles
- Oceans
- Chemistry of the Troposphere and Lower Stratosphere
- Land Surface Hydrology and Ecosystem Processes
- Glaciers and Polar Ice Sheets
- Chemistry of the Middle and Upper Stratosphere
- Solid Earth

Those nations that are technologically capable work together to design instruments and spacecraft and to define data-gathering needs to make this large-scale examination of our planet as efficient as possible. To that end, scientists from one nation may design and build instruments, while a group from another nation designs and builds the spacecraft. Other parties establish data policies that allow the free flow of data across international (political) and physical boundaries.

Some idea of the various groups that have been set up to coordinate the GCRP and its closely related World Climate Programme appears in this organizational flow chart:



The principal players —(nations and organizations) are:

- Canada
- Committee on Earth Observation Satellites (CEOS)
- European Space Agency (ESA)
- EUMETSAT (European Meteorological Satellite organization)
- France
- Germany
- International Council of Scientific Unions (ICSU)
- International Geosphere-Biosphere Program (IGBP)
- Intergovernmental Panel on Climate Change (IPCC)
- Italy
- Japan
- United Kingdom
- United Nations Environmental Program (UNEP)
- United Nations/World Meteorological Organization (UN/WMO)
- United States of America
- World Climate Research Program (WCRP)

The job is so large that no group or nation alone can address it properly. Collaboration is an absolute requirement to do the job properly.

One of the largest and oldest (1982) of these programs, begun by Swedish scientists, is the International Geosphere-Biosphere Programme, an overview of which is found at this IGBP site.

In response to its agreement to participate in the GCRP, the executive branch of the US government established roles for several federal agencies in the USGCRP.

Those agencies and their roles include:

National Aeronautics and Space Administration (NASA): responsible for space-based Earth Science observations, and studies of broad scientific scope that examine the planet as an integrated whole;

National Science Foundation (NSF): promotes basic research in all areas of terrestrial, solid Earth, atmospheric, and oceanic sciences;

National Oceanic and Atmospheric Administration (NOAA): maintains a balanced program of observations, analytical studies, climate prediction, and information management;

Department of Energy (DOE): focuses on carbon dioxide and other emissions from energy supply and end-use systems;

US Department of Agriculture (USDA): conducts research to assess the effects of global change on the agricultural food and fiber production systems and on forests and forest ecosystems of the U.S. and worldwide;

Department of Defense (DoD): conducts mission-related research into environmental processes and conditions that affect defense operations, tactics, and systems;

Department of the Interior (DOI): addresses the collection, maintenance, analysis, and interpretation of short- and long-term land, water, biological, and other natural resources information;

Department of Transportation (DOT): assesses the impact of transportation primarily through the use of fossil fuels on global change;

Environmental Protection Agency (EPA): conducts research to assess, evaluate, and predict the ecological, environmental, and human health consequences of global change.

16-14: There is a hidden "player" in this list, essential to all the above agencies. Who might that be? ANSWER

More information about the above and related programs resides at this Web site: USGCRP .

In July 2003, more than 33 nations participated in an International Summit to discuss an integrated approach to monitoring the Earth's physical "health". This led to the formation of *The Global Earth Observation System of Systems* (GEOSS) initiative. The lead agency for the U.S. in the program is the Environmental Protection Agency (EPA). For an overview of the GEOSS consortium of nations, visit EPA-GEOSS.

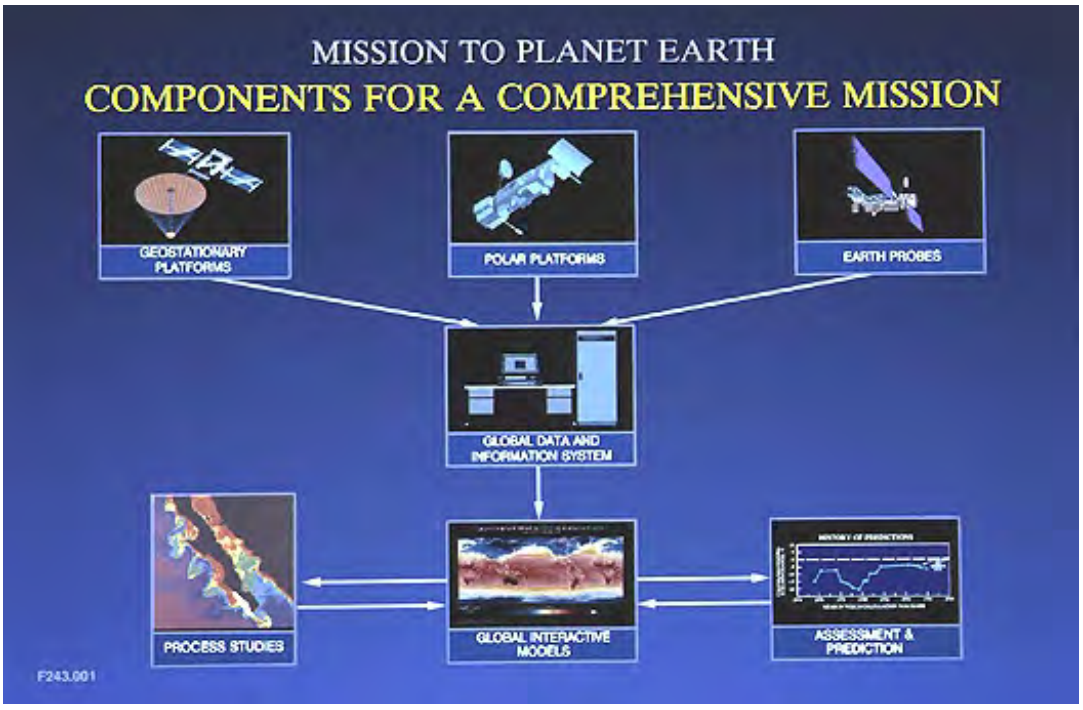
PREVIOUS NEXT

Primary Contact: Nicholas M. Short, Sr.

NASA has been a key player in the ESS programs. It initially defined its program under the title "Mission to Planet Earth". Several years ago that was renamed the "Earth Science Enterprise". A diagram depicts the broad conceptual structure of this ESE. Its principal endeavor is contained in the mission-oriented EOS (Earth Observing System) and the data handling component known as EOSDIS.

Earth Science Enterprise = Mission to Planet Earth

In keeping with its assignment to be "...responsible for space-based Earth science observations, and studies of broad scientific scope that examine the planet as an integrated whole," NASA has organized a multi-year mission, once named the Mission to Planet Earth (MTPE) but renamed to Earth Science Enterprise (ESE) in January of 1998. The prime goal of ESE is to determine the extent, causes, and consequences of global climate change and variability, and the interrelated factors involving these modifications, in order to improve our understanding of the Earth as an integrated system, and to develop and apply the predictive capability needed to aid policy makers.



ESE and its components are not science and/or technology missions *per se*, but rather, they comprise missions that use science and technology to give planners, policy-makers, and decision-makers the tools they need to understand our planet and to manage its resources appropriately. This idea brings a human dimension into the process, i.e., not only what is the role of humans in accelerating global change, but how does global change affect humans?

16-15: In the above diagram, there is one key element that is the potential "bottleneck" or rate-controlling process. Make a stab at what that component of the entire MPTE program might be, and why? ANSWER

ESE includes flying current satellite missions, developing new missions , planning future missions, managing and analyzing satellite and in situ data, and continuing a basic research program focused on process studies, modeling, and data analysis. The space-based component of ESE is a constellation of satellites to monitor Earth from space. Sustained observations allow researchers to monitor Earth's climate variables over time to identify trends. However, space-based monitoring alone is not sufficient. A comprehensive data and information system, a community of scientists performing research with the data acquired, and extensive ground and airborne campaigns are all important components. More than any other factor, the commitment to make Earth-science data easily available to the research and user community is critical to mission success.

To address the need for quantitative analysis of Earth and its various subsystems, ESE's principal element, the Earth Observing System (EOS) provides systematic, continuous observations from low Earth orbit. In so doing, EOS creates an integrated scientific observing system to foster multidisciplinary study of Earth's critical, life-enabling, interrelated processes involving the atmosphere, oceans, land surfaces, and polar regions, and the dynamic and energetic interactions among them. It also develops a comprehensive data and information system (EOSDIS), including a data retrieval and processing system, to serve the needs of scientists contributing to an integrated, multidisciplinary study of planet Earth. It also supports the overall USGCRP by acquiring and assembling a global database of remote-sensing measurements from space, in keeping with the research priorities listed above.

After scientists determine the specifics of the kinds of measurements they must make to address the many questions we face, we have to translate those requirements into a means of obtaining those measurements. This process results in a sensor design, that is, an electronic or optical-electronic device that makes the necessary measurements. Such devices require a great deal of supporting hardware, including physical support structures, power assemblies, data storage, data transmission, calibration, and more. Constraints on many of these come from cost and time, but other matters arise, such as the spacecraft size, power, and data transmission resources available (see below). In addition, if other instruments fly in concert with a primary instrument (as is often the case), we must address fields-of-view and other, interrelated considerations.

A good updated website that summarizes the Earth Science Enterprise program has been produced by Wikipedia.

16-16: Before reading ahead, make an educated guess as to what types of sensors might be prime candidates for the EOS platforms to be launched. ANSWER

PREVIOUS NEXT

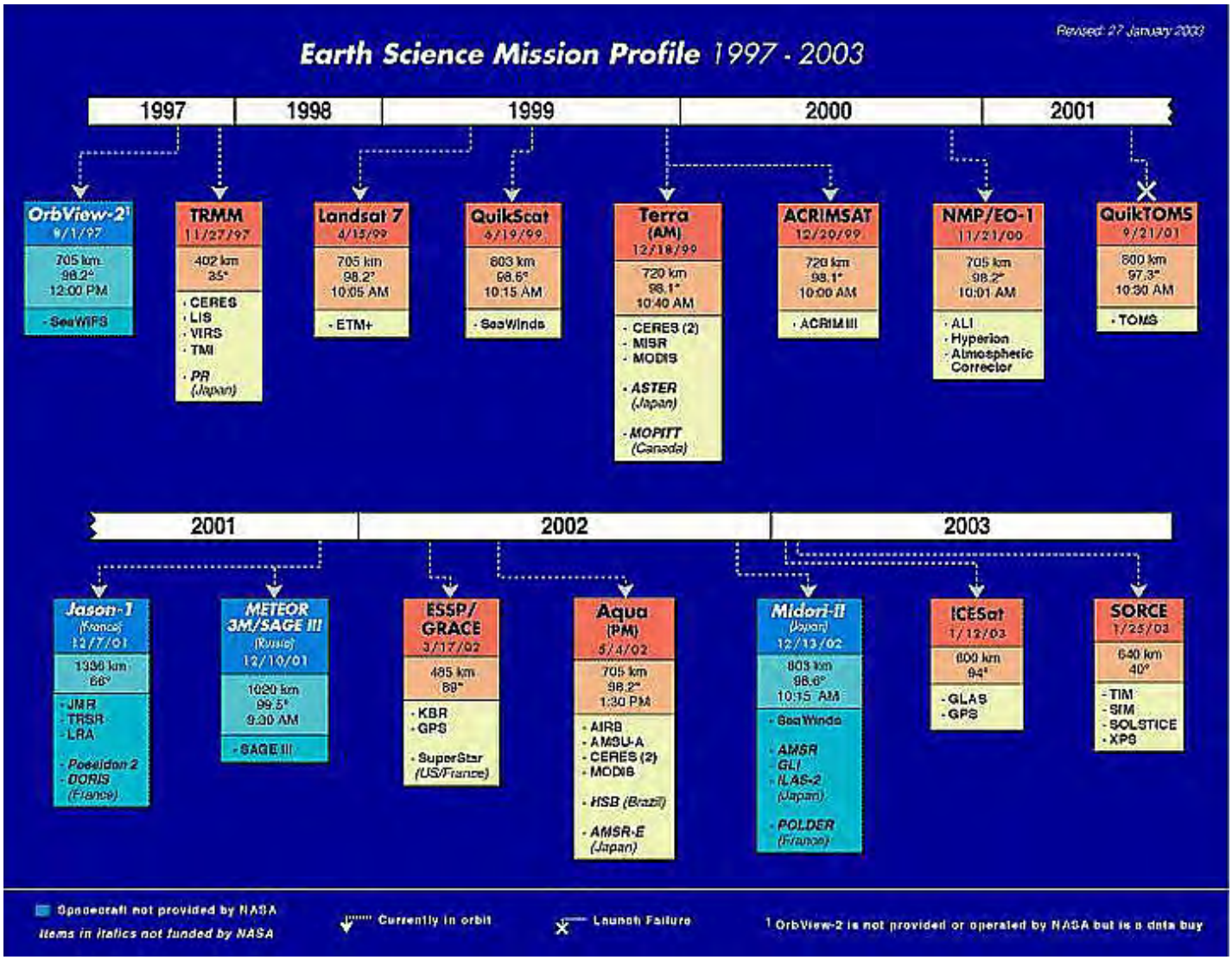
Primary Contact: Nicholas M. Short, Sr.

The ESE program is a highly ambitious series of satellite flights extending well into the first decade of the 21st century. A chart showing these many missions is shown, with the proviso that launch dates can slip and missions may be scrubbed (usually from funding cutbacks). The first EOS launch in the cavalcade to follow is Terra, which is described on this page in its conceptual mode (its actual launch and first results will follow on pages 16-9 and 16-10). Special attention is given to one of its prime sensors, MODIS, with a simulated image promising what to expect. Landsat-7 is recalled again to show how it fits into the grand scheme.

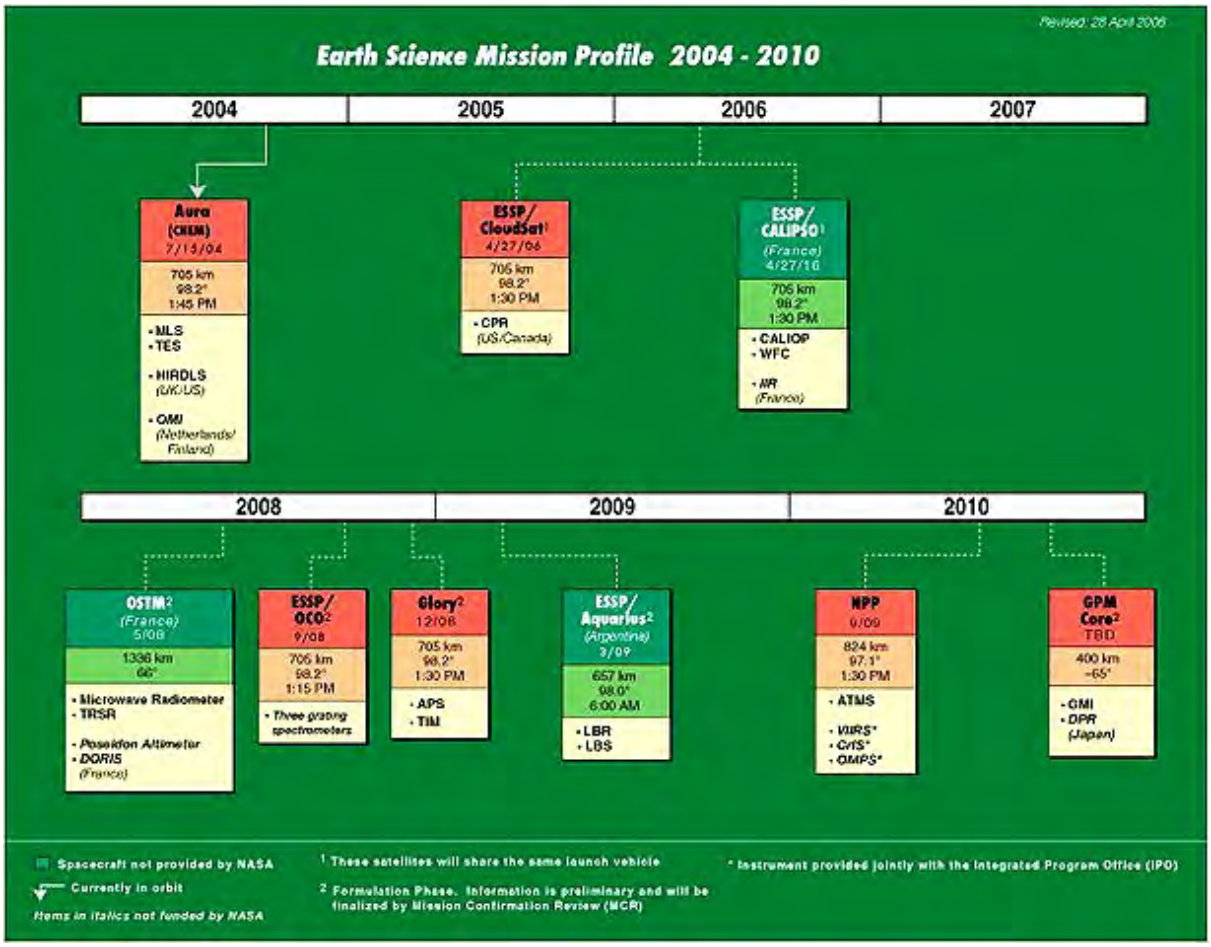
EOS Platforms/Sensors and Mission Profiles

Addressing the scientific questions posed on the preceding pages requires that the ESE program utilizes a variety of instruments–each addressing a different but related issue–to work together. We mount some combination of suitable instruments on a spacecraft that launches into a carefully-designed Earth orbit. The instruments dictate the spacecraft's size and its design, as well as the size of available launch vehicles, etc. The Earth Observing Satellite (EOS) series meet these specifications (get more details on this and the rest of the EOS program at the EOS Web Site).

The EOS program involves multi-year missions, with twenty-one different orbit-based components, similar in concept (if not in detail) to EOS AM-1. By designing many such spacecraft, based on the guidelines outlined above, we formulate a mission profile. The mission profile, over time, provides the necessary data. We present the updated EOS Mission Profile here (keep in mind that future launch dates occasionally slip):



Although more tenuous (particularly susceptible to budgetary changes), the satellites planned for launch between 2004 and 2010 are specified in this chart:



Some of these missions are described in more detail at the EOS missions website. Also, the above mission profiles may be more sharply displayed on this original site (the download at 72 dpi accounts for the fuzziness in the profile charts on this page) if you wish to see some of the fine print.

To allow for flexibility in the missions, NASA and cooperating agencies haven't assigned developed or proposed instruments to specific flights in some instances. In several places the ESE may fly an instrument as a "flight of opportunity," or "FOO." While the first profile above extends only to 2003, additional flights of instruments and flights will likely continue beyond 2006 to provide a full 15-year-long data set, as required by the scientific drivers to the program, as described earlier.

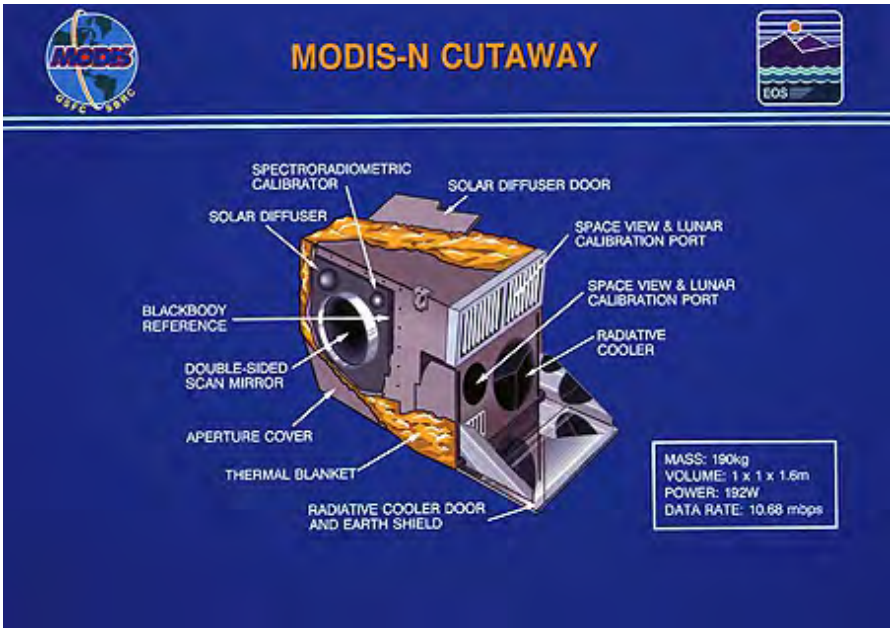
The flagship mission, EOS-AM-1, has been renamed ***Terra***. Its main thrust is to significantly improve our understanding of how the atmosphere works, especially by integrating inputs from the oceans, biosphere, and land surface. It thus is primarily a *climate*-oriented mission but with capabilities for studying other aspects of the Earth System including land surfaces that have now long been the purview of Landsat. After more than a year's delay, it was successfully launched on December 18, 1999 and is fully operational (discussed in detail on pages 16-9 and 16-10). More information about its design, purpose, and instrumentation, along with more sample imagery, can be found on the Terra Home Page.

Designed to have an orbital lifetime of at least six years, Terra (EOS AM-1) provides data on clouds, aerosols, and radiative balance. It allows us to characterize the terrestrial surface and secure data on the global carbon cycle.

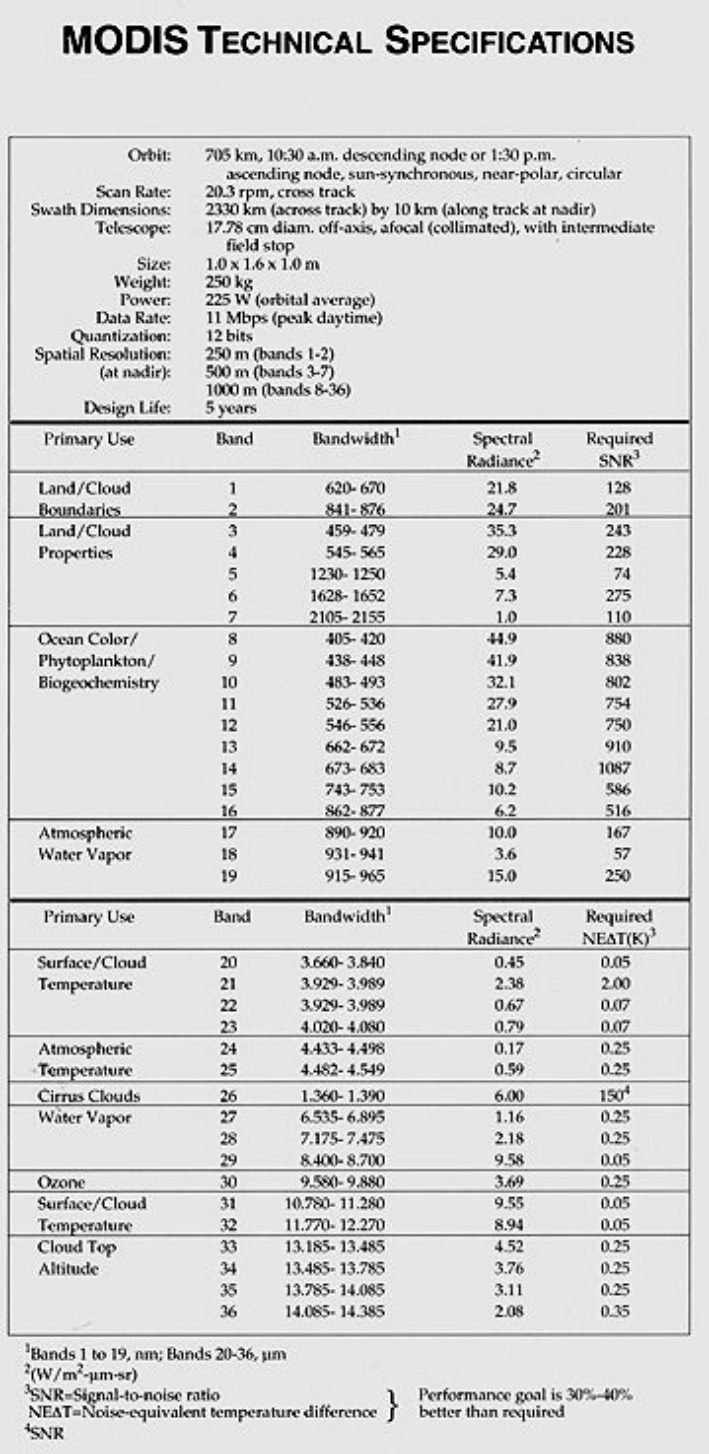
Because Terra is complex, it has neither the proper instrument complement nor the longevity to handle many of the other issues addressed by ESE and related programs. How, then, do we acquire all the needed data? The answer lies in the mission profile which clearly demonstrates that a fleet of simultaneously operating satellites with different sensors is the approach required to gather interactive data.

An example of a state-of-the-art instrument built for the first EOS flight (Terra), is the Moderate

Resolution Imaging Spectroradiometer, or MODIS. MODIS measures biological and physical processes on a global basis, every one to two days. The instrument provides long-term observations, from which we enhance our understanding of global dynamics and processes occurring on the Earth's surface and in the lower atmosphere. The MODIS instrument employs a conventional imaging radiometer, consisting of a cross-track scan mirror and collecting optics, and a set of linear detector arrays with spectral interference filters located in four focal planes. The optical arrangement provides imagery in 36 discrete bands (hyperspectral mode), selected for diagnostic significance in Earth Science. The spectral bands have spatial resolutions of 250 m, 500 m, or 1 km at nadir (the point directly below the instrument). The instrument's data rates range from 6.2 Mbps (average) to 10.6 Mbps (day) to 3.2 Mbps (night). It is temperature-controlled by radiative cooling. The instrument's mass is 225 kg, it has a 100% duty cycle, and requires 175 W (average) and 225 W (peak) power. In its stowed (launch) configuration, its dimensions are about 95 x 158 x 134 cm.



Because MODIS is the pivotal instrument on EOS, it is appropriate to list in detail its specifications.



MODIS TECHNICAL SPECIFICATIONS

Orbit:	705 km, 10:30 a.m. descending node or 1:30 p.m. ascending node, sun-synchronous, near-polar, circular
Scan Rate:	20.3 rpm, cross track
Swath Dimensions:	2330 km (across track) by 10 km (along track at nadir)
Telescope:	17.78 cm diam. off-axis, afocal (collimated), with intermediate field stop
Size:	1.0 x 1.6 x 1.0 m
Weight:	250 kg
Power:	225 W (orbital average)
Data Rate:	11 Mbps (peak daytime)
Quantization:	12 bits
Spatial Resolution: (at nadir):	250 m (bands 1-2) 500 m (bands 3-7) 1000 m (bands 8-36)
Design Life:	5 years

Primary Use	Band	Bandwidth[1]	Spectral Radiance[2]	Required SNR[3]
Land/Cloud Boundaries	1	620-670	21.8	128
	2	841-876	24.7	201
Land/Cloud Properties	3	459-479	35.3	243
	4	545-565	29.0	228
	5	1230-1250	5.4	74
	6	1628-1652	7.3	275
	7	2105-2155	1.0	110
Ocean Color/ Phytoplankton/ Biogeochemistry	8	405-420	44.9	880
	9	438-448	41.9	838
	10	483-493	32.1	802
	11	526-536	27.9	754
	12	546-556	21.0	750
	13	662-672	9.5	910
	14	673-683	8.7	1087
	15	743-753	10.2	586
	16	862-877	6.2	516
Atmospheric Water Vapor	17	890-920	10.0	167
	18	931-941	3.6	57
	19	915-965	15.0	250

Primary Use	Band	Bandwidth[1]	Spectral Radiance[2]	Required NEΔT(K)[3]
Surface/Cloud Temperature	20	3.660-3.840	0.45	0.05
	21	3.929-3.989	2.38	2.00
	22	3.929-3.989	0.67	0.07
	23	4.020-4.080	0.79	0.07
Atmospheric Temperature	24	4.433-4.498	0.17	0.25
	25	4.482-4.549	0.59	0.25
Cirrus Clouds	26	1.360-1.390	6.00	150[4]
Water Vapor	27	6.535-6.895	1.16	0.25
	28	7.175-7.475	2.18	0.25
	29	8.400-8.700	9.58	0.05
Ozone	30	9.580-9.880	3.69	0.25
Surface/Cloud Temperature	31	10.780-11.280	9.55	0.05
	32	11.770-12.270	8.94	0.05
Cloud Top Altitude	33	13.185-13.485	4.52	0.25
	34	13.485-13.785	3.76	0.25
	35	13.785-14.085	3.11	0.25
	36	14.085-14.385	2.08	0.35

[1]Bands 1 to 19, nm; Bands 20-36, μm
[2](W/m²-μm-sr)
[3]SNR=Signal-to-noise ratio
NEΔT=Noise-equivalent temperature difference
} Performance goal is 30%–40% better than required
[4]SNR

16-17: In the right column are two abbreviations: SNR and NE deltaT. What do these stand for and what do they measure? ANSWER

The instrument will provide long-term observations from which to derive an enhanced understanding of global dynamics and processes occurring on the surface of the Earth and in the lower atmosphere. The MODIS instrument employs a conventional imaging radiometer concept, consisting of a cross-track scan mirror and collecting optics, and a set of linear detector arrays with spectral interference filters located in four focal planes. The optical arrangement will provide imagery in 36 discrete bands (quasi-hyperspectral mode) selected for diagnostic significance in Earth science. The instrument's data rates range from 6.2 Mbps (average) to 10.6 Mbps (day) to 3.2 Mbps (night). It is temperature-controlled by radiative cooling. In its stowed (launch) configuration, its dimensions are about 95 x 158 x 134 cm.

16-18: Looking over the channels listed in the "specs", and their intended uses, how would you categorize MODIS as a satellite class? ANSWER

To ensure that measurements provided by the instrument are accurate, MODIS uses a complex

series of on-board calibrators. As the instrument's scan mirror makes a revolution, these calibrators provide information on a per-scan basis that places the measurements in a trustworthy framework.

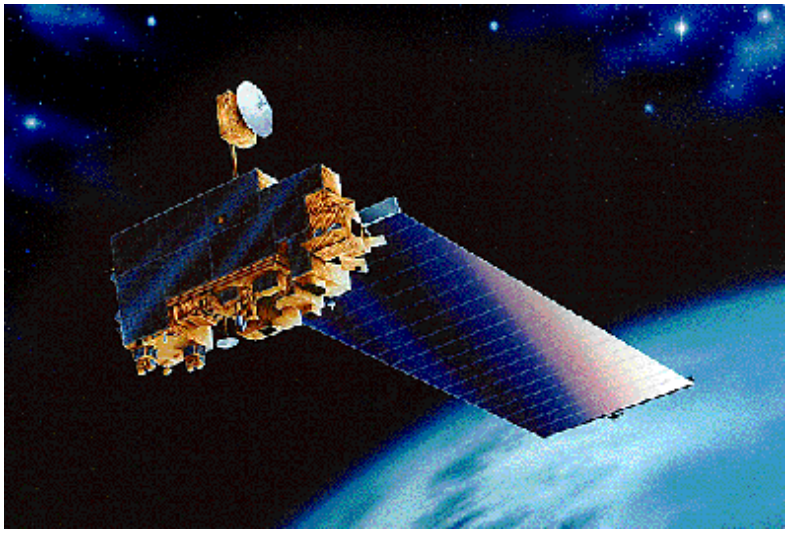
Here is an example of a characteristic 3-band color composite made with the MODIS Airborne Simulator (MAS) sensor using selected MODIS bands, showing a scene in Central Brazil taken in August of 1995.

MODIS has its own Home Page that gives updated information about that sensor.

In the case of the first EOS flight, earlier known as AM-1 because of its designated morning equator-crossing time, four other instruments join MODIS, two of them coming from some of the program's International Partners:

- CERES - Clouds and the Earth's Radiant Energy System
- MOPITT - Measurements of Pollution in the Troposphere (Canada)
- MISR - Multi-angle Imaging Spectroradiometer
- ASTER - Advanced Spaceborne Thermal Emission and Reflection Radiometer (Japan)

These instruments are on the Terra spacecraft, as shown here (the order in the list conforms to the instruments in the diagram from left to right, plus MODIS on the far right):



Of special interest in this battery of instruments is ASTER, shown here.

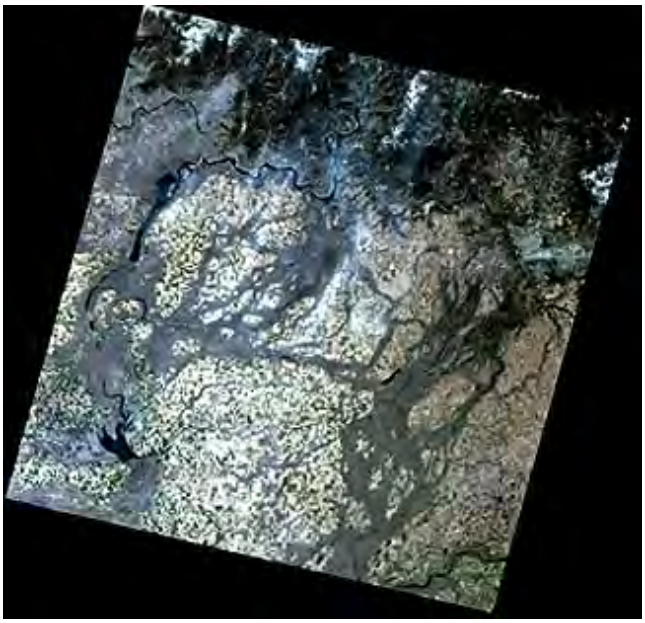
A combination of wide-bandwidth coverage, high-resolution mapping, and stereo DEM is unique to ASTER. It operates in tandem with MODIS. ASTER contains a radiometer in the VNIR, which images in three bands (between 0.52 and 0.86 µm (15-m spatial resolution). A second radiometer uses six bands in the SWIR region between 1.6 and 2.43 µm (30-m resolution). The TIR (thermal emission) radiometer has five bands between 8.3 and 11.3 µm (90-m resolution). There is also a separate single band at 0.76-0.86 µm, inclined backward at 27.6Á from nadir that provide along-track imagery, capable of stereoscopic viewing with one of the VNIR bands to permit images in the DEM (Digital Elevation Model) mode. The VNIR bands measure snow, water, vegetation, and degree of oxidation of the land surface. The SWIR bands are optimized for mineral identification, especially hydrated minerals in clay-rich soils. The TIR bands identify major rock types and also take land-surface temperatures.

16-19: What groups of scientists and technologists are most likely to use ASTER?
ANSWER

Each instrument on Terra has a mathematical (computer-based) model that allows scientists and engineers to determine its response to operating conditions. In addition, because each data-collecting measurement goes through an algorithm (a set of mathematical and operational instructions), a detailed analysis of all algorithms is available in an Algorithm Theoretical Basis Document (ATBD). The ATBD presents the scientific and physical basis for the measurement, and documents how to translate that basis into computer code that converts the physical measurement into data products. The data products give the geophysical parameters required to answer the questions that scientists and government planners ask.

Landsat-7 launched on April 15, 1999. It operates both independently and in synergistic support of EOS. It is now considered to be an integral part of the ESE program. It mounts only one sensor, the Enhanced Thematic Mapper (ETM+) that has identical spectral wavelength coverage as earlier TMs (with the same 30-m resolution, except that the thermal band improved to 60 m). It also has (the +) a panchromatic (0.5-0.9 µm) band with 15-m resolution. The special feature of Landsat-7 is its launch and orbital parameters, which allow the space vehicle to trail (follow) the EOS-AM platform (within its swath) by 15-30 minutes each day. Thus, we can integrate observations from both systems, and we can use atmospheric data from EOS sensors to correct for the effects of water vapor and temperature on Landsat images.

The first image from Landsat-7 is included on page I-21 of the Introduction. Here is another: a cropped full scene of eastern Washington state, with Spokane in the upper right (northeast) corner, mountain ranges near the top, the Columbia River, fields still mostly barren in April, and some large dark gray areas along flowing or inactive streams which comprise part of the "Scablands" formed when a great Pleistocene lake burst sending cascades of water that strongly eroded the landscape.



PREVIOUS NEXT

Primary Contact: Nicholas M. Short, Sr.

Every satellite has a mission control unit and normally a data handling center. That developed for the ESE program, designed so far to handle Terra data, is known as the EOSDIS. The volume of data received from Terra will exceed 3 terabytes daily, a huge load that will require dedicated and efficient facilities. Its main components are diagrammed. On this page also are comments and two diagrams that review what is known about the Earth's energy budget and how physical climate and biogeochemical systems are intertwined. The page ends with a simplified table of some of the satellites of relevance to the ESE that have been put into orbit or will soon.

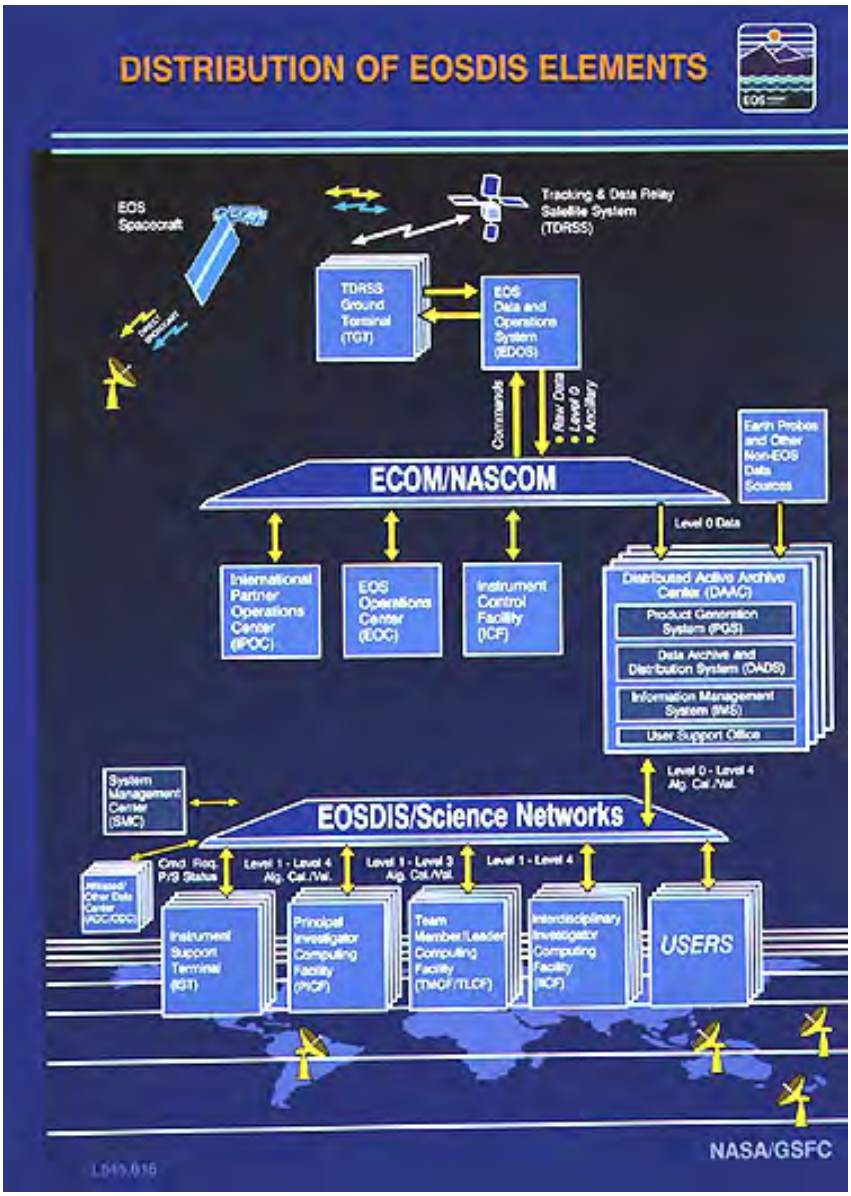
Data Handling: EOSDIS

Availability of data is critical to the success of NASA's ESE. The EOS Data and Information System , as described on this EOSDIS home page site, must be responsive to this requirement. EOSDIS manages data from NASA's past and current Earth science research satellites and field measurement programs. During the coming EOS era, EOSDIS will allow command and control of satellites and instruments, and will generate data products based on orbital observations. In addition, EOSDIS will generate data sets made by assimilation of satellite and in situ observations into global climate models.

EOSDIS's services will include:

- User Support
- Data Archive Management and Distribution
- Information Management
- Product Generation
- Spacecraft Command and Control
- Data Capture and Telemetry Processing

NASA planners base the present plans for EOSDIS on an open system, with distributed architecture. This permits data controllers to allocate EOSDIS elements to various locations to take best advantage of different institutional capabilities and science expertise.



EOSDIS consists of several components including:

1. A core system (ECS), which provides the Science Data Processing Segment (SDPS), the Flight Operation Segment (FOS), and the Communications and System Management Segment (CSMS).
2. Eight Distributed Active Archive Centers (DAACs);
3. Scientific Computing Facilities (SCFs);
4. the EOS Data and Operations System (EDOS); and EOS Networks

Together, these components will address all command and control; data acquisition, transport, reduction, storage, and visualization; and user access needs.

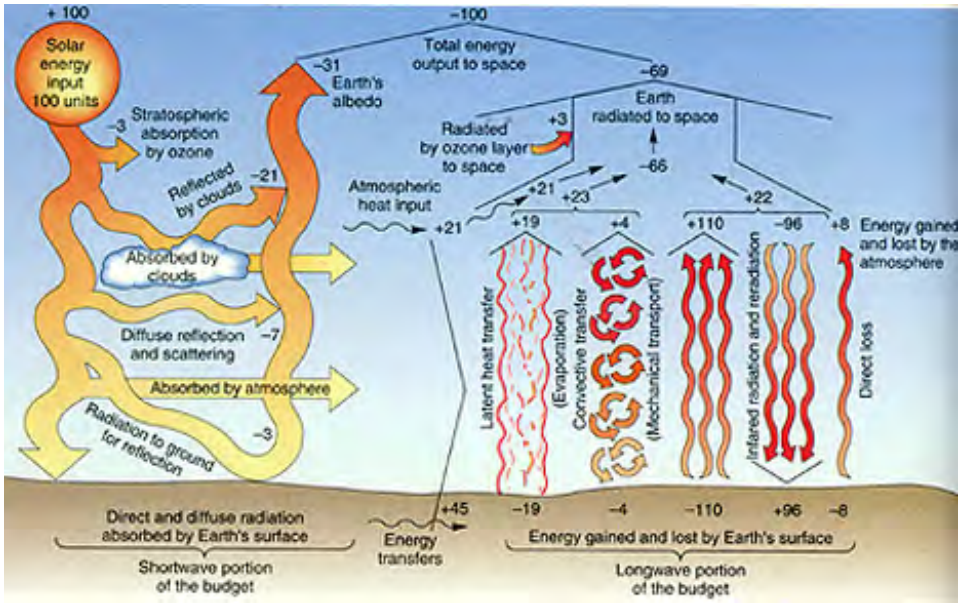
16-20: As was noted in the answer to question 16-15, the daily load on EOSDIS will be the largest yet faced by any NASA or other space agency mission - on the order of 3 terabytes, Can you think of some way to ease or lighten that load? ANSWER

Further Comments

As we noted earlier, the basic impetus for the kinds of studies we described is to provide input to planners and policy makers. Having access to raw (or even processed) data is of little use to such groups. What they need is information for making decisions about allocating and, perhaps, regulating resources. This information answers questions, such as, "If we do not curb automobile pollution, what may be its effect on atmospheric composition and temperatures over the next 5-

10 years?" Or, "If we continue to deforest the Pacific northwest, what may be the effect on biodiversity, land cover, land use, and water run-off patterns?"

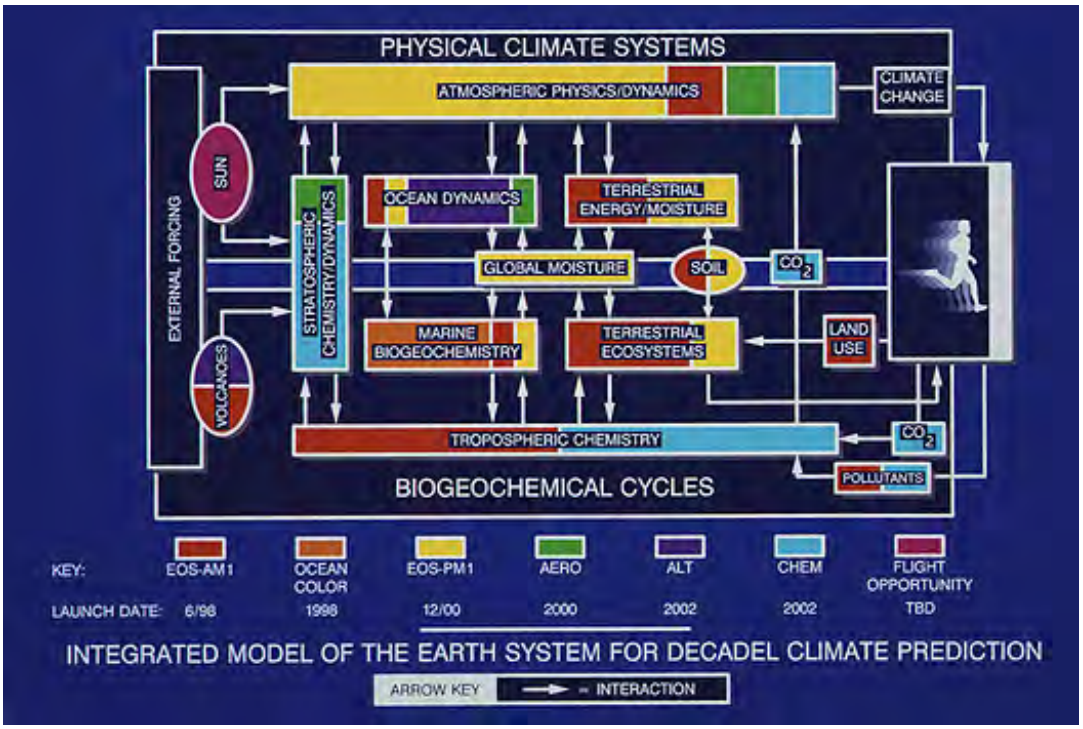
No one can actually predict the future accurately, but we can create mathematical models that generate useful predictive functions. Such models can be as simple as that shown in the next chart, which describes the distribution of incoming solar energy through major portions of the Earth's systems. It's similar to balancing the books: The sum of all components that use solar energy must equal the amount of incoming solar radiation. If they don't add up, our model is wrong, or we're missing something.



Christopherson, R.W., GEOSYSTEMS: An Introduction to Physical Geography, 2nd Ed. © 1994. Reproduced by permission of Prentice Hall, Upper Saddle River, New Jersey)

16-21: This diagram, while not constructed to explain or support the EOS program, does suggest that the various sensors in the EOS missions are all measuring aspects of the same phenomenon. What might that be? **ANSWER**

But the models we need for an Earth-system, predictive capability are much more complex than what we show here. Such models must account for energy budgets, sources and sinks of carbon and other biogeochemically related materials, effects of temperature on wind speed and direction, precipitation patterns, land use patterns, speed and direction of oceanic currents, increases in so-called greenhouse gases, and more. A full understanding of how these components interact with each other supports our willingness to trust the output from such models. We show an example of the complexity of these models, along with a color-coded indication of how EOS and other MTPE platforms contribute to a better understanding, in the figure below, which is based on the Bretherton diagram shown on page 16-4:



Because of the complexity of such models, EOS users must accumulate multitudes of timely data from many different sources, over a relatively long period, by a highly advanced data and information system that ingests, processes, and distributes those data to interested parties around the world.

Based on the predictive models described above, we must make decisions based on our understanding of the potential magnitude of global change, in order for planners and policy-makers to define strategies for mitigation or adaptation. These strategies may have widely different economic and societal impact, involving health, standard of living, and quality of life. EOS studies address how climate changes affect water resources, agriculture and ecosystems, and provide fundamental data sets on land-cover change, and measures of sea level change. These EOS measurements and the resulting predictive models address such topics as marine productivity, ozone depletion, air quality, and resources monitoring. These activities give us the tools we need to understand our Earth system and its many subsystems, and to understand the role we play in modifying such systems, and the roles they play in our daily lives.

We bring to your attention another related program that will rely on a variety of satellites and sensors, and the participation of many nations, to monitor and assess aspects of the environmental that relate to natural and manmade disasters. This program is an outgrowth of agreements worked out during a May 1996 Workshop on *International Cooperation in Space*. We summarize the results of their recommendation to use remote sensing for this purpose in the table shown here. Other satellites have since been added to the list, and EOS will contribute to the monitoring capabilities. (See also, the summary of satellite programs planned over the next ten years given in Section 21 [page 21-1]).

System	Status	Observation capabilities	Types of applications
Weather satellites	many existing	global day and night observations	prediction/monitoring of hurricanes, typhoons, tornadoes, volcanic eruptions
Landsat-7	operational	visual 30 m and multispectral 80 meter land observations	land use, flood extent, environmental monitoring
SPOT 1-4	operational	visual 10 to 30 m land observations	3 dimensional mapping, flood extent, damage assessment, crop identification

IRS-1C	operational	visual 6-30 m land and sea observations	3 dimensional mapping, oil spill detection, flood extent, damage assessment
RESURS-O1	operational	visual and multispectral 160-600 meter land and sea observations	regional environmental mapping, monitoring of coastal zones, crop development, drought, flood areas and fires
ERS	operational	all weather 25-500 m land and sea observations	3 dimensional mapping, oil spill detection, flood extent, damage assessment, night coverage
Radarsat	operational	all weather 10-100 m land and sea radar observations	3 dimensional mapping, oil spill detection, flood extent, damage assessment, night coverage
JERS	operational	all weather 18 m land and sea observations	3 dimensional mapping, oil spill detection, flood extent, damage assessment, night coverage
SeaWiFS	launched in 1997	multispectral 1 & 4 km sea observations	oil spill detection, ocean pollution monitoring, algae detection
Cosmos, KVR-1000	operational	visual 2 m land observations (not near-real-time)	high-resolution mapping, infrastructure identification, terrain analysis
Space Imaging IKONOS-2)	first launched in 1999	visual 1 m and multispectral 4 m land observations	high-resolution mapping, infrastructure identification, terrain analysis, crop identification
OrbView-2	launched in 1999	visual 1 & 2 m and multispectral 4 m land observations	high-resolution mapping, infrastructure identification, terrain analysis, crop identification
QuickBird-2	launched in 2001	visual 1 m land observations	high-resolution mapping, infrastructure identification, terrain analysis
SPOT 5A	launched in 2002	visual 5 m land observations	high-resolution mapping, infrastructure identification, terrain analysis, crop identification
OrbView-3	launched in 2003	visual <1 m and multispectral 4 m land observations	high-resolution mapping, infrastructure identification, terrain analysis, crop identification

Examples of images from several of the above systems are presented on page 21-2.

Now, with these eight pages of background, we can now actually see in the next two pages the first, and most impressive indeed, results from the flagship EOS satellite, Terra. Aqua, and other EOS satellites, are then considered.

PREVIOUS NEXT

Primary Contact: Nicholas M. Short, Sr.

The ESE era opened with the successful launch of Terra on December 18, 1999. After a period of degassing the spacecraft and instrument readiness tests, data began to be taken in February, 2000. All onboard worked perfectly. Terra is now the pivotal spacecraft to demonstrate the concept of making many kinds of measurements with multisensors at the same time, so that correlative data pertaining to a wide range of natural phenomena in the Earth's "spheres" will allow testing of the concept of an Earth System - a network of interrelated and interacting physicochemical and biological processes which tend to affect one another in feedback loops. This page looks at results from the MODIS and MISR sensors on Terra.

TERRA IS NOW OPERATIONAL:

MODIS and MISR

Terra, formerly EOS-AM-1, was successfully launched on December 18, 1999. Thus commences one of the most ambitious and important programs of the space era, in view of its goal to obtain a variety of data sets simultaneously as a means of conducting integrated environmental studies on a near-global scale. After several months of orbit adjustments, instrument degassing, and other preparatory actions, acquisition of preliminary data from the the ASTER, CERES, and MOPITT sensors began in early February with the last two, MODIS and MISR, activated on February 24, 2000.

An artist's sketch of Terra was included on page 16-7. Here, we show the actual spacecraft (3.5 m [11 ft] high) in its "clean room" facility before it was shipped to the Vandenburg (Calif.) Air Force Base Western Launch Facility for mounting on an Atlas IIAS rocket.

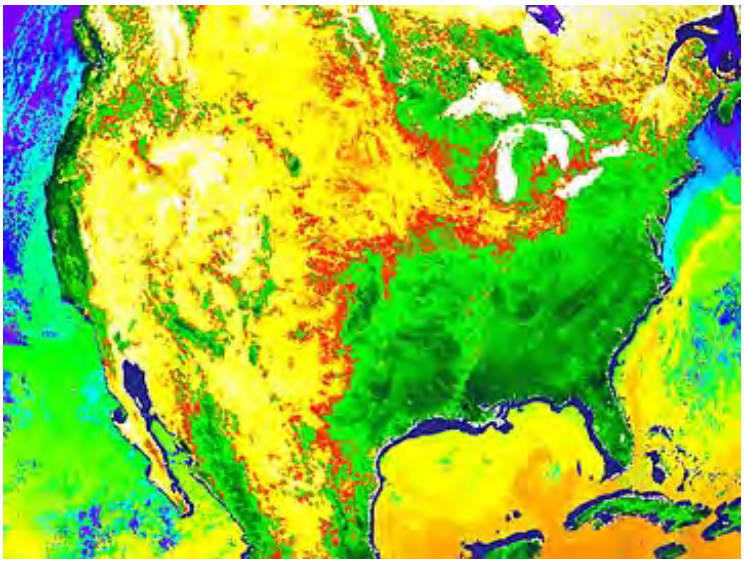
The images that follow are among the first obtained and have not been corrected - calibrations up to that point had not been completed. As new and more informative images are placed online on the Terra home page, a few of these will be added to this Tutorial page, perhaps replacing these now shown. Also, as the principal investigators gain familiarity with handling and interpreting Terra results, important findings will be reported on that Page. Each of the instruments described below have their own Home Pages, with many more illustrations and scientific results; specifically: MODIS (Moderate Imaging Spectroradiometer); MISR(Multi-angle Imaging Spectro-Radiometer); ASTER (Advanced Spaceborne Thermal Emission and Reflection [radiometer]); MOPITT (Measure Of Pollution in the Troposphere); CERES(Clouds and Earth's Radiant Energy System) .

MODIS and MISR

The first is part of a MODIS image set that includes the Mississippi Delta and the Gulf of Mexico. This is a natural color version.

By combining swath images from different successive orbits (close in time) much larger areas, such as this covering the U.S. and adjacent parts of North America, can be assembled into mosaics:

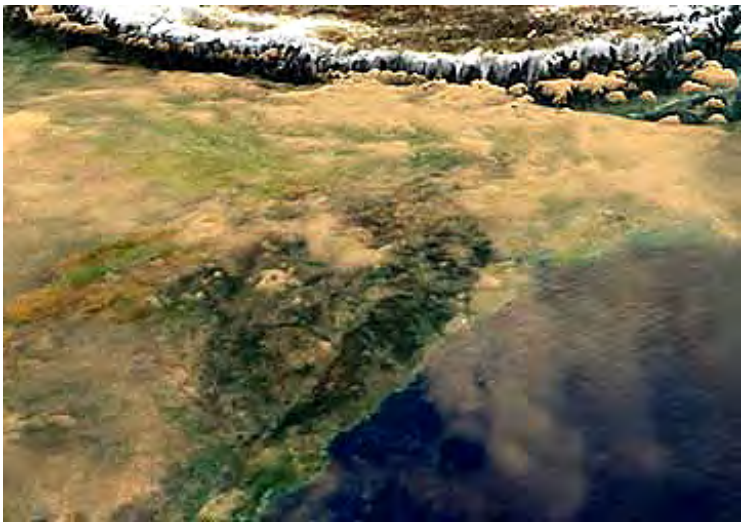
MODIS data allow calculation of the "NDVI" parameter (Normalized Difference Vegetation Index, obtained by dividing a Near IR minus Red radiance by a Near IR + Red value) and sea surface temperatures, as displayed here:



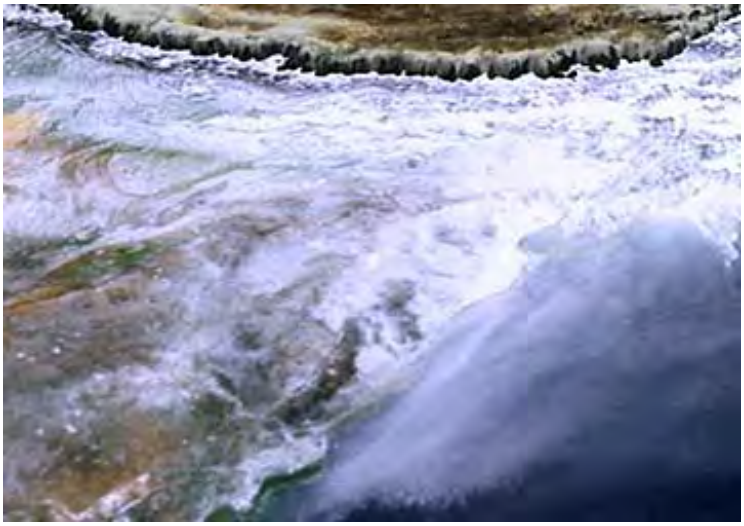
The color green has been assigned to land vegetation - in this case (in early April) for areas in the early stages of the green wave; yellows depict lower amounts of vegetation (low productivity). In the oceans, green and blue are cool whereas yellow and orange are warm.

One of the important EOS joint study programs is the Indian Ocean Experiment. The next image is a MODIS oblique view (constructed with the aid of elevation data) of part of the Indian subcontinent, with the Himalayas and Tibet Plateau in the background (top), shown in approximate natural color:

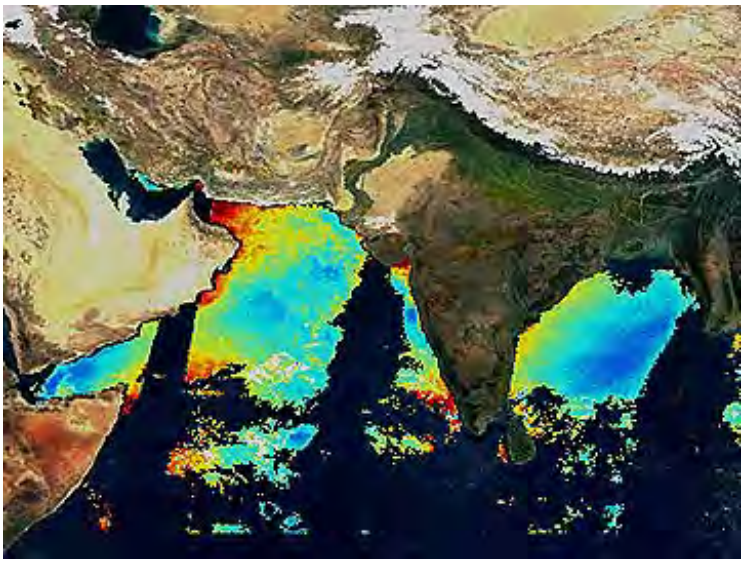
Other wavelengths on MODIS aid in picking up the atmospheric appearance of aerosols, much being the result of pollution:



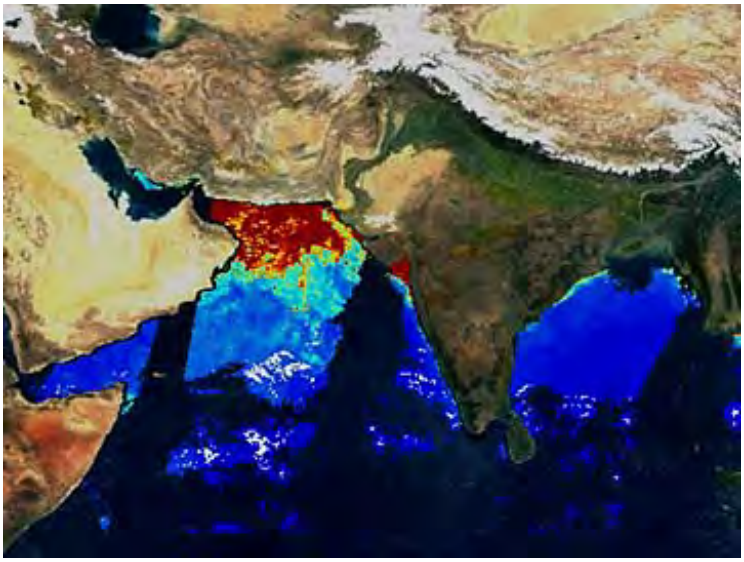
And, MODIS can pick out water vapor, even when heavy clouds are absent or dispersed, as shown in this image:



MODIS can also use its short wavelength bands to measure the fluorescent properties of the ocean, which relate to plankton content. Light not used in photosynthesis is re-emitted as fluorescence and heat, so that higher levels of fluorescence indicate lower photosynthetic activity. This is evidenced in this image (red is highest amount) that includes Africa, Saudi Arabia, Iran, India and Pakistan, and parts of the Indian Ocean and the Bay of Bengal:



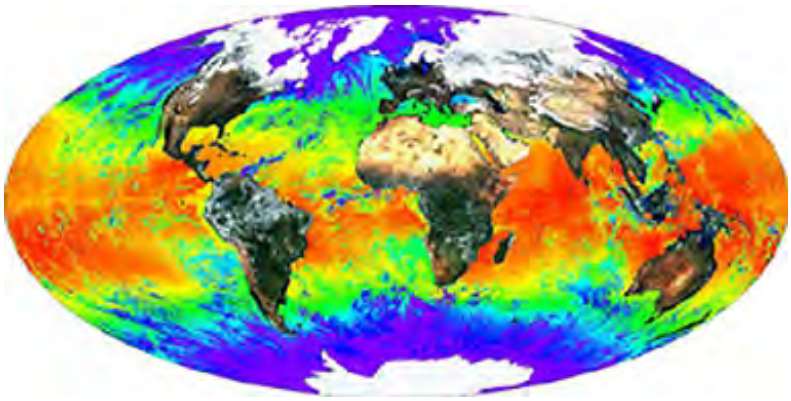
Other MODIS wavebands in the visible are suited to picking out chlorophyll content (similar approach as SeaWIFS, page 14-13), with reds being high and blues low concentrations.



Imagery from sequences of swath widths over short time spans can be combined for MODIS, and several of the other instruments on Terra, to produce images that seem to cover a terrestrial hemisphere, much as is seen by GOES and other geostationary satellites. Thus:

By integrating information collected over several days, global maps showing surface reflectances and sea surface temperature variations can be constructed, like the one below that covers a

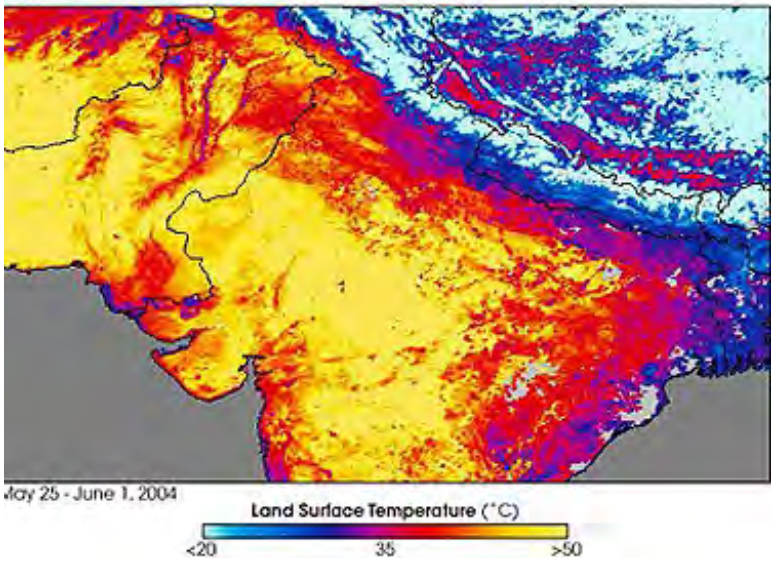
time period in late March. The white shades are mostly snow cover, but can be clouds.



These surface reflectances are suited to calculating NDVI values for the entire global land surface, as shown here:

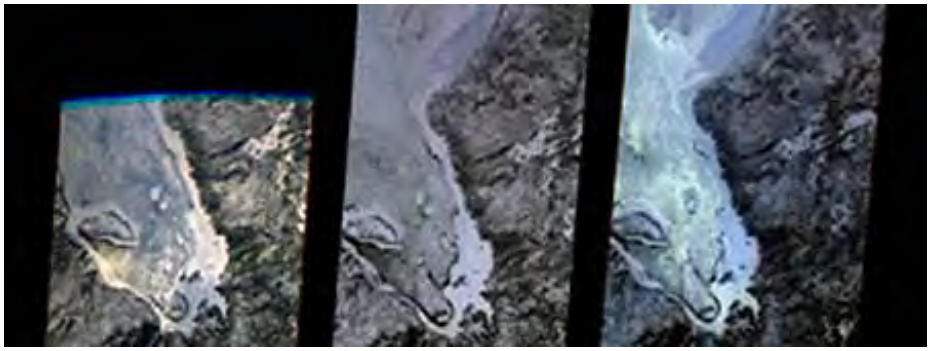
The versatile MODIS also has thermal bands (for surface temperatures from 3.6 to 4.1 µm). The image below shows the entire Iberian Peninsula (Spain and Portugal) during one of the hottest heat waves ever recorded in that region. On July 1, 2004, one locality reached a ground (not air) temperature of 59°C (138°F). The only cool area in the scene are the Pyrennees Mountain chain dividing Spain from France (in greenish-blue).

India is frequently subjected to intense heat waves just before or during the summer monsoonal rains. In 2005, such a temperature extreme (up to 115 °F) persisted from several weeks, leading to hundreds of deaths. Here is a MODIS image of temperature distribution:

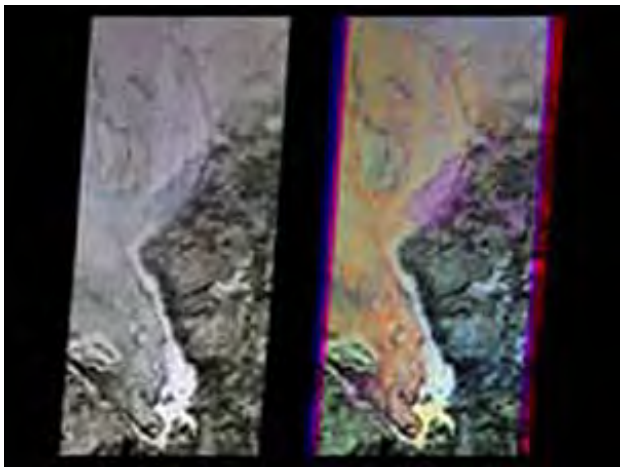


Turning to the next Terra instrument, the MISR spectroradiometer produces multiangle images over short time intervals. There are 9 separate cameras that look out at different angles. Swath widths are 400 km (250 miles) and resolution in a vertical (straight down) image is 275 m (894 feet). Other resolutions, commandable from the ground, are 550 and 1100 meters.

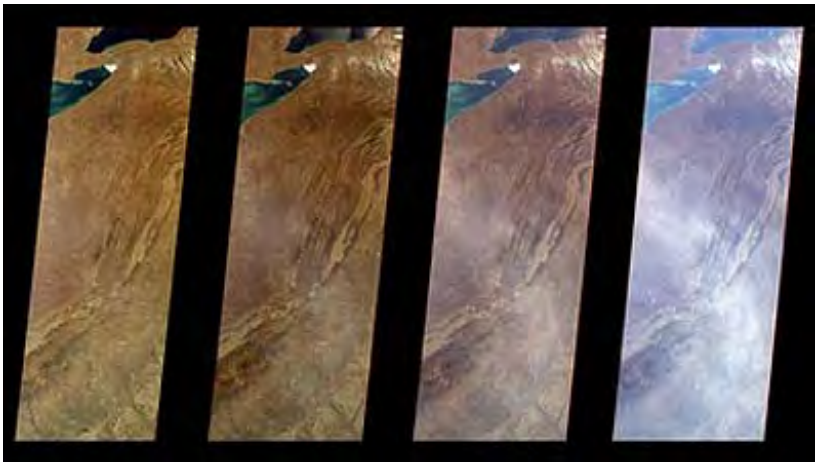
Thus, slightly different images result when the camera array captures a scene over a range of angles . This is exemplified by these three views of James Bay, off the southern Hudson Bay in Canada. This example actually shows the first scene taken and processed when Terra's instruments started gathering data. The left view is a forward image; the center is a nadir image; and the right is from an aft camera. This scene size took 7 minutes to acquire by scanning as the spacecraft moved south.



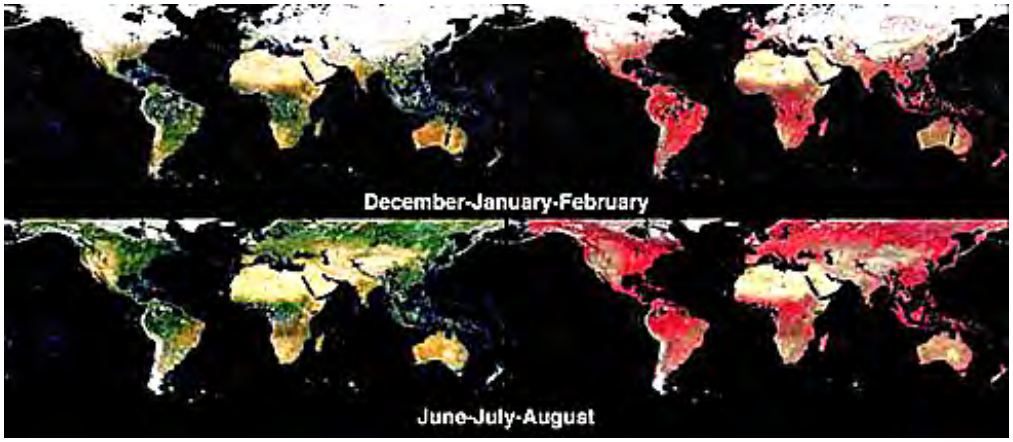
An interesting "trick" can be done with three images from cameras looking at different angles: They can be combined as a color composite. The image pair below shows on the left a single nadir image of the James Bay just examined. In the right display a forward camera view is projected through blue, the nadir image through green, and an aft image through red. Different resulting colors separate snow and ice (smooth is blue; rough is orange) from clouds (purple).



To illustrate the benefit of looking simultaneously at an area from multiple angles, consider these four views of part of the Appalachians; because of the more oblique look angle the image farthest to the right shows an aerosol haze that is absent from the nadir and close-in images.



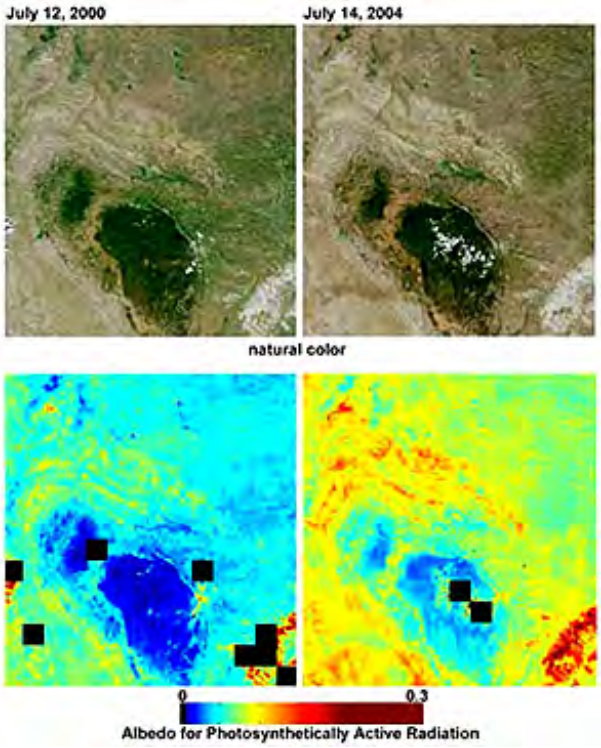
One of the capabilities of MISR is to measure albedos (degree of total reflectance over wide areas. One such measure is the Directional Hemispherical Reflectance (DHR). The four panels below show an average DHR over the non-polar region of the Earth for the time periods indicated:



The pair on the left are rendered in true color; on the right in standard false color. Note that the southern hemisphere images show much less of the characteristic seasonal signatures noted in the northern hemisphere. Thus, most of the vegetation in southern Africa, South America, and Australia does not reach the stages of dormancy as in continents of the northern hemisphere. In part, this is due to the reduced areas of climate-affecting landmasses in the austral hemisphere.

MISR is well-adapted to determining the extent of drought conditions by measuring over time the albedo changes of vegetation. Much of the western half of the United States has been in a

worsening drought since the 1990s. Below are two MISR natural color images of the Black Hills and surrounding plains that are largely grass covered. The rise in albedo from 2000 to 2004 indicates higher reflectance in these lands as vegetation is diminished and turns into more reflective brown shades:

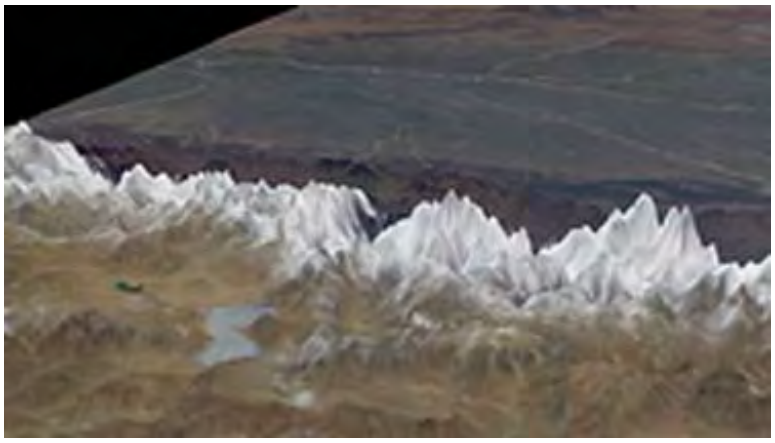


MISR images can be processed and joined to make nearly "seamless" mosaics. Here, 600 images taken during summer passes are combined to make a natural color mosaic of the United States (and part of Canada and Mexico, and Cuba and the Bahamas). This clearly indicates where most of the vegetation occurs as the dominant surface cover. Compare this with U.S. mosaics appearing elsewhere in the Tutorial (e.g., Section 7).

This next MISR scene shows a vertical (nadir) view of the coastline of northern Australia, not far from the Pilbara district we looked at in Section 6, page 6-15. This true color image, using the

blue, green, and red bands, shows the Joseph Bonaparte Gulf, the manmade inland Lake Argyle, and the Ord River. Although this part of Australia is tropical in vegetation, the effects of the underlying complex geologic structure is expressed topographically.

By combining a MISR image with (digitized) topographic map elevations, views that are perspective can be derived. This image looking south shows northern India, the Himalayan mountains, and the Tibetan Plateau in the foreground (bottom of image).



On the next page we will examine images from the remaining three Terra Instruments.

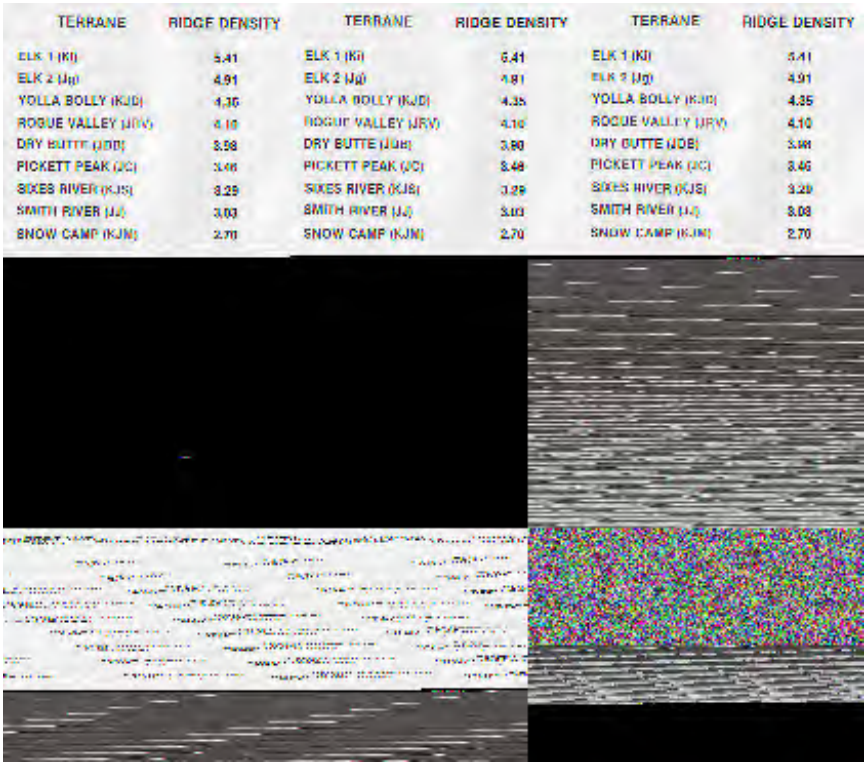
PREVIOUS NEXT

Primary Contact: Nicholas M. Short, Sr.

Both ridge densities and azimuthal orientations of the two most prominent ridge trends showed notable differences among the terranes examined, but this was not subjected to rigorous statistical testing. A causative model for the differences in trends is proposed. The Klamath terrane study is summarized as follows: 1) there are real and meaningful differences in some geomorphic parameters when these are associated with individual terranes - this confirms the hypothesis that such differences should in fact be expected owing to the natural variations in rock type/structure in diverse terranes that come from different parts of the mobile plate system and arrive at different times; 2) Landsat could distinguish some of the Klamath terranes simply by visual inspection; and 3) stereo achievable with SPOT imagery allows certain measurements otherwise obtainable only from topographic maps to be made.

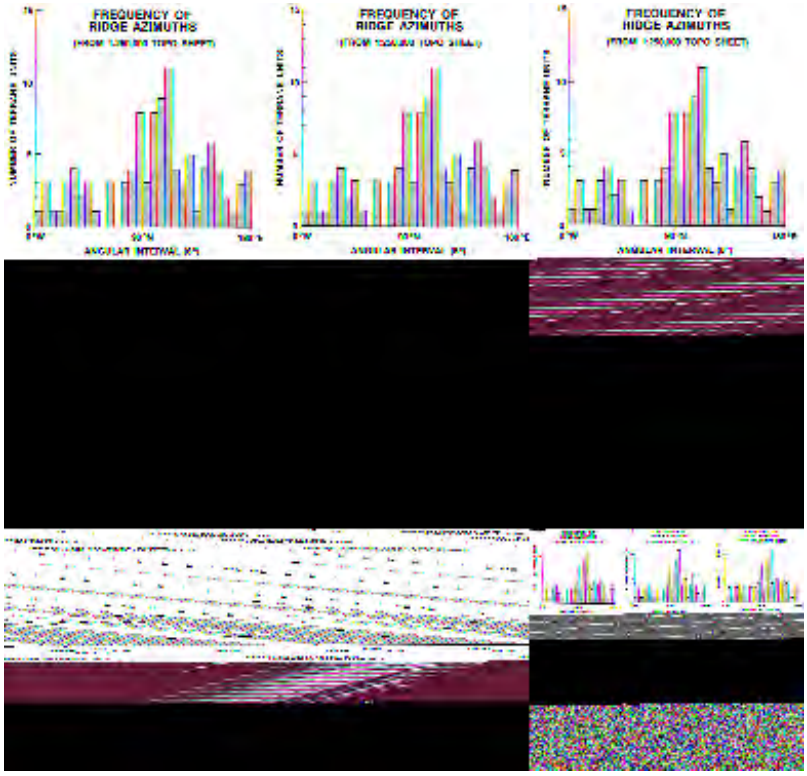
Ridges as Indicators of Terrane Differences

Earlier, we introduced the idea of ridge density (defined as total length of ridges per unit area). For the study, I produced a regional map of ridges from the 1:250,000 topographic sheet, from which I calculated densities using a computer program. This list shows that densities vary by a factor of almost two among the major terranes in the Klamaths:

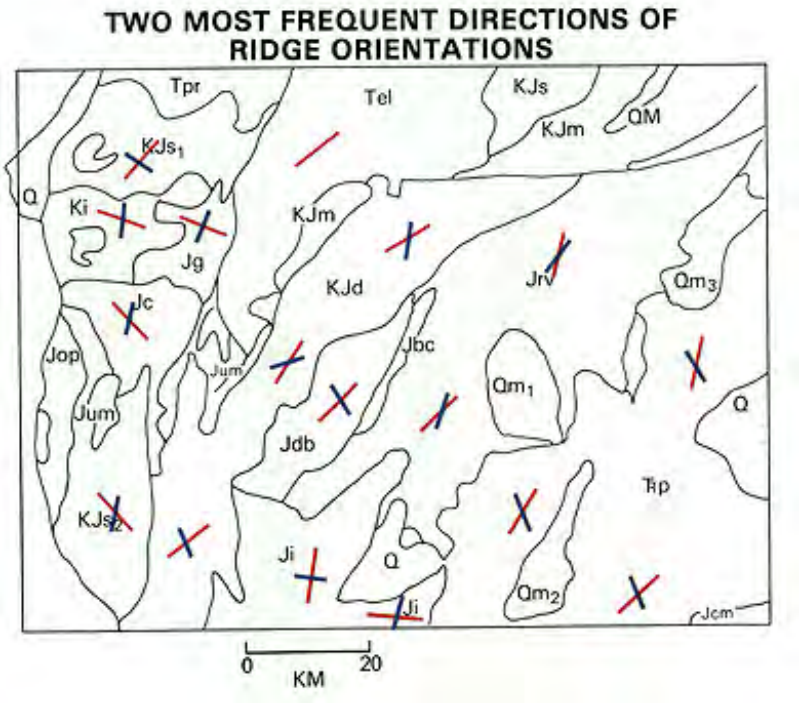


This density value relates to several factors, chief being the "fineness" of the drainage network.

From this map, I plotted the orientations of ridges (normalized to lengths) for the entire region:

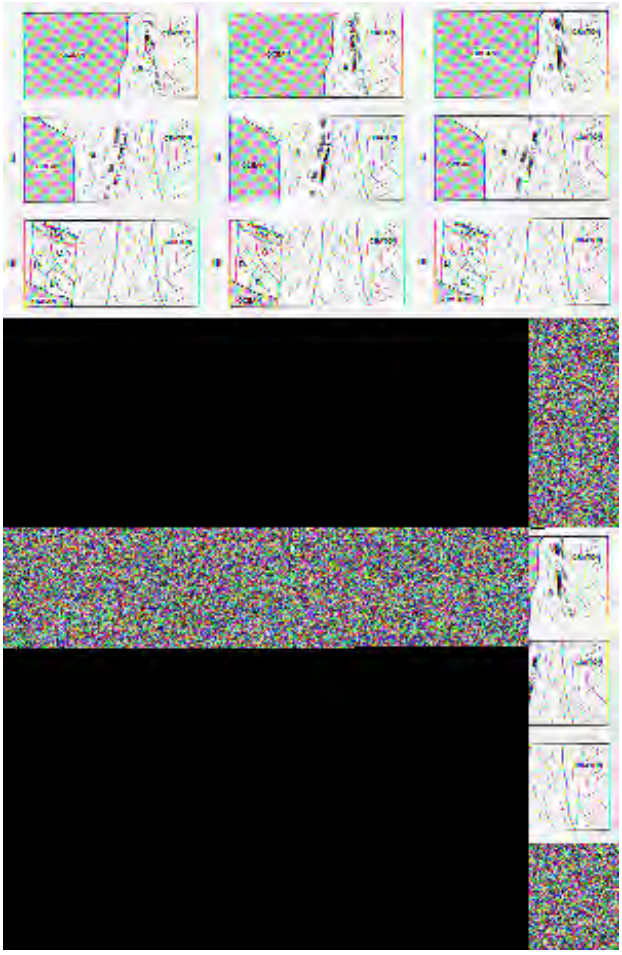


Then, using the above-mentioned program, in which the ridge data are entered as points on a digital plotting board that converts them to lines, their average trend directions were then calculated. These orientations of the ridges for any chosen area (such as a terrane or fraction thereof) were then plotted in rose diagrams similar to those shown above for the Elk and Sixes Rivers terranes. Below is a map, on which is drawn the most common orientation interval (in 5° units) in red and the second most common in blue.



The prevailing direction for terranes including and east of the Yolla Bolly terrane is north-north-east, in keeping with the regional plot just shown. But, in the Elk and Sixes River terranes that direction becomes northwest.

What might account for this shift in trends? This next diagram is my speculation:



In upper panel I, a new terrane smashes into the continental craton along a subduction line of a given geometry (including orientation). As the terrane slides under (or, it can override the continent), it takes on a fabric whose principal direction(s) is governed by the stress field imposed during emplacement. Now, in panel II, the subduction zone location and geometry have changed and a somewhat different fabric orientation results. In panel III, a new subduction zone (bounded by a transform fault [not shown]) adopts a notably shifted orientation and this passes on to the docking terrane. Then, as erosive processes gradually carve out topographic expressions for the terrane group, streams, whose paths tend to be controlled by the underlying structural fabric, produce valleys and intervening divides that reveal the effects of the stress fields that gave rise to the noted preferred orientations.

17-22: Before you go on to the (next) summary statement, from what you have read what seems to be the best parameter(s) for separating terranes by geomorphic properties? ANSWER

Summary of the Klamath Terrane Project

The foregoing report on the writer's research into terrane/terrain interrelations is, as stated, incomplete. But, it warrants several conclusions:

I. Real differences in geomorphic expression exist among the terranes. We can separate the Klamath terranes best by ridge (and stream) density differences.
II. Variations in hypsometric interval (curve) and related parameters among terranes may indicate relative resistance to erosion. That resistance depends on rock types and locations within the Klamath "dome".
III. Principal directions of ridge orientations seemingly differ in individual (or clusters of) terranes. This is probably structurally controlled and could relate to subduction/accretion geometry.
IV. The principal value of space imagery is that it provides a visual overview of differences in landforms/topography expression. It helps to pinpoint areas for quantitative analysis by

conventional morphometric methods.

V. We require stereo imagery for elevation-based morphometry and structural pattern analysis. SPOT-type stereo imagery reduces the illumination bias.

17-23: What, in your opinion, is (currently) the principal value of space imagery in geomorphic analysis of accreted terranes (and perhaps landforms in general)?
ANSWER

The bottom line: In this study, most of the useful discriminatory data were extracted from topographic maps and, to a lesser extent, some aerial photographs. Satellite remote Sensing has a limited but positive role in the analysis of the Klamaths in particular and similar landforms in general, by virtue of its excellence in creating synoptic images that highlight regional relationships that we can interpret visually. This role will increase significantly as high resolution stereo becomes commonplace from space sensors. Ultimately, this imagery will lead to the topological expression of surfaces in a quantitative mode that will prove its worth again, as it has from aerial photography, as a prime approach to geomorphic characterization.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

In much of the preceding 16 sections, the impression is gained that remote sensing is directed towards applications. But, in truth, sensors detecting electromagnetic radiation have been one of the dominant data acquisition and analysis devices in a wide range of (basic) scientific research. Witness the vital role remote sensing is given in observing and measuring the planets in our Solar System, and even more so, in all other things astronomical. This will be addressed in detail in Sections 19 and 20.

When the writer (NMS), a geologist, first started working with Landsat imagery, I was struck by the synoptic overview that the images gave of the Earth's natural landscapes. I realized then that a whole new branch of Geomorphology - the study of landforms - which he termed "MegaGeomorphology" could be developed from space imagery. This led to a symposium on this new subject, followed by the NASA publication of a beautiful and informative book, "Geomorphology from Space" (co-authored with Robert Blair, Jr., who joined NMS as co-editor during a year's sabbatical at Goddard). The first half of this Section attempts to outline the advantages of Mega- or "Regional" Geomorphology, illustrated by examples taken from the book. Because these landform and geologic process images are so effective at the scales afforded by satellite imagery, many examples will be included in the next five pages. The second half of Section 16 reports the principal findings of a practical study by NMS of the geomorphology of the Klamath Mountains in southwest Oregon, in terms of their terrane accretion origin, in which the usefulness of having space imagery to supplement interpretation was assessed.

BASIC SCIENCE STUDIES 1: GEOMORPHOLOGY

Mega-Geomorphology Defined and Geomorphic Maps

The primary thrust of the Tutorial to this point has dwelled on applications - the practical use of remote sensing to oversee conditions or solve problems dealing with the Earth's environments and manmade infrastructures. The emphasis has therefore been on the status of land and water surfaces, and on the overlying atmosphere, insofar as these affect human activities. The payoff from this monitoring is a better understanding of and interaction with (usually changing) events which influence our lives - food production, deforestation, urban expansion, mineral extraction, movement and severity of major storms, flooding, ocean nutrient detection, industrial site selection, to name a few - occurring at both regional and local scales that are readily observable from space.

We don't classify most of the applications, such as those above involving use of remote sensing, under a heading of "Basic Science" studies, although principles of science, such as the physics of electromagnetic radiation, are involved. Purists may argue that information about space "observables," such as determining ozone distribution worldwide, measuring heat flow in the oceans, and observing the biomes and their seasonal variation over the continents, augments the subject matter normally associated with basic science.

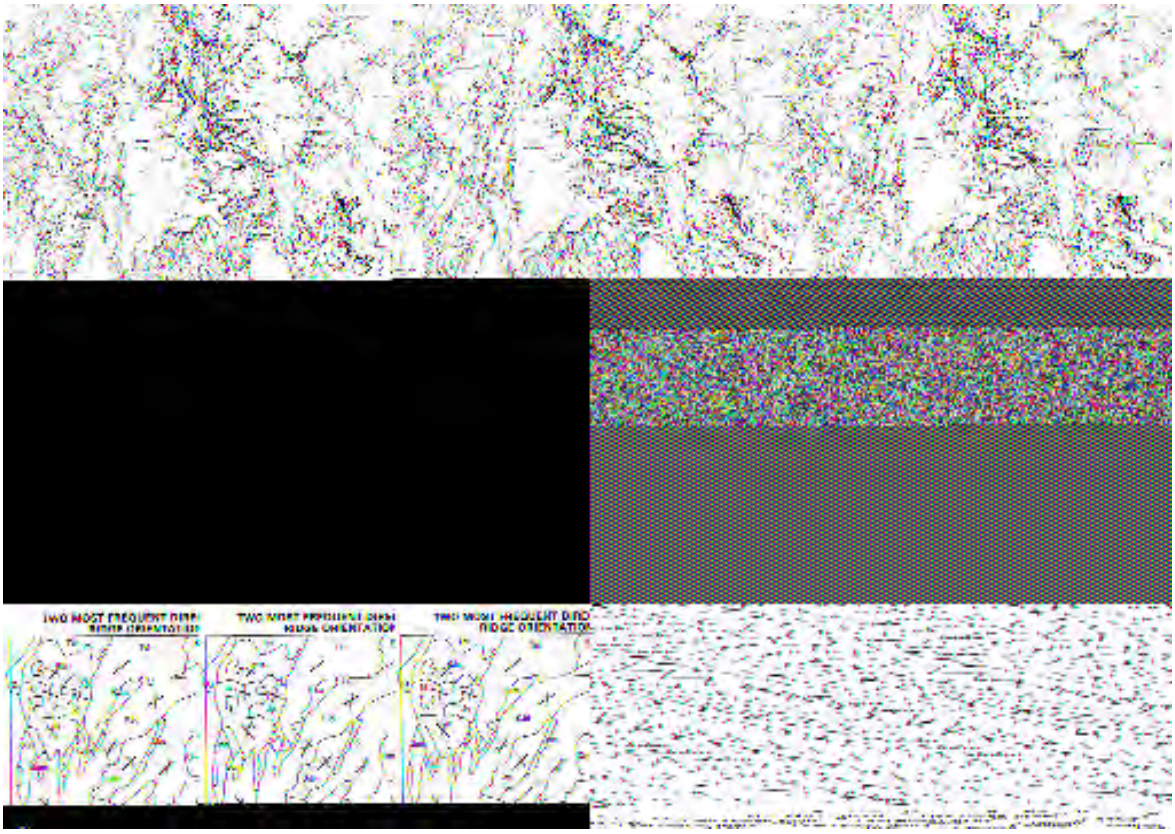
Perhaps the best examples of contributions of remote sensing to the fundamentals of science are those associated with exploring planets and astronomy. Similarly, some observations of a geological nature fall more into this "science" niche than into some applications discipline. In this

Section, we illustrate that idea by surveying how remote sensing has been a large factor in creating a new subfield in the science of Geology: Global Mega-Geomorphology.

17-1: In the fields of chemistry, physics, botany, and geography, list at least one example of how remote sensing contributes to what might be considered a " more scientific" area of study. ANSWER

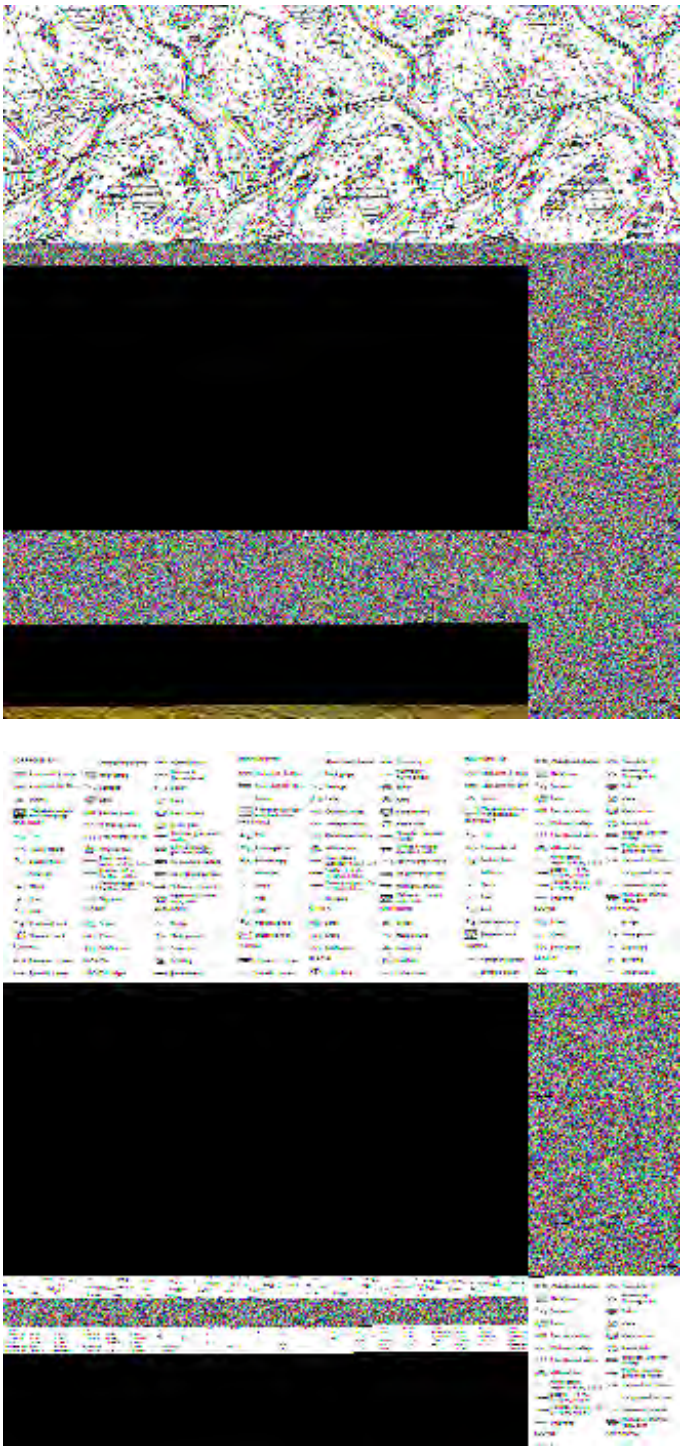
Geomorphology is the study of the shape or configuration of the Earth's (or, by extrapolation, any other planetary body's) solid surface, above and below ocean level, involving the classification of landforms and the processes by which they develop. The term "physiography" relates closely, referring to the description and evolution of land forms. By implication, water bodies, from oceans to small creeks, are participants in geomorphic processes, as major formative agents, even though their shapes tend to be transient and flat in detail.

Until the airplane and then the space age, scientists conducted most geomorphic analysis by on-the ground mapping; generally deriving the input information from topographic (for shape) and geologic (for structure and underlying rock types) maps and by direct field observations and measurements. The observation scales tended to be local or over relatively small areas. Eventually, aerial photographs became a prime tool for mapping and interpreting. The most common format depicted landforms as physiographic maps, using an artist's sketch technique to convey a three-dimensional oblique perspective (exemplified in the top image, below, for part of Utah), or, less often, as geomorphic unit maps

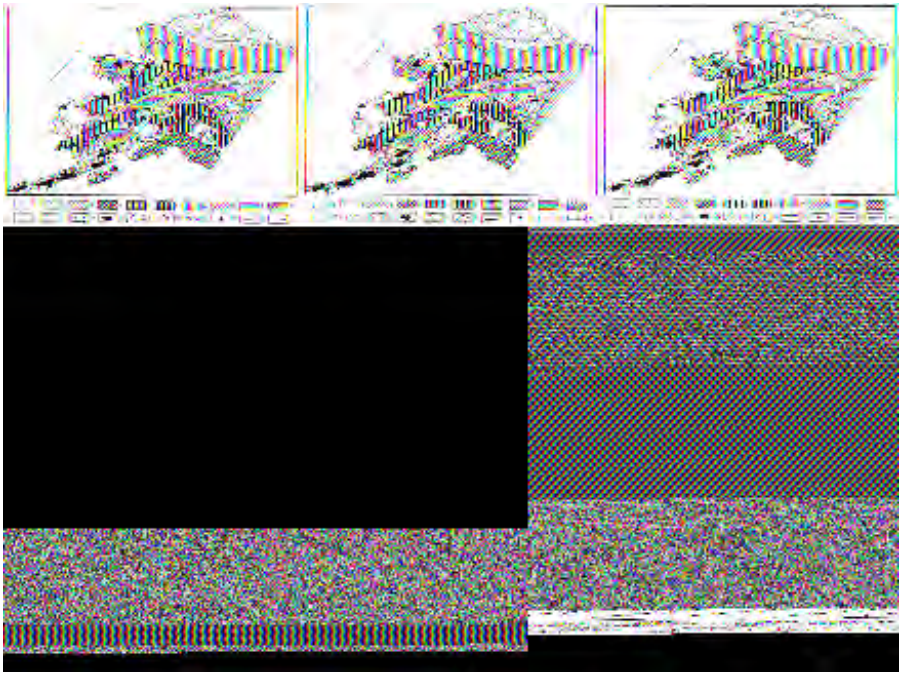


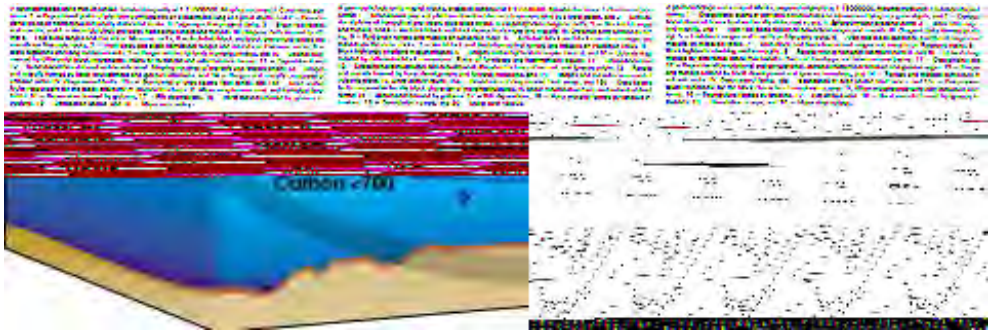
17-2: **Find the Waterpocket Fold (page 2-2) on the above sketch map; hint: it is near the Henry Mountains.** ANSWER

A variation of this is the Geomorphic Units map. Here is an example of one made for an area in Scotland:



Geomorphic maps covering large areas–regional in scale–have not been commonplace. In the 1980s, a consortium of geomorphologists produced multi-colored maps of Europe and North Africa. We reproduce below an example of a regional map, showing much of Alaska, made by Soviet scientists.





However, in general, geoscientists, especially geographers, did not produce studies and analyses of landforms at large scales, even after Landsat and SPOT products became widely available. Within a few years after launch, these satellite imagers had achieved nearly cloud-free global coverage (about 11,000 individual Landsat frames are needed to cover the continents) of the Earth's land surface (exceptions: the polar regions and some oceanic islands). The ability to combine Landsat or SPOT images into mosaics, such as the one of southern Asia, shown in Section 7, greatly aided the pictorial expression of regional landforms. In effect, we can represent anywhere in the world in picture rather than sketch format, so that we see exact landforms (generally the most obvious features at small scales) of truly regional dimensions.

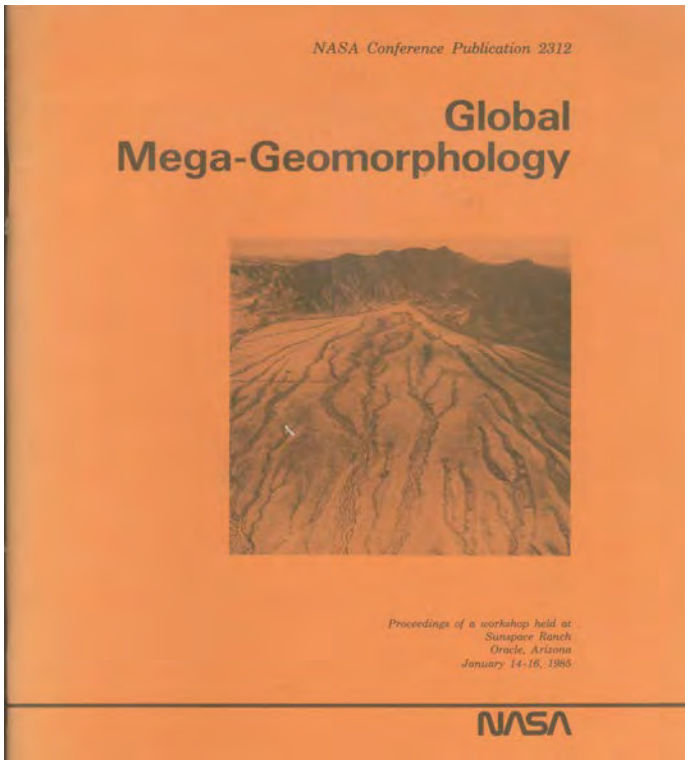
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

This page reviews the background (history) of the Workshop on Global Geomorphology, held near Tucson, AZ in 1985. From this came a commission from NASA Headquarters to develop and publish Geomorphology from Space. Instructions on how to obtain the CD-ROM of that now out-of-print book, or access it on the Web, are given. As part of the book, Robert Hayden prepared several geomorphic maps made exclusively from Landsat imagery, an example of which is shown.

Geomorphology from Space

During the mid-1980s, the writer (NMS) began research on space imagery for producing small-scale geomorphic maps and for investigating some specific geologic problems involving interrelations of landforms and tectonic processes. Aware of this, the Geology Discipline Leader at NASA Headquarters saw the potential benefits and decided to fund a conference on Regional Geomorphology. I was charged with organing it. Thus, on January 14-16, 1985, we held a *Workshop on Global Mega-Geomorphology* at Oracle, Arizona. Thirty international participants spent the first two days presenting papers and engaging in lively discussions and then met in four working groups on the third day to consider the role of space imagery in these thematic frameworks: 1) Global Geomorphology, 2) Evolution and Inheritance of Landforms, 3) Process Thresholds, and 4) Planetary Perspectives. This workshop (summarized in NASA Conference Publication 2312) laid a foundation for future research in a field whose name– Global Mega-Geomorphology, first proposed as a new concept in 1982–was certified by the attendees. The Proceeding from the Conference are recorded in NASA Conference Publication 2312:

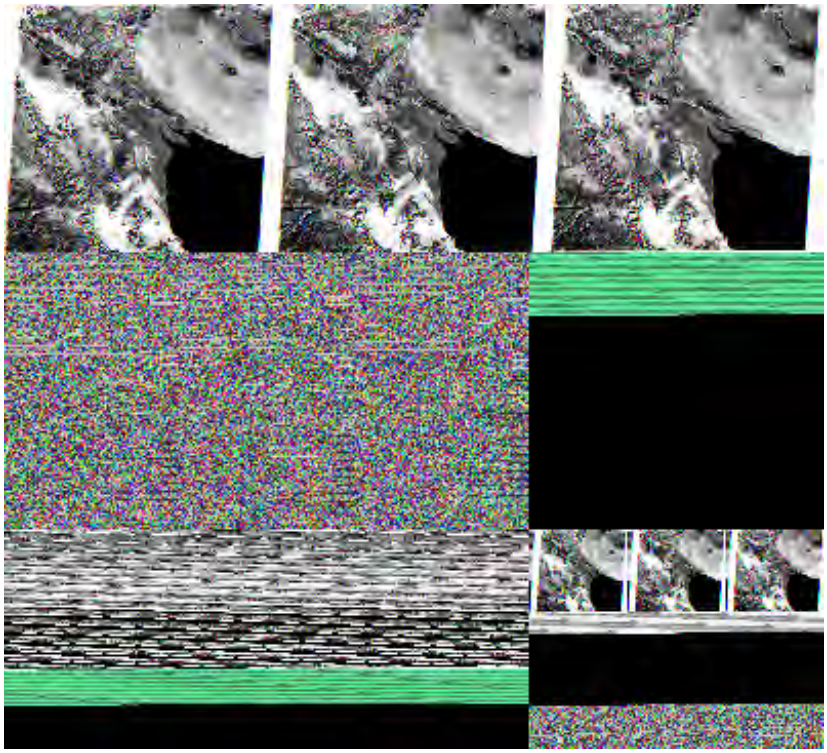


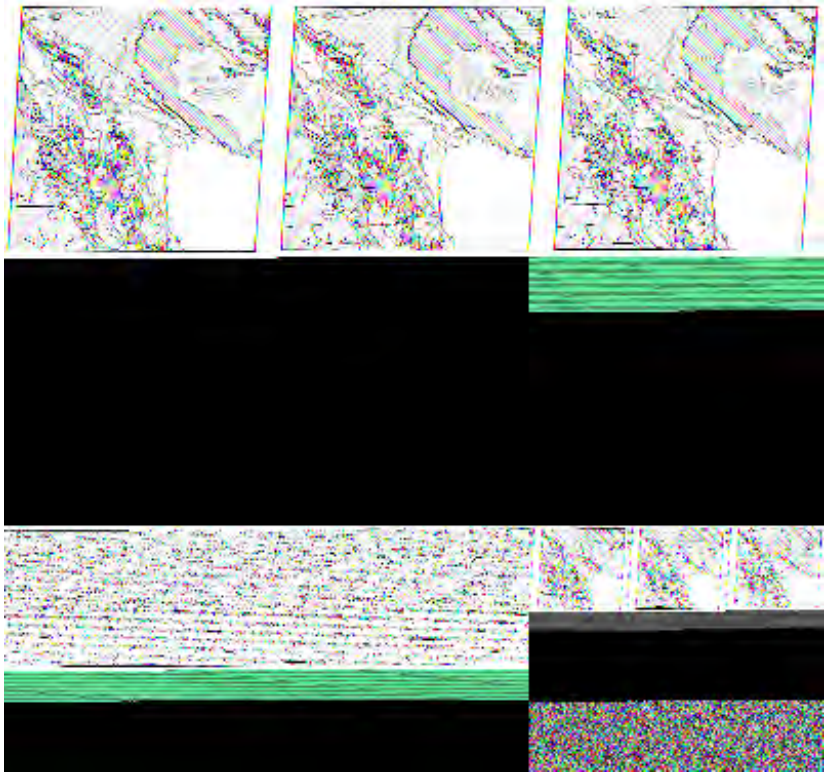
There was an almost immediate payoff. The writer (NMS) received NASA funding to prepare a publication that focused on using space images to characterize landforms. Joined by Robert Blair, Jr of Fort Lewis College, Durango, CO, as co-editor, we assembled fourteen scientists

(including ourselves) to write twelve chapters, each covering a different geomorphic theme, in a 717 page book entitled: *Geomorphology from Space: A Global Overview of Regional Landforms* (NASA SP-486), which appeared at the end of 1986. Although we distributed 4,000 copies, the book did not, at first, receive adequate publicity. In recent years, authorities have acknowledged it as an invaluable compendium of geomorphic information, as well as a comprehensive survey of the world's geology, and it has led to renewed interest in regional analysis.

Unfortunately, there was no second printing, so NASA and the U.S. Government Printing Office (GPO) cannot fill requests for copies (used copies are available from Amazon). But, NASA leaders recognized its value, so people in the educational program at the Jet Propulsion Laboratory (JPL) decided to add it to their growing set of sponsered CD-ROMs, available at low cost to professionals and the public (it is distributed through the Goddard DAAC; contact through daacuso@daac.gsfc.nasa.gov [this phone number may still lead to someone who can help: 301-614-5224]). A staff at NASA's Goddard Space Flight Center did much of the scanning and layout, with this version of the book now accessible on the Web at this Goddard website.

One of the serendipitous outcomes of this book is that Dr. Robert S. Hayden (at the time a Research Fellow at Goddard) was persuaded by the writer to prepare several geomorphic maps from full and partial Landsat images, as a demonstration of the possibilities of mapping at this scale from space imagery. We reproduce here a Landsat-2 image of the upper Gulf of California (mostly in Mexico) and the beginning of Baja California. Below it is his geomorphic map, with key:





"Geomorphology from Space" contains nine thematic chapters: Tectonic Landforms, Volcanic, Fluvial, Deltaic, Coastal, Karst/Lakes, Eolian, Glacial, and Planetary Landforms. These chapters contain many stunning and revelatory images, all open to inspection on the CD-ROM and the Internet. To whet your curiosity, we now present two or more representative examples for each theme (except Planetary, which we thoroughly treat in Section 19), along with a brief description of what can be extracted from the image regarding Geomorphology and general Geology.

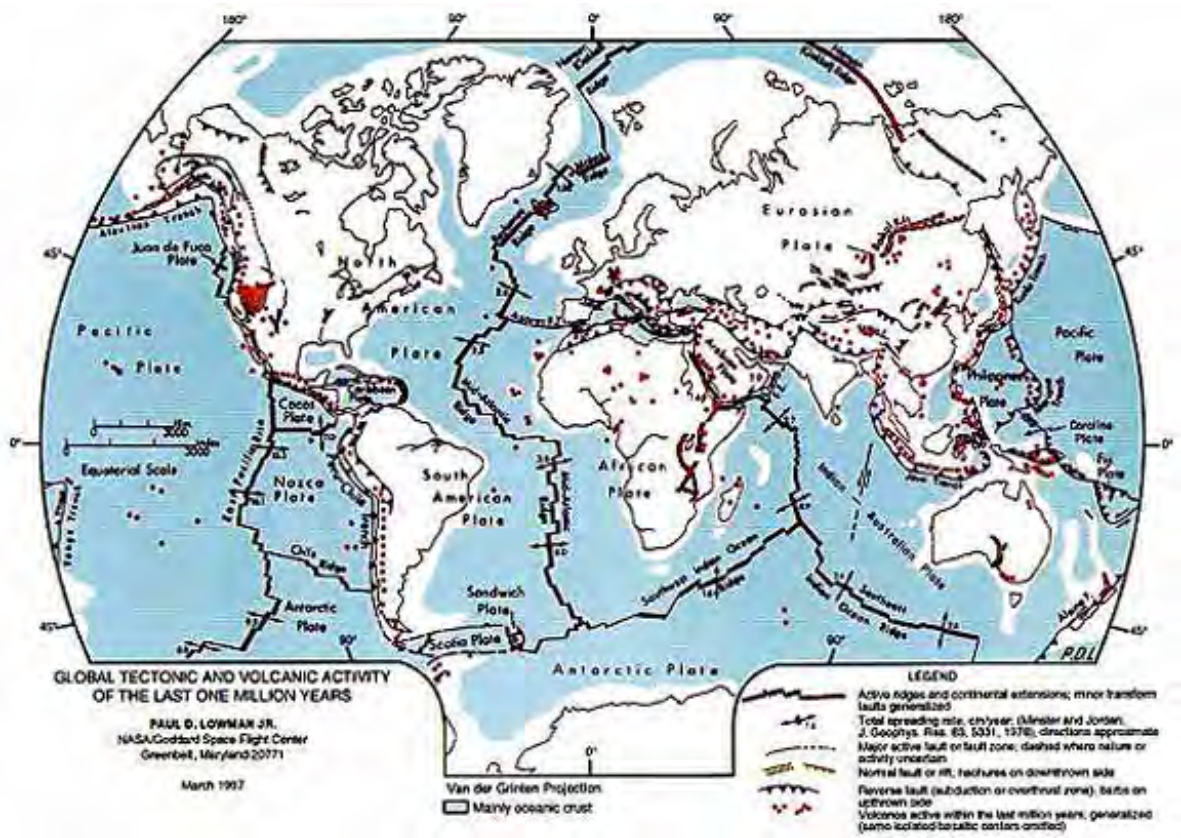
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Perhaps no other types of landforms so dominate the landscape at regional scales as do those formed from tectonic folding, regional faulting, and intrusion/extrusion of magmas. Space imagery, especially when mosaicked, is an exceptional way to portray the end results of the orogenic processes that produce mountains (always eroded to some degree to carve out the landforms that typify them) and other manifestations of tectonism. One type of mountain is the volcano - usually a stratocone or a bulging upwarp. Tectonic activity also leads to outflows of lavas that can cover vast areas and may result in ususual landscapes.

Tectonic/Volcanic Landforms

Tectonic landforms usually dominate the scenery in any region that has experienced significant crustal disturbances, and this activity often shows as truly spectacular expressions in remote sensing images. For this reason, the theme chapter by this title in "Geomorphology from Space" is by far the longest. These landforms frequently reveal surface manifestations of the type of underlying deformation caused by plate tectonic interactions. Some of these interactions characterize orogenic (mountain) belts at subduction zones (convergence of two or more plates) or pull-apart regions where plates diverge. For anyone unfamiliar with the first-order framework of the global tectonic system, examine this map produced by Paul D. Lowman, Jr. (author of Section 12) of the lithospheric plates, spreading ridges, transform faults, and other tectonic features. Consult any introductory Geology textbook for more information on the Plate Tectonic paradigm. Or, better yet, work through this Website: USGS Tutorial.



We described some exceptional examples (drawing upon mostly Landsat images) of tectonic

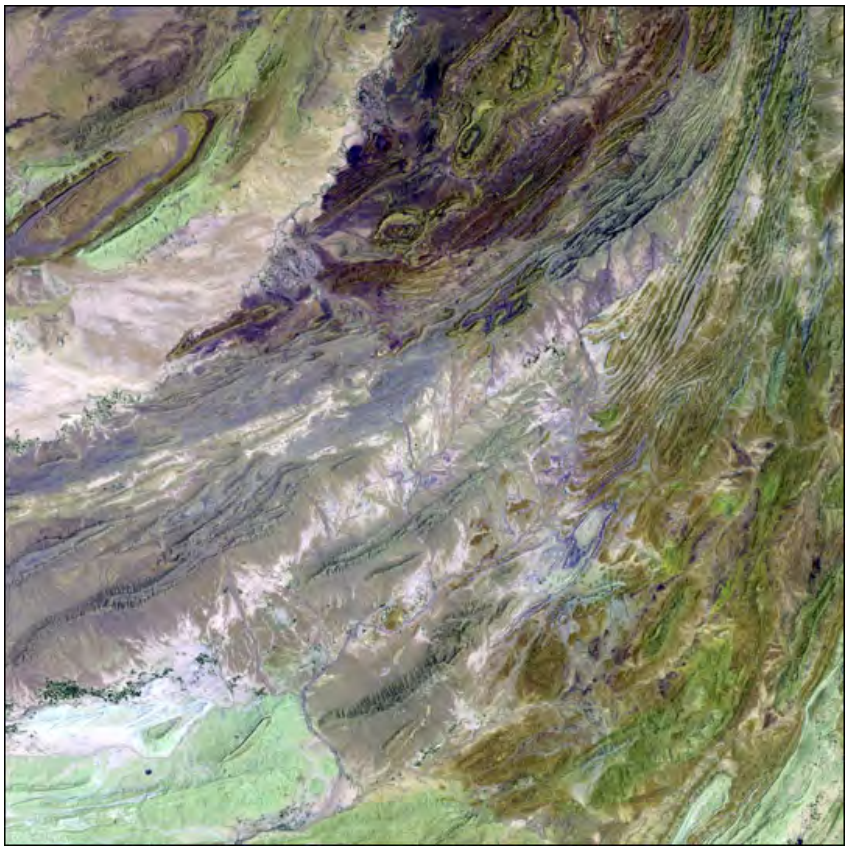
landforms in Sections 2, 6, and 7, which you can review (look particularly at the Zagros folds, the Pindus thrust belts, the Atlas Mountains, and the Altyn Tagh fault in Section 2 and the Appalachian folds and Basin and Range block fault mountains in Section 6, and the European Alps in Section 7.

These images focused on folds and faults, the most common types of tectonic deformation. The resulting landforms commonly have elevation differences (relief) that may be sufficient to change ecosystems developed at these heights. Thus, mountains in a semi-arid climate may be heavily vegetated (dark toned in visible band images) and adjacent basins less so (light), thus, showing strong contrasts in black and white images (the Nimbus 3 image of the Wyoming mountains in Section 14 is a good example). Mountainous terrains appear clearly in Landsat, HCMM, and radar images by virtue of shadowing, which causes tonal variations related to slope/sun positions.

One of the simplest tectonic landforms is the "hogback" - a ridge (straight to curved) made up of harder, more slowly eroding inclined rock units, that lags behind as the general surface level is worn down by erosion (as is found along the front of the Colorado Rocky Mountains; see Section 6). This aerial photo shows a typical hogback:

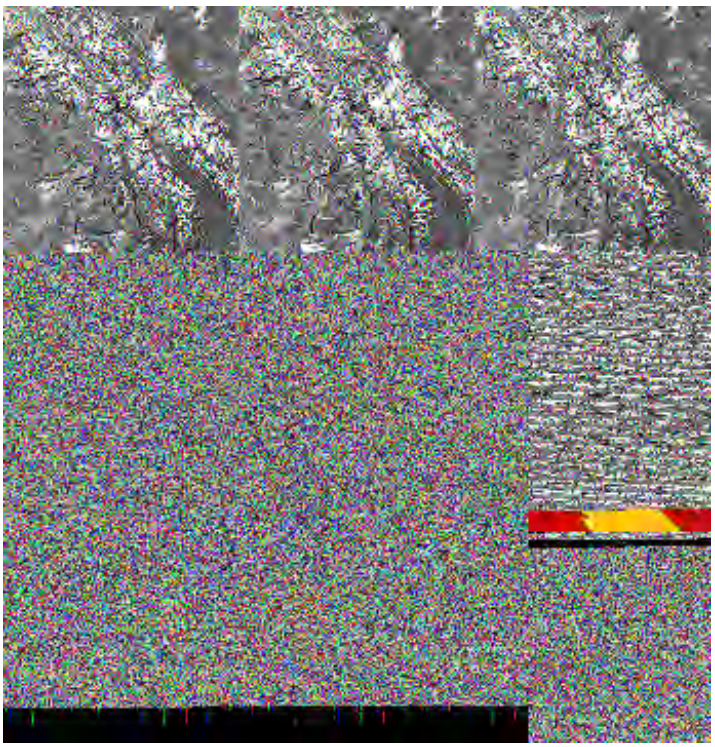
Here are a sequence of three parallel hogback ridges in Mauritania, imaged from space by SPOT-1:

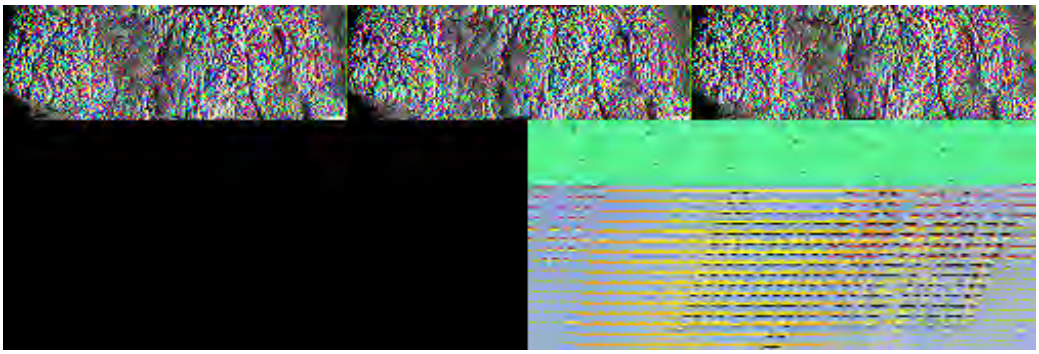
As you saw on page 2-6, space images in arid lands are particularly suited to showing folds - mainly anticlines - as they are expressed by the contortions revealed as erosion cuts into the inclined strata. We show five examples of this; the first four are SPOT images, the last is a Landsat image covering a larger area:



HCMM is especially suited to showing large segments of a mountain belt, providing a small-scale overview. Perhaps the most famous in the world, in terms of how its origin has been interpreted to lead to some earlier hypotheses on formation of orogenic belts, is the Appalachians. Examine the HCMM mid-Appalachians image found on page 6-3 (while there, scroll down to see the frontal hogbacks referred to above).

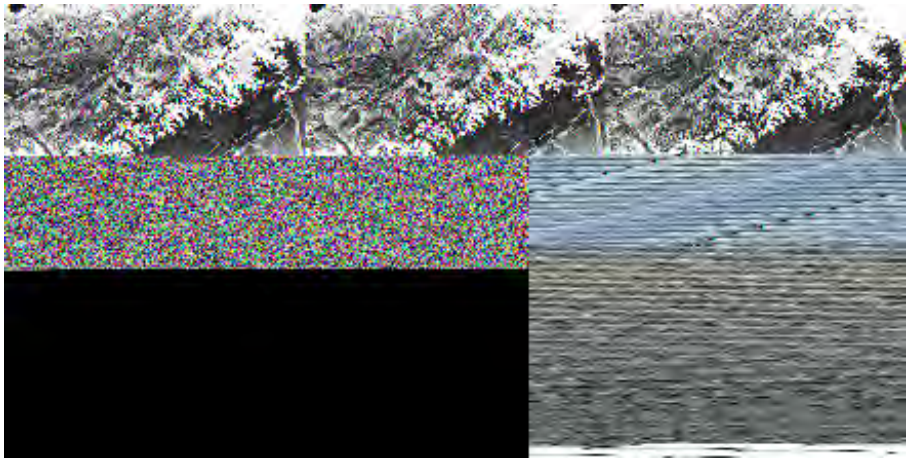
The Rocky Mountains in the U.S. were examined on page 6-6. They continue into Canada and in Alberta and British Columbia almost merge with the Coast Range and other Cordilleran mountain chains. Here is a Landsat mosaic that shows some of the Canadian Rockies. Below it is a strip across those Rockies made from Radarsat imagery.





And here is a aerial oblique view of typical mountain terrain in part of the Canadian Rockies; the broad valley has been widened by glaciation and backfilled with post-glacial deposits:

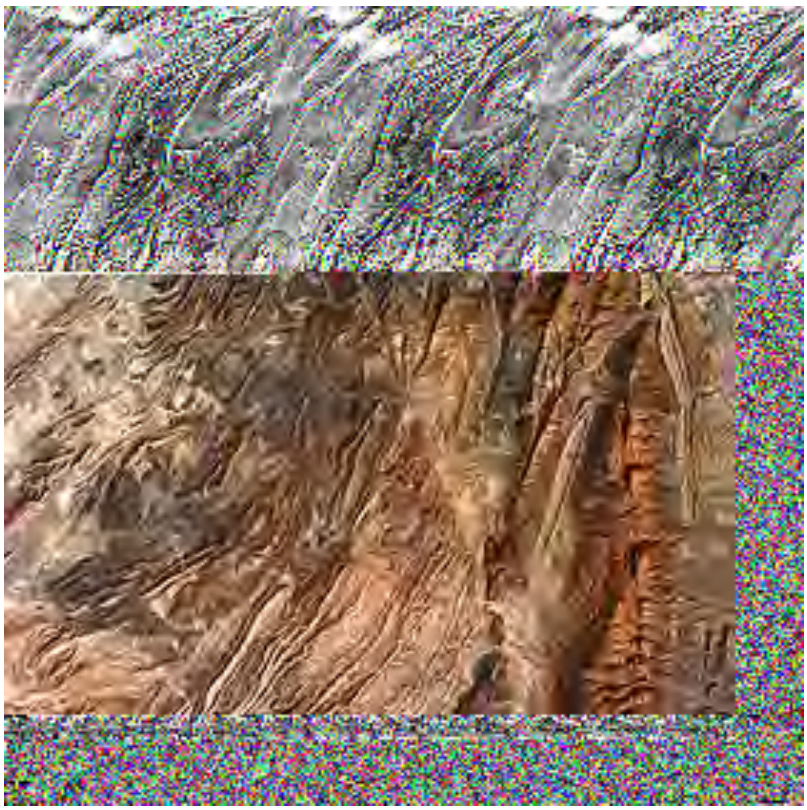
Recall the Landsat mosaic that showed much of the vast chains of interrelated mountain belts in southern Asia where the "crash" of the Indian subcontinent over the last 40 million years created the Himalayas on the north and folds in Pakistan and Iran in the west and others in Burma (Myomar) (see page 5-5) and Malaysia in the east. A spectacular oblique view of the main Himalayas, taken by an astronaut using a film camera, was shown on page 12-4. Here is another astronaut photo, made with the Large Format Camera, covering much of the same scene, including the Siwalik Hills (dark, near bottom right), the snow-covered main high Himalayas, and the southern Tibetan Plateau (on average, the highest generally flat landmass in the world).



To see more detail in the flanking mountains in western Pakistan, here is part of the fold belt that was shoved up by the huge collision between the Indian subcontinent and southern Asia (the context of this is evident in the mosaic examined earlier in Section 7). The scene shows the Sulaiman fold belt, consisting of echelon (offset) anticlines (some closed), making up the ridges (flat valleys occupy intervening synclines). The Kingri fault passes through the image center (look for an abrupt discontinuity). The crustal block to its west (left) has moved northward relative to the block on the east.

17-3: As a generalization, would you say that the "style" of deformation in the Anti-Atlas and Pakistan scenes is similar or dissimilar? ANSWER

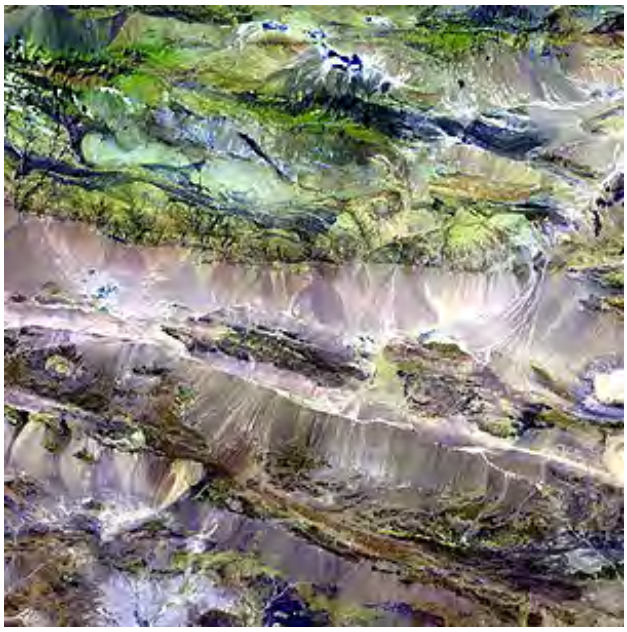
The tectonics of southern Asia is dominated by the Himalayan docking event. Subsidiary tectonic disturbances occur beyond the Himalayas. In central China is this scene (which includes the through-flowing Yangtze River) of what are known as *decollement* folds formed within thrust sheets (like wrinkles on a sliding rug).



Major strike-slip faults occur in the mountains of southern Asia, as land units moved laterally as a result of the Himalayan collision. This mosaic shows part of the 1500 km (1000 mile) Altyn Tagh (Altai) fault, which serves as a zone of crustal rock erosion that forms a valley:

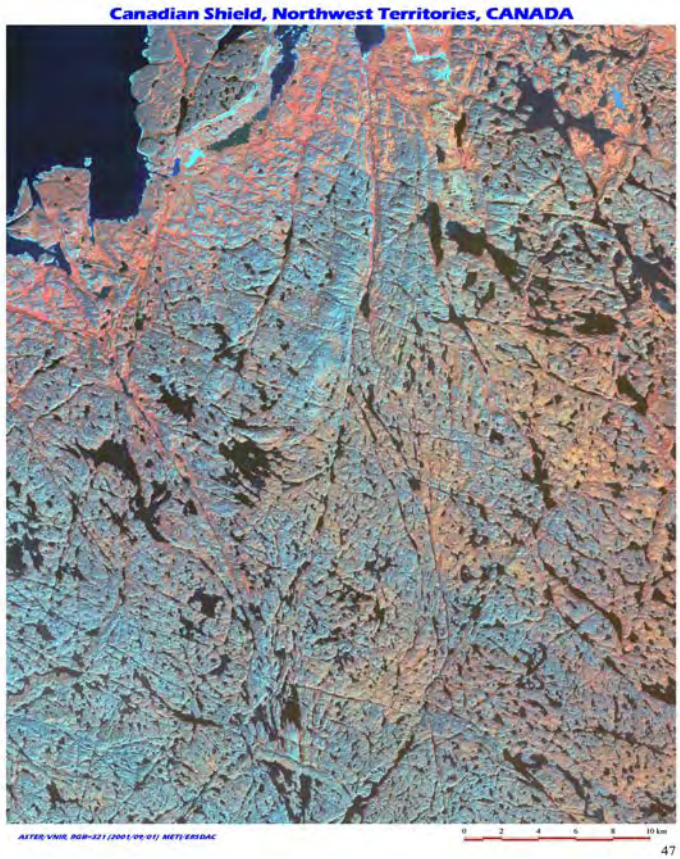
Similar slip zones - the Kunlun and Karakorum faults - in the Kunlun mountains also form valleys, as seen in this astronaut photo.

Western China, in and around Sinkiang Province, is mostly arid lands marked by deserts and mountains. Some of the mountain terrains are bounded by faults as in this Landsat image that includes part of southern Mongolia:



Crystalline (igneous and metamorphic) rocks make up distinctive terrains and landforms. Intrusions of granite, in particular, often accompany orogenies that produce extensive mountain chains. Granites are the most common rock type in batholiths that can form the core of a mountain system. In the United States the best known batholiths are in the Idaho Rocky Mountains and the California Sierra Nevada. Here is a Landsat image that shows the terrains of notable relief (emphasized by glaciation) in the High Sierra; beneath it is a ground photo that typifies the mountainous character of this terrain:

All continents have crystalline igneous-metamorphic rock masses that make up the Shields or Cratons (multiple intrusions and deep burial metamorphism over extended spans of geologic time) around which the continents have grown. In North America, the Canadian (subset Laurentian) Shield is the core around which the continent has grown by accretion and marine overlap of sedimentary rocks on this crystalline basin. In northern Canada, near Churchill, is a classic exposure of barren granitic/metamorphic shield rocks:



To the west of the Indian subcontinental plate is the Arabian tectonic plate, caught between the African, Eurasian, and Indian-Australian plates as they move in different directions. The western part of this plate is a crystalline shield (a continental nucleus containing ancient igneous and metamorphic rocks). Below is a mosaic (from 12 individual Landsat scenes) of the shield as exposed in southern Saudi Arabia and the Yemen Arab Republic.

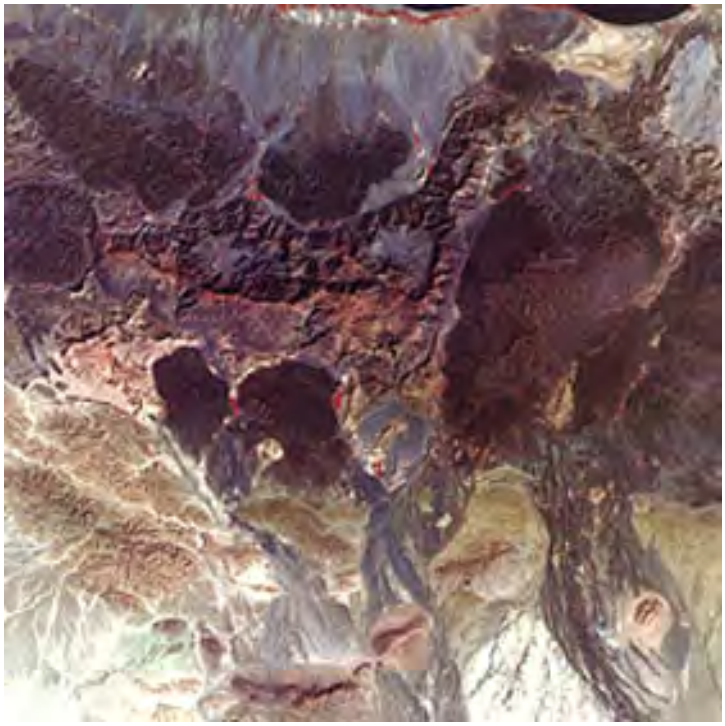
Dominant features in this scene are the numerous granitic intrusions, whose boundaries show as distorted oval shapes. The shield is a region of low mountains separated by valleys, many of which are sand-covered. A prominent escarpment (near the upper, left edge) bounds the western edge of the shield. The coastal plain is bounded by a fault-controlled scarp. Another scarp (lower right) also relates to the fault.

17-4: Broadly speaking, how does the tectonic "style" of this Arabian Shield scene differ from that of the previous two images? ANSWER

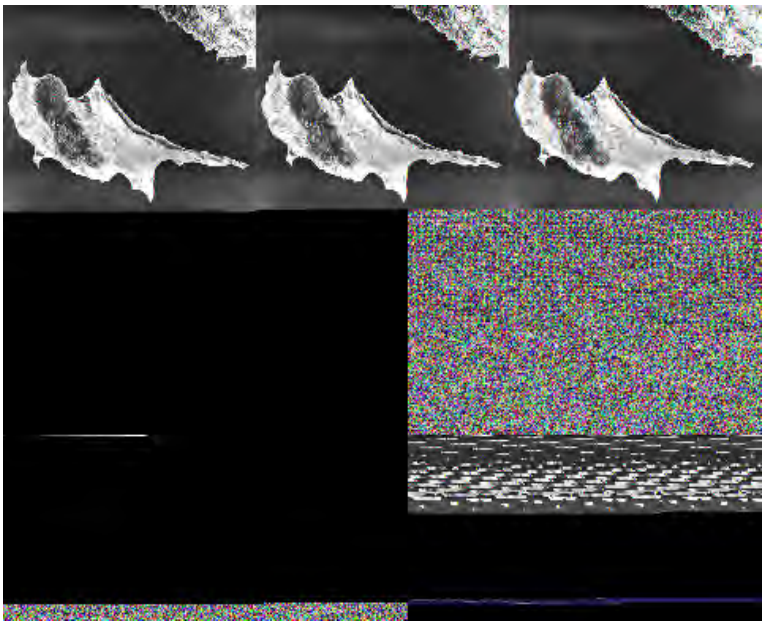
The Arabian Shield is part of a Precambrian craton that extends beyond Saudia Arabia. In Egypt it forms part of the Nubian Shield that is exposed on both sides of the Red Sea. The Southern Sinai Highlands is an offshoot that is found on both sides of the Gulf of Aden, seen here in an astronaut photo. A ground photo indicates how rugged the Sinai terrain can be; the rocks shown are granitic:

We've seen several other examples of shield terrane on this and other pages. Here is another, the Nigerian Shield (astride the Nigeria/Cameroon border). Like most shields, large criss-crossing fractures cut into the surface and the hills tend to be of low relief. The red indicates vegetation; the blue is sedimentary rocks with barren surfaces.

On the east side of the Arabian Peninsula, in Oman, are the Oman mountains, large parts of which are composed of ophiolites. These are ultramafic igneous rocks, first extruded as lavas with shallow intrusives below (peridotites; some gabbros), that made up new ocean floor that moved away from a spreading ridge. On contacting a continental mass at a subduction zone, the ophiolites may subduct but otherwise can also be thrust on (obducted) to the continental edge. In this Landsat image the ophiolites are the dark bluish-black masses.

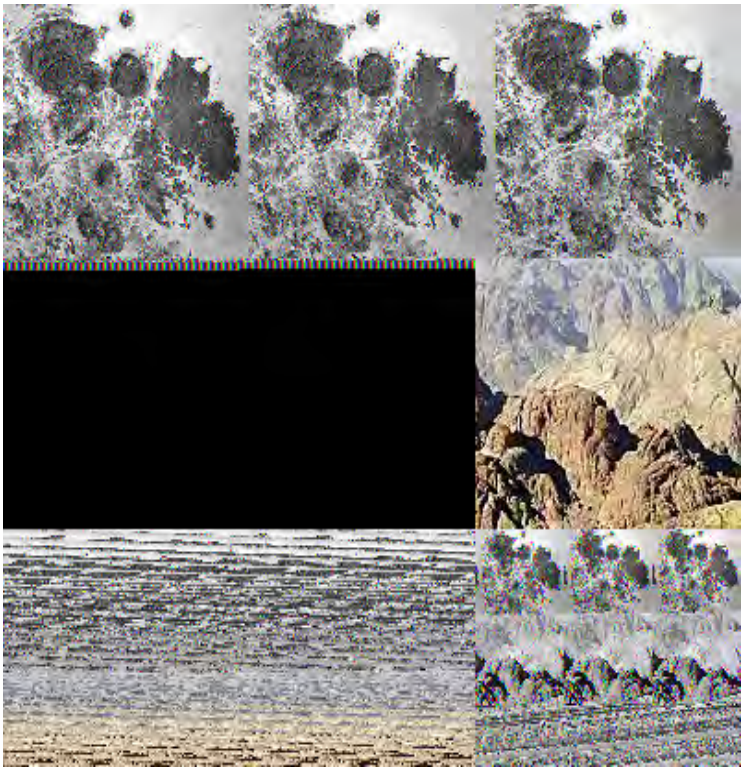


The island of Cyprus in the eastern Mediterranean contains a distinctly different wedge of subducted basaltic rock making up an ophiolite inclusion within rocks of a different lithologic nature. It is prominent in this (cropped) Large Format Camera photo:



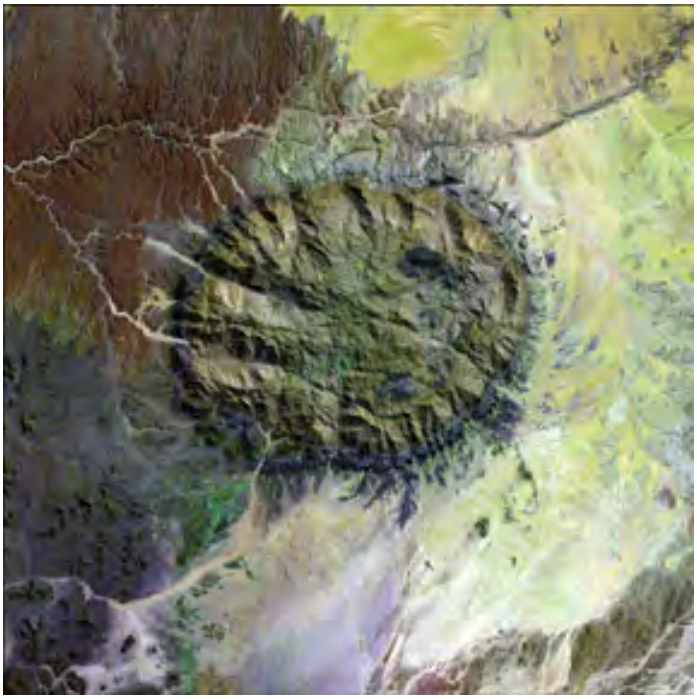
17-5: What part of the image is the ophiolite itself? ANSWER

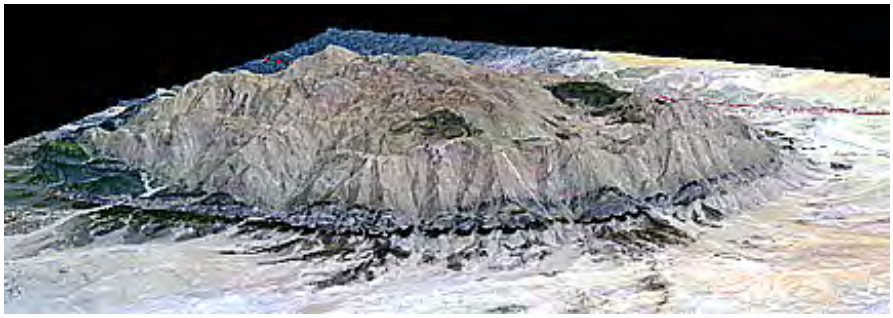
In Africa, Precambrian mountains often stand out in stark relief as topographic highs midst lowlands covered by sand. Such mountains are found in parts of the Sahara Desert. Here are the Air Mountains in Niger, consisting of peralkaline granite intrusions which appear dark (unusual since most granitic masses are light-toned in the field):



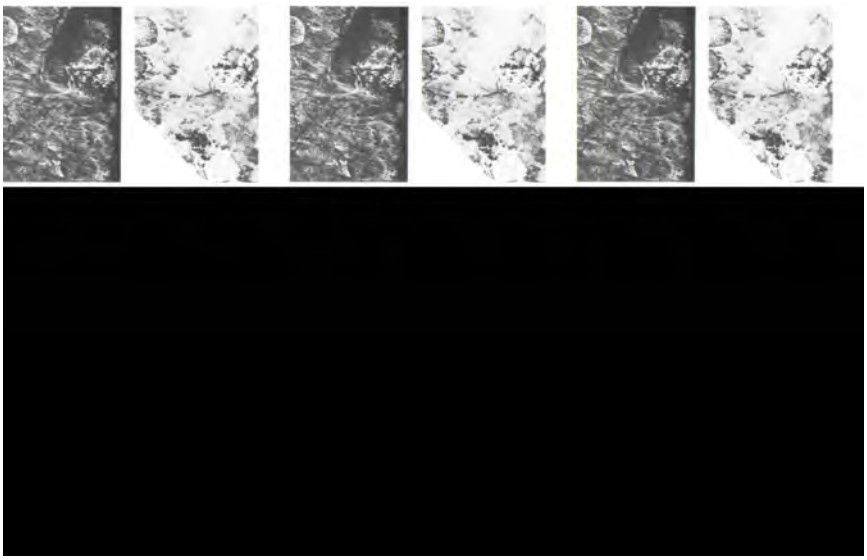
Another swarm of mountains made up of crystalline rocks occurs in the Namib Desert of Namibia. Here is a Landsat-1 image of isolated mountains, some rising to 1800 m (5300 ft) above the desert sands:

Most striking of these is the Branberg Massif (located beyond the above subscene), a Precambrian intrusion, now weathered out, of near-circular shape (occupying 650 km^2) that reaches an elevation of 2573 m (8480 ft). It is seen here in a Landsat-7 image and then a perspective view (January 3, 2001) made from ASTER imagery and its own DEM measurements.





Plutons of the West African Shield in Mali show up differently depending on the type of sensor used. In the left image they are evident in this radar view but do not contrast with the terrain outside the intrusions. In the right Landsat image, the plutons are dark in sharp contrast with the desert sands that occupy the lower terrain (the plutons are low hills not easily covered by sand).

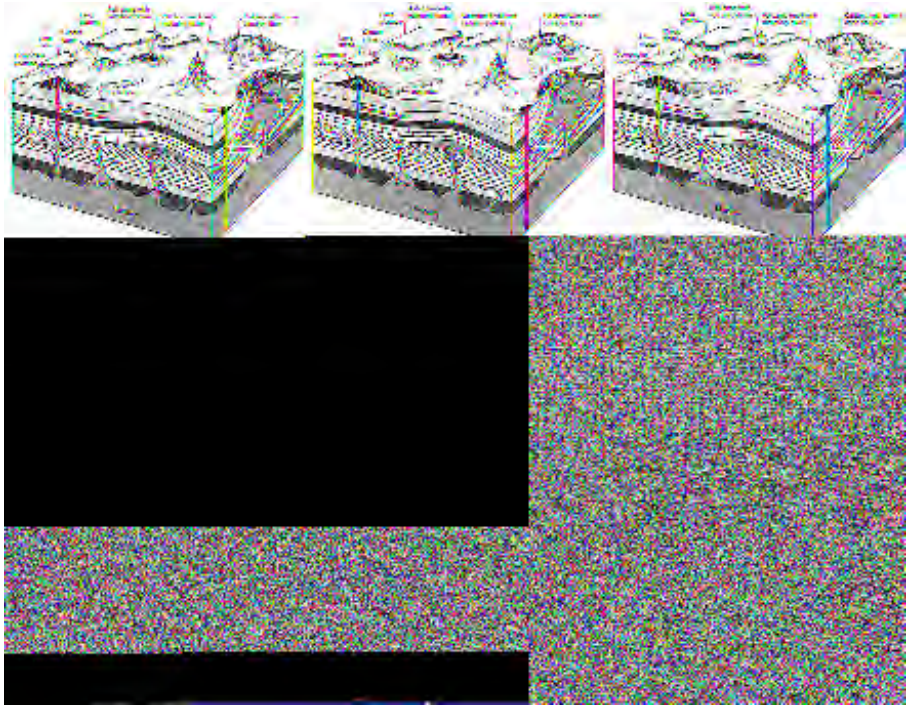


A landform that has aspects both of a structural and a igneous nature is produced by intrusion of magma or lava into a major fault or fracture in the crust to form what is called a *dike*. By far the biggest such feature known on Earth is the Great Dyke of Zimbabwe in Africa. The dike is more than 450 km (280 miles) in length and up to 15 km (10 miles) wide. It is filled with coarse-grained gabbro rather than the customary basalt, which serves to identify it as the feeder vent that produced several lopoliths (now eroded away) that formed in the Zimbabwe craton. Here is part of the Great Dyke as imaged by Terra's ASTER:

A landform that is both tectonic and volcanic is the ring dike. It forms by intrusion of lava into a conical fracture that is circular in plan cross-section. As the surface is eroded, the dike, being more resistant than surrounding sedimentary rocks, stands out as a ring rising above its

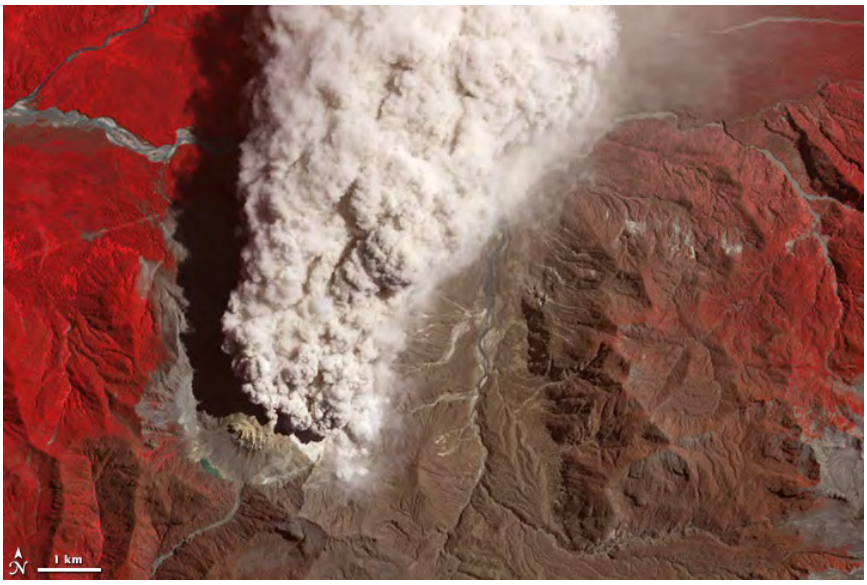
surroundings. A classic example is the Kondyor massif, in which the intrusive igneous rock is an alkaline dunite which hosts platinum minerals and gold, in eastern Siberia, seen in this ASTER image:

Turning now to volcanic landforms, these are especially evident in images made from space. Volcanism is widespread worldwide and eruptions have been monitored in near real time using satellites. Compared with most other landform types, those resulting from eruptions tend to form over short timeframes. Volcanic activity produces a wide variety of surface structures and features. Examine this well-known block diagram prepared by the U.S. Geological Survey that sketches surface volcanic landforms and ties some of these to subsurface structures that serve as the magma chambers and feeders that bring the molten rock (lava) to the surface:



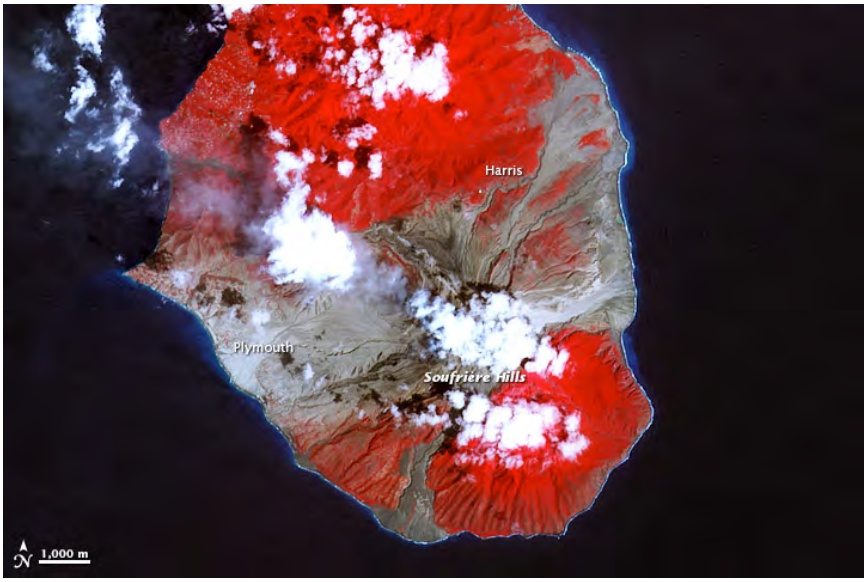
This image introduces how volcanic landforms often stand out in space imagery. Try to identify specific landforms in the scene and name them from the above sketch. The scene includes the Harrat Khaybar in Saudi Arabia:

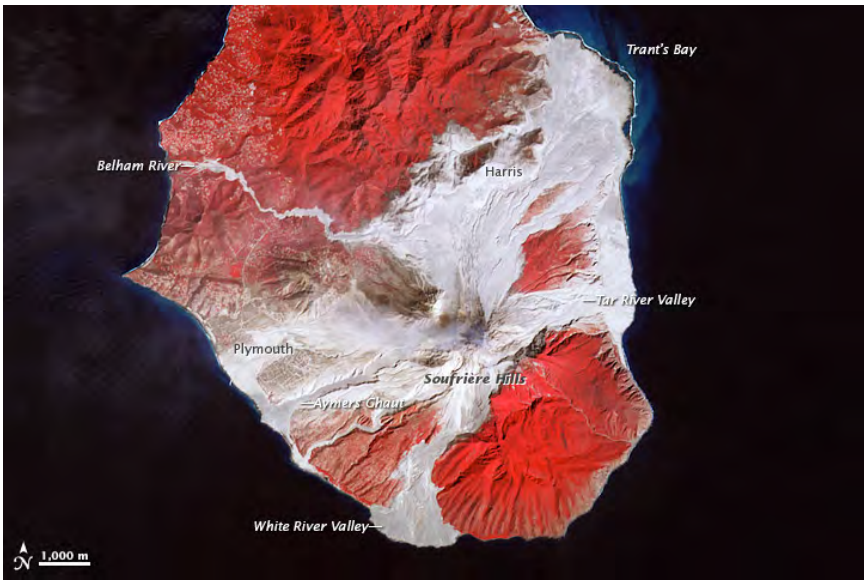
Volcanic landforms are continuously being produced on Earth, since active volcanism is widespread today. Here are two images made by Terra sensors that caught the eruption of the Chaiten volcano near the coast of Chile south of Santiago, plus a ground photo taken at the same time. This volcano had been dormant for nearly 9000 years.



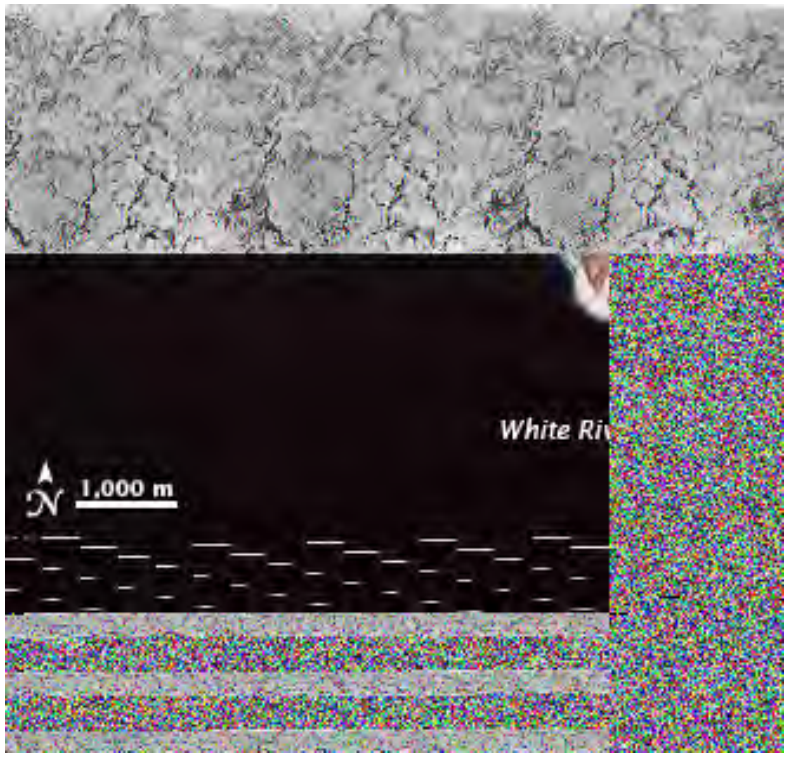
Volcanic eruptions tend to be visually spectacular. As an example, look at the volcano Sakurajima, at the southwestern end of the main islands of Japan.

Satellite imagery is particularly useful in change detection studies because of the short times involved in most eruptions. These two ASTER images show the Caribbean island of Montserrat, with its Soufriere volcano, one of the world's most active. The top scene shows the dark gray lava flows produced during the 20th century; the bottom scene shows the lighter-colored ash from an eruption in January of 2010.

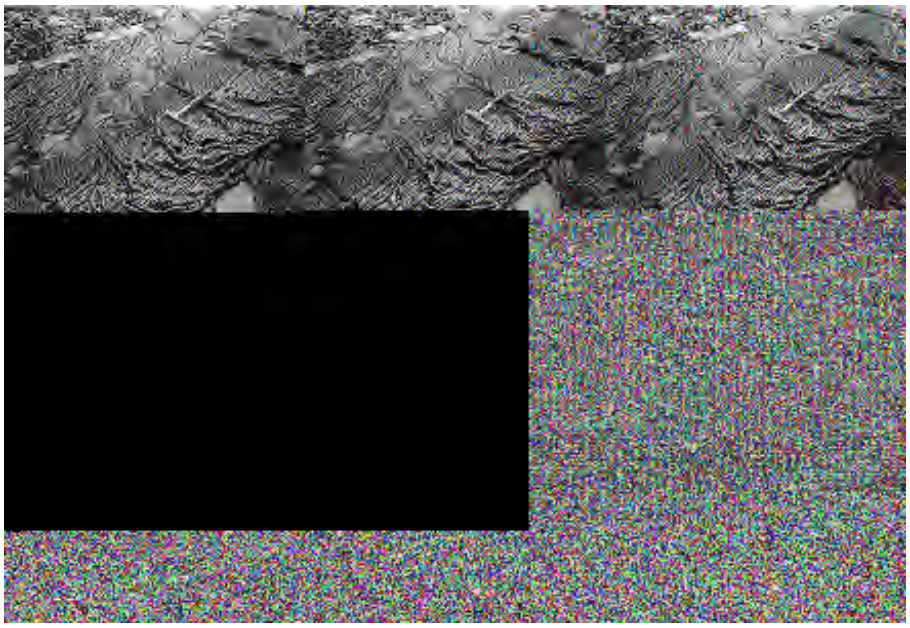




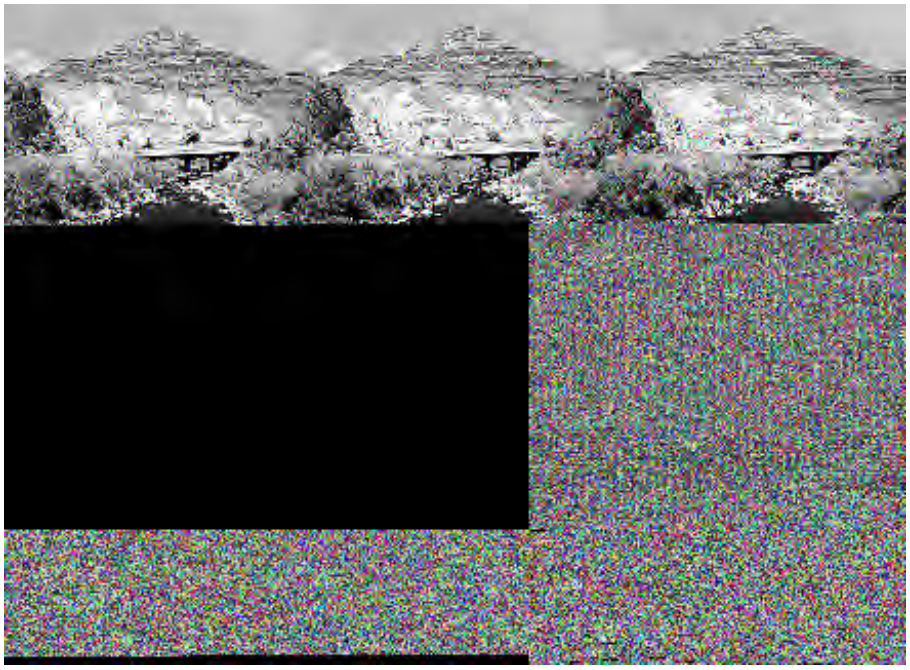
In the diagram depicting volcanic landforms, shown above, note the feature labeled "sill". This is caused by lava intruded along a bedding plane (it thus is parallel to beds above and below). In South Africa are curious landforms, shown in this Landsat image, that result from erosion of the layers of sedimentary rock in which the basaltic rock (dark) has intruded, exposing the lava mainly at the rim edges of low mesalike landforms:



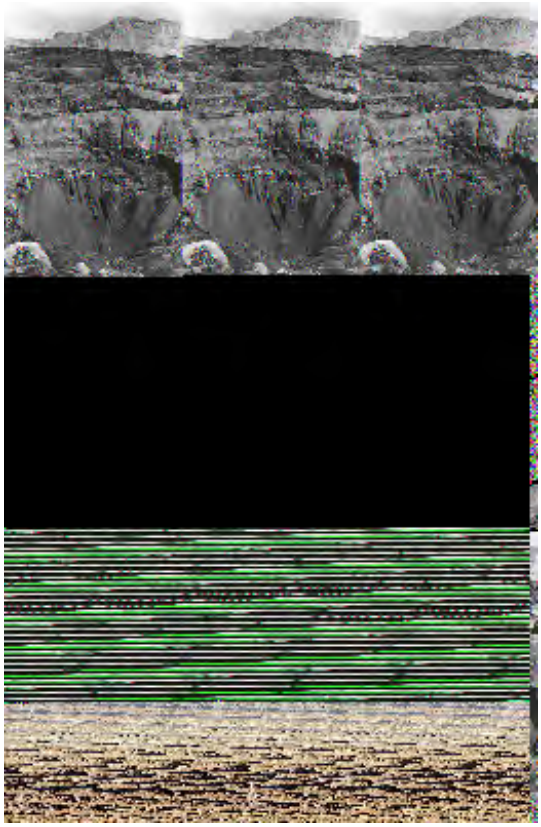
The most common lava is the basaltic type, fluid melted rock high in iron and magneium oxides and with a silica (SiO_2) content around 50% by weight. With this composition it is less viscous and flows easily after reaching the surface. But as it cools, it can make strange surface forms, as seen in this close-up photo of pahoehoe (ropy basalt lava).



Basalt flows often build up thick individual units. These may be separated by volcanic ash deposits. This is the case for the John Day region in central Oregon where alternating flow and ash make up the upper part of the distant mountain and a very thick continuous light ash unit from a volcanic explosive eruption to the west lies in thick beds below it:



In general, volcanic units are composed of layers of ash and/or solidified lava wherever these are emplaced. In the interior of the vent at the top of Mount Vesuvius the layers of this stratocone that built it up over time are exposed, evident in this photo:



We look first at these next five images that represent two major types of volcanoes. Then, we look at terrains carved into vast sheets of volcanic flows or flood basalts.

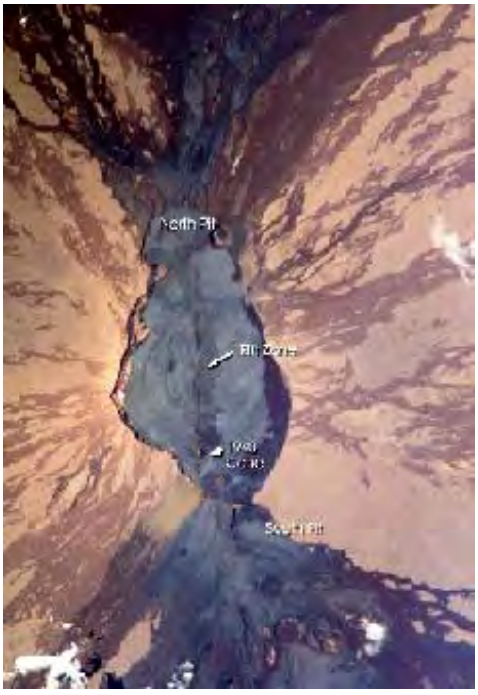
The Hawaiian islands are entirely volcanic, rising as basaltic volcanoes from the ocean floor, reaching heights that carry them above sealevel. The Big Island of Hawaii is the latest (youngest) in this series of volcanic islands formed from melted lower crustal rocks as the Pacific plate moves northwestward over a fixed hot spot in the Earth's mantle. A newer submarine, volcanic complex, now forming southeast of Hawaii, will eventually surface and replace the Big Island as the center of activity. The Islands to the northwest, including Oahu and Maui, were formed earlier as the Pacific plate passed over them in succession. The next image is a Terra MISR scene that includes all of the larger Hawaiian islands:

(As an aside, note that the green signifying vegetation is much more profuse on the right [eastern] side of the islands. The prevailing winds are easterlies; they come from the east. As winds moving water clouds pass over the islands, the precipitation is confined mainly to the

eastern slopes. With much of this water thus lost, the western slopes tend to support considerably less vegetation.)

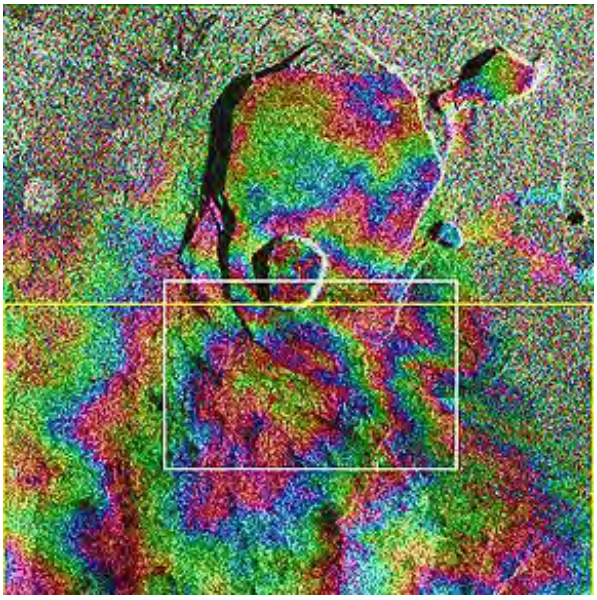
This early Landsat image of the Big Island shows details of the recent (past few thousand years) volcanic flows (see pages 9-7 and 14-11 for other renditions):

Mauna Loa, near the center of Hawaii, is the central part of a huge shield volcano, which comprises the entire island. Its summit crater, a collapsed caldera named Mokuaweoweo, lies beneath a crest at 4,135 m (13,563 ft). Its base lies about 4,000 m (13,120 ft) below sea level, which makes it the tallest single mountain in the world (Everest, while higher, rises from the valleys of the Himalayas that are thousands of meters above sea level, so its relief is less). Mauna Kea, a crater on the north section of the island, is now extinct. But the most active volcano in the world, Kilauea, lies along the east side of the island and is visible here as a dark patch. This island is quite young, consisting of multiple layers of basaltic flows built up in the last one million years. Numerous lava flows (dark basalt), many extruded over the last few centuries, emanate from Mauna Loa, as seen in this photo taken by astronauts aboard the International Space Station:

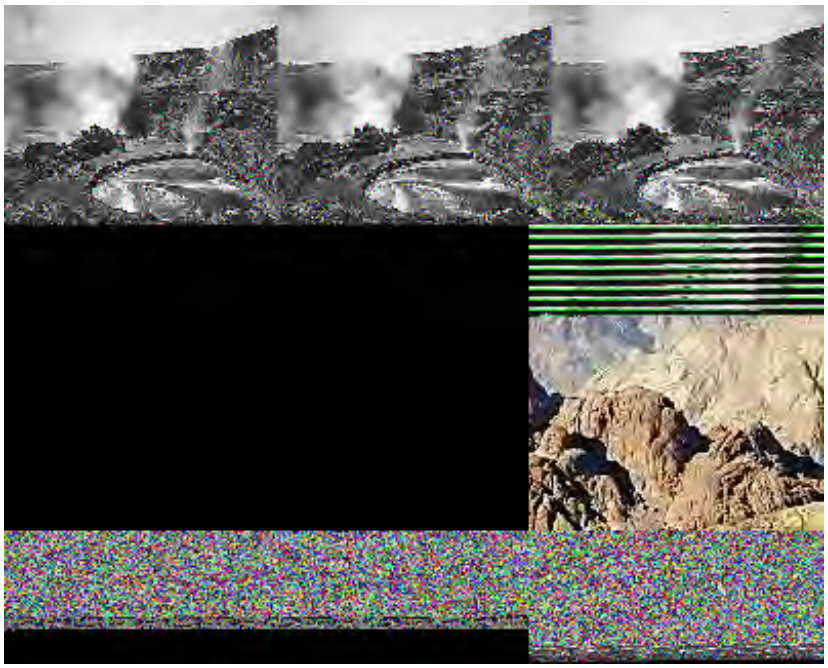


A trip to Kilauea at any time stands a good chance of showing some eruption, either along a rift zone connecting it to Mauna Loa or from the caldera that has developed. The next two images were made by SIR-C X and C band radar; check the captions for more information:

The main caldera is evident in this radar interferometric image of part of the Kilauea volcanic edifice.

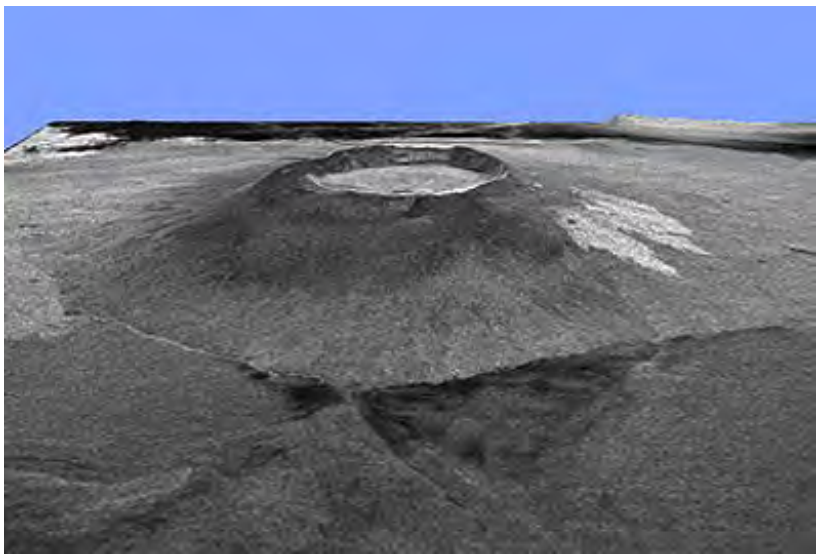


This caldera is steep-walled with a smaller sink as seen in this ground photo:



Calderas occur wherever large volcanoes either have a collapse of their peaks or eruptions have "blown off the top", as at Mt. St. Helens. Here is a large caldera in the Sudan of eastern Africa:

Transitional to the stratocones described below are the steeper-sided volcanoes that make up the basaltic Galapagos Islands (see page 6-10), some 500 miles (800 km) west of Ecuador, in the eastern Pacific Ocean. This next illustration shows Volcano Darwin on Isabela Island in a perspective image made from SIR-C radar data and elevation data acquired by TOPSAR (an aircraft-mounted radar system designed to measure topographic elevations):



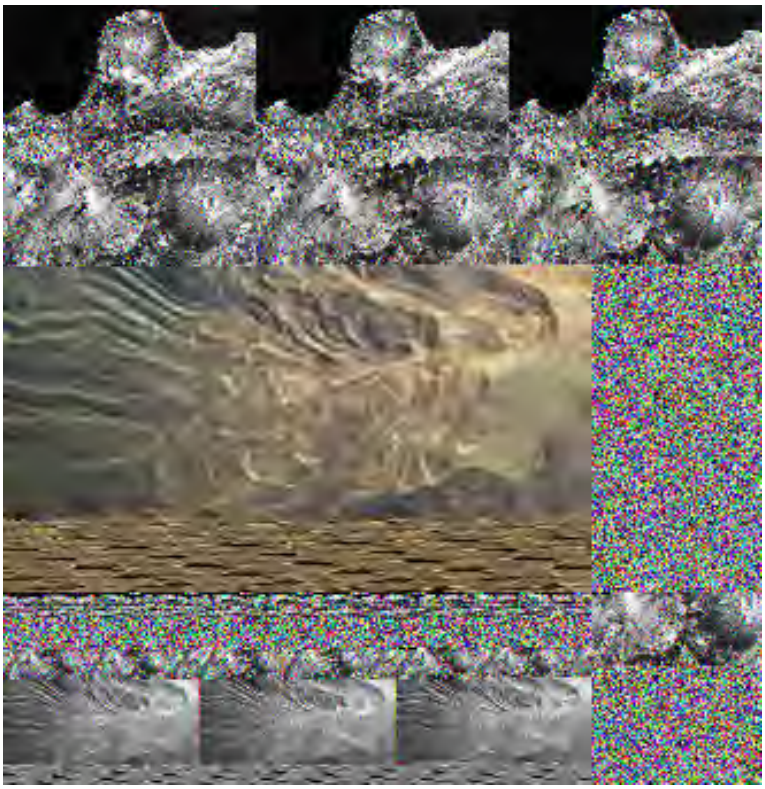
The other spectacular type of volcano is the stratocone, noted for its steep sides and, often, its symmetrical form. In the U.S. Mainland, the most photogenic stratovolcanoes are in the Cascade Mountains (from Northern California into Northern Washington), and in the Aleutian Islands of Alaska. The photo below shows the north side of Mount Rainier, a massive, still active volcano rising to 4300 m (14411 ft) to the southeast of Seattle, WA. Like most other Cascade stratocones, this volcano is superimposed on older, much eroded volcanic rocks from earlier periods of volcanism. Below the photo is a view from space made from multiband radar imagery acquired by SIR-C.

The classic example of a stratocone is Mount Fujiyama west of Tokyo. It rises to 3776 meters (12460 ft) from sealevel. Here is a vertical photo taken by an ISS astronaut:

Mt Fuji is considered to be the most perfect (symmetrical) stratocone presently on Earth. Judge for yourself from this ground photo:

Another photogenic volcano is Teide in the Canary Islands along the Mid-Atlantic ridge. In this astronaut photo, almost as striking are the numerous lava channels, many bounded by lava levees:

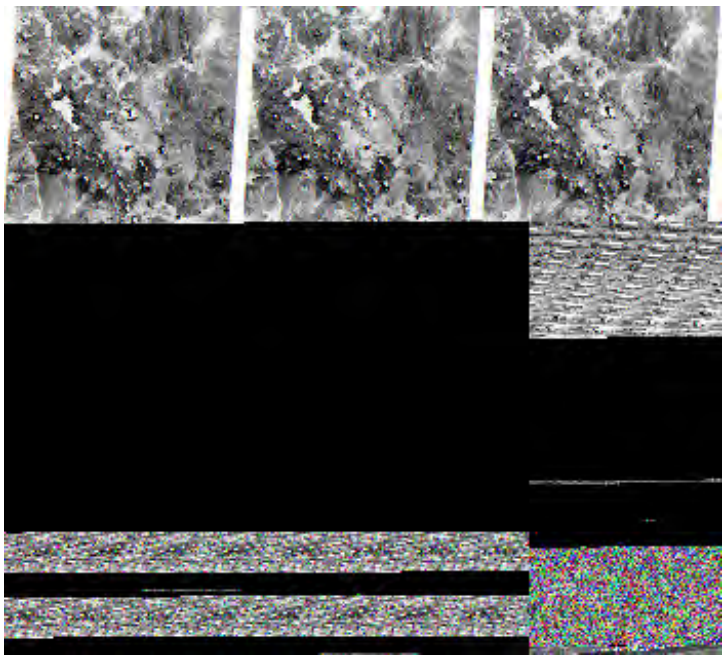
Below is a Landsat view of a segment of Java, the main island in the Indonesian archipelago, a prime example of an island arc terrain still evolving. Nine stratocones are in the scene; the three most prominent are Muria (top center); Merapi (lower left), and Lawu (lower right).

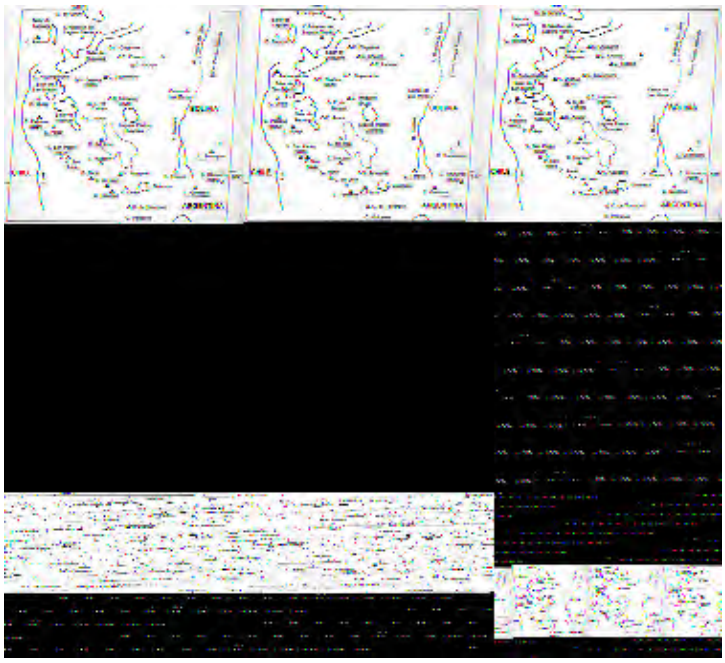


In the midst of thick sequences of geosynclinal sediments are a series of large composite stratovolcanoes, developed from crustal melt induced by frictional heat, as the Indian-Australian plate dives in subduction below the southernmost extension of the Eurasian plate (see the tectonic map at the top of this page). The stratocone on the north peninsula near the Java Sea is Muria. The highest (2,910 m, 9,545 ft) volcano is the active Merapi, which stands out as the lower of two in the left center. To its right is Lawu. Six other large volcanoes are mainly to the west (left) of Merapi.

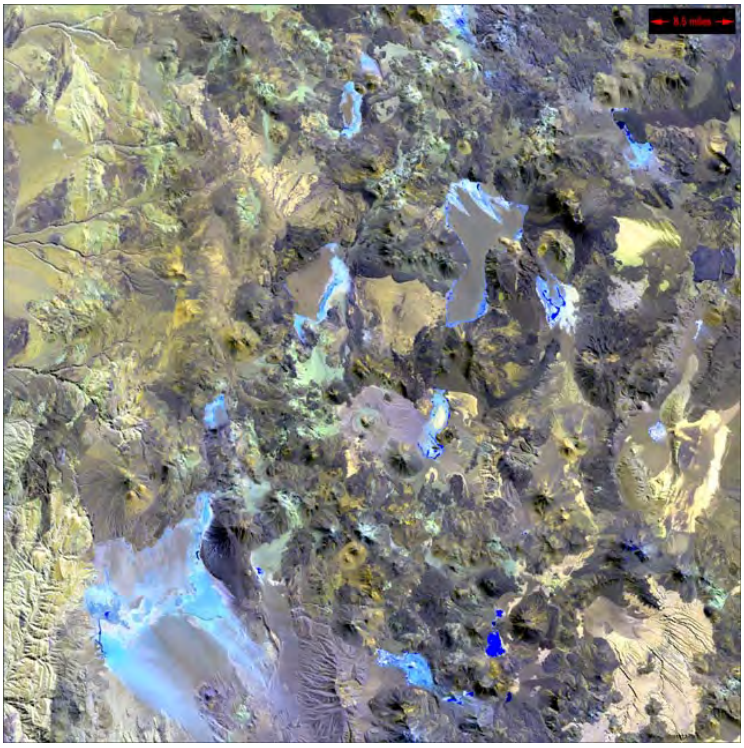
17-6: What is missing volcanically in the Java image that is present in the Hawaiian scene? ANSWER

Stratocones come in various sizes and can occur in dispersed swarms. This is particularly a hallmark of volcanoes in the South American Andes Mountains as seen below. Use the sketch map as an aid to picking out the individual cones, many of which are snow-capped in this southern Fall image, owing to the high elevations of the flanks of the High Andes.





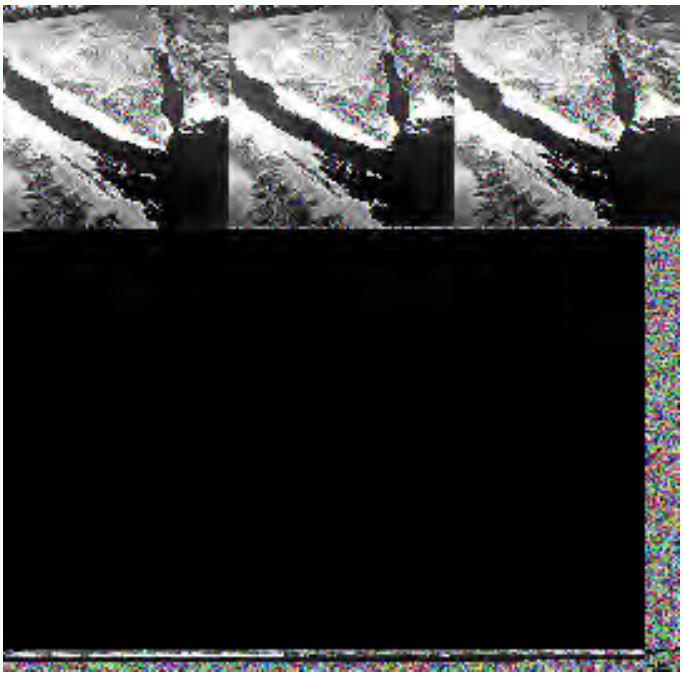
This color version of similar volcano coned-dotted terrain in Chile is interesting:



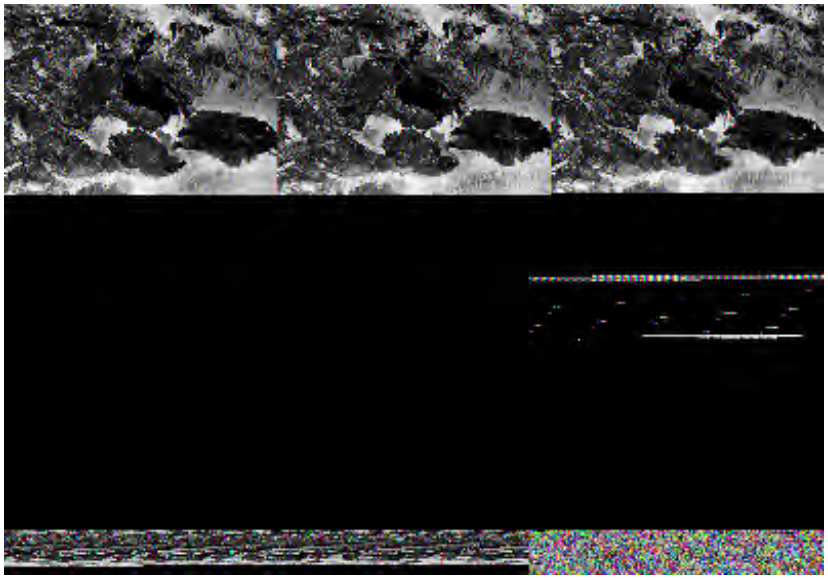
On a much smaller size-scale are cinder cones - in some respects miniature stratocones. Here is a swarm (right side) of cinder cones on the flank of a caldera (left) on the Isla de la Palma, one of the Canary Islands off the Mid-Atlantic Ridge:

This photo taken from the International Space Station shows cones in North Africa.

Lavas (magmas that reach the surface) extrude not only from discrete individual volcanoes but from deep-reaching fractures in the crust that can tap into the upper mantle. The result is widespread flows covering large areas. We saw one example in Section 3 of basaltic flows in the East African Rift. This huge fracture zone runs across much of the eastern side of that continent as one of the "arms" of splitting tectonic plates. Two other arms or dividing zones, where Africa is breaking off from the Arabian plate and from the Australo-Indian plate, meet the newly developing East African arm at a "triple junction" located in the Afar of Ethiopia. This junction was captured photographically by astronauts on the Earth-orbiting Apollo 7 (pre-lunar) mission; the same region was shown earlier on this page:

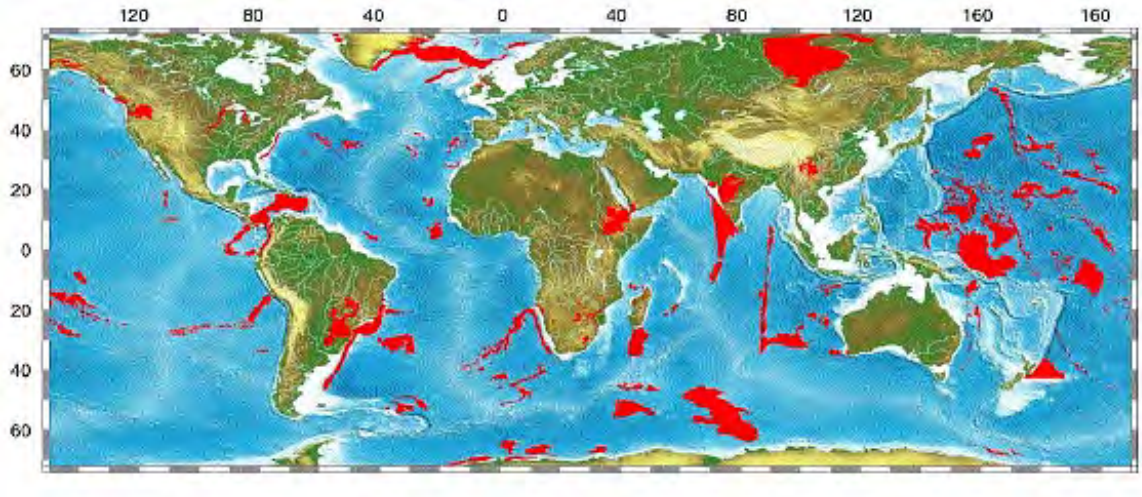


Associated with the rifts, beyond the apex of these spreading centers, great quantities of basaltic lavas are pouring over the surface in the Afar Triangle. This locale was photographed with the Large Format Camera, from the Shuttle, as seen here:

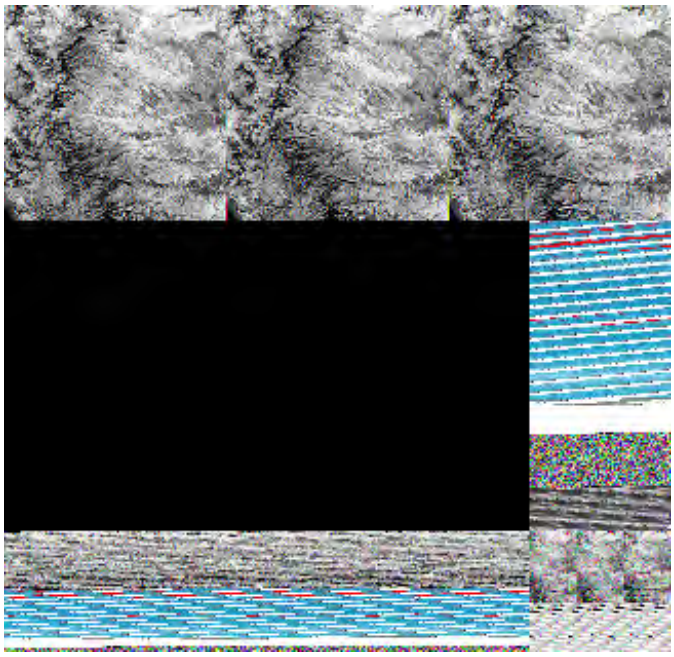


Well south along the East African Rift, the fault zone in the basalts narrows. Here is that segment, part in Kenya and the southern part in Tanzania. In the upper right is the great stratocone of Mt Kenya; its larger companion, Mt. Kilimanjaro is in the lower right:

Flood basalts, extruding from many fissures, and moving out to cover 100s of thousands of square kilometers, are found on several continents. They occur mainly in active tectonic zones. Here is a map showing the global distribution of flood basalts and similar oceanic basalt fields, of various ages, but occurring at the crustal surface:



As examples of flood basalts consider first the Deccan plateau basalts in western India, whose extrusions for more than 70 million years are related to the collision of India against the southern margin of the Asian plate. A Landsat view shows the landscape, often barren, mountainous, and with a lower population density. The flows spread over a wide area The photo below it reveals the expression of the flow layers.



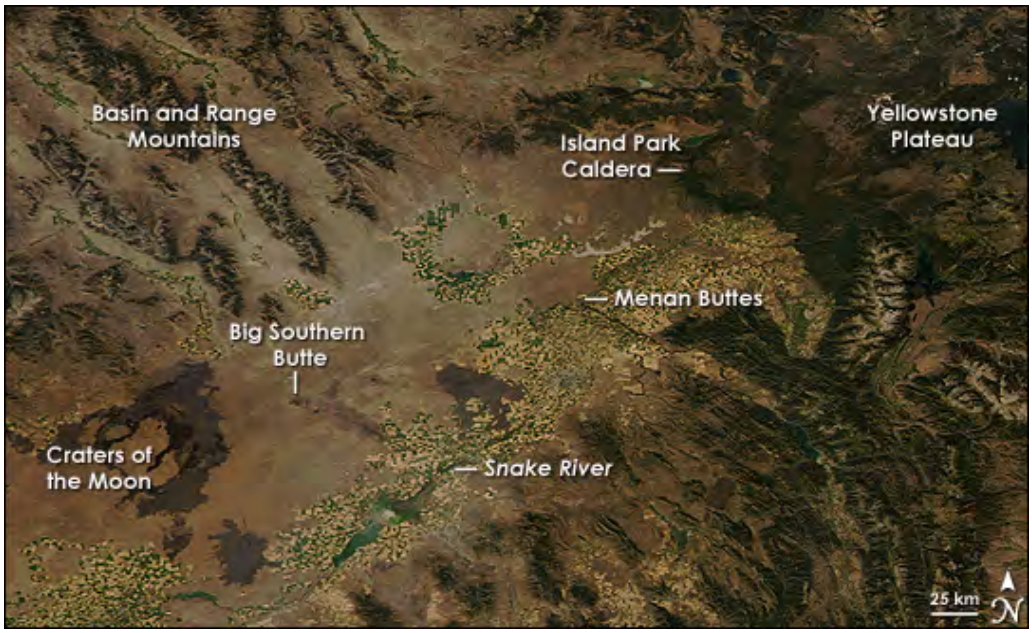
In the United States the largest outpourings of basic lava are the Columbia River and Snake River basalts of Oregon, Washington, and Idaho. They form widespred plateaus built from lavas piled on lavas, the many successive outflows producing distinct layering. This map shows the three major volcanic provinces of the Pacific Northwest:

Lava Plateaus

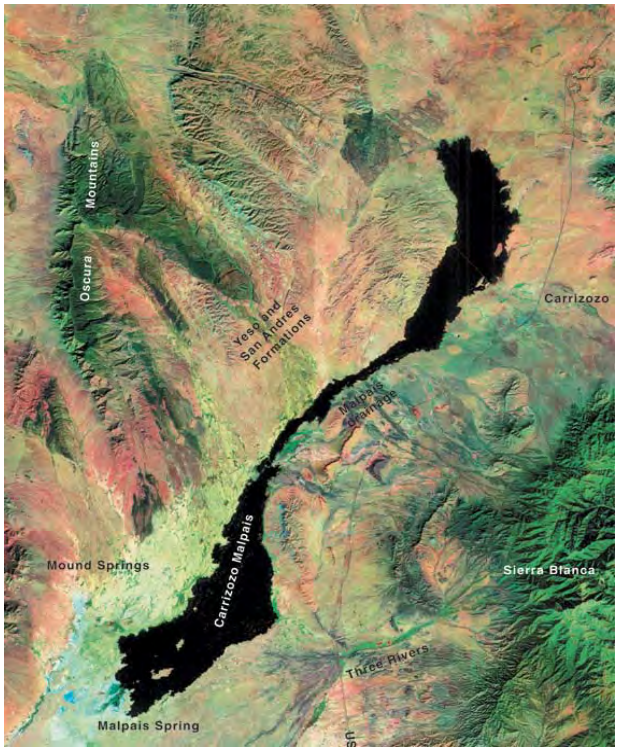
Figs. 4.19

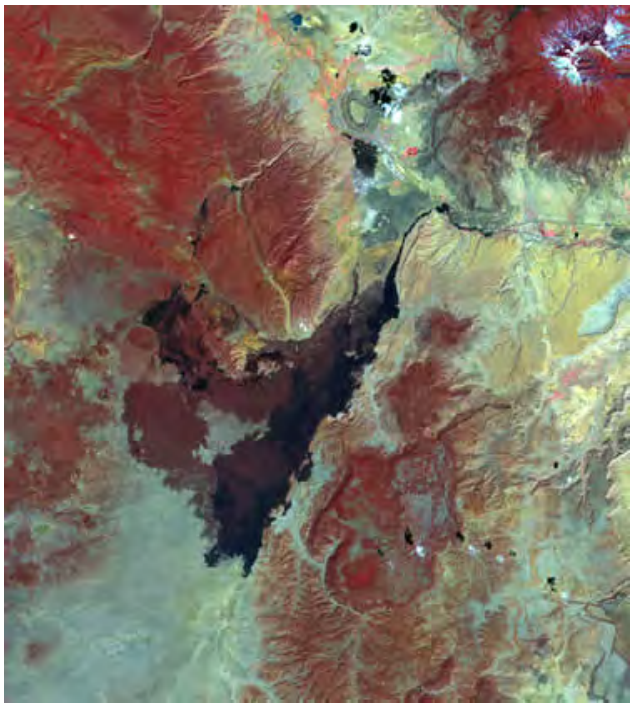
Here is a field photo showing a series of flood basalt flow layers in the Columbia Plateau.

Although extensive is areal distribution and thick (multiple flows), these basalts are not readily apparent in satellite imagery. They tend to have formed soils that support agriculture. The Snake River Plains of Idaho are typical but the role of volcanism is revealed by some recent lava flows, such as those making up the Craters of the Moon National Monument:

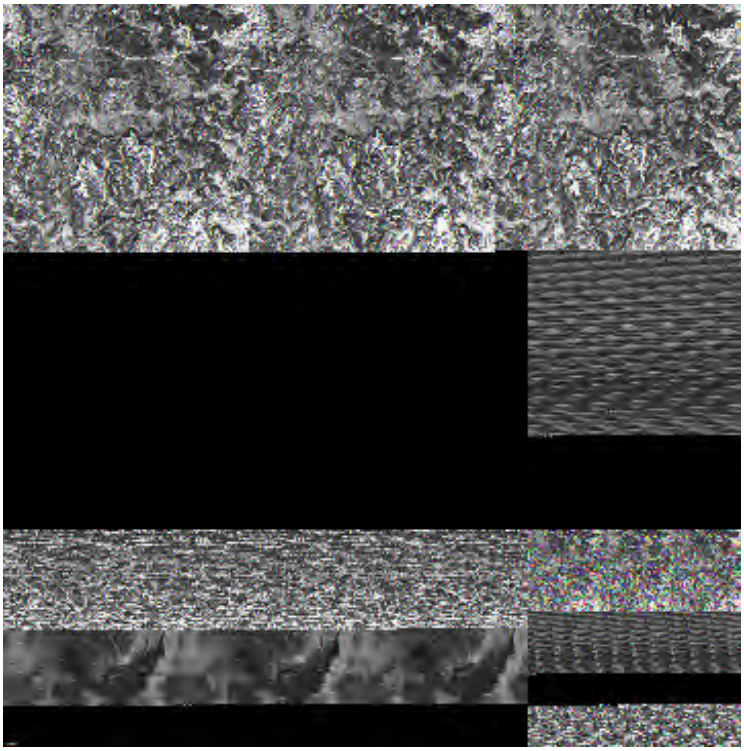


Localized flows such as those at Craters of the Moon often occur in apparent isolation from their source volcanoes. Here are two small flows in the arid mountainous terrain of New Mexico (see captions for their names); both stand out because they are dark basalt:

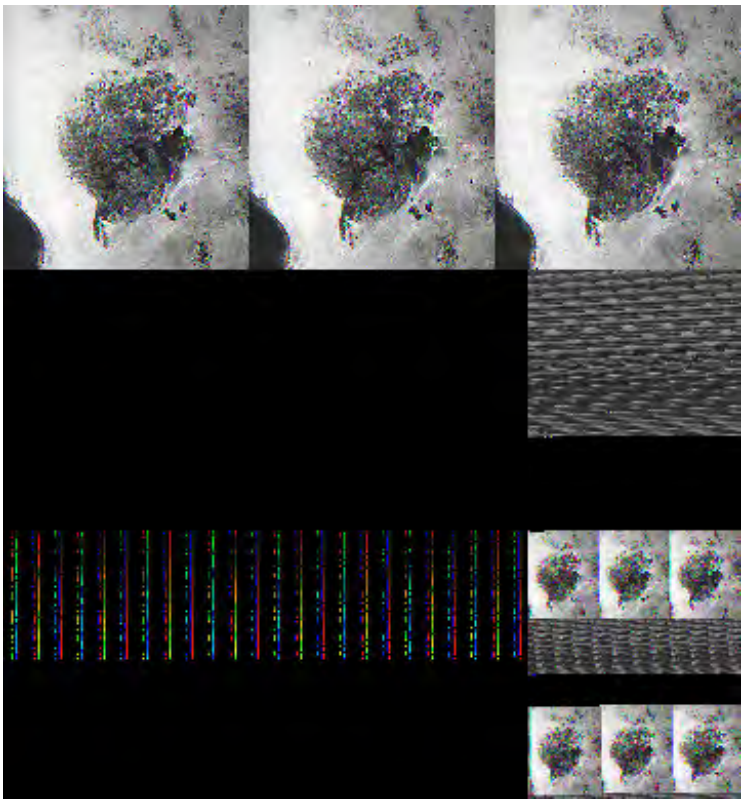




Large areas covered by basaltic lava will develop distinctive terrain forms. This next Landsat image shows flat to somewhat rounded hills in thick flood basalt flows ("trap rock") overlying Permo-Triassic sedimentary rocks in the northwestern Siberia Plateau.

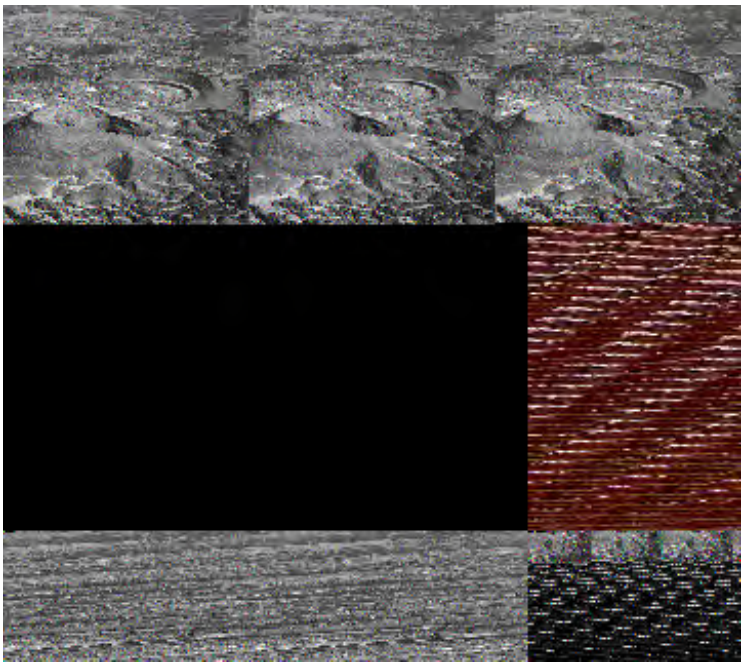


Let us look at a small volcanic field in the Mexican state of Sonora just south of the border with Arizona. Little was known of it among American volcanologists until this Gemini photo was taken in the 1960s. The Pinacate field, covering an area of 1500 km^2, consists of basaltic flows coming from more than 200 vents:

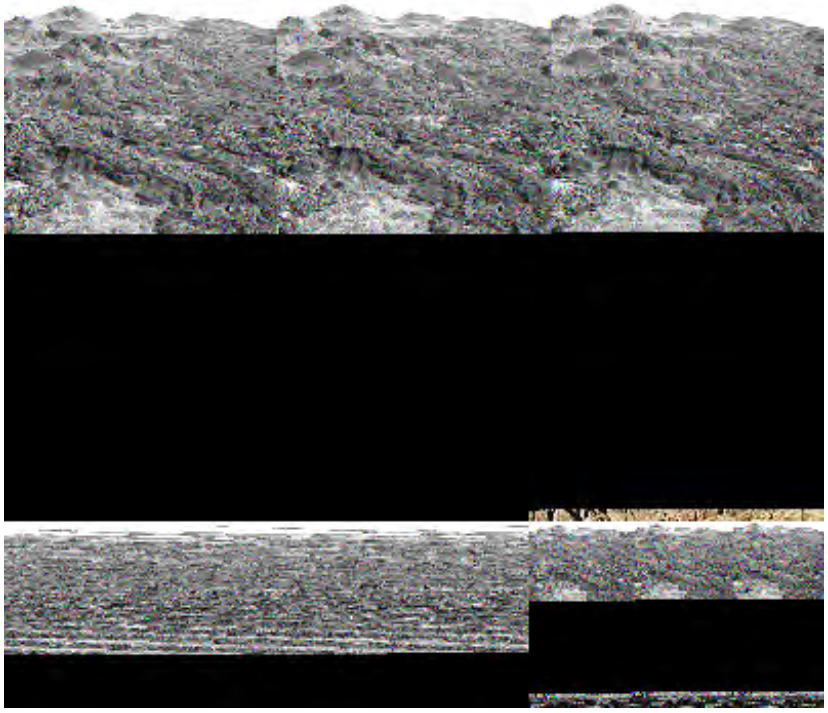


The same field is shown in color made from individual radar bands in this SIR-C image:

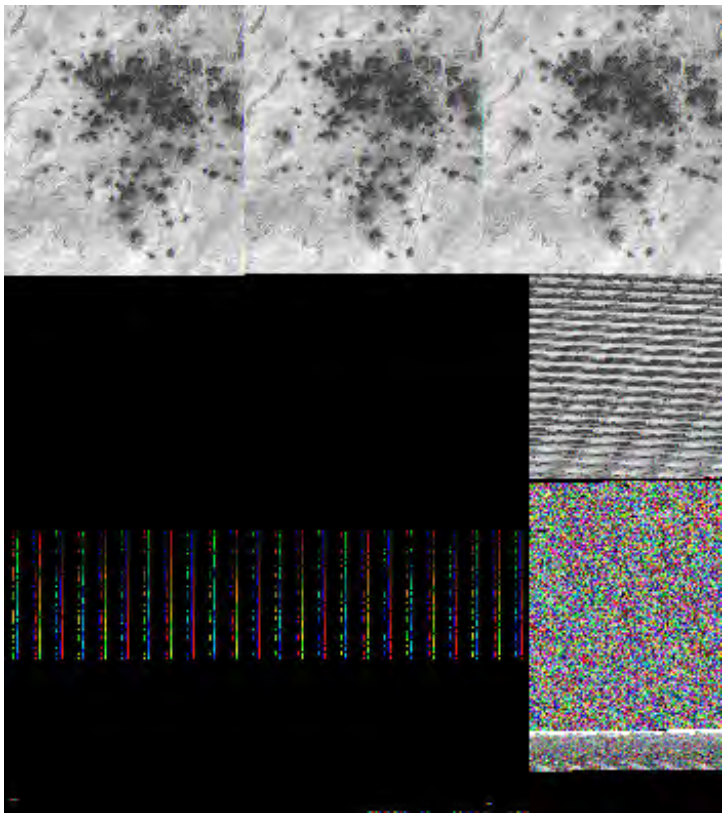
Cinder cones and eroded maars (volcanic structures partly blown away when lava encounters water which turns to steam; another example from Pinacate is seen on page 18-1) are characteristic of this Field.



Taller volcanic cones and lava flows are evident in the eastern part of the Pinacate Field, as seen in this photo:



Finally, we examine the Hopi Butte volcanic landforms that show up as small dark flat-topped hills in a swarm of more than 200 individuals in the Painted Desert of the Colorado Plateau in northern Arizona. This is a subscene extracted from an early Landsat image:



Seen from the ground, the distinctive flat-topped nature of individual structures makes this, along with Monument Valley to the north, a popular spot for western movies (especially favored John Wayne):

Photo credit: Louis Maher

Most of these structures are diatremes (steam-driven volcanic material that punches its way through sedimentary rocks); some are the maars (crater-shapes developed when the surface rocks are expelled) that form above the intruding diatremes which eventually are exposed to make the volcanic necklike surface shapes. Some are volcanic caps on sedimentary rocks.

It is worth commenting to close this page that much/most of the exteriors of the other inner or terrestrial planets, and our Moon, are surfaced by countless basalt flows. (And, of course, this applies to the bedrock below marine sediments on the Earth's ocean floors.)

PREVIOUS RST NEXT

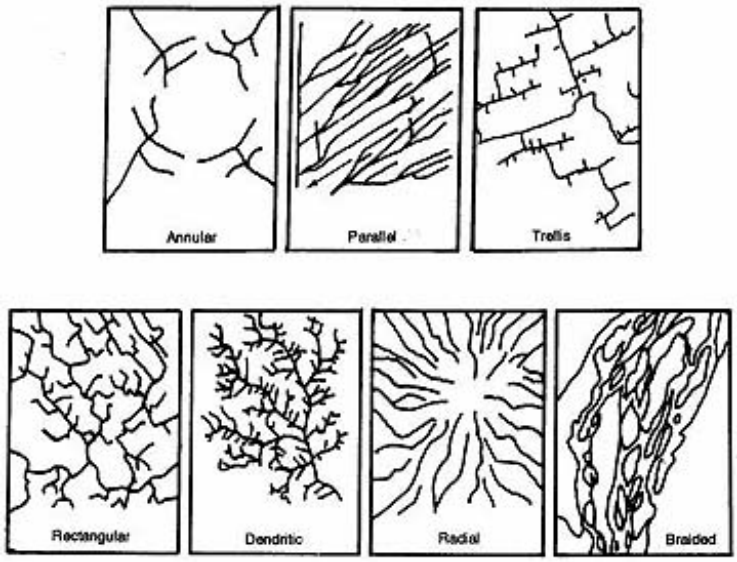
Primary Author: Nicholas M. Short, Sr.

The next three landform categories have in common an origin that involves running water (streams and runoff) or water in lateral motion (waves) on top of oceans or lakes. Drainage patterns established by rivers/streams show considerable variety. One common type is dendritic. The influence of stream erosion and associated slope wash can determine the shapes and heights of the intervening land, whose character ranges from mountains to plains. Streams form deltas and alluvial fans when they drop their particulate loads on reaching stiller water. Marine waters produce a variety of shorelines whose outlines are partly controlled by the nature and elevation of the land affected by coastal erosion.

Fluvial/Deltaic/Coastal Landforms

Switching to fluvial landforms, of which there are many varieties, we start by showing the MacKenzie River in Canada, which shows many typical characteristics of a river. The river itself is fairly straight and broad. Its tributaries display a common condition in which those rivers - in order to lengthen their path and thereby lower their gradient (rate of decrease in elevation per unit length of the stream) - develop curves called meanders.

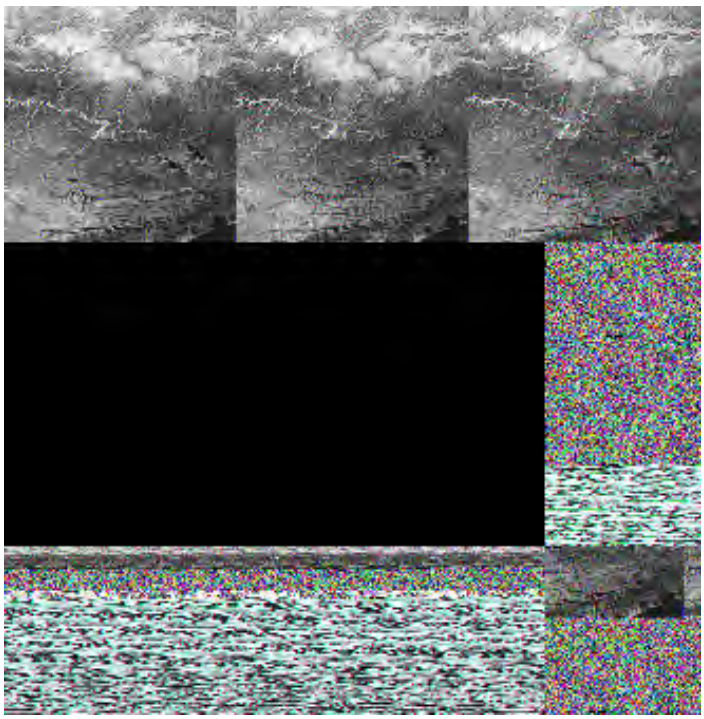
In addition to meandering, rivers show a variety of drainage patterns. This sketch shows the principal types:



In many space images, the type(s) present are not that easy to discern and classify. Tracing the rivers present usually reveals what the pattern type is. By far the most common type is the dendritic pattern. This Landsat subscene of the Dirty Bend, Utah drainage illustrates dendritic drainage:

The appearance of this type of drainage, so distinctive, can be enhanced by spatial filtering. This Landsat image of a highlands area west of Riyahd, Saudi Arabia, illustrates the effect:

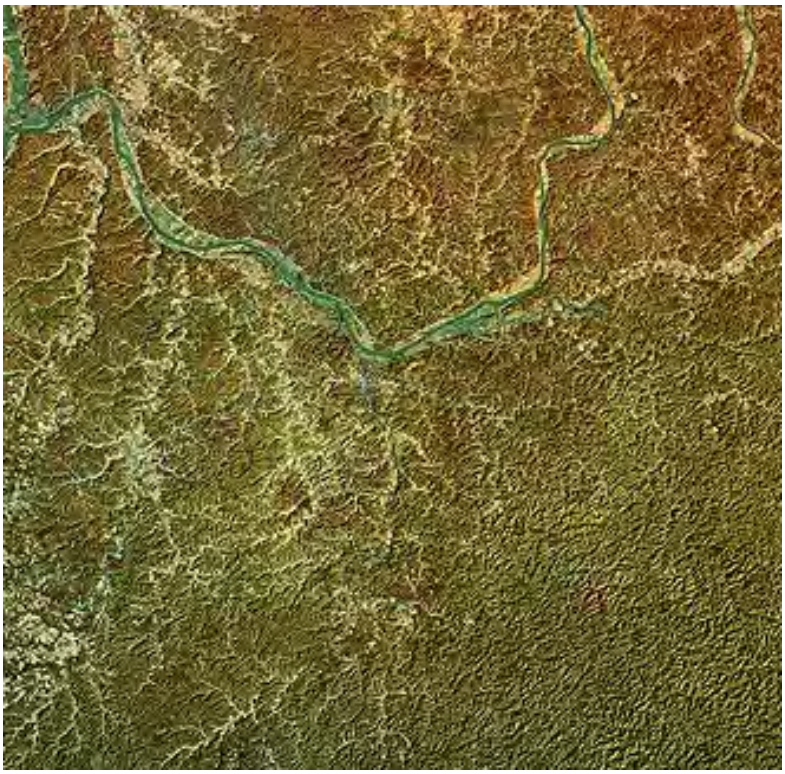
This Landsat scene of South Yemen shows that the dendritic drainage has highly dissected the topography in flat-lying rocks. The subscene below it further amplifies the drainage style:



The South Yemen Landsat scene above extends over part of the Hadhramut Plateau, an uplifted section of Tertiary limestones and shales, which experienced folding into a broad

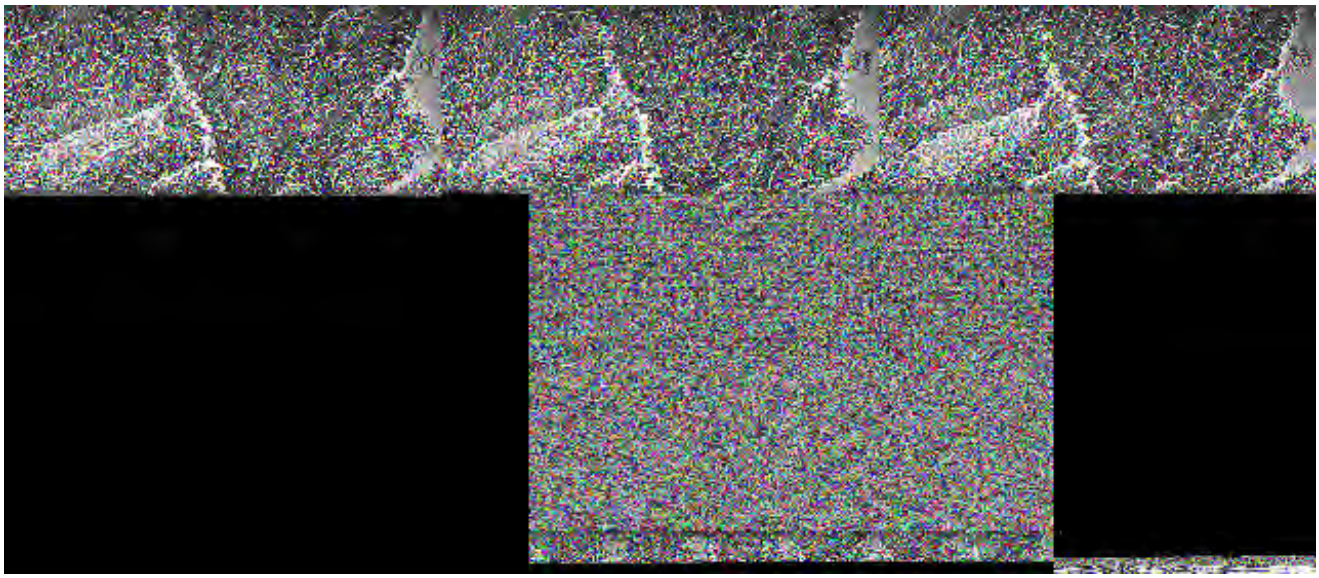
syncline (center) and two anticlines (top and bottom of the image, but not visibly evident). Typical dendritic drainage develops as the Wadi al Masilah, which, along with its tributaries, is a sometimes ephemeral stream (its valley is sand-filled) that obtains flow mostly after large storms but maintains enough water to support local farming. Much of this drainage likely developed during a wet period before the regional climate shifted to its present arid state. Note the typical headwaters pattern for dendritic drainage in the system in the upper right. The drainage pattern near the coast has become trellis-like.

17-7: The term "dendritic" was introduced above, but not defined. Using the image as a clue, what do you think this word means, in a geomorphic sense? ANSWER



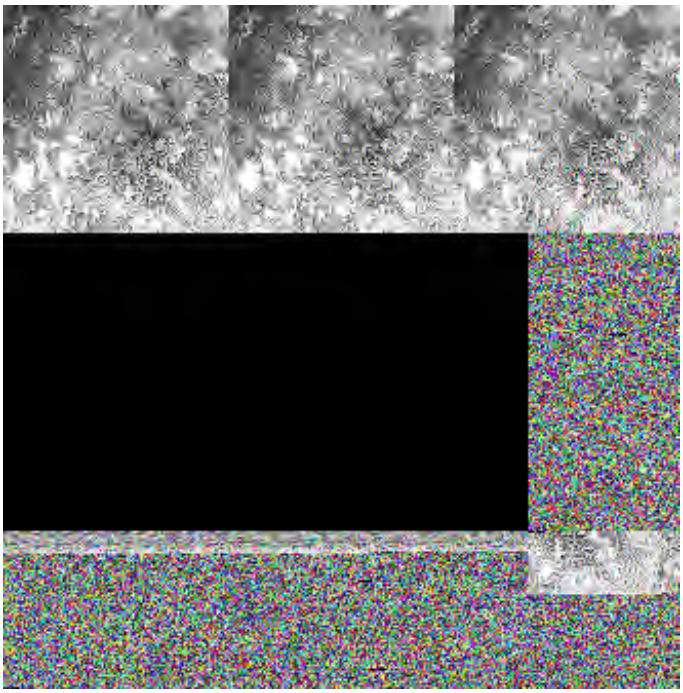
The above subscene (about 100 km [62 mi] wide) is in West Virginia, next to the Kentucky border. The Ohio River flows just to the north of the image top. The area is part of the Appalachian Plateau, an uplifted sector of the crust involved in the Appalachian orogeny that did not experience folding, so that the rocks remain horizontal. The drainage has reached a level referred to as mature, in which a distinctive high-density dissection has reduced divides to sharp ridges, with little of the earlier uplands remaining, and narrow valleys. Although this pattern leads to maximum relief, the differences in elevation are seldom more than about 200 meters (656 ft).

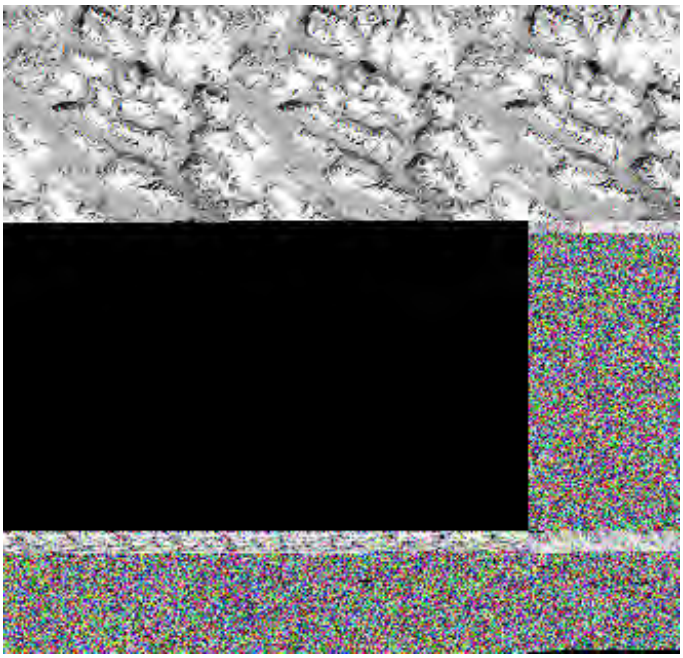
Another example of dendritic drainage, this time imaged by the SIR-A radar, is this scene in east-central Columbia. The region is one of tall grasslands and forests.



A fourth example is also in South America. This STRM topographic image shows rivers with conspicuous tributaries. The area covered lies in western Brazil where it meets parts of Bolivia and Peru. Note the strong expression of the Trans-Amazon Highway (I interpret it to be made of concrete). There is a small impact crater seen in the lower left.

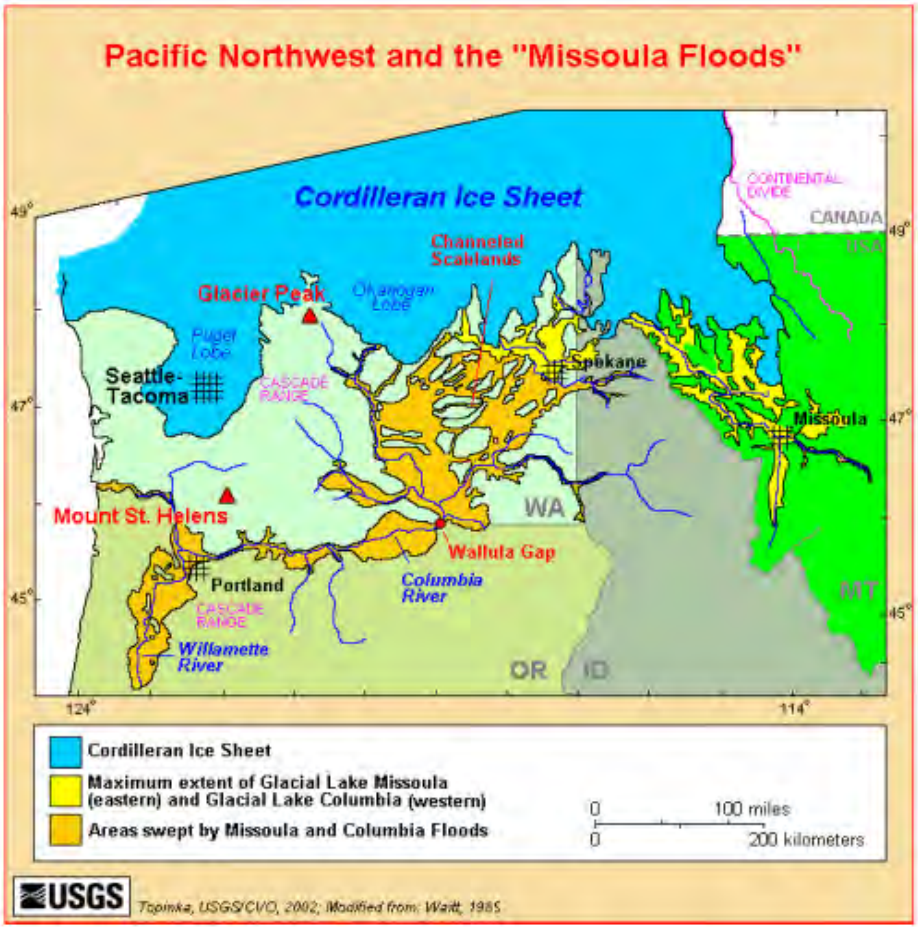
Both parallel and trellis drainage patterns are usually structurally controlled. Here is an example of each; see caption.



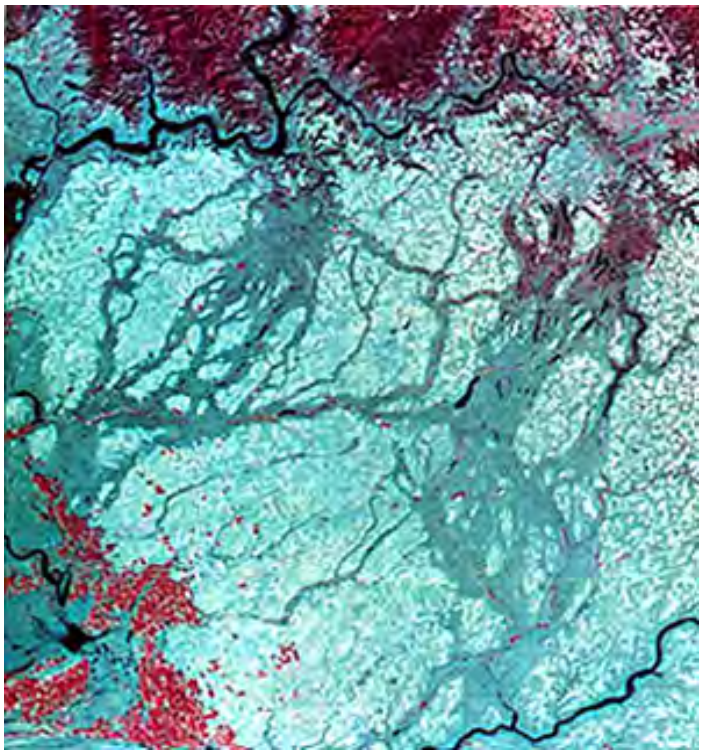


Radial drainage tends to develop on structural domes. It is even more common on slopes of stratocone volcanoes, as seen on the Bromo Volcano of eastern Java.

Normally, the sculpturing of landforms by running water is an imperceptibly slow process, taking millions of years to bring about notable changes. There is a rarely occurring exception to that: if the water is released suddenly, as from a dam burst which empties the backfilled lake, the huge volume thus emptied can carve into the downstream landscape in hours to days. This happened on a grand scale during the ice age in western Idaho, much of Washington, and a part of Oregon. The result is a modified landscape known as the Channeled Scablands. Its location (shown in gray) and the inferred origin are indicated on this map:



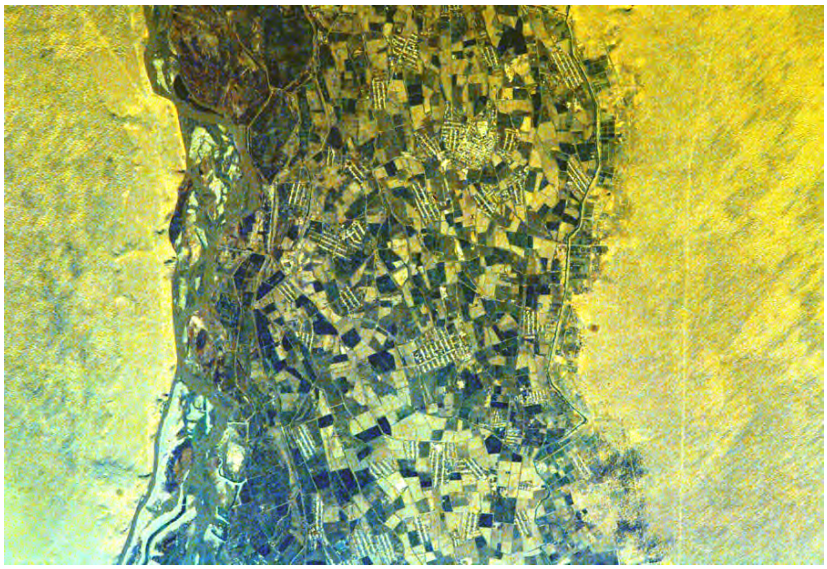
From space, in this Landsat image, the affected areas are darker gray in the color composite:



The present day Columbia River runs across the top of the image; Spokane, WA is near the upper right corner. The farmlands in the right third are in the Palouse country, a region of mainly wheat crops. Most of the lowlands are underlain by Columbia River basalts, but these are covered by windblown loess (fine dust) derived from Pleistocene glacial deposits. Such material is easily eroded by the rush of waters. Most of the waters had formed large glacial lakes (Lake Missoula in Idaho; Lake Spokane in Washington) developed behind natural dams made of glacial ice. When these burst, they sent cataclysmic floodwaters out over wide areas, largely confined to pre-existing valleys, that cut into the loess and basalt (the latter is responsible for the darker grays that denote the channels). Distinctive landforms resulted; one example appears below. Other examples are displayed in this USGS web site.

Many rivers have distinctive floodplains - flat areas affected both by erosion and by deposition - which extend beyond the river channel, often to bluffs that mark the edges of higher land being dissected. A typical example is the floodplain of the Brahmaputra on the Indian subcontinent:

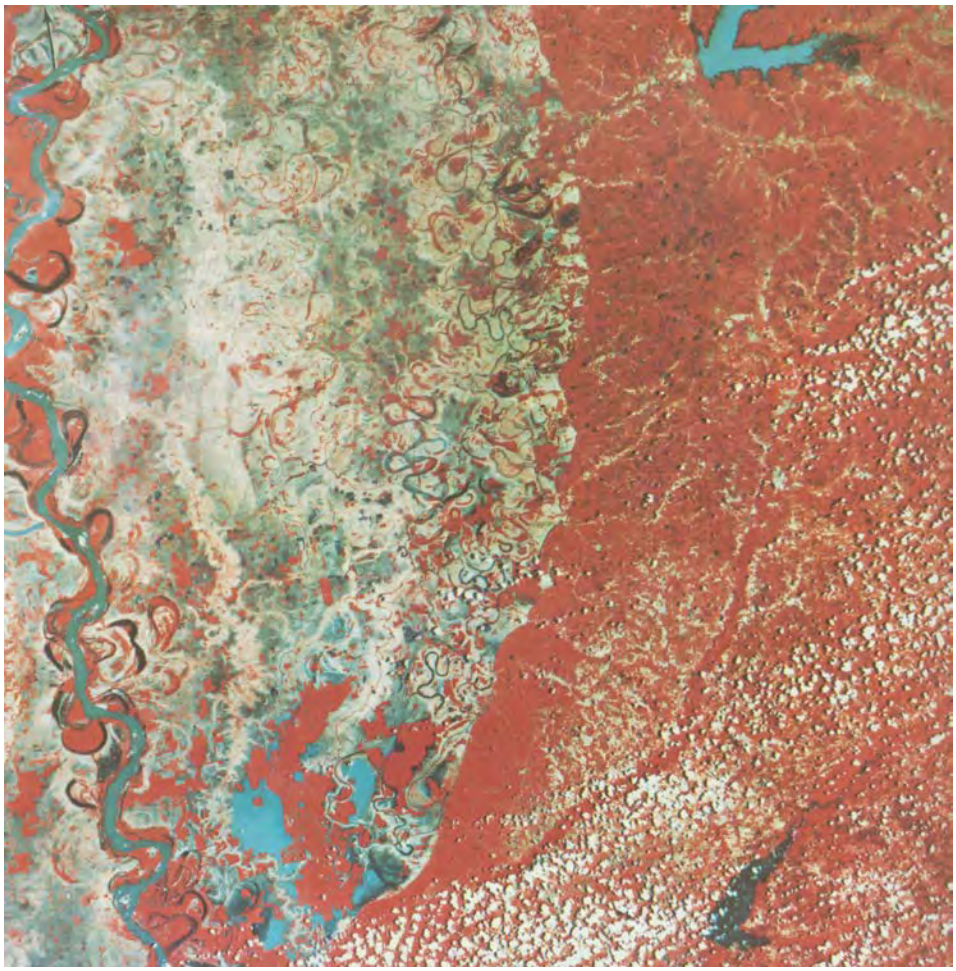
Floodplains have great importance in agriculture. The river provides water that supports farming, usually with the aid of irrigation canals. This is the case for the Amu Darya river south of the Caspian Sea; beyond its floodplain are sand-covered uplands:



This next SPOT image focuses on the Shebele River in Ethiopia. Its floodplain extends well beyond its immediate course. The bluffs in this case are steep cliffs; the uplands are true mesas (flat surfaces):

Mature river systems, with wide, flat floodplains and not too high above sealevel, over extended time periods tend to develop laterally shifting river channels - the so-called *meanders* mentioned at the beginning of this page. Especially during flooding, the swollen river is often able to move straight ahead abandoning the meandering course leaving behind a cutoff channel segment called an *oxbow lake*. Meandering is a natural process that allows a river to adjust its gradient by lengthening its course. The photo below shows a typical meandering river; note that on the left two meanders are nearly touching so that in a flood they might be joined:

Meandering is strikingly illustrated in the floodplain of the lower Mississippi River, seen below in two images. In the upper one, the floodplain edge is marked by bluffs in the state of Mississippi, on the right, with forests covering the uplands. Many oxbow lakes, some crescent-shaped, are evident in the floodplain. The silt-rich (blue) river west of the Mississippi is the Arkansas River. The lower image zeroes in on the meandering and shows now detached cutoff scars.

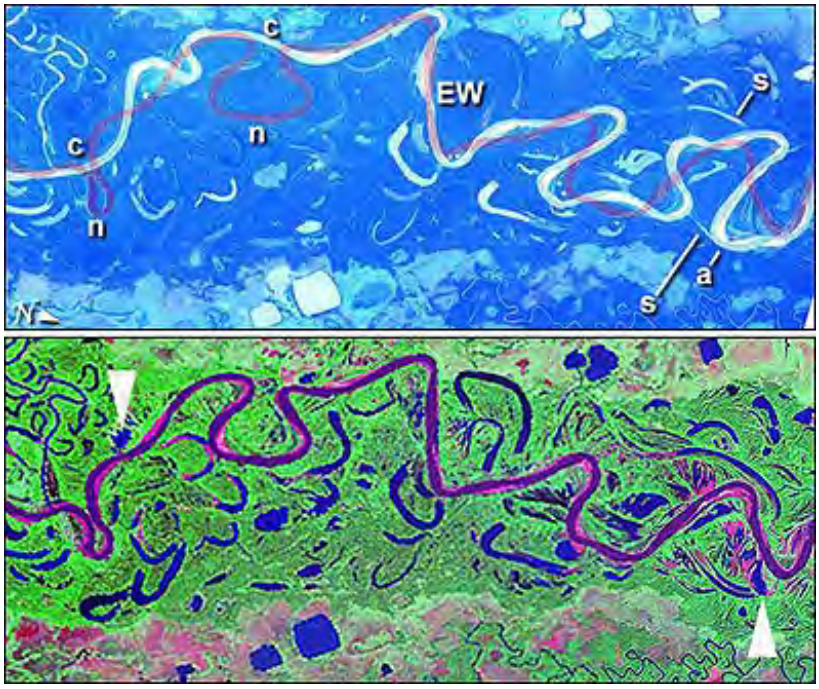


This map shows a history of meandering over a small stretch of the Mississippi River

Another river displaying numerous meanders is the Songhua River in the Manchurian plains of northern China:

The Rio Negro in South America also shows extensive meandering within its floodplain:

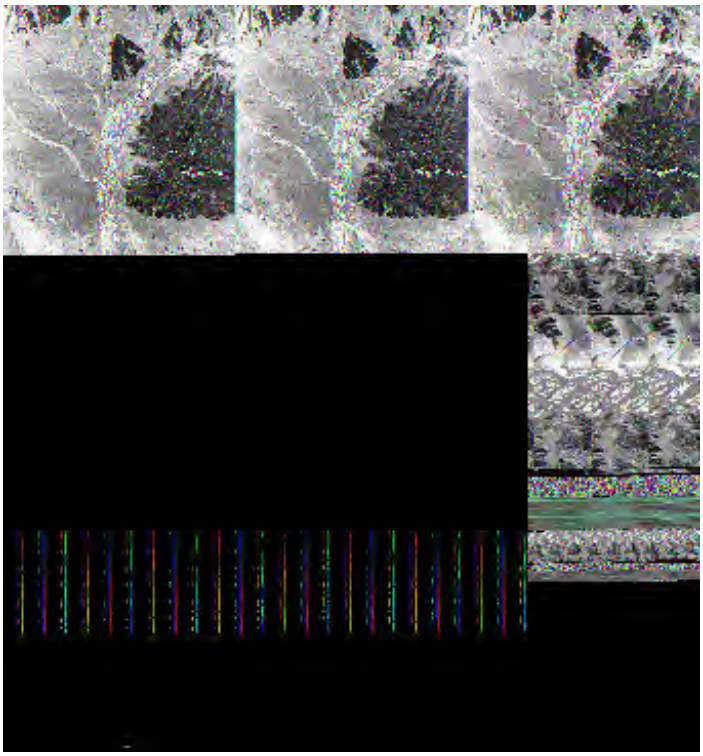
Many rivers show almost no significant changes in channel configuration (mainly by meandering) in a human lifetime but some undergo large shifts on a time scale ranging from a few years to several decades. This latter case is illustrated by the Mamore River in Bolivia northeast of the Andes. The top illustration is an astronaut photograph taken in June, 2003 from the International Space Station (ISS); the bottom is a Landsat-7 image of the same scene in 1990. The red line in the top photo marks the 1990 centerline of the river's flow superimposed on the present course. One can readily note the new meanders, several small oxbow lakes, and the disappearance of several lakes.



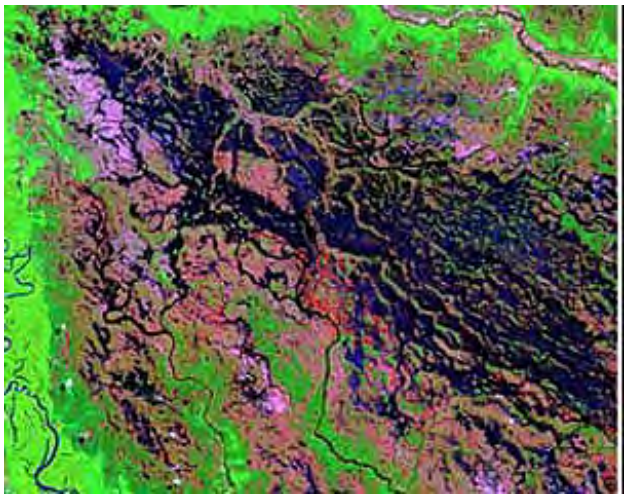
In Section 3 (page 3-7) we showed an image of the "Goosenecks" of the San Juan River in Utah. We repeat this image here with a further explanation. The San Juan is one of many rivers found now within the Colorado Plateau. Back 10 million years ago the region was much lower in elevation. Its rivers, with low gradients then, tended to meander. But the region has since experienced extensive uplift of a mile or more, at a geologically rapid rate. The streams continued to flow in their developed channels but the higher gradients caused them to increase their downcutting while maintaining their meandering. The result for some rivers was to form canyons that are curved, preserving the meandering path.

Streams laden with sediment can produce distinctive deposits in their channels and onto adjacent floodplains when conditions force deposition of their load. The continuing flow of water may thus be broken up into a network of intersecting branches producting a pattern known as a braided stream. This is well illustrated by this Eosat image of a mountain valley in Tibet, east of Lhasa, through which flows the Brahmaputra River:

This river flows south from the Himalayas through Bangladesh into the Bay of Bengal. It retains this braided appearance over most of its length, when seen in the dry season (it floods during the monsoon season), as in this Landsat image, and in more detail from SPOT:



Braided conditions represent deposition from streams choked with sediments. The resulting "islands" within the channel produce multiple segments of individual channelets within the broader master channel. A similar multichannel condition forms where a great deal of flow water (but not laden with much sediment) traverses flat lands and splits into individual channels, again separated by "islands". This occurs along several rivers in the Amazon Basin. Here is an example: the Demini River in northwest Brazil.



A similar condition can develop in deltas (see below)

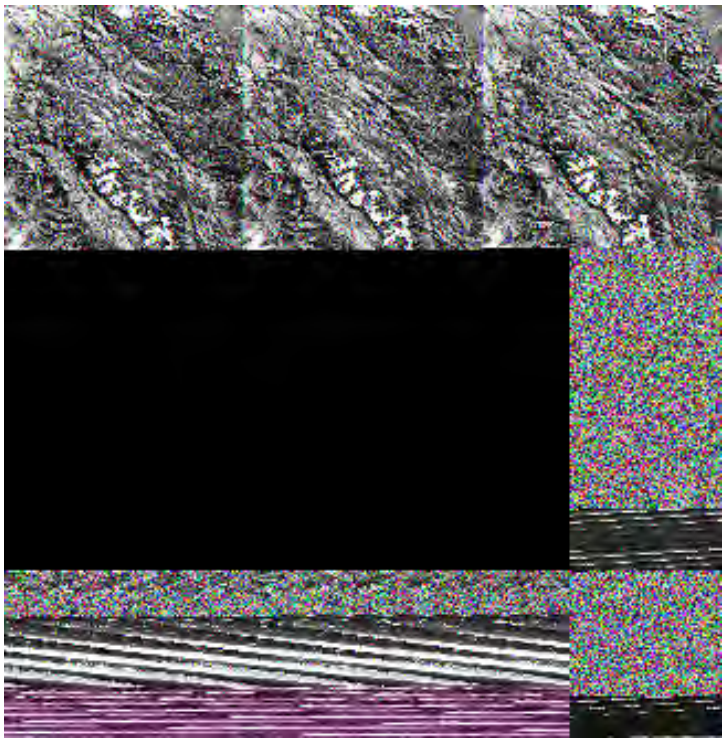
One landform associated directly with the river that makes it is a *waterfall*. This may be hard to detect from space owing to its verticality. Nevertheless, water in the stream may spread out the channel just before or just after it pitches over the drop - Victoria Falls in Africa (see below) is a good example. At the U.S-Canada border, the Niagara River, coming from Lake Erie and emptying in Lake Ontario, behaves similarly as it plunges over the regional Niagara Escarpment. Here is an Aster image of this popular tourist attraction:

Horseshoe Falls on the Canadian side (west; left) shows off its spectacular visage in this IKONOS image:

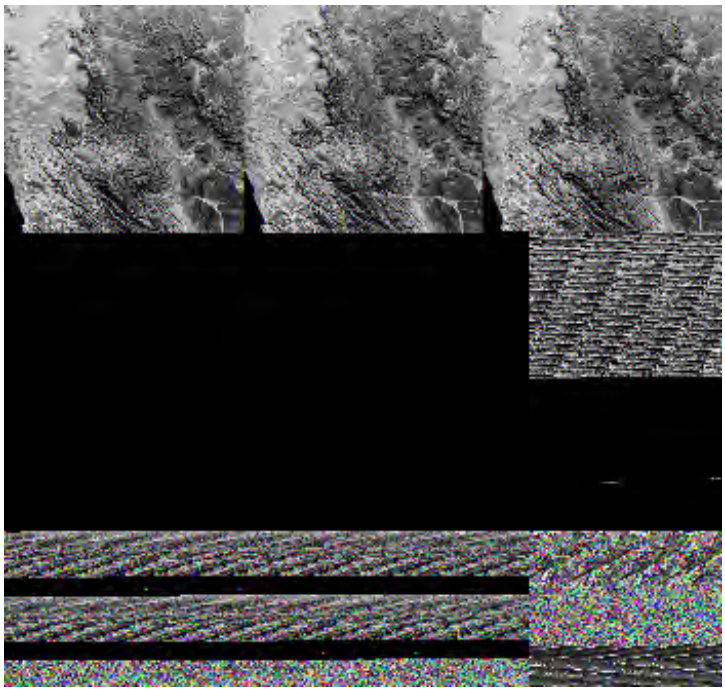
Another spectacular water cascade is the famed Victoria Falls in Zambia in central Africa. It extends for 1700 meters (more than a mile as water from the Zambezi River flows over a plateau in a broad channel (vegetation-covered) until it drops precipitously (113 m [373 feet] on average) into a savannah landscape below. Here are two views in a photo taken by an astronaut from the International Space Station.

Victoria Falls is one of the 7 Natural Wonders of the World. It is an impressive sight as seen from an airplane:

When we discuss fluvial landforms, we should remember that two aspects must be considered: first, as shown above, the drainage patterns. But, the nature and shape of the land between streams is governed (at least in part) by the combination of stream action and slope adjustments. We saw in Section 2 (thrust belts in Greece) that different tectonic units - in terms both of the rock types involved and the structural styles - will control the expression of the individual mountains or groups thereof in their gross landform character. Here is another example: in the High Cordillera of the Andes Mountains, in this Landsat image you should be able to differentiate four tectonic zones whose individuality owes much to the response of their rock types to the effects of stream erosion at high elevations:



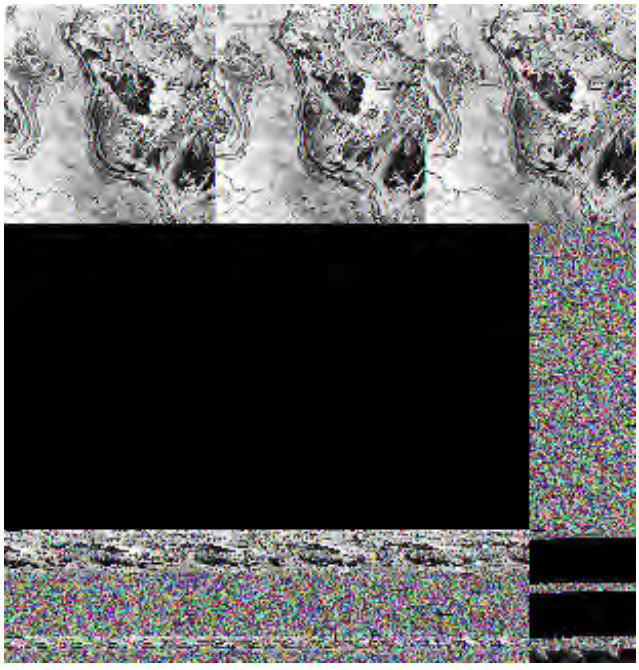
The story is different where flat-lying sedimentary rocks are involved. In 1967, M.L. King of South Africa proposed multiple planation cycles, based on his studies in southwestern Africa. These denudation surfaces occur in steps at different elevations and represent remnants of stream-eroded landscapes (similar to the peneplain concept) formed at different times. Dr. King would have been excited to see this next Landsat image, which shows in a single image four of the planation levels he proposed (the one along the coast is cut into an ancient granite surface):



Streams cause a wide variety of rock-based landforms: ridges; plateaus; mesas/buttes, canyons, etc. These landforms are often both spectacular and picturesque when developed in semi-arid and desert landscapes, as we have seen in the Southwest U.S. (pages 2-2 and 6-7.

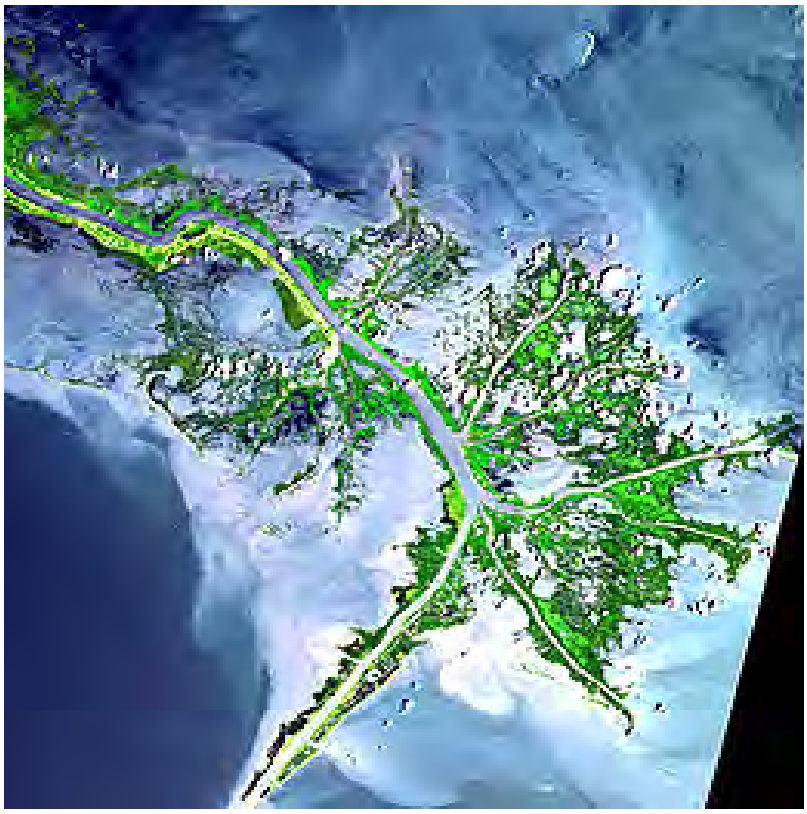
This next scene is a splendid example of a large plateau as depicted in a Landsat-1 image. Situated on the Bolivian-Brazilian border, the plateau is called Serrania de Huanchaca. The cover is dense vegetation. The plateau is almost 2000 m (6600 ft) higher than the meandering Guapore River to its east.

Below is another plateau example, this time in the Guiani Highlands of northeast Venezuela. Although not obvious in this black and white version, the region is heavily vegetated. The Rio Caroni is the major draining river. Bedrock of nearly flat sedimentary layers rests on Precambrian crystalline rocks. Over the eons the back and forth course shifts through meandering of the Rio Caroni and other rivers have led to the upper layers being eroded back, in a manner similar to the King planation surfaces, into a series of stepped plateaus and mesas (the dark ones are heavily vegetated). The highest, Auyan-Tepui (dark area near center), tops at 2950 meters (9739 ft). Over its rim spills Angel Falls, highest in the world (979 m; 3212 ft), seen looking up from its base.



One of the most unusual fluvial-caused landforms is the so-called natural bridge, made by stream undercutting and penetration through a rock unit. One of the most famous in the western U.S. is the Rainbow Bridge in the Glen Canyon National Park.

Most primary rivers end up entering large bodies of "standing" water, thereby losing velocity and hence load-carrying capability. There, a delta forms at the mouth of a stream where the load of sediment carried by running water is dumped as the stream empties into less mobile water (lake or ocean). In Section 4, we showed one example of a river delta: the bird's foot delta of the Mississippi River as it builds into the Gulf of Mexico in Louisiana, south of New Orleans. This ASTER image shows the tip of the delta in which several *distributaries* have built up deposits above sea level, giving the bird's foot effect:



As seen by Landsat, this enhanced image brings out more details in the sediments:

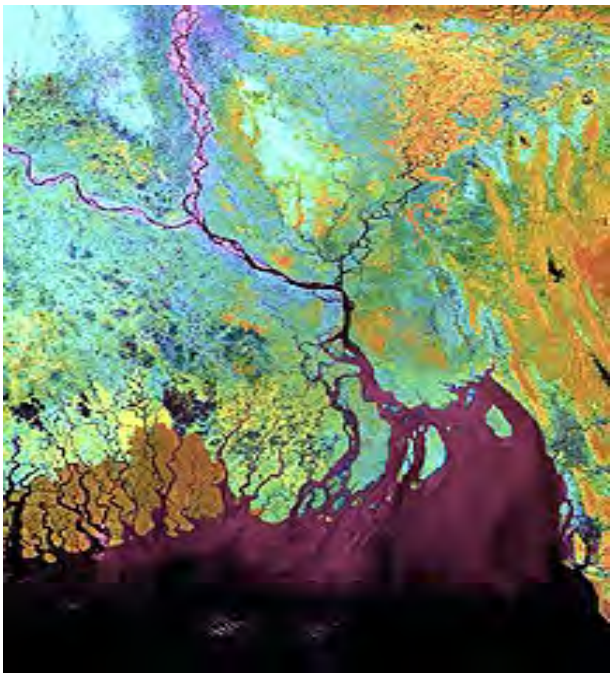
A classic example of a "bird's foot" type of delta is seen in this astronaut photo taken from the ISS while it flew across Canada. This is the delta of the Saskatchewan River made as it empties into Cedar Lake in Manitoba. The delta is in fact much smaller than the Mississippi River delta. It is made up mostly of muds and clays.

Here are six more classic examples of distributaries:

The first is the delta formed by the now combined Tigris and Euphrates Rivers as it empties in Iraq at the head of the Persian Gulf. The river is building a dominant finger into the sea, but other delta land is being created by tributaries. The river is lined with farmlands growing a variety of crops including wheat, rice, sorghum, cotton, and millet. The dark areas in the upper left are swamps. Part of Kuwait is near the bottom:

Great rivers drain from the Himalayas and flow to the Indian Ocean on either side of the Indian subcontinent. On the west the Indus River that flows through Pakistan forms a delta at the ocean, as seen in this astronaut photo:

The next Landsat image shows about half (the western part) of the Ganges Delta in Bangladesh (the India border is near the left edge, and Calcutta is at the center of the left edge). This is the world's largest delta, being more than 200 km (124 mi) in straight distance along the Bay of Bengal. The delta results from deposition of heavily silt-laden waters of the Ganges and the Brahmaputra Rivers, transporting sediment from the Himalayas far to the north. Below this image is a Landsat subscene that covers a larger area. The image has been reprocessed to convert the RGB color co-ordinates to the Intensity, Hue, Saturation (IHS) system of color expression:



The present-day Ganges drains southward just off the image to the right. That segment is now the active delta region. In both scenes above, we see the so-called abandoned delta which formed in the past when the Ganges flowed in various positions and shifted gradually eastward. Other rivers still flow into the Bay adding somewhat to the delta, as seen in the light blue (red in lower image) sediment flowing into marine waters. These quasi-distributaries become tidal channels that tidal currents highly influence. The dark red tones (brown in lower image) along the coast are mangrove forests and swamps. In the upper left quadrant, the area is part of the depauperate delta, where the clay soils now support sporadic agriculture. People seeking farmland removed much of the forest that was once there. The entire region, especially the low flat areas near the coast, is vulnerable to frequent cyclones (hurricanes) that cause widespread damage and loss of life, because of high winds and tidal surges.

Similar in appearance and ecology is the Irawwady Delta in southern Myanmar. Here it is, first in a regional setting as shown on Google and then in a Landsat-7 subscene:

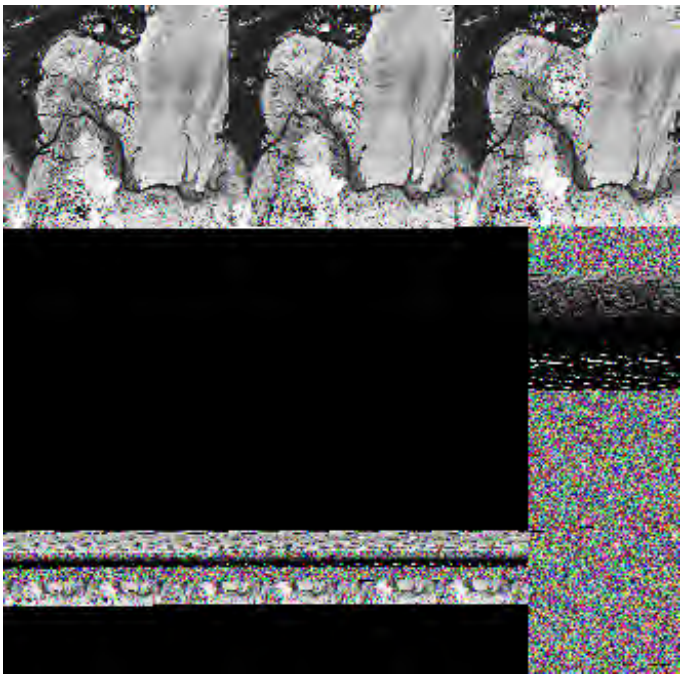
The distributary system of the largest river in Madagascar, the Betsipoka, an island nation off the east coast of the African continent shows an interesting characteristic. It drains from the Madagascar highlands which include red soils. During the rainy season, the muds in the river give it a distinctive brownish-red tone, as shown in the upper scene. In the dry season the stream is normally dark, but the land between retains the reddish mud deposits. The insets show the corresponding effects of the red muds in seawater off the coast.

March 25, 2004 (ISS008-E-19233)

September 4, 2003 (ISS007-E-14344)

Because there is sufficient water available from a river entering the ocean to support vegetation even in an arid climate, tropical forests can develop in such areas. This is the case where the Gambia river has built a delta that does not extend much into the Atlantic in southern Senegal (African West Coast), as imaged in near natural color by the MERIS sensor on ESA's Envisat.

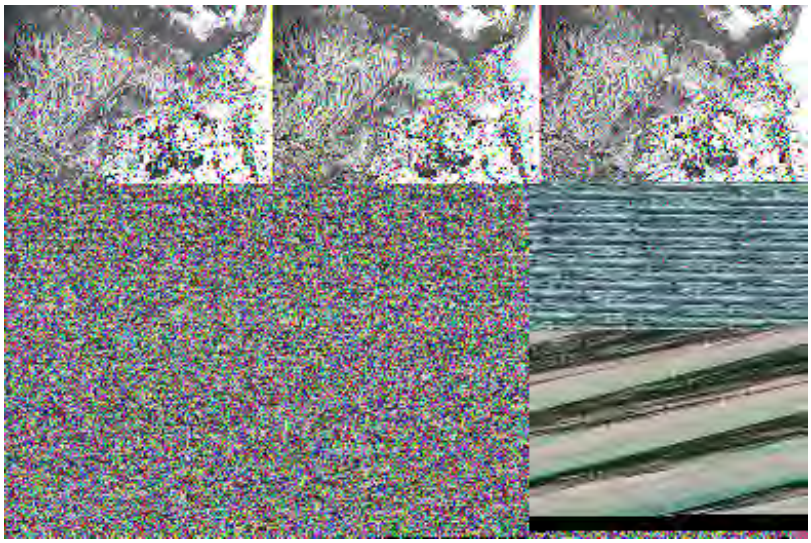
This next Landsat MSS Band 7 (IR) image covers the central west coast of Alaska. Here the Yukon River flows into Norton Sound on the northeastern Bering Sea, forming a distinctive semi-circular delta.



This color composite reveals the extent of vegetation on the Yukon Delta.

The main branch now carries sediment to the south end of a large semi-circular delta. It is actively extending the delta but an offshoot tributary is doing much of the deposition in the central part. This present delta is young (perhaps only a few thousand years since its start) that began with a major shift of the Yukon from a location not in this scene. Within and inland from the delta are numerous small lakes of ice origin. Along the coast at the bottom of the image are linear bands, which are beach ridges, developed when sea level was higher.

The limited MSS image resolution causes informative details about the geomorphology of a feature such as the Yukon Delta region to be missed. This ASTER image, covering a smaller area at finer resolution, shows the many sediment-choked distributaries within the delta proper as well as other features marking the influence of interactive glacial processes:



17-8: What is/are the main difference(s) between the Ganges and Yukon deltas?
ANSWER

Similar to the Yukon Delta is the Lena Delta in eastern Siberia formed where the Lena River empties into the Laptav Sea north of the Arctic Circle.

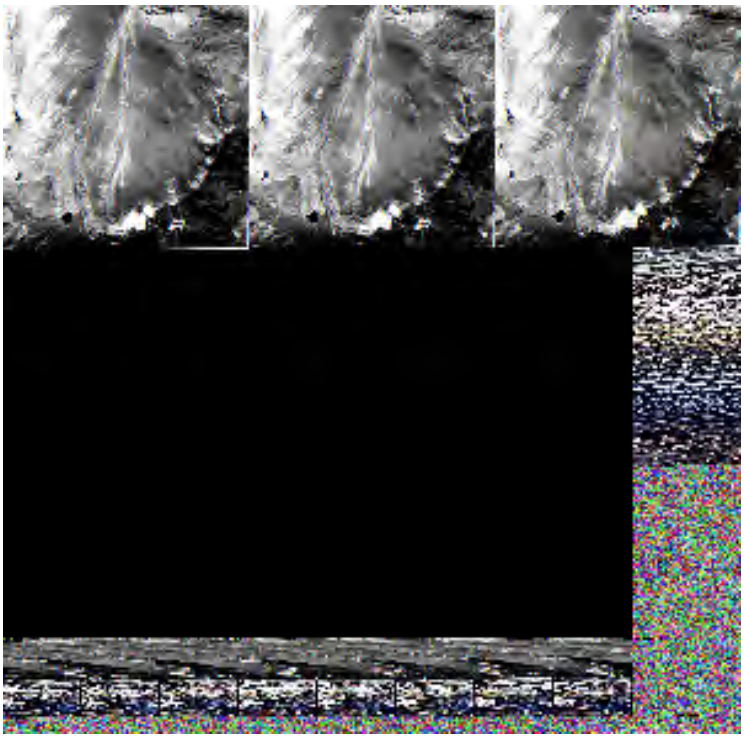
Another rather exotic delta is that of the Parana River between Uraguay and Argentina:

Rivers can produce land "deltas", widespread deposits that build up as streams carrying heavy loads down steep gradients then encounter flatlands, with low gradients that cause the load to drop and spread out as the system meanders. These are called alluvial fans. Here is an individual fan, formed off a mountain range in southern Iran; note that, as is typical, the stream splits into several distributaries. These multiple streams have carried beyond the fan where the occasional water has allowed green vegetation and agriculture to flourish.

Here is an alluvial fan in Tibet, which is adjacent to a lake:

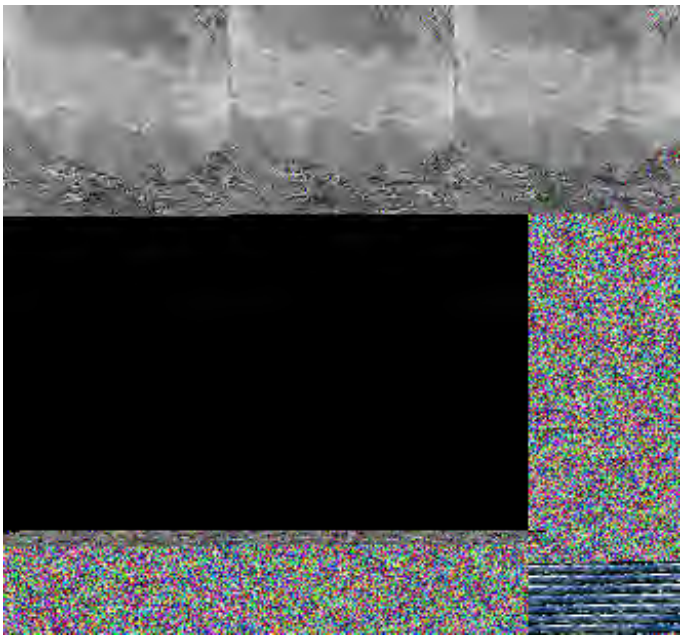
In mountainous, often semi-arid terrain, erosion cuts away the uplands and deposits the debris in the lowlands, as streams flow over pediments and fill the basin. This ASTER view of part of the Andes in Chile, a narrow chain of mountains made up of Cretaceous sedimentary rocks, shows the Altiplano that has received both sediment waste and pyroclastic fallout from volcanoes deposited as fans. Note how numerous streams seem to start at the contact between basin and mountain2 (this is a modern example of an unconformity) but actually are a continuation of uplands drainage that stands out especially in the white band facing left (a dissected pediplain, constructed from the fans).

One of the biggest alluvial fans in the world occurs in the Badain Jaran desert of east Asia, where the Ruo Shui River drains north from the Nan Shan mountains:



Rivaling that in size (and not too far away) is this fan in the southern Taklimakan Desert in Sinkiang Province, western China. This ASTER image, in near natural color, shows active (water-bearing) distributaries in blue and abandoned distributaries in black.

Alluvial fans usually extend into lowlands or basins, those commonly bounded by mountains on either side. This is typical of the Basin and Range topography of Nevada, shown on the previous page. A comparable situation occurs in the Zagros Mountains of southern Iran. In the Landsat scene below the E-W trending basin is more than 200 km in extent. Multiple fans approach the basin deposits on both the north and south sides.



We turn now to landforms that have a very limited expression, in that they are confined to narrow strips of land adjacent to water bodies such as the oceans or lakes. These are collectively known as coastal landforms. A coast is technically the strip of land (usually a beach) that is the interface between land and water; this interface's exact position varies with the tides. But in a broader sense geomorphologists include among the coastal landforms features behind this interface (the strand line) influenced by the water (e.g., lagoons, marshes; dunes, etc.); also involved because of its influence on waves and currents is the offshore shelf.

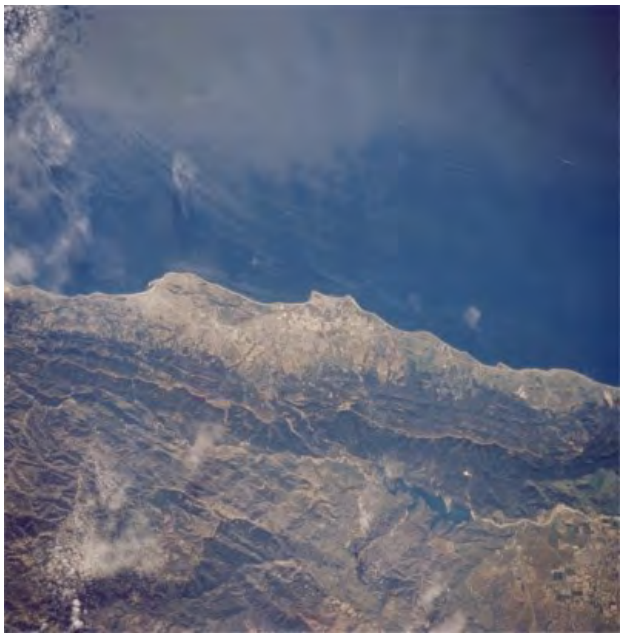
Classifications of coastlines can be tricky. One grouping is coastlines of submergence and emergence (although these terms are now held to be obsolete and misleading by some geomorphologists). This takes into account the long term effect of rising or falling sealevel (which has in the past been affected mostly by water being transferred from the oceans into masses of ice at the continental scale - lowering the general level of the oceans relative to continental heights - or melting of glacial ice - raising the general level. Another classification recognizes the interrelation between a subaerial continent's relation to the two chief components of an active plate in the plate tectonics model. On one side of a continent the ocean may extend to a spreading ridge. The continental margin in this case is called passive - the Atlantic coast of North America falls into this category. On the other side there is extensive tectonic activity associated with subduction or transform fault displacement. This is the active case: The Pacific coast, with its system of mountains adjacent to the coastline is an example.

These two photos (the first from space) give a good idea of typical Atlantic coast landforms which include low, relatively flat coastal plains, drowned river estuaries (the river mouth), coastal sand bars, and lagoons; the offshore continental shelf can extend out to sea for more than 100 km (62 miles). These views of coastal North Carolina illustrate this:

Distinct land forms are associated with Atlantic coast types of landforms, such as those illustrated in this diagram:

However, under the right circumstances cliffs can be produced in otherwise flat coastlines not involved in mountain building. In this photo, the horizontal Cretaceous chalk beds that make up the White Cliffs of Dover show such a setting:

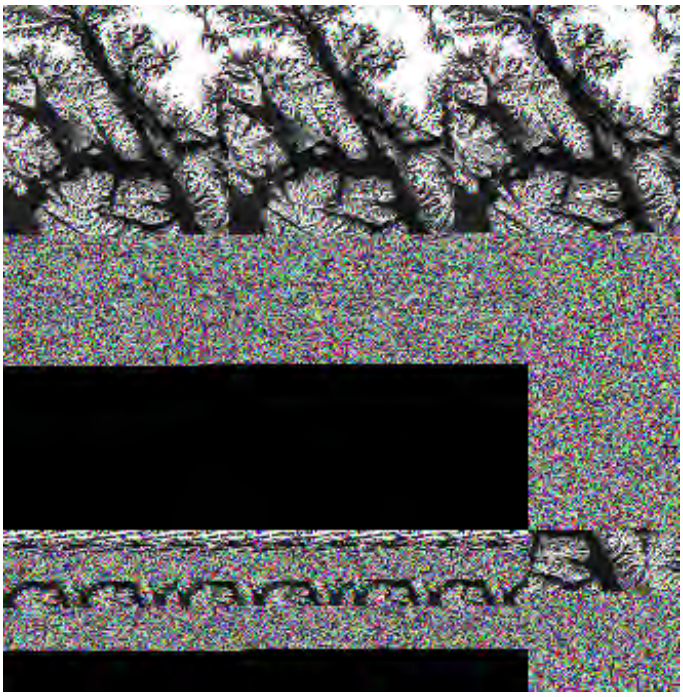
Pacific coast landforms are characterized by (often) irregular coastlines, embayments, narrow beaches, often rugged topography (such as mountains) on the land side, and rapid deepening of waters in the narrow continental shelf. These views are examples.



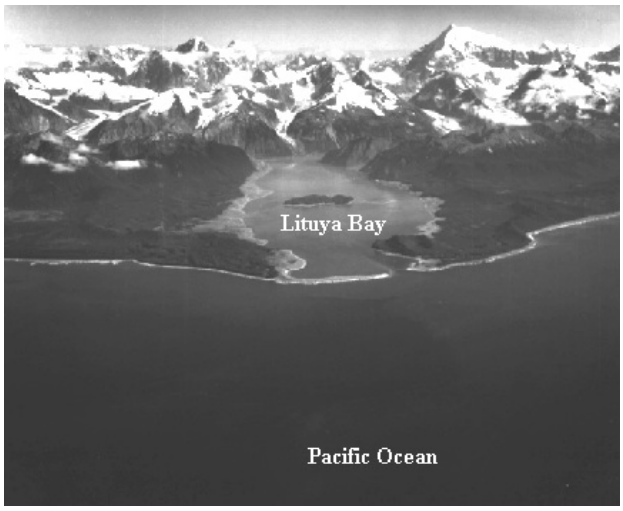
Where the mountains are actually encroached upon by the ocean, a rugged and irregular coastline is the norm, as shown here.

In between the mountains are narrow beaches:

Space images don't always do justice to the coastal landforms they may contain. This is due in part to the fact that the landforms tend to be concentrated along thin linear strips rather than spread over much of the images. Still, as the examples that follow illustrate, much of the larger scale features of this group of landforms can be expressed in images. We will start with more images along the western coast of North America.



The above scene lies in the Coastal Ranges along the Pacific Ocean in the region where the Alaskan Panhandle extends along Canada (near the top of the image). Juneau, Alaska's capital, is near the center. The region is tectonically active, with major faults separating individual crustal units known as terranes (see below). These faults and other structural features served as lines of weakness for erosional attack by streams and glaciers, which together carved out deep valleys. Some present day glaciers are visible in the Glacier Bay National Park area northeast of Juneau and elsewhere. Narrow patches of flat land almost at sealevel can develop midst the high ranges that meet the waters, as shown here:

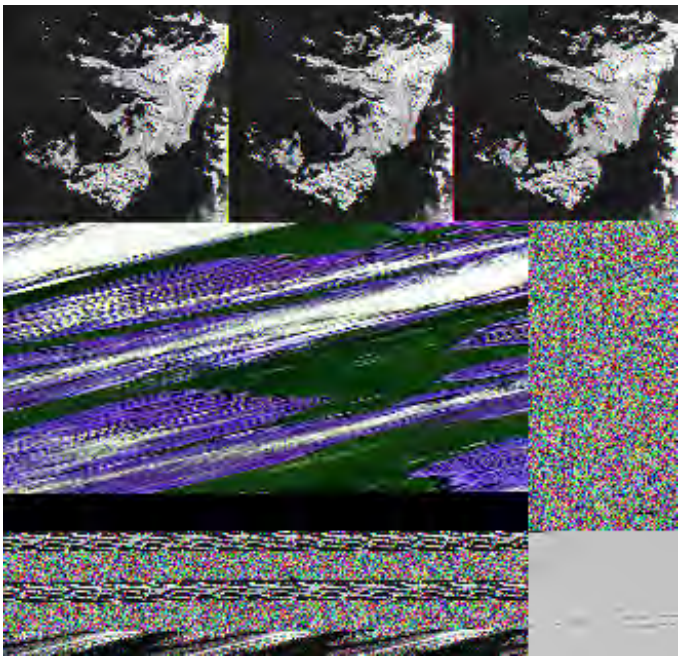


After the close of the last major glaciation, melting glaciers are now in retreat to the extent that, as sea level has been rising, the ocean flowed into some large valleys cut earlier to below sea level by the ice, effectively drowning them. The resulting landform is a *fjord*–a Norwegian name assigned to submerged coastal valleys once occupied by ice.

17-9: How do you think the scenery shown in the above image has strongly influenced the economy of the region? ANSWER

Fjords abound along the coastline of Iceland. Here is a peninsula in the northwest that shows glacially-sculpted valleys now submerged as sealevel rose following the last Pleistocene glaciation.

Another example of a rugged, embayed coastline, shaped in part by drainage off higher land, and now influenced by the rise of sea level, is that of the West Falkland Island in the South Atlantic. A bit of the East Falkland Island is at the right edge of this Landsat subscene.



The Atlantic seaboard, seen below, is generally now a coast of emergence associated with regional uplift. Over the past 50 million years or so, seas have lapped well onto the eastern North American continent, laying down thick, subhorizontal sedimentary layers, but the ocean has been gradually retreating eastward. In the last few thousand years, a rise in sea level, resulting from glacial ice melt, has reversed this trend as marine waters drown coastal valleys (e.g., Chesapeake Bay) and push shorelines inland. Along much of the Atlantic coast from New Jersey to Florida, thin narrow lines of sand deposits, built up above sea level by deposits from ocean waters encroaching on shallow bottom slopes, form barrier islands. The image is a photo taken by an Apollo 9 astronaut of the famed Outer Banks of North Carolina. The point farthest east is Cape Hatteras, and the southern point is Cape Lookout. The wide stretch of water towards the mainland is Pamlico Sound, which, with Albemarle Sound inland to the north, we term a lagoon. Offshore, submerged sandbars–incipient islands–form hazards to shipping. Because of the irregular, cuspate coastline west of the barrier, geomorphologists argue that the island had already formed prior to current onlap by ocean waters, thus protecting the inner shores from wave erosion.

17-10: Why is it risky to live on the Outer Banks? ANSWER

Much of the Eastern United States, as well as other parts of the world, where topography is low

and flat, are becoming coastlines of submergence, as sea level slowly rises with the current melting of glaciers, sea ice, and continental ice sheets. The Chesapeake Bay (drowned Susquehanna River) in Maryland and Virginia and the Delaware Bay south of Philadelphia are classic example, as seen together in this MODIS image.

The Chesapeake and Delaware Bays are both examples of estuaries. When rivers meet the ocean or large lakes they either form estuaries or deltas. We saw examples of deltas earlier on this page. Let us now consider marine estuaries, which usually consist of brackish water (salty ocean water mixed with fresh river water). Several examples are shown next, with their special characteristics described in the captions:

Previously on this page we saw the region where the Tigris-Euphrates River(s) empty into the Persian Gulf, as an example of a delta. This area is shown again below to illustrate with labels the multiple coastal characteristics that include estuarine deposits:

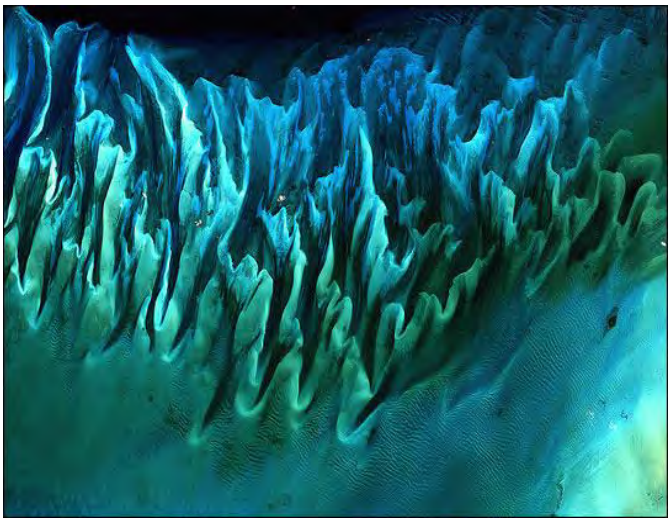
One of the more striking features found off coastlines are tropical islands or *atolls* built around a central emergent landmass, and fringed by, reefs. One extreme form of an atoll is the type that has just a single, narrow band of coral reef, enclosing a lagoon that comprises almost all the area of the island. The top of the submarine land mass (volcanic) is now completely submerged. The Arno Reef in the Marshall Islands and the Oeno atoll are two prime examples:

One of the classic "Paradise" islands in the Pacific Ocean is Bora Bora, 240 km 150 miles) northwest of Tahiti, in the Society Islands of French Polynesia. This is a narrow reef built out from a central island of volcanic origin, separated by a sparkling lagoon (the blue-green portion is very shallow water covering a white coral limestone floor). More than 4500 people live on this island, which is one of the most popular resort destinations in the South Pacific. First look at this Quickbird image of the whole island:

Because of the special beauty and romance of this type of oceanic island, we show two aerial oblique views of Bora Bora:

Close to the U.S, the best known of these are the Grand Bahamas, imaged here by SeaWiFS. Both Florida and Cuba are included. The light blue-green color is close to true, owing to the presence of stirred up calcium carbonate muds derived from the breakup of corals. No wonder the Bahamas are so popular - this looks like Paradise!

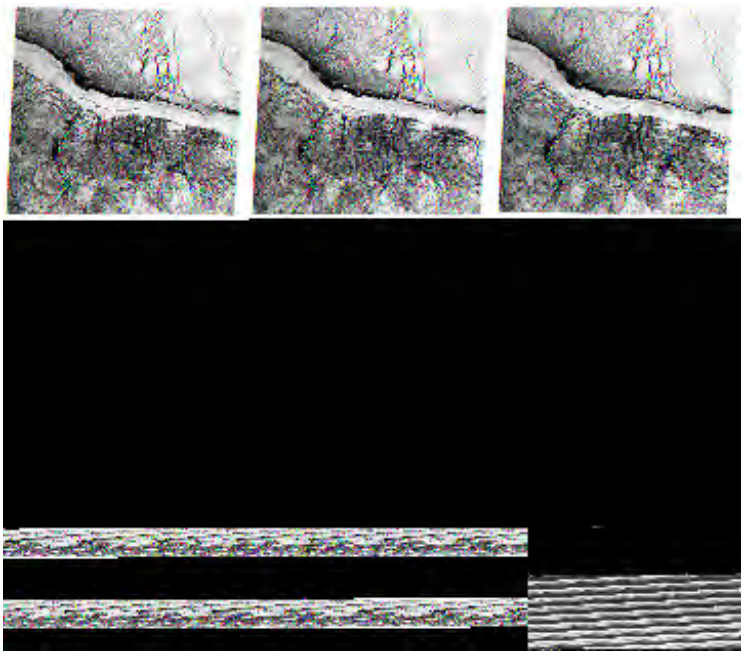
The white beaches that make the Bahamas so attractive to winter visitors are made up of carbonate sands derived mainly from precipitation of the $CaCO_3$ by organisms. The ocean on the leeward side of the Bahamas is quite shallow, with clear water. These submerged sand shoals are easily visible from above, as shown in this Landsat-7 ETM+ image:



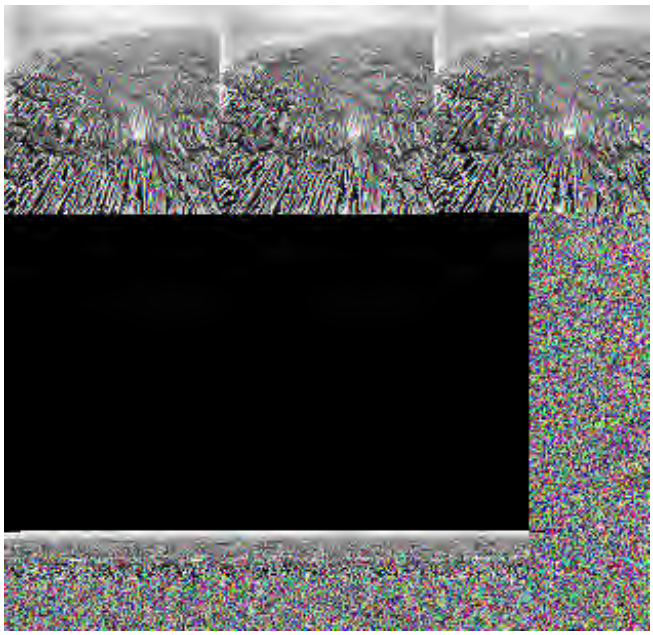
Probably the best known reef complex in the world is the Great Barrier Reef off the northeast coast of Australia. It is a very popular destination for "snorkelers" and scientists studying the

habitats of underseas life. Here is a view made by Terra's MISR of much of this chain of coral islands.

Since the melting of the last glacial icecap that covered northern North America, parts of the Canadian Shield have been gradually rising, owing to the principle of Isostasy (rebound of depressed land after load removal to maintain gravitational equilibrium). But the ocean levels have also risen and fallen depending on the amount of ice locked up in glaciers and ice caps. Thus, at earlier times during the last few million years of off-again/on-again glacial episodes, shorelines have been established at various higher levels as sealevel fluctuated. The shifts in water level and the depressions and rebounds have combined to produce stranded deposits and erosive features on the present land surface In the Hudson Bay-James Bay region this is marked by successive shore line beaches and ridges, each indicating the position for a time of the water's edge, long enough to establish deposits. In this area, the set of higher shorelines is probably more the effect of rebound than of water level rise..The resulting beach lines are quite evident in this Landsat image:



An aerial oblique photo defines the succession of ridge/beach lines in more detail:



One of the consequences of continental glaciation is that sea water becomes converted to great masses of ice on both continents and open seas. Sea levels fall when this water is withdrawn. They rise again during interglacial melting. As we saw around Hudson Bay, this can give rise to terraces and cliffs. Areas away from the glacial ice, such as near the equator, are susceptible to sea level fluctuations. This is nicely illustrated by the several terraces on Isla Blonquilla, offshore from Venezuela, in the Caribbean:

Traces of ancient strand (beach) lines owing primarily to sealevel changes, but also often influenced by tectonic or isostatic adjustments, are found throughout the world. This next image indicates a succession of shorelines masked partly by vegetation in the Mosquito Coast (so-named from the Moskito River rather than a surfeit of mosquitos) along the Gulf of Mexico where Honduras and Nicaragua come together at the Coco River:

Three of the next four landform groups are also tied in with water as a prime formative agent.

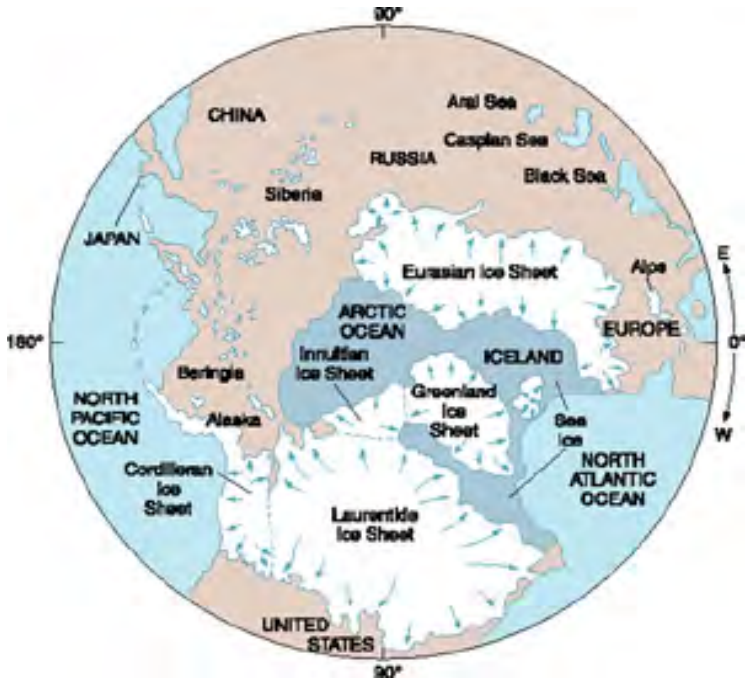
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Great accumulations of ice flow under their own weight, carving into previous stream courses in mountainous terrain to make Valley or Piedmont glaciers. Even thicker ice can cut away at the regional surface as continental-scale ice sheets spread across great distances. Parts of continents subjected to Pleistocene glaciation also have distinctive landforms due to deposition of debris eroded by these ice masses.

Glacial landforms

Now on to the main snow- and ice-generated features that are associated with glaciers. In the last 2 million years, ice centers in high latitudes and in high mountains several times have undergone expansion and, particularly in the northern hemisphere, ice miles thick moved from the polar regions into Siberia, northern Europe, and in the U.S as far south as Cincinnati and into the northern Great Plains. This map summarizes the distribution of continental ice coverage in the northern hemisphere at the height of the Pleistocene age:



Before proceeding, you have the option of visiting these Wikipedia websites that cover the topics Glaciers and http://en.wikipedia.org/wiki/Ice_age Ice Age.

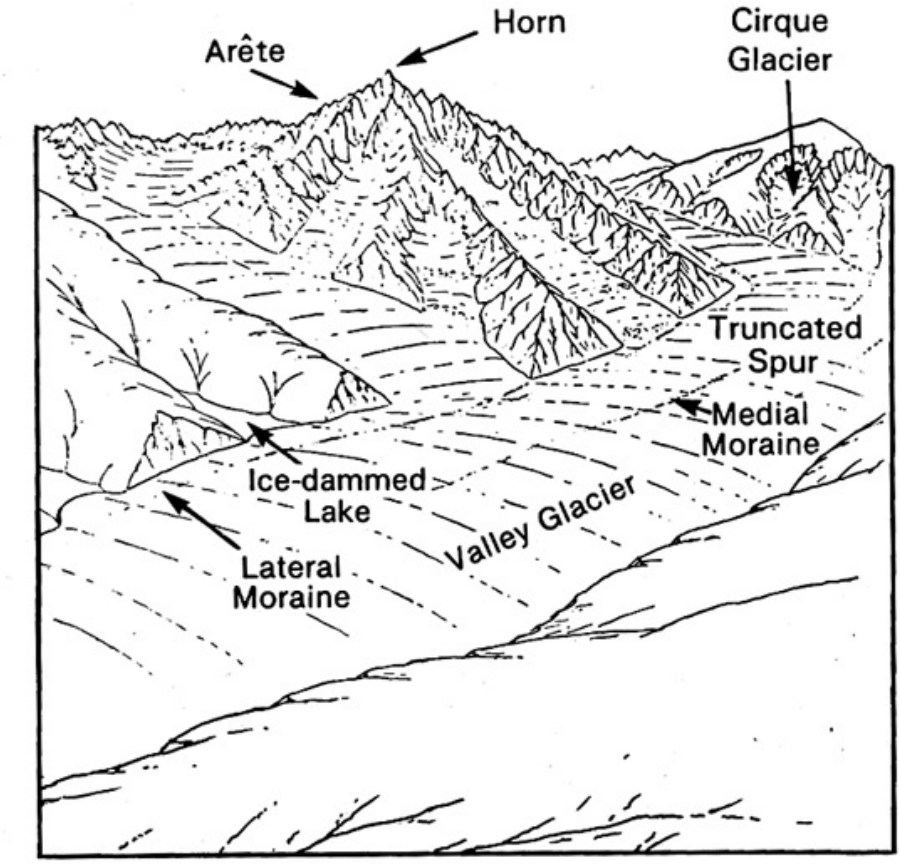
The last major advance of Pleistocene continental glaciers over North America is the so-called Wisconsin event, shown in this map:

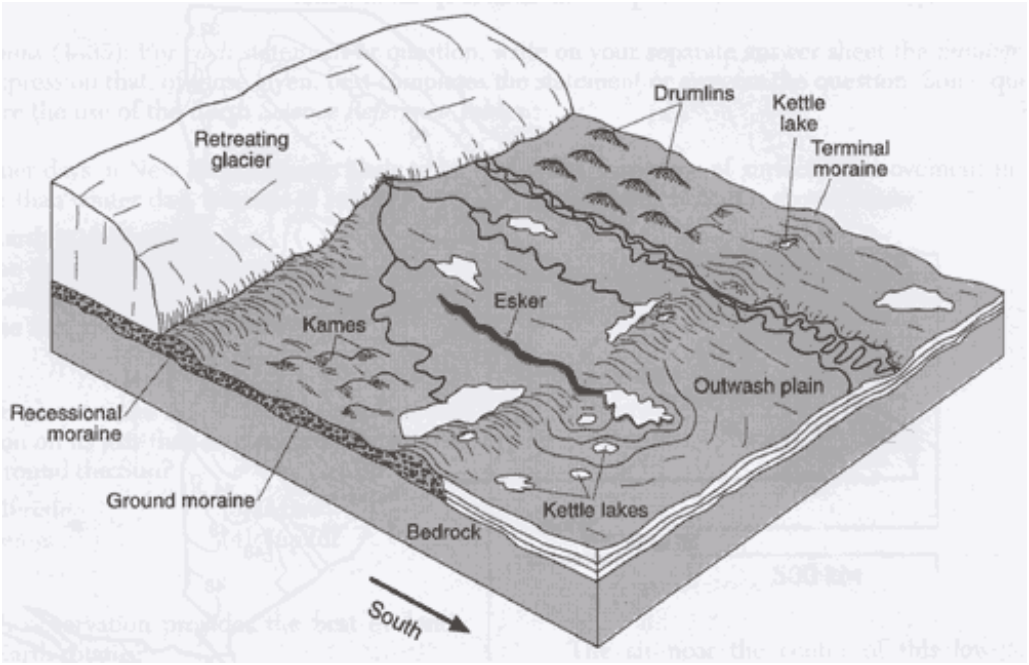
In North America, there have been several (perhaps 6) significant advances and retreats (stages) during the Pleistocene. Generally, the last advance (Wisconsin) overrode glacial deposits and landforms of the earlier glaciations, wiping them out in some places. This excerpt from Wikipedia summarizes the current status of the nomenclature used to specify these stages: "The major glacial stages of the current ice age in North America are the Illinoian, Sangamonian and Wisconsin stages. The use of the Nebraskan, Afton, Kansan, and Yarmouthian (Yarmouth) stages to subdivide the ice age in North America have been discontinued by Quaternary geologists and geomorphologists. Those stages have all been merged into the Pre-Illinoian Stage in the 1980s."

Glacial landforms are widespread in the northern hemisphere, being found in lowlands of Europe, parts of Asia, and North America, all of which were visited by continental glaciers in the Pleistocene. Likewise, glaciers are found outside these continental regions wherever mountain conditions permit extended cold climates.

Glaciers appear wherever it is cold enough year round for periods of extensive snowfall. As the snow compacts, pressure aids in its melting and refreezing as ice. Almost all modern day glaciers appear either as extensive ice sheets in the cold high latitudes of both global hemispheres or more localized where mountains rise to heights that remain cold much of the year (glaciers occur near the top of Mount Kilimanjaro in Kenya near the Equator).

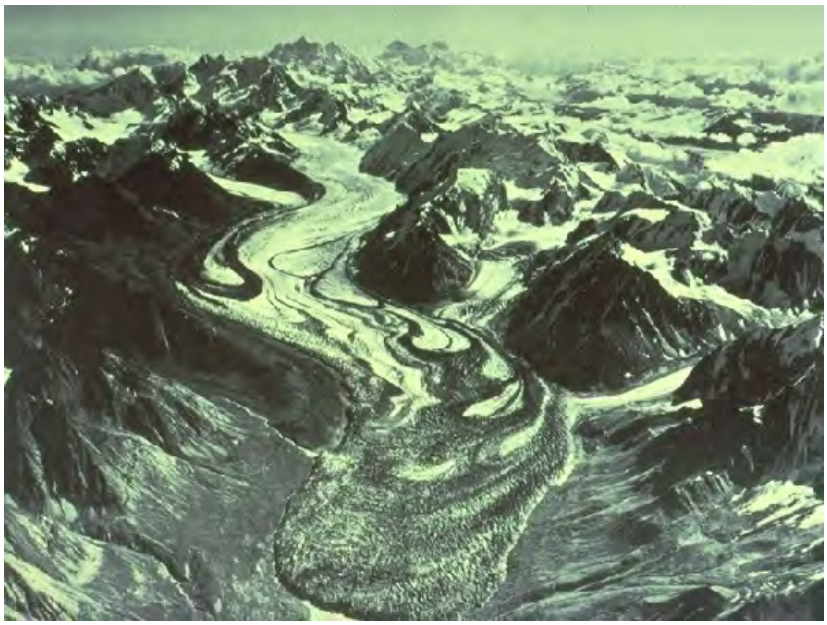
There are many specific landforms associated with both mountain and continental glaciation. Some are detectable in satellite imagery. Others are hard to recognize except in high resolution images or aerial photos. The next two illustrations depict most of the common landforms (some have French names, since Louis Agassiz and other pioneer glaciologists were frenchmen); they are taken from Geomorphology from Space:



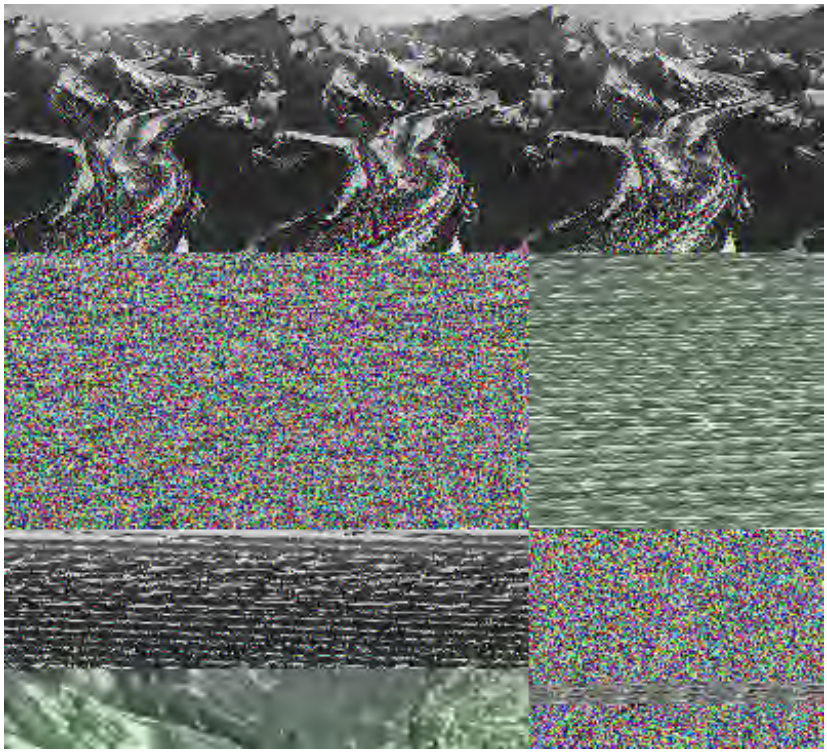


Before we embark on a series of satellite images showing glaciers and glacial terrain, lets illustrate some of the above landforms depicted in the preceding diagrams as they are visible from the ground or air. Some are erosional while others are depositional; some occur mainly/only in mountain glaciers while others are more typical of continental glaciers.

This ground view is typical of a mountain glacier, with its medial and terminal moraines (see below), and the jagged topography of adjacent mountains, all brought about by glaciation:



Glaciers tend to have black streaks on their surfaces - these are medial moraine deposits formed as rock falls loose at the head of the glacier and continues this process over the years, so that the debris are strung out as streaks that move over the years. Here is an aerial oblique view that shows this morainal pattern in the Steele Glacier of Alaska:

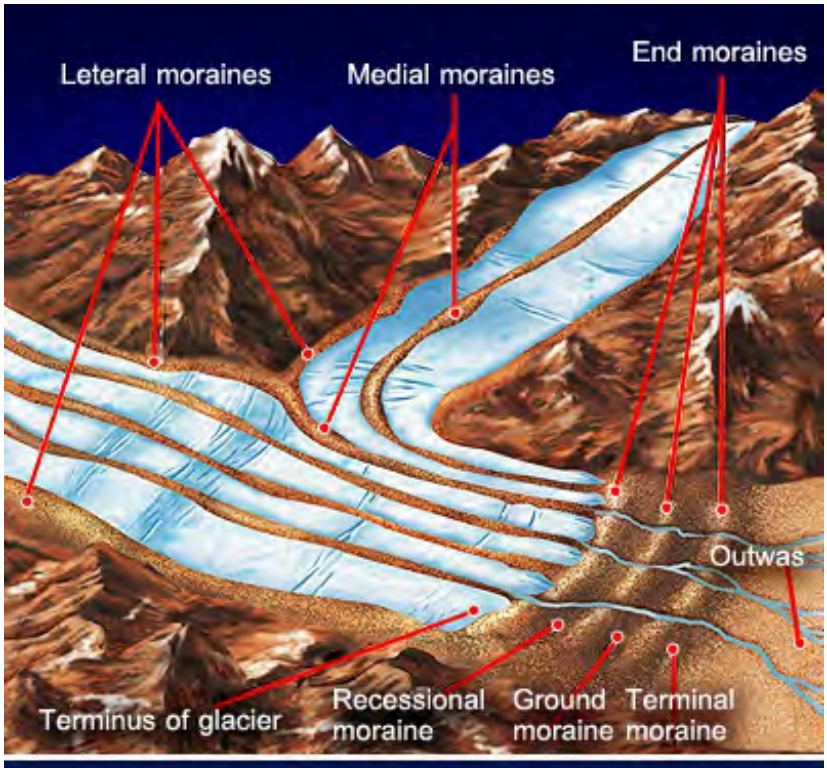


In the upper reaches of mountain glaciers, the ice plucks away at the bedrock. At the head (top) of a glacier, an amphitheaterlike erosional depression called a cirque is produced.

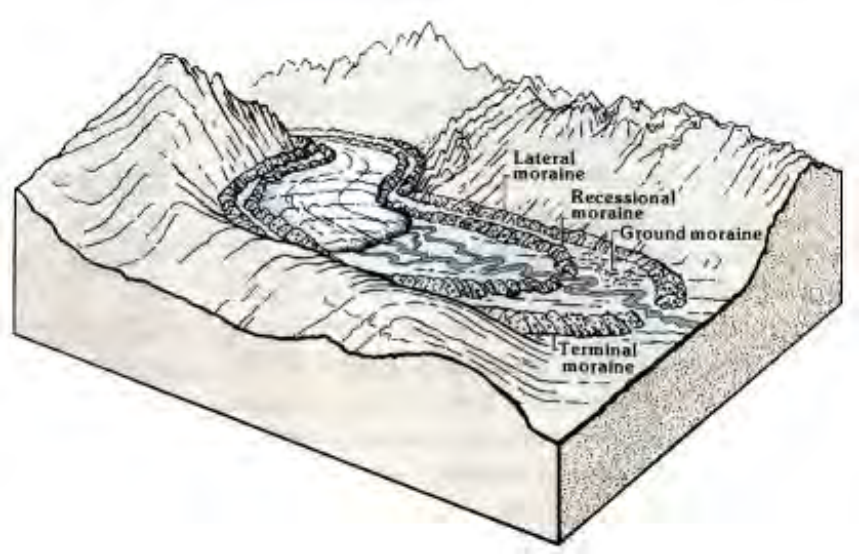
Where multiple cirques coalesce, knife edgelike ridges called aretes are the result:

Aretes and horns in the Grand Tetons.

Glacial ice is a powerful agent of erosion. At the contact surface where the glacier meets the bedrock of the mountains it occupies, water from melt mixes with the rock and plucks away, grabbing the rock and incorporating it in the ice. Other rock gets on the surface by means of avalanches or gravitational weathering. Deposits of the incorporated rock make up "**moraines**", a major feature of glaciers - both mountain and continental. These deposits are usually dark (most rocks are nonwhite) and stand out in contrast to the white tones of the ice. Some rock is located at the glacier's base (*basal moraine*); some rock is found within the rock (*englacial*); some rock is built up along a glacier's sides (*lateral*); some rock is collected on the glacier's surface. Where glaciers meet, surface deposits coalesce to form *medial* moraines. As glaciers melt from the limit of their farthest advance, they produce *terminal* (alternately called *end*) moraines; as they retreat they leave behind morainal deposits that are spread out as *ground* or *recessional* moraines. Two general terms refer to unsorted, heterogenous glacial deposits: **drift** and **till**.



Types of Moraines



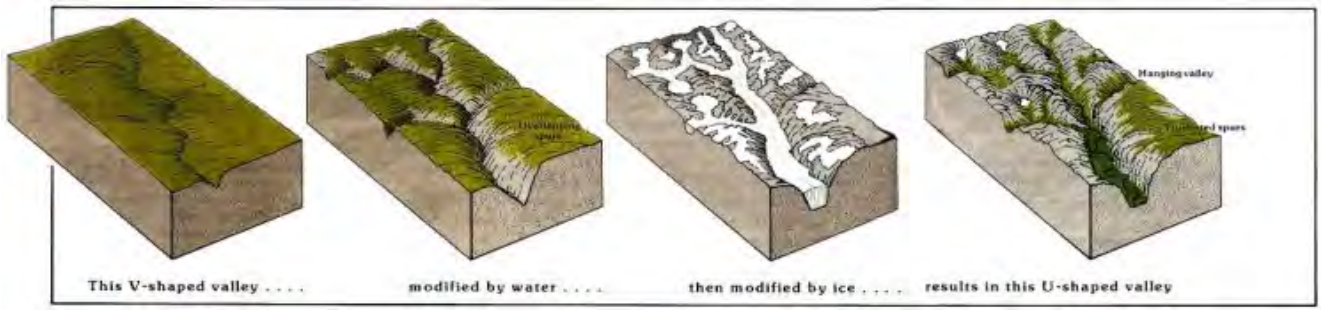
Next, consider these photos that focus on the deposits of unsorted loose rubble (with a wide range of sizes) that comprise moraines

Hummocky unvegetated morainal deposits.

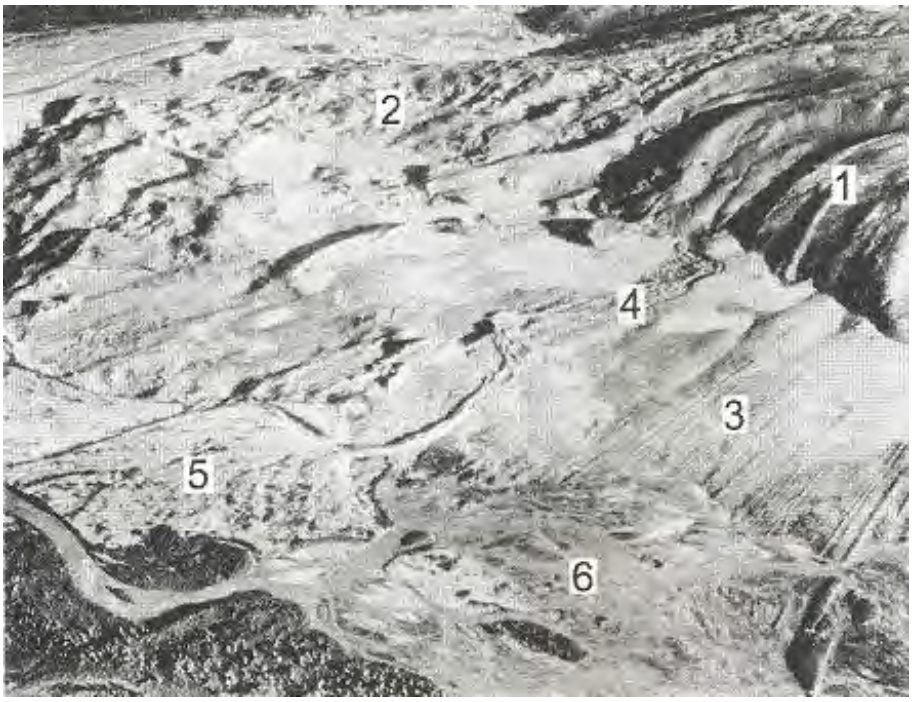
Moraines can become vegetated as soils develop on them. Here is a ridge made up of a terminal moraine deposit.

What does a landscape once dominated by valley glaciers and ice sheets look like when the ice cover has melted? In mountainous terrain, a glaciated valley is both deepened and widened, giving rise to a characteristic U-shaped cross-section:

The sequence of stages that end in a U-shaped valley after the glacial ice has melted is shown in this diagram:



In the lowlands beyond a retreating mountain glacier, some of the landforms that develop appear in this oblique aerial photo (note: several also are characteristic of continental glaciation):



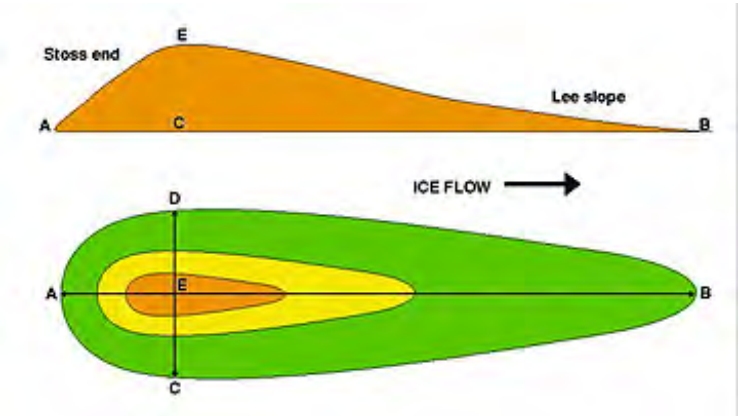
This photo shows how each landform is interrelated to one another through glaciation. (1) Glacier, (2) ice cored hummocky end moraine, (3) fluted till, (4) esker, (5) fan and (6) braided streams. (Martini, 2001; Canadian Landform Inventory Project)

One of the more expansive landforms is the outwash plain, seen in this example from Alaska; beneath it is a braided stream within an outwash plain:

Eskers form when a subglacial ice tunnel is filled to choking with rock debris, followed by melting of the glacial ice leaving sinuous low hills that are distinctive. Here are two examples:

Drumlins are elongated hills of glacial till molded by ice sheets as these paste the subglacial load against obstructions. They have distinctive shapes, as indicated in the aerial photo of a drumlin field in Manitoba:

This cross-section and plan view of a drumlin show its streamlined shape; compare this with an aerial shot of a single drumlin:



A drumlin

Another aerial view of a swarm of aligned drumlins emphasizes the distinctive shape of this intriguing glacial landform.

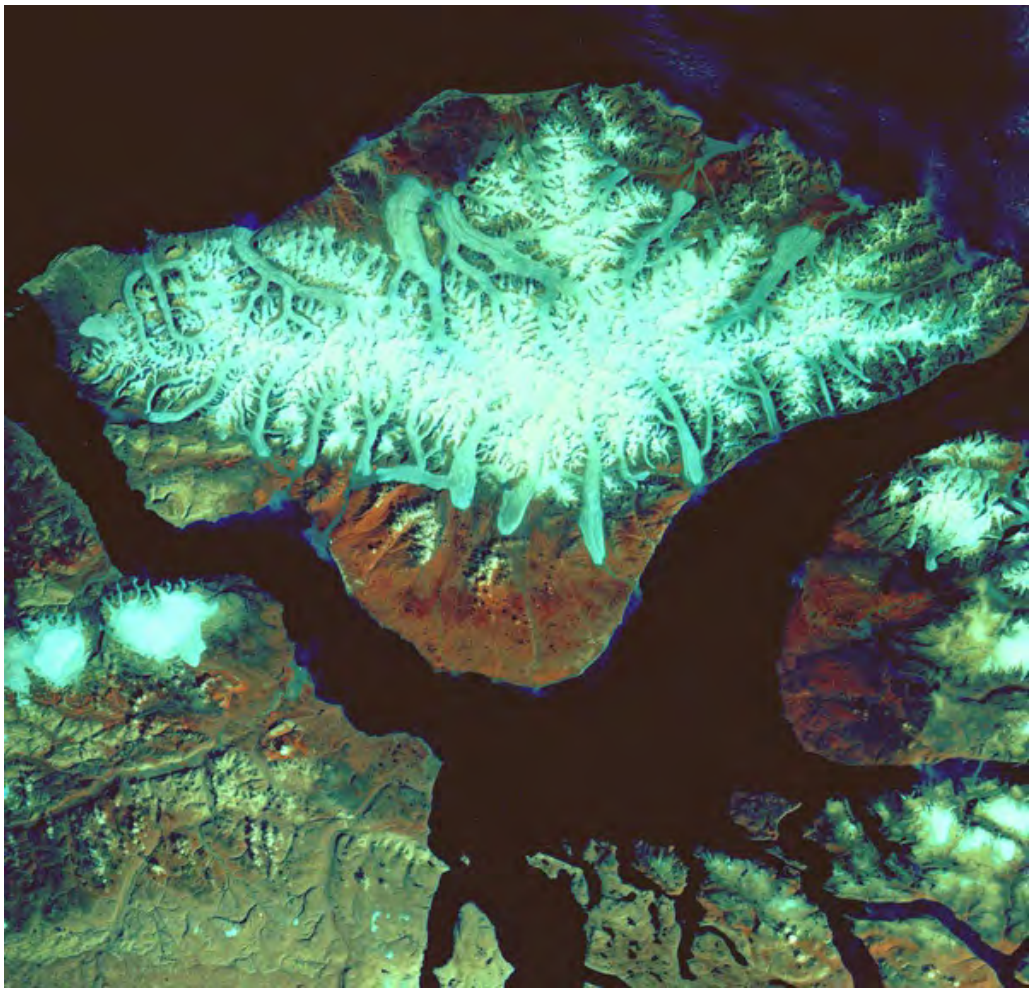
Kames form as till is washed into holes in retreating ice. Low, moundlike hills result:

Kettles are depressions in till plains that form where large chunks of ice melt, leaving a "hole" in which water may fill into a lake.

Now, on to examples of glacial features as seen from space, from the air, and the ground. Glaciation is commonplace in cool regions of the globe, at high mountain elevations and in high (polar) latitudes. Ice is conspicuous from space because of its whiteness and high reflectivity (albedo); many of the landforms ice movements produce are also distinctive. Ice can cover nearly an entire continent, such as Antarctica, or a huge island such as Greenland. When ice begins to flow by gravity to lower elevations, usually along distinct, often narrow paths, it becomes a glacier. Here are an example of ice accumulation and glacial movement in an icecap in the south of Iceland:

The above image is a special, false-color image, using yellow, red, and blue filters, to enhance various types of ice and glacial zones. It shows the thick ice cap (defined as a dome-shaped ice mass with radial flow) known as Vatnajökull, in south-central Iceland which has 43 outlet glaciers (light blue), many with lobate termini. Areas in yellow-orange are vegetated, while reds associate with basaltic rocks and sparse vegetation. Green in the ocean is sediment, and black around the ice cap is a zone of ground soaking from glacial meltwater.

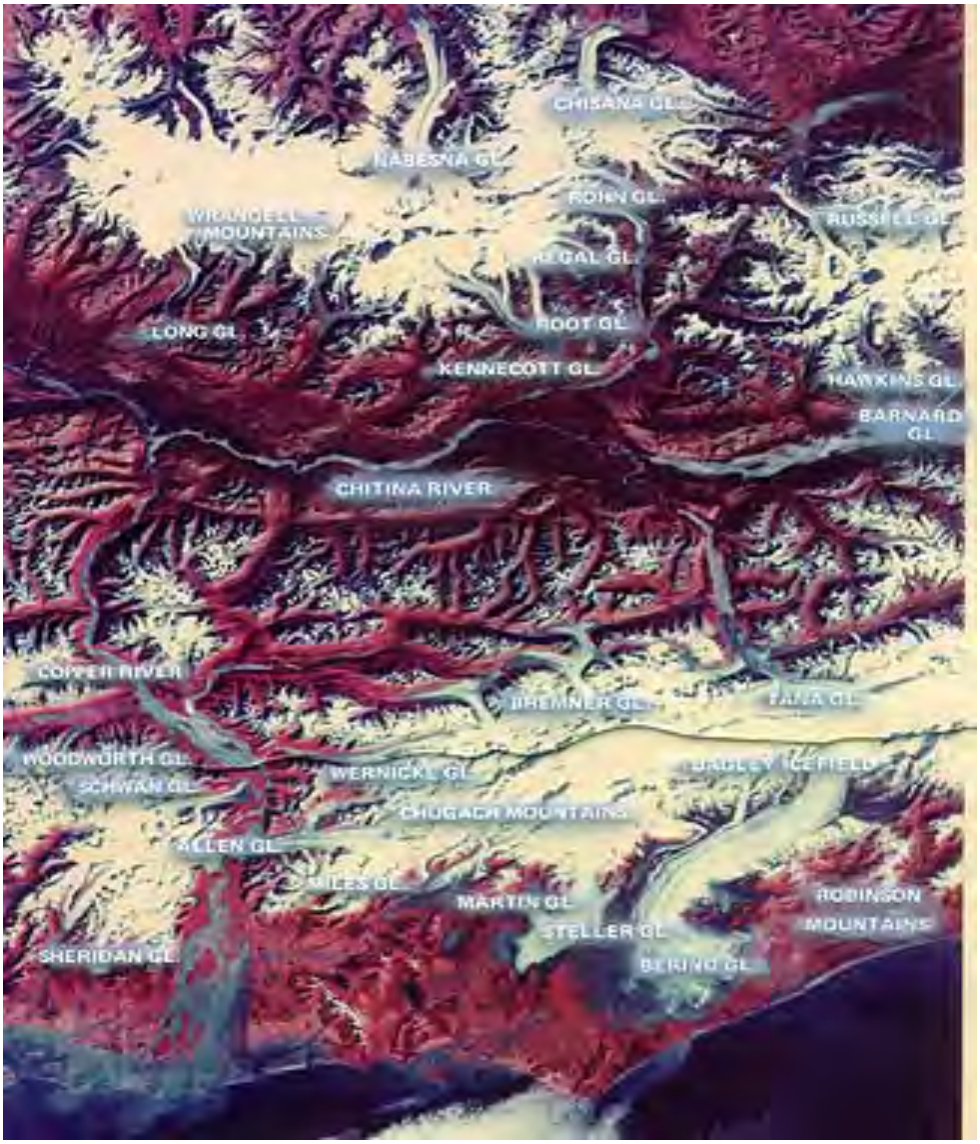
On a smaller scale, Bylot Island in the Arctic Sea of northern Canada is capped by a small permanent icefield, from which mountain glaciers are flowing downslope on both the north and south sides:



These glaciers have been shrinking steadily in the last 40 years. Here is an aerial oblique view of several:

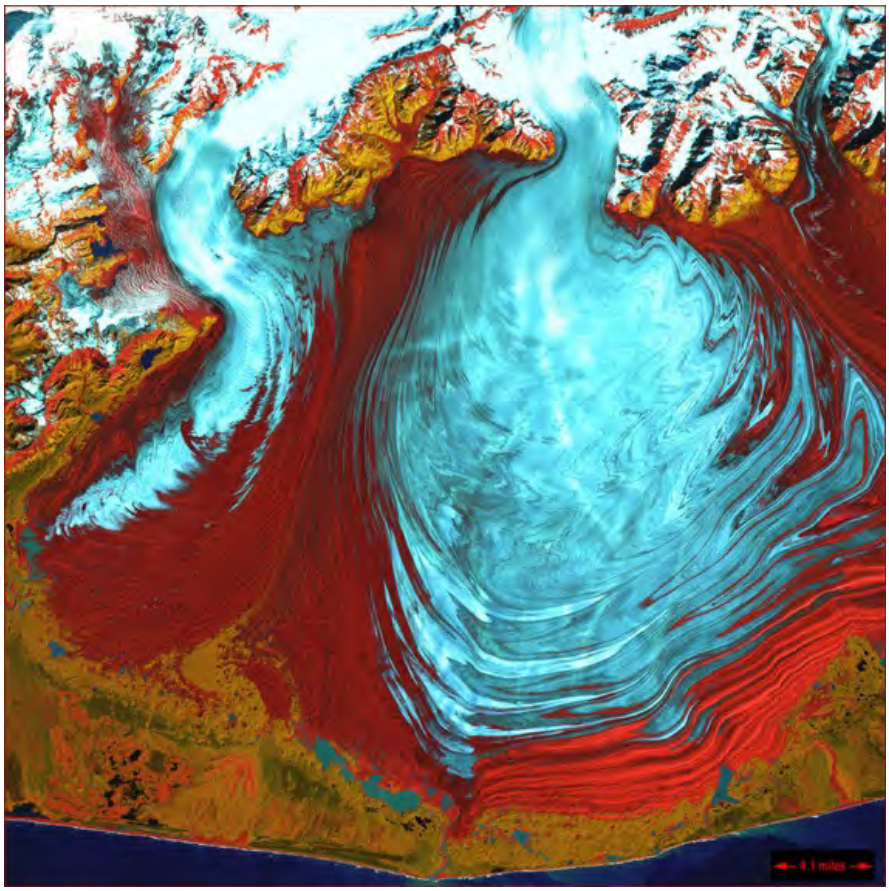
The image mosaic below shows typical alpine or mountain glaciers developed from snow fields covering the higher elevations of parts of the Wrangell and Chugach mountains of southeast Alaska. The first range rises to 5,043 m (16,541 ft) at Mount Kennedy. Among the large glaciers are Nabesna (to the north), Kennecott and Rohn sending meltwater into the Chitina River, and Russell and Barnard coming off the eastern segment of the Wrangell group. In the Chugach

mountains, within the scene, the highest peak is Mount Tom White (3270 m; 10630) but higher peaks are found both to the west and east. The Bering glacier is the largest in the Chugach group, although not as broad as the Malaspina glacier off the image to the east. Note the dark streaks in some of these piedmont glaciers. These are medial and lateral moraines (glacial debris that may become rock deposits). The reds in the scene are mostly tundra vegetation.

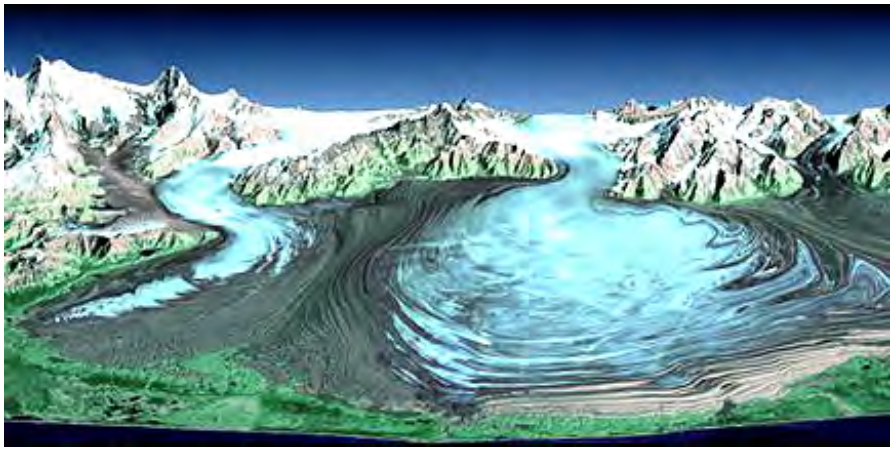


17-15: What would be the appearance of the terrain if all the glaciers in the Chugach Mountains were to melt and disappear? ANSWER

Along the same southern coast of Alaska is a famed active glacier known as the Malaspina Glacier. The Landsat subscene below shows a series of internal moraines on the glacier and a prominent wide lateral moraine to the west (left).



The perspective view of this glacier as it creeps towards the Gulf of Alaska, made by combining Landsat and STRM data, shows it in context with its mountain source:



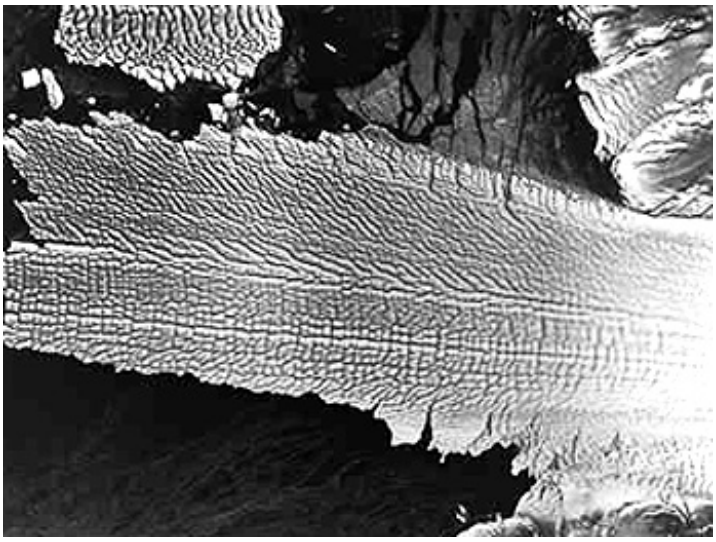
Mountain glaciers occur in groups largely owing to the preglacial prevalence of multiple, often parallel river valleys, all filling with ice when the climate turns cold enough to support extensive ice accumulation. Several glaciers can occupy tributary valleys that coalesce into a single larger valley, as seen here:

The Walsh glacier in Alaska, imaged from space by the KH-7 military satellite, shows this group of englacial morainal features; each represents deposition of rock debris off the side of the glaciated valley:

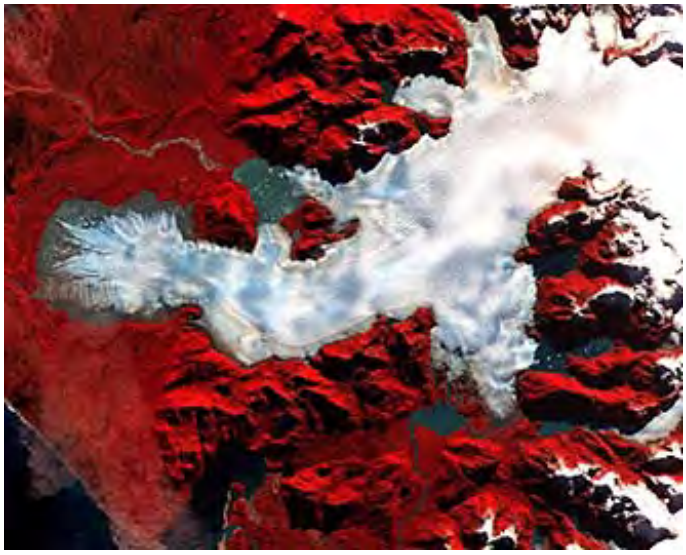
Note that part of the above glacier is almost completely covered with rock debris (eventual till). Although uncommon, this degree of coverage is unusual. An example is the Imja Glacier flows through eastern Nepal, part of a glacier network that ultimately feeds the Ganges River. On October 4, 2010, the Advanced Land Imager (ALI) on NASA's Earth Observing-1 (EO-1) satellite captured this natural-color image of Imja Tsho and surrounding glaciers. Dirt and debris coat these rivers of ice; like the glaciers feeding it, Imja Tsho appears dull gray-brown.

Dark streaks in advancing ice can have other causes and in other modes of ice movement. This EO-1 ALI image shows the edge of the Barnes ice sheet on Baffin Island in the Canadian arctic. The dark lines are produced by seasonal dust accumulating as the sheet grows.

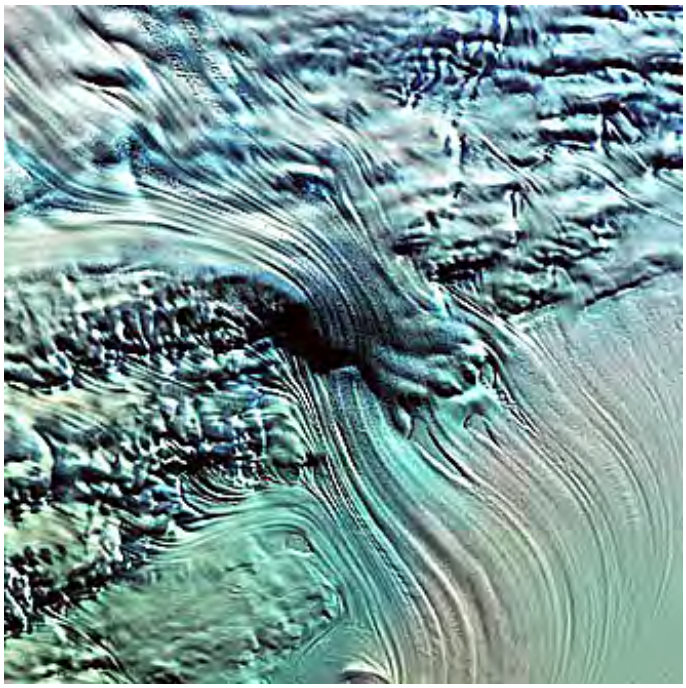
This JERS-1 SAR (radar) image of an Alaskan glacier brings out the details in the cracking of the ice as the entire mass moves slowly down gradient.

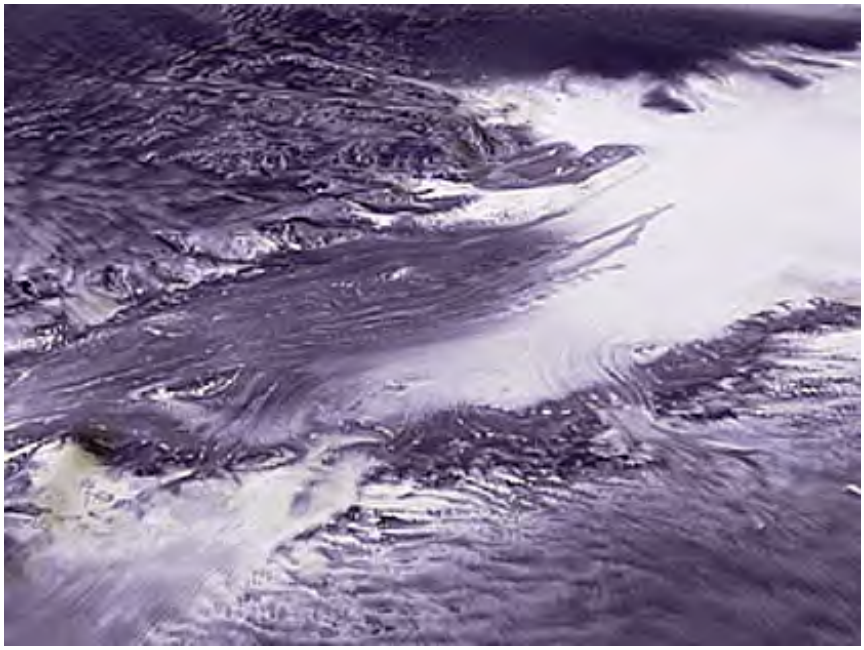


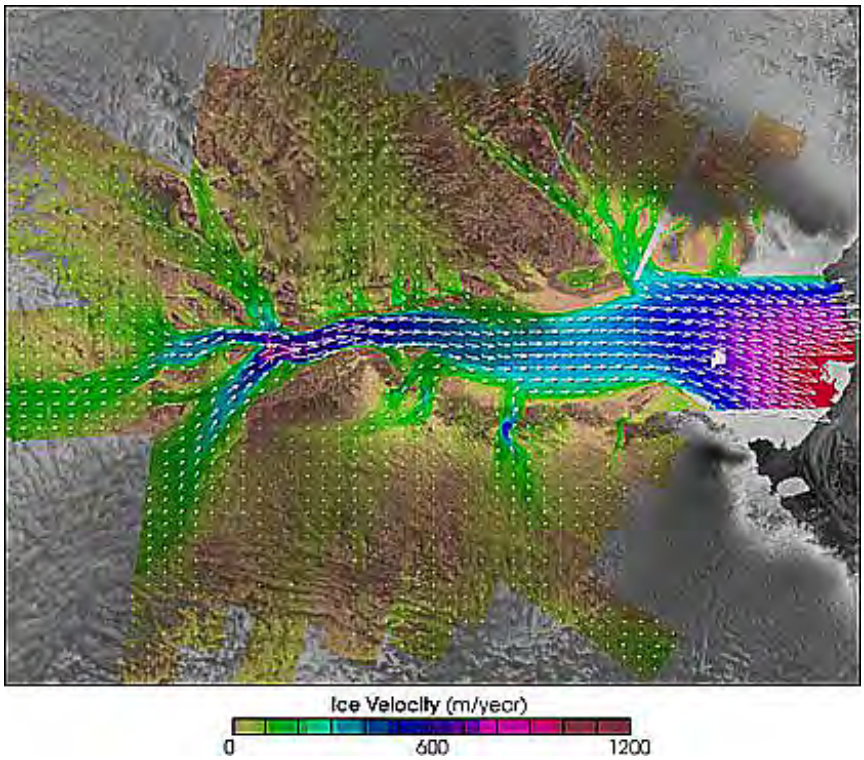
In the southern hemisphere glaciers occur in New Zealand and in South America, as this ASTER view of a mountain (valley-filling) glacier in the Andes displays so well (the time of the year is the southern summer [around December]):



The largest and longest glacier in the world is the Lambert Glacier, in northeastern Antarctica (Australian sector), which meets the sea at the Amery Ice Shelf. Its length is given as 403 km (250 miles); it width reaches to 64 km (40 miles). Below are three space images: the top shows part of the glacier in a Landsat image; the center is a perspective view made using DEM data; the bottom displays rates of flow determined from radar data taken over an extended period:



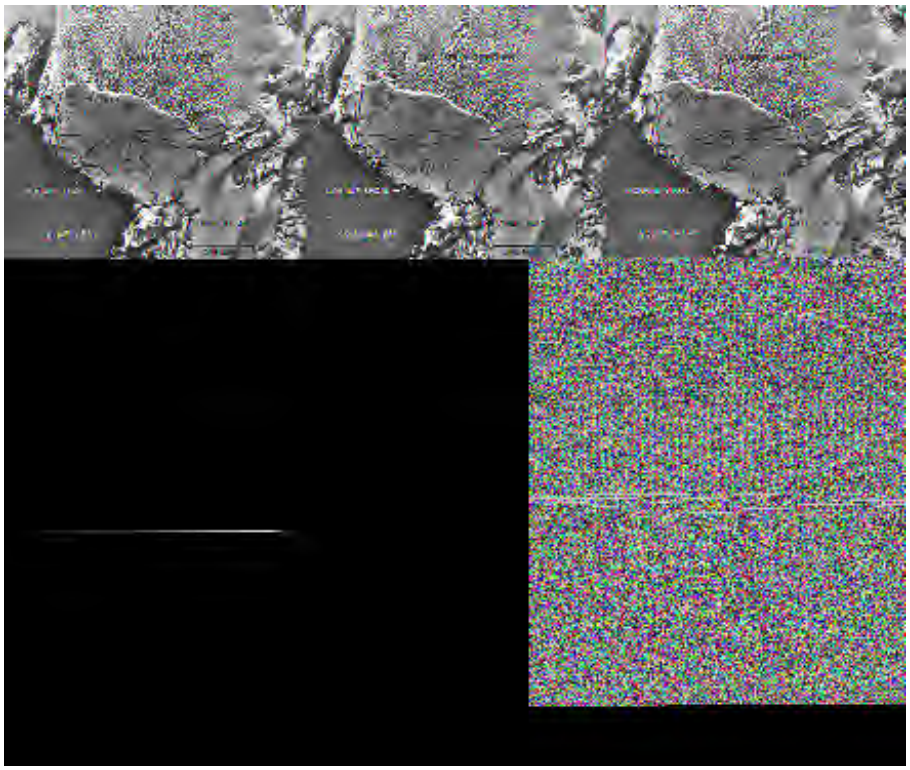




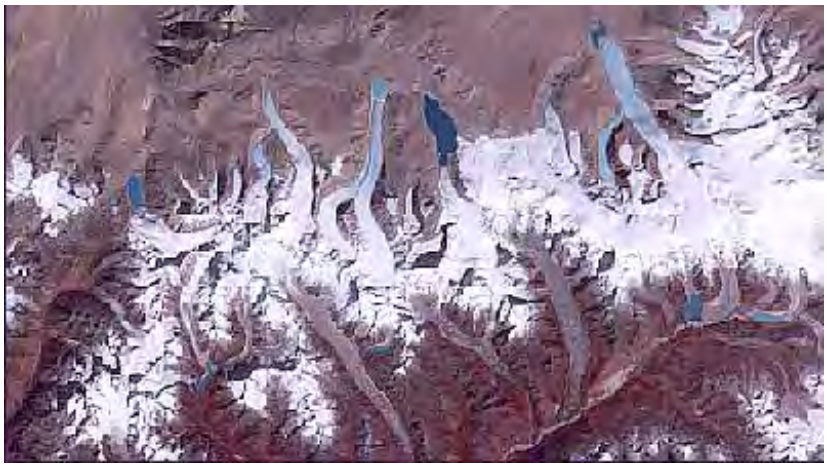
Another Antarctic glacier is Byrd, named after the famed explorer Admiral Richard Byrd, who explored both the North and South Polar regions:

This glacier is advancing at a moderate rate. But the current fastest-moving glacier is the Jacobshavn Glacier in Iceland. Here is a Landsat-7 image on which is superimposed lines for the glacier's toe or front in the years indicated.

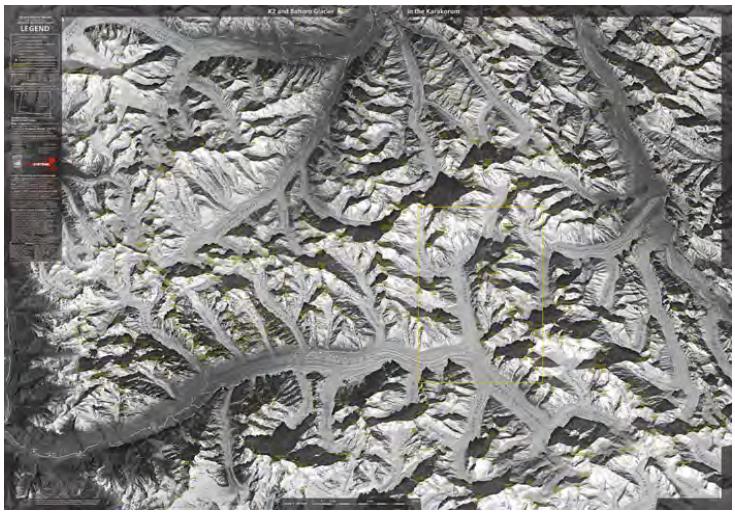
However, evidence is growing that most of the world's glaciers have been receding (retreating) in the last 30 or so years, probably in response to either a natural interglacial warming or man-induced global warming, or perhaps both. The new land being exposed is very evident in the Columbia Glacier in Alaska along its coast, as photoed from the air:



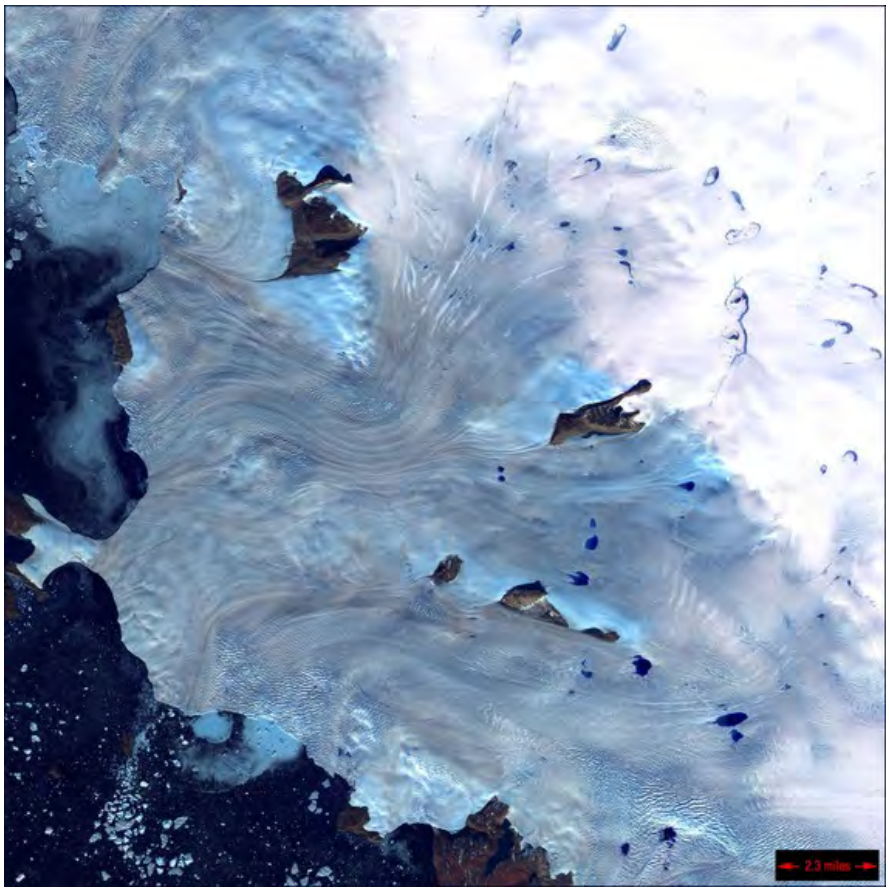
Glaciers in the Himalayas show this continuing condition of retreat. The mountain glaciers in Bhutan shown here are shrinking and developing new lakes at their termini - solid indications of diminishing glacial active in that region.



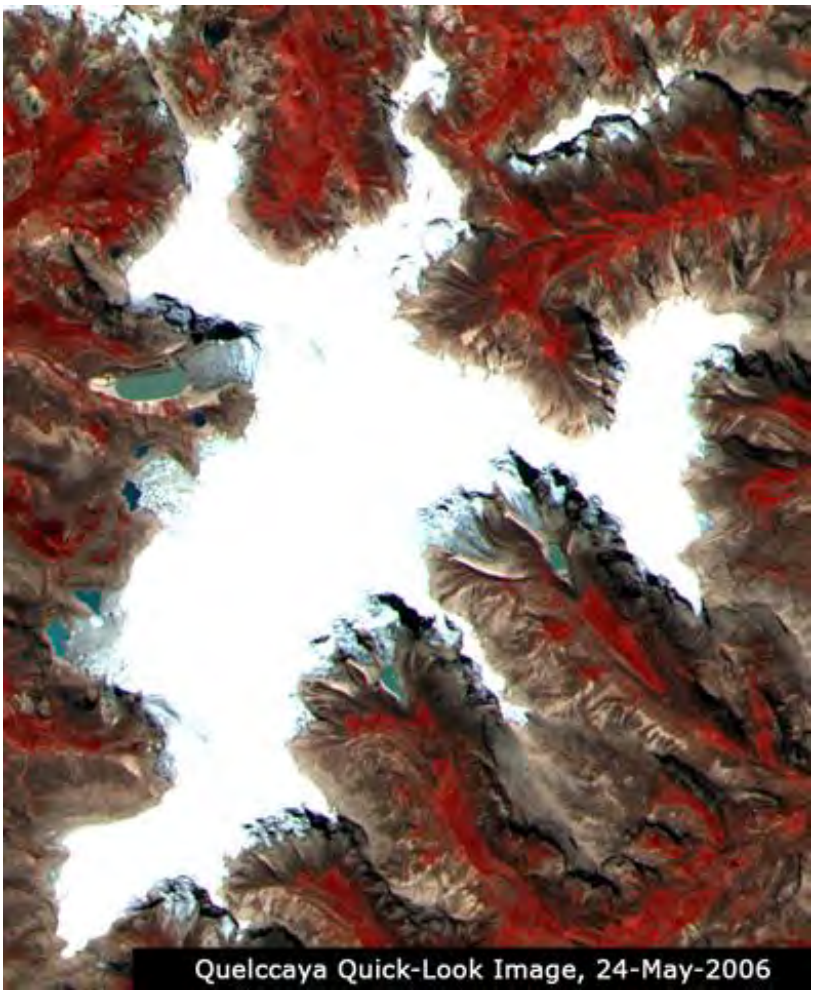
Glaciers in the Kharakoram Range in the western Himalayas show similar slow recession:



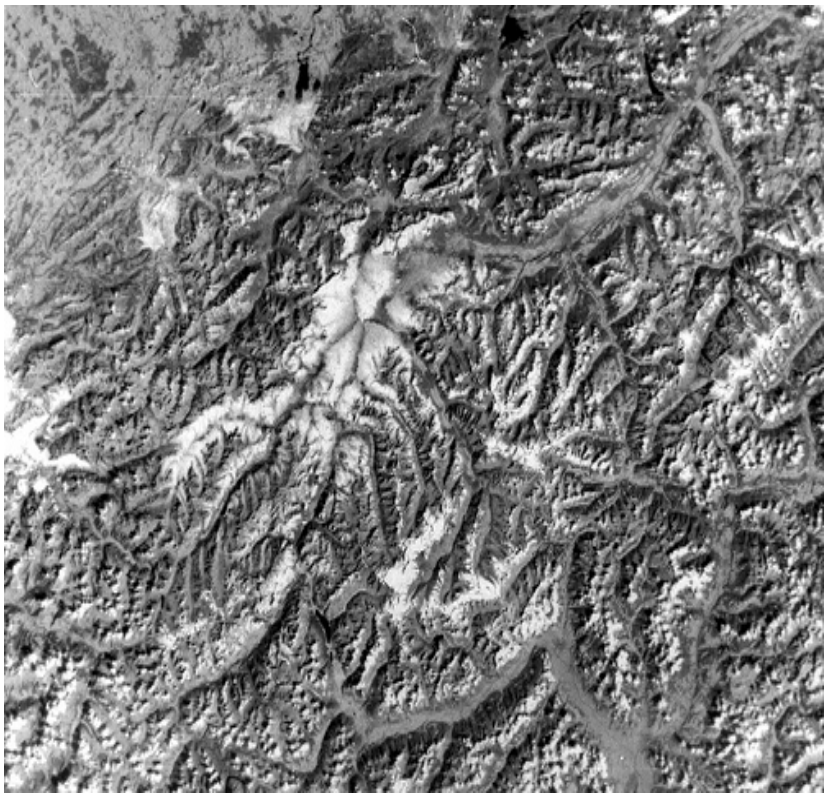
Glaciers also come off great ice sheets such as the one covering most of Greenland. This view (ASTER) shows the sheet's edge and flowing glacial ice emptying into Baffin Bay on the northwest corner of this subcontinent (across from Ellesmere Island and north of the Thule Air Base).



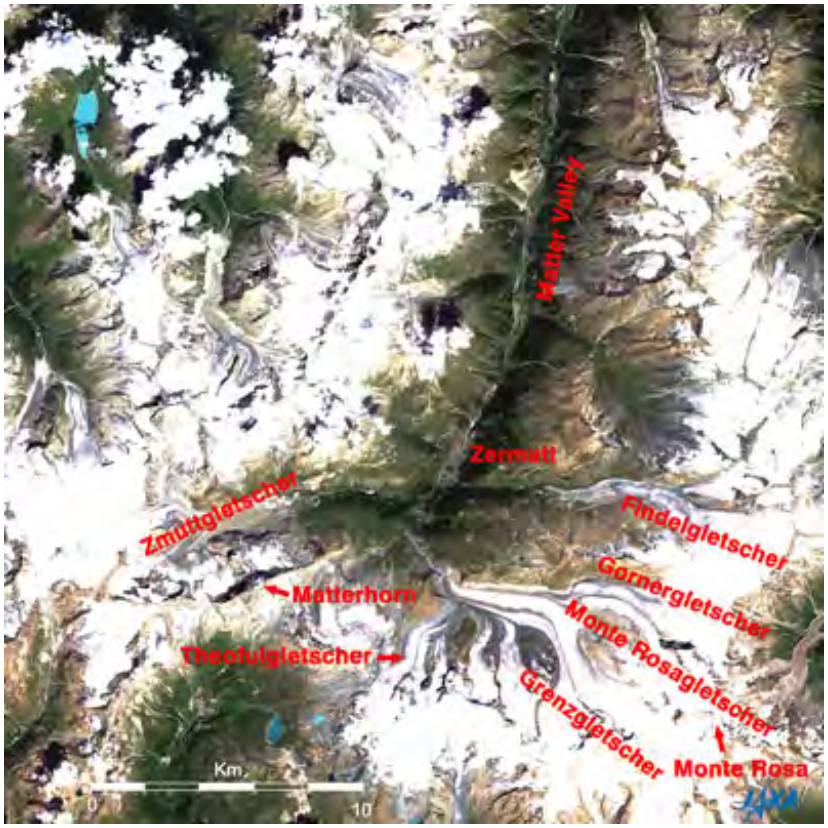
Generally, when one thinks about where glaciers are likely to occur on Earth, the first response would be in the colder regions, such as the Arctic and Antarctic. But, as evident from several of the above examples of glaciers they can also occur in the higher altitudes of mountain ranges. It may seem surprising to learn that glacial ice can be found even in the tropics and temporal latitudes. This next image shows an ice field not far from the Equator. This is the Qualccaya Ice Cap in northern Peru.



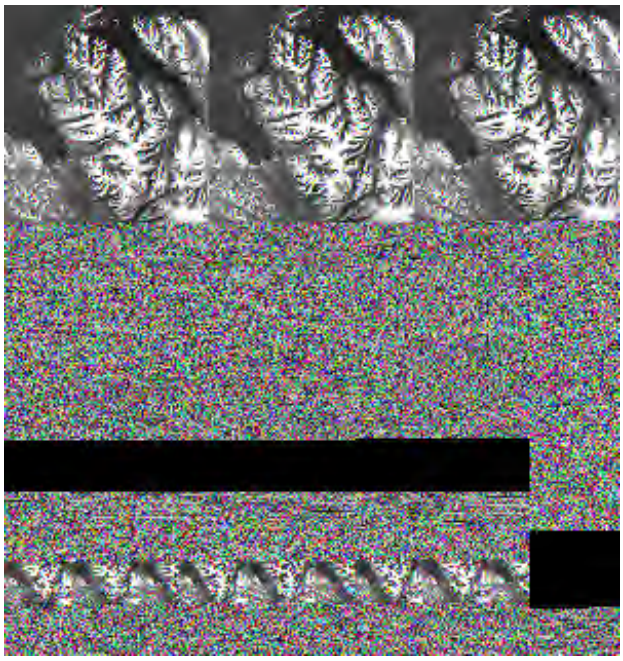
U-shaped valleys are commonplace in the Alps of Switzerland. This Landsat scene shows widened valleys in the alpine region of that country:



The Alps have been extensively glaciated. Some spectacular scenery is the result. Probably the best known single glacial landmark in Switzerland is the famed Matterhorn (which the writer first viewed from nearby Zermatt). It results from the meeting of aretes produced by cirque coalescence.

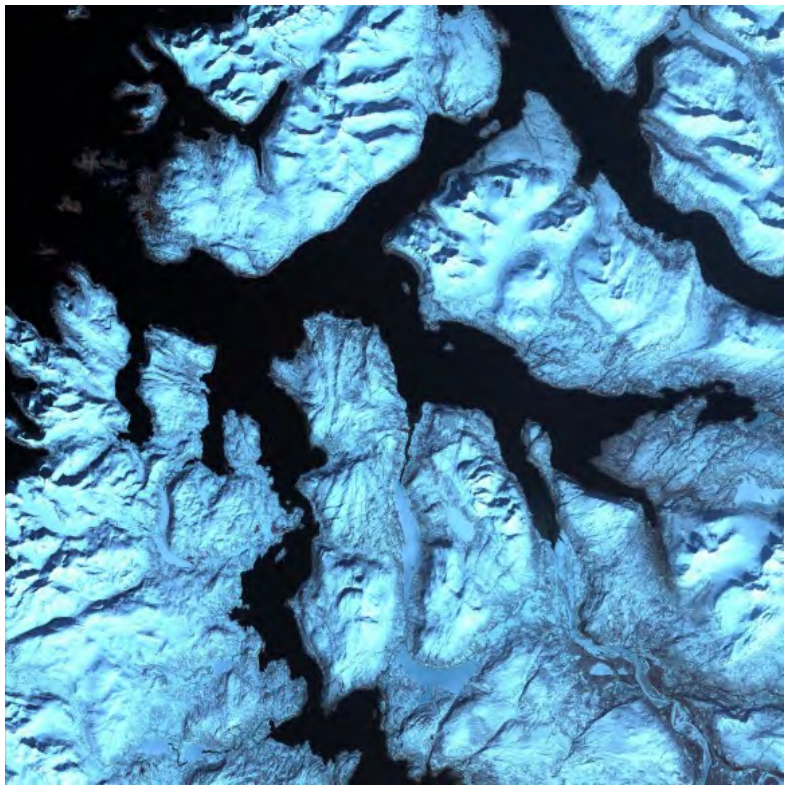


This partial Landsat scene in the Trollaskagi region of northwest Iceland provide a post-ice view of an ice-sculpted mountainous area where glaciers, now gone, have left conspicuous troughs (wide valleys) with intervening ridges now sharply creased into aretes. Several large fjords are formed where seawater has encroached into larger glacially-scoured lowlands.



Fjords were treated briefly two pages back in the segment dealing with coastal landforms. But fjords are primarily expressions of glacial action in which the U-shaped valleys carved by now retreated or vanished glaciers that had reached the coastline are now drowned by marine waters. In the next sequence of images we will look at several more examples of fjords:

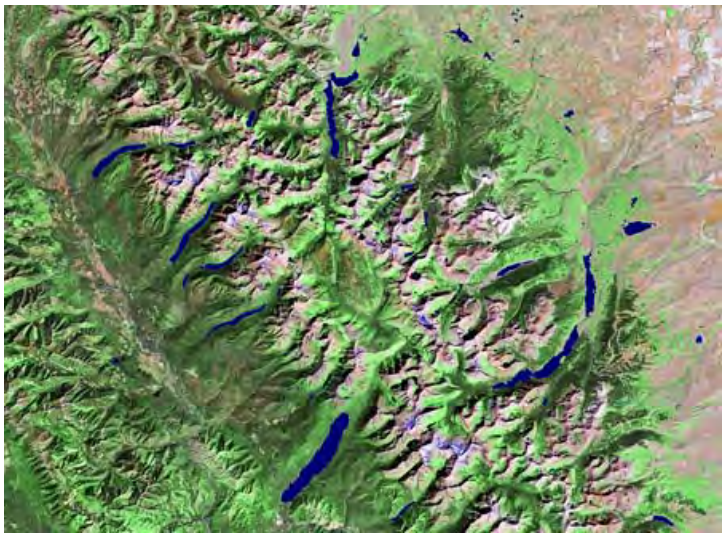
First, another look at Norwegian fjords:



Similar fjords are found in Alaska.

In Patagonia (Chile) the Andes reach to the Pacific. A long line of fjords develop. Some of these occupy valleys that originally were developed along structural lineaments.

Mountain glaciers once were common in the 48 U.S. states. Most have now melted but conspicuous glacial landforms remain. Here is a satellite view of glaciated valleys in Glacier National Park in northwestern Montana:



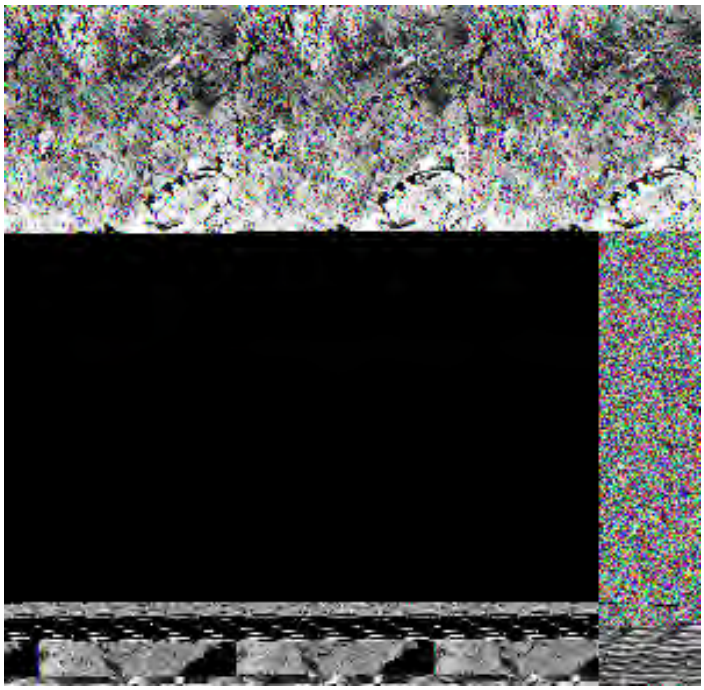
The Uinta Mountains of Utah were also glaciated. In this view, aretes (thin ridges) and cirques (ice-sculpted amphitheaters developed at the highest points in individual glaciers) are evident:

Several advances and retreats occurred (at least four, perhaps more in North America). Today with that ice almost gone from North America, the land once covered has been notably modified by both erosional and depositional glacial landforms. Southern New England is a good example of a land surface affected both by erosional scouring and deposition:

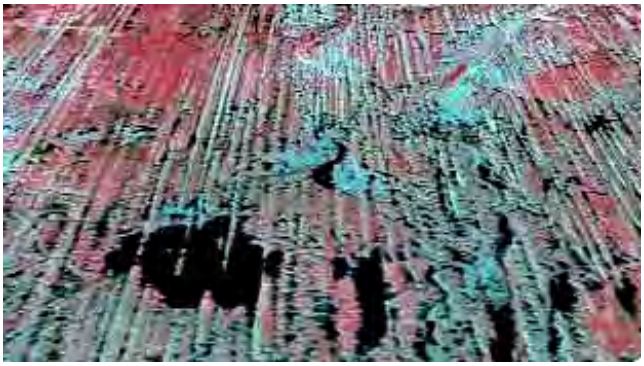
Deposits of moraine and till in places are several hundred meters thick. Depositional landforms are harder to make out in many space images but here are two examples, both from Canada.

The first shows glacial scouring, fluting, alignment of lakes, and morainal surfaces and till plains in

the Northwest Territories.



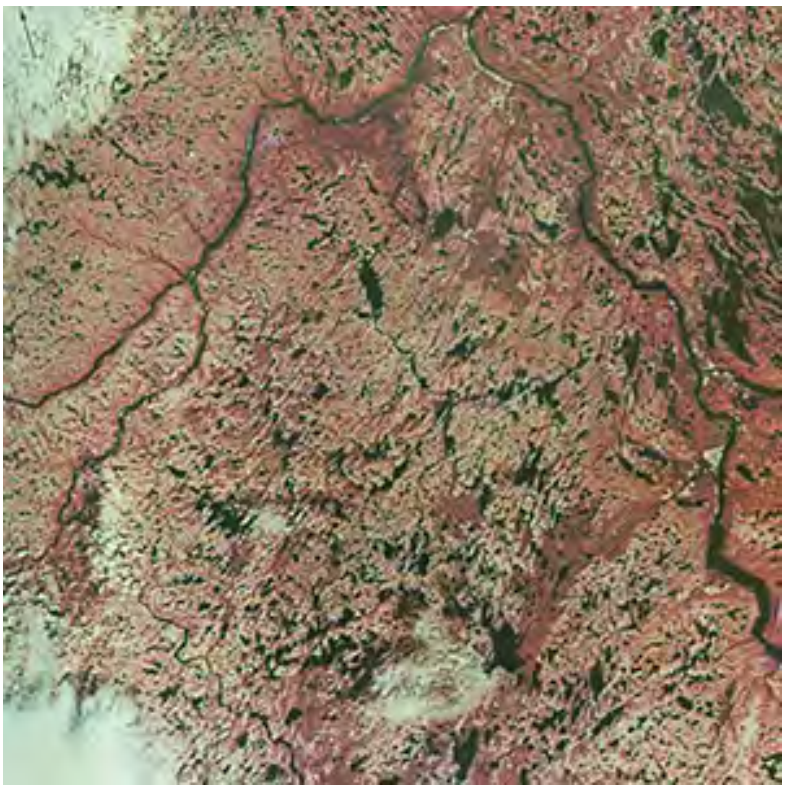
This second image is a Landsat subscene that shows *megafluting* (areas of scraping kilometers wide) in the Nunavat Territory of Arctic Canada.



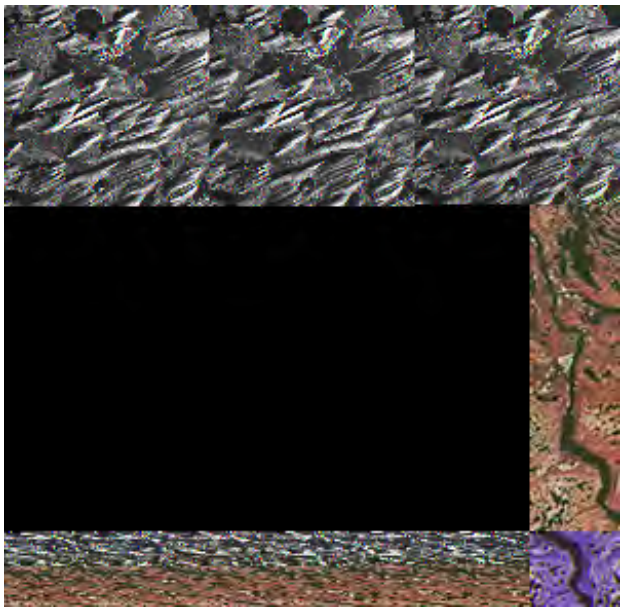
This satellite image shows megafluting in more detail:

Satellite image showing megagrooves.

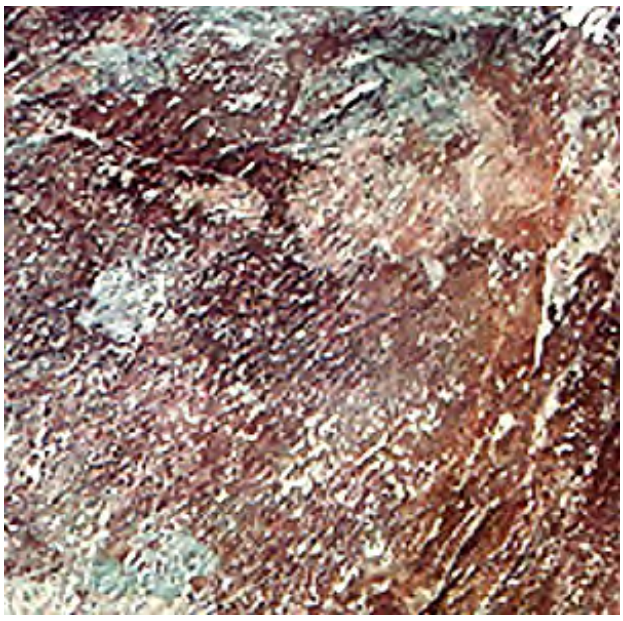
Another example of ice flow features are these fluted and ribbed moraines and crag-and-tail hills in the Ungava Bay district of northern Quebec. These are subtly expressed in the "grain" of this Landsat-1 image; the ice flowed in a southwestern direction:



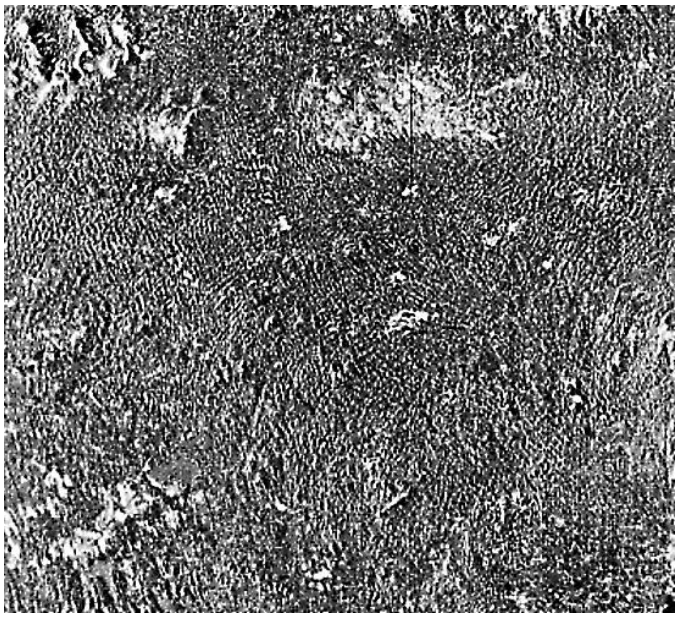
Now we turn to drumlins again. As a reminder, here is a detailed look at drumlins within an aerial photo



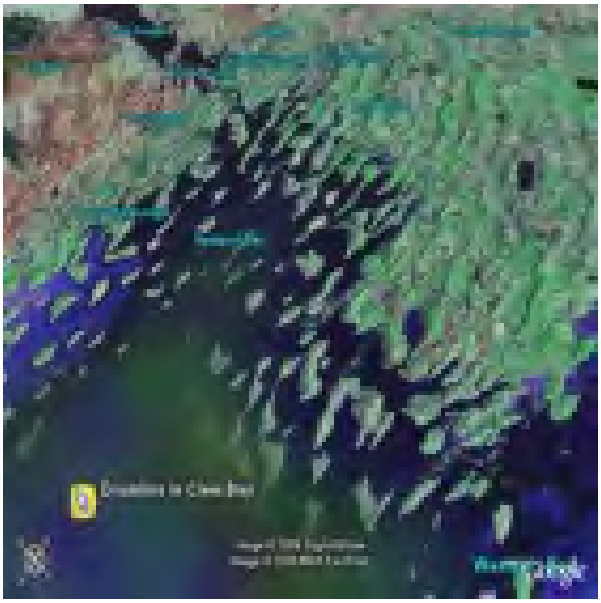
This next image is a Landsat MSS subscene that includes a field or swarm of drumlins, elongated low, cigar-like hills of glacial till that were molded into this shape during a moderate glacial advance. They appear after the ice melts and retreats. Note the direction of motion from the northeast.



Seasat viewed a large drumlin swarm in northwest Ireland, as seen here. The term "drumlin" is Irish, being given long ago to these peculiar hills.



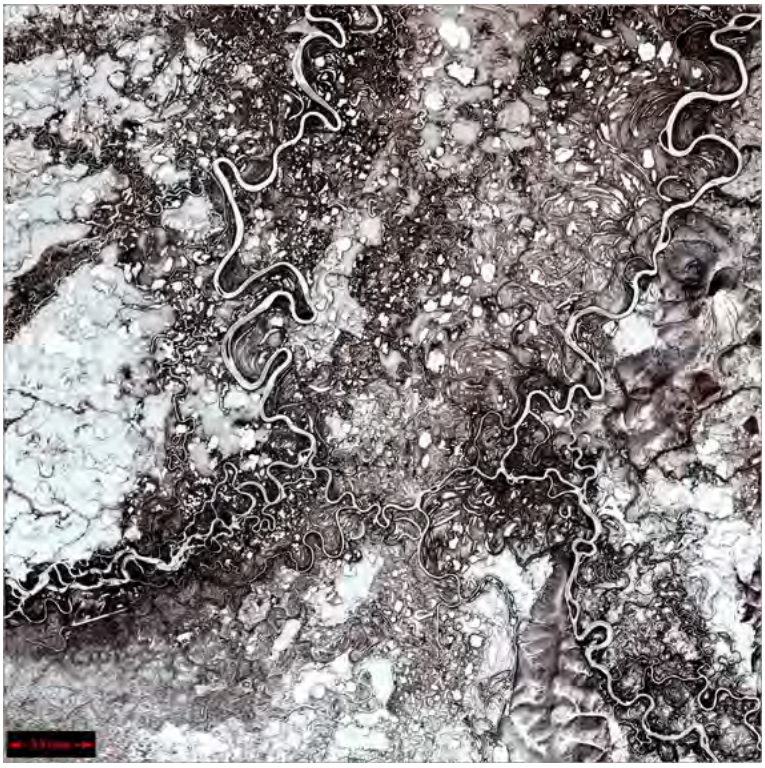
Another satellite image shows drumlins in Clew Bay, Ireland.



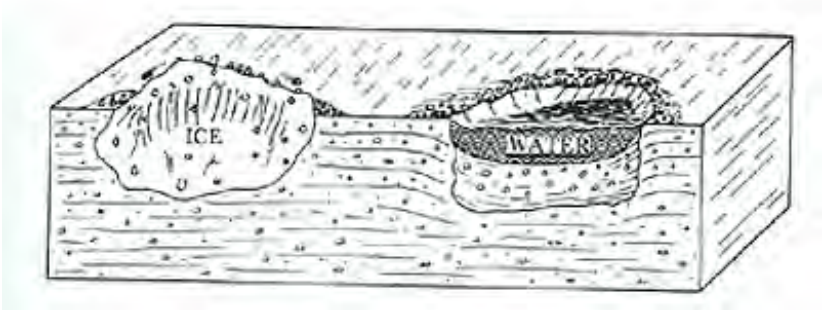
Now, on to a final topic for this page: A survey of landforms associated with dominantly subfreezing high latitude terrains in the northern hemisphere. Large sections of Alaska, Canada, and Russia lie at latitudes where temperatures move above freezing for only part of a year. The ground remains frozen just beneath the surface all year round - this is referred to as **permafrost**. The term is sometimes interchanged with "**periglacial**" which has the more general connotation of "regions beyond moving ice bodies where freezing is prevalent most of a year" The ecology of such regions is distinctive, with the term "**tundra**" being restricted to the particular vegetation - lichens, small plants, limited numbers of trees - that make up this biome. The modern day distribution of permafrost is shown in this map:

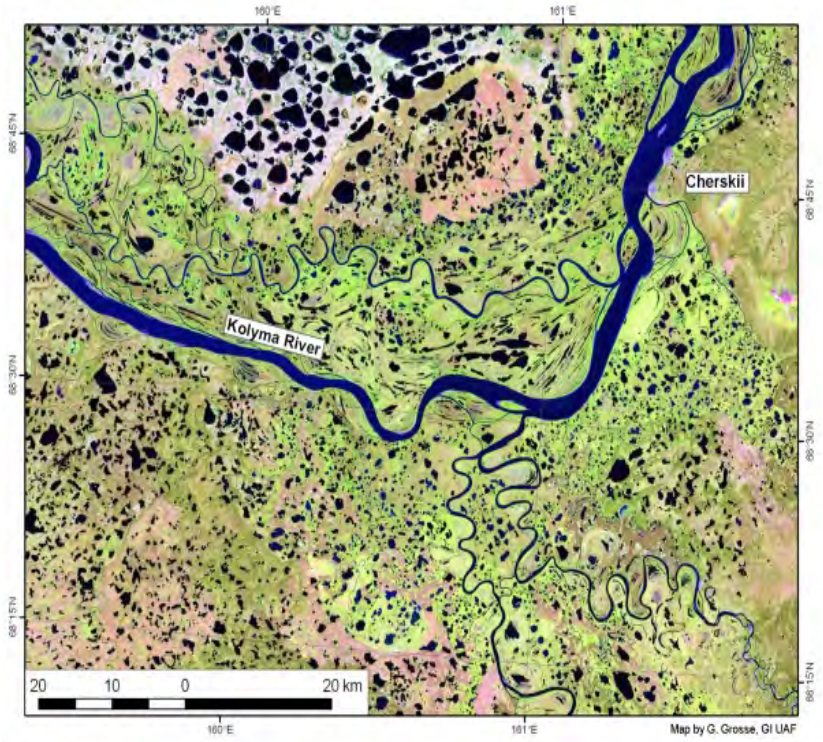
Here are ground scenes typical of periglacial terrain:

>Here are two satellite views of tundra terrain in Siberia:

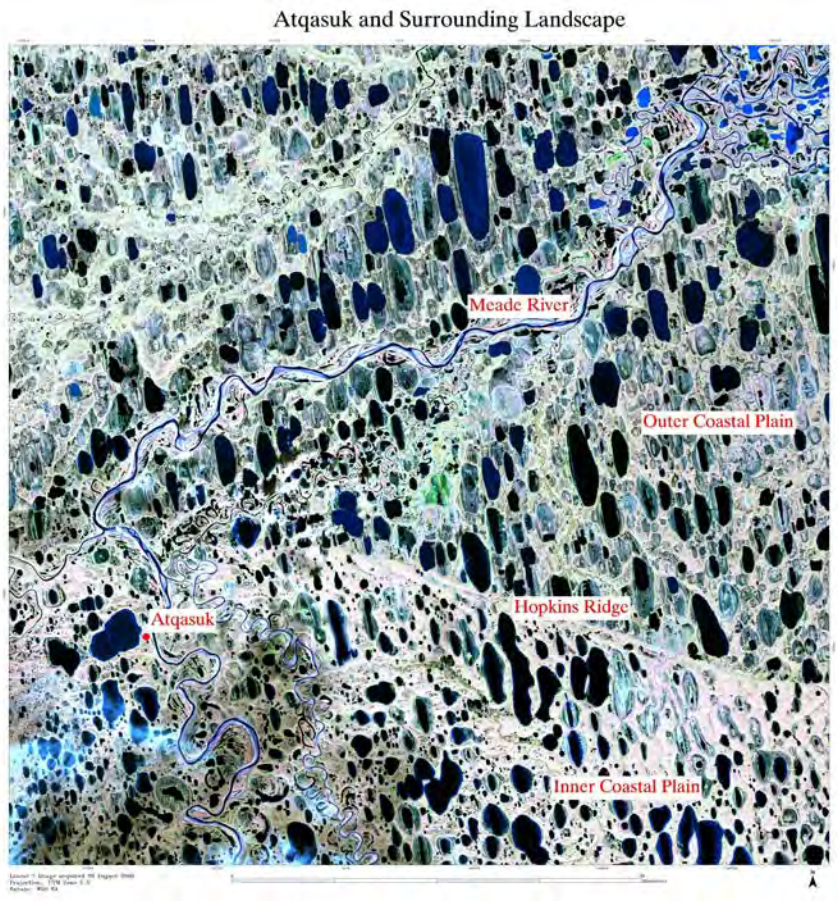


One of the hallmarks of periglacial terrains is the common occurrence of numerous small lakes. Some, called kettle lakes, are water-filled depressions made by the melting of blocks of ice.

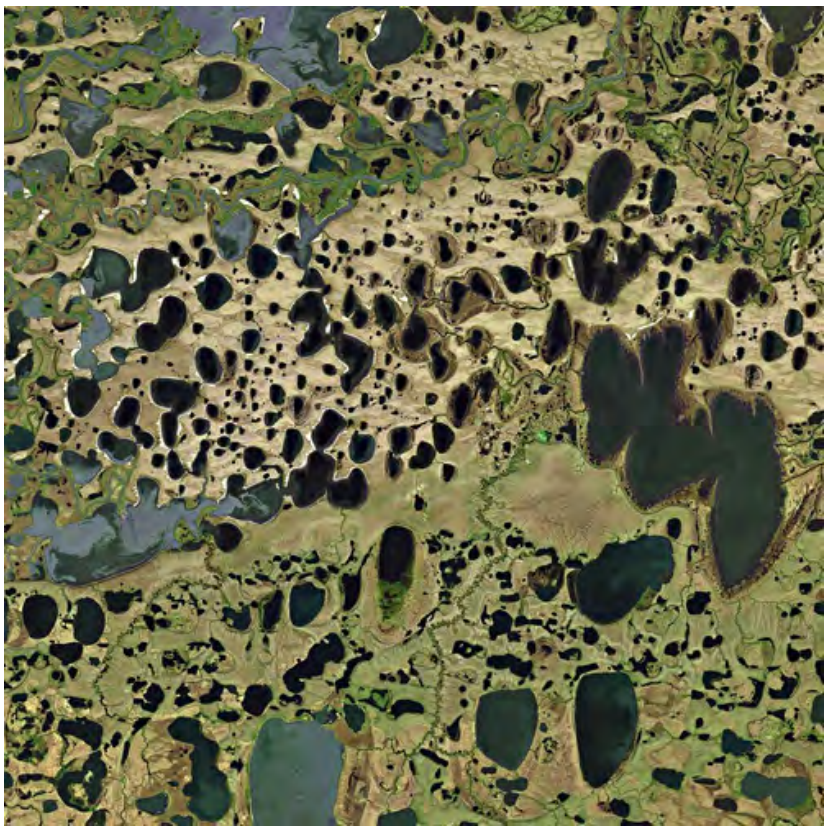




Also sometimes called thaw lakes, many kettle lakes are aligned in a directional orientation, such as seen here:



Lakes are common in the Siberian tundra:



Another hallmark is called 'ice wedge polygons' or 'patterned ground'. It is fairly widespread and certainly distinctive. Look at three photos which display its characteristics.

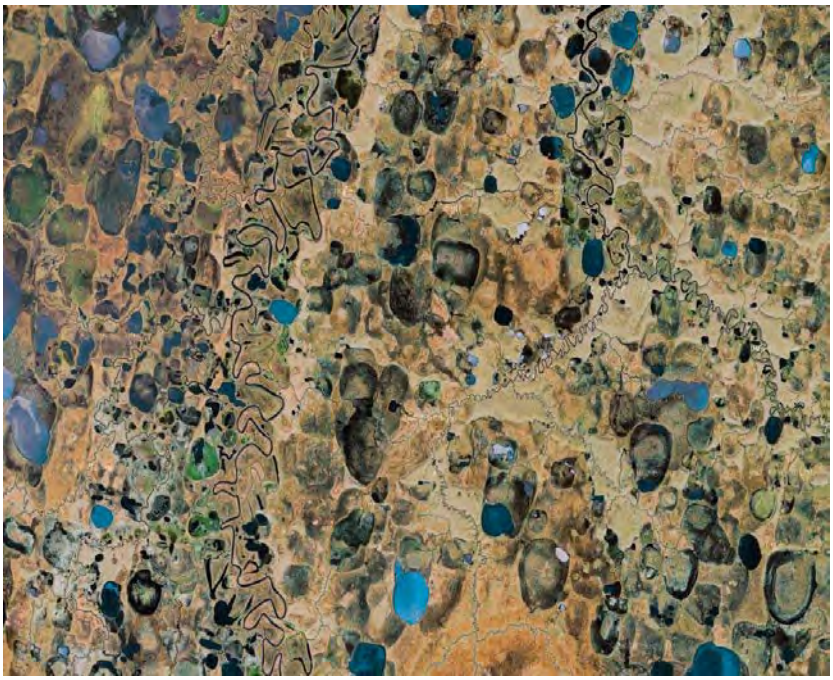
One would expect this feature to be well documented by aerial photos and satellite images. But, the writer spent almost two days looking for just one such illustration on the Internet. Result: one meagre possibility (but patterned ground images for Mars were located). Here it is (where and what unknown, as Internet hit was poorly documented).

The last landform to be considered is called a "pingo". Here is an example:

The mode of formation of pingos requires the action of water from beneath the permafrost layers. The process is summarized in this direct quotation from the Wikipedia site for that name:

"Hydrostatic-system pingos form as a result of hydrostatic pressure on water from permafrost, and commonly form in drained lakes or river channels. Permafrost rises to the drained body's former floor. Pore water is expelled in front of the rising permafrost, and the resulting pressure causes the frozen ground to rise and an ice core to form. The shape and size of a hydrostatic or closed system pingo is often similar to the body of water that it originated from. They can vary from symmetrical conical domes to asymmetric, elongate hills. Hydraulic-system pingos result from water flowing from an outside source, subpermafrost or intrapermafrost aquifers. Hydrostatic pressure initializes the formation of the ice core as water is pushed up and subsequently freezes. Open-system pingos have no limitations to the amount of water available unless the aquifers freeze."

Pingos and kettle lakes are present in this satellite image of permafrost terrain in Siberia.



This ends our tour of highlighted, thematic images from "Geomorphology from Space." You can see more, with detailed captions, on the Internet Site mentioned above or in the CD-ROM prepared by JPL. If you study the book, the Web Site, or the CD-ROM, please look at Chapter 11 that presents an effort by Dr. Robert S. Hayden to demonstrate the value of Landsat imagery in geomorphic mapping. Although he was not the first to produce this type of map solely from space imagery (C.F. Pain's 1985 map of an area in New South Wales, Australia is one early example), he has put together excellent examples of this capability.

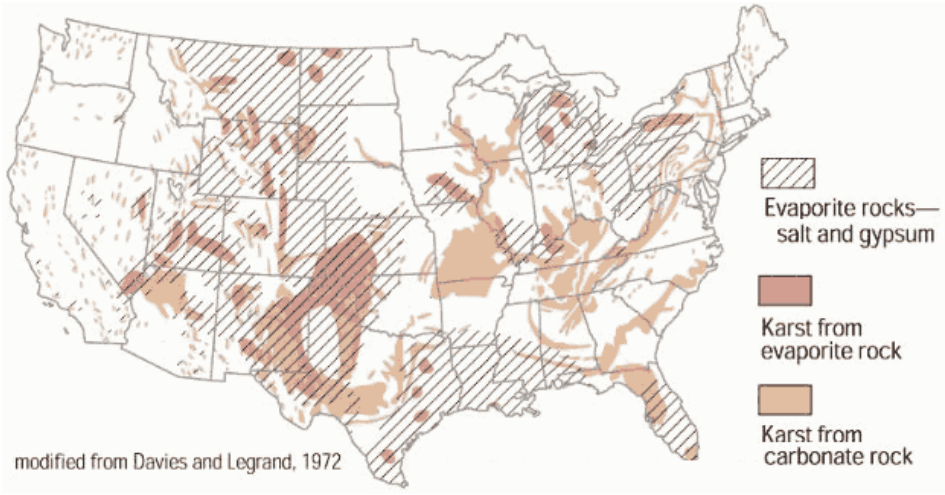
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Other processes besides running water modify the Earth's surface. Percolating ground water can dissolve sedimentary carbonate rocks, leaving sinkholes and other bedrock irregularities that fall into the category of Karst topography. Where water is impounded into lakes, distinctive lacustrine landforms around the shores are produced. Wind can alter surface materials, especially in desert climates, both by erosion and by deposition. Thus, aeolian landforms are widespread across large regions of the Earth's land surfaces.

Karst/Lacustrine/Aeolian/ Landforms

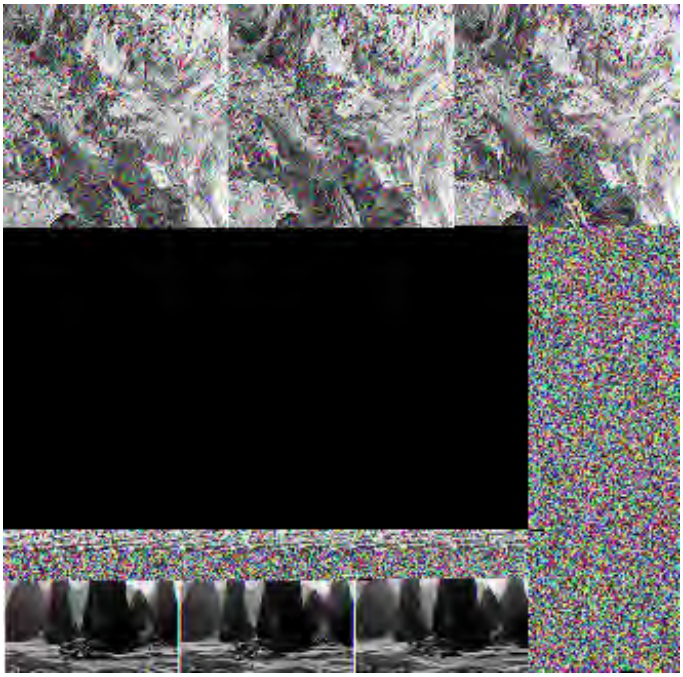
Depressions caused by solution of near surface rock, by groundwater action at greater depths, and by other collapse mechanisms (such as is caused when subterranean mine openings induce failure at the surface [coal mines are particularly vulnerable]) produce distinctive landforms including those collected under the term "karst". This solution occurs mainly in regions where carbonate rocks (mainly limestones) make up much of the strata in the top 1000 meters or so (where groundwater can circulate), but collapse can also occur where salt beds (evaporites) are dissolved. This map shows the potential for carbonate/evaporite solution in the United States:



Surficial landforms developed from subsurface solution of halite and gypsum are fairly common worldwide. A search of the Internet failed to find good examples to show on this page. The one exception is this satellite image of the Isachsen salt dome in the Canadian Arctic. Note the rough surface which is solution sculpturing of the salt itself as exposed at the surface:

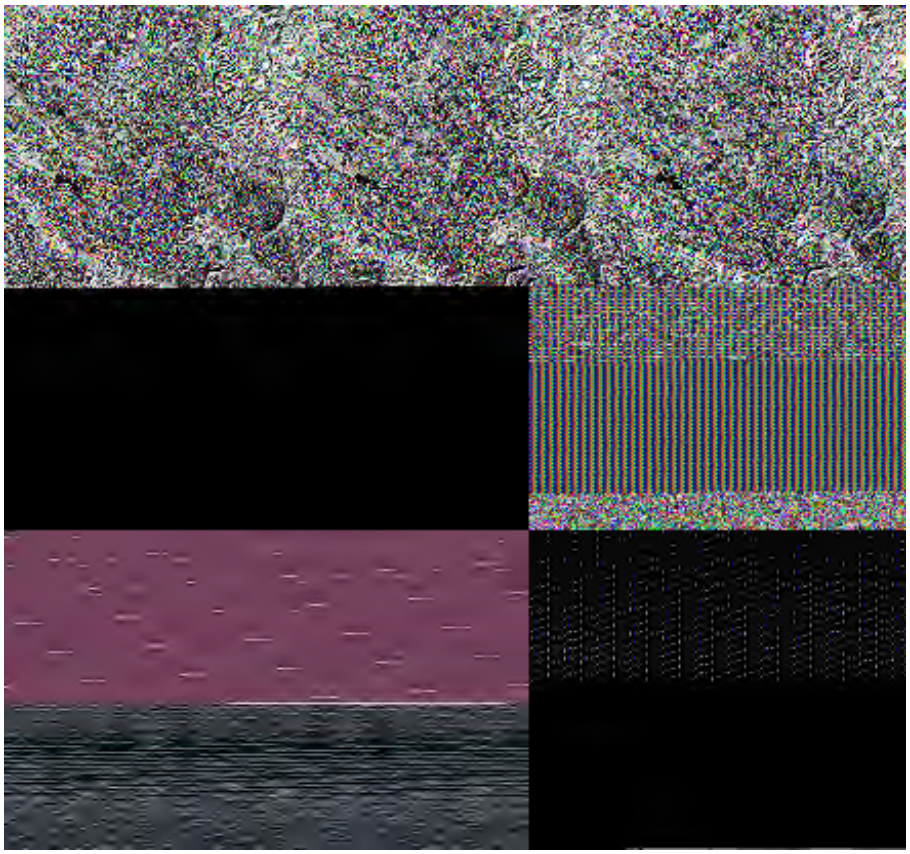
Landforms resulting from carbonate solution are even more widespread and numerous examples on the Web were found, as now described. Carbonate Karst topography encompasses a variety of landforms, such as solution valleys, sinkholes, subterranean caves, and towers, which develop largely by chemical dissolution of limestone rocks. Solution often begins in and extends from structural joints. As these enlarge, they can turn into valleys that may widen and coalesce to leave the terrain pockmarked with depressions (sometimes called "cockpits", a term applied to Jamaican karsts). The "type locality" for karst terrain is in the Dinaric Alps of Croatia, where Mesozoic limestones are notably pitted by solution, as seen in this astronaut photo:

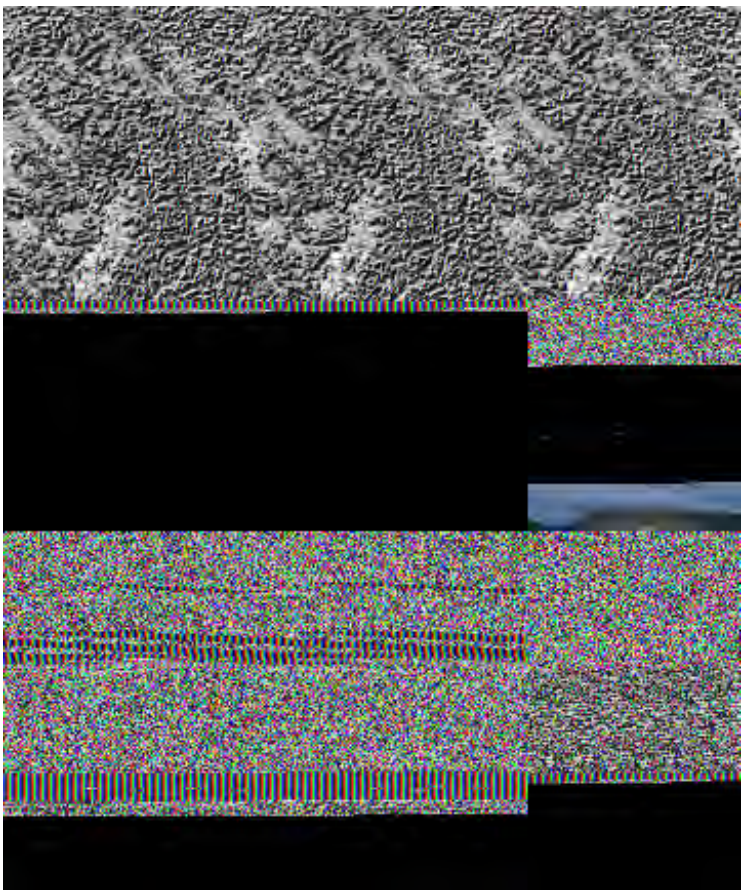
Expansion of dissolved volumes may leave the carbonate rock units between them as residual peaks or towers. Nowhere in the world is this more spectacularly developed than in the Guangxi Province of Southern China. In the image pair below, a group of towers appears at the top and a Landsat image of a region west of Guilin, a popular tourist center in China, at the bottom.



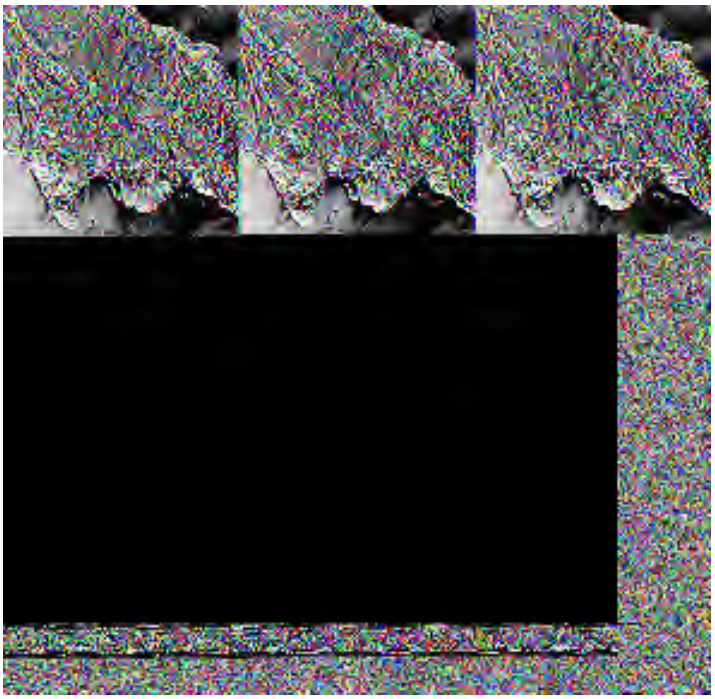
The karst topography in this scene appears in the darker-toned surfaces. These karsts are a thick series of carbonate rocks that elsewhere in the image have their outer rock removed to expose older non-carbonates underneath. Two main joint sets criss-cross the carbonate sequence and enlarge into intersecting valleys. Drainage across these plateaus is now mostly internal, so that the region also contains numerous caves. This scene offers a remarkable illustration of how dissimilar rock types (which host different joint spacings) give rise to strikingly varied topographic expressions and resulting landforms

The next pair of images are from an area next to the above scene. The first new image is an enlarged subscene containing mostly Karst topography. Below it is a further enlargement which shows the individual hills making up this distinctive landform.





The island of Jamaica in the Caribbean is composed mostly of limestones. As such, it is subject to intensive karst development in response to high annual rainfall. This Seasat radar image shows the typical karst topography developed under subtropical conditions.



The image shows a variety of landforms: some of mountains that are foreshortened, areas of land use, and roughened offshore waters. Limestone beds govern much of the island's underlying geology. In this area of heavy rainfall, these beds readily dissolve to form typical karst topography. Doline depressions (solution pits) occupy large sections of the scene. These pits are obvious in the aerial photo below, but in the Seasat image, they form a fine texture that may not show well on a monitor. If it does, look also for a faint but visible lineation caused by fractures that have a preferred orientation because of the look direction.

This next image is a perspective view of Jamaican cockpit karst using IKONOS and DEM inputs:

17-11: What common feature of internal drainage (both groundwater and underground streams), quite popular to visit, is frequently associated karst topography? ANSWER

Another landform found almost entirely in terrains made up of limestone at the surface is the sinkhole. There are three types: Solution; Collapse; and Subsidence. Sinkholes are more common in humid climates because of the increased rainfall. But they can develop over time in arid climates. Here are sinkholes in the Kaibab limestone strata in northern Arizona; note their morphology:

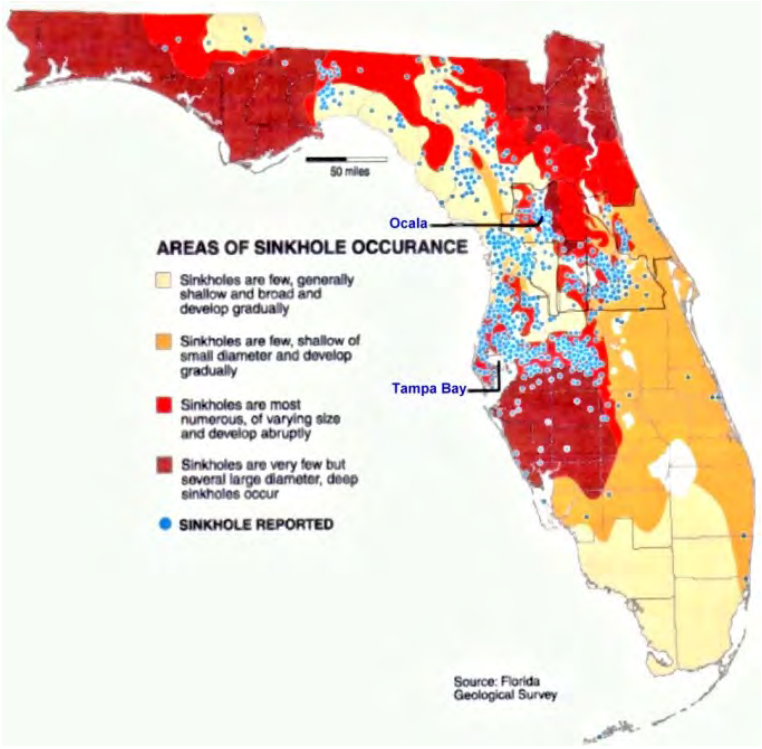
Sinkhole like depressions are common on the surfaces of large areas in states making up the Interior Lowlands. Carbonate rocks (unfolded; not involved in tectonic mountain-building) are prevalent in these supracrustal rocks on the central craton. Here is a view of a landscape in Kentucky that shows distinct pockmarkings from sinkhole-related solution:

Large volumes of carbonate rocks can be dissolved away to form caves. In Kentucky the most famous cave is Mammoth Caverns, which is a National Park. Its main entrance is not particularly impressive but once inside there are miles of passages, some open to the public but other parts less familiar. The first figure shows that entrance.

Claims are made that Mammoth is the largest cave in the world. More than 570 km (350 miles) of openings have been mapped. Some of the rooms are quite large:

The aerial photo of the surface above part of Mammoth Caverns shows topography similar to the cockpit terrain of Jamaica:

As we learned earlier, Florida's surficial bedrock is almost entirely limestones of Cenozoic age. Hence sinkholes are in abundance, especially in the central part of the state. They can form slowly over time but some appear in days if collapse follows long term solution. Here is a map by the Florida Geological Survey that shows sinkhole distribution in the state:

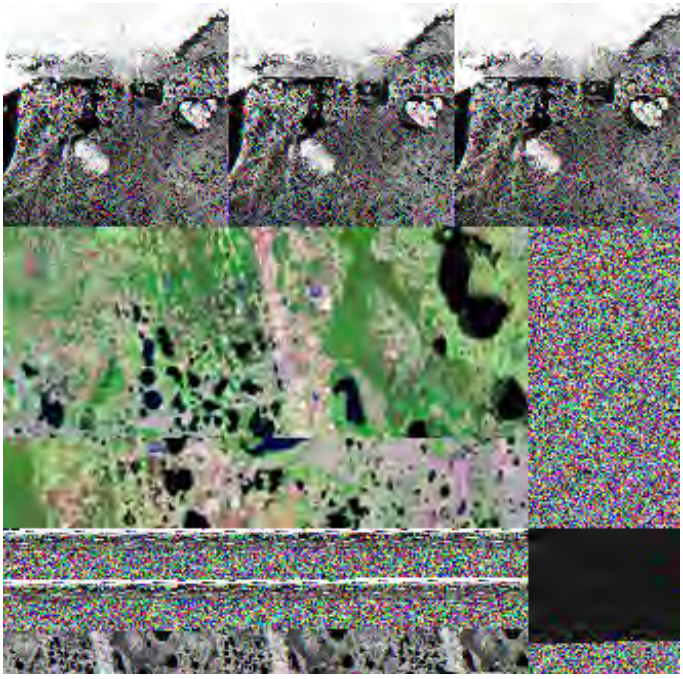


Sinkholes can occur anywhere in the state, sometimes in populated areas with calamitous results:

Many sinkholes fill with water to become lakes. Here is one such feature:

Satellite images of central Florida show the widespread distribution of sinkhole lakes, seen first in this Landsat subscene around Orlando and then in an aerial photo:

These last scenes transition us to lacustrine (lake) landforms. Lakes come in many shapes, sizes, depths, and origins. Most are freshwater but some may be saline because of evaporation that concentrates dissolved salts. Glacial activity is a particularly common cause of lake development:

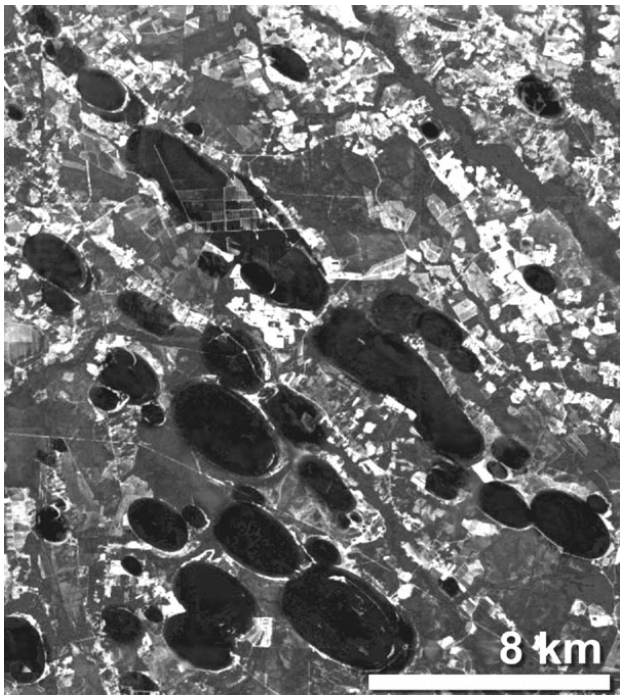


The first image shows elongated, thermokarst lakes in the coastal plains ending at Point Barrow in northwest Alaska; the second is an ASTER image of similar lakes on Alaska's northern slope. The lakes tend to be ovoid and elliptical and are generally less than 6 m (20 ft) deep. They result from collapse, slumping, and caving during the summer thaw of soils and sediments affected by permafrost (a condition in which vadose water in upper layers above the (ground)water table freezes permanently to some depth). The upper layers in the permafrost will experience some pore water melting when heated in the warm season. The cause of elongation is uncertain but may be due to paleowind control at a time when winds came from the northwest, rather than today's northeast.

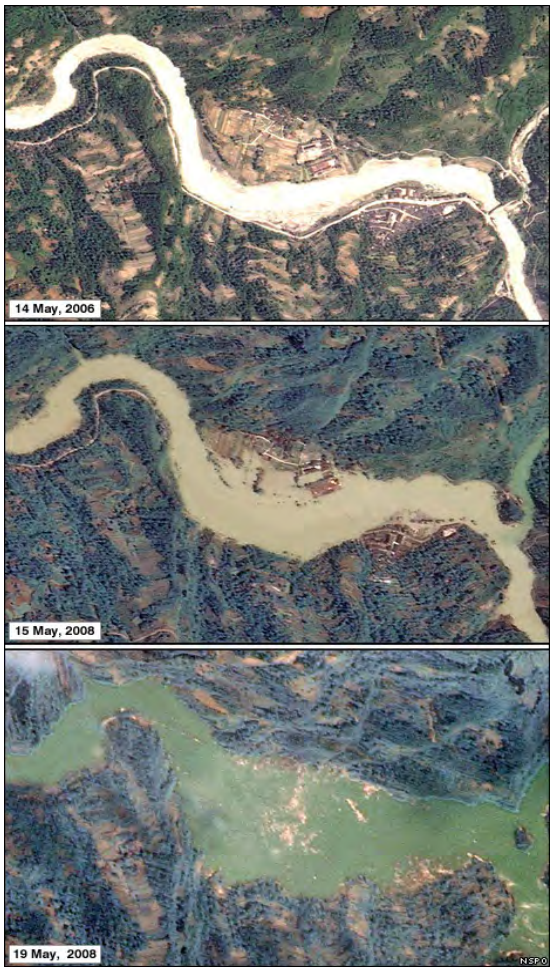
17-12: Much of Alaska and Canada were glaciated in the Pleistocene. Suggest another way in which glacially-related action can cause elongate, aligned lakes. ANSWER

Lakes are a widespread feature on the Earth's land surfaces. They develop usually in depressions that collect water from runoff or, more commonly, from streams that empty into them. Generally, lakes survive for only a few million years at most, as they tend to disappear by evaporation or erosion.

The thermokarst lakes shown above are aligned, owing to directional movement of the ice. Lakes that have a similar appearance and are commonly aligned occur on the coastal plains of North America. They have been named the Carolina Bays but have been found from New Jersey to Texas. Ones like those shown below are shallow and usually only a kilometer or so in length. Their origin is still uncertain: the impact crater explanation is totally unsupported; processes associated with the influence of Pleistocene climatic conditions are proffered as a plausible cause but the exact mechanism remains uncertain (something akin to solution collapse is one hypothesis).



Here is an example of a fast-forming lake. An earthquake in the Sichuan province of China caused a landslide that dammed up a river. Water quickly accumulated, as seen in these Chinese satellite images:

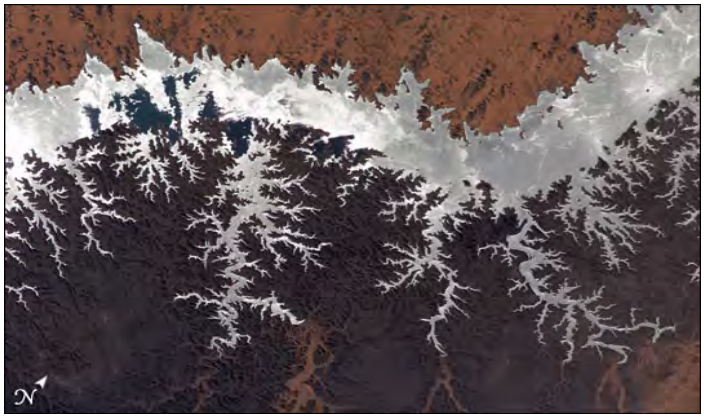


Lakes can undergo significant area and volume changes in short time spans. A drought in the first decade of the 21st century in the desert Southwest of the U.S. has produced the changes observed at Lake Meade (formed by the Hoover dam) in Arizona:

April 19, 2003

May 5, 2000

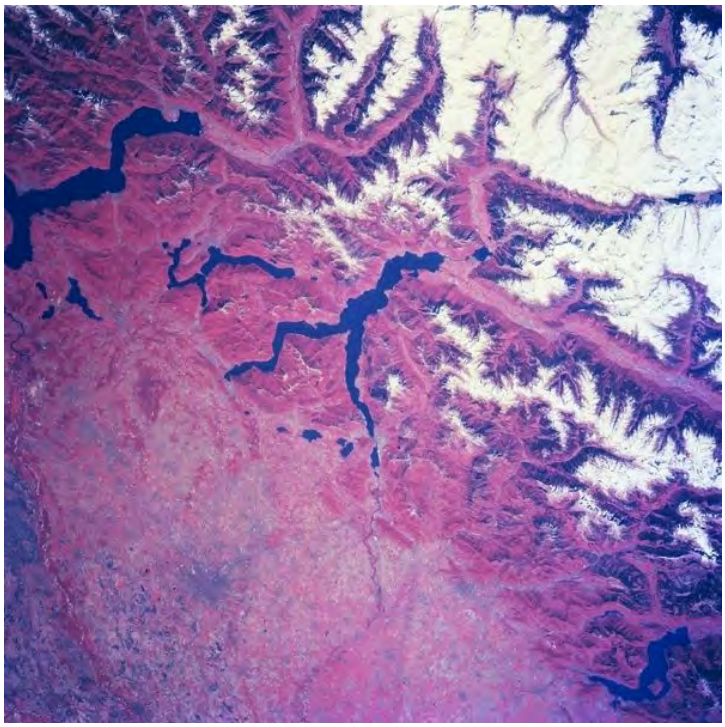
Lakes forming by manmade dams are common. One is Lake Nasser in Egypt, formed by the Aswan Dam. In this unusual photo taken by an astronaut, the lake appears white instead of blue. This is actually sunglint, white provides stark contrast with the rock at the shoreline:

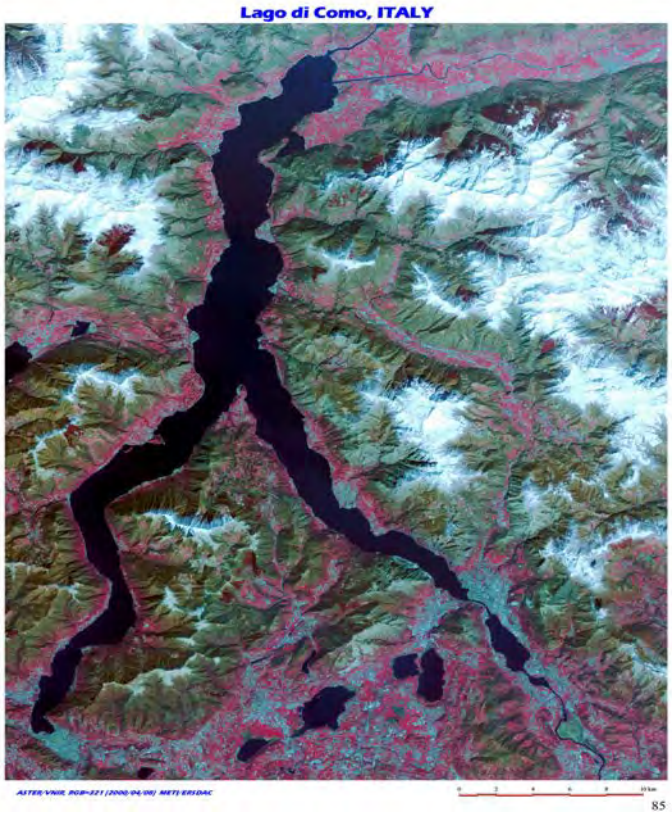


Lakes can form as an intermediate body of water along streams. Lake St. Clair lies just northeast of Detroit, Michigan. The St. Clair River drains out of Lake Huron to the north and into Lake St. Clair; the Detroit River empties from Lake St. Clair into Lake Ontario. Lake St. Clair was produced by glacial scouring along with the larger Great Lakes (some geographers include it among these lakes).

The continental glaciers of the Pleistocene Epoch reached into today's northern states east of the Great Plains. Deep gouging of the relatively soft sedimentary rocks produced big and variably deep depressions that have filled with water since the last glacial retreat some 8000 years ago. These produced today's Great Lakes, which collectively are the largest grouping of fresh-water lakes in the world. Here they are in this Landsat mosaic (lakes identified in the caption):

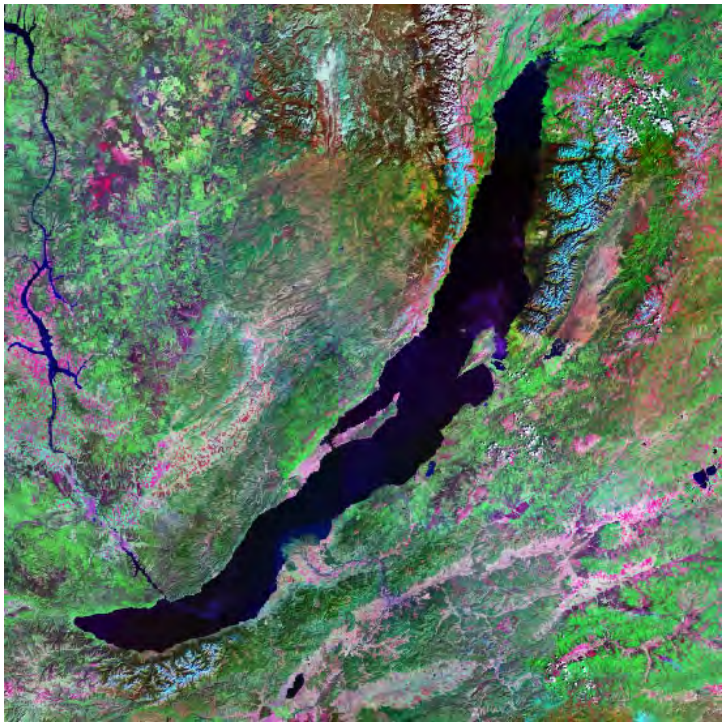
Lakes are formed by mountain glaciers as well. They may result from accumulation of meltwater behind a terminal moraine. The term "alpine lakes" applies. Here are a group of such lakes in the Italian Alps; they include Lake Como, Lake Lugano, and Lake Maggiore:



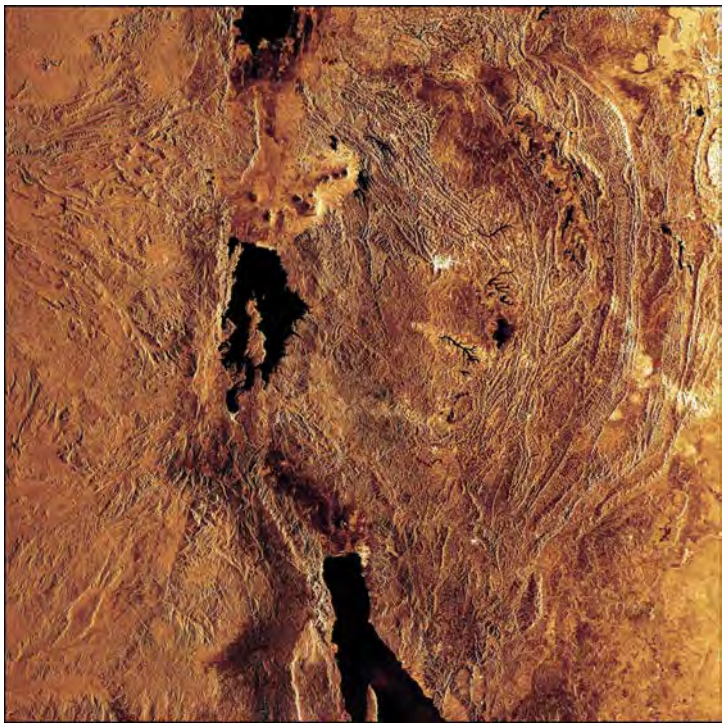

Lago di Como, ITALY

The Finger Lakes of Alaska owe their origin to glacial action in mountains and to damming by moraines (see next page):

Lakes can be directly or indirectly related to tectonic activity. The deepest lake in the world is Lake Baikal in southern Siberia. Formed in a rift valley, it bottoms at 1637 m (5370 ft) and is also the largest (volumewise) freshwater lake in the world.

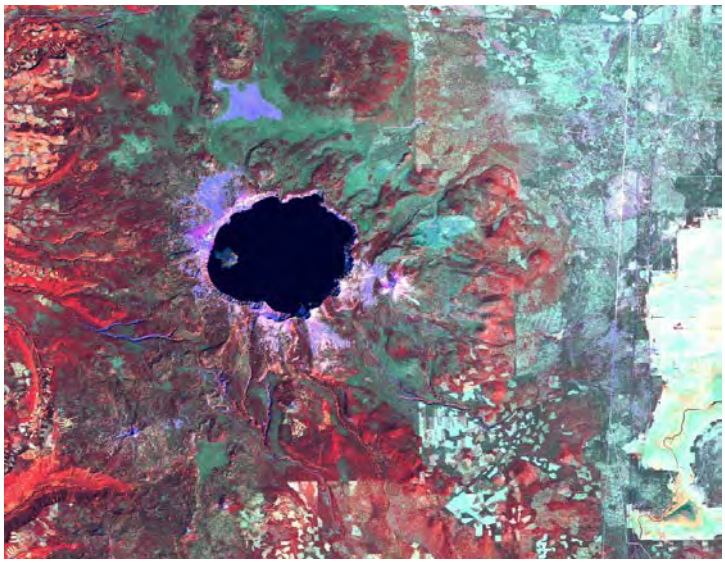


Smaller lakes abound in the African Rift Valley such as seen in this Envisat SAR image:



Some of the African lakes associated with volcanism can hold poisonous (toxic) gases such as methane, or CO_2, that escape into the air when thermal overturning occurs. This has resulted in deaths among villagers living on the lake shores. One such lake is Lake Kivu in the African Rift on the border between the Democratic Republic of the Congo and Rwanda:

Lakes are commonly formed within volcanoes whose peaks have collapsed or were blown off. Among the best examples is Crater Lake in Oregon which was formed during an eruption of Mount Mazama that resulted in the loss of its upper structure. Here are space and ground views:



As will become evident in the next Section, on impact craters, lakes often form in the depressions made when such craters are formed. The best example are the two Clearwater lakes in Canada:

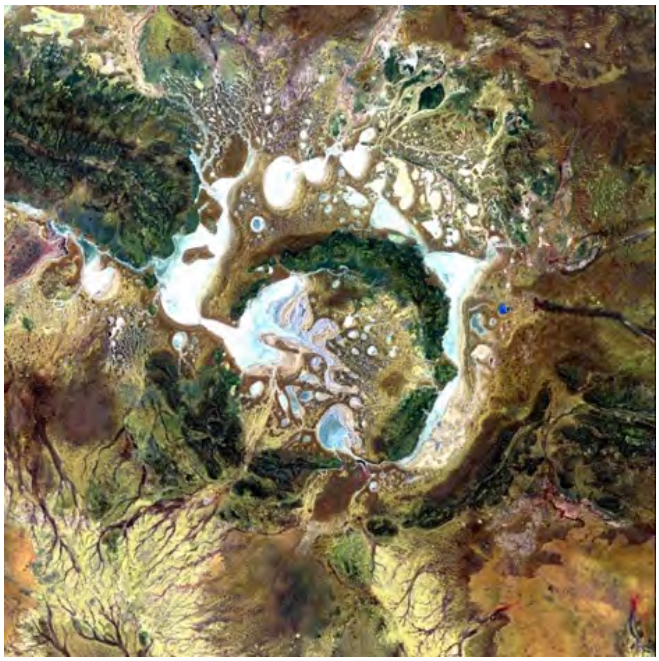
In arid climates, lakes may be ephemeral to the extent that they contain water only part of each year during the rainy season. A playa is a lake bed formed in an arid climate by rapid evaporation

during dry season of wet-season stream inflow causing sediment and salt deposits (evaporites) to accumulate over time. One of the best known large playas is Owens Lake on the east side of California's Sierra Nevada:

Changes are frequent in playa lakes, as rainy seasons vary. Note the difference in the location of the shoreline in Australia's Lake Yamma Yamma:

March 3, 2004 February 22, 2001

Chemicals such as sodium carbonate (Natronite) can collect in playas. Here is an example in the lake associated with the Shoemaker impact crater in Australia (described in the next Section).



This playa lake is famous:

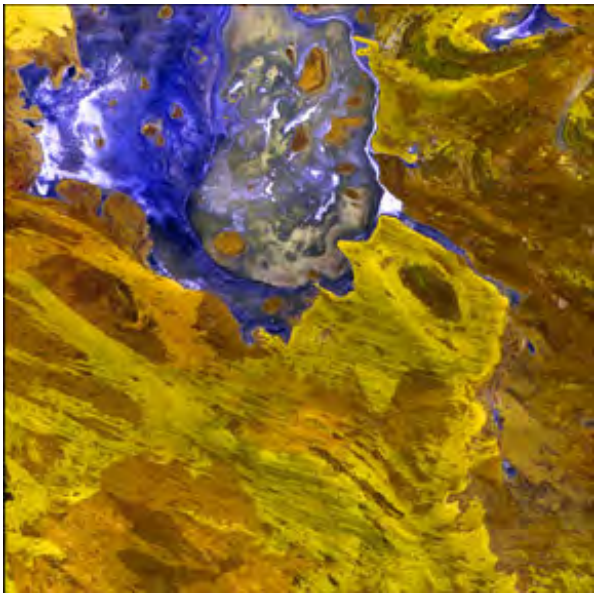
The huge lake (more than 9,000 km^2 [3,475 mi^2]) in the image above is the Salar de Uyuni in Bolivia, the largest salt playa in the world. In this March 1973 image, the salar (salt pan) has about a meter of briny water (blue). Some of the surface is uncovered salt (white), which, in addition to sodium chloride, NaCl (rock salt) and calcium sulfate, $CaSO_4$, also contains lithium chloride, LiCl, making this the biggest source of lithium in the world. This, and the Salar de Coipasa (upper left), were once part of a much bigger lake that had formed in this part of the Altiplano province of the Andes, during a rainier climate in the past.

17-13: What is the eventual likely fate of these salars? ANSWER

The most caustic (Sodium carbonate-rich) lake in the world is Lake Natron in Tanzania. In the view below, the red in this lake is due to an alkaline-loving algae, Spiricum, which attracts more flamingoes to this lake than any other place on Earth:

Similar to some playas/salars are ephemeral lakes - those that fill to shallow depths during extended periods of wet seasons but during droughts dry up until the next periods of rainfall. A good example is Lake Carnegie in Australia; here is one of its filling phases but at other times of the year it dries partially into marshland.

Lake Disappointment in Western Australia is another ephermeral lake that can contain salt brine for years:

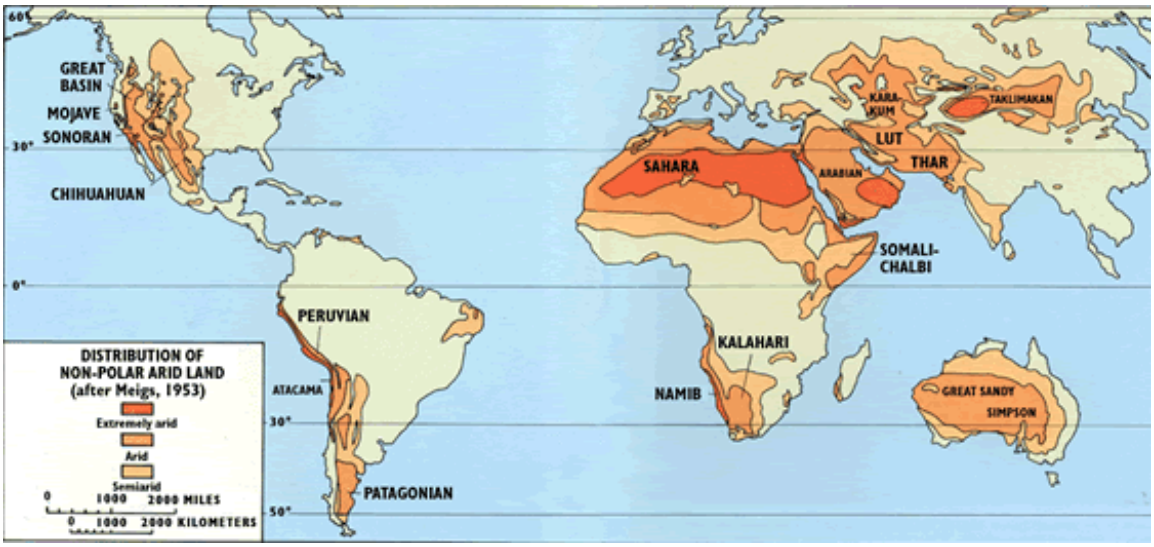


In an odd twist of vegetative activity, look at these two Landsat images of the Great Artesian Basin in east-central Australia: On the left, waters from the gently flowing Barcoo River has flooded the arid land to form a lake-like spillover, known when it is there as Lake Yamma Yamma. The land shows signs of widespread grasses and acacias. But in the dry season, the vegetation

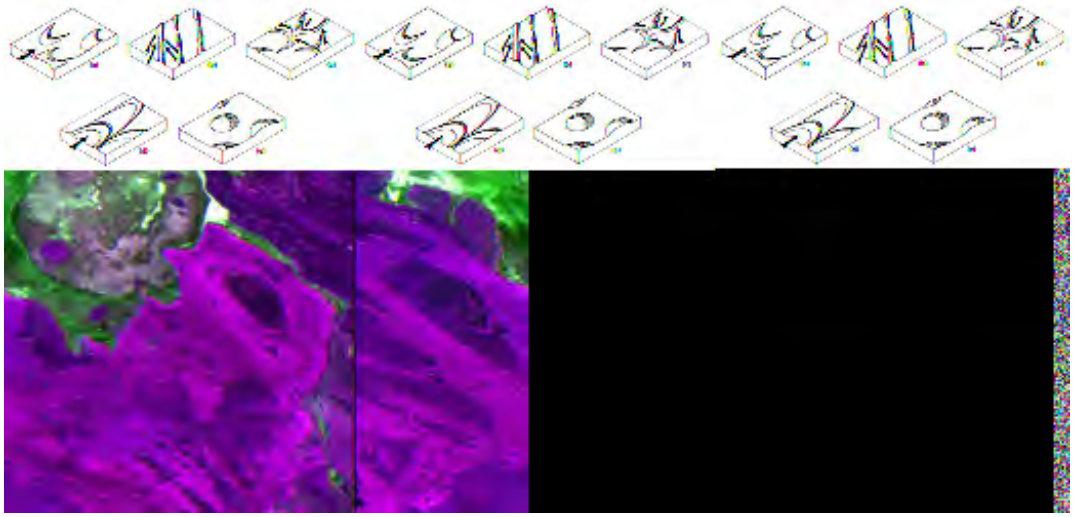
vanishes overland but widespread thick marshy vegetation now fills the lake:

Another lake that is dry most of the year is Lop Nor (nicknamed "The Great Ear of China" by the writer after seeing it in a Landsat image) in the Takla Maklan desert of Sinkiang Province in western China. The area is extremely dry but meltwater from the Tien Shan mountains to the north partly fill the depression each year. Multiple shorelines help to similate an earlike appearance:

A significant fraction of the Earth's land surface can be described as arid. This is primarily a climatic term, implying lower rainfall and distinctive vegetation adapted to the sparsity of water. Much of the arid terrains can also be called deserts. Deserts do not necessarily consist mainly of sand; some are mountaineous but with usually well exposed rocks and thinner soils. This map shows the main deserts of the World; note that most of Australia and a large part of Africa are arid lands:



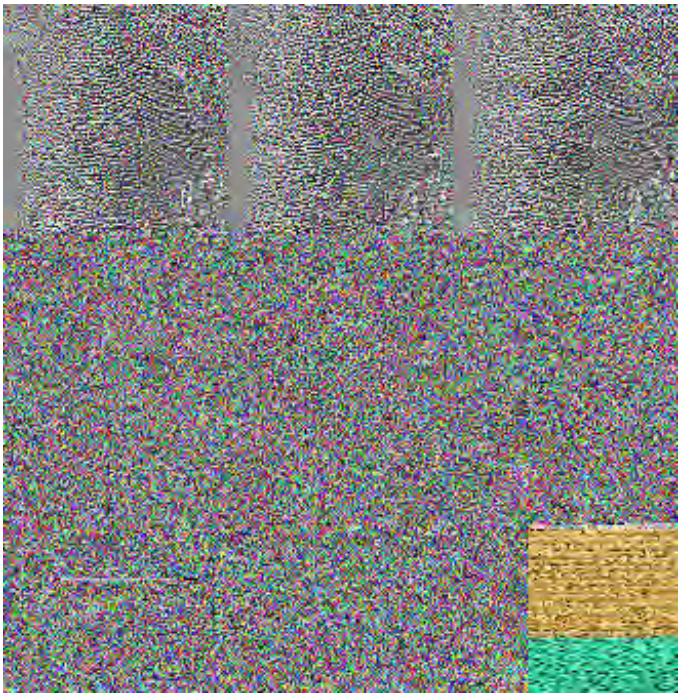
Many landforms characterize arid regions and we would need numerous images to show their variety. We elect to start by showing 8 examples of what many people visualize as the hallmark of desert terrain - sand dunes, which people usually associate with the Sahara Desert that has been the locale for many movies (perhaps the best known: *Beau Geste*). But first, consider this sketch which shows the major types of dunes - given specific names - and the relation of their shapes to prevailing wind directions.



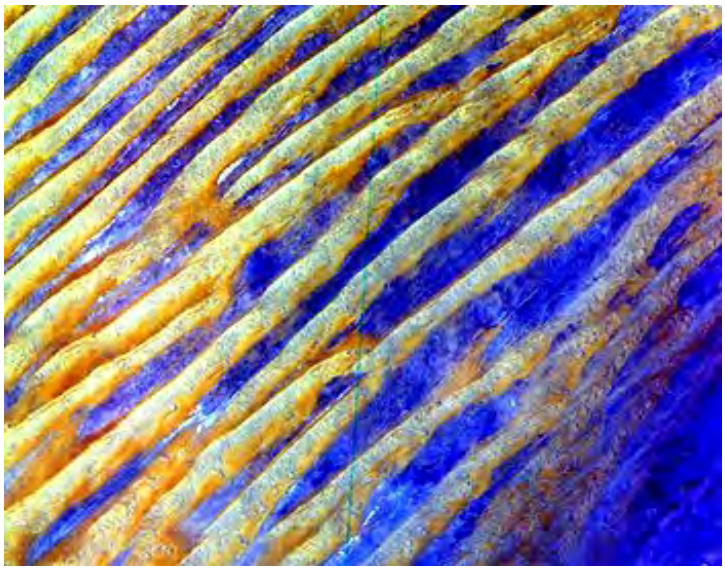
The dune forms are as follows: a) Crescentic dune; b) linear or longitudinal dune; c) star dune; d) parabolic dune; e) dome dune.

Our first Landsat example shows star dunes is found in the Grand Erg Orientale, a Sahara Desert member here at the border of Algeria with Libya. These appear as roughly equi-sided hills rising in some cases more than 100 m (330 ft) above the sandy base:

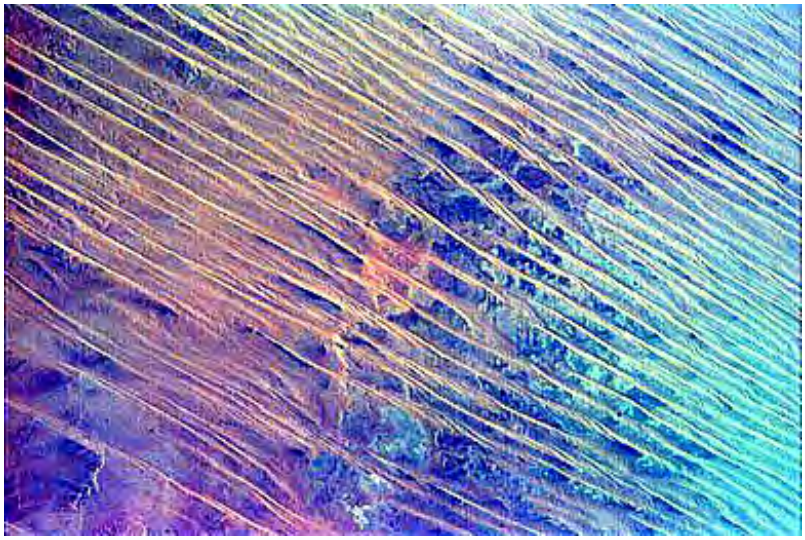
The largest active sand sea on Earth (560,000 km^2 [216,237 mi^2]) is the Rub'al Khali or Empty Quarter, shown above, in the southern Arabian Peninsula. Loose sand, up to several hundred meters thick, covers most of the solid rock landscape in this region. In the Landsat scene below, six distinct sand fields, each characterized by particular types of dunes (distinguished mainly by shape), occupy the entire area. Of these, we mention four. Most of the center consists of Complex Crescentic Dunes on the left and Complex Linear Dunes on the right. In the lower right corner are Star Dunes, and along the left is a separate sea consisting of small Linear and Complex dunes.



The ASTER sensor onboard Terra has generated false-color images of some of these Empty Quarter dunes)longitudinal type) in southern Saudi Arabia. In the version below, dunes are bright yellow and brown whereas the blue relates to the spectral response of interdune clays and silts.

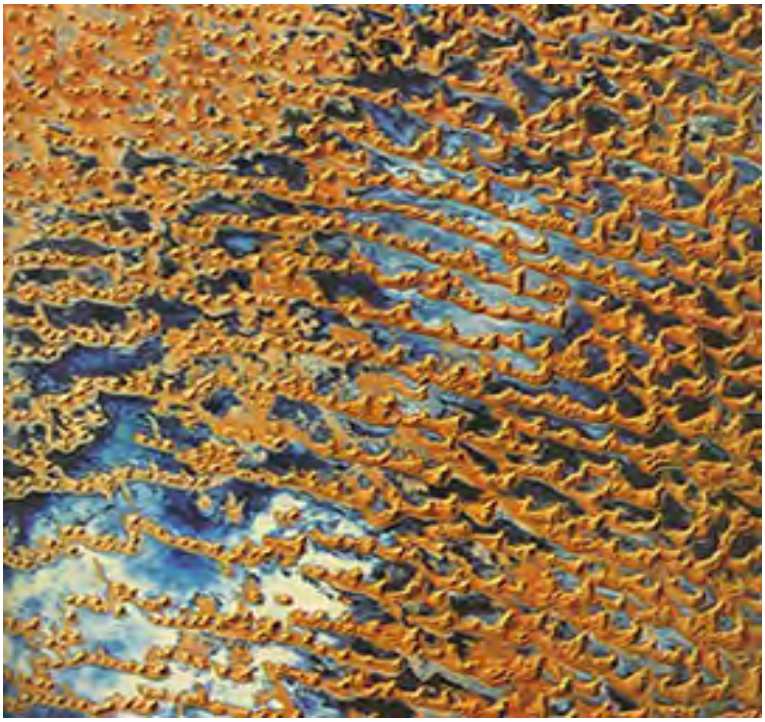


For comparison, look at these more widely spaced longitudinal dunes in the Egyptian desert

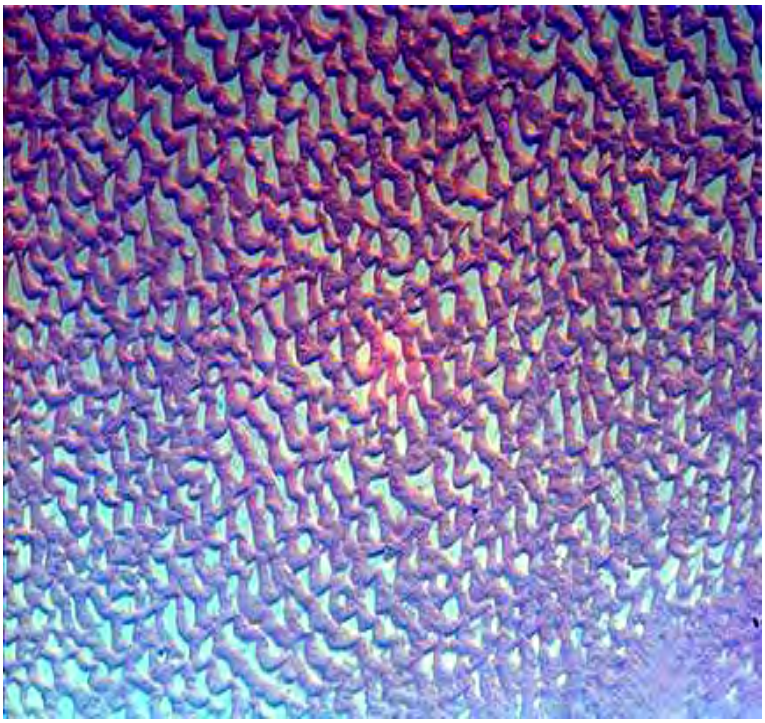


Close-spaced longitudinal dunes are shown here; they are found in Oman in the eastern part of the Arabian Peninsula:

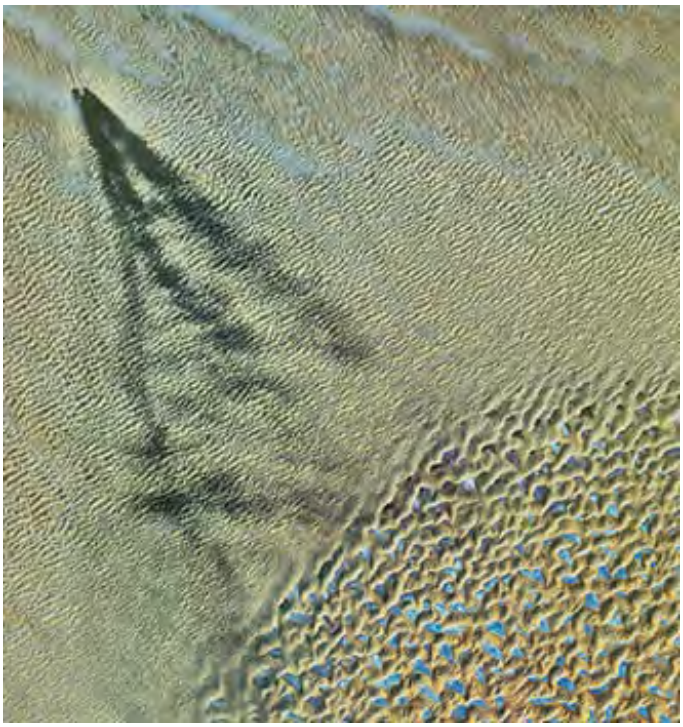
Elsewhere in the Rub' al Khali are these unusual seif dunes, with arrowlike point upwind, and again blue area which denote some surficial moisture.



A variant of these seif dunes, also found in Saudi Arabia, shows them to be extensively interlocked.



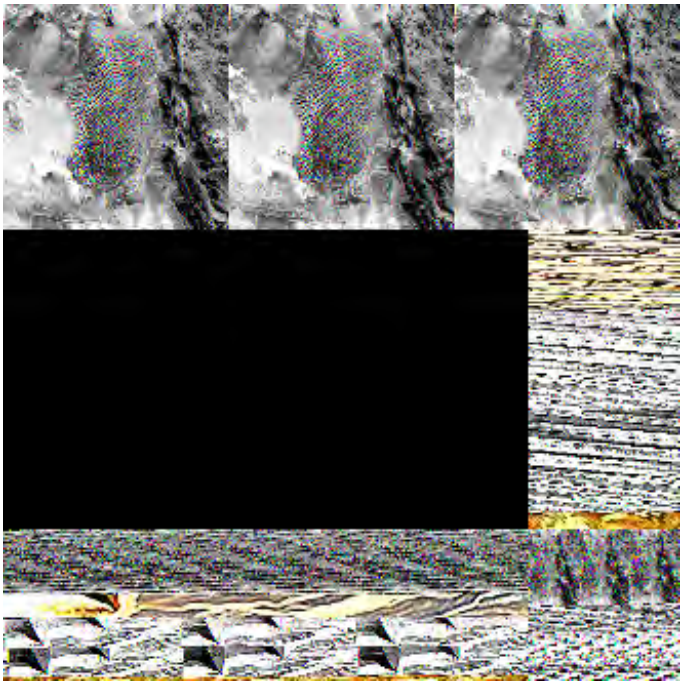
This next Landsat image shows three separate dune fields (note part of one in the upper right) in the Arabian desert. Each is a longitudinal type (the one in the lower right shows some curvature). The black smudges are smoke plumes from an oil well.



Not far away is still another example of multiple types of dunes in an image of Landsat-1 size:

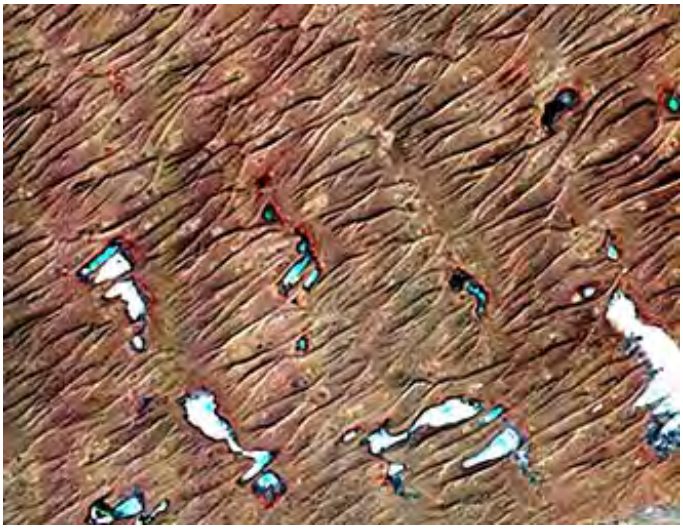
Perspective images of sand dunes provide a different impression of their appearance. The dunes in Namibia are believed to be the oldest field on Earth. They also can show up to 100 m (330 ft) of relief. Here is an oblique view made by combining ASTER imagery with DEM data:

The image below is part of the Dasht-E-Lut, or stony desert, located in southeast Iran. Dominating the scene is a large, but isolated, sand sea containing linear or elongated star dunes with a prevailing northeast trend. The surrounding lowlands have gravelly surfaces. Along the lower left edge of this Landsat image are dark streaks that orient towards the southeast. These streaks are the eastern end of a big complex of long, narrow ridges called yardangs, left behind when winds blew out soft clay/sand sediments, leaving trenches. As the trenches elongated, they controlled the eventual scooping out of the linear valleys that defined the residual yardangs. In this scene, exposed bedrock comprises the low mountains to the east of the sand sea, from which streams carry and deposit dark rocks in a series of distributaries and fans.



These two ASTER images show details of the yardang field (left) and the dunes (right; some as high as 300 m - among the tallest in the world) intermingled with white salt pans:

Further to the east, in the Thar Desert of western India, strongly crescentic dunes have formed, as well displayed in this ASTER subscene:



17-14: Can sand dunes ever be vegetated? ANSWER

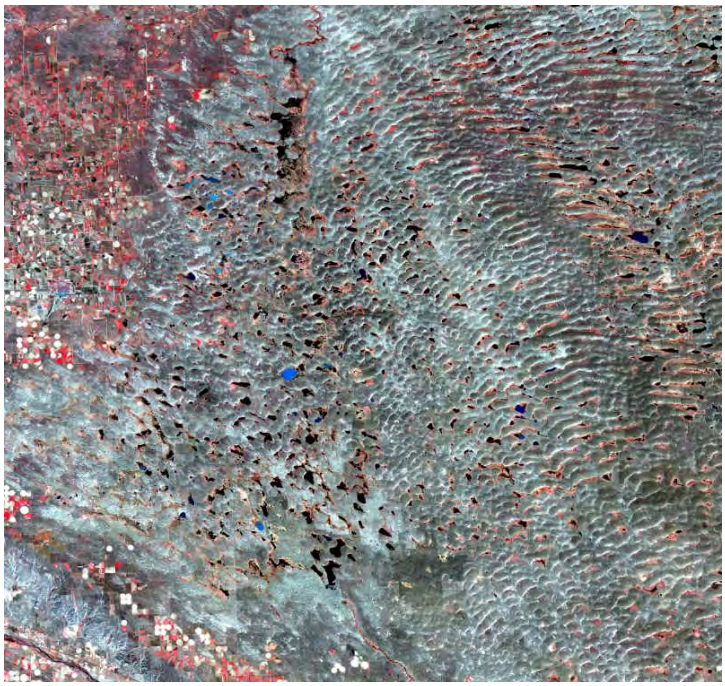
Here are two Landsat images of dunes in Africa that have enough vegetation to support wildlife:

One example of dunes hosting vegetation is found in the Kalahari Desert in Namibia. Here dunes are migrating northward, overriding natural vegetation and farms, leaving some growth to still persist as more sand continues to move to the grasslands beyond the image top.

Another dune region, in Mali, is crossed by the Niger River and its secondary streams, which provide enough water for vegetation to develop locally (red patches in this Landsat TM scene).

A similar situation exists within dunes around the rapidly drying up Lake Chad. Vegetation now has profuse growth that has stabilized the dunes.

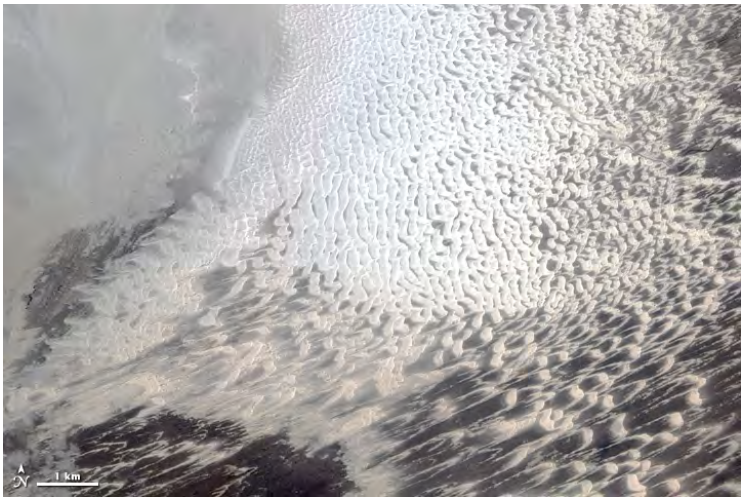
We saw an example of vegetation gaining a foothold among dunes in the Great Sand Hills of Nebraska (Section 6). This Landsat image shows the same scene, now in color, showing the distribution of vegetation (the reds) within the Sand Hills; an aerial photo is beneath it:



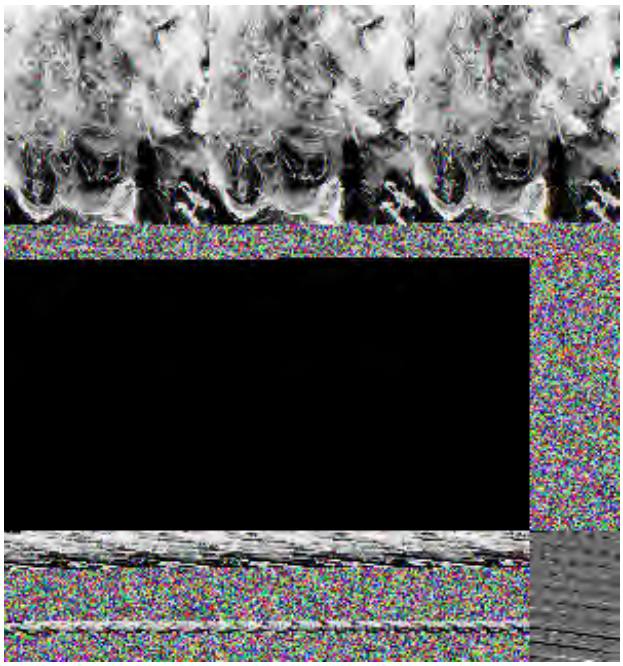
Here is a similar case: the vegetated dunes are near the Caspian Sea where the Volga River is

emptying its flow:

The vast majority of dune fields are made up of quartz sand. But there are exceptions: dunes made up of clastic particles of composition other than SiO_2. Best known are the White Sands (National Monument) of New Mexico (see page 6-6). These sands are composed of Gypsum (hydrated Calcium Sulphate). At one time the gypsum deposits were laid down in lake beds. Once the lake dried up, this mineral was broken into sand-sized grains that have been dispersed by the wind into the deposits shown in this image made by the ALI (Advanced Land Imager) sensor on NASA's EO-1:

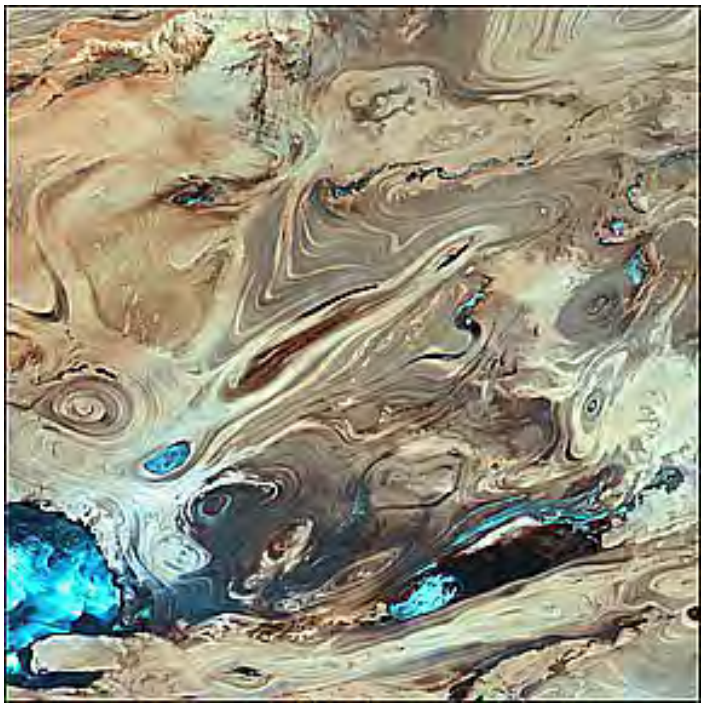


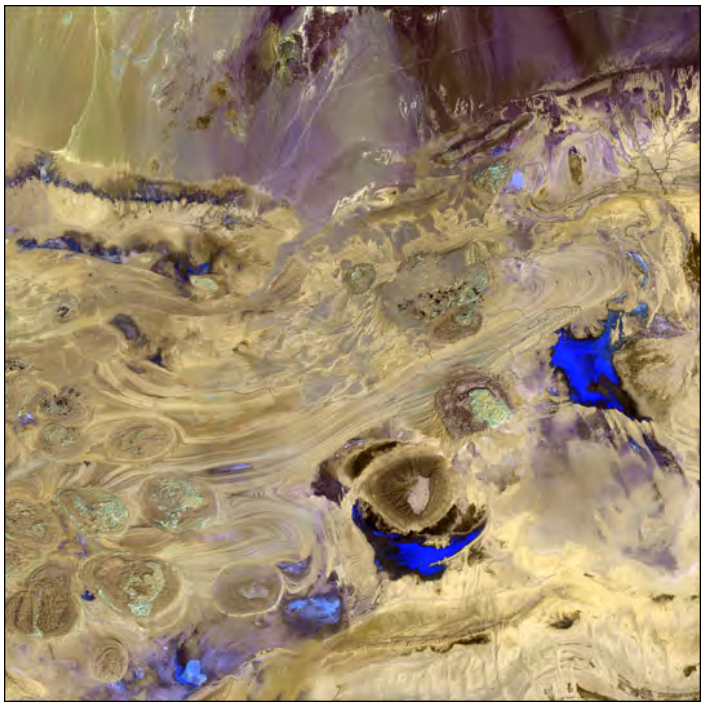
The above aeolian views feature sand dunes. But many have the misconception that deserts are mostly sand. In fact, most of the desert terrain in the world is rocky, often with gravel surfaces or thin fragmental soil. Some weird terranes can form in arid lands where the wind blows persistently. Sand blast sculpturing is a common process, leaving odd-shaped forms that in space imagery really stands out. Consider this Landsat image of the Tanezrouft Basin of southern Algeria. The Sahara, of which this a part, is mainly rocky rather than sandy in cover. Here sculpturing has carved out deflation basins within horizontal to gently dipping sedimentary rocks. Patterns that tend towards circular are the mark of this kind of terrain, but on the ground the effect would probably go unnoticed.



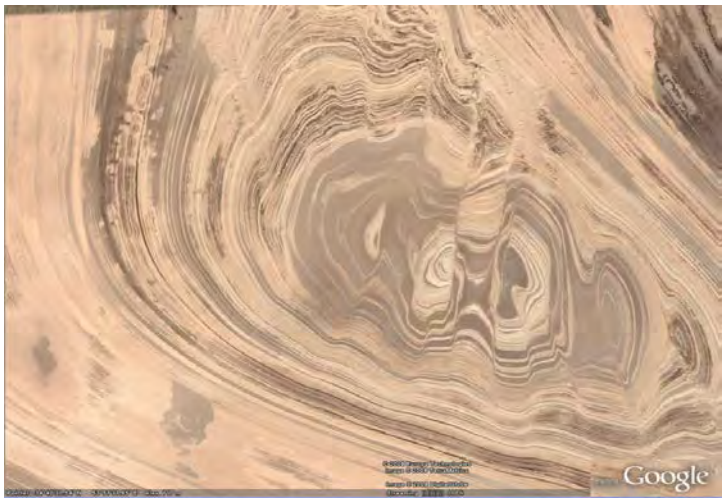
This color TM image is typical of the Sahara Desert region, being a mix of sand-filled valleys and low hard rock mountains:

Deserts can also contain other substances besides sand. Much of Iran southeast of Tehran is a desert region in which large stretches of land are periodically under water in a brief wet season. Upon drying, playas and salt flats result. Over time, some of these erstwhile lakes are marked by wide bands of sediments, as seen from space. The salt marshes are known locally as kavirs. Although rocky and sand surfaces exist in Iran's Dasht-e-Kavir (see above), salt (NaCl) is mixed with muds in the lower, playalike areas. Here are two Landsat images of this colorful land.

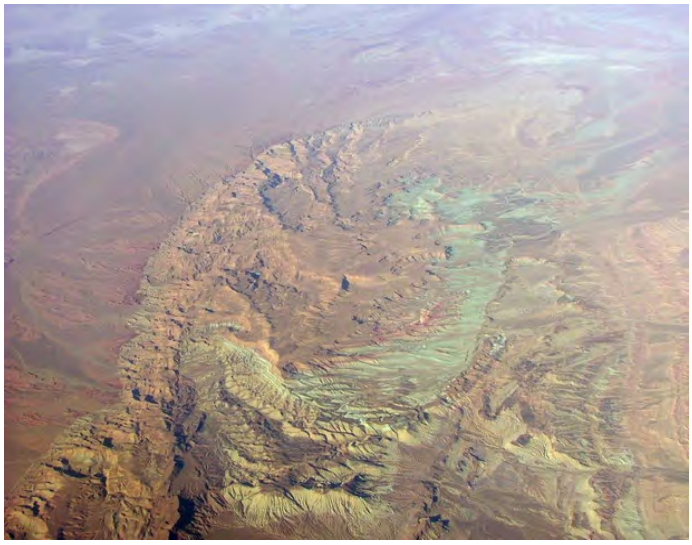




The next image is controversial. Its Internet source states the view to be part of the Kavir Buzurg and claims the feature to be a faulted tectonic fold. While this may be the case, the writer (NMS) believes it to be the markings of a dried-up lake rather than sedimentary layers; there is a fault sliver in the center but it seems likely to be the result of underlying ground displacements that have affected the lake beds. Judge for yourself.



There is obvious tectonic activity in the Kavir region. This is a typical landscape.

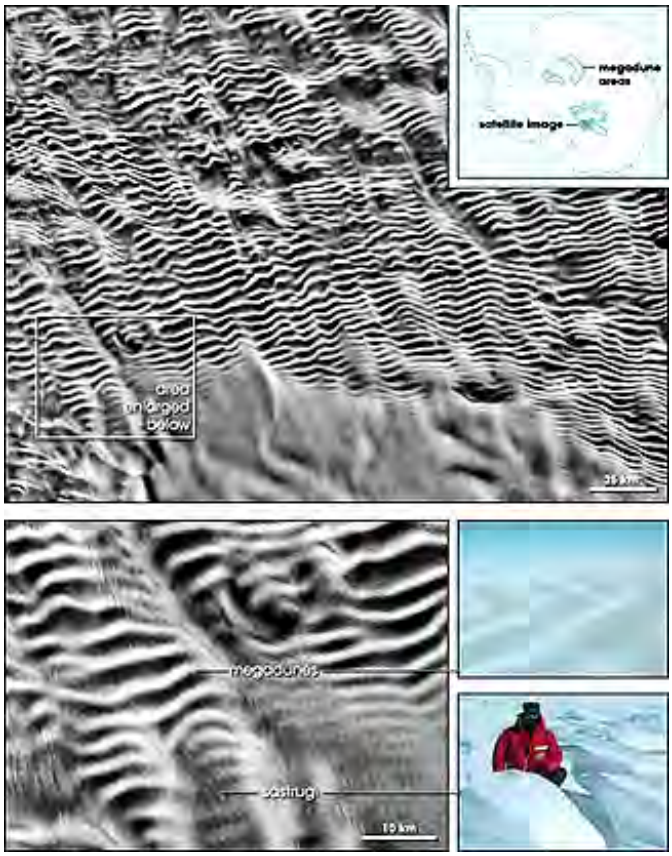


This next image contains a real oddity. The locale is the Atacama Desert that was discussed on page 6-11. Make an educated guess as to what the features within the rectangle might be. Hint: images elsewhere on this page and on page 6-10 afford clues. Look now:

These are salt pans similar to those in south San Franciso Bay, in terms of colors. The salars shown above also display evaporite beds. The greens are caused by algae.

Sand dunes usually are made up of sand- and silt-sized particles, most commonly of quartz. But it is possible for winds to produce regular-spaced dunes composed of snow flakes and particles, as shown in this image of a snow field (the full extent of which is almost as spacious as California) in the East Antarctic Plateau. The dunes are up to 8 meters (25 ft) high and 2 to 6 km

(1.2 - 3.7 miles) from crest to crest. Although the region was traversed for years prior to this image, the true nature of this drift field was not realized until space imagery brought out the pattern seen below:



This last image takes us to the next topical landform: glaciers and glacial landscapes.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Following completion of Geomorphology from Space, this Tutorial's prime writer (NMS) felt it would be instructive to further test the scientific utility of space imagery in analyzing the Earth's landforms. He decided to conduct a pilot study of the Klamath Mountains of southwestern Oregon. The project was dubbed "Terranes as Terrains". A terrane is a structural geology term that applies to large masses of rock - island arcs, subcontinental fragments of crust broken loose from tectonic plates, and other rock units embedded on a moving plate - that move along a spreading plate until they collide with the upside rock bodies at a subduction zone. Instead of going under in subduction, they are "welded" to the body (it may be a continent) and are thus added to a continental margin. The continent grows by accretion in this manner. Along western North America, where subduction has been active for hundreds of millions of years, there are many accreted terranes. Since each is generally different from the others in rock type and structural style, any landforms (terrains) imposed on individual terranes by the usual geomorphic processes are likely to be different (and probably thus distinguishable) from one another. This is the hypothesis that was tested. This page expounds upon this background and shows the terrane makeup of both the West Coast (up to Alaska) and the Klamath Mountains (composed of many terranes).

Terranes as Terrains: The Klamath Mountains Oregon Study

As stated earlier in this Section, in 1985 the writer (NMS) initiated a research project while still at NASA's Goddard Space Flight Center, on which I continued to work until retiring from there in September, 1988. The supporting field effort was conducted while on sabbatical from Goddard between June and September of 1985. Although presented as a paper twice at professional meetings, this work remains unpublished because I deem it to be still incomplete, with some open questions that would benefit from more research (unlikely, as funding is no longer accessible). However, the study did produce enough valid and meaningful results to justify putting forth the main data displays and analytical inferences as a model for the role of space imagery in a basic geoscience investigation.

The title of one of the presented papers, "Terranes as Terrains," aptly summarizes the essence of the study. Freely translated, this refers to an attempt to determine whether certain kinds of geo-structural units (*terranes*) have distinctive and diagnostic landscape features (*terrains*) that make them easy to identify and that help us understand their origin, significance, and history. Three goals prompted this investigation:

I. To define and describe, qualitatively and quantitatively, the distinguishing geomorphic characteristics of known (previously mapped) terranes,
II. To determine whether individual terranes in a group have geomorphic signatures sufficiently different to permit us to recognize and separate them in aerial/space imagery, and to better delineate boundaries between terranes based on terrain differences,
III. To assess geomorphic information about terranes as clues to modes of emplacement and structural imprints.

The study applies morphometry to conventional and digital topographic databases and, wherever useful, Landsat/SPOT imagery.

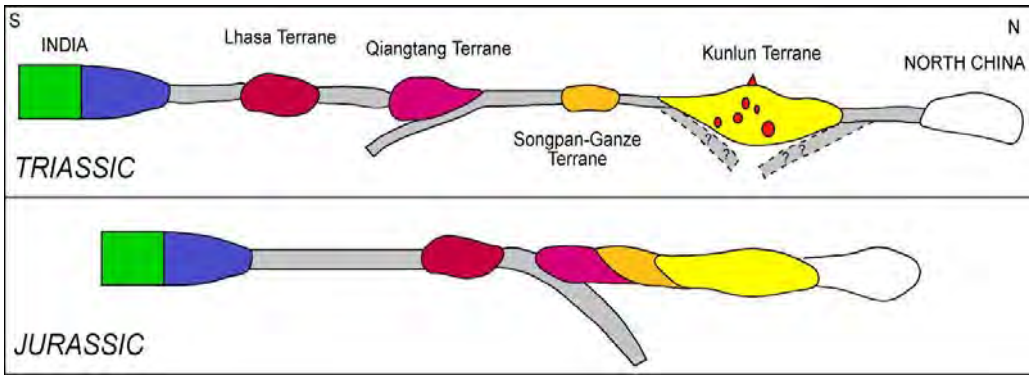
The word "terrane" originally referred to "a crustal block, usually bounded by faults, with a geologic history distinct from histories of juxtaposed blocks of differing characteristics." With the

advent and acceptance of plate tectonic theory, users modified this concept with new insights into certain crustal blocks that they called "tectonostratigraphic", "accreted" terranes. They have also been referred to as "suspect" or "exotic" terranes but these variants are not now commonly used. Terranes are considered to be diverse segments of crust that originated and developed often far from their present locations and over time traveled on a moving plate that converged with another plate, eventually colliding with a continental margin (along or within this second plate). There, because these crustal bodies are usually of moderate densities, they are obducted (shoved onto and overriding) against the margin, rather than more or less subducting (diving under) the second plate, thereby becoming accreted on or against that plate. A brief overview of terranes is given on this Wikipedia website.

. Also, you may have already encountered the "terrane" concept earlier if you read through page page 2-1b; revisit the part of that page dealing with terranes if you need to refresh your memory.

17-16: To help you visualize what happens in continental breakup, drift, and reassembly, imagine (or you can actually do this) a spherical globe (like the one used in school to show the geography of the world), and ignore the actual continents on it. Place on it a piece of writing paper cut into a circle (about 4 inches in diameter). On the opposite side place another circular piece, perhaps of smaller diameter. Now, cut the paper into 5 to 6 individual pieces. Have each piece then slide across the face of the globe but in different directions, all moving simultaneously. Eventually, any of the pieces will meet the other smaller circular paper (assume it to stay fixed in place, but it could itself move as a unit or could be cut into several pieces that move [drift] apart - but this gets complicated even if it depicts a real possibility). As the spreading pieces of the first circular paper meet the second, they can be envisioned as lapping onto the second paper and becoming part of it, i.e., accreted by analogy onto what is a separate continent. In the accreted terrane paradigm, this process goes on continuously and involves many terrane units,large and small, repeated collisions, and eventual rupture of any continent that gets quite large. The continents, however, never get so large and stable that they cover the entire Earth, but instead, the spaces between them are made up of non-continental oceanic crust which is mainly basaltic lavas released to either side of mid-oceanic ridges that build up from lavas derived in the mantle; movement of these lavas away from the ridges leads to ocean "spreading" and drives plates against each other. ANSWER

To set the stage for the following review of the Klamath Mountains as terranes, we present this pictorial example: Here are two stages of the growth of the Himalayas. Individual land masses (these can be large, hundreds to thousands of kilometers on a side) riding on an oceanic plate that is subducting into southern Asia are in succession added to the Asian continent. Eventually in Cenozoic time the entire subcontinent of India will be welded against the Asian plate:

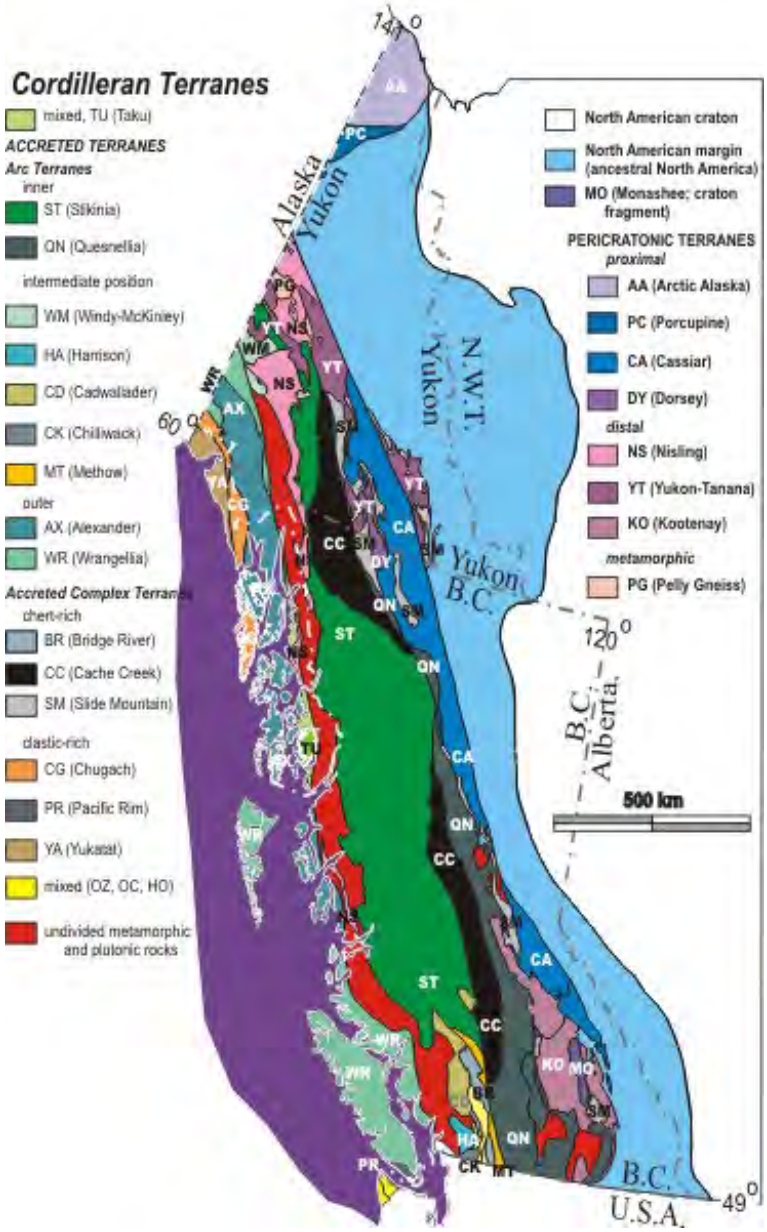


Thus, masses such as microcontinents embedded in oceanic crust moving to a subduction zone may be unable to follow that crust downward and will be "scraped off" along thrust faults to emplace along the growing edge of the continental crust on the other plate. Or, a variant: if an island arc complex (e.g., Java, on the upper plate, above a subduction zone gets squeezed between two converging plates as a continent on the lower plate eventually reached the zone,

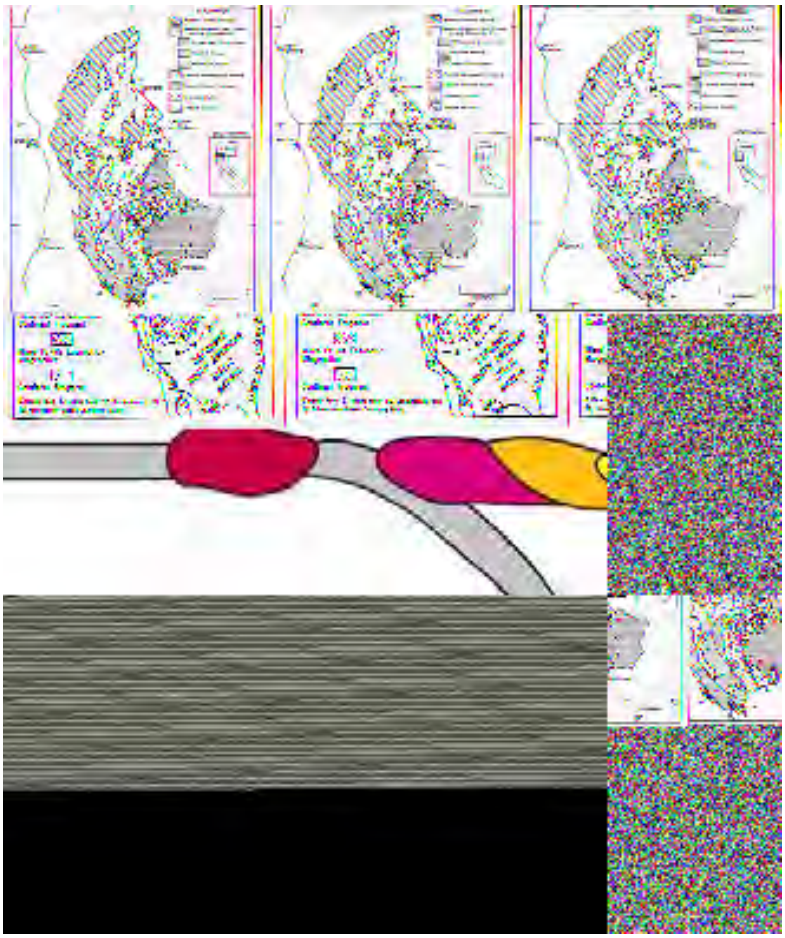
much of the arc would be caught "in the crunch" to be driven on and welded to (accreted) the encroaching continent. The usual criteria for terrane recognition come from stratigraphic, structural, and paleomagnetic discontinuities between adjacent terranes. (

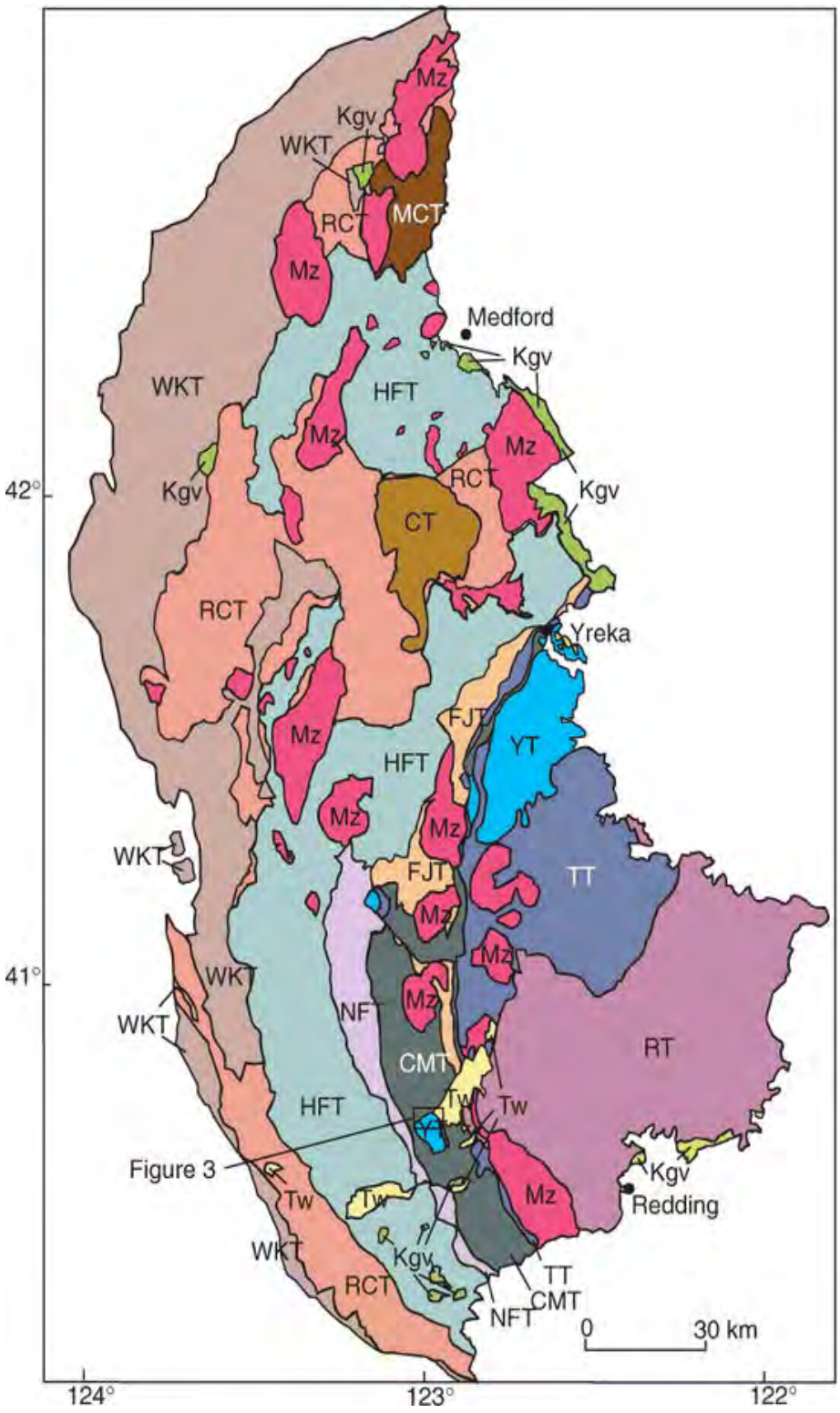
Whereas in the thinking about growth of continents prior to the accreted terrane concepts, scientists explained complex geo-crustal units in the continents as the consequence of fold-squeezing and fault transport during mountain building at geosynclinal sites, without much geographic displacement. Now, the new view recognizes distant site origins and relocations because of long movements on the "conveyor belts" of diverging plates. Continents thus grow by progressive accretion of numerous plates over time onto early cratonic nuclei. Since this concept first gained favor in the 1970s, geoscientists have recognized hundreds of terranes on all continents, in the sense that we re-interpreted older ideas about mapped geologic units to fit the terrane model. As an illustration of the terrane addition/continental growth version, consider this map that portrays a collage of the major terranes emplaced along the western margin of North America.

This map concentrates on that part of the western Cordillera (mountain system) that is found in Canada; the terranes shown lie against the older North American craton:



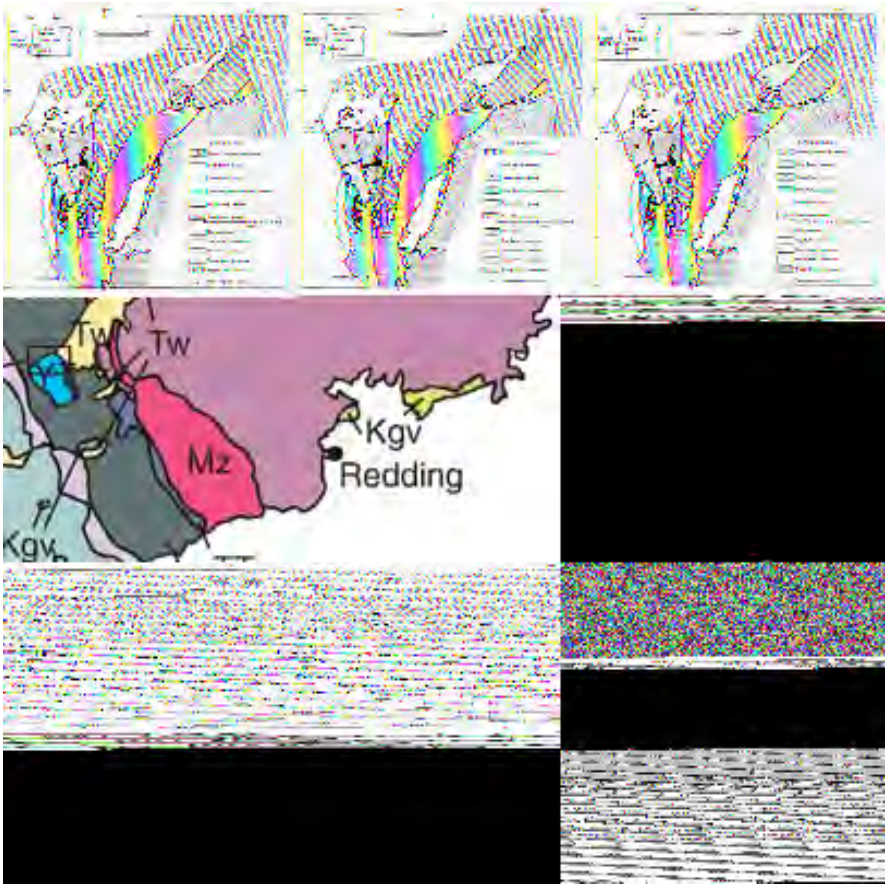
The Klamath Mountains of northwestern California-Southwestern Oregon were the type of locality for developing the concept of accreted terranes, because of the pioneering mapping and interpreting by Irwin, Blake, and others. Early, somewhat generalized, maps of the two adjacent areas display the first interpretations of these groups of terranes (the upper map that follows shows the terranes in northern California and part of the Oregon Klamaths, below it is a colorized version (but without a key) that helps to single out individual terranes:





The above two maps have similarities but the various units seem to differ. Neither map (taken from the Internet) has a usable legend. Among the major terranes found in northern California are the Rattlesnake Creek terrane, the North Fork terrane, the West Hayfork terrane, and the Salmon River terrane.

The map below shows just the Oregon segment of the terrane complex; this is the reference map we will use in the geomorphic study of the Klamath terranes.



This last figure also has a poorly reproduced legend. So, we have constructed a list of the terrane units as follows; not that none of these names seem to have any counterpart on the first two maps shown:

BC = Briggs Creek
DB = Dry Butte
ER = Elk River
GB = Gold Beach
PP = Pickett Peak
RV = Rogue Valley
SM = Smith River
SR = Sixes River
YB = Yolla Bolly

Each of the named Oregon terranes is a unit which, after travelling some distance, "docked" against the paleo-North American continent, during some specific period. Most of the inner (continentward) units are progressively older than those outside, although thrust faulting can sometimes carry a younger unit over part of older ones. Each terrane maintains some internal structural integrity and consists of a series of stratigraphic units (formations), many not having counterparts in other nearby terranes. Those that do were likely deposited across boundaries after terrane emplacements.

Because each terrane seems to have originated from a different part of the globe, has different lithologies and ages, and "docked" at a different time, the writer reasoned that its topographic/geomorphic expression ought to differ from its neighbor. This was the premise that led to the investigation that is presented in the next four pages. The "short answer" previewed here is that real differences were determined and quantified.

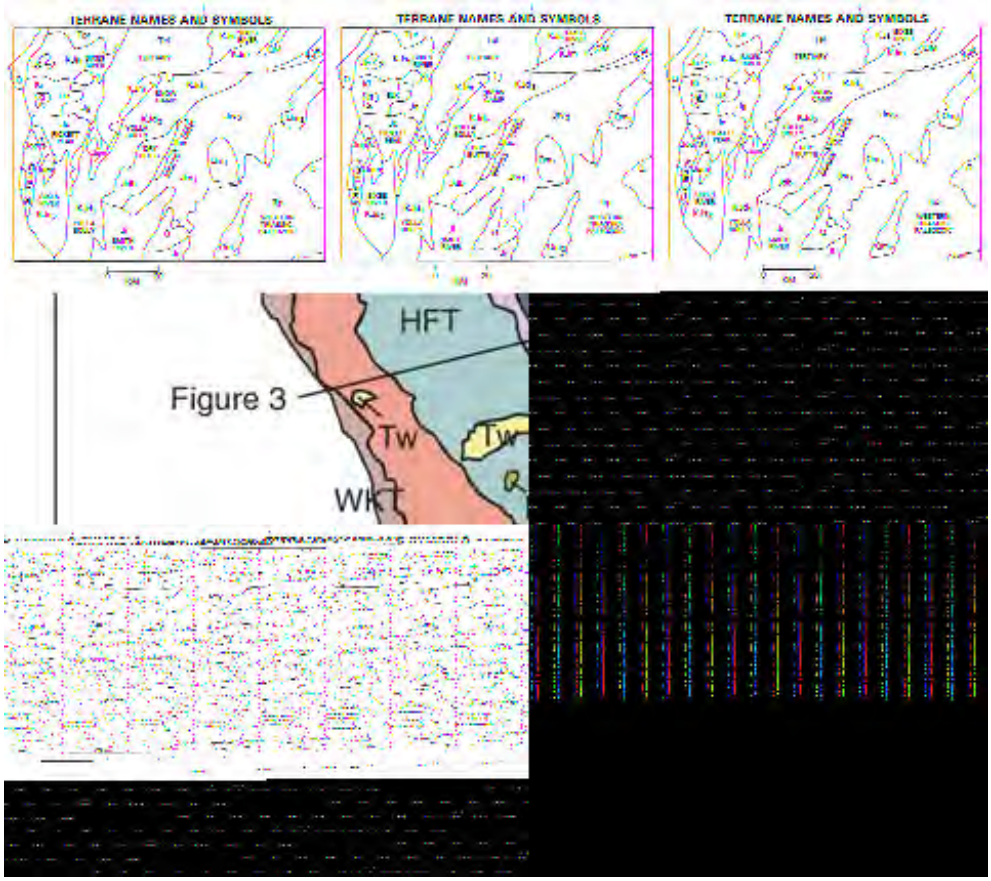
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

The Klamath Mountains of Oregon are described and visualized through ground photos. The conventional geologic map (rock units are mapped but the context is not that of terrane units) is shown and the map that characterizes the assemblage of stratigraphic units as distinct terranes is presented. DEM data are used to make shaded relief portrayals of all of Oregon and of the Klamaths within. The strategy guiding the study is stated in terms of the principal parameters to be measured, both by standard geomorphic methods and through use of space imagery.

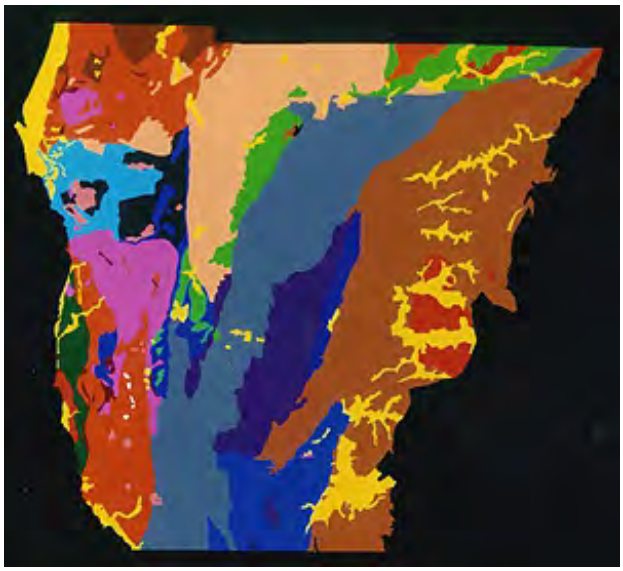
The Klamaths from Space

For this geomorphology research project, I selected a segment of the Oregon Klamath Mountains for prime attention, although for some tasks I also looked at adjacent areas. The following map displays the area, on which I drew terrane boundaries, labeled the terranes, and assigned stratigraphic symbols to each (the first letter(s) denote(s) their ages in terms of geologic periods, with Tr = Triassic; J = Jurassic; K = Cretaceous; Tel = Tertiary; Q = Quaternary). I digitized the boundaries on the map to allow "cookie-cutting" of geographic data in other maps, so that he could group the data into categories for terrane analysis.



As a visual aid to realizing that there are distinctly different terranes, we reproduce here a simple computer-generated map (unlabelled) that assigns a different color to each terrane. Match with the map above. The easternmost terrane shown is J_{rv}, which stands for Rogue Valley (Western Triassic-Paleozoic may not be a terrane). As a guide, the Brown stands for Rogue Valley, the Dark Blue, Medium Blue and Blue Gray denote Dry Butte, Smith River, and Yolla Bolly respectively; the Light Blue marks Elk River, the Purplish tone specifies Pickett Peak, the

Brownish-red unit against the ocean is Sixes River and the Medium Tan represents the Tertiary units. Yellow denotes Quaternary deposits - both river and lowlands fill; Red near the right corresponds to intrusive granites (Quartz Monzonites).



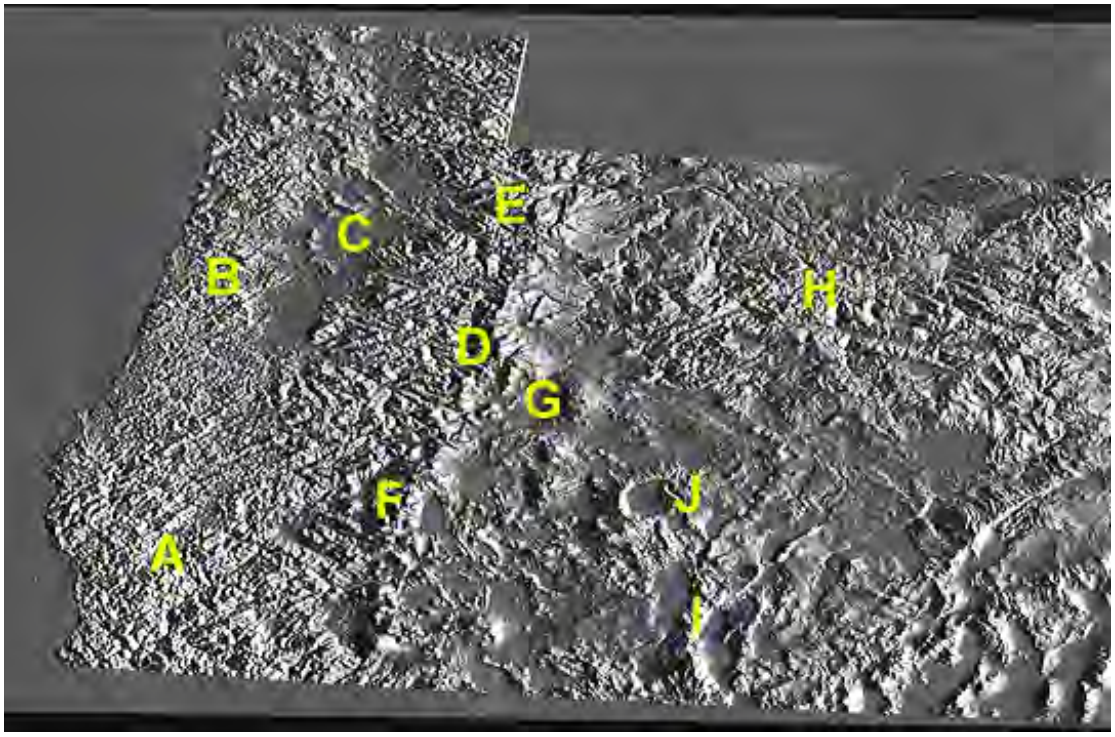
By comparing the above terranes map with the geologic-units map for all of western Oregon (below), you may be able to see that there is some correspondence between some units and their enclosing terranes, a broader correspondence for others, a number of individual units that are scattered about within the terranes, and some units that seem to cross over or overlap terrain boundaries.

Because there are so many units [generally, at the formation level], I don't show the map's legend. Also, in the legend sequence are units that aren't present in the Klamaths. Two examples of exceptions are the deep purple unit (J_{ui}), which is actually outliers of ultramafic (ophiolitic) lavas, once part of oceanic crust, which appear similar in the field but originated locally within their terranes, and the pinkish-red unit, which consists of intrusive rocks that invaded several terranes, already in place. One expects this general agreement between terranes and their constituent stratigraphic units (i.e., internal consistency), along with the exclusivity of some

units to single terranes, inasmuch as terranes typically develop from rocks formed in different source areas at different times.

17-17: Although it will require considerable scrolling up and down on this page, go ahead and try to match the different color units in the geologic map with the terrane units in the map above. ANSWER

The Oregon Klamaths are among the most rugged landforms units in the state. We can gain a sense of their appearance by inspecting this 3-D perspective image of nearly all of Oregon, generated from DEM data:

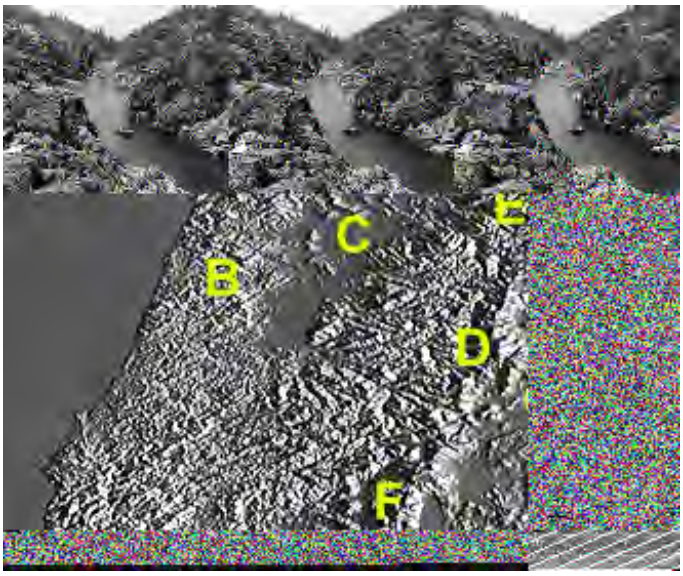


Features of interest in this shaded map are: A = Klamath Mountains.; B = Coast Ranges; C = Willamette Valley (Portland at the north end); D = High Cascades; E = Mount Hood; F = Crater Lake; G = Paulina Mountains (Newberry Crater); H = Blue Mountains.; I = Abert Rim; J = Summer Lake area (the circular area is somewhat an artifact of the illumination)

You can get a feel for the appearance from the ground of several of the Klamath terranes through these photographs taken during the writer's time in the field:

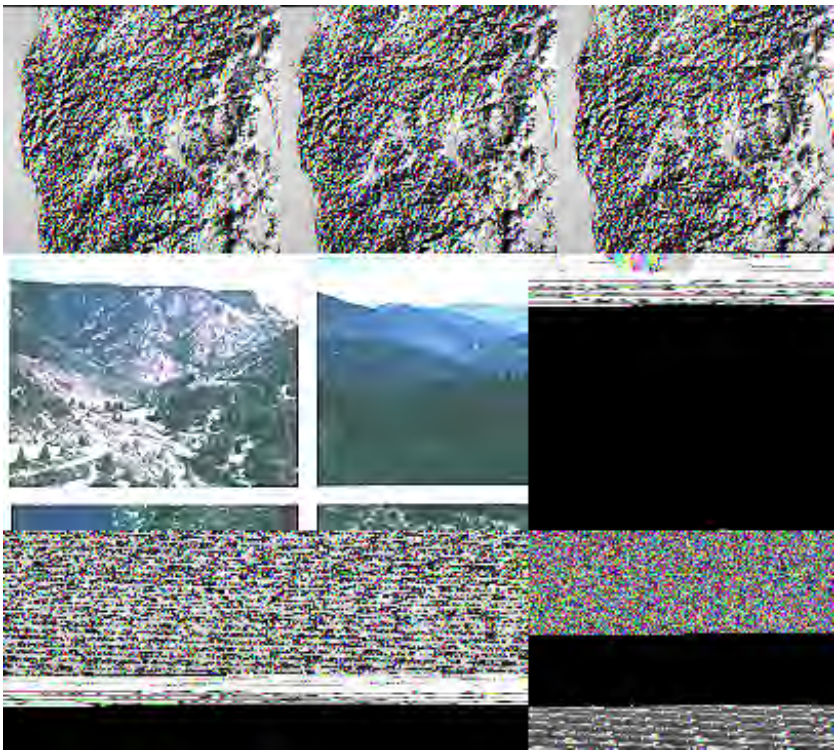
The upper left scene shows terrain around Canyonville, with Yolla Bolly terrane in the foreground and Sixes River terrane in the distance. Next to it is a view from the coast of the Elk terrane. In left center is a segment of a canyon dissected by the Rogue River, in which sedimentary units of the Rogue Valley terrane (J_{rv}) are exposed. To its right is landscape typical of the higher elevations in the Klamath Mountains. The lower left scene is an exposure of serpentinite (the metamorphosed ophiolitic basalts associated with oceanic crust). The lower right shows alternating layers of siliceous (cherty) shales that constitute island arc sediments. Here they are part of the Yolla Bolly terrane.

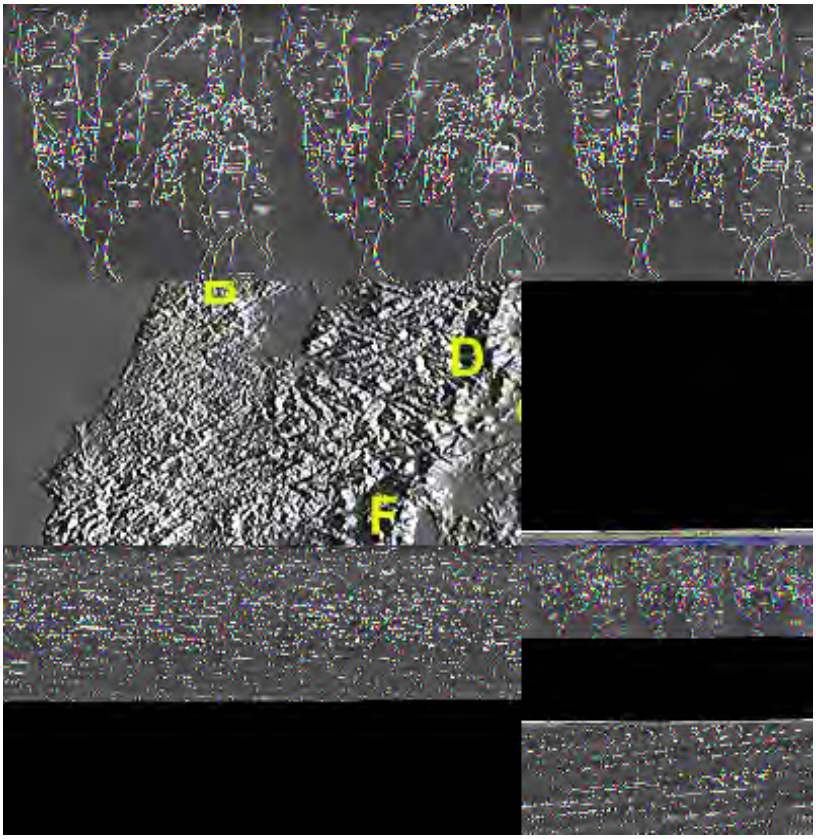
The wildest part of the Klamaths is the Rogue River which cuts deep canyons into the rocks. This is one of America's premium whitewater rafting streams. This photo captures it ruggedness:



We proffer a preview of the topographic expression of the Klamath Mountains, which will be presented in Landsat format shortly, by this enlargement of the area made from the same DEM data. Notice Crater Lake (nearly circular on top of a mountain) in the upper right part. Another simple terrane map is placed below the DEM image. This map is a photo negative of a transparency (terrane boundaries in black) made to superpose onto a Landsat mosaic

(discussed later).

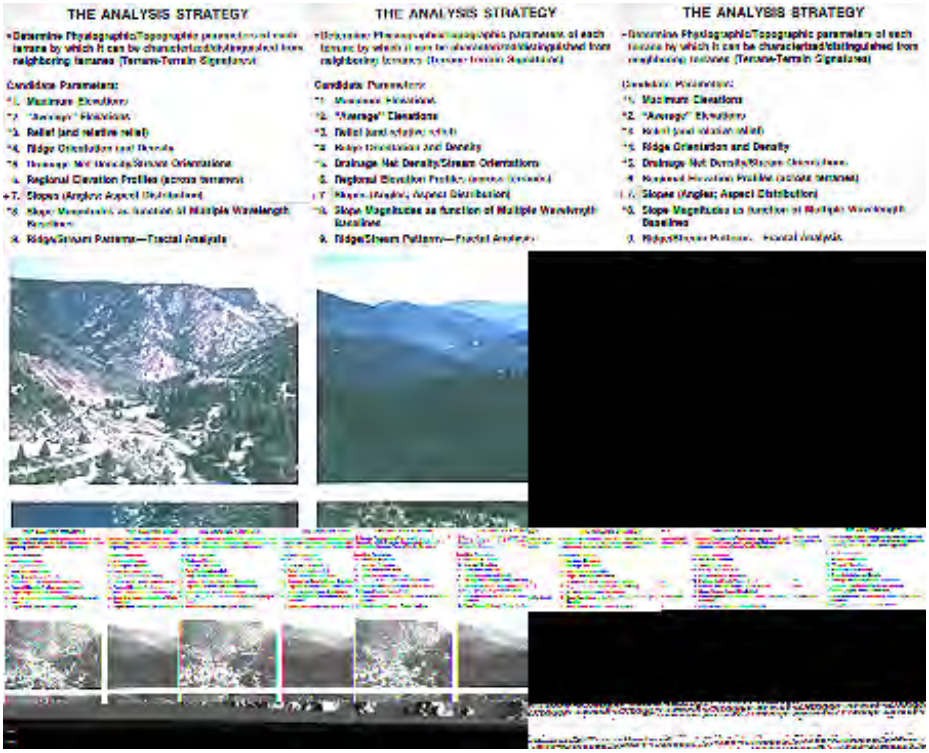




In this visual version, we have trouble seeing any significant variations in terrain that call attention to noticeable terrane differences. Of course, we can vary the illumination directions to emphasize contrasted ridge and valley orientations. We did vary the illumination directions on this data set, giving several distinct expressions but, again, obvious patterns that correlate with the various terranes as mapped did not stand forth. However, when one learns where to look, after familiarization with terrane locations, strong hints of certain differences for several of the terranes are discernible. These differences can also be disclosed by appropriate analysis of topographic data in map-sheet or DEM formats and, in fact, this approach proved superior to visual differentiation.

The reason we would postulate for expected geomorphic differences is that each terrane consists of a group of rock types that normally differ from other nearby terranes - this is known from field work. Fault discontinuities also bound each one, at least partly. A terrane will respond to regional erosive action according to its mix of lithologies. Thus, any one terrane may develop landform characteristics that differ from its neighbors and can appear visually as separable. We should be cautious about expecting the differences to stand out sharply across terrane boundaries, because, as the landscape develops within any one terrane, the equilibrium forms tend to exert some influence on terrains outside the boundary. Nevertheless, the hope in testing the hypothesis of distinctive terrane-controlled topographic expression is that real differences do occur when each terrane is considered in its totality as compared to its neighbors.

In planning the geomorphic analysis, I set forth this strategy:



This effort is predicated on my being able to make a series of quantitative measurements that can be tested numerically and statistically to ascertain valid differences in the topographic character of each terrane. As it turned out, most of the measurements could not be made from Landsat as a test bed. In general, it is quite difficult to extract any of these measures with confidence from Landsat imagery. However, one overlapping SPOT image pair was available, so that a stereo image is achieved, allowing limited recovery of measurable variables that depend on topographic variations. So, in the exposition that follows, Landsat's prime role is to help to confirm the terrane-terrain association by visual recognition rather than extraction of terrain parameters.

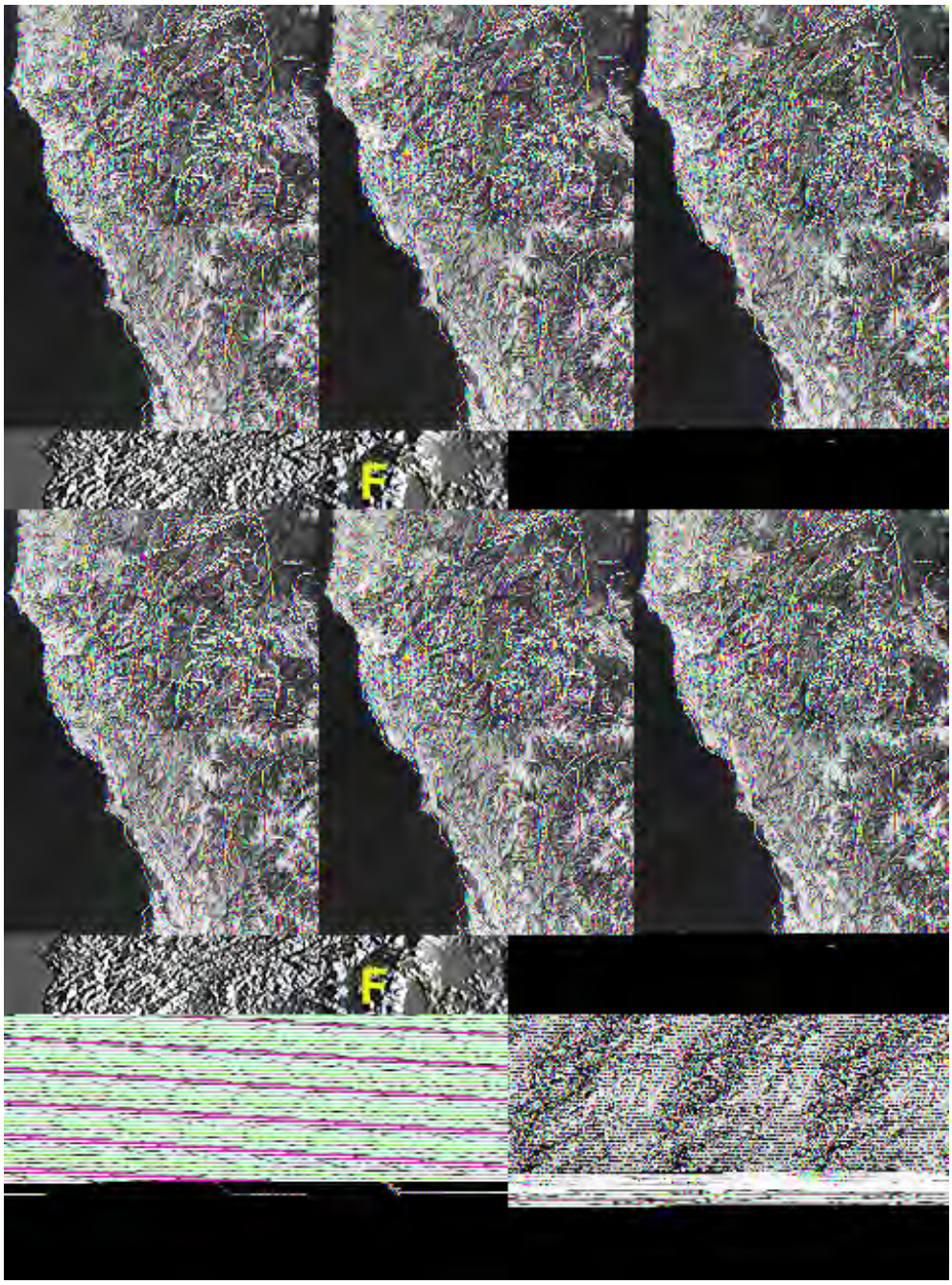
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

This page reviews some results of the analysis of the Klamath terranes. In general, most of the data used in the study had to come from already existing sources - principally published topographic maps. Landsat mosaicked images show the study area - close inspection suggests (but does not definitively emphasize) differences in the visual appearance of terranes in juxtaposition. One subscene near the coast, however, clearly displays strong differences, as supported by elevation maps that show one terrane higher than its neighbor. Directional trends of ridges also are different. This was supported by a SPOT stereopair, on which ridge lines could be mapped. Two inland terranes could also be separated visually. Maximum elevation differences also distinguish some of the terranes.

Terranes in TM Imagery; Ridges and Elevations.

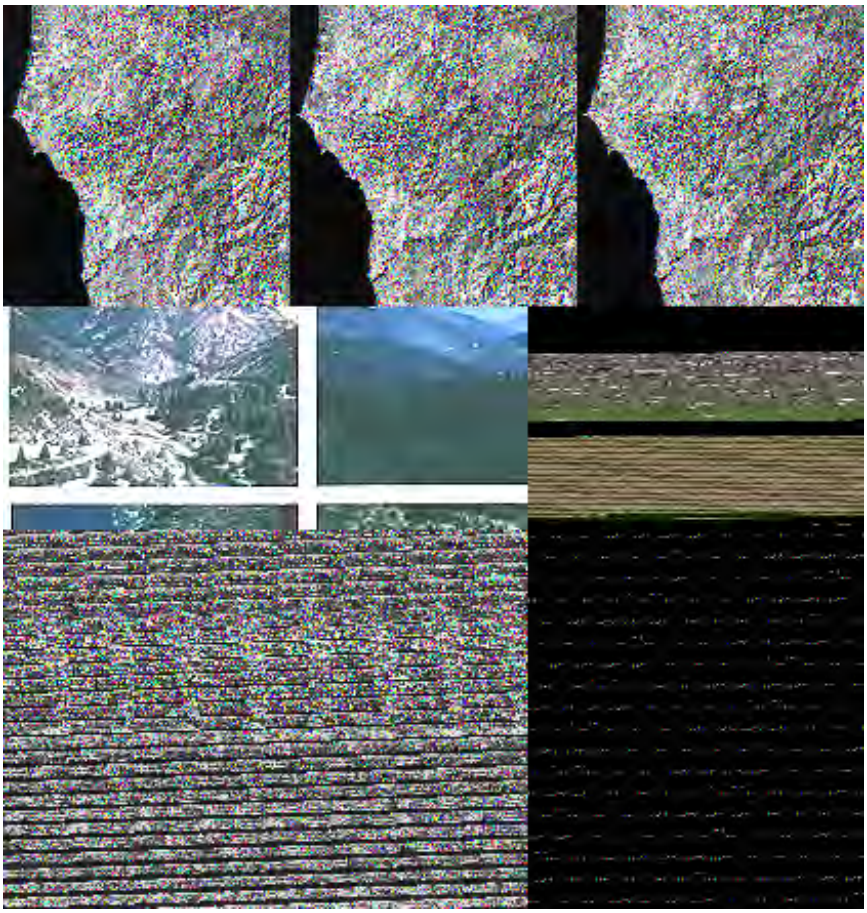
Without further ado, we now present a two-image (September 1984), Landsat Thematic Mapper Band 3 mosaic that includes the study area and mountains to its south, all the way to the California border. We overlaid the terrane boundaries, redrawn from a regional map, prepared by the U.S. Geological Survey, in white.



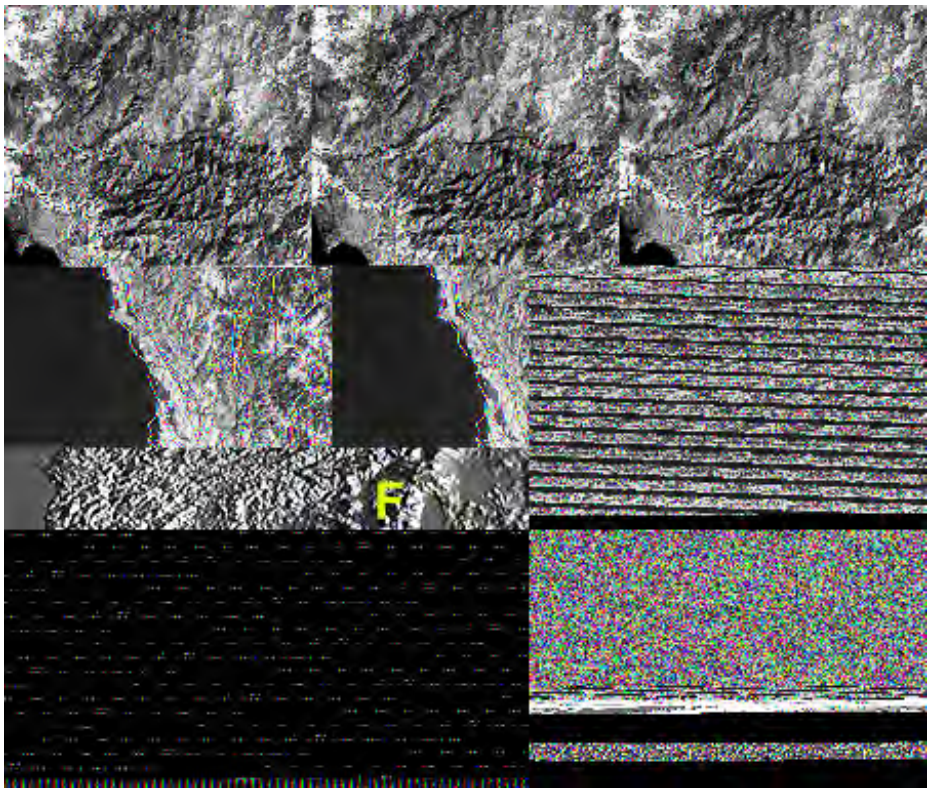
At first glance, and especially with monitor limitations in mind, you may decide that differences in topographic expression are hard to identify. It is true that this mode of display subdues the variability; looking at the actual mosaic laid across a table yields more conspicuous differences. But, to illustrate that these differences are really there - and can show up on your screen - for some terranes, look at the two in the upper left, listed as Sixes River and Elk. These terranes are clearly dissimilar, and, as we shall see in detail further down the page, they are the best case supporting the terrane-terrain hypothesis. Other terranes with apparent differences relative to their neighbors include: Rogue Valley, Smith River, and Yolla Bolly. Sixes River and Pickett Peak look too similar to identify any distinctions.

The Smith River and Yolla Bolly terranes also display differences in topography within themselves. Several areas within the Smith River terrane near its upper left (northwest) border have apparently high elevations and have more broadly spaced divides. These divides coincide with a lithologic unit known as the Josephine ophiolite, which consists largely of metamorphosed basalts or greenstones that represent parts of oceanic crust that failed to subduct. Consider also three arbitrarily selected segments of the Yolla Bolly terrane designated later in this survey as YbNorth (next to the Snow Camp upper label), YbCentral (next to the Dry Butte label), and YbSouth (next to the upper Smith River label). You should easily see differences in the terrains (e.g., look for ridge or valley spacings). We will quantify these differences later.

Let's take a closer look at the northwest section of the mosaic, in this enlarged image. Let your eyes adjust to this image and then try to draw imaginary boundaries around parts of it that seem separable. Once more, the Sixes River-Elk pair of terranes leap out as different.



Now, we will enlarge the image even further to focus on just the Sixes River and Elk terranes.

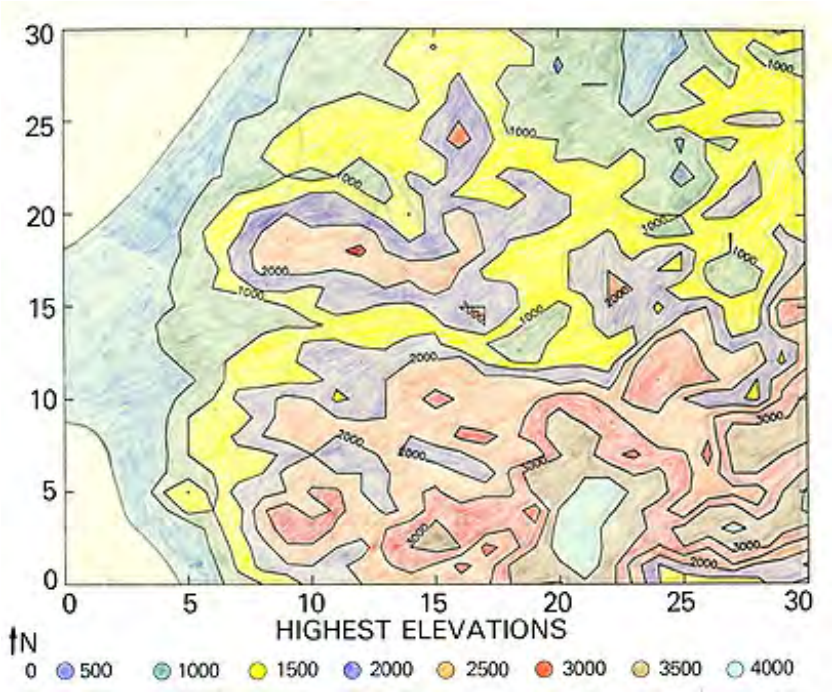


The upper and lower halves on the right and center of this enlargement are conspicuously different. They are indeed separable visually on the basis of topography.

17-18: What seems to be the most obvious difference in this visual depiction of the two terranes? ANSWER

Geomorphic Parameters from Maps

What is responsible for the obvious differences? First, when we use DEM elevations (in feet) for both terranes to compute the maximum height value in each square mile and then fit a surface to these maxima, this plot results:



The boundary between the Sixes River (north) and Elk terranes runs horizontally across this map at approximately the middle. Clearly, the Elk terrane is on average at a notably higher elevation than the Sixes River terrane.

17-19: **Which terrane is higher? ANSWER**

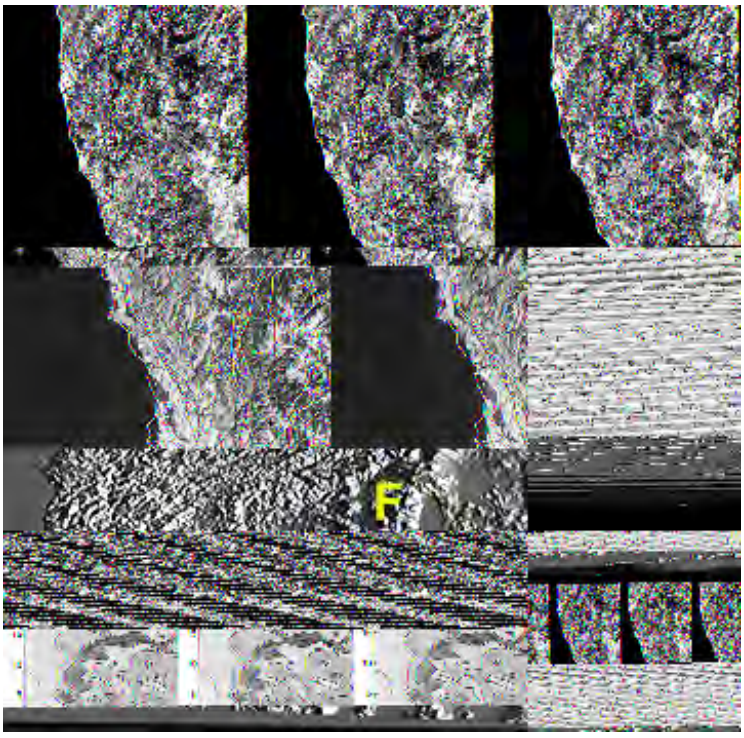
Then, using a map (not shown) that plots the mountain ridges between valleys, initially drawn from contours on individual 15 ft topographic quadrangles and then merged into a single regional map, I calculated the total length of ridges per unit area (each a square mile). This is an indirect measure of ridge density. The average of two subunits in the Elk terrane is 5.16 (miles/square mile), whereas that for the Sixes River is 3.29. This suggests that the ridges are more closely spaced in the Elk terrane than in the Sixes River terrane, a conclusion borne out by simply looking at the enlarged Landsat images.

A third difference is not readily apparent from the images. But when the regional map of ridges are analyzed in terms of average direction of trend or orientation of ridges in square mile sampling units (a special computer program did this), resulting rose diagrams for each terrane indicate strikingly dissimilar patterns:

Most ridges in the Elk terrane trend northwest in contrast to the northeast trend dominating the Sixes River terrane. I'll give reasons for this later (for now, accept the idea that both ridges and stream valleys reflect structural control).

These ridge measurements, when made from topographic maps, are relatively unbiased, i.e., with careful work, we should record all the ridges that are bisectors of crest-proximate contours. This eliminates the inherent problem in space imagery and aerial photography of missing some linear features because of Sun directional bias. Re-examining the Landsat imagery that we already saw indicates that the shadowing in this high-relief terrain, due to the morning Sun direction from the southeast, does emphasize the northeast trend of those ridges. But, when one looks at enlarged Landsat imagery, many of the ridges in other orientations can be detected by using tonal patterns related to moderated illumination differences along slope pairs that are not optimally oriented for easy discrimination. Still, inevitably some will be missed.

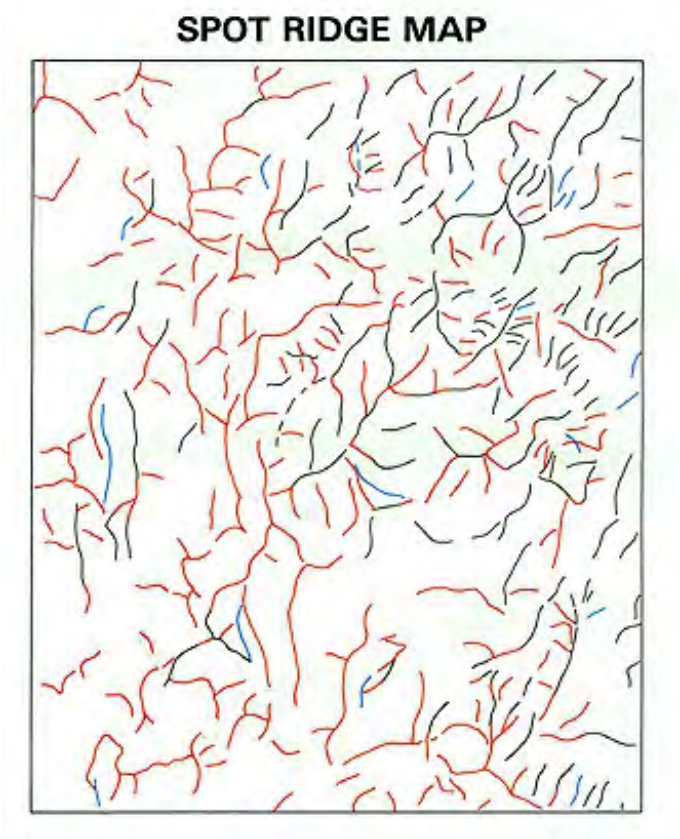
Stereo imagery pairs should reduce this bias effect. Two SPOT images made with the tiltable HRV scanner (see page 3-2), pointed at different angles and look directions on different dates, allowed me to map ridges directly from the imagery. Here is one of the (15 m [49 ft] resolution) panchromatic images (60 km [37 mi] on a side), covering the Gold Beach, Pickett Peak, and segments of the Sixes River and Yolla Bolly terranes:



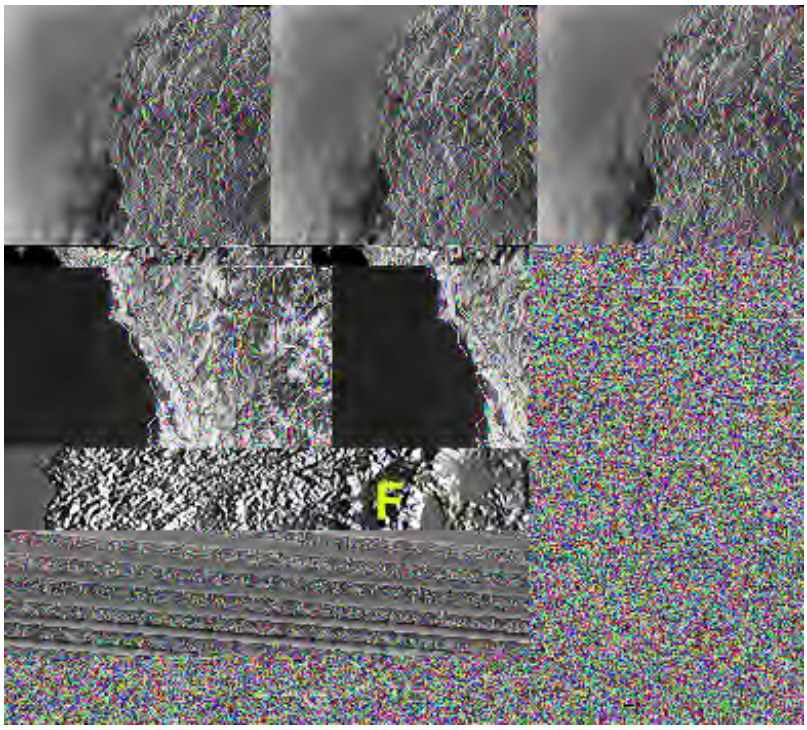
Although the higher resolution leads to more detail, the image appears flat, with the lack of strong shadowing inhibiting the expression of much of the topography. When viewed stereographically, the ridges associated with mountains of various heights do stand out well enough to foster tracing them onto a map.

17-20: The SPOT image contains a large number of light-toned, squarish "spots" (no pun intended). Any idea of what they might be? (Hint: think "forests"). ANSWER

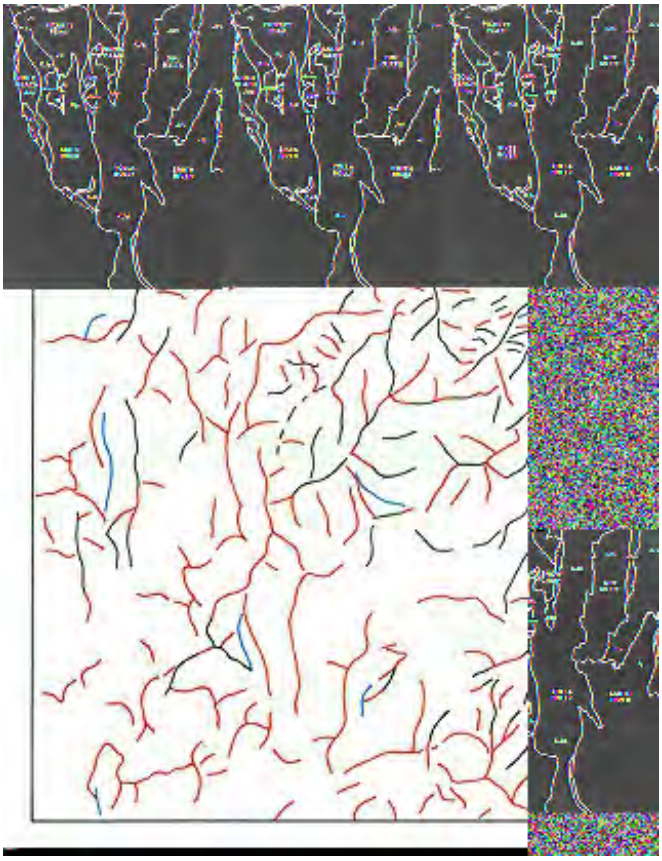
This next map shows all the ridges mapped in SPOT and Landsat images in red. Green denotes extensions of ridges selected from Landsat alone, and blue records images found only in SPOT. Surprisingly, the use of SPOT does not seem to improve ridge detection by much. Perhaps the Klamath terrain is well suited to depicting ridges effectively, even under the morning sun conditions.



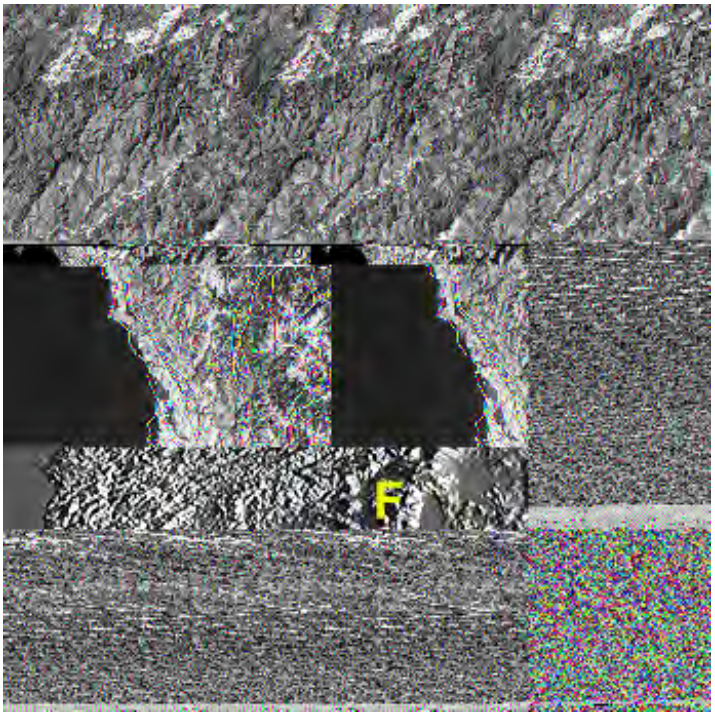
The same cannot be said for Seasat radar imagery. This digitally correlated Synthetic Aperture Radar image, taken in August of 1978, over part of the same areas as the SPOT image, indicates the sensor is looking east, and as such, there is a pronounced slant-range distortion that results in foreshortening layover. The image introduces a notable bias, that highlights the bulk of ridges oriented north-north-east and misses many others completely. A data product in this form is not as well-suited to orientation analysis.

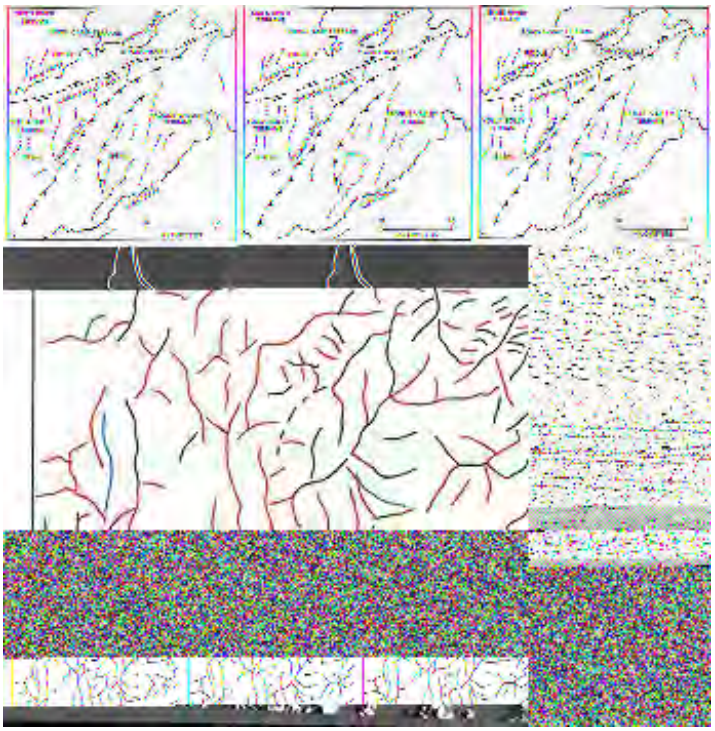


The land area in the SAR image is approximately fitted to this part of the terrane overlay map. Try to fit this overlay to that portion of the SAR scene; it proves hard to single out the several individual terranes present by visual inspection but the Yolla Bolly terrane seems to be expressed differently.



Look now at two terranes - Rogue Valley and Yolla Bolly - that appear similar at their northern end, as evident in this enlargement (with a locator sketch map) (we don't comment on the Sixes River and Snow Camp terranes north of the Canyonville fault):

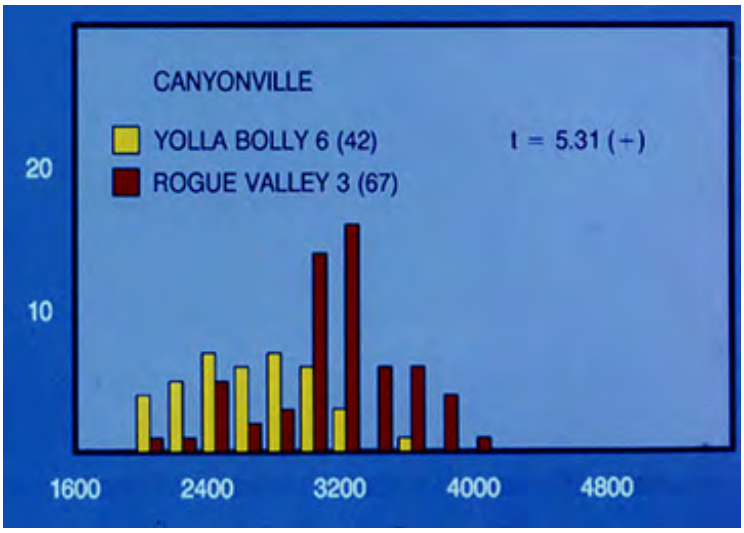




A right lateral strike-slip fault bounds both terranes (i.e., the northern block moves right, relative to slippage to the left in the lower block) near Canyonville. The terranes are juxtaposed along a large thrust fault, inclined downward to the east, resulting where the Yolla Bolly terrane slid under the Rogue Valley. At a quick glance, the ridge spacing and relief for the mountains in each terrane near this fault seem similar. The ridge orientations of the Rogue Valley terrane are more northward than those of Yolla Bolly.

17-21: In visually examining these two terranes in the above Landsat subscene, do you notice any difference(s) besides ridge orientation? ANSWER

One difference is not evident from visual inspection but comes out when one makes measurements from the Canyonville topographic sheet. Arbitrarily, I read the highest elevation in each township section (1 square mile) from that map. When plotted as a histogram (frequency of maximum elevations), the following distribution results:



Allowing for the difference in the number of data points involved (67 versus 42), there still is a notable variance between height distributions. In this topo sheet, the average maximum elevation in the Rogue Valley terrane is 984 m (3,226 feet), whereas in the Yolla Bolly terrane, it is 845 m (2,774 ft). Treating these two sets of measurements as two distinct populations, a statistical device known as Student's t-test gave a value of 5.31, which is significant (i.e., shows

a real difference) at the 99% confidence level. I surmised that this dual distinction results from differences in dominant lithologies. The Yolla Bolly is characterized by siliceous shales, and the Rogue Valley includes more resistant rock types.

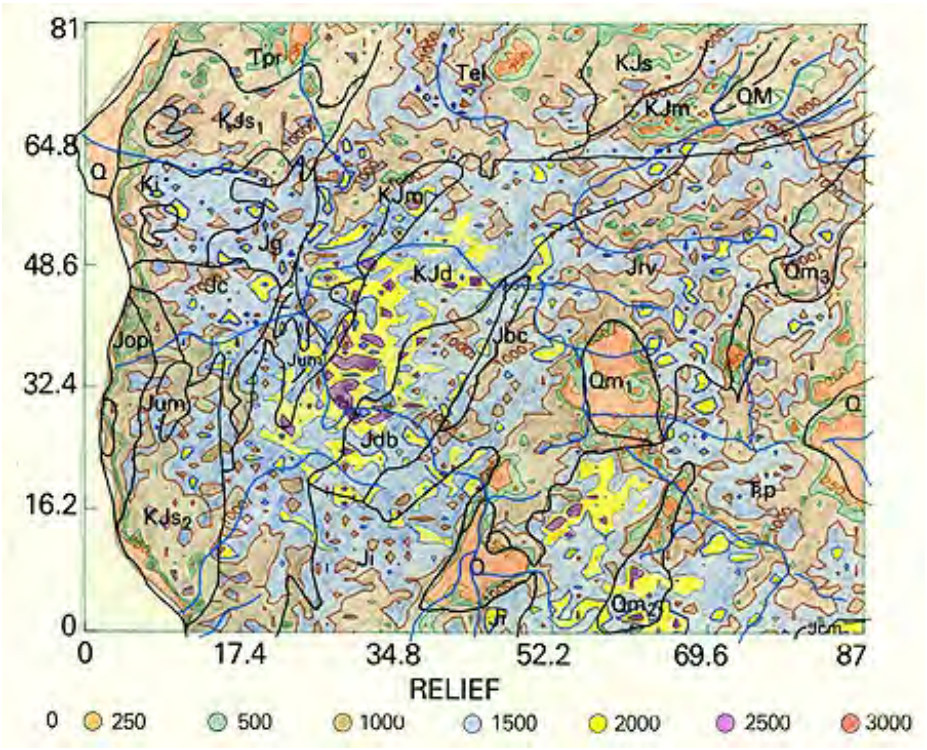
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

On this page, all measurements refer to those extracted from topographic maps. Relief and hypsometry (frequency distribution of elevations) data are analyzed. Relief differences were notable but did not help to differentiate terranes as terrains. But, hypsometric curves showed real and provocatively distinctive differences.

Relief Measurements

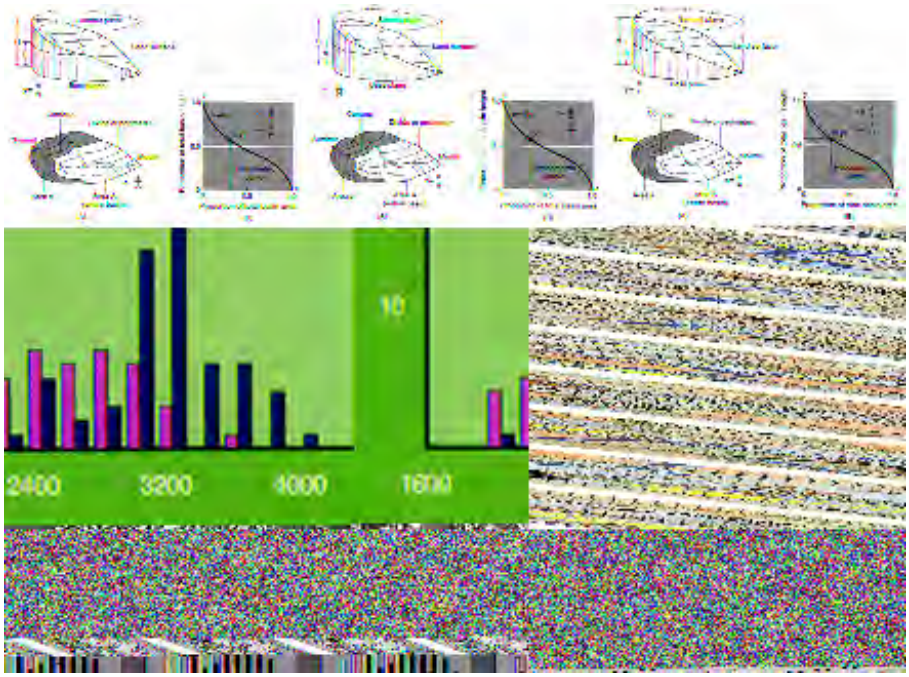
We turn now to a final consideration of some geomorphic properties that apply to the entire study area. We made a map showing relief (in feet) relative to terranes from digitized data in the Digital Elevation Map 1:250,000 sheet for the Klamath Mountains:

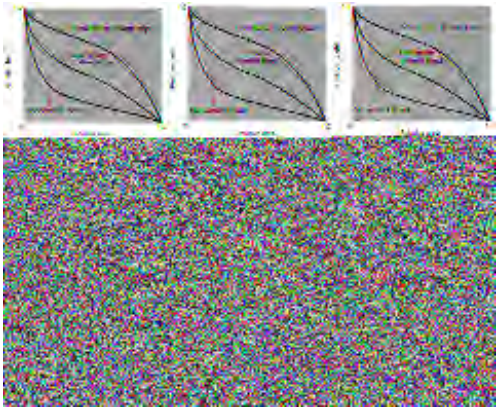


Maximum relief occurs in the central Yolla Bolly terrane. This is also the area (in purple) of highest maximum elevations (1,174 m [3,850 ft]). This height and the ease of downcutting into the shales account for the relief. We show the lowest inland areas in orange-red. These areas are valleys, one containing Grants Pass, the largest town in the Klamaths, that Quaternary deposits cover. The Grant Pass valley is underlain in part by a granitic rock, which weathers into a thick "rotten rock" (saprolite) soil and erodes easily to leave a lowlands.

At this point, with some trepidation, I introduce the concept of hypsometery. This rather complicated idea must be treated in a general way because, when it was applied to the Klamath terranes, some interesting results ensued (but their significance remains unclear).

Hypsometry refers to finding the distribution of elevations as a function of area occupied by each contour interval within some geographic unit (such as a topographic map sheet). This process determines the fraction of a surface bounded by specified elevations. These diagrams (from Process Geomorphology, by D.F. Ritter) depict the concept:



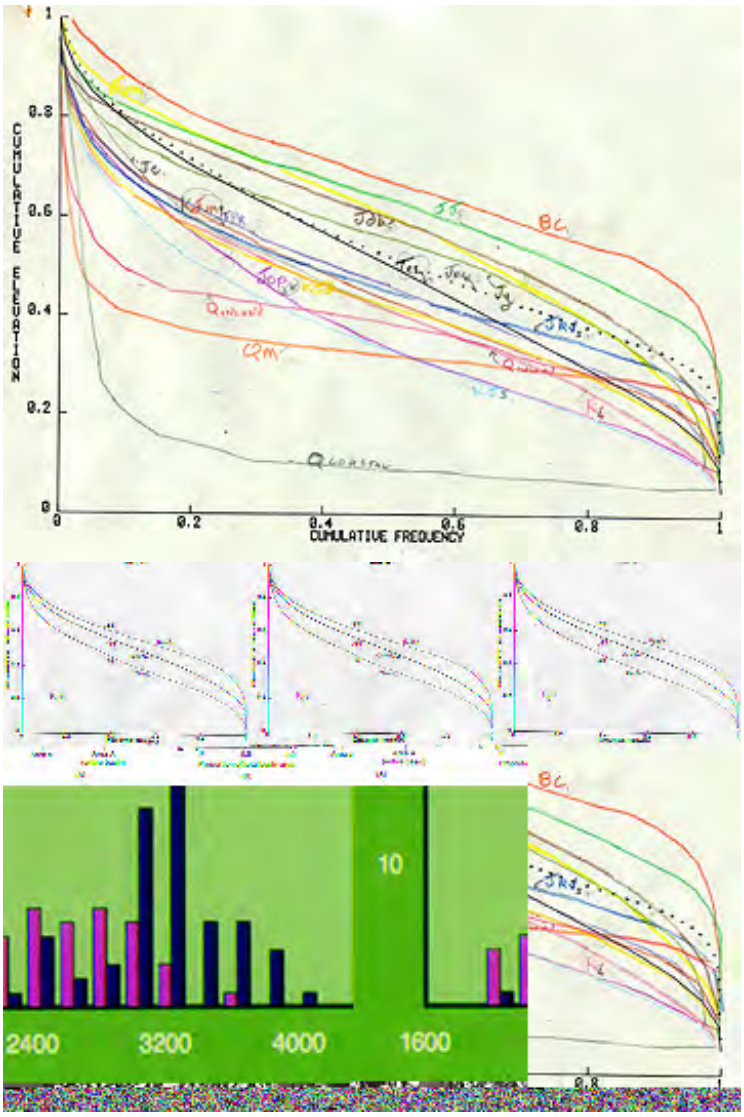


The prime hypsometeric parameter is defined as a volume bounded by the base and summit planes, of a given area A, and the total height (H) from the base to the highest elevation. As we move from the lowest elevation (base) upwards on the map progressively through increasing values of h, we measure all the areas within each contour interval. With each successive increase in h, the area increases by some new value of a. This progression is cumulative. When plotted on an x-y diagram, the relative values of h/H and a/A form the hypsometric curve (center right). Note that x and y values are dimensionless, representing proportions of the total area and height. Thus for y = 0, all heights are above the datum plane and therefore, lie within the total area, i.e., x = 1. We can calculate the area below the curve as the hypsometric interval (HI). As applied to a drainage basin, HI represents the unconsumed volume as a percentage within a geometric figure delimited by base plane, summit plane, and perimeter areas.

Various factors influence the shape of the hypsometric curve and its position in the x-y plot - hence its HI value. Direct interpretation of a family of curves, and their possible interrelationships, is not straightforward, but we can make some generalizations. Overall, the HI broadly denotes the degree of uplands dissection. Look at the bottom diagram above. Its top curve can typify a relatively high terrain (keep in mind that the geomorphic unit, the drainage basin, consists of valleys and intervening divides and all slopes in-between), which may be in the early stages of downward erosion and slope wall development. The lower curve is characteristic of the late stage of erosion, marked by broad, low valleys, with occasional higher promontories ("monadnocks"). Of course, various combinations of surfaces can lead to similar curves in proximate positions.

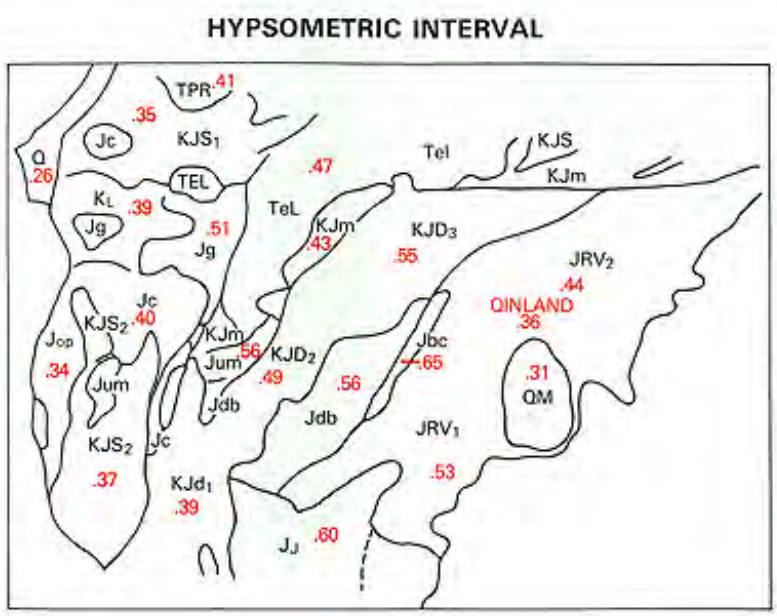
Again, using digitized topographic data, I produced hypsometric curves for individual Klamath terranes and, for several segments thereof. I show hand plots of these segments in the top

diagram, and plots for the three aforementioned Yolla Bolly segments are on the bottom.



I do not expect you to grasp in detail the relations among these hand-drawn curves (it is hard to identify some of the terrane symbols), because I have not derived specific interpretations that have obvious meaning. Suffice to comment that the one on top associates with the sliver of Briggs Creek terrane, which is one of the highest. The Jj (Smith River) and J_{db} (Dry Butte) terranes also have high elevations and rugged topography. But, the Yolla Bolly terrane as a whole (K_{Jd} in the top diagram) occupies a lower position in the set of plots. From the bottom diagram, it is evident that there is moderate variation within this terrane, although less than the range among all terranes in the top diagram. The three map units–none are actual terranes–with the lowest curves (and thus smallest HI's) are the lowlands associated with the Quaternary coastal plains and inland areas, and the granite lowlands.

When I plotted the values of HI (in fractions of 1.00) on the terranes map, a pattern emerged:



The highest values of HI cluster in the central terranes. This pattern suggests that these are also topographic units that have not been significantly downcut towards sea level. Those units near the coast, with lower elevation levels, even though younger (docked later), may have eroded more completely. Because their stream heights and gradients are closer to sea level, one would expect a smaller amount of relief.

My interpretation of these curves is simplistic: Those near the top of the group would describe terrains that are equivalent to what Davis called Youthful topography; those in the middle are progressively more mature; the lowest still suggest rugged terrain that has more subdued relief. The lowest curve is in Quaternary sediment-filled valleys that shows little relief (as would a peneplane). While this deduction is plausible, I cannot discount the (major?) role of another factor: the differences are due largely to relative resistance to erosion. But, I stop here without further speculation, because these hypsometric data, while tantalizing, are not equivocally interpretable.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

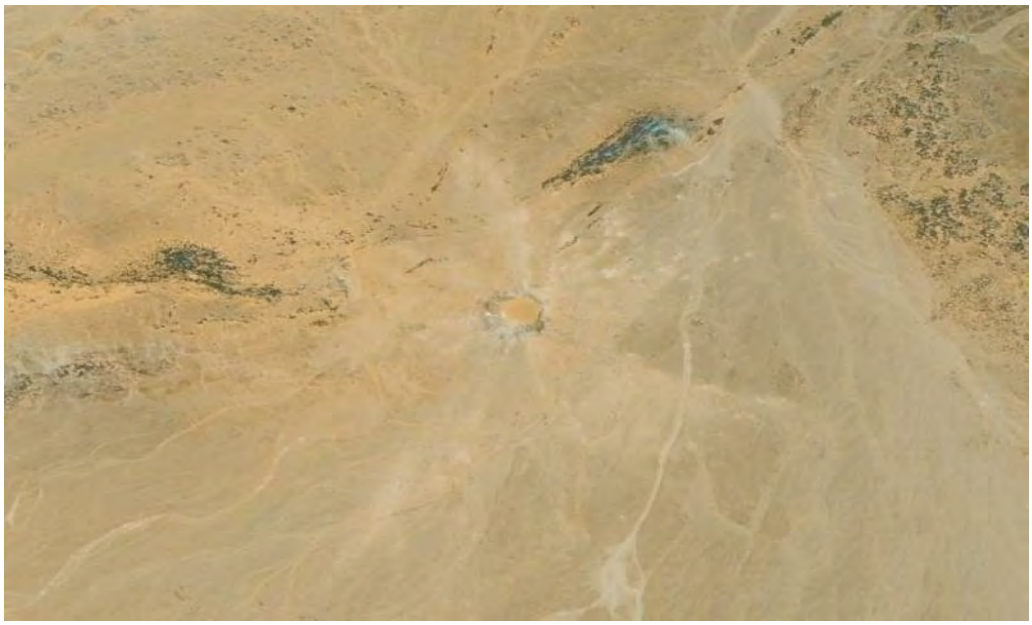
From the status 50 years ago of "minor curiosity", impact cratering has now been elevated to be one of the three most important geologic processes. Impact craters are common on all four inner Solar System planets - Mercury, Venus, Earth, and Mars. About 200 craters are known on Earth (compared with millions on the Moon that survive because tectonic and fluvial erosion processes have not destroyed them). This second section on scientific uses of remote sensing will delve into three topics: 1) what happens to make a crater; 2) how are impact craters found and recognized, and 3) can remote sensing help to find more? To date, several individuals have found previously unknown craters directly through space imagery. Special processing of images also can reveal subtle details about known craters that add to our knowledge of their size and deformation characteristics. Impacts on Earth are now especially important (to study) because they may be responsible for (or, at least, a major factor in) mass extinctions of life (both animal and plant).

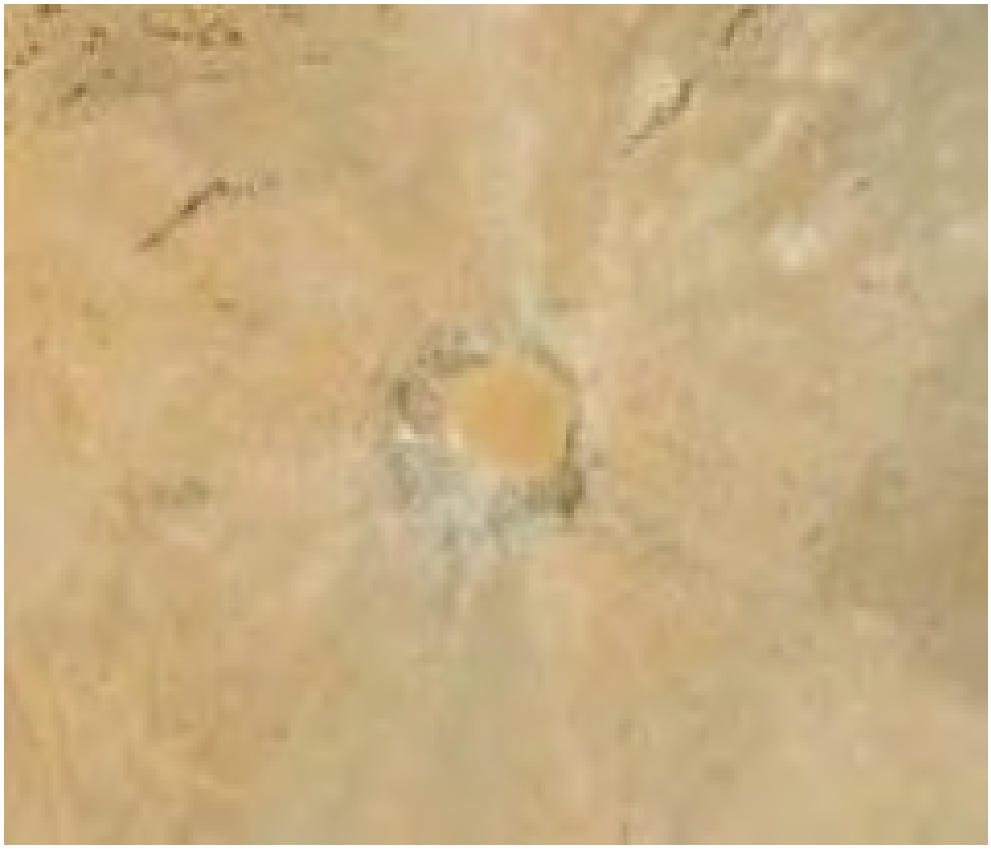
BASIC SCIENCE STUDIES II: IMPACT CRATERING

Basic Science II presents another geological subject that also was once the writer's (NMS) principal specialty - the study of nuclear explosion craters and their natural counterparts, meteorite (asteroid) impact craters. For years, impact cratering was esoteric. Now, impacts are hot topics, especially since they have been identified as the likely immediate cause of dinosaur extinction. Where in the 1960s the number of scientists working on impact craters were probably less than 20, now there are hundreds. More to the point, scientists and the general public now realize that impacts by extra-terrestrial bodies (asteroids, comets, meteorites) are genuine dangers to mankind, capable of causing catastrophes greater than any other natural process known to affect the Earth's surface. Plus, they are almost certainly to occur (as they already have) as huge events sometime in the future of civilization. This subject of impacts is pertinent and interesting, and most impact craters are detectable by remote sensing (but newly discovered ones require ground truth and examination of their rocks to prove they are of impact origin). -

Remote sensing has played an important role in the study of impact craters. Since Landsat, and now a bevy of new, higher resolution earth-observing satellites, the entire land surface of Earth has come under surveillance. Previously known impact craters are being re-examined in the context of their surroundings. But, the scientific "fun" has been to search for new ones, especially in isolated regions not easily accessed or populated. The basic strategy in the hunt is to look for distinct circular structures or features, the hallmark of craters, and then to find confirmatory evidence that impact was involved. New craters have been discovered in satellite imagery, as we shall see. Let us start this Section by showing a crater discovered entirely with such imagery, and first reported on in 2010.

This is the small (45 meters), fresh, young (perhaps less than 5000 years old) crater Kamil in the southern Egypt desert. It was first noted in high resolution GeoEye (Digital Globe) imagery but although still with distinctive morphology easily missing unless one were actually perusing such images specifically to find impact craters. The crater is imposed on sandstone, covered by loose sand, that has distinct ejecta blocks. The presence of iron meteorite fragments confirms its impact origin. Here are the "discovery" images and a ground photo showing the team of geologists who went into the wilds to analyze the crater.





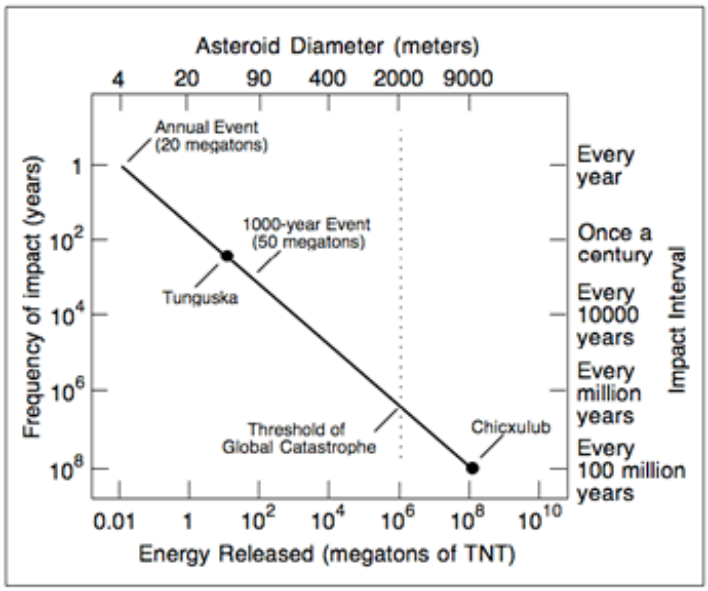
There are several reason why both scientists and laypersons have come to know about impacts and their consequences. One is the realization that the Moon and all the Solar System's rocky planets and many of their satellites have circular structures, the majority of which are impact-produced. Impact seems responsible for events in the Earth's past that made significant changes in the evolutionary history of animals and plants. The catastrophic damage done by impacts has become a popular topic for TV and Hollywood movies. Perhaps most compelling is the knowledge that at least one major extraterrestrial object has caused widespread damage in the lifetime of people still living on Earth. The message is: impacts still happen and may do so again..

This was the Tunguska event of 1908 that took place over east-central Siberia. This was manifested by an explosion (equivalent to about 15 megatons of TNT) in the lower atmosphere of an incoming comet or meteorite. This occurred in a then uninhabited part of Siberia. Here is its location:

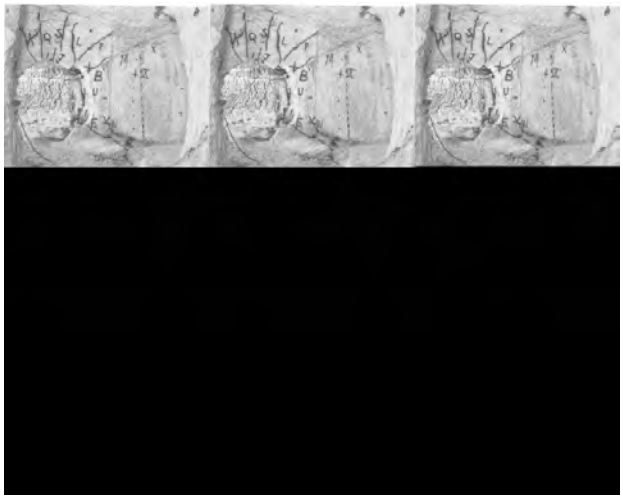
The explosion was heard by many and barometric gauges over much of the world registered overpressures from compressivive atmospheric waves. Years later explorers reaching the region below the blast found millions of downed trees:

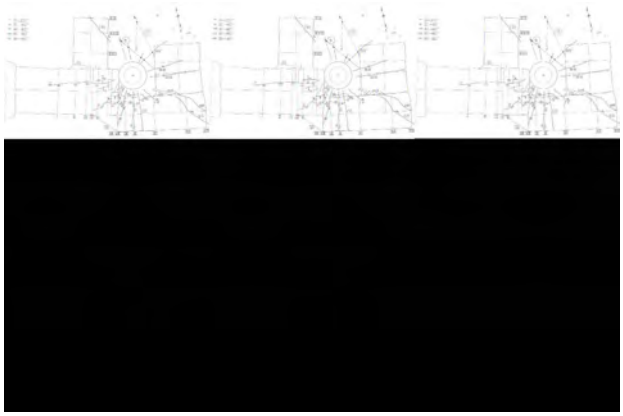
More information about Tunguska is found at this Wikipedia web site.

If the ferocity of the Tunguska event didn't get your attention, the graph below should. It shows the energy release caused by asteroidal impacts, expressed in terms of megatons of Hydrogen bomb equivalents, and the likelihood in terms of average number of years between cratering events of different magnitudes (frequency of recurrence):

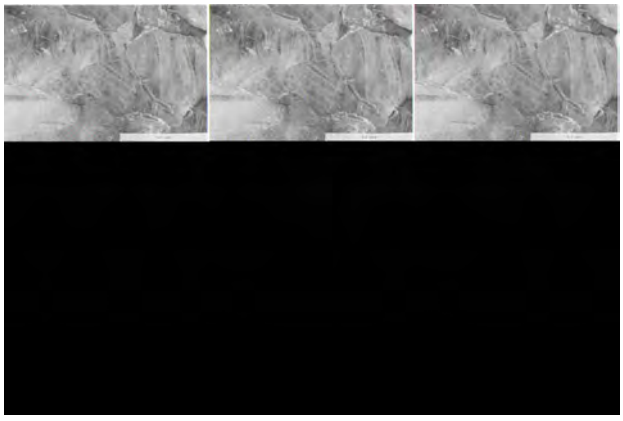


The writer's interest in impact craters began in 1959 when I joined the Lawrence Radiation Laboratory in Livermore, CA where my job was to support the Nuclear Weapons Division inasmuch as the testing of nuclear devices went underground, so that geological factors now had a major role in controlling and interpreting phenomena associated with the explosions. In April of 1960 I was assigned a most peculiar job: to supervise mining around a 1000 pound detonation within a salt dome outside of Winfield, LA so as to assist the physicists who had calculated the effects of this explosion as a simulation of a contained nuclear detonation. The model developed for the computer run had predicted fractures that would radiate uniformly from the cavity that held the Pelletol (dynamite in pellets) charge. I found this model to be somewhat wrong - the fractures were confined to a narrow sector of about +/- 60 degrees because of the influence of gypsum bands within the salt. These are two illustrations from the first paper I ever had published, about this follow-up study of the Project Cowboy event:

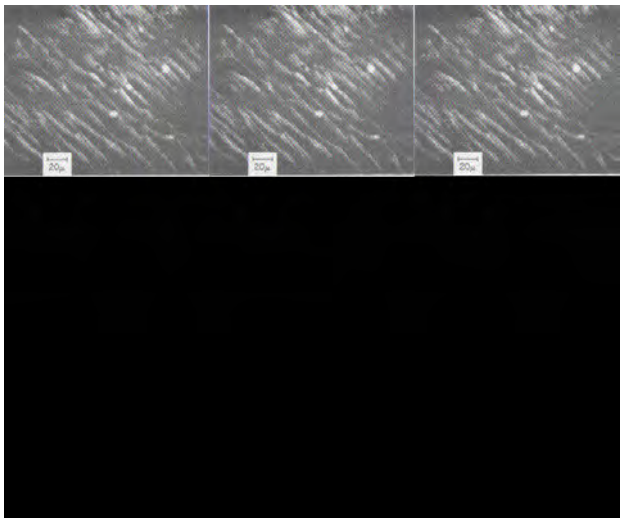




The salt immediately beyond the cavity walls showed numerous small fractures which actually were slip planes that allowed the halite crystals to deform by a quasi-plastic flow. Under a binocular microscope I could see that tiny spheres which were bubbles containing carbon dioxide gas had all been deformed into (strain) ellipsoids whose long axes were all pointed in the same direction, i.e., were aligned. When this next photo was taken, the gas had entirely leaked out but the cloudy appearance of the salt results from the close-spaced slip planes:



I was particularly intrigued by one sample. The salt had been so severely shocked that the halite crystals were partially isotropized but retained residual strain bands:



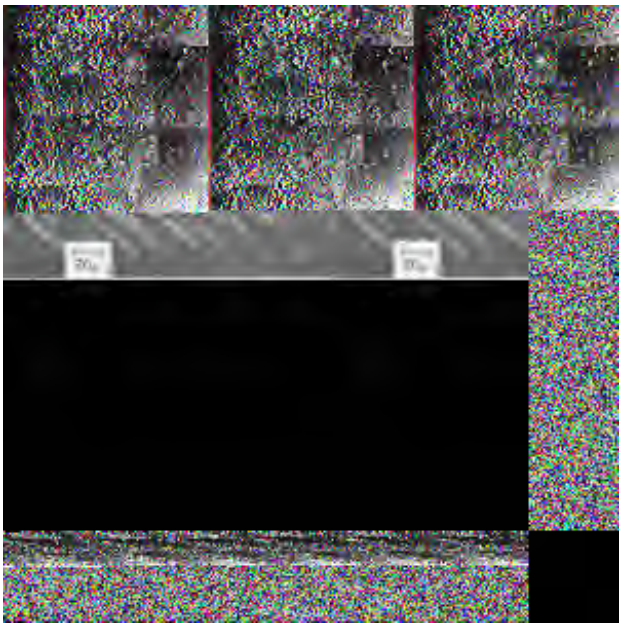
It was this project and in particular the effects seen in the salt crystals that started my interest in shock effects in rocks that grew into more studies of nuclear and chemical explosions on the rocks themselves and eventually into my major scientific endeavors concerning the rocks involved with meteorite and asteroidal impacts.

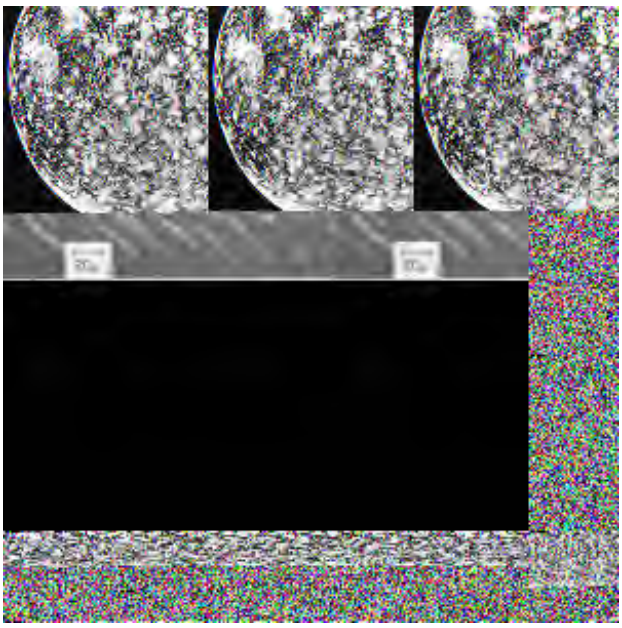
(An exceptional summary of impact cratering, and especially the record of these events imposed on the target rocks, is a 120 page treatise written by a colleague, Dr. Bevan M. French, and entitled *Traces of Catastrophe: A Handbook of Shock-Metamorphic Effects in Terrestrial Meteorite Impact Structures*, published by the Lunar and Planetary Institute [Houston, TX]; LPI Contribution No. 954, 1998.)

The Nature of Craters

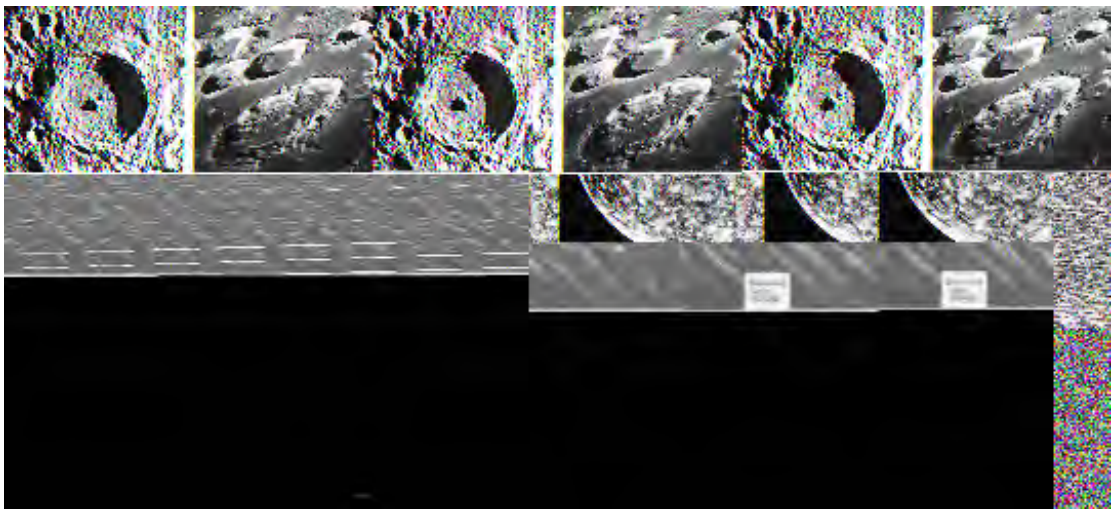
After more than 50 years of impact cratering investigations, we may now assuredly say that IMPACT is one of the three fundamental processes involved with planetary and asteroidal bodies (the others being VOLCANISM and CHEMICAL/PETROLOGIC DIFFERENTIATION; on Earth, TECTONICS is a co-equal fourth). Excepting probably Jupiter's Io, whose young surface seems devoid of impact (but not volcanic) craters, all other bodies have varying numbers of craters (Saturn's Titan remains uncertain, and Pluto still hasn't been imaged at resolutions that would detect craters), from the 200+ on Earth to millions on the Moon, Mercury, and most of the satellites of the Giant planet.

Thus, one of the indirect payoffs of the space program is that much of the general public now realizes that fast-moving asteroids, comets, and large meteorites collide with growing or stable planets during their formative period. (Asteroid catastrophe movies are fashionable.) Moreover, impacts persist in modifying such bodies throughout their history. The surfaces of some terrestrial-type planets and many of their satellites display craters that totally pockmark their surfaces: Mercury (top) and Callisto (fourth satellite out from Jupiter) are prime examples.





Earth's Moon has conspicuous craters with generally circular depressions, ranging in size from less than an inch to more than 1,200 km (746 mi) in diameter. Many of the larger ones are visible at full Moon through a pair of binoculars. The most conspicuous lunar impact structure is Tycho (below, on the left) located on the Moon's southern hemisphere. We can readily see that it is the source of great streaks (rays) almost visible to the naked eye that result from deposits of ejecta hurled across the Moon.



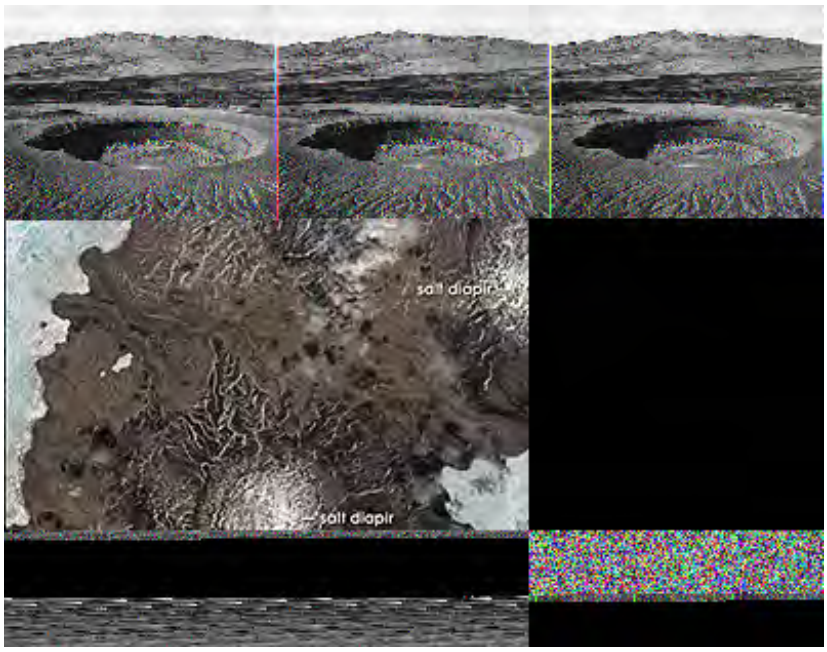
The Tycho crater is the classic exemplar of a large impact structure, with these hallmarks: circular raised rim; concentric nest of slumped walls inside this rim; central (uplift) peak; rough, irregular crater floor (here a mix of fragmental ejecta and lava extrusion); and exterior ejecta in hummocky deposits. Typical of farside craters is Goclenius (55 km, 34 mi wide) and several smaller ones (right) as they appeared to Apollo 8 astronauts circling our lunar neighbor. Note their flat interiors filled with mare lavas.

On Earth, circular structures are rare but, as mapping parts of the continents since the early 19th Century led to discoveries of such features, a number had been identified by the middle of the 20th Century. One of particular note is the series of concentric ridges in Mauritania (West Africa) known as the Richat structure (about 40 km outer diameter). The structure was first discovered *from space* by Gemini 4 astronauts McDivitt and White. This structure results from doming of sedimentary rocks by an igneous intrusion; erosion has left the quartzite ridges as hogbacks (inclined strata that are raised relative to softer rocks between circles) that outline the "bulls-eye" appearance of the structure. Here is a view taken by the ASTER sensor of the Richat structure which briefly was thought to be of impact origin until a visit by R.S. Dietz found field evidence that precluded that possibility and confirmed a purely internal origin. Below it is a perspective oblique view made from an SRTM-Landsat composite to show this remarkably circular basin as it fits next to a mountain range

Other types of circular depressions, craterlike, can be produced by normal geologic processes. A sinkhole is one example. Another depression is caused by the solution of NaCl in domes that reach the surface after diapiric piercing of sedimentary rocks by plastically-flowing salt. This astronaut photo shows two such domes invading the rocks in Melville Island, near Baffin Island, in the Canadian Arctic:

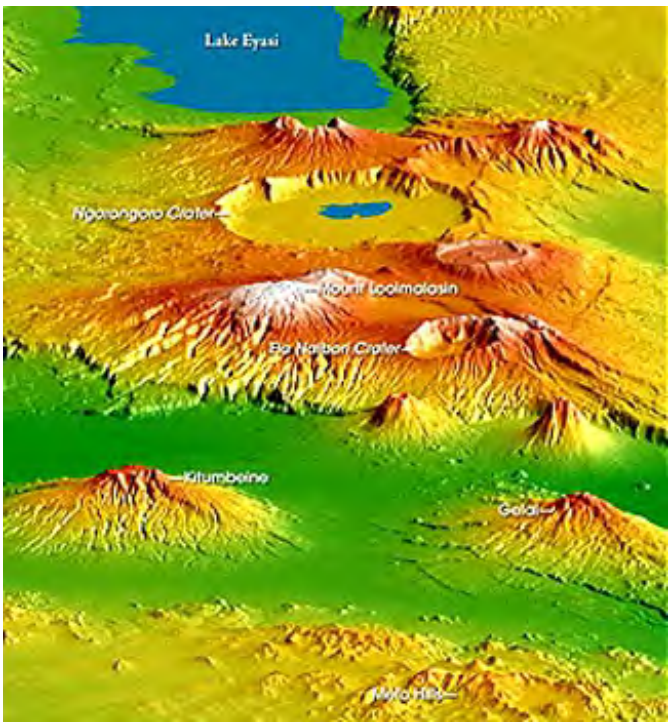
Less than 100 years ago, scientists considered the concept of craters (especially those that remain circular) formed by impacts from meteorites and other extraterrestrial bodies ("bolides"), such as comets, as unrealistic and highly improbable. Several scientists had, by then, suggested that such craters covered the Moon, but the bulk opinion attributed these to volcanic processes. And, indeed, there were depressions in terrestrial volcanic fields that bore sometimes strong resemblance to lunar craters.

One volcanic type that leads to near circularity of some of the resulting craters is the maar crater exemplified below by the Crater Elegante (1.5 km [0.9 mile] diameter) in the Pinacate basaltic field in northwest Mexico (below the Arizona border). This kind of volcanic crater often has a subdued rim. It is formed when lava encounters near surface water which flashes into steam causing rock overhead to be pushed out explosively leaving a depression that is somewhat backfilled by volcanic fragments (and later debris washed in). The gashlike depressions outside the rim are caused by water erosion.

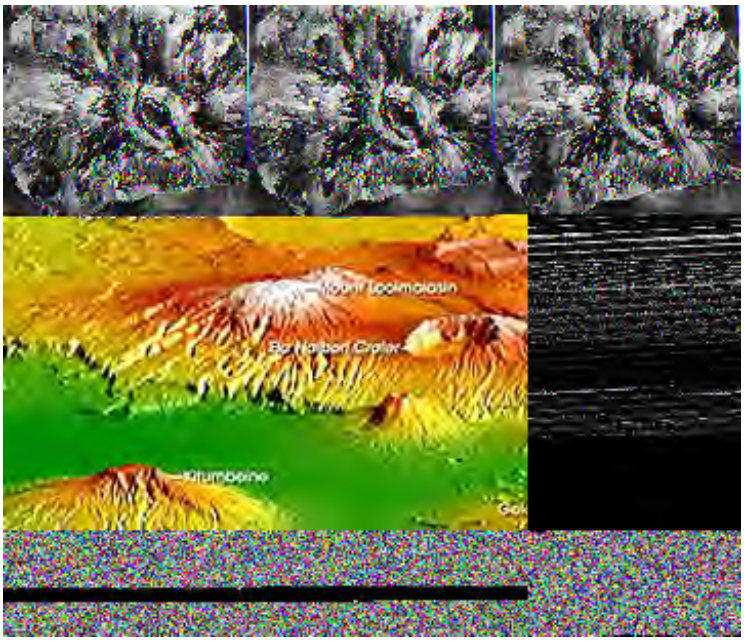


This next example helps to show how impact craters further differ from volcanic craters. The large volcano Nyiragongo in Tanzania, part of the East African Rift system chain of volcanoes, has a large central crater that is nearly perfectly circular. This volcanic crater type differs from impact craters by 1) being atop a built up mountainlike edifice; 2) having steep interior walls

(although some slumping can occur); 3) having its layered units, exposed in the walls, largely at low angles (subhorizontal) rather than overturned as is the usual case with impact craters; 4) having a rim that is only slightly raised; and 5) not showing distinctive ejecta (fragments with a large size range up to massive blocks) beyond the rim (some may exist if the volcano has explosive phases, but multiple, piled-up lava flows can extend across the slopes). A sixth criterion is usually a diagnostic separator: impact craters occur mostly in non-volcanic rocks (igneous; sedimentary; metamorphic) that are exposed in the rims, infill, and ejecta blankets (but a form of induced volcanism commonly results from the impact, leading to volcaniclike rocks in the crater floor).



Another example is the elongate caldera on the summit of the raised shieldlike basaltic volcano on Isle Fernandina in the Galapagos Islands. Note the irregularities in its shape and the numerous lava flows emanating from the side of the volcano's slopes.



18-1: Until the space program in the 1960s, debate over the nature of lunar craters raged as a controversy for more than a century. But with close exploration of the Moon and then planets like Mercury and satellites like Callisto disclosed that most planetary surfaces were heavily cratered. The weight of opinion shifted drastically in favor of

impact as the dominant process creating the myriads of circular depressions spread widely over these surfaces. Can you develop (deduce) some arguments that support this impact hypothesis? ANSWER

Impact as a lunar and terrestrial process was first suggested by European geoscientists in the early 1900s. The proposal that Meteor Crater in Arizona had an impact origin, opened the possibility of impact as the cause of other, similar, circular features. G.K. Gilbert conceived the idea and then supported it by finding iron meteorites around the crater. This "first among equals" crater, about which we will say much more on page 18-5, is shown below, along with a typical meteorite (the Canyon Diablo find) from the ejecta deposits outside the crater rim:

Work by R. Baldwin and others kept the impact process alive as an alternative but the vast majority of astronomers and geologists were highly skeptical. All this changed in 1960 with the classic study of cratering mechanics at Meteor Crater and several nuclear explosion craters by Eugene M. Shoemaker, which opened up the possibility of impact as the cause of similar circular features, of which more than 50 were then known on Earth.

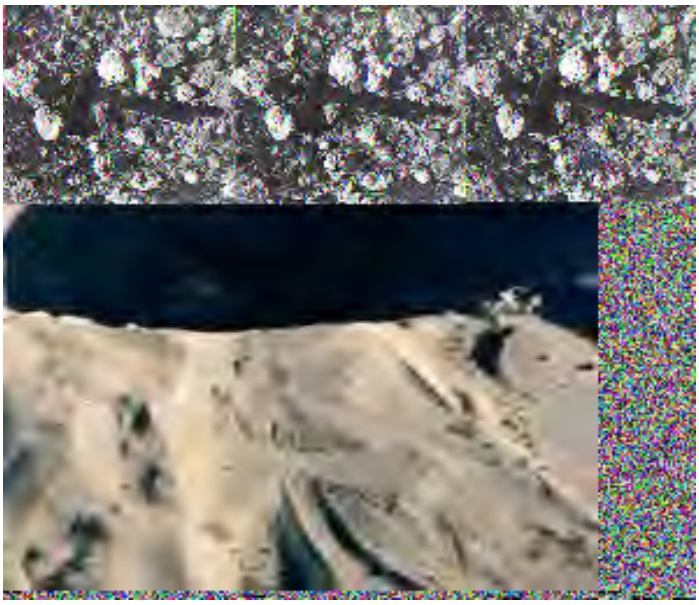
The critical proof that the impact process has occurred on Earth emerged in the 1960s from the study of the effects of the intense pressure (shock) waves generated during impact on the target rocks at terrestrial craters, with key studies by E.M. Shoemaker, E. Chao, W. von Engelhardt, B.M. French and others. These effects are described under the general term "shock metamorphism" which we will treat in detail on page 18-4 and 18-4a. Suffice for this introduction just to list the main effects: Shatter Cones; Breccias (rocks made up of fragments [clasts] held together by a matrix of powder, smaller fragments, and, in some, quenched melt [glass]); Petrographic features in individual minerals such as Planar Deformation Features (PDFs), Deformation Lamellae, Kink banding, and Thetomorphs; Melt (either in discrete layers [that can be many meters thick] or dispersed with other ejecta components; and Injected Veins (some containing Pseudotachylite).

In 1960, the writer (NMS), began my studies of nuclear and chemical explosion craters as part of my job in the Plowshare program (engineering uses of nuclear explosions) at the Lawrence

Livermore Laboratory (LRL) (see page 18-4 for further details). The first underground nuclear explosion with Plowshare ties was the Gnome 10 kiloton detonation in potash salt near Carlsbad, NM. (the [in]famous event that heaved up the desert surface, awakening wintering rattlesnakes that caused a mad scramble by nearby observers). The explosion produced a 70 meter (220 ft) cavity into which the roof partly collapsed; the photo below shows a geologist colleague, Don Rawson (at bottom), inside part of the remaining cavity:

In my work at LRL, I was responsible for determining the changes induced in rocks at the Rainier, Hardhat, Danny Boy, and Sedan nuclear explosion sites. Sedan, in particular, yielded quite interesting results because this cratering event was conducted in alluvium that contained granite, sandstone, and limestone "float" (blocks of rock a few inches to a foot in size, carried in by streams), all of which were subjected to strong shock pressures during the explosion. Sedan at that time was the largest cratering experiment yet conducted. A 100 kiloton nuclear device produced this 372 m (1230 ft) in diameter crater. Shown below are, on top, the ejecta being tossed out of the developing crater at about 10 seconds after the detonation (I was located about 6 km (3.8 miles) east of ground zero, on top of Rainier Mesa [the closest approach of anyone that day], busily snapping color photos that were later helpful in determining the downwind debris history). Below it is the crater itself about a week after the event, when I was allowed to move up to the rim to collect samples:

Here is a typical small patch of ground in the Sedan ejecta zone, showing mostly shocked rocks (mainly melted alluvium, shocked volcanic rocks, and quartzite).



These underground nuclear explosions generate peak shock pressures of the same magnitudes as are developed during impacts. One would therefore expect to find shock effects in the rocks involved that were similar to those noted at impact sites. And indeed I found that shock metamorphic features in these rocks were identical to corresponding features in impactites (rocks from natural craters). This tie-in led to the argument that only impacts could cause high pressures on the order of those in underground nuclear explosions, as the features produced are never found in rocks associated with volcanism, even the types called "explosive".

Gradually, the idea that impact is one of the fundamental formative processes acting on planets won broad acceptance. In fact, scientists have now proven that planets grow by accretion of infalling materials, with the craters representing the last stages of buildup, as the planets reach their full sizes. The nature of impact cratering is important, yet introductory Geology textbooks still treat the subject poorly.

Heavily cratered planetary bodies, such as those shown above, share the impact markings that appear on their ancient surfaces but that have not been fully demolished or masked by erosion, lava outflow, deposition, or obliterating mountain activity. Cratering on their surfaces was most intense during the last stages of planetary growth, very early in solar system time (beginning about 4.6 billion years ago), through a later period of about one billion years, after which, the flux of objects striking those surfaces dropped off notably. Earth, Venus, and parts of Mars, also profusely cratered at the outset, by contrast now show far fewer craters because of the subsequent destructive processes that erased or covered most of the impact evidence left on the primitive surfaces.

With the return of the first lunar samples from Apollo 11, shock effects in moon rocks were observed, supporting the conclusion that the bulk of lunar craters are the result of impact.

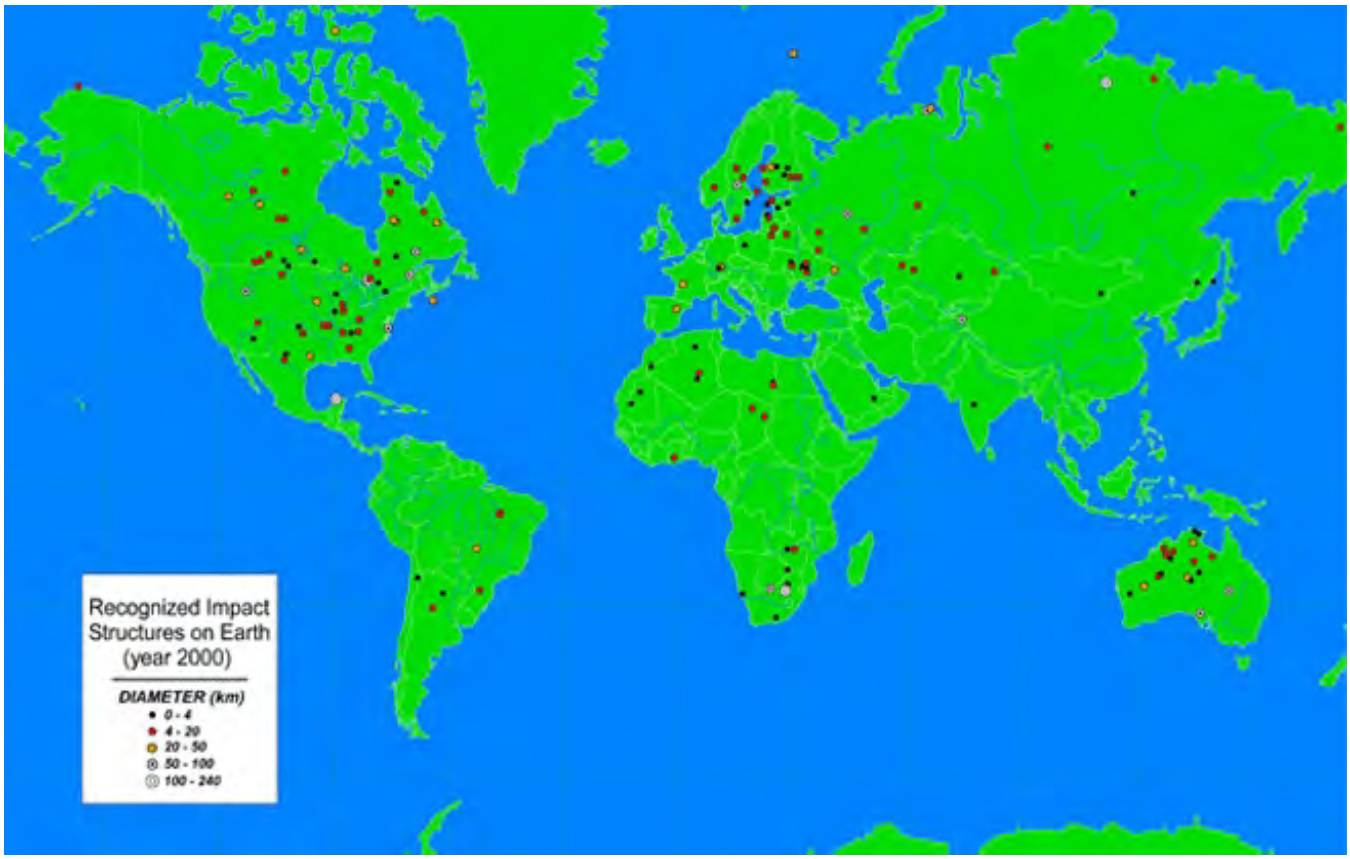
Distribution of Impact Craters on Earth

Earth today, despite its many recyclings of continental and oceanic crust, retains signs of huge impacts imposed in the last two billion years, as well as smaller ones that took place since then even up to historic times. Scientists have now found about 200 surviving craters of definite or probable impact origin. This is well below the estimates of tens of thousands that would be expected if Earth's surface and crust had not experienced such dynamic ruin from plate tectonic action and atmospheric-driven erosion. Also, protection by deep oceanic waters (more than 70% of the planet’s surface) and burial by sediments further account for this deficiency in anticipated numbers. Nevertheless, more craters remain to be discovered, and satellite imagery should be an effective means for conducting a systematic search, as we shall see near the end of this survey.

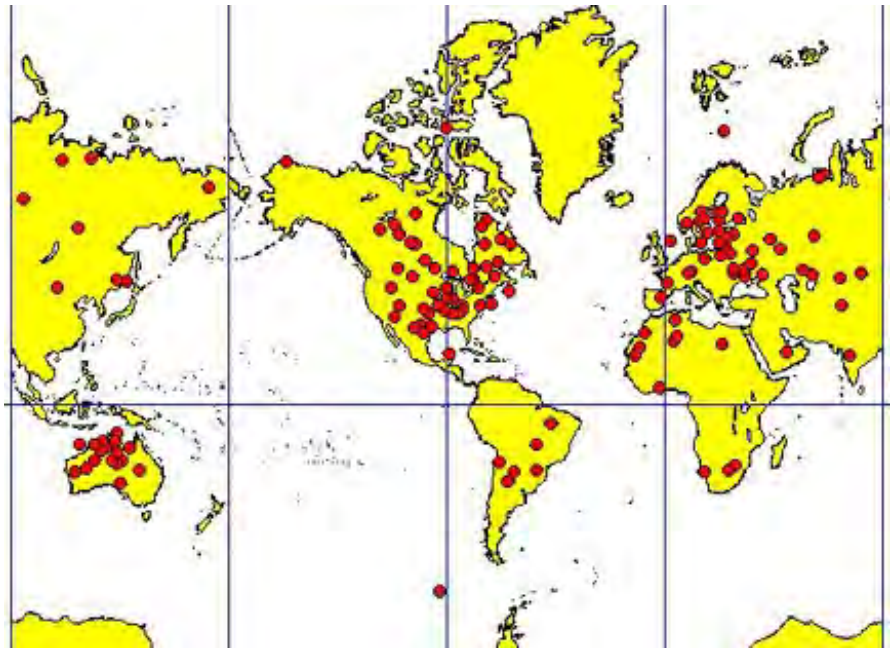
There are a variety of informative Web Sites on impact cratering; most provide a list of known

and probable craters. Calvin Hamilton has prepared a general treatment, supported by selected images of terrestrial and planetary craters, at (Solar System Cratering). Still another explanatory site is this review of impact cratering by Koberl and Sharpe; this article contains a nearly up-to-date geographic distribution map of known craters, plotted by size groups, which is reproduced below. A simple listing of craters, with location, size, and age, can be brought up from the Internet simply by pressing these two sites (active as of March 2005): (1); and (2). Another impact crater data base is maintained by the University of New Brunswick. Surfing the Internet under the topic "impact craters" has led to these interesting sites:ErnstonClaudin; impact structures; TheLivingMoon. The first two are produced by Kord Ernston and Ferran Claudin; The Living Moon site has several links (go to its "Menu") that are appropriate to both this and the next Sections, but the identity of the site's creators is not specified.

This is one of the better maps showing global impact crater distribution:



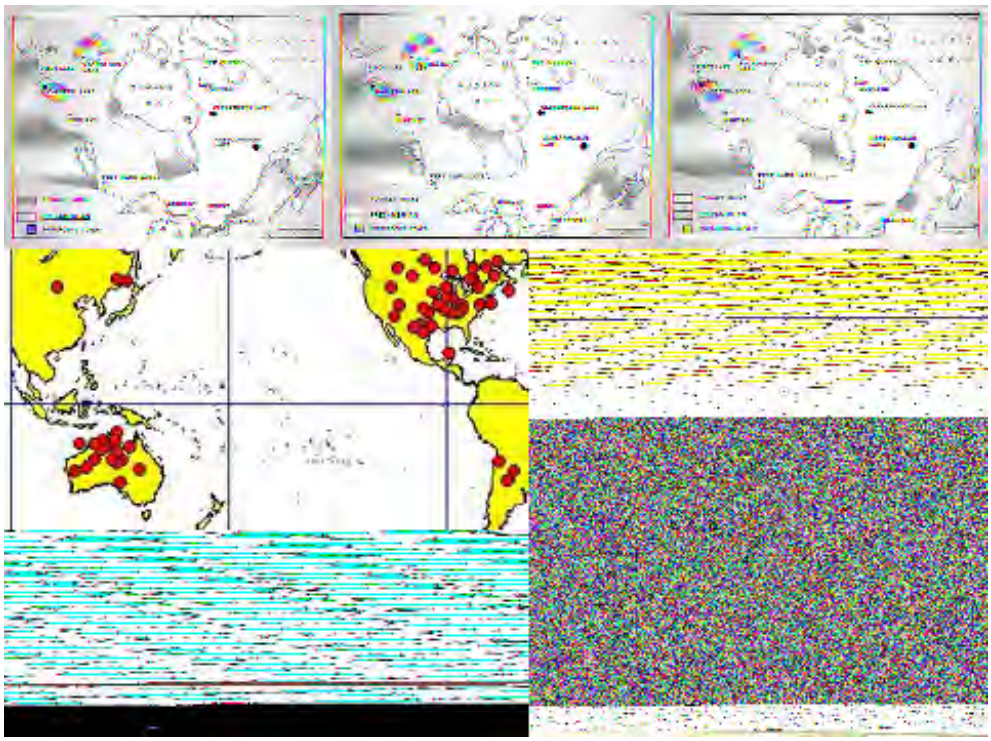
This next map has the advantage of really making the crater locations and distributions quite obvious by using red dots:



Even these maps are a bit out of date. Latest count of confirmed impact structures worldwide is 196.

18-2: Inspection of the above global distribution maps shows large areas on Earth in which confirmed impact craters are absent or sparse. Offer several explanations for this scarcity. ANSWER

We want to call special attention to the impact structures that occur in Canada by displaying this figure which shows locations and names of some of its craters and astroblemes (erosion scars of former craters). The Precambrian Shield of Canada is well suited to preserving craters as old as 1.7 billion years despite erosion by Pleistocene glaciation. It was the first large region of the Earth's continents to undergo organized search and study of impact structures under the direction of C.S. Beals of the Dominion Observatory and his successors, including Michael Dence. The West Hawk Lake structure was the one this writer (NMS) studied in detail through a NASA Research grant. Here is a well-known map, about 30 years out of date, that brought attention to the craters that had been found by 1975 through a systematic, deliberate search of the Canadian Shield:



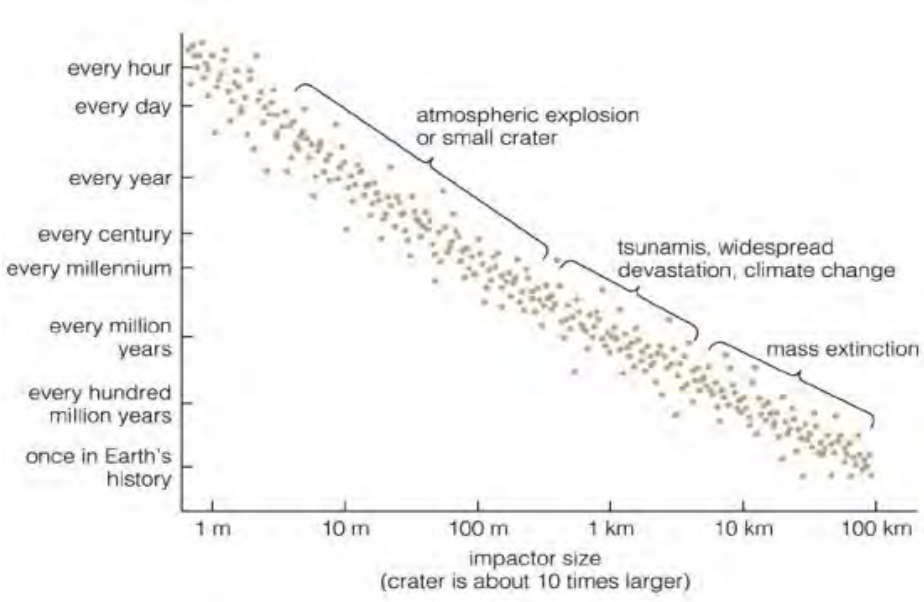
Since this map was drawn 15 more craters have been found in and around the Canadian Shield.

The Importance of Impact Cratering on Earth and Elsewhere

So, why is the study of impact craters, and the mechanisms of cratering (next page) important? Because over the last 40 years scientists have come to realize that this mode of crater formation is one of the three or four most fundamental natural processes affecting the Earth's (and planets' in general) surface. This will become obvious to you as the next Section - on the planets - is visited. The Moon, Mercury, Venus, and Mars, and the satellites of the Giant Planets, are all thoroughly peppered with impact craters. Earth has around 200 known impact craters but once had many more; active degradational geologic processes (stream erosion; tectonic deformation; burial by sediments, glaciation) that have erased most of the earlier ones.

In the last 100 years, several other meteorite impacts leading to small craters have occurred on the land surface around the globe; even more frequently incoming bodies have fallen into the oceans, potentially causing tsunamis. On February 12, 1947 an iron meteorite (possibly as large as a million kilograms) struck the ground in the Sikhote Alin mountains of eastern Siberia (near Vladivostok). Remnants of the meteorite are findable in a 1.3 km^2 strewn field which also contains small craters.

From studies of terrestrial impact craters in terms of size and age, from planetary surface studies, and from monitoring asteroid and comet distributions, a rather accurate estimate of cratering frequency (size versus occurrence in time) has led to plots like this:



This plot would seem to indicate that a 1 km crater would form somewhere every hundred years (had the meteorite above the Tunguska River hit the ground instead of bursting in the air, such a crater would have resulted). Every million years or so, a 10 km crater is a realistic possibility if land is the target. And in time spans of a few hundred million years, the crater so developed by impactors of tens of kilometers diameter (typical of asteroids) would put so much debris into the atmosphere that mass extinctions of many life forms would result. On page 18-4 we will see that this has apparently happened at least twice in the last half billion years.

So, this last conclusion should have grabbed your attention!! Lets proceed to learn more about impact cratering - a universal process - starting with a review of how such craters are generated.

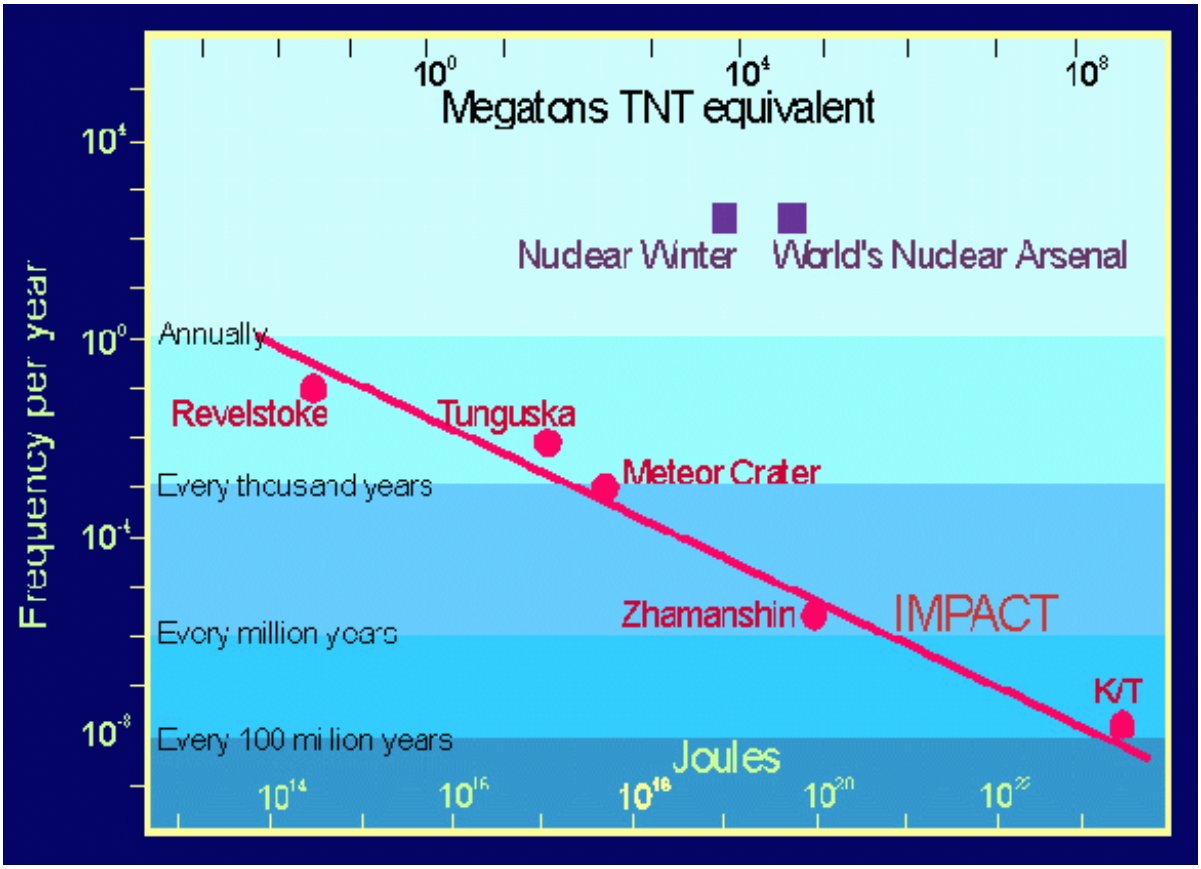
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

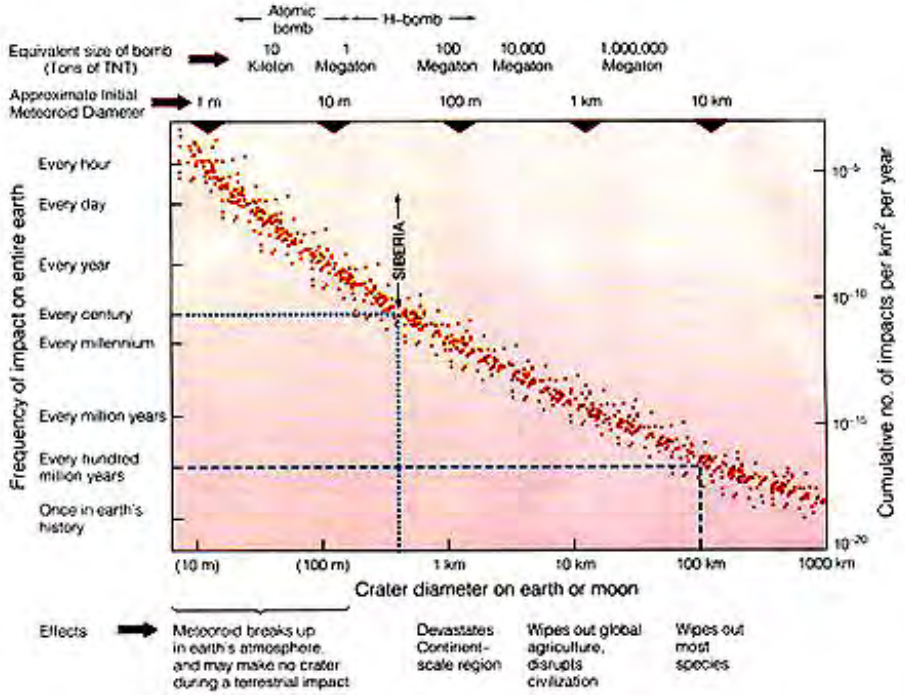
In many respects, impact craters form in a similar manner to explosion craters where explosives are buried at some depth and then ignited. But, the energy needed to form an impact crater begins to act at the point of contact at the surface. As it burrows into the target (rock or water), it has an effective (not real) center of energy release. An impact crater involves energies on the order of hundreds to millions of kilotons (equivalent nuclear detonation energy). A series of panel cartoons, and accompanying descriptions, shows the sequence of formation of an impact crater. The final result is shown as a cross-section.

Cratering Mechanics

No other natural event is as powerful, devastating, or potentially catastrophic as a major impact. Consider one capable of producing a 50 km (31 mi) wide crater, excavated to a depth of 5 km (3 mi): the energy expended is thousands of times greater than the simultaneous detonation at one point of all the nuclear explosive devices (euphemism for bombs) manufactured to date. We gain some idea of these magnitudes from this logarithmic (log-log) plot of the crater frequency as a function of energy at impact (or detonation) given in joules.



Below is a similar diagram with some additional information. The word "Siberia" equates to "Tunguska" in the upper diagram.



To appreciate the magnitudes of large impact events, keep in mind that the 12.5 kiloton device exploded at Hiroshima was equivalent to about 10^{14} J (Joules), the Mount Saint Helens volcanic eruption involved 6 x 10^{16} J, and the largest earthquakes release up to 10^{18} J (note: the relation between energy in Joules and in kilotons[kt] of explosive TNT is given by 1 kt = 4.186 x 10^{12} J). In this context, the impact that produced the Sudbury structure (215 km [134 mi] initial diameter) in Canada released about 10^{23} J, roughly 100,000 times greater than earthquakes of magnitude 9.0 on the Richter scale (Sudbury, then, could have generated an earthquake-like response on the order of magnitude 14). Slightly larger is the Chicxulub crater in the Mexican Yucatan, reputed to be evidence for the catastrophic impact event that hastened the demise of the remaining dinosaurs (many families and types had already diminished or reached extinction before this event). In common, both earthquakes and impacts are the fastest known large geologic phenomena, each causing ground disturbances that last only a few minutes at most after their initiation times.

(Note: there is a log linear relation also between crater size [diameter] and energy release, not shown on the above diagram. The scaling formula relating energy to diameter can be approximated by D = 0.1 times the cube root of the energy E in kilotons. By way of example: the cube root of 1000 kilotons [a megaton] is 10, so the diameter of the crater from an event of that magnitude is 0.1 x 10 or 1 kilometer.)

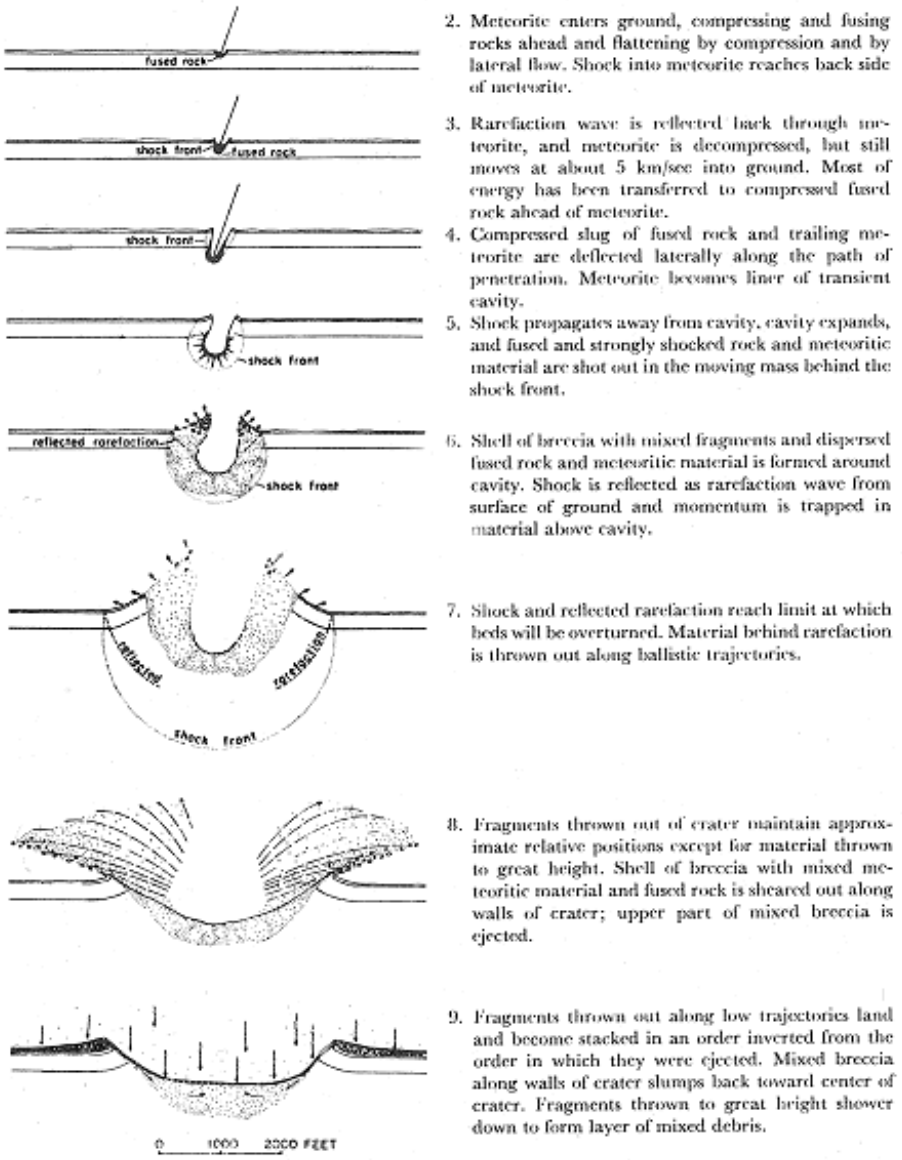
18-3: The Zhamanshin crater, in Asia, is 13.5 km (8.3 miles) in diameter. From the above graph, are there enough nuclear warheads in the arsenals of all nations combined to make a crater of this size if they are exploded simultaneously underground at one place? Are there enough atomic bombs to bring about nuclear winter? Roughly, what is the time likelihood of an impact of the size needed to have something like a nuclear winter forced on the Earth? (And, did you see "Armageddon" or "Deep Impact" in 1998? Does this potentiality for an impact catastrophe worry you?) ANSWER

The source of this tremendous impact energy is the direct consequence of a great solid mass moving at high velocity. *Remember from physics that kinetic energy (K.E.) = 1/2 mv^2, where m is the moving body's mass, and v is its velocity.* To gain a sense of the magnitude involved, consider this calculation. Let a 30 m (98 ft) diameter iron body (in effect, a large meteorite) weighing about 200,000 metric tons (around 440 million pounds) strike Earth at a typical, in-space velocity of 30 km (19 mi) per second (not hours!) (20 mps corresponds to 72,000 mph). This impact would generate about 20 megatons (TNT-equivalent) of energy (about 10^{17} joules) that would cut out a crater about a kilometer and a half (almost a mile) wide and 185 m (607 ft)

deep. This is the size of Meteor (Barringer) Crater, which we will examine later. The ejection process would scatter most of the excavated rocks to a radius of at least 10 km.

The very large kinetic energies (K.E) owing to the great masses (m) and high velocities (v)(remember from physics: K.E. = 1/2 mv^2) involved in impact cratering are converted to dynamic, fast moving transient *shock waves* that diverge hemispherically from the line of impact. These are compressional waves that have high amplitudes. The pressures generated are given in either of two units: Gpa (Gigapascals) or Kb (Kilobars) (a bar is equivalent to 0.971 of an atmosphere of pressure, namely 14.7 pounds per square inch). A kilobar is 1000 kb; a megabar is a million kb; 1 GPa = 10 kb. The shock waves decay (decrease in value) as they spread out from the impact point. When a shock wave meets a free surface, it is in part reflected as a tensional wave.

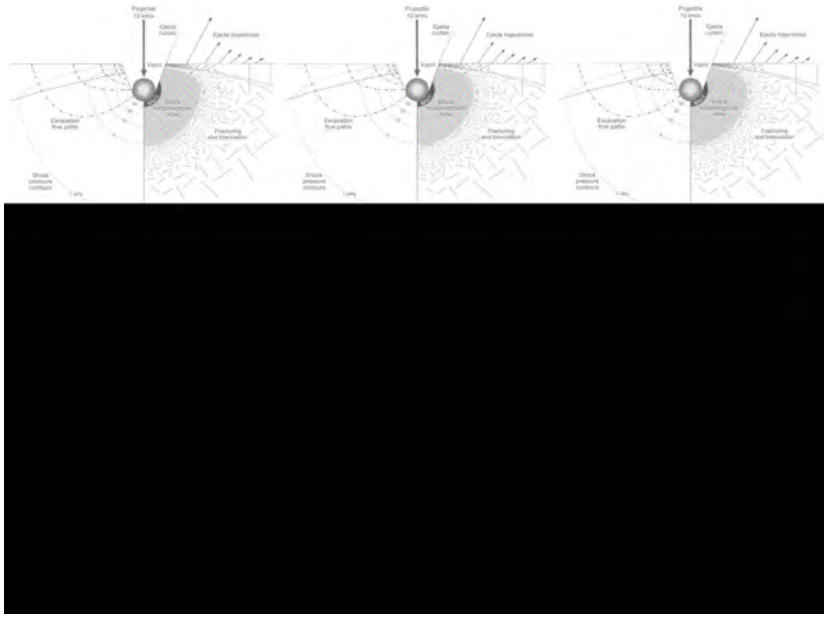
The first truly modern explanation of how an impact crater is formed was presented by Eugene M. Shoemaker in a 1963 paper. The key illustration from this paper is reproduced here:



2. Meteorite enters ground, compressing and fusing rocks ahead and flattening by compression and by lateral flow. Shock into meteorite reaches back side of meteorite.
3. Rarefaction wave is reflected back through meteorite, and meteorite is decompressed, but still moves at about 5 km/sec into ground. Most of energy has been transferred to compressed fused rock ahead of meteorite.
4. Compressed slug of fused rock and trailing meteorite are deflected laterally along the path of penetration. Meteorite becomes liner of transient cavity.
5. Shock propagates away from cavity, cavity expands, and fused and strongly shocked rock and meteoritic material are shot out in the moving mass behind the shock front.
6. Shell of breccia with mixed fragments and dispersed fused rock and meteoritic material is formed around cavity. Shock is reflected as rarefaction wave from surface of ground and momentum is trapped in material above cavity.
7. Shock and reflected rarefaction reach limit at which beds will be overturned. Material behind rarefaction is thrown out along ballistic trajectories.
8. Fragments thrown out of crater maintain approximate relative positions except for material thrown to great height. Shell of breccia with mixed meteoritic material and fused rock is sheared out along walls of crater; upper part of mixed breccia is ejected.
9. Fragments thrown out along low trajectories land and become stacked in an order inverted from the order in which they were ejected. Mixed breccia along walls of crater slumps back toward center of crater. Fragments thrown to great height shower down to form layer of mixed debris.

FIG. 4—*Diagrammatic sketches showing sequence of events in formation of Meteor Crater, Arizona.*

This next diagram (adapted from H. Melosh) summarizes some of the major aspects of the pressure distribution and shock wave-produced phenomena associated with an incoming *bolide* (fancy term that applies to the causative projectile, such as a meteorite or an asteroid/comet):



The incoming bolide penetrates to a distance approximately twice its diameter. As it does, it is itself destroyed by tensional waves, as it vaporizes and breaks up into melt drops and fragments. The ground target is impinged by the generated shock waves. At pressures around 70 GPa, the target material (usually rock) is vaporized; rocks subjected to waves between 50 and 70 GPa are melted. The hemispherical volume of rocks that experience shock pressures between about 5 and 50 GPa is affected by processes involved in shock metamorphism - whose effects are considered in the next two pages. Beyond, to a level around 1 GPa, the rocks are fractured and brokened into pieces that accumulate as breccias (rock fragments [clasts] held in a matrix of finer particles and melt that is quenched into glass) which are deposited both within and beyond the resulting crater. The shock waves thus break up the target rock, setting it into motion along trajectories (the curved lines with arrows; only a few such lines are shown) that carry the now disrupted pieces (some as large as houses but grading in size downward to pea and dust sizes) upward and outward. In this way the crater is excavated by the combination of vaporization, melting, and fragmentation that removes nearly all rock that underwent pressures down to a few GPa and relocates it as ejecta. Rock beyond the final transient crater is displaced and disrupted, with fractures that open receiving injected debris.

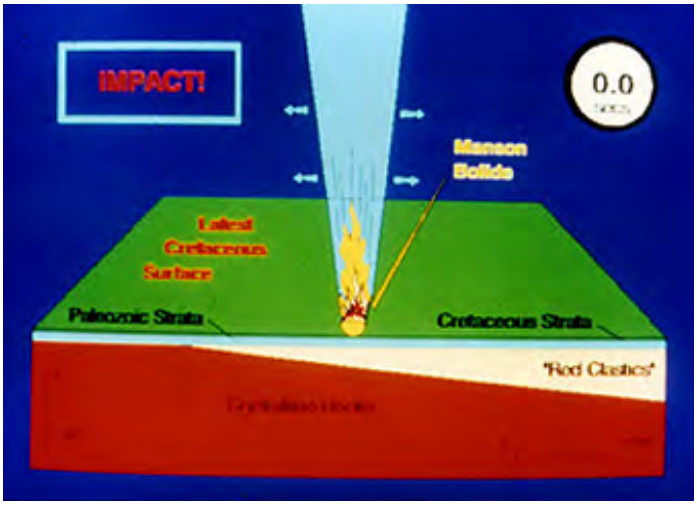
Craters are round or circular (in plan view) and bowl-shaped (as seen in a cross-section) for this reason: the shock wave can be considered as starting from a point (actually a narrow zone of some length related to the penetration of the bolide). This waves travels radially from the point of contact at a uniform velocity. Thus the effects, including excavation, are equal in all horizontal directions (hence, a circle) and approximately so in the directions that define the hemisphere whose outer limit is the true crater.

Impact cratering is unique among natural geologic processes in that very high pressures are attained almost instantly; these decay in minutes or less so that the resulting crater is produced in a very short time. The shock metamorphic phenomena so developed are also unique. Volcanic processes can also produce craters but the processes generally occur over much longer time spans and do not achieve pressures above about 2-3 GPa and hence do not impose the shock metamorphic features in **impactites** (general term for any rock that is affected by shock waves from an impact) that are diagnostic of impact events. Pressures within the crust and mantle can reach well above 5 GPa with increasing depths but the processes involved (such as the weight of the overburden and superimposed tectonic stresses) are applied much more slowly, so that there are no shock wave features imposed.

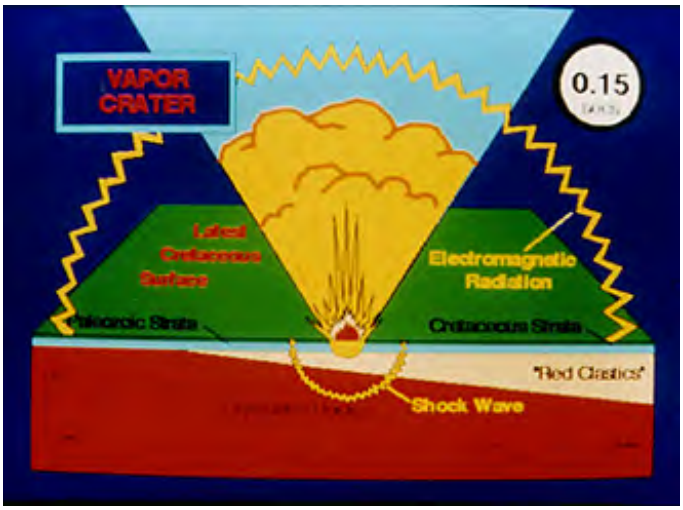
Let us now follow, second by second, the formation of a large or complex crater (one greater than about 5 km [3 mi] wide that has a central peak and concentric slump walls). We use a series of sideview panels created by Dr. Raymond Anderson of the Iowa Geological Survey Bureau (and used here with his permission) to show the steps in developing the Manson structure (see page

20-3a for an in-depth survey of this Iowa crater). The writer, during the 1960s, when he was working primarily on impact structures, is generally credited with "proving" the impact origin of this very large crater, which, at one time, many thought was the "smoking gun" that killed the dinosaurs about 65 million years ago. But a new age date found it to be 74 million years old, which disqualified it as the culprit. This circular structure, 32 km (20 mi) in diameter, whose centerpoint is some 130 km (81 mi) northwest of Des Moines, Iowa, is largely intact but now buried under 30 m (98 ft) of glacial debris.

Each of the following schematic diagrams represents a stage in the sequence of mechanics of formation of a large (complex) crater; read text for description; the number in the upper right circle indicates the time in seconds or minutes after the initial moment of contact between the incoming bolide (probably an asteroid) and the ground surface. While these diagrams (prepared by R. Anderson) apply specifically to the Manson structure, they apply to the cratering process in general.

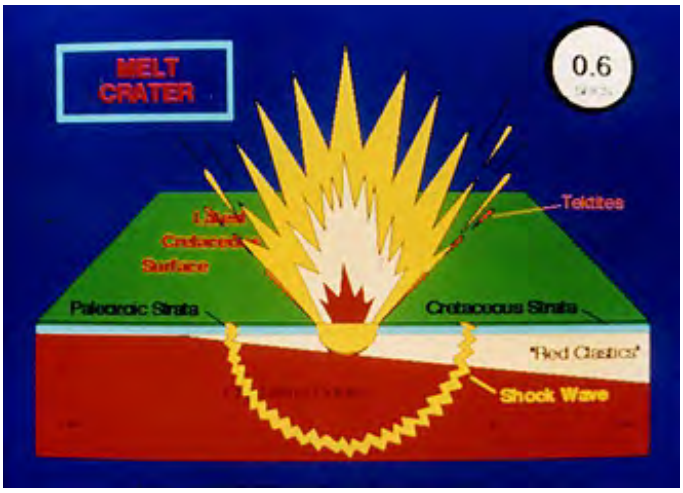


At the instant of impact (0.0 sec), the target consisted of an average of 90 m (295 ft) of Mesozoic sedimentary rocks (mainly Cretaceous in age) (in green) underlain by 495 m (1624 ft) of Paleozoic sedimentary rocks (light blue). These rocks lie unconformably on top of Proterozoic sandstones and other red clastics (yellow), whose thickness increased to nearly 3 km (1.9 mi) to the southwest. This entire section rests on top of Precambrian crystalline (granites and metamorphic) rocks (red) buried at depths to almost 4,600 m (15,088 ft).

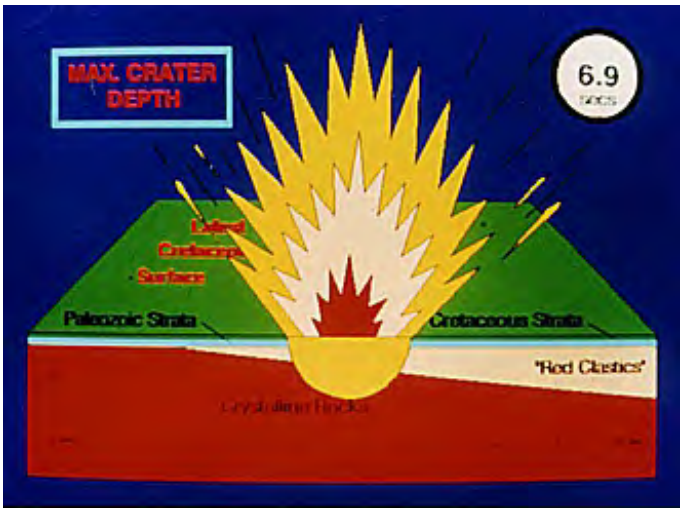


As the incoming impactor (or bolide, a general term that includes both comets and asteroids) impressed onto this late Cretaceous surface, at 0.15 seconds, it was totally fragmented and vaporized. At it penetrated into the rock, it imparted its energy (about 2×10^{23} J) in the form of supersonic shock waves that generated compressive pressures ranging up to a megabar

(1,000,000 atmospheres). We usually find such pressures only at depths well into the Earth (100s of km). Rocks just beyond the point of impact vaporized. An initial curtain of ejecta, consisting of gases and melted rock, streamed upwards in a steep cone, within which is a momentary partial vacuum caused by the projectile passage. The energy released also generated electromagnetic waves that extended into the atmosphere.

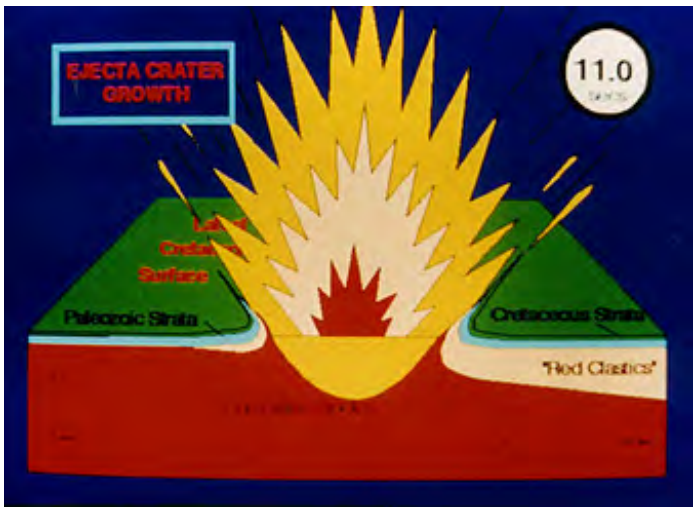


At 0.6 sec the shock wave had progressed along an enlarging hemispherical front well into the target, severely transforming rocks at pressures ranging to about 600 kilobars (kb) (or 60 Gigapascals [Ga], a fashionable new pressure unit) close to the line of penetration. A fraction of the target (up to 10% of the total that the impactor eventually displaces) melted. Some of that molten rock carried downward along with the now-compressed and mobilized rock underwent fragmentation. Some of it pushed out of the crater and fell back nearby, and some literally squirted out as tiny blebs that might have traveled hundreds of miles out of the atmosphere, and then returned to Earth as tektites (glass "pebbles"). A fireball, similar to that caused by atmospheric burning at surface detonations of chemical or nuclear explosions, started to form. Within a few seconds, the excavation phase of the crater commenced, where the shock wave first compressed the rock and then a trailing wave (rarefaction wave) moved through, causing tensile fragmentation. As the waves spread outward and down, decreasing in intensity, peak pressures dropped to a few 10s of kilobars.

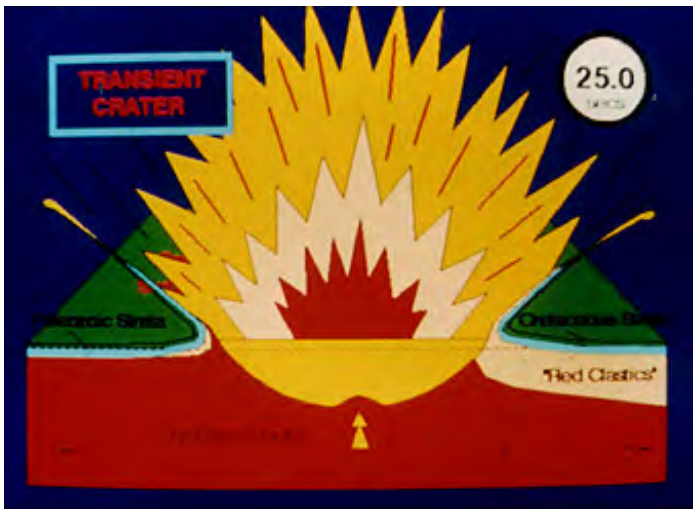


By 6.9 seconds, the initial or transient crater, arising from vaporization, melting, and direct ejection and from centrifugal "shoving" of the target matter outward under compression, had reached its maximum depth. At Manson, this rapidly growing crater front cut down through the Mesozoic, Paleozoic, and Proterozoic sedimentary overburden, well into the Precambrian crystalline rocks. Most of this earthen material received shocks to varying degrees and the

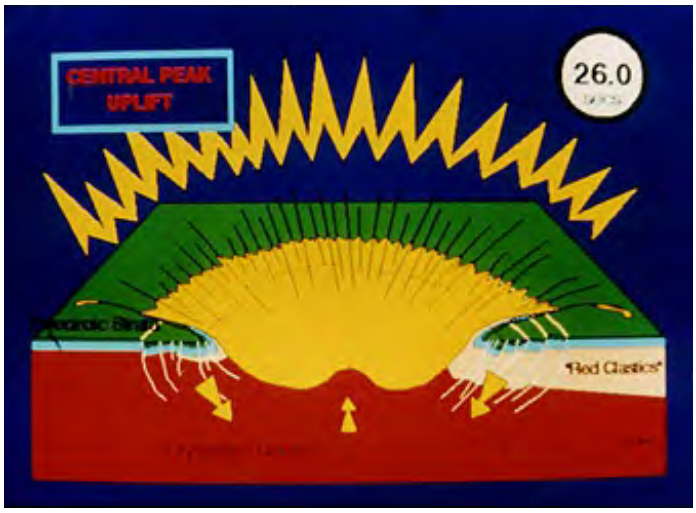
effects of these pressure waves were permanently imposed on the rocks. Trailing tension (rarefaction) waves continued to decompress these rocks and break them into fragments ranging from microscopic in size to objects bigger than a house that eventually came to rest as deposits called breccias.



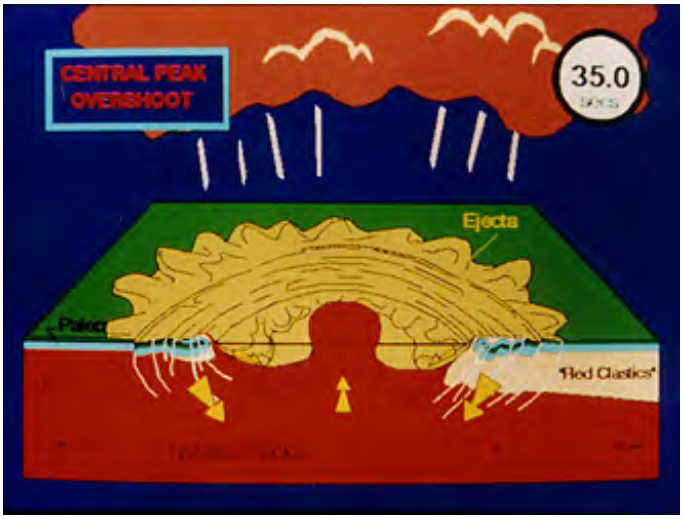
By 11.0 sec, as excavation continued, the peak shock pressures at the wave front had now decayed to under 20 kb (2 Ga). As the pressure waves advanced outward from ground zero (point of impact), they kept breaking the rocks into fragments and blocks, launching them into ballistic trajectories that started particles downward and then swung them up, above the still growing crater along arcuate paths. Most of the material left the site along low to moderate angle paths. Generally, particles deeper and farther out from the impact center left later and usually fell on top of particles that left earlier from near-surface positions. Thus, the ejecta layers tended to deposit in reverse order relative to their initial position in the stratigraphic sequence, with ones lower in the target falling on top of upper ones (although some mixing occurred). Beyond the edges of the crater walls, rock units experienced faulting and folding. Along the upper walls, sedimentary (layered) rocks pealed back so that the layers might even overturn.



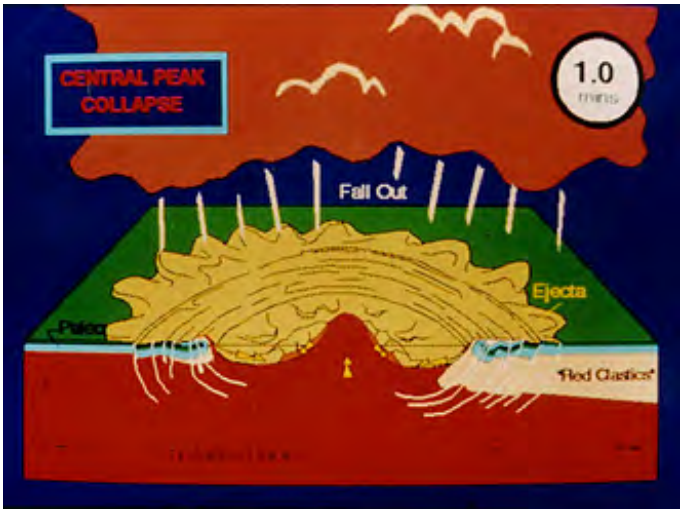
For Manson, the crater reached its maximum excavation diameter around 25 sec, as the last voluminous ejecta emerged. Its upper walls were especially unstable and began to fail along steep concentric fractures and faults. At the central bottom of the crater, the rock below began to rebound upward.



At the transitional 26 second mark, the last ejecta were well into flight. Slices of rock just past the walls now began an inward sliding along faults. The crater base had started an upward movement that soon led to a central peak. This rock material probably behaved plastically as it almost flows upward (a good analogy is the inner blob of water that shoots up into a momentary "crater" forming by dropping a stone into a pond). The collapse of the upper walls may have aided in this effect by pushing downward toward the center.

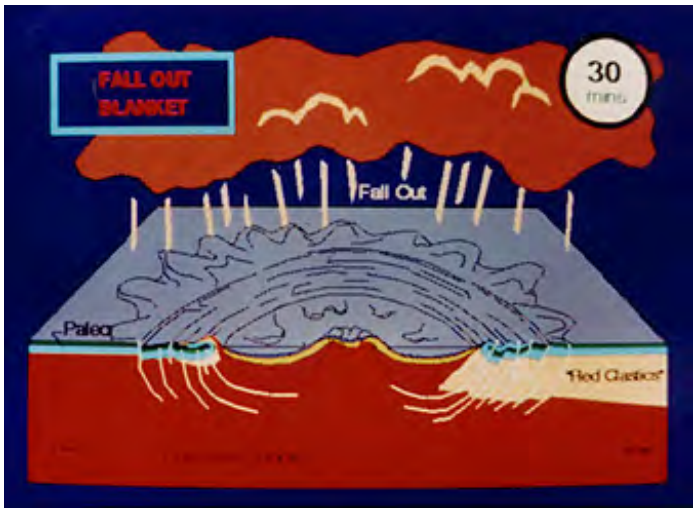


By 35 seconds, the central peak had attained its maximum height (overshoots) and began to founder in collapse.



By about a minute after the impact started, the central peak was well into subsidence towards

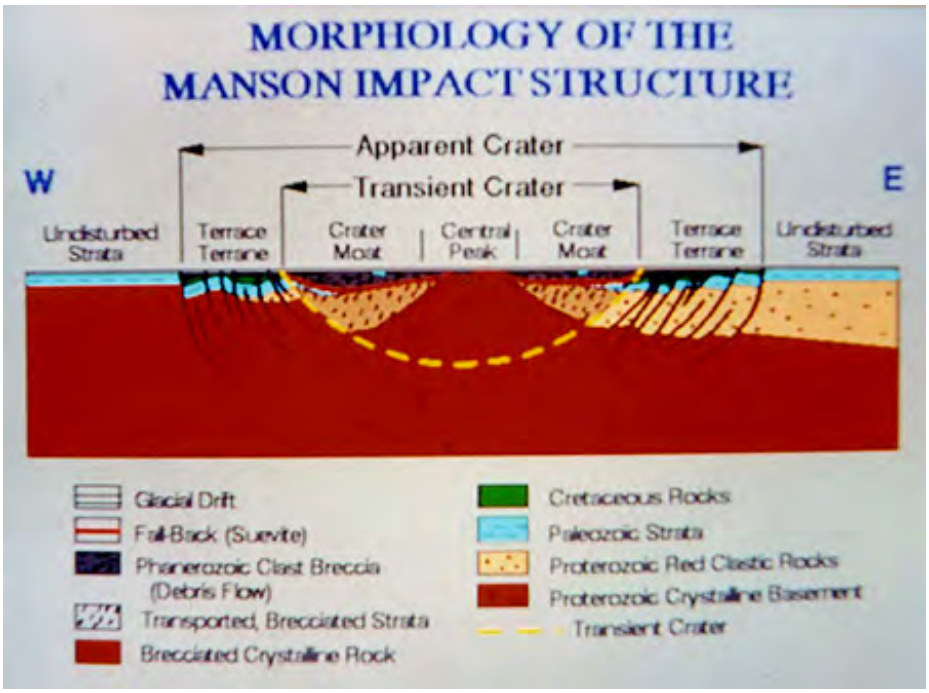
its final position and the crater walls had commenced to slide and tumble inward to form nested or terraced rings (see the Tycho image for an overhead view of these conditions). By this time, some of the material that left at high angles directly above the crater began to descend. The heavier, larger particles settled first, because they passed through the atmosphere more quickly, due to their momentum.



Over the next 30 minutes or so, this fallout piled up in a continuous blanket, inside and outside the crater. Other materials expelled at lower angles formed a wider apron of ejecta that these later fallout deposits covered. Small particles and dust from the event carried hundreds of miles. Manson material has been found in a thin layer at sites in South Dakota, up to 500 km (311 mi) away and the finest sizes traveled in the stratosphere probably well beyond North America (likely global in extent).

18-4: To recapitulate, specify the time or time interval at which each of these stages in the Manson crater formation was important: a) Maximum melting of rock; b) Maximum depth of transient crater; c) Moment when shock wave had decayed to about 20 kilobars (roughly the lower limit at which signficant shock features are produced in the rocks; d) Maximum excavation of fragmented rocks; e) Inward failure of crater walls; f) Start of central peak rise; g) Collapse of central peak; h) Deposition of fallout. ANSWER

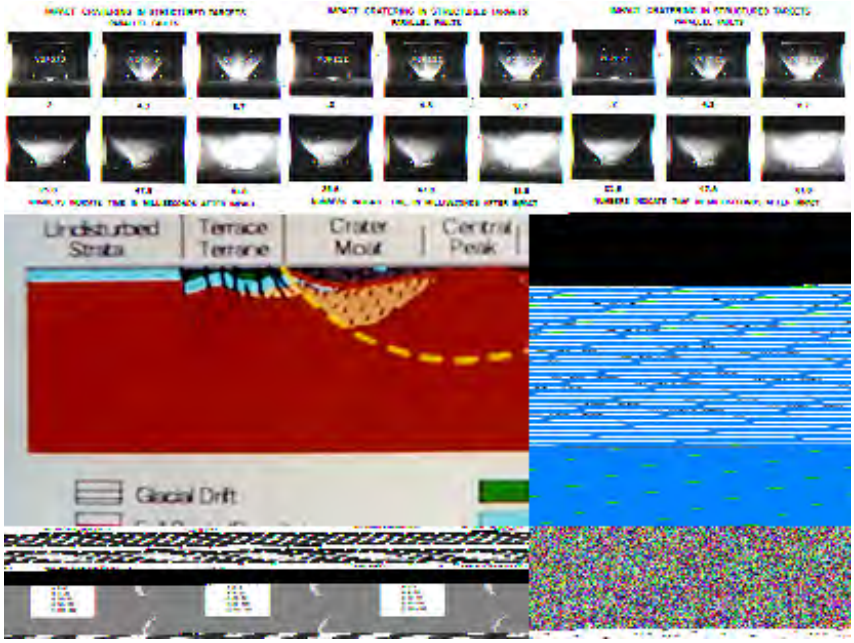
At the time of impact, 74 million years ago, the Manson area was almost certainly under water, because the region lay within a shallow sea. This impact should have produced a tsunami-like disturbance (steep fronted waves that travel at velocities >800 kph [about 500 mph]). Large-body impacts into the open oceans probably spawn huge waves, whose initial heights may exceed 325 m (1,066 ft). If this was so at Manson, the various ejecta deposits would not have formed in the usual sense, because they would have entered disturbed waters and would be irregularly deposited or stirred up by waves moving back into the crater area. The seas retreated a few million years later, leaving the land exposed to erosion. About 10% of the crater's upper structures and deposits eroded. Now glacial deposits of the Pleistocene Age cover the crater, which has no surface expression at all. We show a cross-section (side view) of the Manson crater, as it remains today (glacial cover is a thin gray line):



The dashed yellow line marks the boundary of the final transient crater, modified upwards centrally by the rise of its surface along the central peak. The curved concentric maroon or black lines are fault planes, bounding slides of bedrock that dropped downward to help create terraces. Their outer limits define the maximum (apparent) crater diameter.

18-5: Drillers at the surface above Manson cannot see what they are "aiming" for because of the glacial cover. But, suppose geophysical surveys have outlined the main elements of the crater, so the drill team knows where the center and the rim are located underneath. What would they encounter, as evident in the recovered drill core, if they drilled a) at the center; b) half way out (in the "moat"); and c) into the rim? ANSWER

Much of what has been shown in the above panel cartoons that follow the sequence of events during impact cratering can be reproduced at laboratory scales. This next set of sequential panels are photographs taken by a high speed camera of the development of the ejecta curtain from an impact of a small (centimeter-sized) projectile fired into loose sand from a gas-gun that accelerates the projectile to high velocities.



In the above experiment, vertical black-painted sand was inserted as columns (using thin celluloid tubes to contain this sand) in the target material. After the impact the target sand was

sealed by a liquid glue and then sectioned, one of which is shown here:

The black sand markers just below the crater base show an abrupt bending towards the rims on either side. This confirms that the shock waves induce motions in the sand that roughly parallel the growing surface of the forming crater. Thus, transport of ejected sand is outward at angles; below the final crater base the deformation broadly follows this motion.

For anyone seeking a comprehensive and technical treatment of cratering mechanics, we recommend the book "Impact Cratering: A Geologic Process," by H.J. Melosh, 1989, Oxford Monographs on Geology and Geophysics No. 11, Oxford University Press.

Having now grasped some idea of what happens when impact craters are produced, it could prove interesting to you to run your own calculations in determining Impact Cratering Effects. This Web site brings up a page that allows you to enter various parameters to determine what is predicted to happen when an incoming asteroid or comet strikes the Earth.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

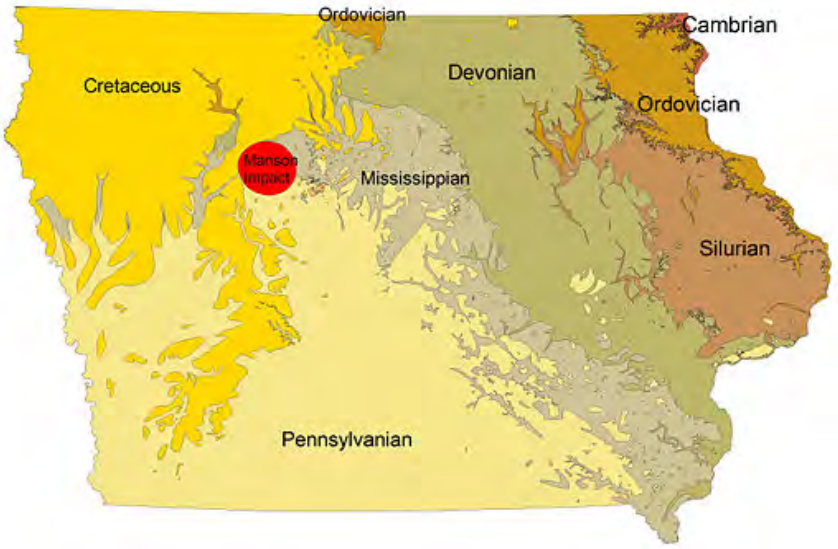
The Manson Impact Crater

The writer first encountered rocks from the Manson Impact Structure in 1965 when I stopped off in Iowa City, IA enroute to my field trip to the West Hawk Lake structure which I was about to study as part of an NSF grant that took me to Ottawa, Canada afterwards to collect samples from drill core into the structure's breccia deposits. While in Iowa City I met with Dr. Richard Hoppin who had first proposed Manson as an impact crater based on samples obtained from the structure as part of a deep well drilling program for groundwater recovery. Upon seeing thin sections of Manson breccias I immediately confirmed that those rocks had undergone high pressure shock wave damage. I published this information a year later in an article entitled "Shock Processes in Geology". Although Prof. Hoppin is credited with "discovering" Manson's apparent identity as an impact crater, I was (years later) credited at a Meteoritical Society meeting as the person who first described the shock effects.

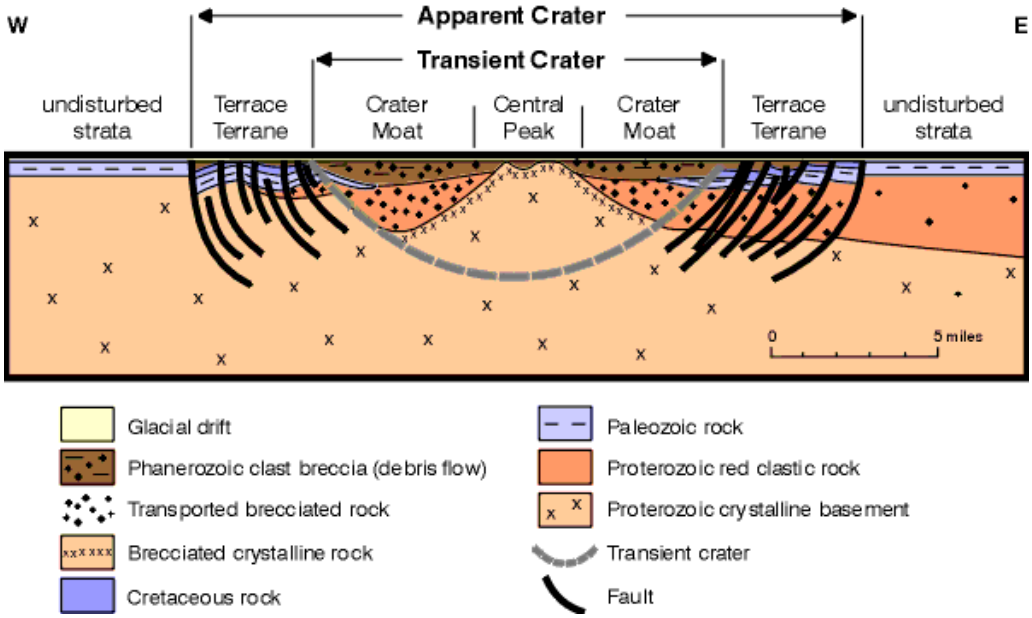
Manson is still the second largest (31.5 km diameter) impact structure in the 48 United States. Interest in it was rekindled after publication of an age of formation near 65 million years. Thus it could have been the prime candidate for the impact event postulated to have killed the dinosaurs. (A more exact age of 74 million years, obtained later, disproved that hypothesis which was good in that the crater seemed too small to have had a worldwide effect.) Because of this interest, a consortium of investigators was formed in the 1990s to renew study of Manson. I was contacted to urge my participation. At that stage I had recently retired from my brief stint at Bloomsburg University, so I had abundant free time. I borrowed a petrograph microscope and Universal Stage from a colleague at NASA Goddard, set them up in my basement, acquired samples from Ray Anderson of the Iowa Geological Survey, had thin sections made (using a small grant from the Barringer Foundation), and spent many exciting hours doing a complete analysis of those rocks from a purely petrographic approach.

So, about 33 years after my last work on shocked rocks (the lunar samples in 1970), I conducted one more study of these fascinating phenomena. This was summarized in a paper, "Petrography of Shocked Rocks from the central peak at the Manson impact structure", published in 1996 with other Manson papers in Geological Society of America Special Paper 302 entitled "The Manson Impact Strucure: the Anatomy of an Impact Crater". Some of the main results I reported are restated below.

At the time of impact, Iowa was covered by Paleozoic rocks, as shown here:

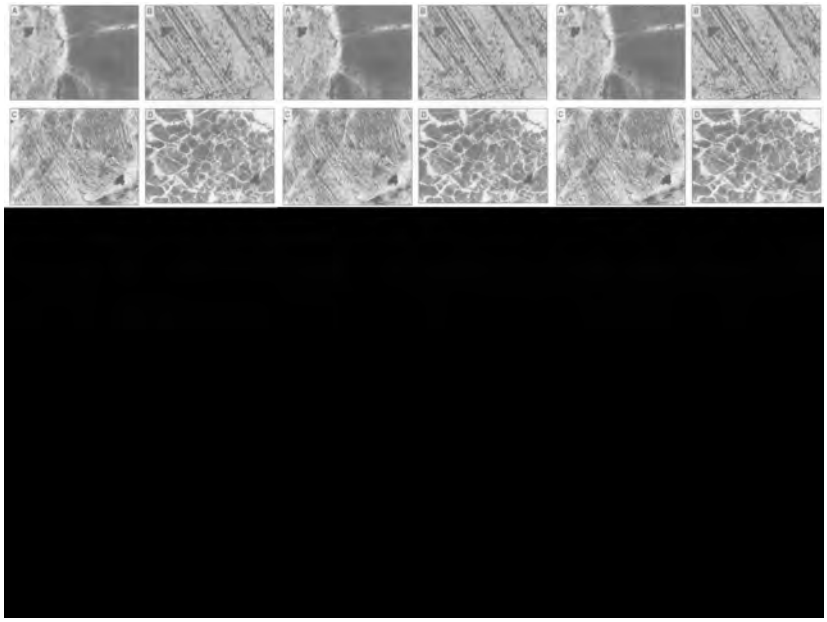


Recall from page 18-2 that Manson has no surface expression since the sedimentary bedrock shown in the above map is entirely covered by masking glacial drift. As a reminder of the geology and morphology of Manson itself, here is a variation of the cross-section shown near the bottom of page 18-2:



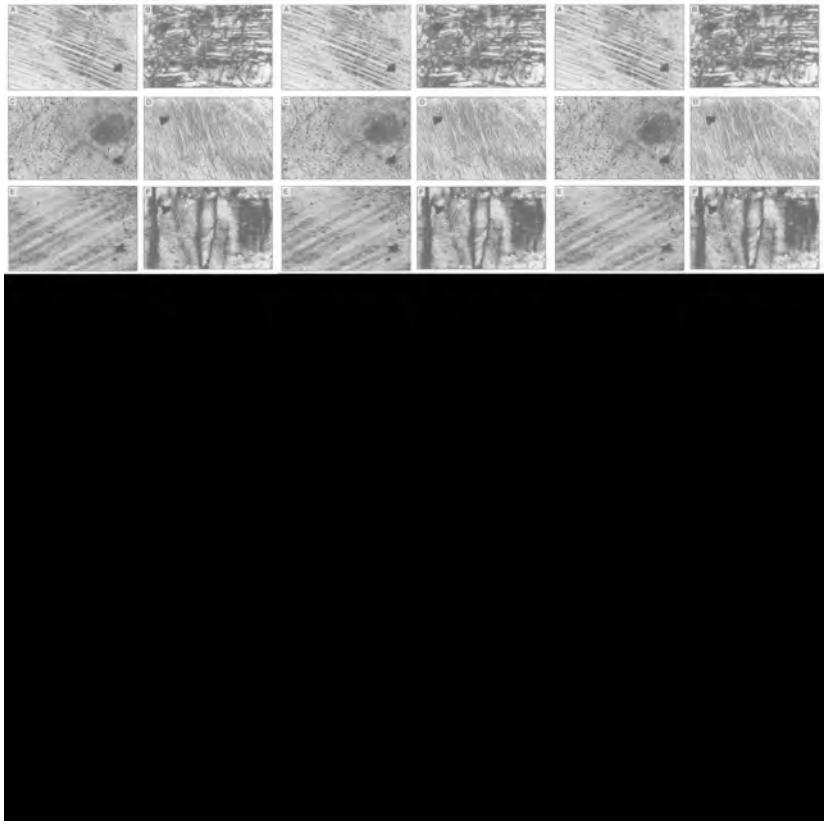
The first (1953) drill hole, 2-A, went to a depth of 145 meters. A second drill hole (M-1), which in 1992 went down to 210 meters, penetrated into another part of the Manson structure; this also recovered core. Most of the core consists of breccias made up of clasts of Proterozoic arkoses, siltstones, and shales which filled a late Precambrian graben and of biotite granites, granodiorite, and gneiss that made up the crystalline basement rocks at impact time; in hole 2-A the lower samples are from blocks of basement rock carried up as the central peak. Occasional fragments of sandstone and shales from the thin supracrustal sedimentary cover were also noted.

Planar features are pervasive through most of the quartz-bearing rocks, attesting to the huge energy release from the impact (in millions of megatons). In the four photomicrographs shown next, these PDFs are evident. The quartz grains in the lower right are "toasted"; this refers to a condition that results in an optical appearance much like the color of white bread toast.



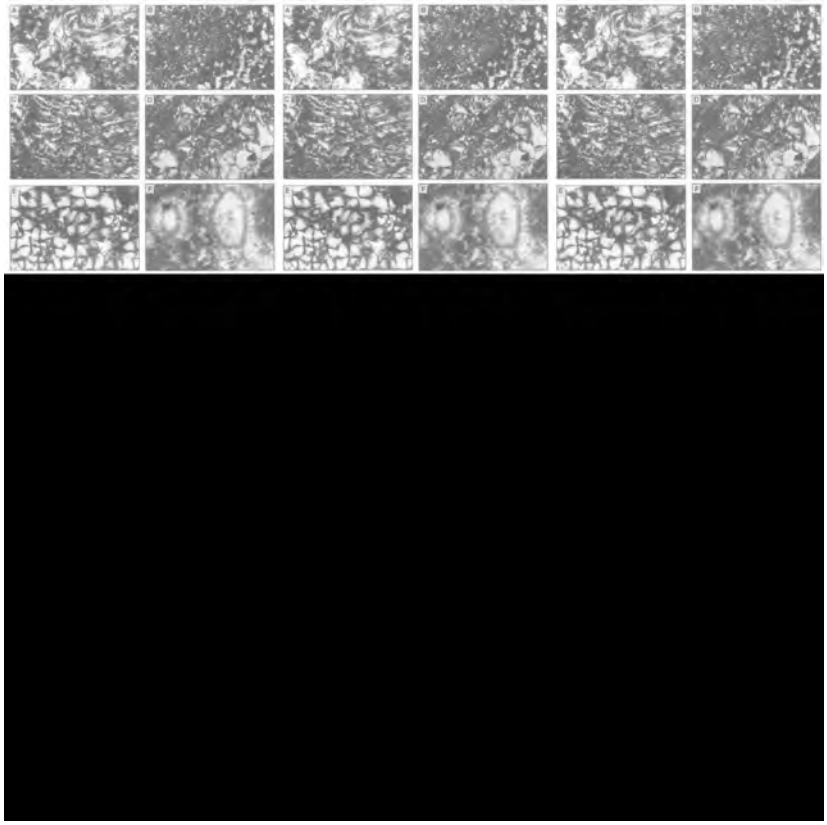
These planar features are mostly oriented along the so-called omega crystallographic plane, in which the pole (line) normal to the plane makes an angle of 23 degrees with the c-axis of quartz.

The next set of six photomicrographs show planar features with various orientations within the several feldspar minerals found in Manson samples.



Structural shock effects in the feldspars: A. Albite twins with alternate pairs showing brown alteration; B. Microcline, with twin units bent and offset by microfaults; C. Single set of PDFs in albite twins; poikolitic quartz (darker gray oval) also has PDFs; D. Bent twins containing PDFs; E. Diagonal (en echelon) PDFs in altered alternate twins; F. Deformation bands with single set of PDFs in a K-feldspar crystal.

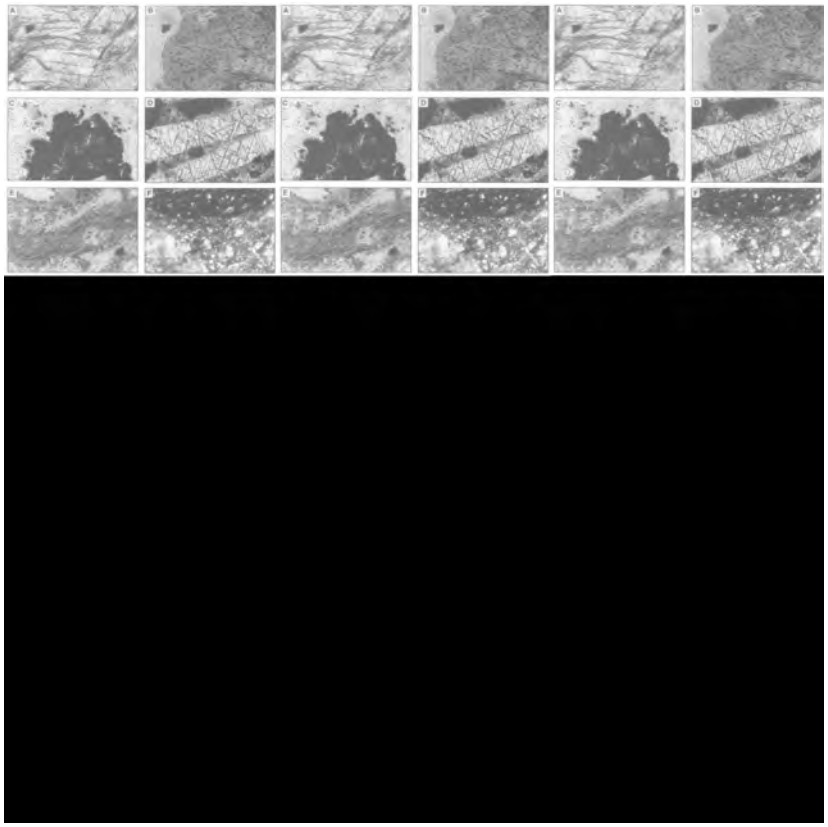
The feldspars show other distinctive shock features, shown in these six photomicrographs:



For the above six: A. Flow bands in a plagioclase crystal that appears to have experienced intracrystalline melting; several quartz crystals have embayed boundaries; B. Recrystallized

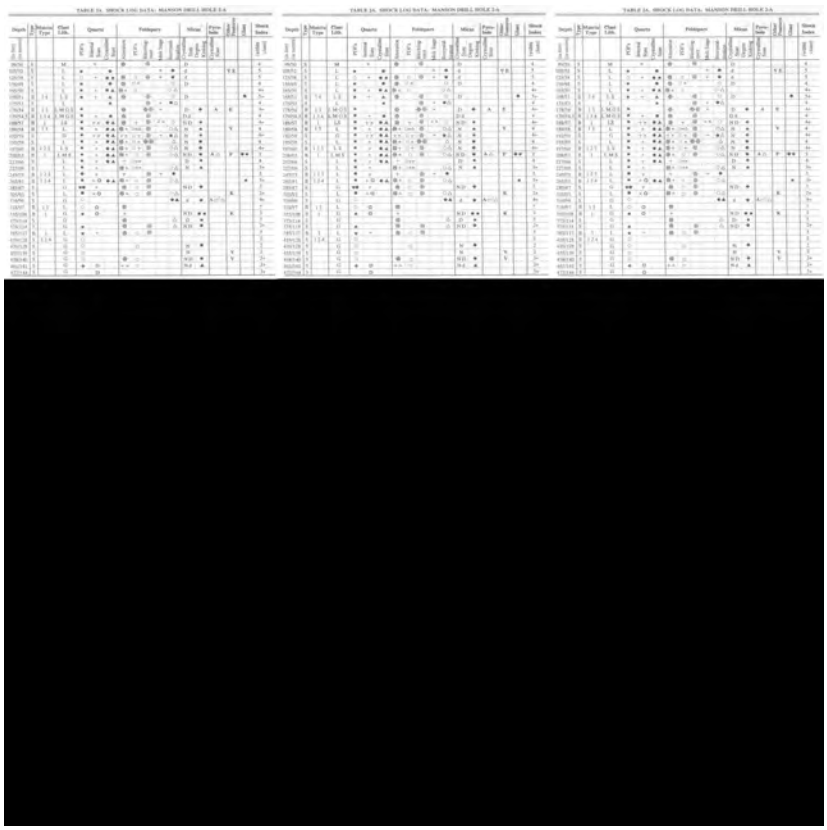
feldspar on left and quartz (right); C. Elongate feldspar microcrystals developed by recrystallization of partially isotropized single feldspar crystal; D. Clusters of feldspar with spherulitic texture; E. Spherules of recrystallized feldspar with internal radiate texture; F. Reaction rim "coronas" of clinopyroxene and plagioclase around quartz grains in a glassy (devitrified) matrix.

The next six photomicrographs show various shock effects in other minerals and in the matrix:



For the above six photomicrographs: A. Lenticular kink bands in muscovite; B. Brown hornblende with some type of lamellae (shock-induced); crystal has abnormally low birefringence; C. Titanite (dark) with thin PDF-like bands (shock-induced mechanical twinning?); D. Three sets of cleavage traces in elongate apatite crystals; E. Flow-banded glass selvage within fragmental matrix; F. The lower part consists of a clast composed of apparently shock-lithified quartz, feldspar, and other mineral fragments held in a dark matrix.

The writer attempted to produce another shock log (different from the West Hawk Lake log shown on page 18-2) for drill hole M-1 samples. It is shown here without its corresponding legend.



The Manson study, synopsized here, proved an exciting adventure for the writer after a 34 year hiatus from working directly on shock metamorphic signatures in impactites. Manson proved to be the equal of the Ries and Canadian impact structures in petrographic phenomena.

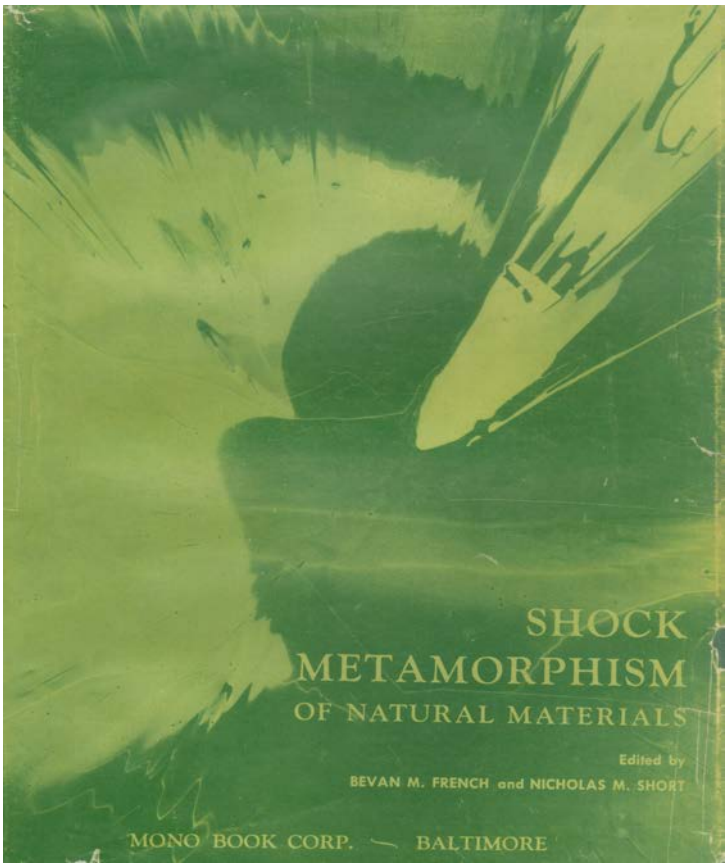
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Before much was known about impact craters, and their mechanics of formation, most geologists considered terrestrial craters since proved to be impact to have a volcanic or some other purely terrestrial origin. But, the pressures involved in any known endogenic process that occurs in the upper crust were quite low. Pressures produced in underground nuclear explosions (or generated in the laboratory by firing high speed missiles at rocks) as determined by calculations and direct measurements proved to be very much higher than those associated with volcanism, mountain building, etc. The rocks affected by these explosions or firings had unique effects in their minerals. Rocks found in impact structures had the same effects and many peculiarities not observed in any other kind of crater. It was concluded that impacts and explosions alter the rocks they act on by what is now known as shock metamorphism. That is the topic of this page. Its applicability to the origin of the black glass tektites is considered.

Note: The writer (NMS) is one of the first geoscientists to work in and develop the specialty field of *shock metamorphism*. This happened serendipitously. While working for the Atomic Energy Commission at the Lawrence Livermore Laboratory (California), in 1960 I became inquisitive about the effects of nuclear explosions on the rocks surrounding them. Using mainly a microscope to examine thin sections made from these rocks, I discovered many phenomena that were not described in the literature (some results of similar effects in impactites had been published but in journals not known/accessible to me). Then, I attended a Conference on Cratering at the New York Academy of Sciences in 1964, presenting the results of my studies of nuclear explosion-induced shock features in granite, basalt, and quartzites, which proved of great interest to attendees who were studying impact craters. I likewise heard/saw several papers describing effects on rocks found at supposed impact craters - these were the same features as I had noted in the nuclear explosion rocks. I realized that the impact cratering process generated huge pressures of the same order as calculated from the nuclear explosion process. My contribution to the field was thus to provide direct experimental evidence that impact sites had indeed experienced great pressures (up to one half to one megabar), beyond levels known from any other terrestrial near-surface process (volcanic explosions produce overpressures of less than 50 kilobars). In effect, I provided the "missing link". To reiterate - since the pressures calculated (but obviously not actually measured) for shock metamorphism imposed on the impact rocks corresponded to that observed in nuclear explosion rocks on which the causative pressures (hundreds of kilobars) had been directly measured by instruments emplaced around the nuclear explosions, the argument is strong that structures containing these shock effects can only have been produced by high pressure-generating events such as impacts (no other mode of origin has proved plausible; this conclusion must stand until geologists find another process that reaches such pressures by purely terrestrial mechanism[s] - pressures of that magnitude exist deep within the Earth but rocks from such depths never can reach the surface).

As a bit of serendipity, I met Dr. Bevan French at that Conference. In 1967, we two happened to sit together on a field trip that was visiting impact structures in Missouri. We began to discuss the need for bringing the researchers into impact cratering and nuclear explosion effects together in some type of meeting. Thus was born the eventual Conference on Shock Metamorphism of Natural Materials, which in turn led to grouping the papers presented into a Book by the same name, published in 1968, which has remained the reference "bible" for this new field.



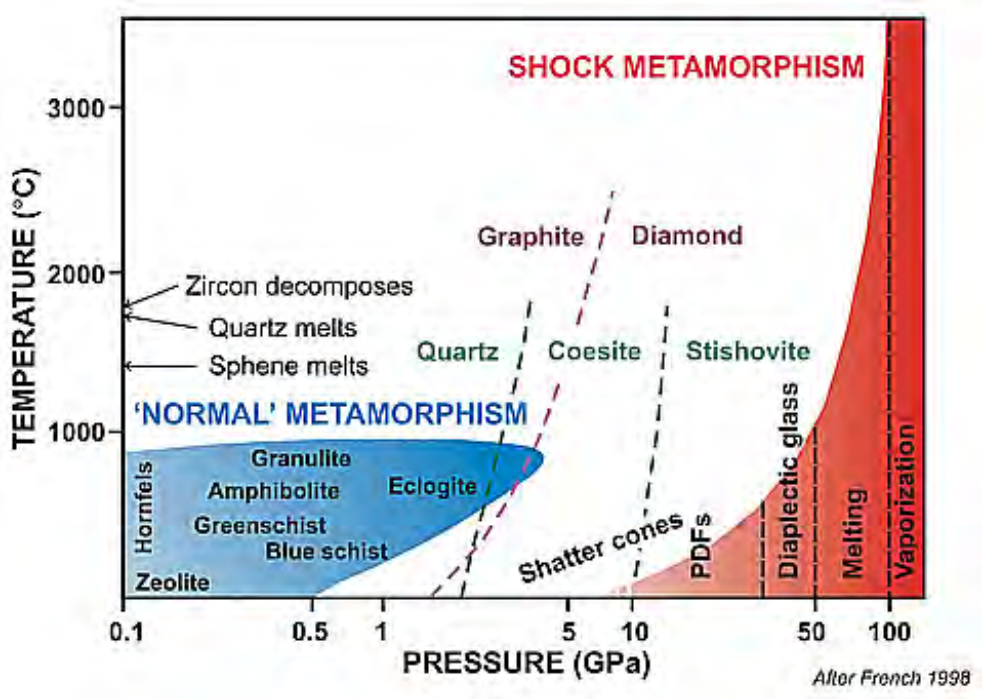
Shock Metamorphism

By far the best indicators of an impact event are preserved in the rocks that were close enough to ground zero to experience shock pressures of 20 to 500+ kb. A kilobar (kb) is the pressure produced by the weight of a thousand atmospheres, or about twice that exerted by water at the deepest ocean bottom. It's also equivalent to the weight effect of about 3 km (2 mi) of overlying rock; those pressures are usually static. Shock pressures are dynamic, with rapid, almost instantaneous rises in a state of compression as the shock wave passes. These pressures are greatly in excess of those that occur in upper crustal rocks from internal forces bringing about conventional metamorphism. The rocks undergo unique changes or alterations described as produced by a process termed **shock metamorphism**.

The features associated with increasing shock pressures are summarized in this table (from C. Koeberl):

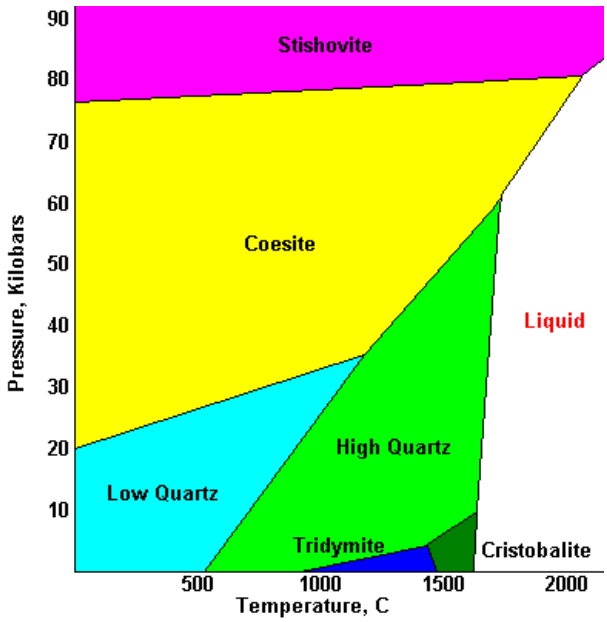
With increasing pressures (and corresponding rises in temperatures resulting from "energy deposition" in the rocks associated with compression), the tendency is for individual minerals to undergo phase changes (into higher density forms) and then to melt, while a fraction of the rock experiencing the highest pressures is vaporized. Note that the post-shock densities of the still solid rocks decreases to the right.

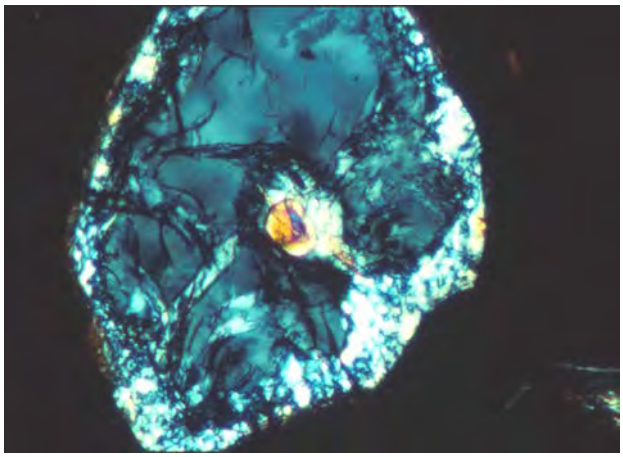
A general plot of shock phenomena as functions of specific temperature (T) and pressure (P) appears here:



18-9: What is the lower limit of shock pressures at which some physical change of state (including phase transformation of one mineral to another) defining a stage of shock metamorphism occurs; the upper limit(s)? ANSWER

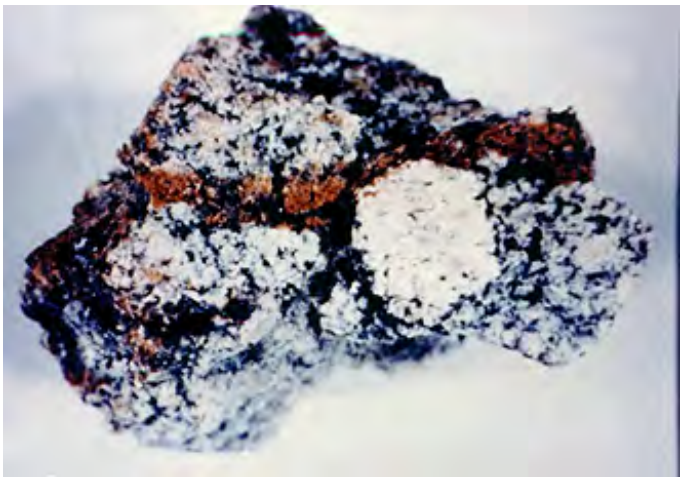
Pressure and heat generated by the shock waves transform the crystal structures of individual minerals in spectacular ways. The common mineral quartz (which crystallizes in the hexagonal system) is perhaps the best recorder of shock-induced changes. *Planar deformation features* (see below) develop in quartz over a wide range of pressures, from 2 to 7 GPa (they also form in other minerals, such as the feldspars, but at somewhat different ranges). Under high pressure quartz transforms to a phase called *coesite* (crystallizing in the monoclinic system). At even higher pressures, another form of silica, SiO_2, known as *stishovite,* (tetragonal system) occurs, although it may be unstable at high temperatures. The phase diagram (stability of a substance at various pressures and temperatures) for SiO_2 is shown below, along with a photomicrograph of coesite.



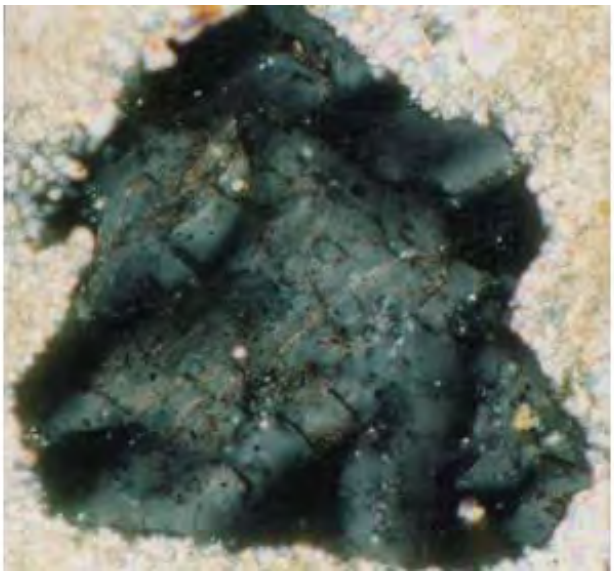


Rarely, diamonds (which on Earth are found in rocks once deeply located where pressures are high or in the laboratory using high pressure presses) have been found at a few impact craters (and in meteorites made up of highly shocked rock; see elsewhere on this page). Here is one from the Ries, made presumably from graphite in pre-impact target rock):

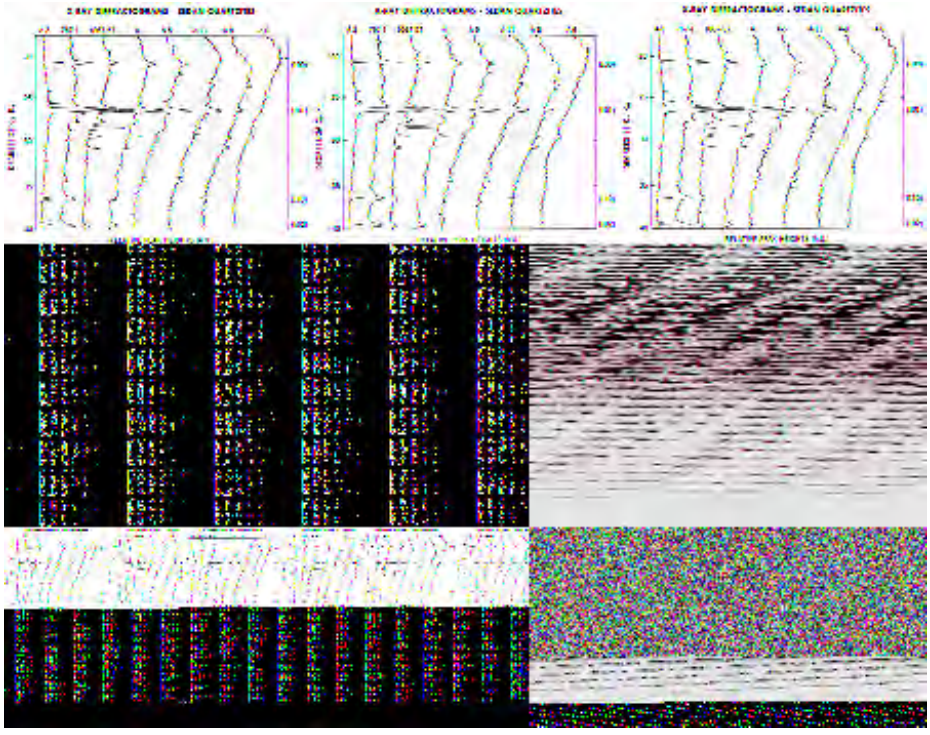
At even higher pressures, crystals may undergo atomic-structural displacements that convert them to glasses without passing through a melt stage. These *diaplectic* glasses usually retain their original shapes (e.g., grains), giving rise to forms known as *thetomorphs* (crystals that retain their original shape while having been converted to the glassy state). The photo below shows a small hand specimen of granite collected by the writer (NMS) from among the ejecta tossed out by the Sedan nuclear cratering explosion (100 kiloton device) within alluvium at the Nevada Test Site. This specimen contains only glass thetomorphs, in which the individual crystals (including the larger multi-sided phenocryst of feldspar) have remained intact without any melt-like internal flow.



Conversion of crystals to glass (discussed in more detail below) is a hallmark of the high degree of damage from strong shock waves. This photomicrograph shows a crystal of quartz converted to a near-isotropic state in a rock collected at the Azuara impact structure in Spain:



Shock metamorphism is progressive, that is, the effects increase or change in style as shock pressures increase. This style change is evident in this series of X-ray spectrometer diffractograms, made from Cu K-alpha radiation on powder mounts of material, extracted from eight quartzite samples, which the writer collected as ejecta from the Sedan nuclear cratering explosion.

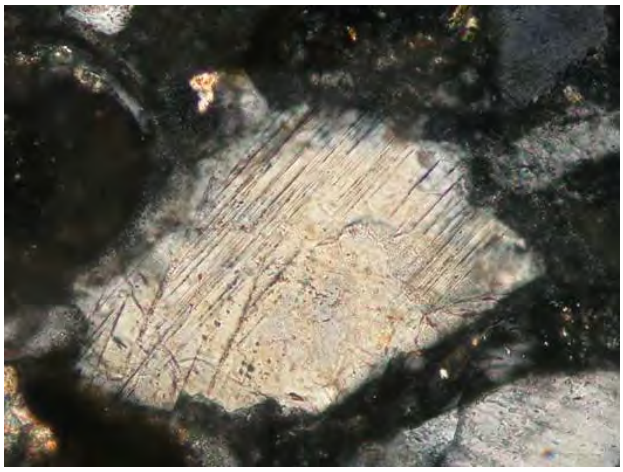


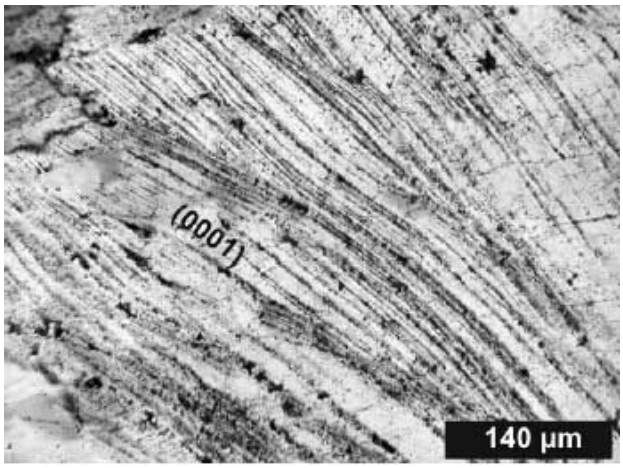
The peak pressures acting on each sample are unknown. I redrew the strip chart record for each sample by arranging the sequence shown from left to right in the order of increased shock damage, based on visual criteria under the microscope. Peaks near 20° , 27° , 36° , and 39° represent quartz reflection planes (crystal indices are on the right). Those peaks near 28° , 29° , and 31° associate with feldspars. The peak at 27° (101 plane) is especially sensitive to the degree of crystal structure integrity. As the level of shock damage increases, peak height diminishes as this structure undergoes progressive disorganization, beginning in the quartz with the development of microfractures (samples A-2 and 767-1) and proceeding to the diaplectic glass stage (samples A-8 and A-6), at which the crystal structure becomes extremely disordered.

We see shock metamorphic effects best in thin sections (thin slices of rock ground to a thickness of 0.03 mm) under a petrographic microscope. In the next series of illustrations, we present these features as photomicrographs. Most sections were viewed in cross-polarized light; PP indicates plane-polarized light.

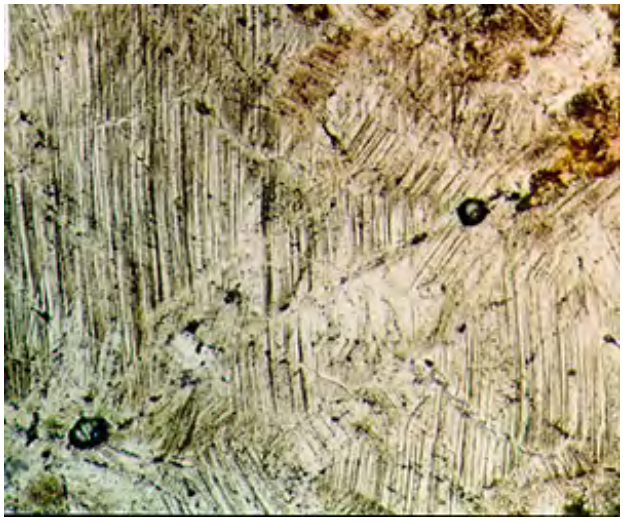
At lower levels of shock pressures, the prime effect on minerals is to shatter them through cleavage and fracturing. Quartz can experience rhombohedral cleavage in non-impactites but it is rare. This cleavage has been observed in shocked rocks, as shown in this photomicrograph of a West Hawk Lake thin section:

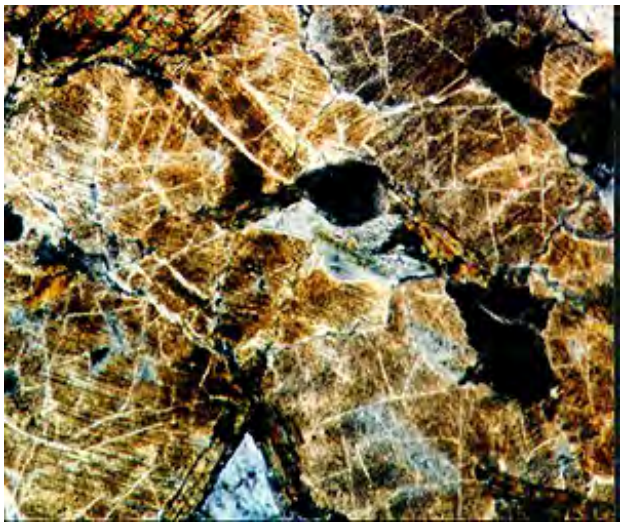
One unique change results from submicroscopic breakdown and slip along crystal planes that produce planar deformation features (PDFs). We show good examples of these features in quartz and feldspar - two very common rock-forming minerals - in thin sections under a petrographic microscope. In thin sections, these may appear as just narrow, usually straight lines (the intersection of the planar feature with the thin section) or they may seem broader because they are "decorated" (darkened by tiny bubbles). Here is an example of each:





Shown next on the top (PP) are decorated PDFs in quartz, within a granitic rock, recovered as core from the Manson structure that was studied by the writer in 1993. Shock damage may be so intensive that it induces a brown discoloration, called "toasting", as seen (bottom image, PP) in this cluster of quartz crystals (interpreted by the writer as caused by the shattering of a single crystal in this granite clast from Manson).





18-10: How many different sets of PDFs (i.e., different orientations) can you discern in the upper Manson photomicrograph? ANSWER

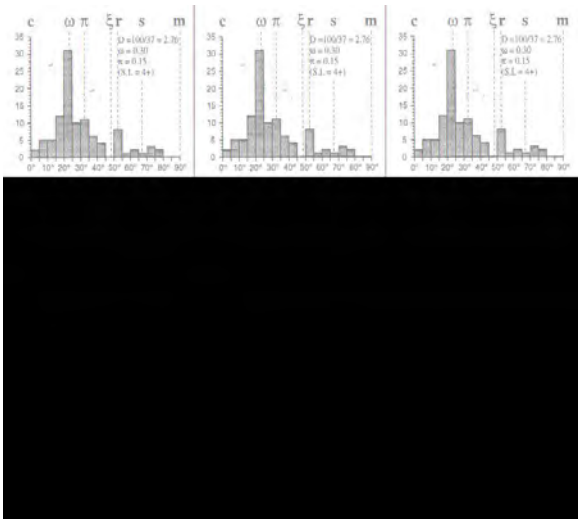
The writer first encountered planar features when I studied rocks involved in the Hardhat 5 kiloton explosion in a small granodiorite pluton near Rainier Mesa at the Nevada Test Site.

Multiple sets of undecorated PDFs in quartz abound within a sandstone (top image, PP), involved in the Sedan nuclear-cratering event. When hydrofluoric (HF) acid etches a slice of

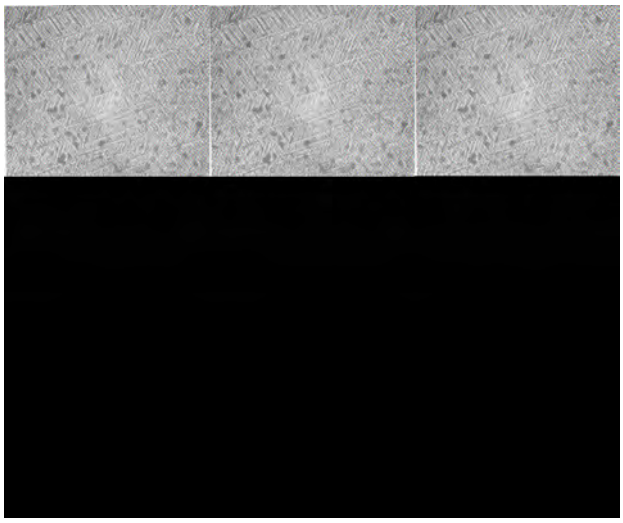
shocked rock, it selectively removes disordered silicate material within PDFs, leaving a gap. In the bottom image is a quartz grain from a Sedan sandstone, as examined at high magnification under an electron microscope, that confirms this removal, suggesting PDFs consist of disordered SiO_2, converted to glass that is more susceptible to etching. Note that the PDFs are indeed remarkably planar.

Planar features resulting from the passage of a shock wave have several distinctive orientations relative to the quartz axis. One such orientation is considered diagnostic since it rarely occurs in tectonites containing quartz, i.e., rocks stressed by mountain building processes; nor has it been found in quartz found in crystalline rocks such as granite and gneiss. This is the so-called *omega* feature, named for the crystallographic plane whose pole (line normal to the plane) is tilted 23.5 degrees with respect to the optical c-axis for quartz. (Omega is a Greek letter, seen in the next

illustration; it looks like a "w") The Miller indices for the omega plane is {1013}, with second 1 being a negative (not showable on this screen). Here is a plot of orientations (made using a Universal stage) of PDFs found in a thin section cut from the drill core recovered at the Manson structure; although a range of orientations is observed, the maximum lies in the 20-25 degree interval, which denotes the omega feature; the second most common orientation is denoted by the Greek letter for 'pi'.



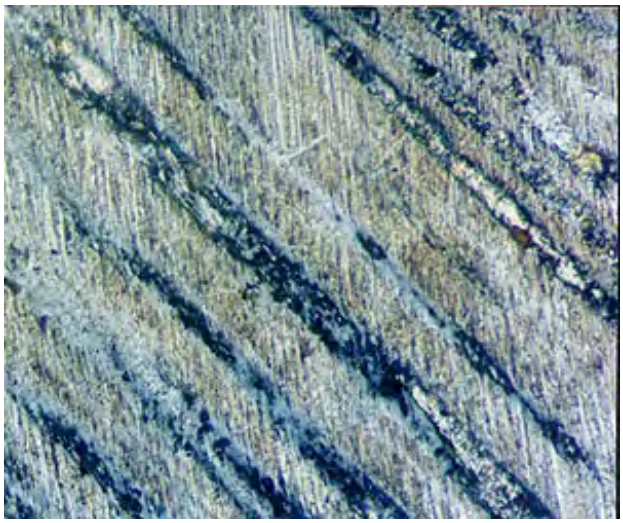
Pi orientation (30-35 degrees) planar features resemble omega features; here are two sets of Pi features from a quartz crystal in a Clearwater Lakes core sample:

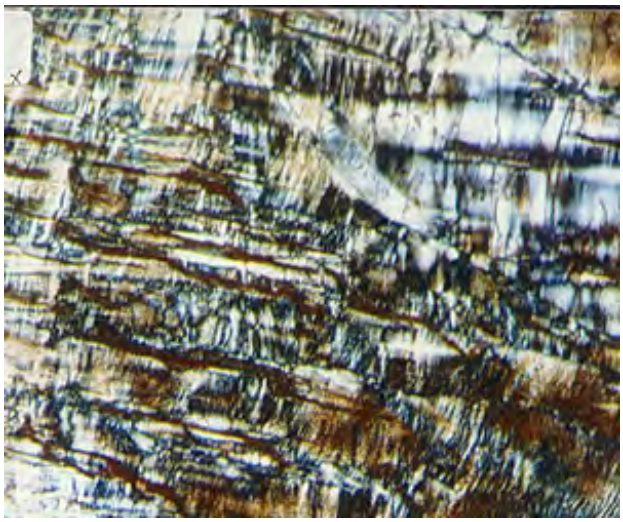


The PDFs do not form in thetomorphs since those result from pressures above the range of PDF formation. But this PDF-bearing crystal from the Azuara structure is an exception. The photomicrograph in cross-polarized light shows part of the quartz grain to have become dark (glassy state) A conjecture: the rock containing it was heated after the breccia was emplaced, to the extent that the grain shown is becoming isotropized.

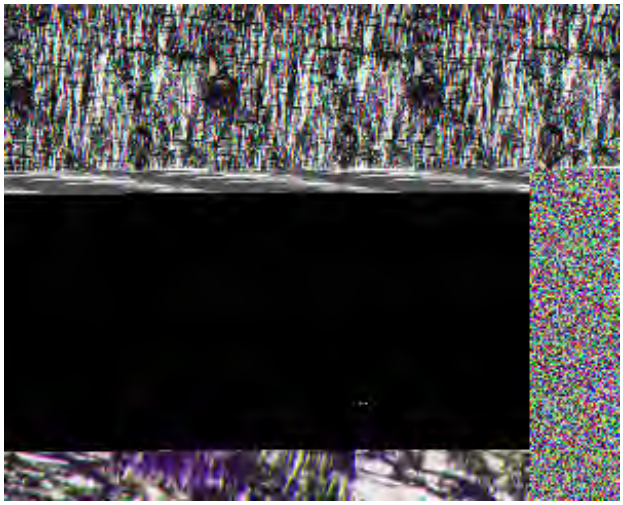
Feldspars, being also tectosilicates, develop planar features as well. Here are PDFs in feldspar within a granodiorite fragment that was shocked during the Sedan cratering event;

In the next pair of photomicrographs, the top image is a single set of PDFs, arranged en echelon (slanted) in alternate twins (gray bands), within a soda-feldspar crystal in granitic rock, taken from Manson. In the bottom image, feldspar within a granite rock, at the Carswell Lake (Canada) impact structure, appears strongly "kinked" (these are also refered to as deformation bands):



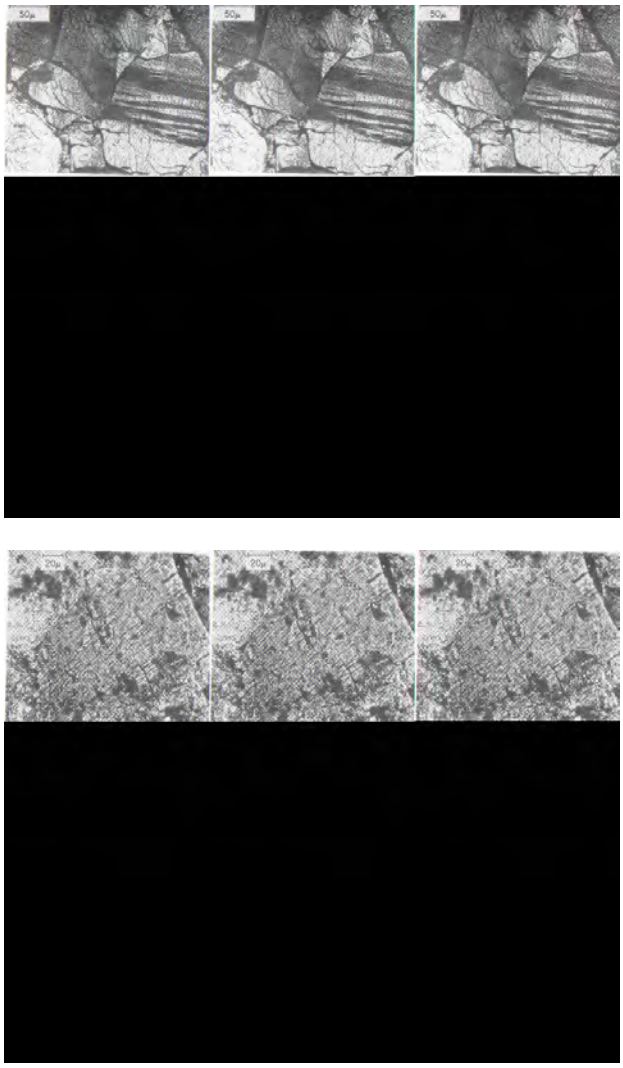


The micaceous mineral biotite, which consists of very thin cleavages, stacked like pages in a book, also kink easily, as shown in the top image (PP) below, for a sample of granite, subjected to a nuclear explosion (the Hardhat event); the lower image shows more detail in another crystal.

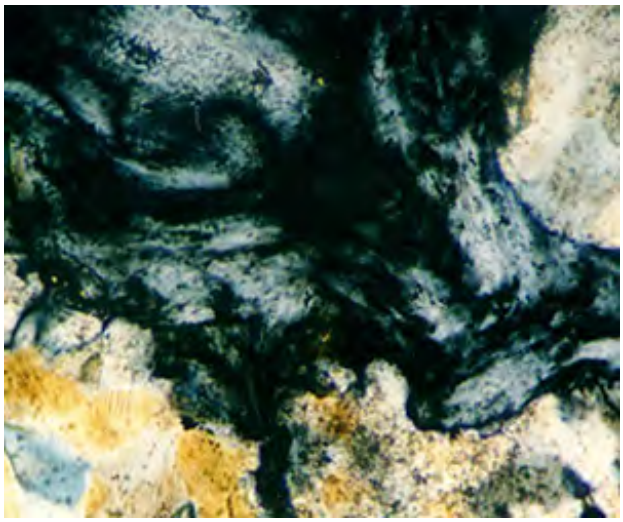


18-11: Visually, what do the biotite kinks remind you of that you have seen before in this Tutorial? ANSWER

Other mineral species experience one or more modes of shock-induced deformation. Below are photomicrographs, the first showing strain bands in an olivine crystal, the second showing lamellae formed in entatite (a pyroxene mineral); the third pictures shock-induced twinning in calcite:

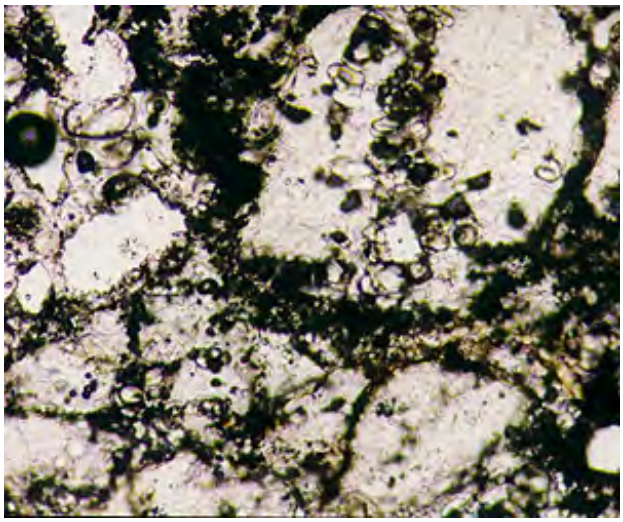


As pressures enter the 40 GPa (400 kilobar) range, feldspar in a Manson granite began to melt, as shown in the bottom photo, by dark and gray flow bands, but the rock remains intact (the quartz is still crystalline).

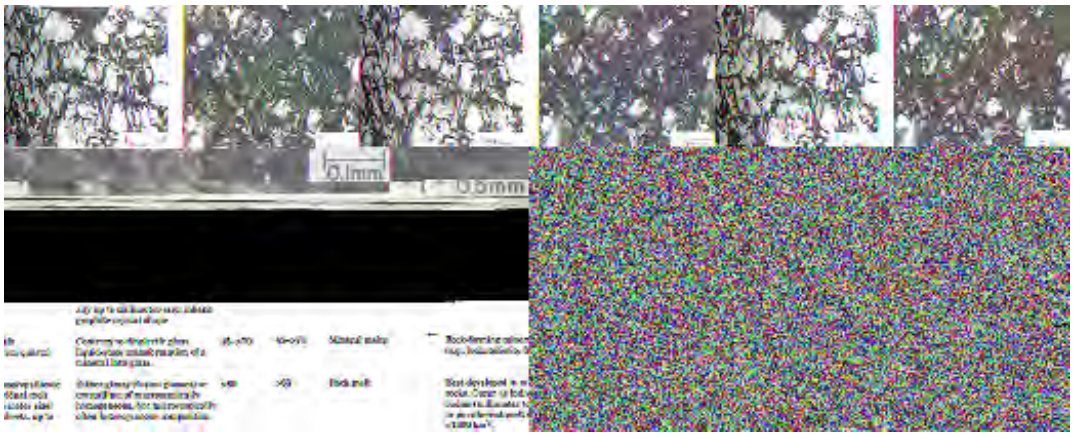


Because of the writer's close association with a major study of Manson in the 1990s, I have decided to add a page, namely page 18-3a, that the reader can access optionally to learn more about this structure from a petrographic viewpoint.

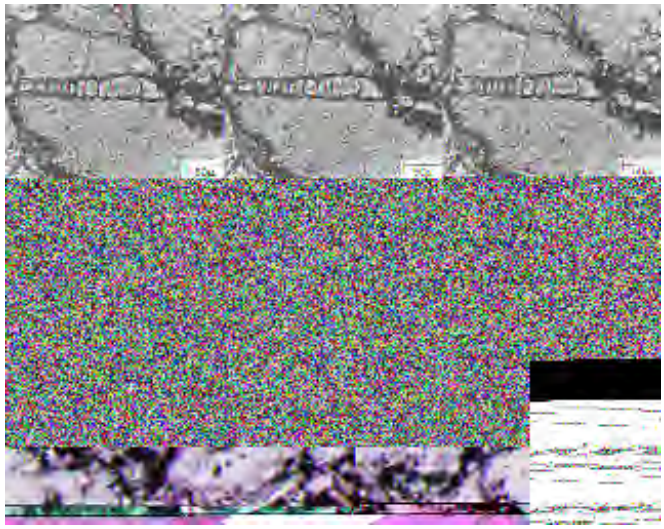
At more extreme pressures, mineral grains may convert into glass without any change in their original shapes, i.e., the texture is preserved, while the composition changes from crystalline to glassy. We show these thetomorphs in a microscopic view (PP) for quartz grains in a sandstone rock, collected from around the Sedan nuclear crater at the Nevada Test Site. The SiO_2 appears to have undergone incipient vaporization, as indicated by the occasional round vesicles.



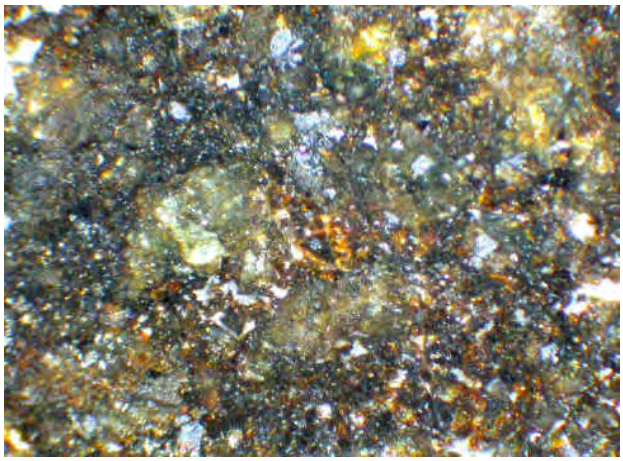
Here is a pair of photomicrographs. The left shows Sedan quartzite in which all grains are now glass; the right image shows the same area of a thin section in which the petrographic microscope's polarizers have been engaged (Crossed Nicols), showing here that the quartz grains are all isotropic as expected from glass (glass under the microscope shows no birefringence and hence will be dark at all positions when the microscope stage is rotated 360°).



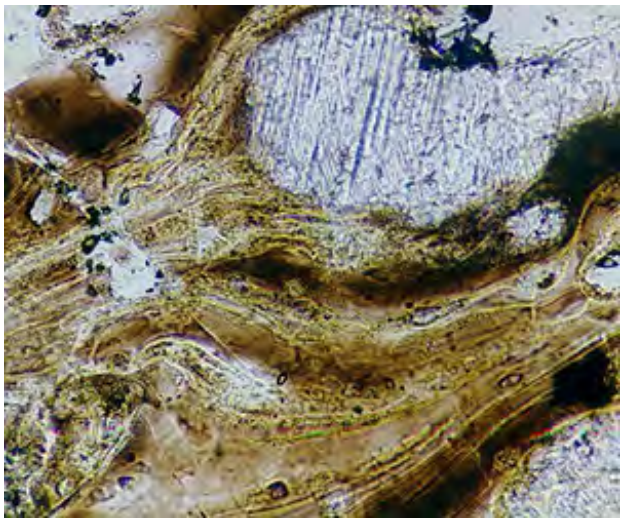
At pressures within the 400-500 kilobar range, rocks melt as though severely heated (above about a half megabar, rocks start to vaporize). In this photomicrograph, a quartz thetomorphic grain has started to vesiculate (silica vaporizes).

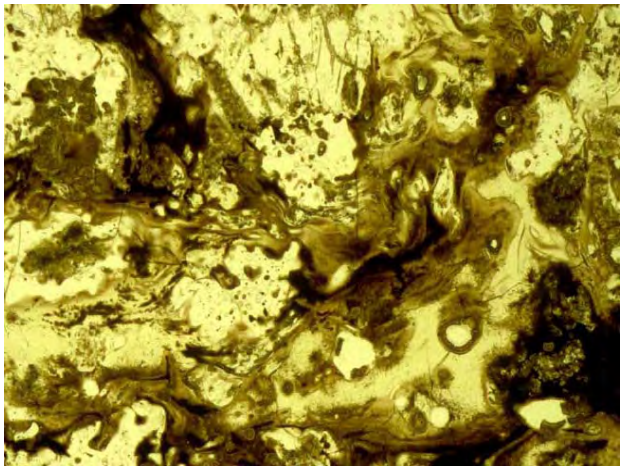


In impactites, the melting quickly quenches into glass and may become singular masses mixed in the breccias, or as discrete layers near the bottom of the final crater. The hand specimens below are examples of breccia from several impact structures, starting with two from the Ries crater in Bavaria (called *suevite,* locally at the Ries crater but the term is now generic and applies to breccias at other impact structures).



In the top below. is a microscopic view (PP) of the breccia from the Ries crater that contains shock-melted rock (brown flow bands) and occluded fragments of quartz with PDFs. At the bottom is a photomicrograph that shows diaplectic glass of two compositions, derived from quartz and feldspar (converted to thetomorphs whose boundaries have been distorted by flow) in a Ries rock that was intensely shocked.

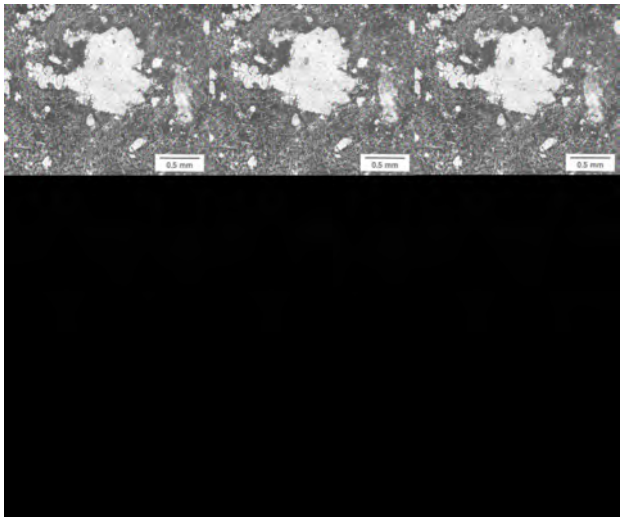


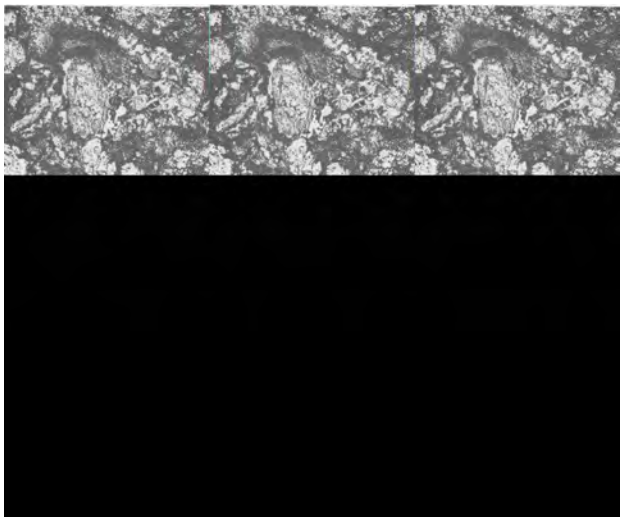


At some impact structures, fractures opened beyond the true crater walls as the event proceeded were filled with injected fragments and melt. Two examples from the Rubielos de la Cerida impact structure in Spain illustrate typical fracture-filling veins;

An impact-injected vein whose filling includes shock melt is called *pseudotachylite*. Here is an example from the Vredefort structure, in which the veins shown are at hand specimen size:

The melts within large impact craters are commonly mixed with fragments of shocked rock. They can, however, be fairly uniform in texture and could be mistaken for volcanic flow rock, except they usually contain some fragments that display shock effects. Here are two examples





The next three show hand specimens of impact melt:

The next photomicrograph shows a melt from the Manicouagan (Quebec) crater, whose composition is close to that of feldspar, in which new crystals of feldspar have grown in place rapidly as the melt quenched.

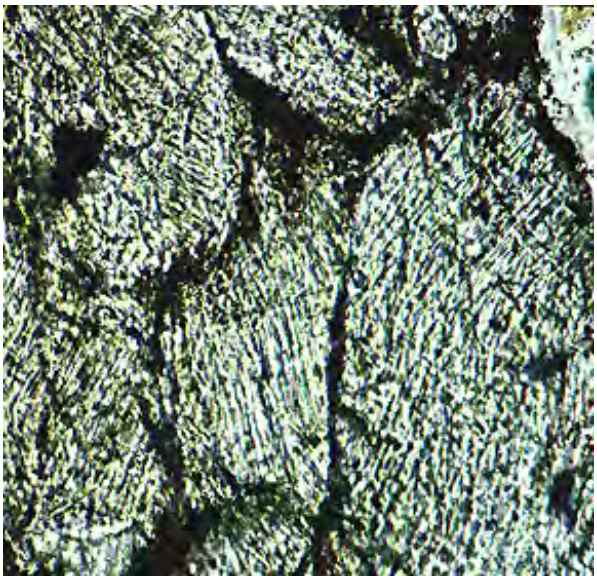
Thetomorphs and the types of PDFs shown above occur in nature only within rocks involved in structures that have at least some of the characteristics of impact craters. They also readily form in rocks surrounding nuclear explosions, where instruments directly measure pressures in hundreds of kilobars. And, we can make them experimentally in the laboratory using controlled explosions to create these pressure ranges, such as in the implosion tube method invented by the writer (see below). They are not present as such in breccia rocks associated with volcanic explosions, where pressures rarely exceed 10 kb. As the writer has already stated, their presence in quartz and/or feldspar is decisive proof of an impact event as the cause of a deformed structure.

The writer, while at Lawrence Livermore Laboratory, took advantage of the facilities at Site 300 which the physicists used to conduct shock experiments on materials. One was a 16-inch cannon barrel and loading chamber obtained from a decommissioned battleship. On several occasions I supervised the firing of flat-nosed shells at rock targets (in the enclosed housing), recovered the samples, and studied the shock effects. Here is that cannon.

The rocks hit by these shells only showed one distinct shock effect: Quartz and feldspar crystals were invariably fractured, even shattered.

In 1964, the writer invented a new way to shock rocks. Using some of the principles by which nuclear devices are detonated, I designed the implosion tube method, whose setup is shown here prior to detonation.

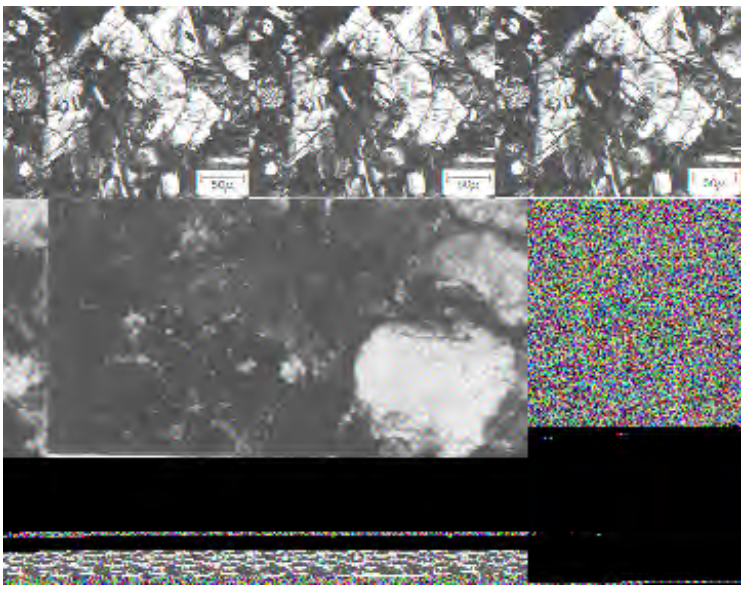
The small tube, shown on the right, is made of brass or steel. It is hollowed in the center. Into this are fitted either cored rock or mineral samples, or loose sand made up of such samples. The tube is welded shut. It is then placed with positioners along the central axis of a large, 4-inch diameter, aluminum tube (left) that is filled with a liquid explosive. Upon detonation, the cylinder is blown apart but shock waves also move inward, squeezing (imploding) the sample container which is recovered. Samples inside experience shock over a range of peak pressures up to 500 kilobars. The samples are released from the tube by sawing or trim-cutting on a lathe, and then made into thin sections for examination under a petrographic microscope. Those containing quartz grains or crystals usually display the same kinds of planar deformation features found in impactites. This next picture is a photomicrograph of planar deformation features developed at an estimated 20 GPa (200 kilobars) in a sample of unshocked sandstone placed in an implosion tube.



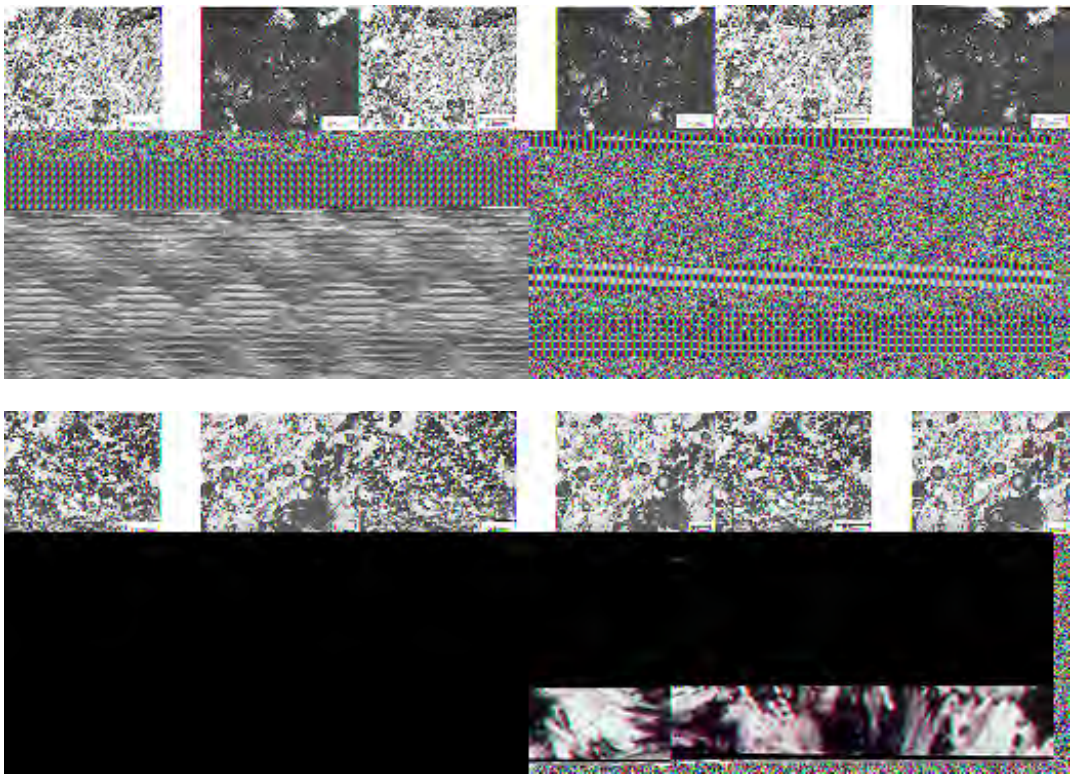
PDFs also formed in feldspar crystals in a granodiorite shocked in an implosion tube:

An interesting effect developed in one sandstone mounted in the implosion tube. Quartz progressively experienced increasing shock damage in grains from the outer edge of the sample (fracturing), through to the interior (PDFs and then isotropization at the center). This photomicrograph shows part of the gradient variation:

Of particular interest is the progressive shock metamorphism of a basalt flow unit that was the site of a small nuclear cratering explosion (Danny Boy event) at the Nevada Test Site. Samples from that event showed mainly fracturing, whose numbers per unit area of thin sections made from samples recovered at progressively closer distances to the crater wall increased as the wall was reached. This is a typical view of fractured plagioclase in the basalt.

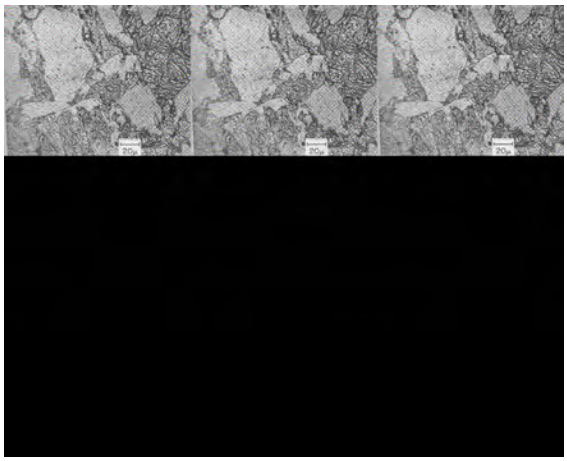


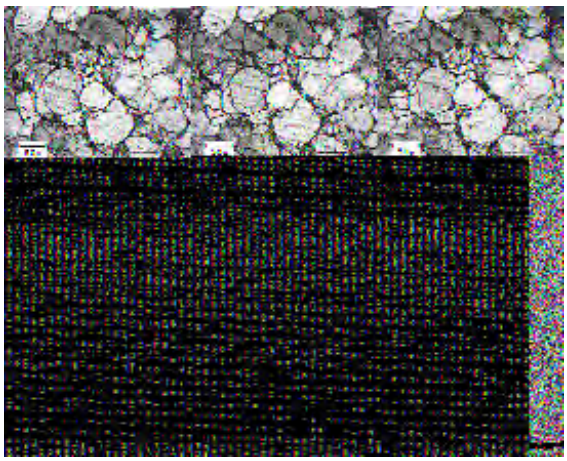
Of more interest is the behavior of basalt at high shock pressures. Samples of the basalt were included in several implosion tube experiments. Below are 4 photomicrographs. In the top two are uncrossed (left) and crossed (right) Nicols views of a thin section of basalt that was shocked to an estimated 300 kilobars. One can see that all the light toned plagioclase laths are converted to glass thetomorphs (Feldspar glass produced by shock pressures is found in meteorites and is known as Maskelynite). As pressures increase (perhaps as high as 400 kilobars), the plagioclase begins to melt (bottom left) and the rock starts to vesiculate. Full melting of the basalt in the implosion tube (also shown in plane polarized light) appears in the bottom right. It was this work that probably led to the writer's (NMS) selection as a lunar samples investigator because the Apollo 11 site was predicted to be basalt that had experience repeated impacts.



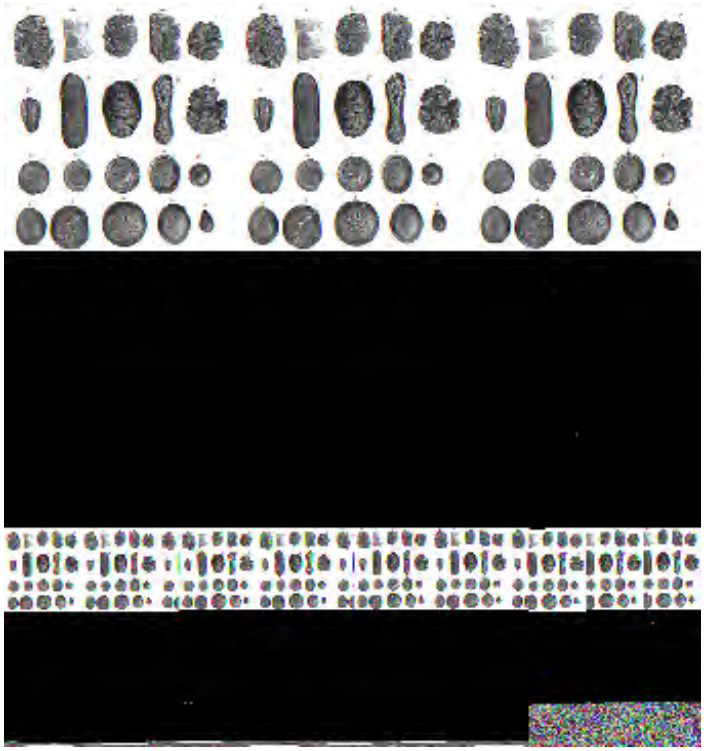
Over the course of a year (1963-64), the writer used this new technique to shock more than 30 rock types, some 25 different minerals, and several mixes of sand (the olivine and enstatite examples on this page were from the implosion tube experiment). Most of the phenomena found in rocks from impact structures were reproduced by this technique (including, at that time, the only known example of calcite glass).

Instant rock, formed by shock lithification, was discussed in connection with the material around the Arabian Wabar crater (page 18-2). Shock lithification has affected much of the lunar surface rubble (regolith and eject deposits). A sample containing loose quartz grains was converted by implosion to a "sandstone" much like Wabar rocks. Another sample containing loose grains of olivine and plagioclase was included in an implosion tube. The implosion converted it to a shock-lithified analog to a basaltic rock, as shown here:





The writer's work on shock processes made a serendipitous contribution to a controversy raging in the 1960s regarding "mysterious" pebbles found widely scattered in Asia, Europe, Africa, and North America. These are called *tektites*, small pieces of black, obsidianlike glass in elongate irregular to "leather button" shapes. Below are some typical examples:



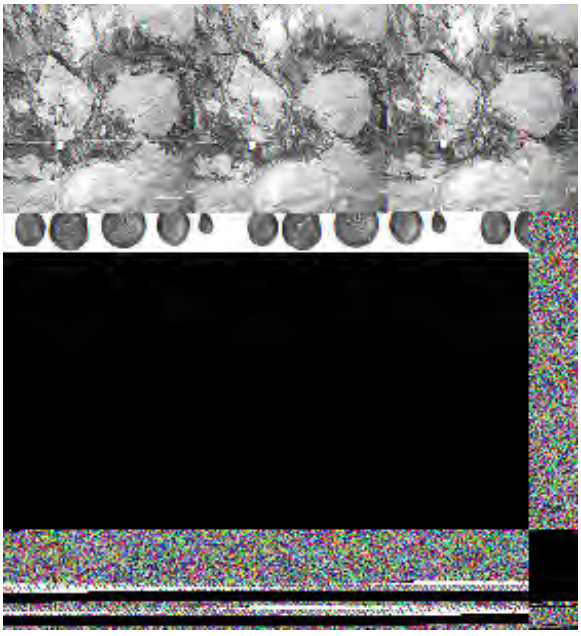
The composition of most tektites ranges from about 60 to 80% SiO_2. No minerals occur within tektites but melted quartz glass (Lechetelierite) smeared out by flow is rarely present. Tektites are totally devoid of water. The button shape in particular is consistent with an aerodynamic reshaping as though the glass was initially molten, rapidly cooled, and surficially remelted upon

passing through the atmosphere.

Then, and more so now, the majority opinion considers tektites to be formed by the high pressure range developed during impacts on terrestrial rocks. At those pressures the rock target would be molten and the blobs of melt tossed far from the crater into, then out, and then back into the atmosphere, falling land or sea (there, the equivalent of the tektites are microtektites, in tiny sphere shapes) scattered over wide areas. Researchers now believe they have found the source impact craters for some of the tektite strewnfields: Bosumtwi crater in Africa; Australian crater(s) for the Indochinites; the Ries crater for the Moldavites (Czech Republic). The American tektites (Georgia) have not been tied to any known crater (the Chesapeake crater is now a possibility).

There has been a small group of researchers with a "minority" opinion as to tektite origin. Led by the late Dr. John O'Keefe of Goddard Space Flight Center, these individuals considered tektites to come from silicic volcanoes on the Moon. However, no volcanic structures of the silicic type have been found there and the composition of lunar rocks does not fit that of tektites. However, a strong argument made lunar tektite advocates is the total absence of water (almost all moonrocks are anhydrous). Since most terrestrial rock contain varying amount of water, some should be present if the tektites derive from rocks on Earth.

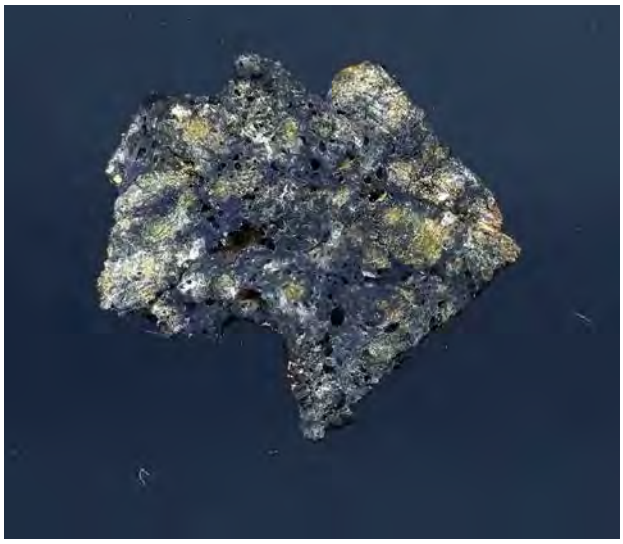
The writer (NMS) believes he has put his finger on the answer to the water quandry. During my first year at Lawrence Livermore Laboratory, I studied the glasses found in the base of the cavity produced by the world's first underground nuclear explosion, codenamed "Rainier", which took place in 1957 in volcanic tuff (containing up to 17% water) at the Nevada Test Site. Three types of glass are present: 1) pumicelike, whitish and vesicular, with moderate water content; 2) obsidianlike, black, dense glass, with less water, and 3) a pinkish dense glass totally devoid of detectable water, found injected into cracks opened up by the explosion. The first two types are shown in this view of the crater base:



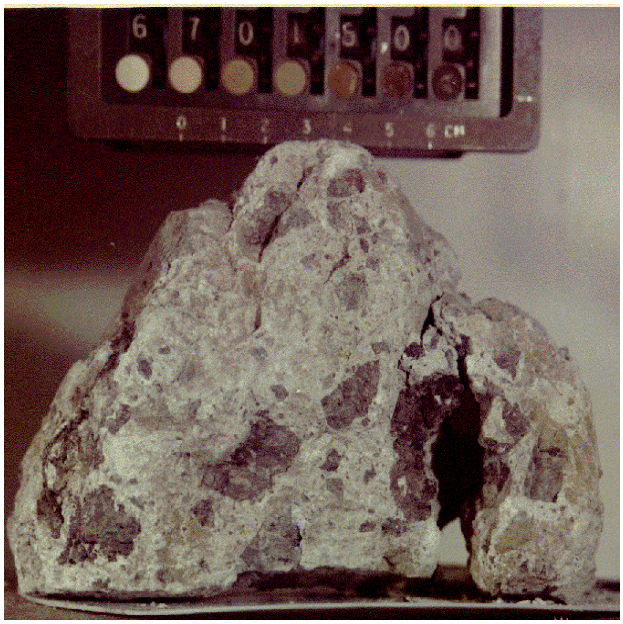
The third type is most like tektites (except for color, an orangy-pink caused by finely dispersed copper from the wiring in the nuclear device). The most potent technique for water detection failed to find any in this glass which was derived from the tuff nearest the nuclear device and hence subject to the highest pressure. My conclusion was that at very high temperatures (at which some vaporization may occur) the shock-induced melting also completely removes any initial water by some separation process yet unspecified. This observation would seem to extrapolate to tektites, as I have proposed. Despite this evidence, Dr. O'Keefe and his supporter, Dr. Harold Urey, never gave up their lunar origin hypothesis for tektites.

While tektites are almost certainly products of a terrestrial shock event, many extraterrestrial

meteorites show some of the shock features discussed on this page. Some of these meteorites are breccias; others are single rock types. Since it is generally believed that at least some meteorites come from the Moon and Mars, those so postulated were probably pushed off their parent body by impact events. Here are two meteorites that are brecciated or otherwise show some signs of shock metamorphism.



As you will see in Section 19, which considers the planets in the Solar System, rocks from the Moon collected during the Apollo missions include breccias that were lithified (some by shock?) or are mare basalts that were shocked during impacts. Here are two examples from the Apollo 16 Cayley site:



Shock effects in lunar samples, which the writer studied as an Apollo investigator, are shown on page 19-6.

We have now presented convincing evidence that impact craters exist and can be properly identified by their characteristic shock effects. We will close this Section with examples of impact craters that have been imaged by remote sensing methods; several were discovered using remote sensing imagery.

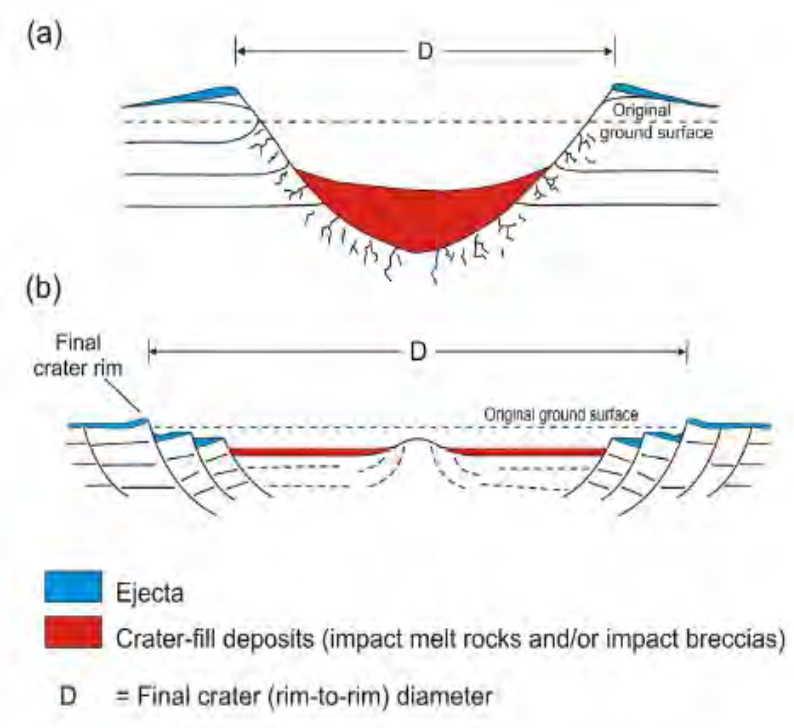
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Impact craters almost always start out as circular structures bounded by a raised rim and bottomed by a depression which may have a central uplift or peak (exception to roundness is the elliptical form that occur when a crater strikes at a very low incidence angle). As crater diameter increases, the ratio of depth to diameter decreases. Crater morphology is altered with time as erosion (mainly by water on Earth and by repeated subsequent impacts and buried by ejecta on the Moon) tends to subdue its topographic expression. As craters wear down to scars (astroblemes) in the bedrock, their initial circularity may still have an effect on drainage. Buried craters are sometimes identifiable by their patterns in seismic or gravity surveys. In situ features of craters include shatter cones and breccias. This page also describes several very large craters whose after effects may have altered conditions for life on Earth.

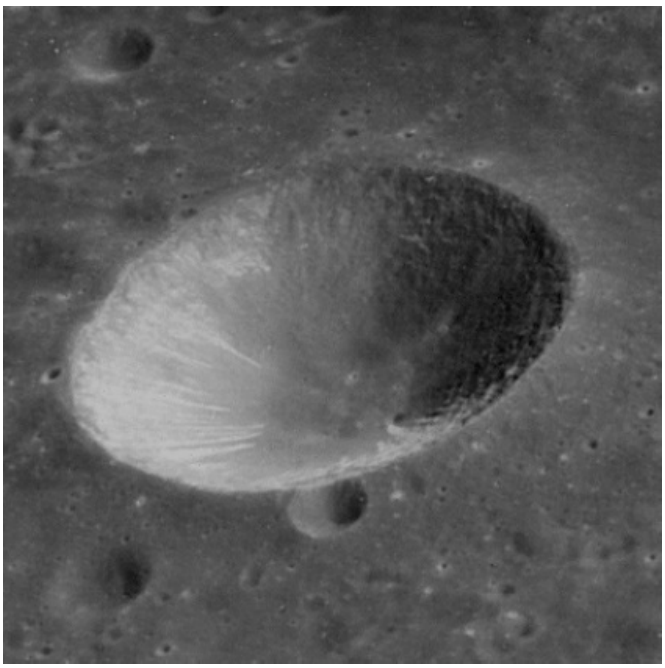
Crater Morphology; Some Characteristic Impact Structures; Impacts as causes of mass extinction of life

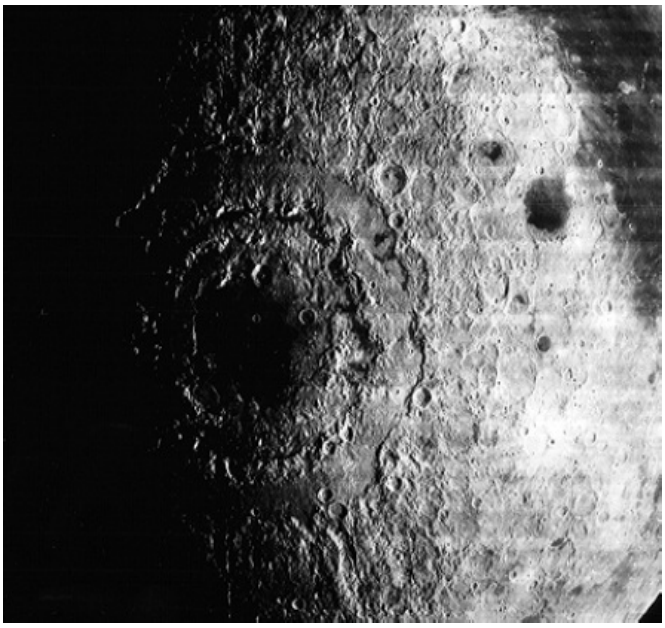
Craters come in three shape/size categories: 1) Simple; 2) Complex; 3) Basins. The difference between the first two is evident in this diagram (Complex craters have a central peak and terraces along their walls, resulting from slumping along concentric fractures.)



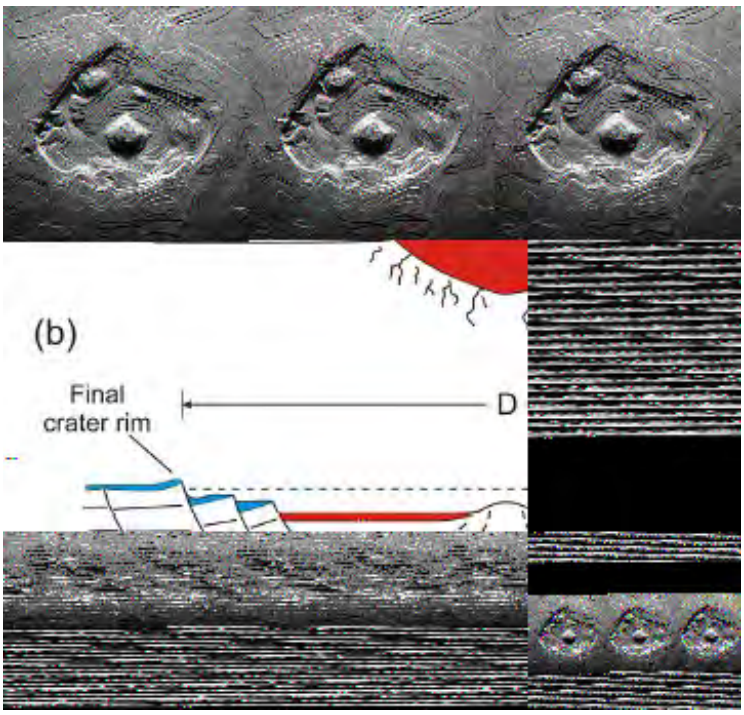
Generally, simple craters are less than 20 km in diameter and complex ones are wider in diameter. Basins, which are several hundred kilometers or more in diameter, may also have central peaks but these can be completely submerged because these big craters can induce widespread crustal or mantle melting that fills them up. In the upper diagram, the top of the red area (the lens of the fall back and slump breccia fill) is known as the surface of the **apparent crater** (what one sees looking into the crater. i.e., the top of the ejecta and/or some solidified melt [caused by shock and/or by subsequent invasion by activated lava from below]); the bottom of the red area is the boundary of the **true crater** (the limit of excavation by the cratering process, below which the country rock is intact although fractured).

We will use three lunar impact structures to illustrate the three types - Simple, on top; Complex, in middle; Basin, at bottom:

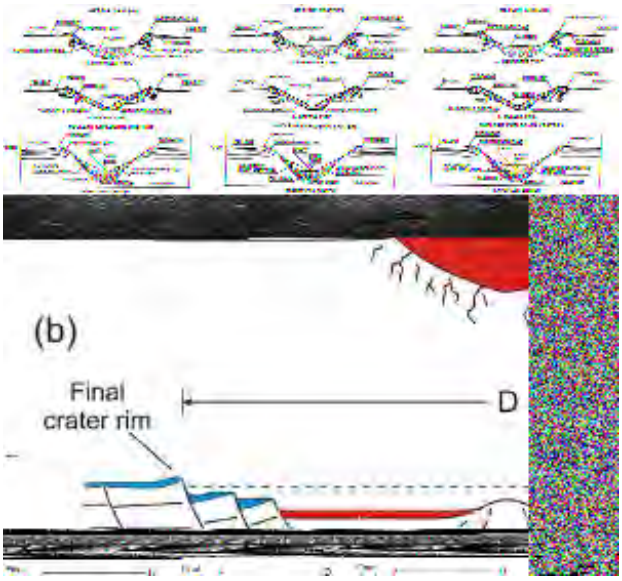




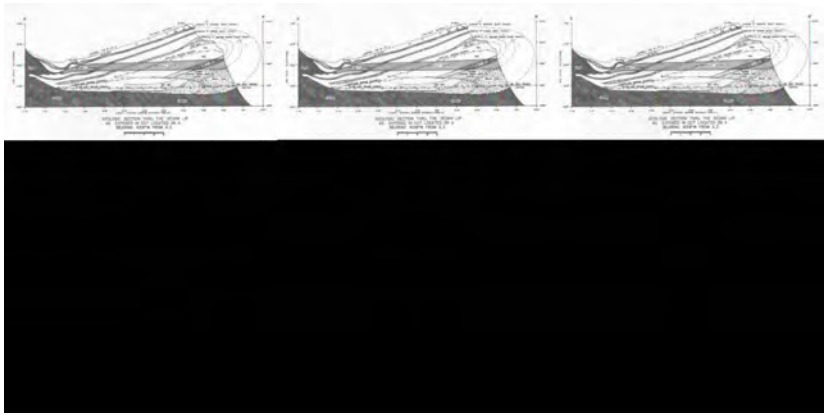
When craters are exposed at the surface, the younger, usually less eroded ones are recognized by their morphology or external form. They are approximately circular (unless later distorted by regional deformation), have raised rims, show structural displacements in their wall rocks, and may have a central peak, consisting of rocks raised from deep original positions. We can emphasize the morphology of these craters in 3-D perspectives (commonly using Digital Elevation Map data) of their contours, exaggerating the elevations and applying shading or artificial illumination (computer-controlled). An example of how this makes craters more obvious, when today they often have low relief, is the Flynn Creek structure (3.5 km [2.2 mi] wide) in Tennessee:



The structural deformation in the affected rock immediately beyond the crater boundary is usually intense and distinctive. Initially flat-lying layers of sedimentary rocks near the surface, beyond the rim, are commonly deformed by upward bending (layers inclined downward [dip] away from the crater walls). Anticlines may be formed or in the extreme the layers are completely overturned producing a flap in which the top layers are upside down, flipped over on top of layers farther out. Modes of layer deformation at two impact craters and one nuclear explosion crater are shown in this diagram (the term "authigenic breccia" refers to fragments (clasts) formed in place with little or no transport; "allogenic breccia" refers to fragments that were broken up elsewhere and then transported to their present sites.

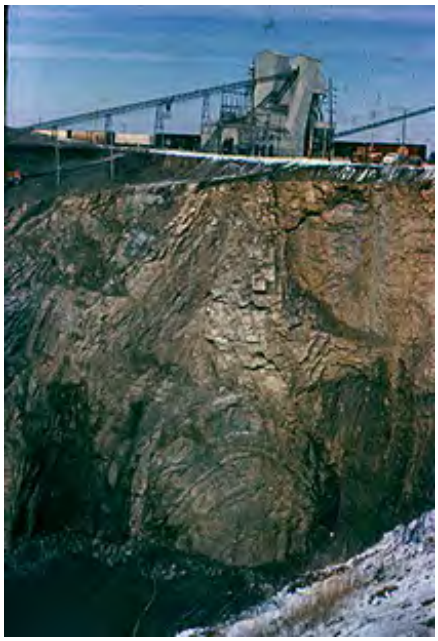


The term "overturned flap" warrants an example. Mapping at the Sedan nuclear crater was particularly revealing about this unusual deformation. Look at this cross-section of the area just beyond the crater:



There were several distinct layers in the Sedan alluvium that could be used as markers. The cross-section reveals that as the crater deformation proceeded, these layers were completely folded back on themselves (overturned) for nearly 100 meters. When the writer (NMS) visited the mapped area, he was shown where an asphalt road leading up to the pre-detonation drill hole that emplaced the nuclear device was flipped over onto itself. Amazing!

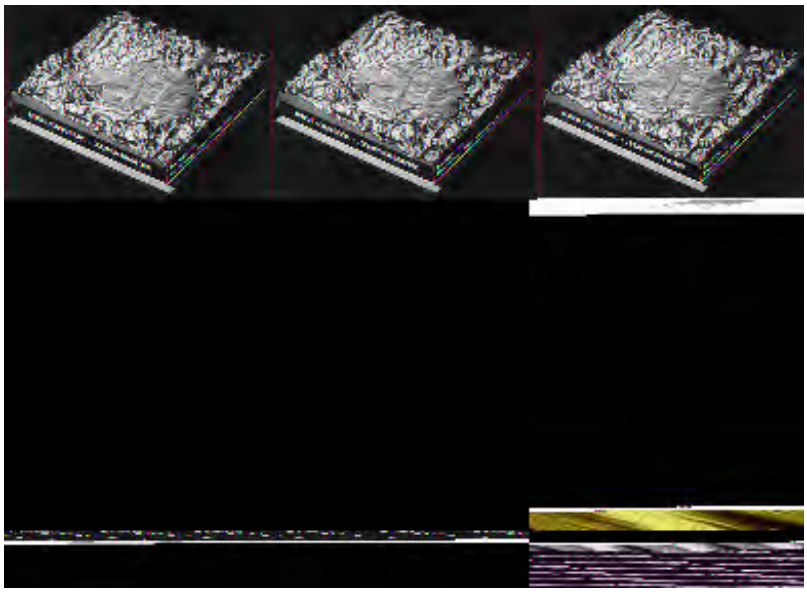
Deformation involving bending and overturning is well exposed in the layered limestones exposed by quarrying at the Kentland, Indiana impact structure:



This structure involves deformation of an area up to 13 km wide in a region where all other rocks are still horizontal. Shatter cones, coesite, and other shock feature provide the proof that this is the remnant of an impact crater.

One of the most famous, and best studied, large complex craters is the the 24 km (15 mi) wide Ries Kessel (also referred as the Nordlinger Ries or just plain Ries) in Bavaria. (An Internet site [in German] that provides more information and field photos is sponsored by the Ries Museum of Nordlingen). Here is a photo montage (made by piecing together several wide-angle lens photos) of part of this structure. (On one of its rim units, thick largely evergreen forests develop, whose dark appearance helps to outline the structure; the occurrence of clouds over this unit appears to result from local evapotranspiration from trees.)

And here is a DEM reconstruction of its generalized subsurface structure:



18-6: In the Ries Kessel perspective view, the crater appears surrounded by mountains. But in reality, the actual landscape is hilly but not mountainous. Explain the illusion.
ANSWER

In satellite images, the Ries structure is not easy to spot. Its interior depression has been backfilled and its rim is now notably eroded. But analysis of its topography using elevation data extracted from ASTER data on Terra brought out the still surviving circularity of the crater, as seen in this bottom view (the top view is an ASTER pseudo-natural color image in which the circularity is obscured but to some degree hinted at):

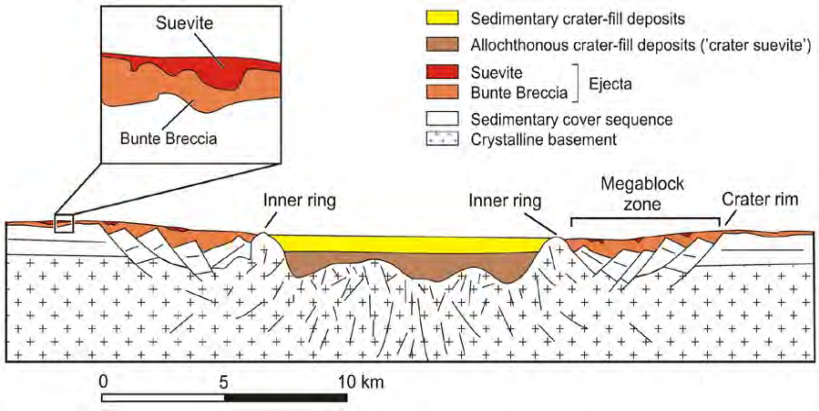
Simulated Natural Color

Elevation

Elevation (m)

300 500 700

The geologic nature of the present, somewhat eroded Ries structure is encapsulized in this cross-section



The Ries is young enough for much of the ejecta that deposited in thick units (when consolidated the general term "breccias" applies; at the Ries the special name "Suevite" is given to this rock) to still be preserved. Here is some field outcrops:

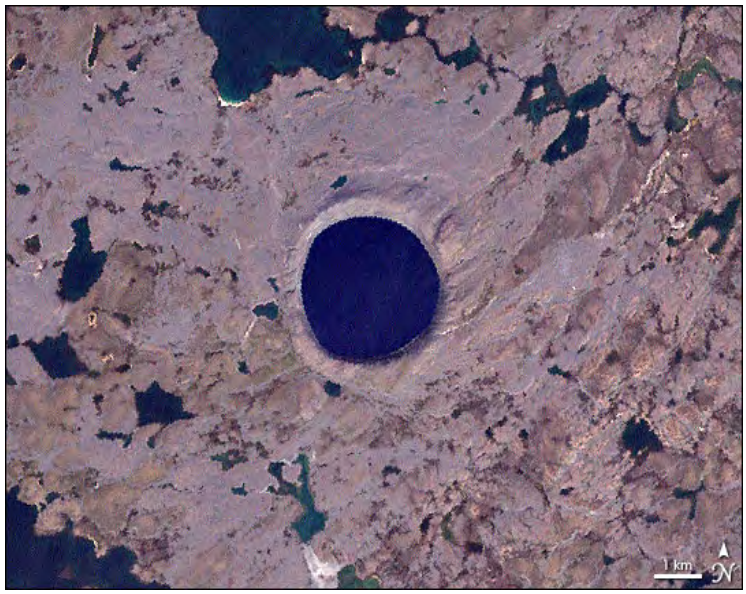
The Ries lies astride "Das Romantische Weg" - The Romantic Way - made up of towns and cities that have preserved much of their medieval buildings. Within the Ries is a remarkable small town, Nordlingen, surrounded by a protective wall. Here is an aerial view of this marvelous throw-back to another era:

Since medieval times, the local residents in Nordlingen quarried some of the breccia deposits that had hardened into rock. This was used as building stone. The Catholic Church near the center of this walled city is made up of this Suevite rock; unfortunately, the rock is easily weathered (because it contains much glass that is unstable over time), so that the Church today is in constant need of repair. Here is this Church:

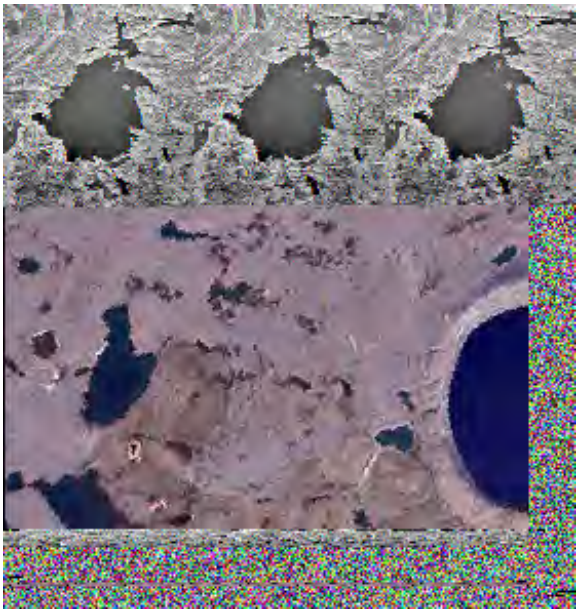
Some of the ejecta "clasts" at the Ries are as big as a house. Megabreccias, similar to those found on the lunar highlands (Section 19), are not uncommon at terrestrial impact structures. A striking example is exposed along a cliff next to a lake inside the Popigai impact crater in Siberia:

Courtesy: Phillipe Claesp

Simple craters (and some larger ones) often have depressions that fill with water. On the top below is an aerial view of the 3.5 km (2.2 mi) wide New Quebec crater (renamed Pinqualuit crater) in granitic shield rock, exposed in Northern Quebec; at the bottom is a view from space.

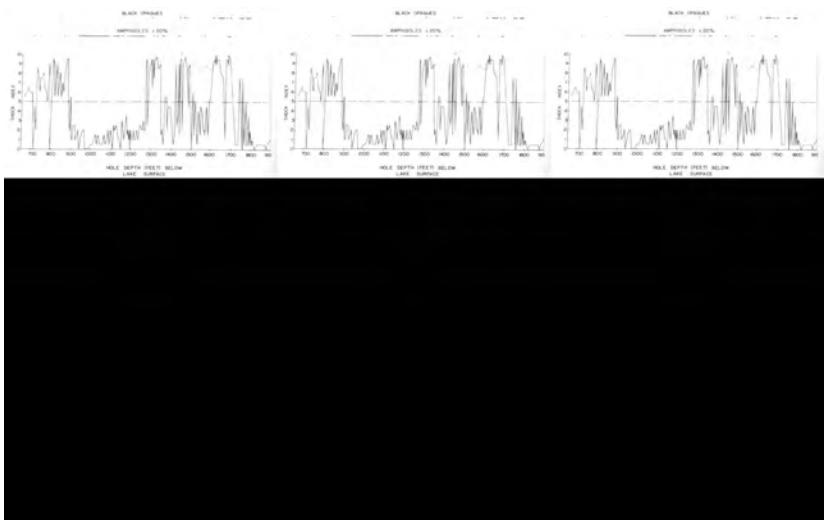


The West Hawk Lake structure (2.5 km [1.6 mi] diameter) formed in metamorphic rocks in westernmost Ontario near the line with Manitoba was the first impact crater studied in detail by the writer [NMS], in 1965; the details were published in 1966 in the Bulletin of the Geological Society of America.



In Canada, and other northern latitude countries, lakes filling impact structures, such as at West

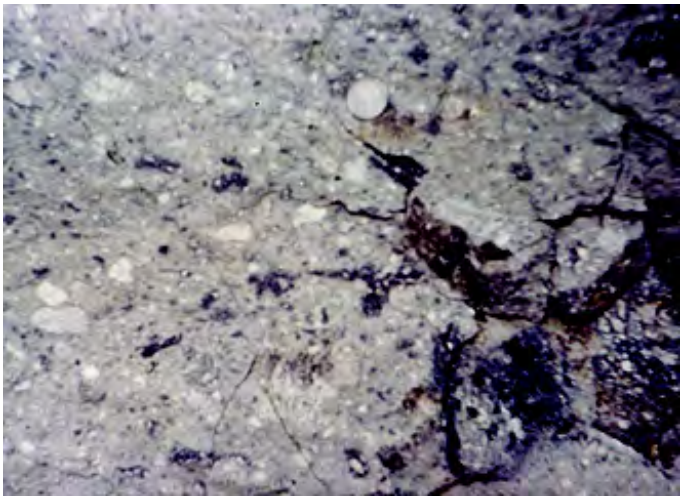
Hawk Lake, freeze in winter, allowing support for drill rigs, so that scientists can explore the crater infill materials by recovering core. One observation from this study was that the distribution of shock effects (mainly planar features and the degree of isotropization) varied rather nonsystematically within the breccia deposits. To quantify this, the writer developed the concept of a "shock log" which plots variations in shock damage as a function of depth; the shock metamorphic features (see next page) used to determine the level of shock damage are given in the key below the plot:



A spectacular complex crater is Manicouagan, a 100 km (62 mile) structure in southern Quebec, Canada, which has a great central peak area of igneous and metamorphic rocks, among which are feldspar-rich rocks. In these rocks, much of the feldspar has transformed by shock into glass, known as Maskelynite. Similar to Manson, there is a depression or moat between the peak and rim (now eroded) that formed annular valleys, which filled with water when a hydroelectric power dam blocked the draining rivers. Because of this contrasting surface expression, astronauts journeying back from the Moon could see this crater from well out in space.

A great deal of impact-produced melt is found along the lake at Manicouagan, as shown here:

Deposits of fragmental rock surround most younger craters. An example (top, below) of such rock , from an outcrop at the Ries crater, illustrates these ejecta deposits (Suevite breccias). A second example (bottom) seen in core from a drilling that penetrated the Manson central peak, shows the diverse nature of the rock types making up these breccia fragments (called clasts).

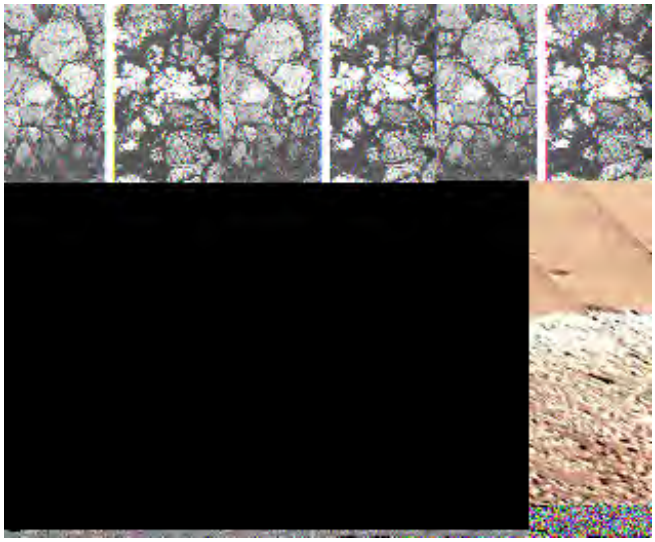


18-7:Suppose a continuous length of drill core consists of first an interval of breccia much like that shown in this figures, then a 10 meter interval of a single rock type, say granite, and followed by more small fragmental breccia. What explains this? ANSWER

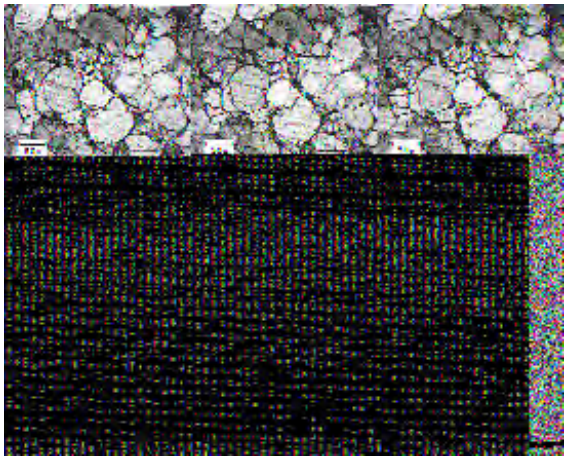
Most ejecta blocks found around younger craters consist of fragmented bedrock derived from subsurface units. There can be exceptions if the surface material is unconsolidated. The writer (NMS) discovered a fabulous example of this, which at first was discounted by other specialists in this field. The crater is the small Wabar structure (there are 3 craters - 2 are small - there). in the sandy desert of southern Saudi Arabia. Around the rim are small fragments of white quartz sand, many coated with a black glass. Here is two views:

A few years earlier the writer had been given small pieces of "sandstone" around chemical explosion craters formed during an experimental program at the Nevada Test Site (NTS), where

white loose quartz sand had been used to backfill the access hole through which the explosives were loaded. He postulated that the fragments were made up of this sand that had been driven together and compressed (a process he named "shock lithification", calling the fragments "instant rock"). He proposed the same origin for the Wabar fragments, namely, that they were desert sand shock-liithified by shock waves from the impact (and many were then covered by shock melt that overtook them). This pair of photomicrographs shows the texture of the NTS instant sandstone on the left and the Wabar lithified fragments on the right.



Below is a second photomicrograph of the NTS instant sandstone, showing more details.

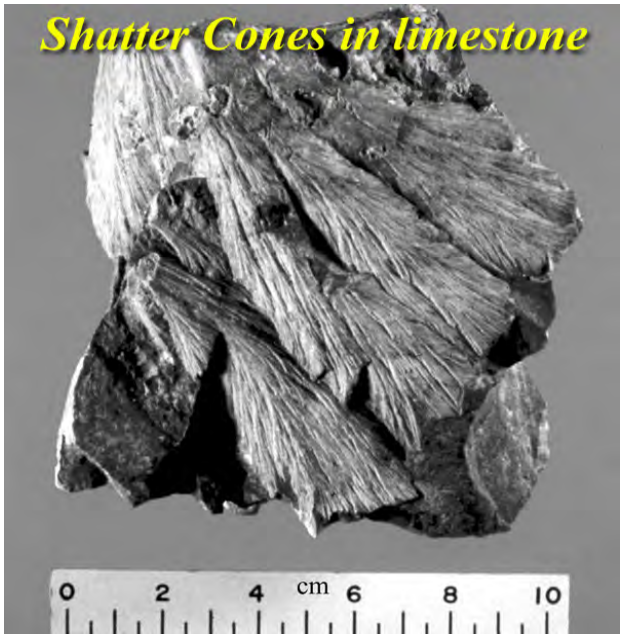


The paper on this interpretation was rejected by *Science* Magazine because the reviewer had been there and thought he had noted thin sandstone layers in the rim. Through a stroke of luck, the writer, telling a colleague at Shell Oil in Houston of the discrepancy in interpretation, was surprised to receive a call later from that friend who reported the loose sand at Wabar was more than 200 meters thick (he had asked a Shell field geophysical crew to run a seismic line next to the site; they determined an accurate thickness). With this new "proof" the paper was resubmitted to Science and was published.

Wabar not only has abundant examples of shock-lithified "sandstone", but it also has numerous samples of shock-melted sand as black glass:

The writer obtained several Wabar samples which were "most peculiar". These consisted of the instant rock but with a thin coating on the exterior of this black glass (the bottom row above may show examples). Odd indeed! My speculation: the instant rock fragments were expelled as ejecta that passed through a "cloud" of the melted sand that made up the black glass. This is speculation but seems reasonable - I can't think of a better explanation. Strange things happen during impact events.

Eroded craters lack definitive external shapes, although the initial circularity may have a persistent effect on drainage, keeping streams in roughly circular courses. Such craters are often hard to detect but the presence of anomalous structural deformation and of brecciated rocks give clues. In rocks that were just outside the original wall boundaries (the true crater), a peculiar configuration, known as shatter cones, commonly develops.



These "striated" conical structures (described as "horsetail"-like in shape) can be very small or can reach six feet or more in length, as seen above in quartzites at the Sudbury, Canada, impact structure. When the folded rocks containing the cones are restored to their original positions in an orientation graph, the cone apices invariably point toward an interior location that lies above the central crater floor. In effect, this denotes that the position where the energy was released was above the floor, a situation incompatible with a deep volcanic source, as once advocated by skeptics. The cones, which also sometimes form in rocks subjected to nuclear explosions, occur in lower (peripheral) shock pressure zones, as the rarefaction phase of the shock waves, spreading outward, places the rock into tensional stress. Many cones appear to originate from point discontinuities (e.g., a pebble) as though the waves were diffracted.

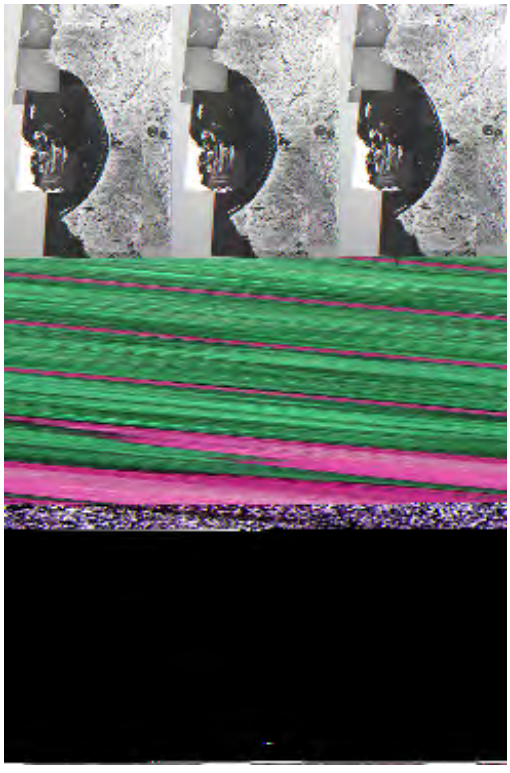
Shatter cones come in a range of sizes from a few centimeters to the 9 m cone at the Slate Island (Canada) impact structure, shown here:

18-8: Try to explain what happens to cause the apex of a shatter cone to point towards the upper center of the crater near the point of impact. ANSWER

Still, the best evidence for a extraterrestrial origin of a crater is the survival of the incoming bolide as pieces of meteorites or asteroid/cometary material. This is relatively rare, although abnormal chemistry (such as iridium and other unusual concentrations of trace elements) in rocks and melt from older structures often can point to the intermixing of the bolide with the target. Iron meteorites are found in and around Meteor Crater, Arizona (see page 18-5). Iron meteorites are present in the small, relatively recent Campo del Cielo craters in South America and also at the

Wabar crater discussed above. Eight small craters formed less than 10000 years ago in Poland near Poznan, with the largest being about 100 meters wide, were identified quite by accident as caused by large fragments of an iron meteorite, pieces of which were discovered by troops digging fortifications in World War I. The depressions are well-preserved, as seen in this photo:

A word of caution: Lest one assume that every circular structure is impact in origin (we've already pointed out circular volcanic craters), here is a case where circularity on a grand scale does not mean a great impact event occurred. Consider the Nastapoka Arc on the eastern shore of Hudson's Bay in Canada, as seen in this Landsat mosaic:



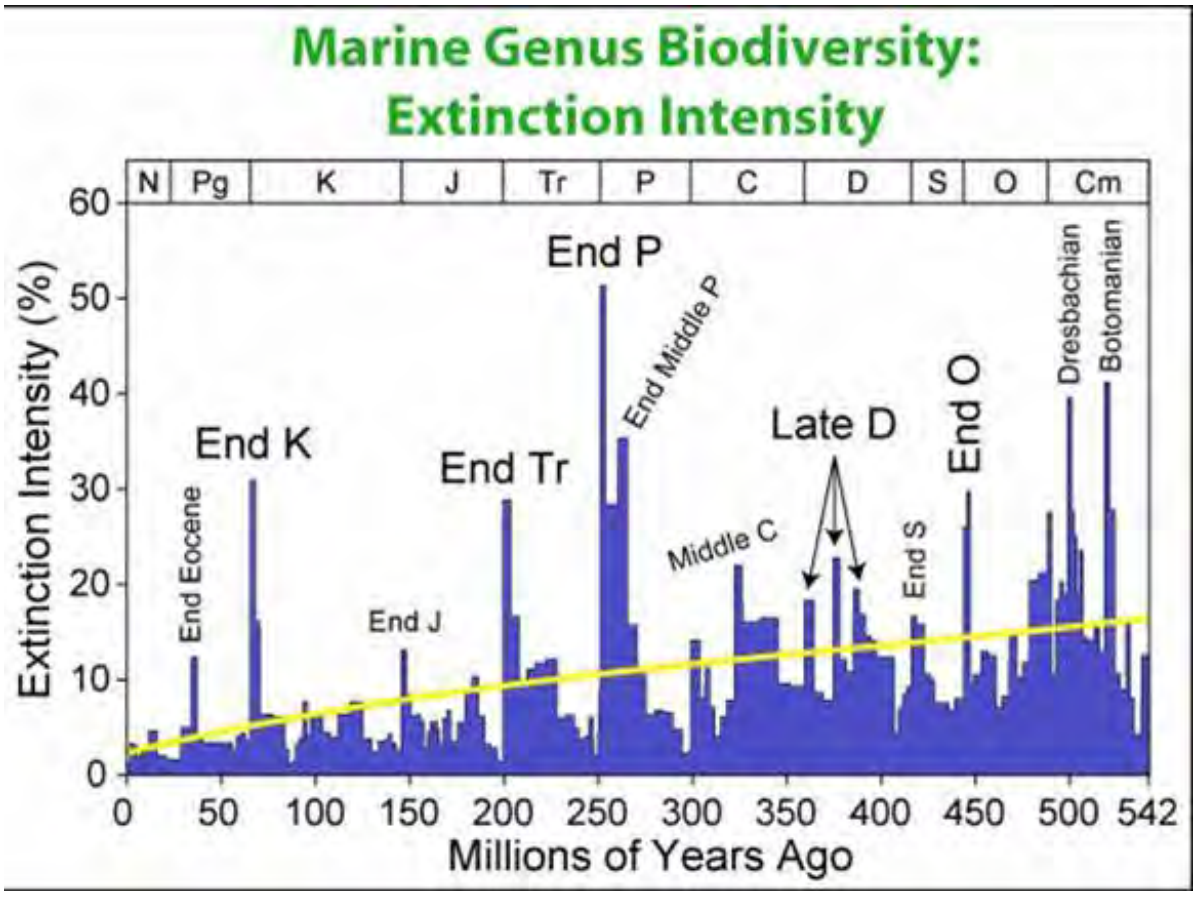
Its circularity is imposing. Many hoped this was the largest impact structure on Earth. Possibly it

may someday prove just that. But all the field evidence so far has not yield a single positive indicator of impact. Granted the surface rocks include some younger sedimentary cover. But no shatter cones have been found. The Belcher islands to the west show no shock features, expected if they were part of a central uplift. Limited deep drilling has not encountered shocked rocks. The circularity may simply be a fortuitous configuration of a sedimentary basin. But note the two circular features, side by side, to the east. These are the Clearwater Lakes impact structures, shown on the next page.

The Nastapoca Arc ambiguity calls attention to the fact that there are still other circular structures whose origins - impact, volcanic, tectonic - have been questioned. A case in point is the Upheaval Dome in Utah, seen here in this annotated astronaut photograph; for a long while this dome was considered to be the surface manifestation of some type of intrusion (perhaps a salt diapir):

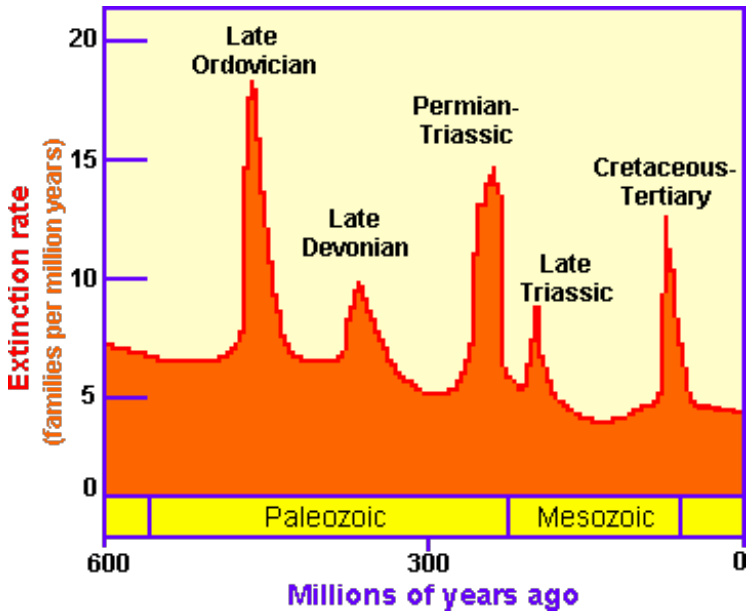
It was this structure that was studied by the Dean of Astrogeologists - Dr. Eugene Shoemaker - when he first worked for the USGS. He concluded that it could (?) be impact in origin. This aroused his interest in impact mechanics that eventually led him into planetary studies. For decades, the consensus disputed this origin. But recently several diagnostic shock features at the microscopic level have been discovered in its rocks, tipping the opinion balance in favor of a meteoritic collision.

Now on to a very relevant subject: Can impact craters, which represent huge releases of explosive energy, affect and even threaten life on Earth? Indeed, large impact craters - big enough to produce "basins" - are, of themselves, interesting to search for and study. But they also have important implications for the history of life on Earth. To determine whether impact plays such a role we start first with a quick overview of what is known about mass extinctions since life in its macroscopic form (large plants and animals) became abundant at the beginning of the Paleozoic. Consider this diagram:



The diagram indicates that there have been many small extinction events during this time span, and a few larger ones. One of these is the famed demise of the dinosaurs at the Cretaceous-Tertiary boundary. Many geologists now believe that the biggest extinction of all ended the Paleozoic at the close of the Permian. Other mass extinctions affected large fractions of animal and plant life at the end of the Ordovician, the Devonian, the Triassic, and possibly the Paleocene.

These extinctions are best defined at the family level, as shown in the graph below, and the accompanying explanation beneath it:



* The late Ordovician period (about 438 million years ago) - 100 families extinct - more than half of the bryozoan and brachiopod species extinct.

* The late Devonian (about 360 mya) - 30% of animal families extinct.

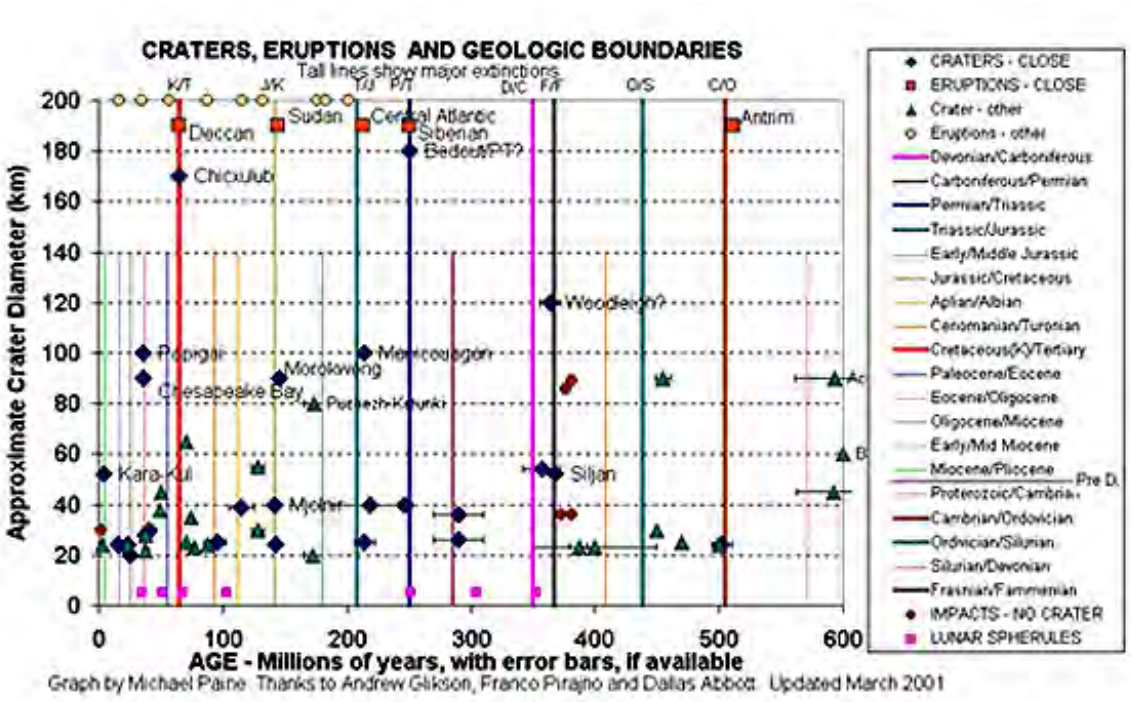
* At the end of the Permian period (about 245 mya) - Trilobites go extinct. 50% of all animal families, 95% of all marine species, and many trees die out.

* The late Triassic (208 mya) - 35% of all animal families die out. Most early dinosaur families went extinct, and most synapsids died out (except for the mammals).

* At the Cretaceous-Tertiary (K-T) boundary (about 65 mya) - about half of all life forms died out, including the dinosaurs , pterosaurs, plesiosaurs, mosasaurs, ammonites, many families of fishes, clams, snails, sponges, sea urchins and many others.

However, the above diagram tends to obscure the fact that extinctions are very common in the geologic record. At various times, the extent of extinction has been less than those associated with "mass extinctions" but in notable numbers nevertheless. Some researchers attribute at least one or more of these destructive occurrences to impact - although other mechanisms have been proposed (see bottom of page).

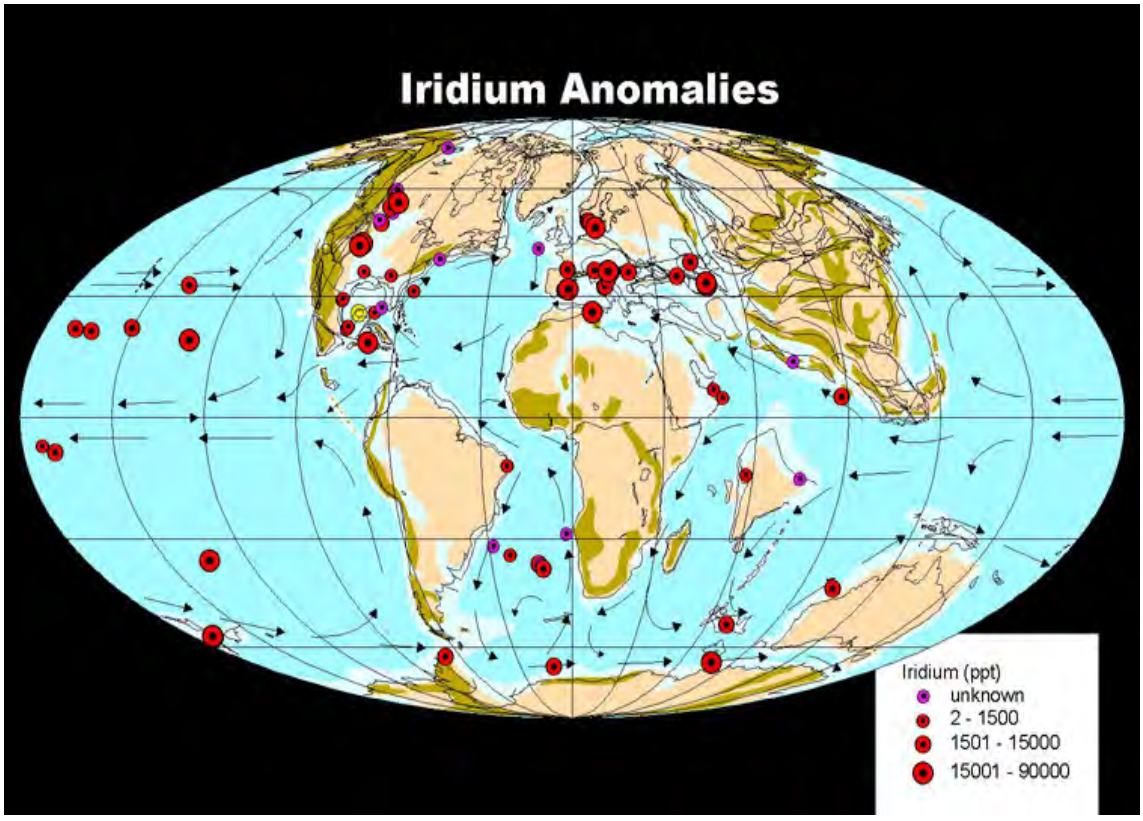
In the context of these known extinction events, can impact be demonstrated as the cause of at least one or more of these? Consider this diagram, which shows most of the extinction events, along with the time estimates of both major impact crater formation and several spans of major volcanic activity, from which you can draw your own conclusion:



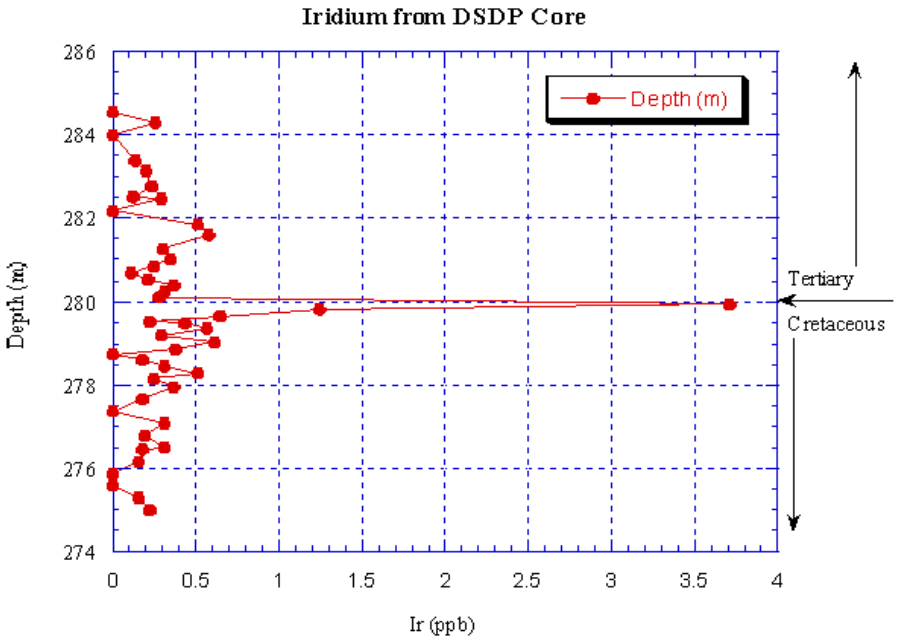
With this background re: mass extinction, let us examine the evidence that includes both impact cratering and volcanism as the most likely causes of the disappearance of life forms on a global scale. The discovery of a highly probable huge impact at the end of the Mesozoic Cretaceous period is now believed to have destroyed worldwide a large percentage of living forms (about 70%, mostly on the lands) including the dinosaurs.

Evidence for an impact catastrophe is best demonstrated by the K/T boundary (time horizon between the Cretaceous [K] and Tertiary [T]). The first important report of unusual features in deposits at the boundary was made by the father and son scientists Luiz and Walter Alvarez, based on studies they made of deposits at Gubbio, Italy and elsewhere:

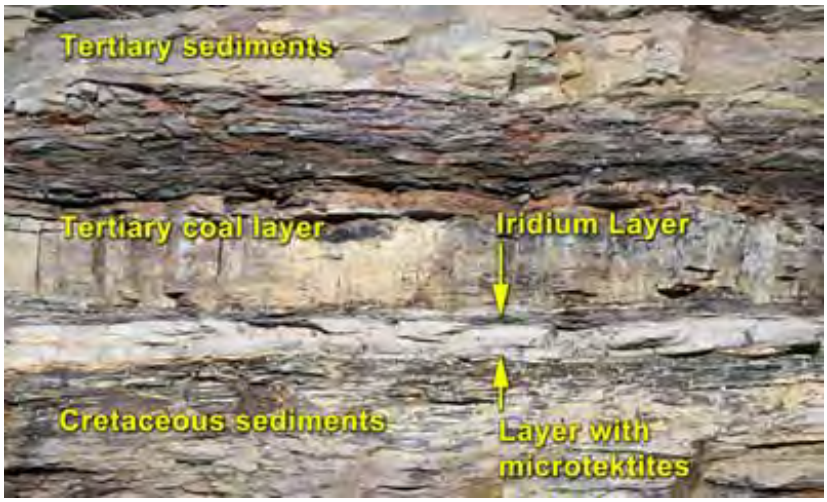
The Alvarez's deduced the nature of these deposits, and their significance, from the presence of Iridium, an element related to Platinum in the Periodic Table. Iridium is rare on Earth, but is found in minute amounts in some basaltic rocks. It is calculated to be much more abundant in the Earth's iron core. Iridium is a notable constituent in meteorites, mainly those composed primarily of iron or with iron inclusions. The Alvarez pair proposed that the Iridium in the K-T boundary layer deposits was part of the fallout of debris injected into the high atmosphere by a huge impact. They, and others, began to hunt for Iridium in deposits considered to be at the K-T boundary as recognized in various parts of the world. Many localities with Ir-enriched K-T layers have been found, as shown in this map:



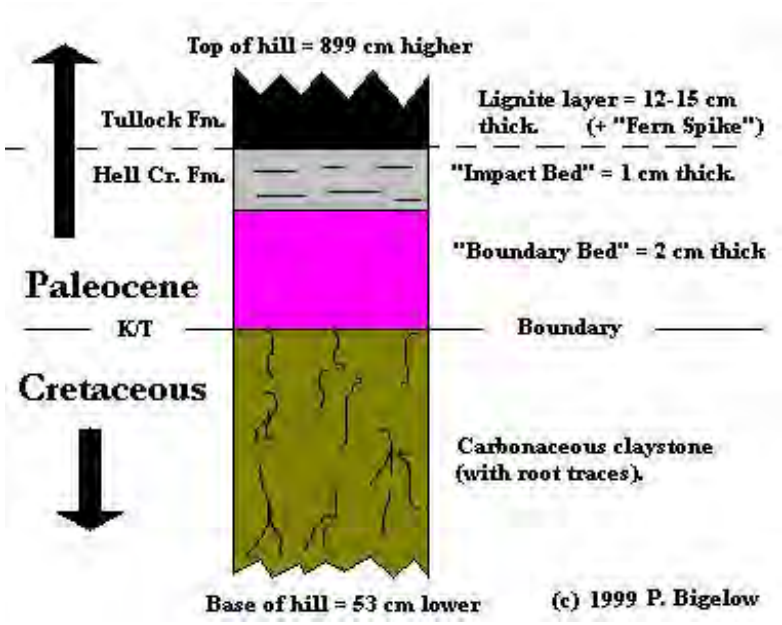
The resulting debris from the reputed impact event was ejected into high altitudes spread around the globe and settled as a thin layer of material that marks the precise K/T boundary between the last rocks of the Cretaceous Period and the first sediments formed in the younger (overlying) Tertiary Period. The deposits contain Iridium, a metallic element related to Platinum present in many iron meteorites, in amounts far greater than can be accounted for by volcanic sources or other terrestrial rocks.



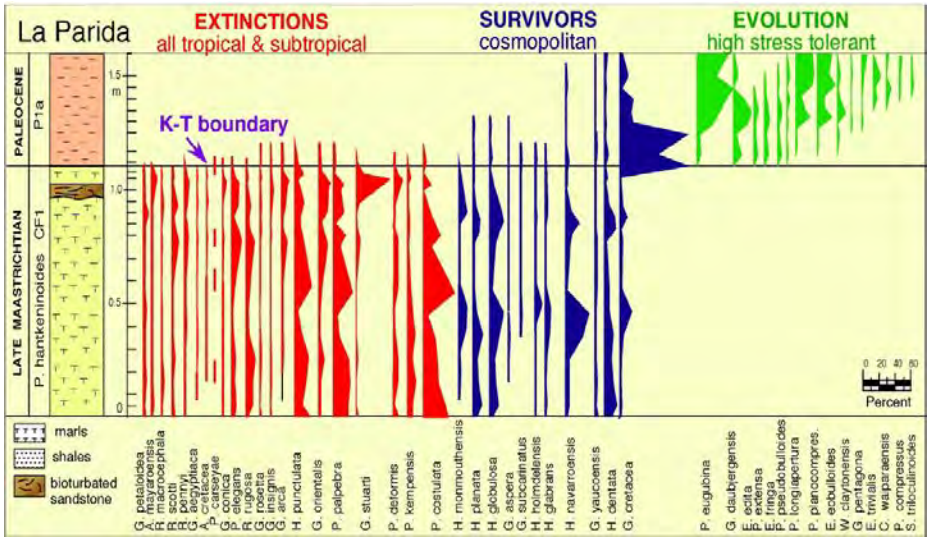
The deposits at the K-T boundary are usually very thin. They represent the fallout layer that may have been worldwide in distribution. Here is two examples of this now-famous layer, which also includes soot particles from the after-impact fires that consumed much of Earth's forests.



The immediate stratigraphy above and below the K-T time plane is quite distinctive. This cross-section for the Hell Creek locality in Montana is typical:

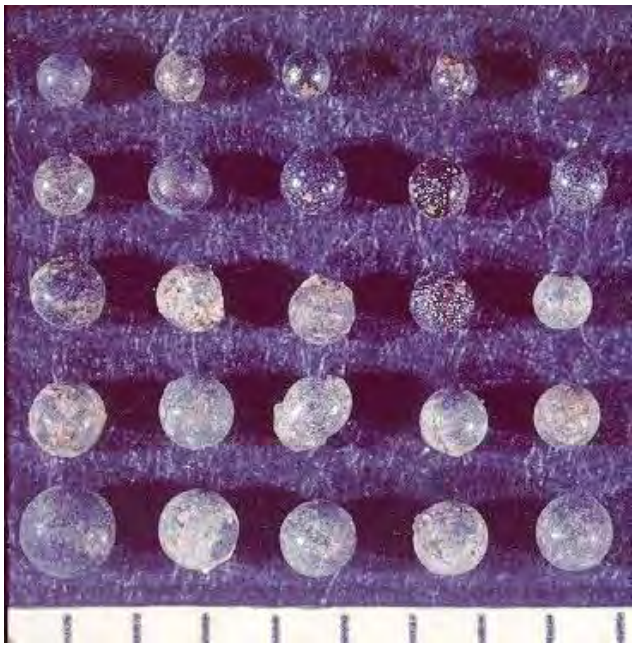


Below the K-T layer, sedimentary rocks are rich in fauna, both macroscopic (such as animals like the dinosaurs) and microscopic (such as plankton and other tiny marine life forms). The Tertiary rocks above the layer, for a few meters at least (representing several million years of deposition time), are missing most of these fauna, although some do carry over from the Cretaceous. This plot shows some of these characteristics:

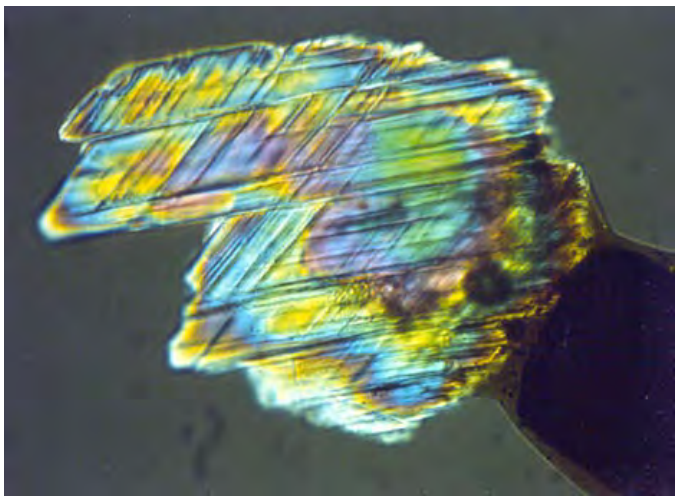


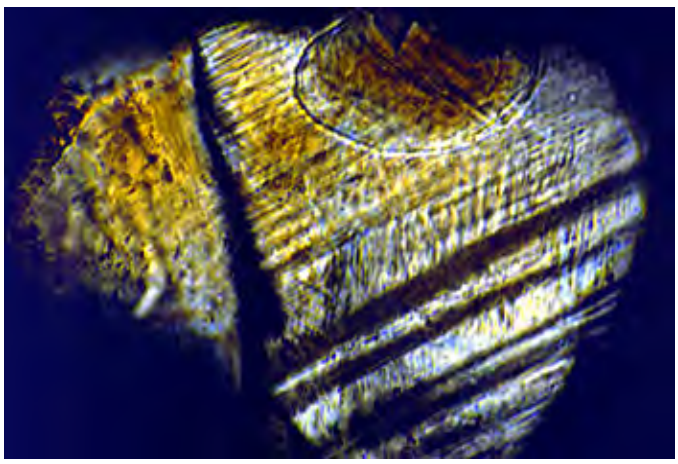
The nature of this abrupt change in faunal speciation and distribution has proved complicated. Suffice to say that some paleontologist dispute the claim of an abrupt disappearance of many fauna just above the boundary. Consult this Wikipedia web site for more details.

Soon after the Alvarez announcement, they and others presented evidence that shock features occur within the clay layers right at the boundary. The deposits usually contain small glass spherules, some similar to the tektites that have been generally accepted as evidence of impact. Here is an example.



Some mineral grains found in the boundary layer bear evidence of intense shock (including quartz crystals with planar features; see page 18-4). Here are two example, the first in quartz and the second in a feldspar grain:





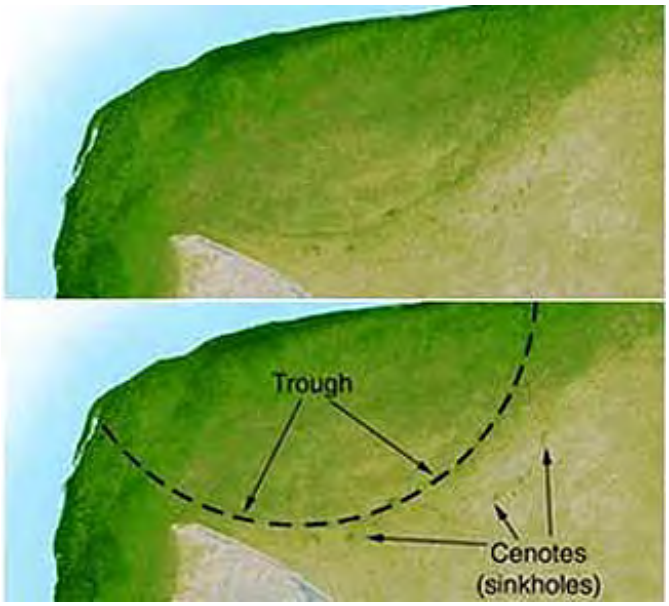
Although Iridium is sometimes spread into sediments by erosion of volcanic rocks that show somewhat larger concentrations, the very large amounts in the K-T layers, together with the shock features just shown (which are never produced from volcanic processes), have proved definitive in the association of the layers with fallout from an impact. The worldwide distribution of the Iridium implies a huge impact event. The question at the time was simply WHERE?

The killer impact site has now been found. In view of the tremendous energies involved, it is no

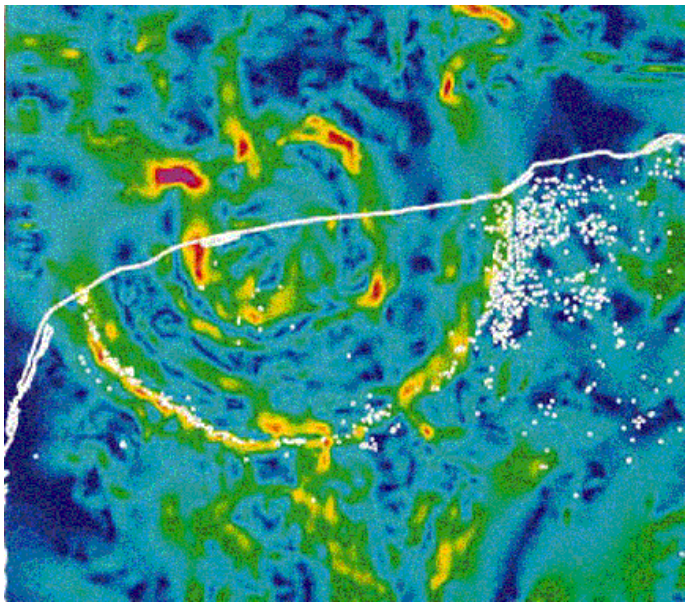
wonder then that we classify the Chicxulub impact in the Yucatan Peninsula as one of the largest short-term natural events known in the geologic record (of nuclear-comparative magnitude in excess of 100 trillion tons of TNT equivalent). It occurred 65 million years ago and led to a 200-300 km (>150 mi) wide (there's still some uncertainty regarding the location of the outer rim) and perhaps 16 km (10 mi) deep depression. Here is its location:

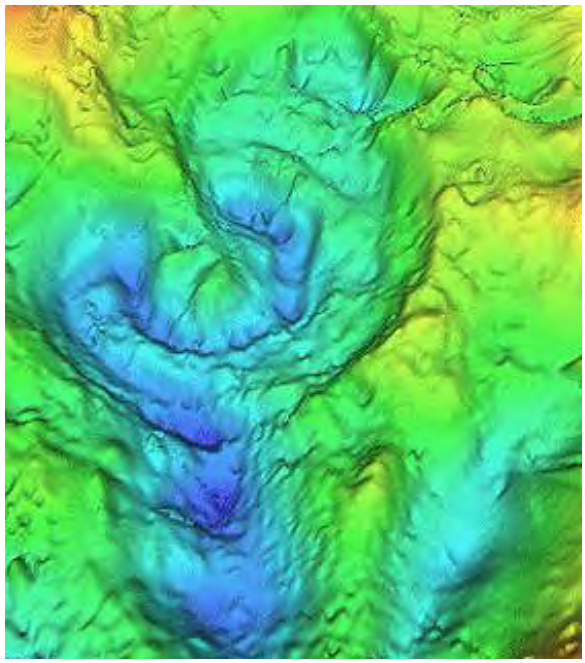
Satellite images of the Yucatan fail to disclose any sign of surface expression of the crater, mainly because the surface is almost completely masked in thick jungle vegetation. However, it does

In case you missed this trace, here are a pair of images derived as an enlargement of the part of the Yucatan containing the Chicxulub crater. On the lower one, the rim boundary has been drawn in and the location of sink holes that seem to relate to subsurface control by the fallback material beyond the rim is marked.

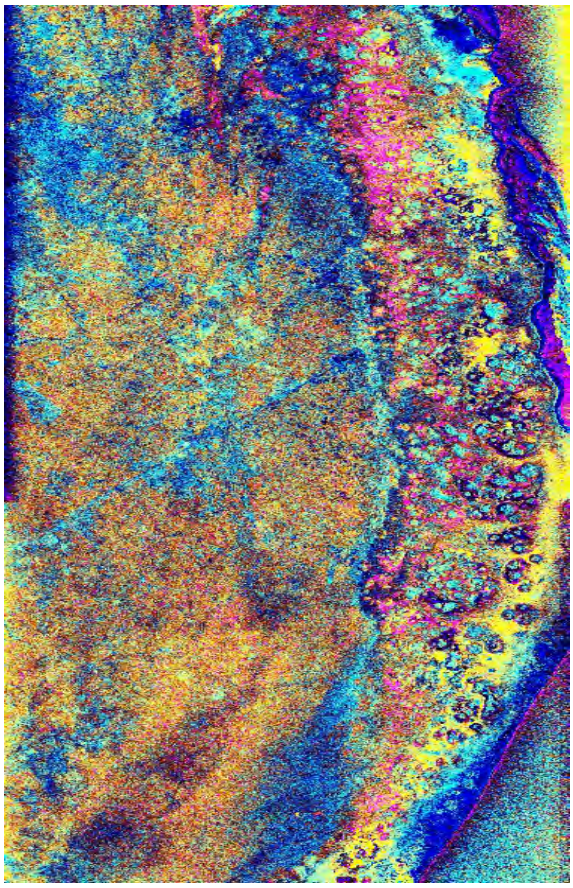


When the effects of vegetation are compensated for, this huge structure has no evident surface expression, being covered by younger sedimentary rocks, but does appear subsurface as a strong gravity anomaly, with a definite circular pattern, as shown below.

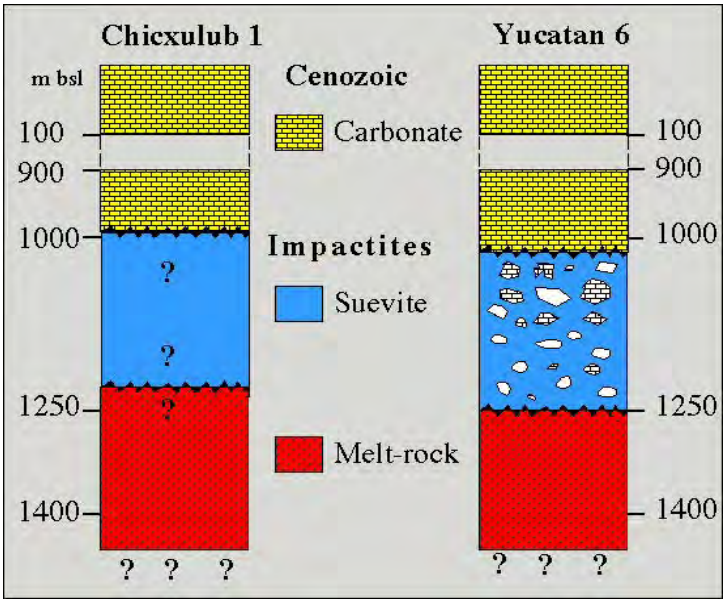




It was thought that Chicxulub could not be recognized in space imagery because of the post-impact sedimentary rocks covering the structure and also because of dense vegetation cover. However, radar data from the SRTM program have been specially processed to bring out otherwise subtle suggestions of the buried crater rim showing its presence by some manifestation at the surface (probably induced by differential subsidence). Here is that SRTM image of much of the Yucatan Peninsula. See if you can locate the rim trace:

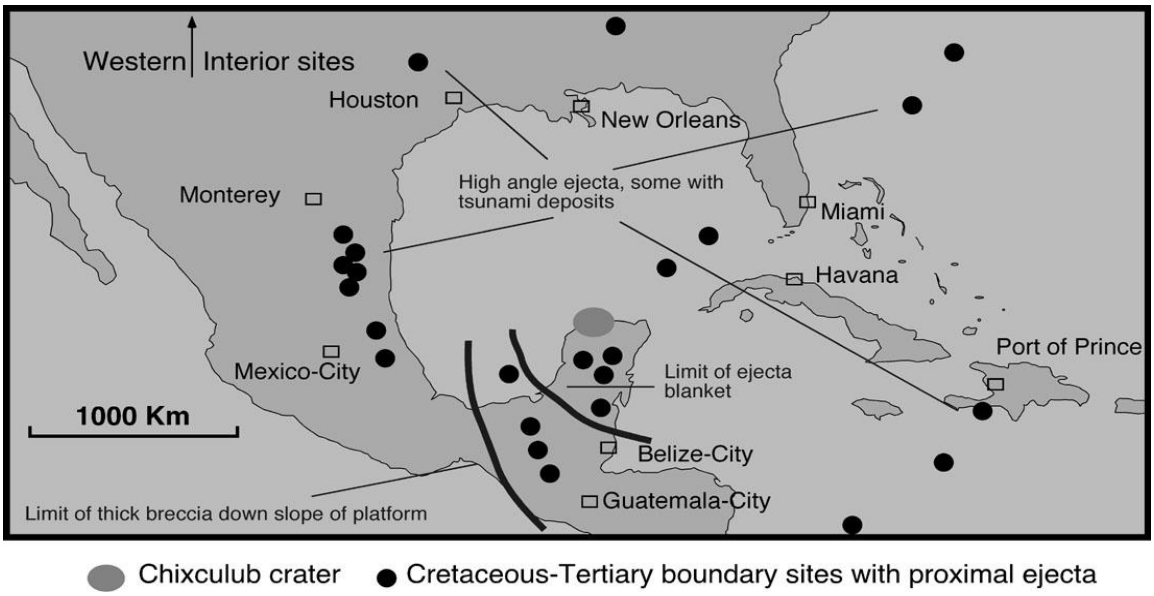


Evidence in the impact-related rocks (bedrock below the crater floor and ejecta within) was then discovered almost incidentally in core samples obtained through earlier exploration drilling for oil. The samples languished for years in the basement of the University of New Orleans' Geology Building, before someone re-examined them and deduced the origin of the anomalous materials beneath the post-impact sedimentary rocks. Intervals with these core samples, containing so-called volcanic rocks (now known to be shock-melted rock), showed distinct shock effects. These two stratigraphic sections, from two drill holes, present this interpretation.



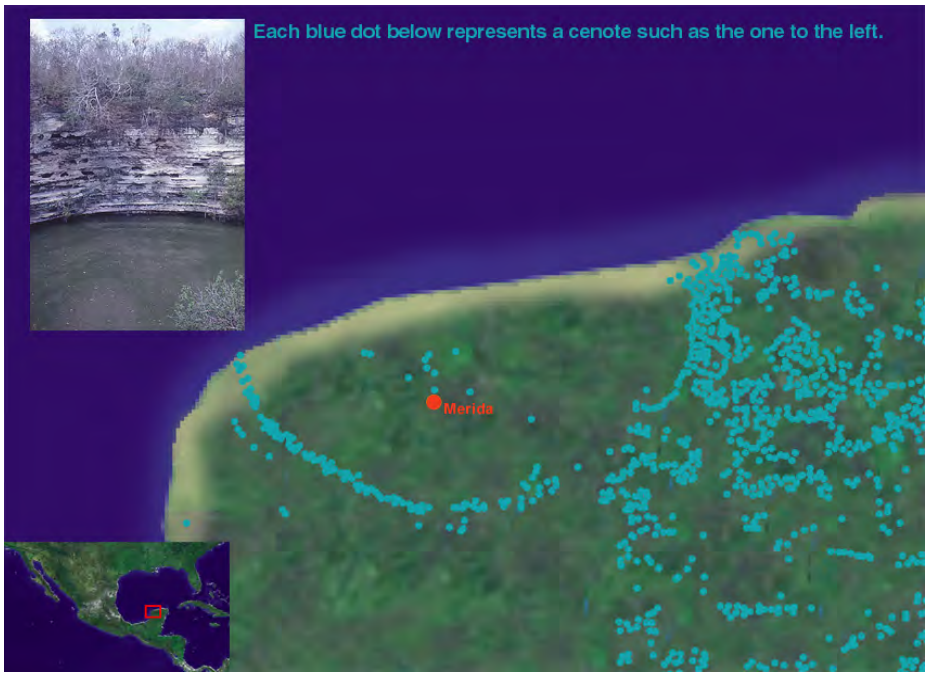
Quenched melt (glass), consistent with an impact event, has been recovered from drill core into Chicxulub; it contains fragments showing other evidence of shock damage:

As interest in Chicxulub grew because of its apparent connection with the K-T boundary extinction, new evidence from its immediate vicinity and from deposits found at various distances in the U.S., Mexico, and Central America helped to confirm the impact hypothesis. This map shows some of the localities where deposits that could be tied to Chicxulub were found:



Ejecta deposits in Belize are definitely at the K-T boundary and show kinship with Chicxulub.

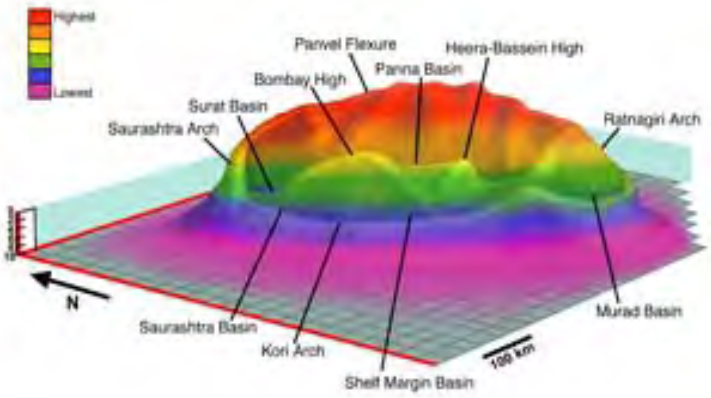
One of the consequences of the crater hypothesis is that the distribution of cenotes (Spanish for "sinkhole") in carbonate rocks could now be explained. This is evident in this map:



Some of the sinkholes were used by the Aztecs as ceremonial sites for killing rituals. This included several cenotes in the semi-circular ring that overlies the crater boundary, which had an effect on locating sinkholes outside the range where most were produced.

To summarize the Chicxulub event in terms of its almost certain relationship with the K-T boundary layers: The Chicxulub impact into shallow waters of the Gulf of Mexico generated huge waves (tsunamis) and, even more destructive to the planet, tossed enormous amounts of hot rock and water/stream into the atmosphere. An immediate result was to set forests and grasslands over much of the globe on fire, in the biggest firestorm in history. These materials, in turn, caused a worldwide "cloud deck" of aerosols, gases (including SO_2) and particulates leading to temperature fluctuations, general darkening, an anoxic (oxygen-poor) atmosphere and reduced photosynthesis that wiped out much of the food chain and provided the "*coup de grace*" to the reduced number of dinosaur families still living then on Earth. Up to 50% of angiosperm (flowering plants) species were destroyed along with many animal families in the sea and on land. Some have estimated that it took thousands to a million or more years for ecosystems to recover. Mammals, inconspicuous before this event, were able to flourish in these restored systems and gradually gain ascendancy during the Cenozoic; this led the way for the eventual appearance of humans. For even more information on the now famous Chicxulub crater, go to this Wikipedia web site.

One could almost predict that once a strong candidate for the impact crater that killed the dinosaurs had been found, others would propose another crater that seems to fit the bill. The Shiva structure off the west coast of India in the Indian Ocean has been offered up as either an alternative to Chicxulub or in the view of some a second crater formed either at the same time or as a separate unrelated event very close in time to the Mexican crater and the K-T event. The structure is huge: some 650 by 400 km in dimension (this asymmetry suggest a glancing or low angle impact). It is still known only from geophysical evidence since it is under water and is topped by younger sedimentary material. But drill core (justified as exploration for oil) has returned brecciated rocks that date close to the K-T event. Its chief advocate is Prof. S. Chatterjee of Texas Tech University, who claims that rocks recovered from the site contain both shock features and high concentrations of Iridium. He notes the association of the event with the Deccan Trap basalt flows and suggests that outpourings of that lava were accelerated by the impact. You can learn more about Shiva crater at this Wikipedia web site.

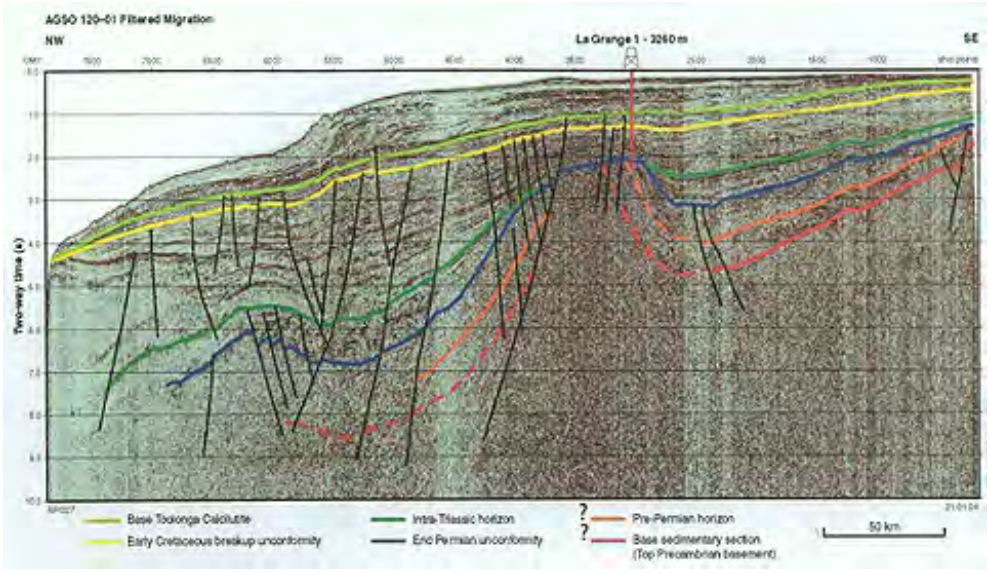


However, many impact advocates have questioned the Chatterjee observations and some, including Christian Koeberl, declare the conclusions to be unproven and possibly erroneous. One problem currently complicating the postulate that Shiva is the main culprit in the K/T extinction is that age dates for it may be different from that for Chicxulub by as much as 30000 years. If this holds up, and Shiva is indeed the largest impact crater on Earth, this would mean that the two biggest such events took place almost (but not quite) simultaneously - a remarkable (but plausible) coincidence.

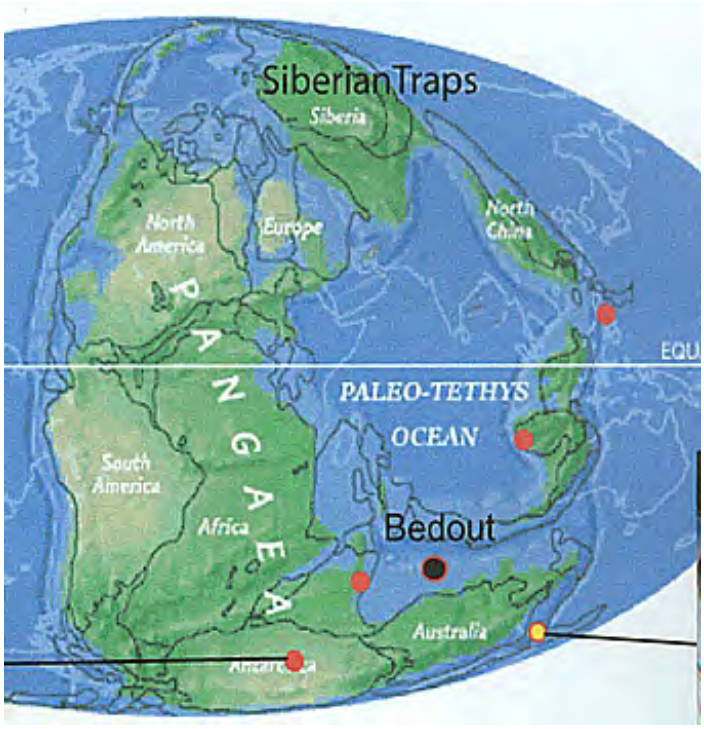
Now to other craters that could trigger extinction consequences. A report in November, 2003 has presented evidence from rock deposits in Europe and China of a very large impact, about 251 million years ago, that coincides with the end of the Paleozoic (Permian-Triassic boundary), a time in which about **90% of marine species living then** are estimated to have vanished - a time now referred to as the "Great Dying". This is the greatest mass extinction in Earth history - it set the stage for a new burst of now different life in the Mesozoic. The hunt for this super crater has been a top priority in the last five years.

Most destruction of life was in the oceans. But reptiles occupying the land also were largely compromised (but not completely wiped out or there would have been no dinosaurs or mammals in the Mesozoic). Among these was Dimetrodon, a reptile with a sail-like dorsal fin, who disappeared completely at the Permian's end.

In May 2004 announcement was made that the Permian killer crater may indeed have been found. Marine geophysics surveys off the northwest coast of Australia turned up a distinct anomaly buried under shallow seas that was promising enough to drill two deep holes in search for oil (several impact craters have served as petroleum traps). Here is a seismic refraction survey map that shows the buried structure including its central peak (40 km [25 miles] diameter}.



The structure has been identified by Professors Luann Becker (UCSB) and Robert Poreda and Asish Basu (University of Rochester) as an impact structure, and named the Bedout (pronounced "bidowe"; French word) crater. Most of the Bedout images shown on this page were extracted from their online paper (no longer available). The figure below shows the paleogeography at the close of the Paleozoic, in which most continents were grouped together in Pangaea. Bedout lies just off the soon to become Australian block, in shallow waters of the PaleoTethys Ocean.



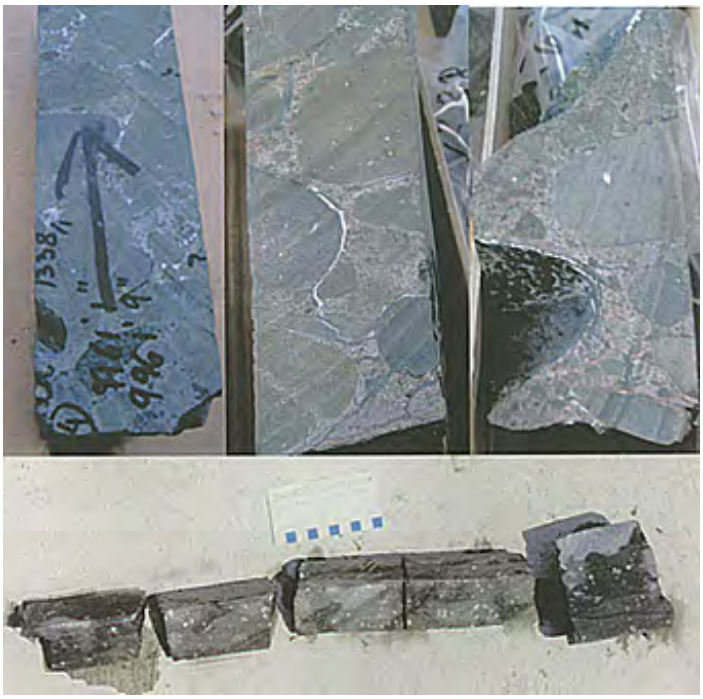
The yellow-centered circle is the Fraser Park, Australia locality containing what they (and others) interpret to be fallout debris - small fragments of rock and glass. The red dots are other localities - each containing a layer of highly probable impact ejecta containing such evidence as shocked meteoritic chips, glass spherules, Fe-Ni-Si grains, and fullerenes (so-called "buckyballs" made up of carbon arranged in a "geodesic" structure). Note the occurrence of extensive volcanism, the "Siberian Traps". (This outpouring of basaltic lava supports an alternative hypothesis for the demise of so much life, namely the accumulation of harmful gases and particulates in the atmosphere may have created adverse conditions [such as acidic seawater] that diminished life.) Basaltic volcanism also covered the Bedout structure on the sea floor; afterwards about 3 km (2 miles) of mid-Triassic through Cretaceous, and then Tertiary-Quaternary sediments were formed.

This montage of photos shows a land outcrop traceable to the Bedout event, a guide fossil, Glossopteris (a plant), which disappeared immediately afterward, and a grain of quartz which shows planar features - a key indicator of impact as discussed below and on the next page.

In the May 2004 paper, the above mentioned investigators claim to have now found actual pieces of the asteroidal impactor (estimated the size of Mt. Everest) itself. They note too that at the close of the Permian, Australia was part of Pangaea, so the asteroid might have struck on this supercontinent but the land containing the resulting crater has since detached and become buried by marine sediments and ocean water.

Core recovered from the structure, which presently has the dimensions of 200 km (125 miles) but may be even larger, produced fallback breccias in a zone more than 300 meters thick. This was initially interpreted as a volcanic breccia. One core segment shows the recovered breccia with a Chicxulub example along side.

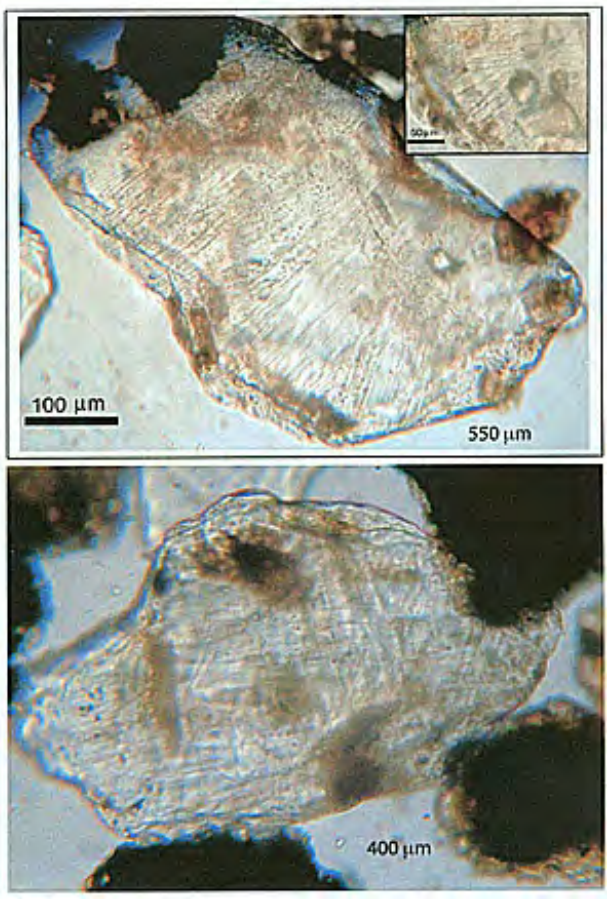
Another group of Bedout cores and core from Chicxulub (bottom) at first glance looks almost like gray sediment of clay or fine-crystalline limestone nature:



When the Bedout core was examined in detail, the breccia clasts were determined to be mostly made up of largely devitrified glass, containing plagioclase, iron oxides, iron-titanium oxide, and recrystallized chlorite. This is a typical petrographic microscope view of a Bedout clast, which resembles impact melts seen at various accepted impact structures:

Very strong proof of impact origin for Bedout is the shock metamorphic phenomenon of conversion of plagioclase crystals (as laths) into the glass known as Maskelynite. This is evident in this pair of images showing photomicrographs of a sample in which (in the top view) laths of plagioclase showing a brown tone are set in a crystalline albite (white) and titanite (black) matrix. In the bottom, the laths are now revealed to be totally isotropic (remain dark in crossed-polarized light when the microscope stage is rotated) as bespeaks of a glassy state.

Some Calcium-Magnesium Carbonate material in the core has been converted to glass (this is very rare but was first produced by the writer in an implosion tube experiment). Another indicator of shock at the Bedout site is the "toasted" quartz grains with single to multiple sets of planar deformation lamellae found rarely in the core but more commonly in the fallout layer tied to this event. Here are two examples from the Fraser Park, Sydney basin site:

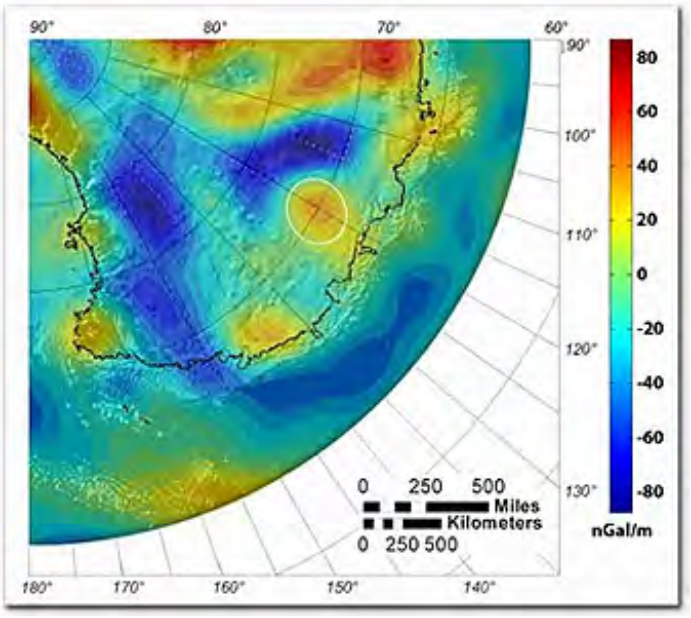


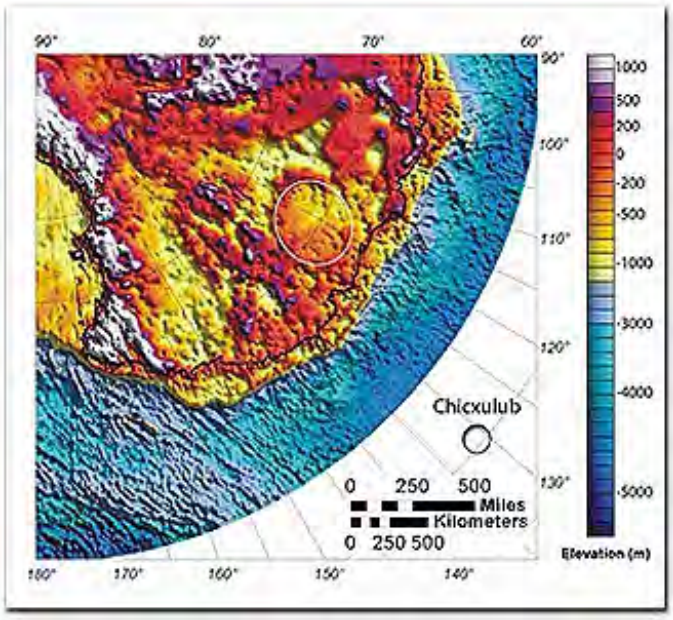
The Becker team and others have done age dating on the core materials from the crater and

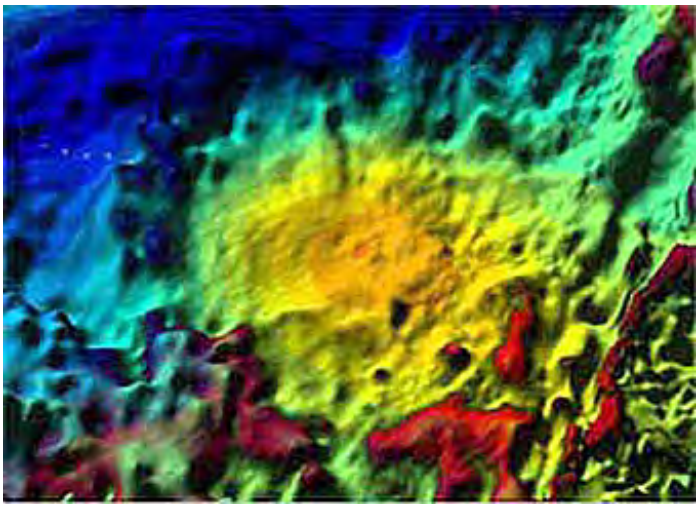
from fallout layers elsewhere. The Ar^{40}/A^{39} dating yields ages of 250.1 +/- 1 million years, almost exactly the time assigned to the Permian/Triassic boundary (end of the Paleozoic).

At the moment, several impact specialists have disagreed with the interpretation of the evidence cited so far as proof of an impact origin for Bedout, so the inevitable dispute from multiple interpretations has commenced (the concept of "multiple hypotheses" was first promulgated by a geologist). But the study is just beginning. If proved to be a huge impact, Bedout probably is at least one, and maybe the exclusive, factor in the Great Dying.

As so often happens in Science, just when a definitive answer seems to have been found, something new is discovered that clouds the issue. A competing event, in some ways superior because of size, has been announced as a leading candidate of the Killer Crater. This - as yet unnamed - lies under 1.5 km of ice in Wilkes Land in the Antarctic. The buried structure may be as wide as 500 km (300 miles); indirect evidence suggests it may be about the same age as Bedout (but drilling to sample its rocks will be needed to confirm this). It shows up as a gravity anomaly (mascon; positive mass caused by upward movement of denser subcrustal rocks) and has a crude circular outline. Inspect these three maps:

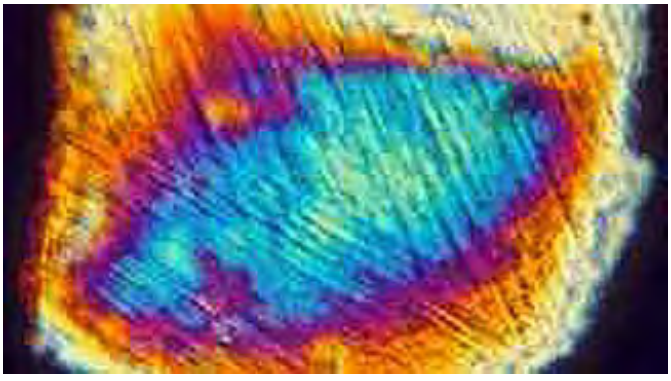






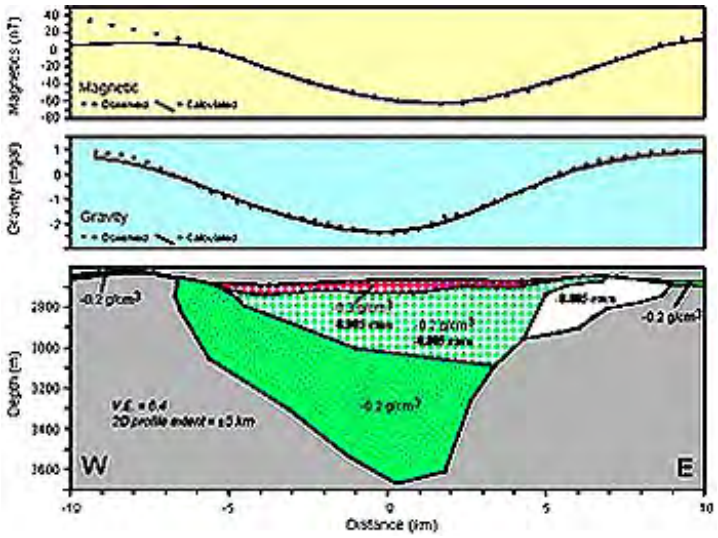
A buried crater under the Chesapeake Bay may have been large enough to kill off some life during the Tertiary. This structure has a diameter of at least 90 km (56 miles), with a central peak. It is located as shown below:

Like some others, the structure was first found during a geophysical survey. It lies buried beneath both coastal and estuarine sedimentary materials. It has been drilled; another drilling is underway. Recovered core shows extensive breccia units. Age dating of these deposits places an age of 35 million years for the breccia matrix. Proof of an impact origin includes telltale planar deformation features in quartz within breccia clasts, as shown here:



Chicxulub, Chesapeake, Bedout and Wilkes Land are among a growing number of impact structures that are buried and have been discovered during geophysical surveys. Another - almost a type case - is the Ames structure (~13 km diameter) in northern Oklahoma, found

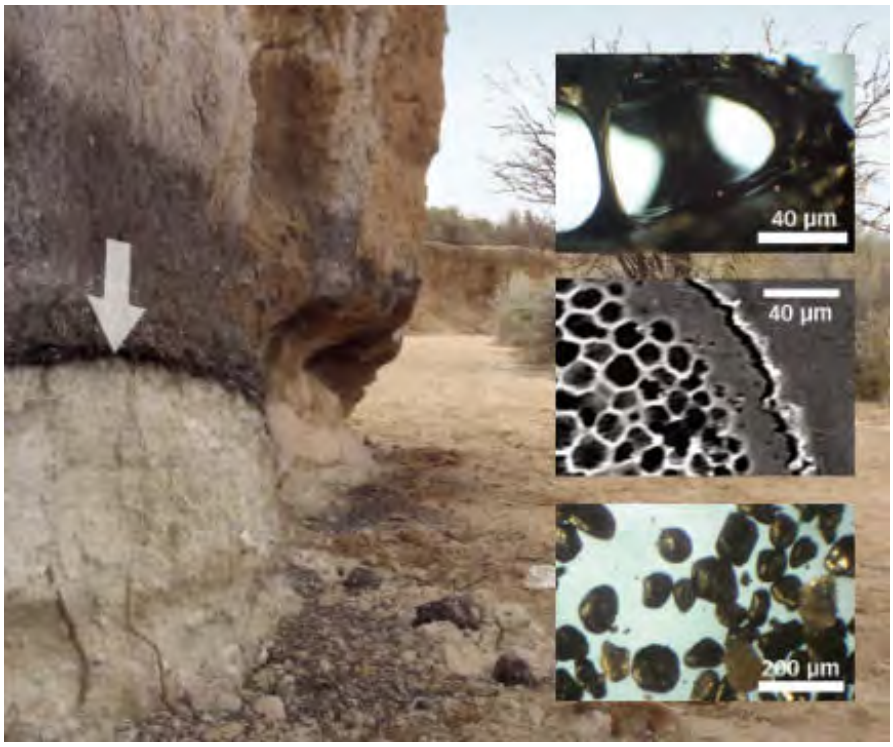
when it was picked up by gravity and magnetic surveys and then drilled in search for oil. Here is a cross-section prepared by Prof. Judson Ahern showing the crater, its distribution of materials of different densities, and survey results:



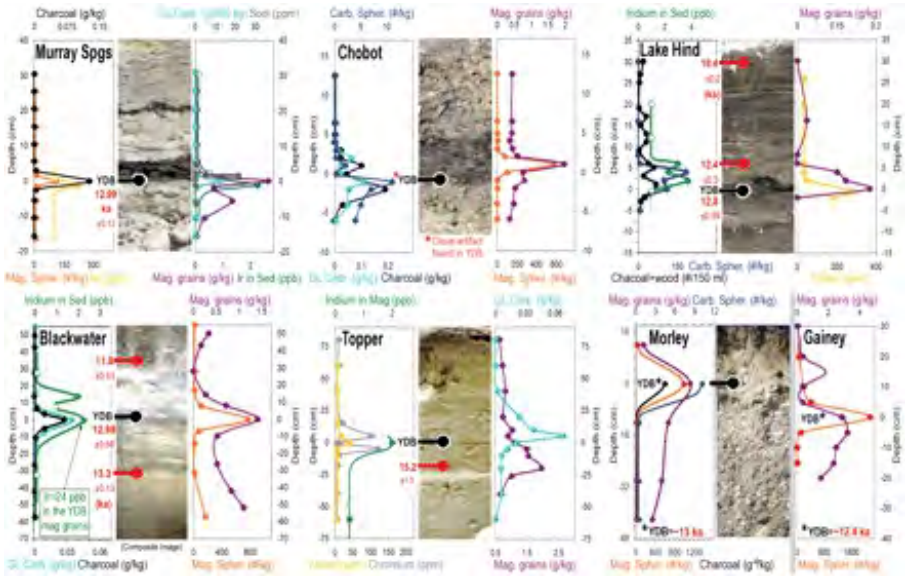
The above proposed impact events extend into geologic time, well before humans entered the evolutionary picture. But the threat of world-level catastrophes owing to impacts has now been accepted as real and concerning. Has there ever been such an event? The answer seems to be 'yes'.

Some 12,900 years before the present, there was a sudden, abrupt disappearance from North America of at least 35 species of large mammals including the Wooly Mammoth, the Saber-toothed Tiger, and the Giant Sloth. Primitive Indians known as the Clovis people also seem to have vanished as a culture. The climate too became colder (about an average of 8 degrees Celsius) and glaciation, already covering most of Canada, expanded. What could have caused this?

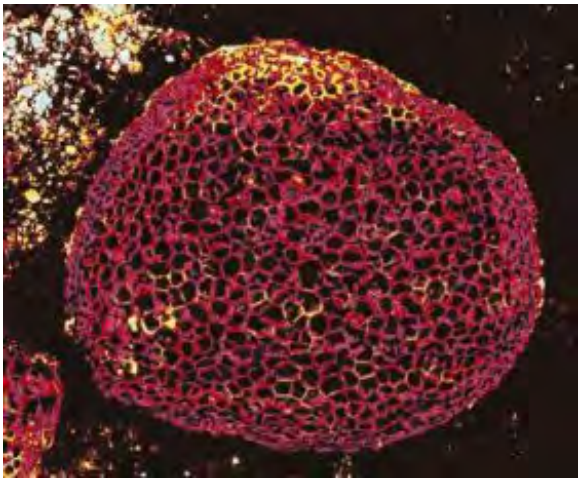
Although several explanations have been advanced, the best fit seems to be a comet impact. The Clovis comet, as it has now been named, is proposed as the culprit. There is abundant evidence for a major explosion, either in the atmosphere or in the Canadian ice (no crater resulting from impact has yet been found). A layer, named the Younger Dryan, of black soot, about 4 cm. thick, is found widespread across North America and even into Europe. Here is an outcrop in Arizona with this layer:



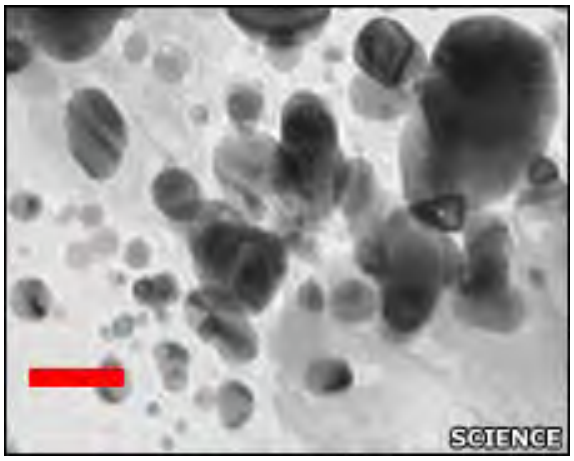
The layer contains several hallmarks of impact, including iridium anomalies, nanometer diamonds, and fullarenes. The iridium spikes are not as pronounced as those associated with the K-T boundary /layers.



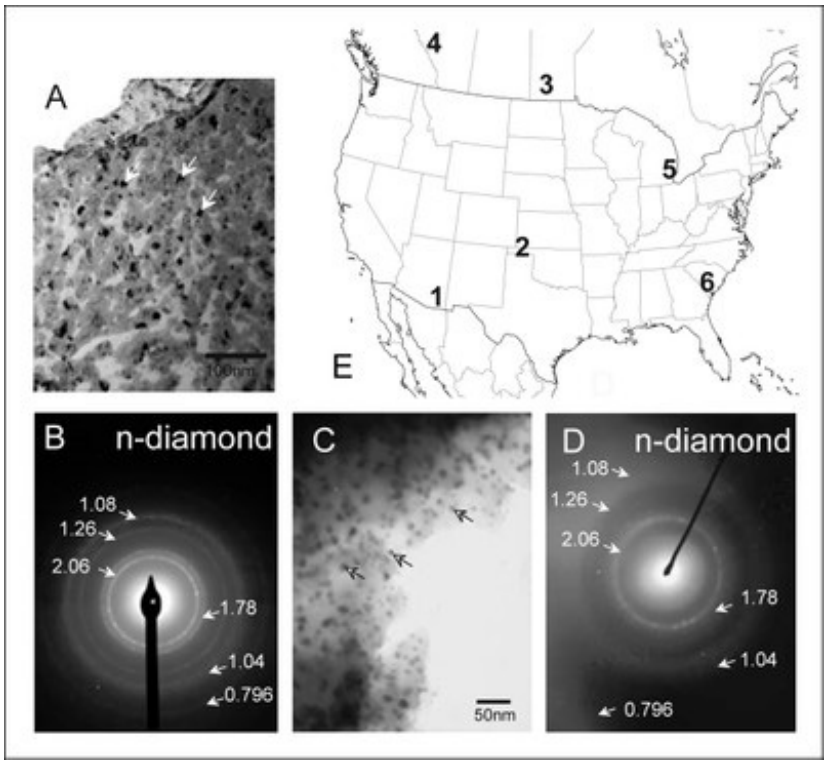
The spherules are distinctive, as shown here:



The diamonds are typical from 10 to 100 nanometers in width. Their crystal structure is hexagonal, not the isometric diamonds found in intrusive pipes on Earth; hexagonal diamonds are known only from settings that involve shock pressures, such as impacts.

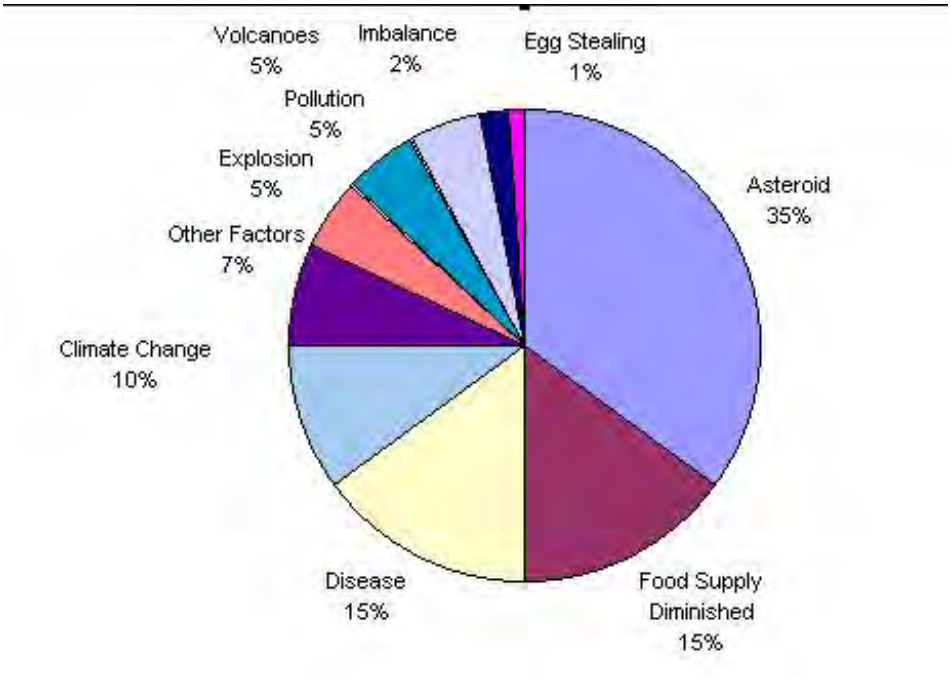


These diamonds have unique signatures when examined by transmission electron microscopy:



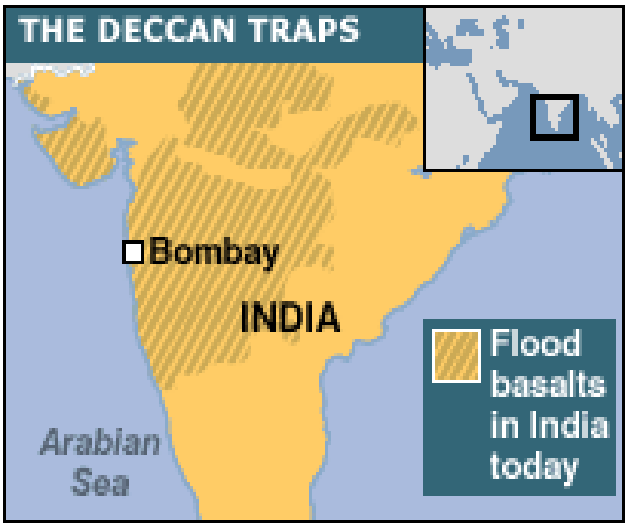
But as is usually the case in Science, hypotheses as "outrageous" as killer impacts have their challengers. Despite models that show how an impact crater hundreds of kilometers in diameter could affect the atmosphere and surfaces worldwide, there are those still skeptical of this mechanism. These scientists propose other reasons for mass extinction, chief of which is intense regional to worldwide volcanism which at its height could saturate the atmosphere with light-blocking ash and moisture.

So, where does the scientific community stand on the general topic of mass extinctions, both from the standpoint of impacts and the other possible causes of disappearance of entire classes of animals and plants. Comments above indicate that there is a difference of opinions about the actual reason why the dinosaurs did not survive past the K-T boundary. Whether this was due to one short-lived event or was the result of a set of circumstances spread over a gradual period of time is still being debated. Multiple explanations have been proposed. This chart is the plot of a survey among involved scientists as to the primary cause as well as possible secondary or multiple causes:

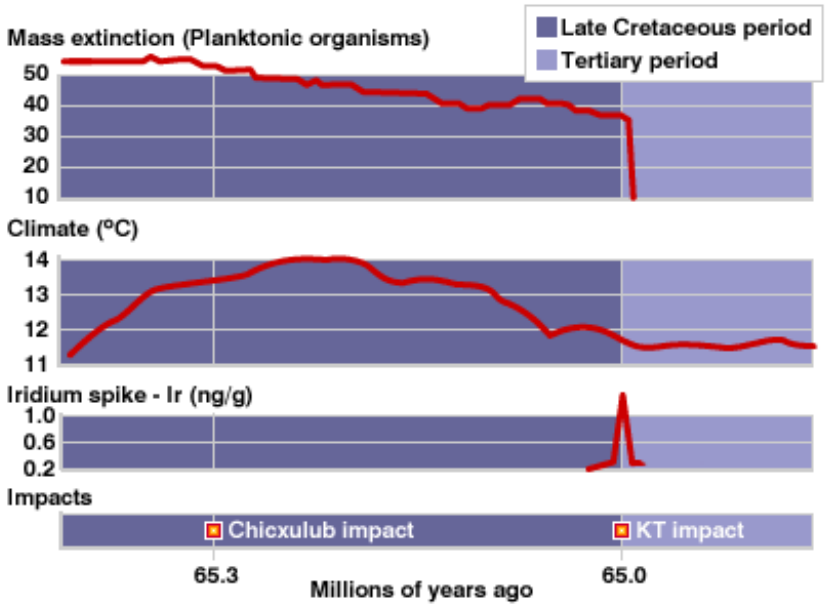


The impact hypothesis remains the currently favored cause of the extinction that prevailed at and after the K-T boundary layer was deposited. But, the mechanisms responsible for the extinction are still being debated. Some scientists claim they have found contradictory evidence pertaining to proposed causes. For instance, if global fires were instigated by the fallout from the impact, the atmosphere would have held huge quantities of soot that would lead to deposits of charcoal in the boundary layer. In fact, only traces of charcoal are present. Therefore, such extensive fires seem unlikely. Again, atmospheric chemistry would have been affected by the impact, with some models leading to acid rain, which would have severely stressed vegetation. But evidence for this is inconclusive. Various models result in the possibility of overall cooling and/or warming (in succession) of the atmosphere. This is to be expected from the vast amounts of dust particulates that an impact would eject into the atmosphere. So, once more the evidence is frought with uncertainty for either modification of temperatures. The presence of fern remnants just above the boundary layer is consistent with the postulate that, although vegetation was severely diminished by whatever was happening to the environment after the K-T event, plant life bounced back rapidly. Animals - mainly small mammals - also reappeared soon thereafter, as these apparently survived the adverse effects in the global environment. Needless to say, much more research must be done to resolve these anomalies.

Another "bone of contention" is whether a Chicxulub-sized crater tosses enough material into the atmosphere to have had a global influence. To go global would require efficient transportation and distribution by the atmospheric circulation system operating at the time. In today's world, the present circulation would likely not carry particulates across the globe; transportation into the southern hemisphere from a northern hemisphere event would be difficult. In the writer's opinion, this is a major stumbling block to the idea that the Chicxulub impact (**for which the evidence is very solid**) was the sole cause of the general extinction that did happen. It seems more likely that there were other contributing factors. One, often cited, is the likelihood of extensive atmospheric contamination by chemicals (including sulphur compounds) introduced from the vast outpourings of lava that formed the Deccan Traps (basaltic flows) of India. These lava deposits extend over much of the Indian subcontinent and the sea floor to the west, as depicted in this diagram:



Another problem with the Chicxulub model may have merged. Examine this diagram.

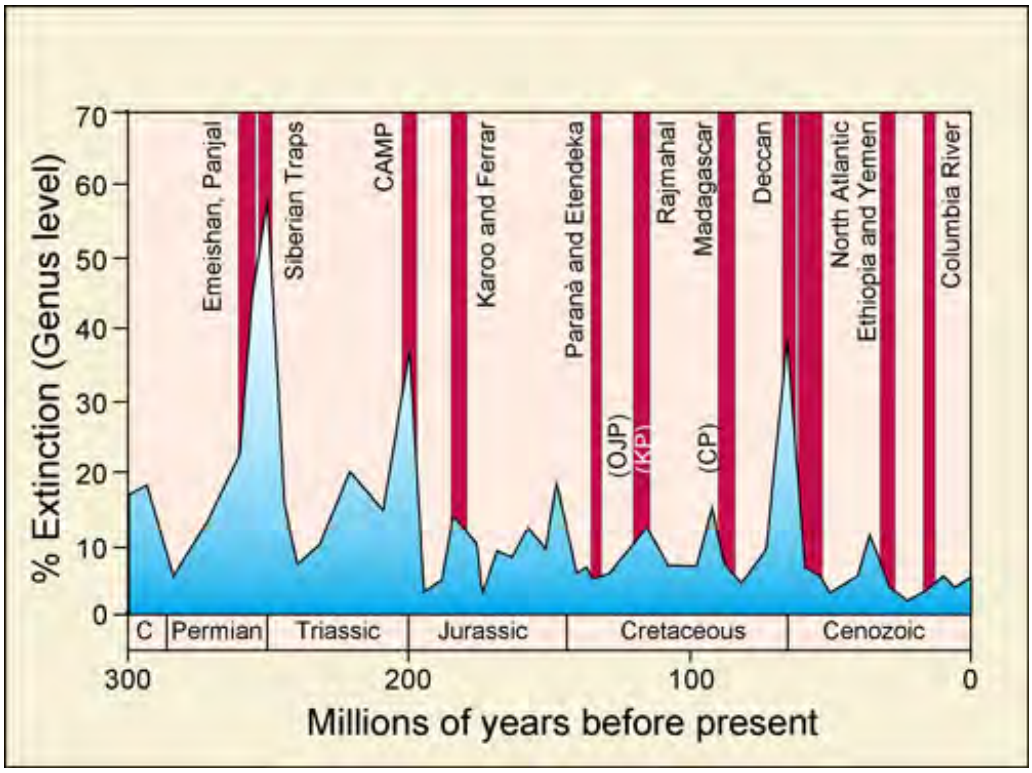


The bottom chart is our "talking point". A group of scientists who are studying the Chicxulub material obtained through drilling have acquired radioactive dates from the volcanic rocks that indicate the rocks are 300000 years older than the K-T boundary layer age. If this is valid (but the error range for the dating might overlap the two ages), then the implication is that there were two separate large impacts closely spaced in geologic time - not impossible, but not of high likelihood.

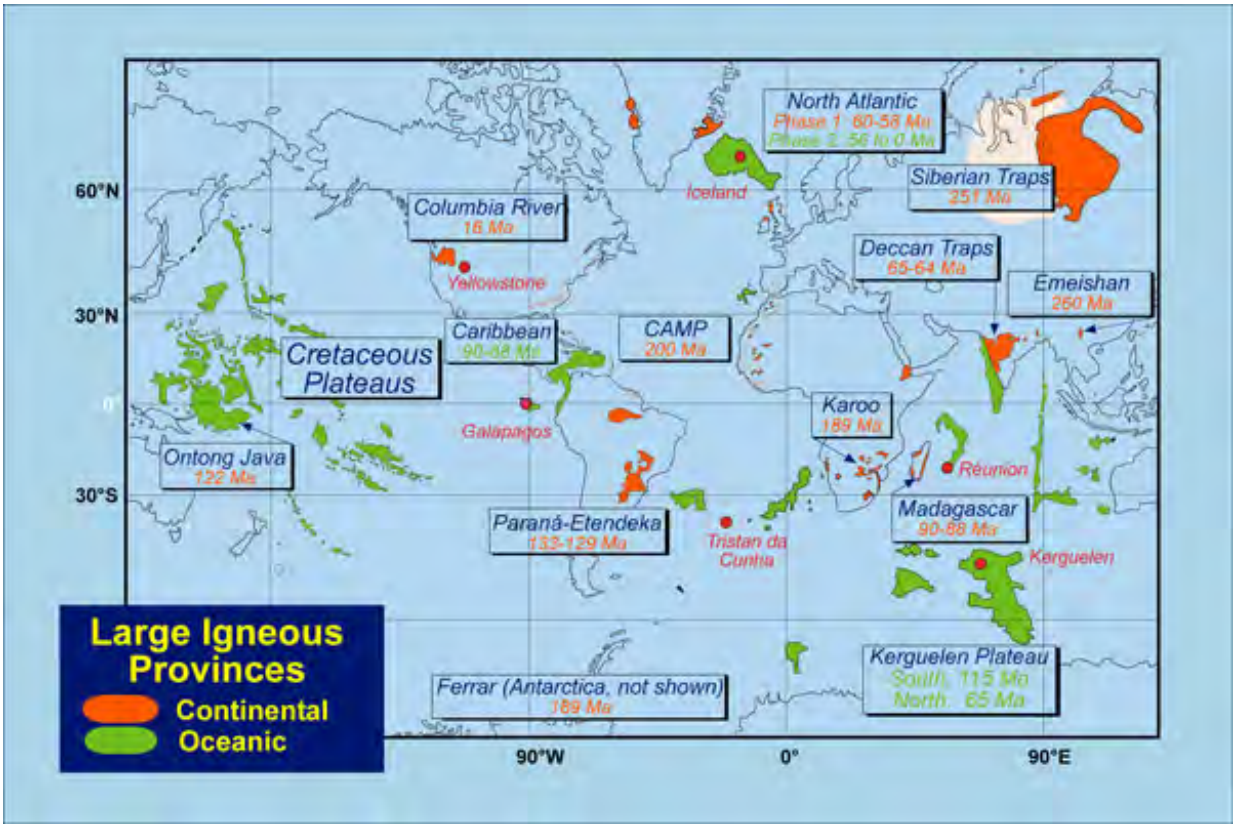
There is another, perhaps more imposing, objection to impact as the main reason for dinosaur (and other animals) extinction. This is the mounting evidence that the dinosaurs were faced with adverse conditions over millions of years prior to the K-T event which were making their environment increasingly unstable. Evidence in upper Cretaceous rocks show a general trend of aridity (desertlike conditions) across much of the world. If this indeed did happen, it could account for the decrease in the number of dinosaur species observed in Cretaceous rocks prior to the K-T event. So, the dinosaurs were being stressed and heading for extinction anyway, according to some views. The impact(s) at the close of the Cretaceous would have some (probably serious) influence on the dinosaur demise - they would have administered the "coup de grace", but were not the sole cause.

The most frequently cited alternative to impact as the primary cause of mass extinctions is significant volcanism. Although stratocone eruptions can introduce a wide variety of toxic gases into the upper atmosphere, these tend to be localized and of short duration - one at a time.

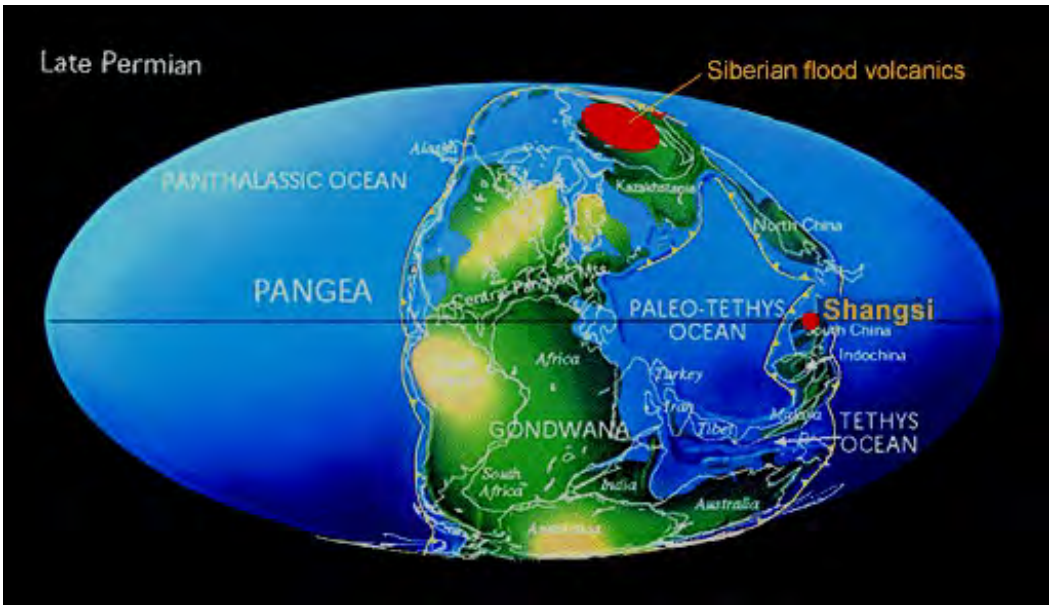
Flood basalts cover a much wider expanse, and calculations show that the amount of gases from such long-lasting events will contaminate the atmosphere (probably globally) to a much higher degree. This next diagram plots extinction events and major flood basalt volcanism:



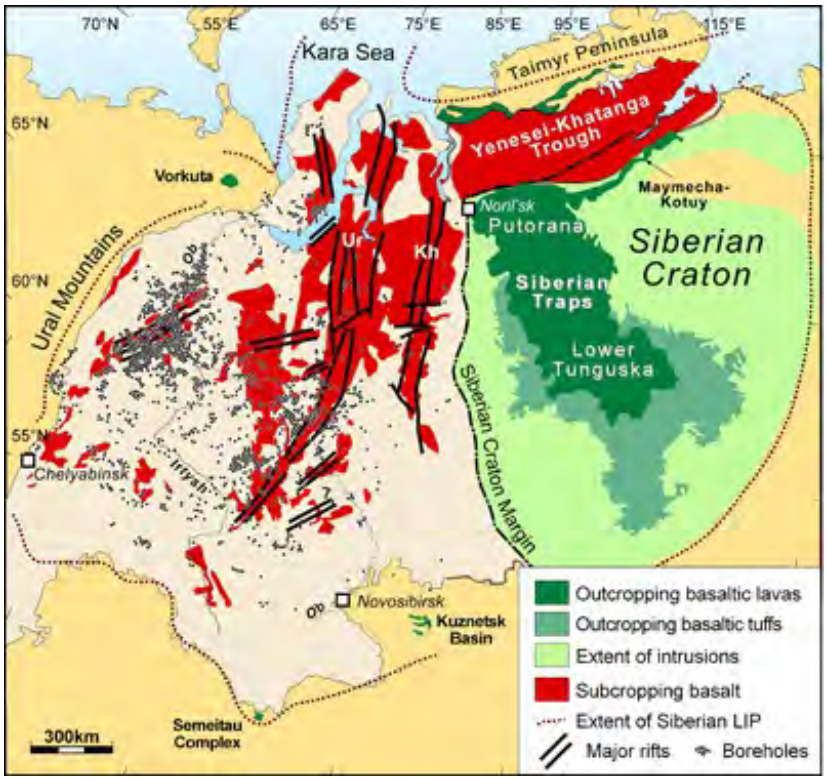
It is evident that there is some definitive correlation - at least some volcanism on a grand scale occurred during periods of mass extinction. This diagram shows the geographic location of major flood basalt outpourings, along with other extensive volcanic deposits:



But keep in mind that the distribution of continents was much different in past times. For the calamitous Permian extinction event, the Siberian Trap basalts - which many believe were the prime cause of the extinction - were located in the northern part of the supercontinent Pangaea.



Details of the Siberian Trap deposits are evident in this map:



The Siberian Trap basalts cover a large area of north-central Siberia. We saw a Landsat image showing Trap terrain in Section 6. The image used there is repeated here:

Two websites bearing on the Siberian Traps are found at Wikipedia and University of Glouchester (be sure to click on "Volcanism").

So, in sum: Like the K-T extinction, the cause(s) of the other extinctions is still being debated. Large impacts certainly can be detrimental to the animal and plant species of the time. But volcanism, climate change, and other factors must be considered. However, the mechanism that is most likely to force the various environments at any given time into "catastrophe mode" is worldwide atmospheric contamination that brings about temperature and precipitaton modifications so drastic that a multitude of species cannot adapt. Impact and subcontinental scale volcanism are the two causitive leaders.

A good treatment of the K-T event, with links to web sites that consider mass extinctions is presented on the reliable Wikipedia K-T Extinction and Mass Extinction Internet connections.

Now that you are familiarized with the general appearance of impact craters and their possible role in influencing life on Earth, we next will consider the best evidence for proof that a circular structure has been caused by a cratering event - shock metamorphism.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Prior to satellites with high resolution, much of the searching for impact craters in more desolate parts of the world used aerial photography, if it existed. The Canadian group at the Dominion Observatory found a number of craters that way. With Landsat, SPOT and now much higher resolution IKONOS and other systems, and worldwide coverage, the possibilities for finding new craters have notably improved. To date, at least 10 more have been found with the aid of space imagery. And, special processing such as the enhancements described in Section 1 can bring out new information about a crater imaged by a digital system. This page contain examples of impact craters as recorded in both aerial and space imagery that are found in North and South America. Particular attention is given to Meteor Crater in Arizona.

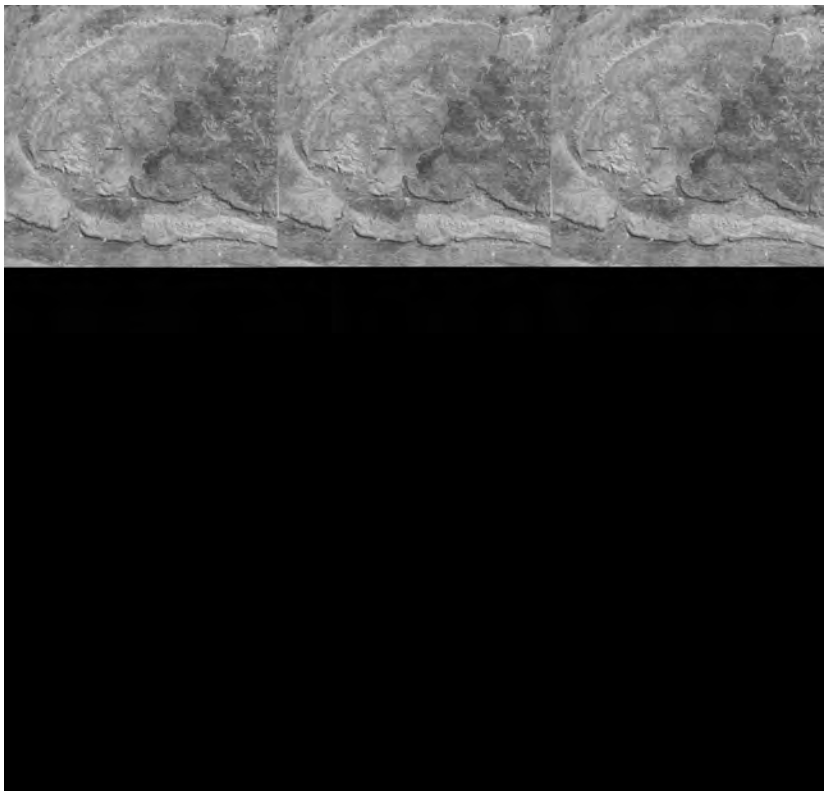
Remote Sensing of Craters

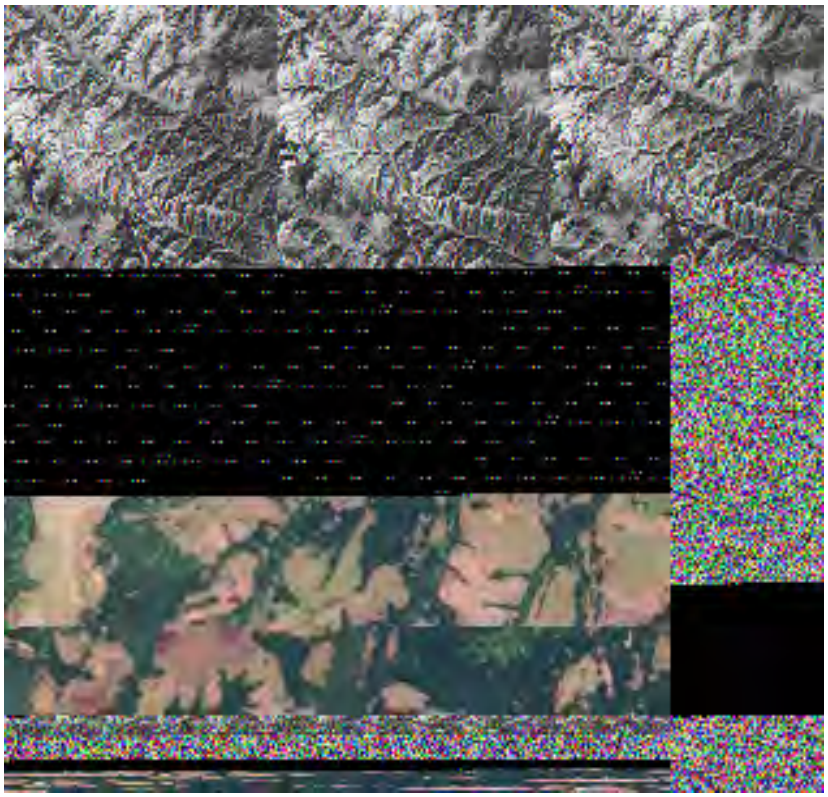
How has remote sensing played a role in the search for or verification of a supposed impact crater? In two ways: 1) recognizing morphological features that are compatible with an impact origin (but this does not rule out certain volcanic craters) and 2) detecting shock-induced changes as spectral changes (this is usually difficult to do). The strategy is one of simply using remote sensing to identify landform anomalies consistent with an impact origin and then confirming them by on-site inspection and examination of rocks for shock metamorphic effects. The absence of shock effects leads to ambiguities, making volcanic craters an alternative hypothesis.

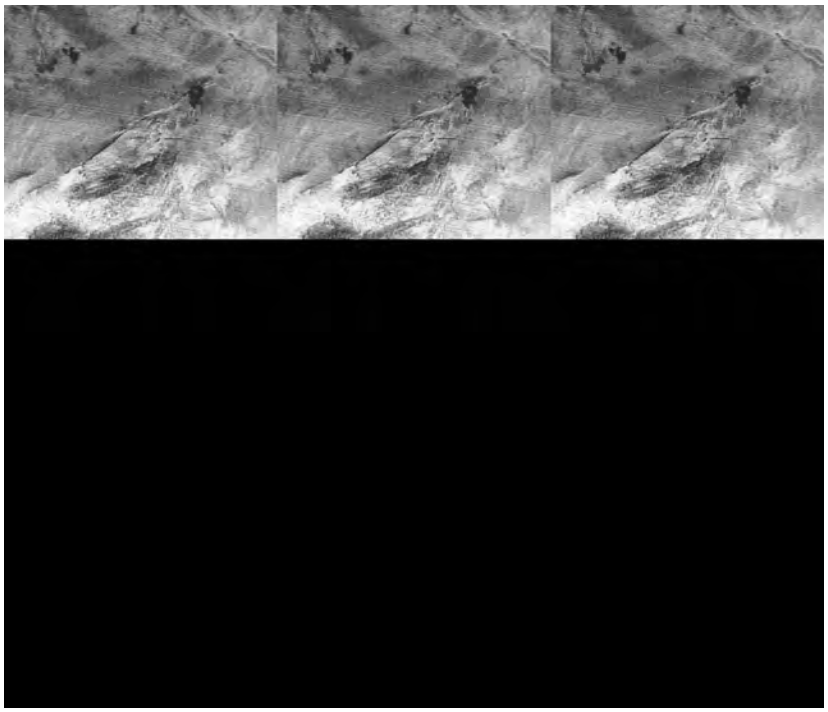
Impact-minded searchers have discovered at least ten new impact structures through satellite remote sensing. Before Landsat, and to some extent since 1973, observations from airplanes, and in particular aerial photos, were a source of information about possible craters. Many impact craters are fairly large, and a few are huge, so that in favorable settings these survive various forms of erosion (terrain containing such structures that becomes involved in major tectonic activity, widespread volcanism, burial by thick sediments, or glaciation may experience obliteration of distinctive crater morphology, but even eroded ones [astroblemes] may have subtle circular geometric expression). Thus, in large parts of the world the more obvious craters have been found by conventional techniques. But in parts that were poorly mappped or explored evidence from satellite imagery may be the first overview type looking that discloses heretofore unnoticed impact shapes or scars.

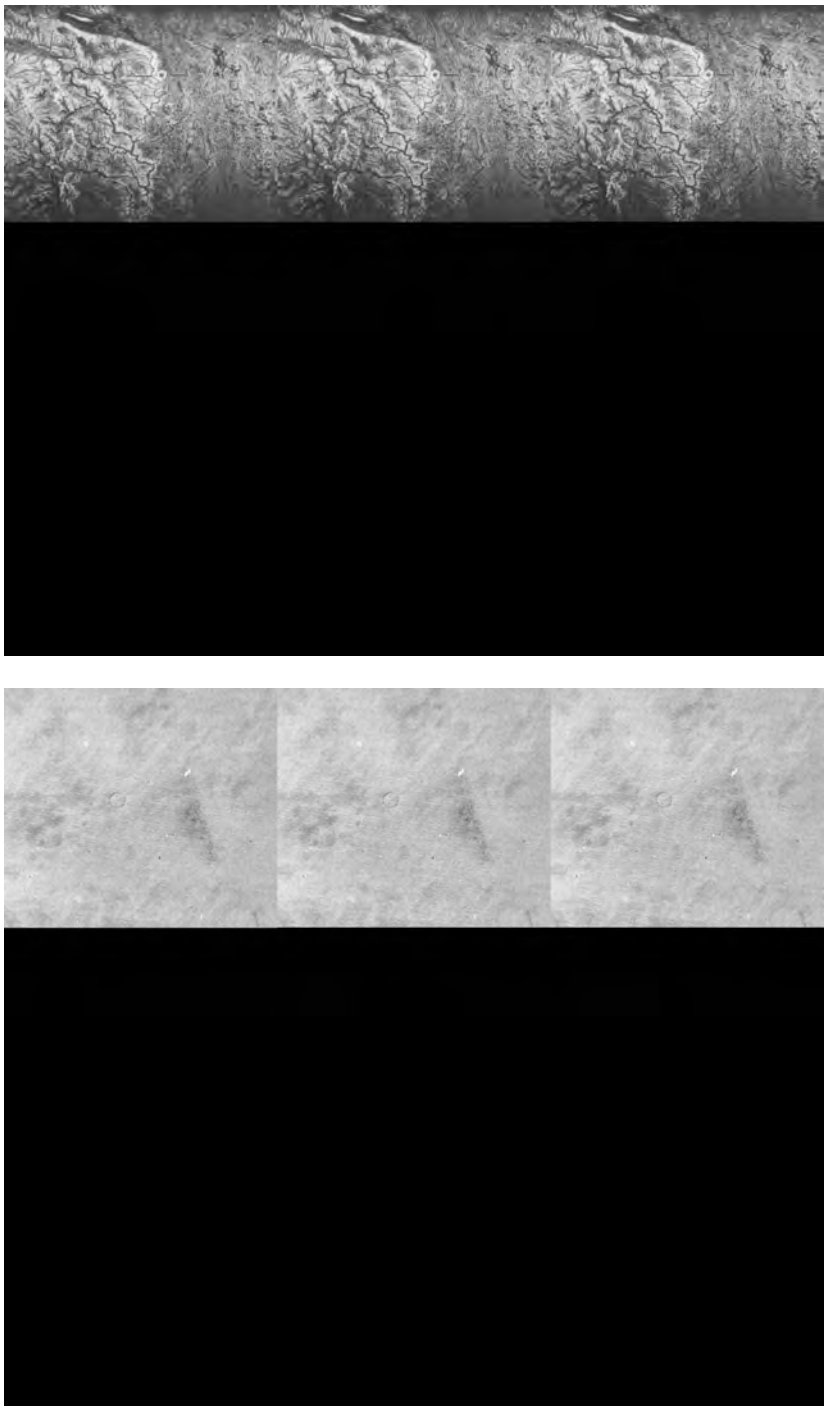
As seen from above, in photos and images, impact craters have several hallmark signatures. Circularity is the prime sign. If young, a crater will show a rim, and rarely an ejecta field. Most older craters have been worn down by erosion, so that morphology may not disclose them. One distinct feature, shown in several illustrations in this Section, is a perturbation of stream drainage. Rivers may find weak rocks related to the structural deformation caused by the impact, and thus adjust their course, leading to an expression of partial circularity. One such example is the Wembo Nyama structure in Africa, shown below, as discovered in space imagery. But so far no verification of an impact origin has been reported. This is a good example of how one must be cautious in any claim of impact as the origin of a feature. Other geologic situations can produce circular features.

Prior to Landsat, searches for impact craters using remote sensing were confined to aerial surveys. However, discoveries were usually serendipitous, since these surveys were expensive, were normally confined to small areas, and were initiated for other reasons, such as the hunt for oil or mineral deposits. Landsat afforded worldwide coverage, and has been the prime imagery used in looking for craters. As this and the next page will demonstrate, many of the larger craters, whose locations were already known, do indeed show up well in Landsat imagery. But systematic searches for new craters, most of which would likely be small since larger one could typically have been found by other means. Only a sharp, and suspicious, eye could spot new ones. To substantiate this statement, we show here four craters (three previously known) in rugged or bland terrain, that are hard to see (on two images black horizontal lines [hard themselves to see] have been drawn to pinpoint the craters; the caption may give further clues):

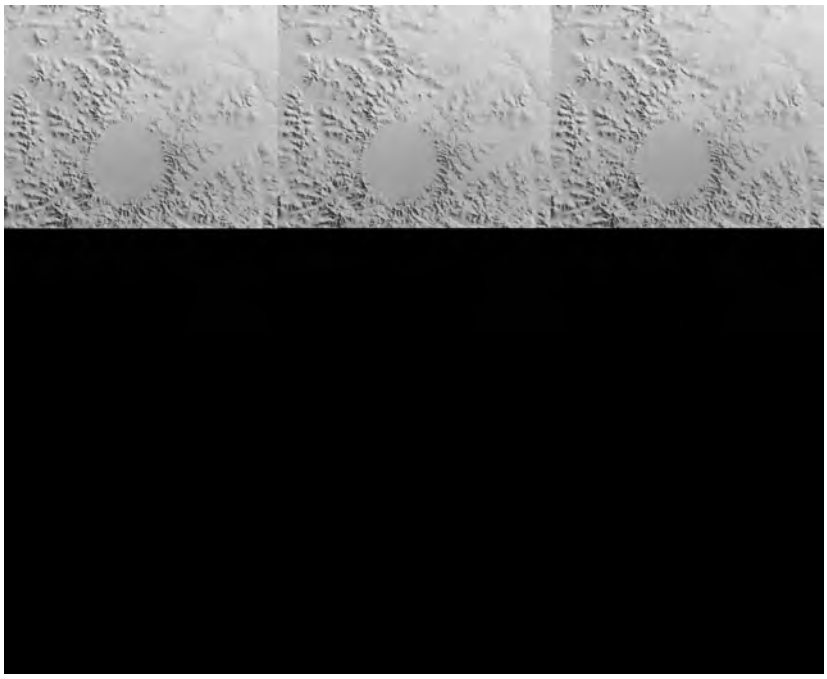








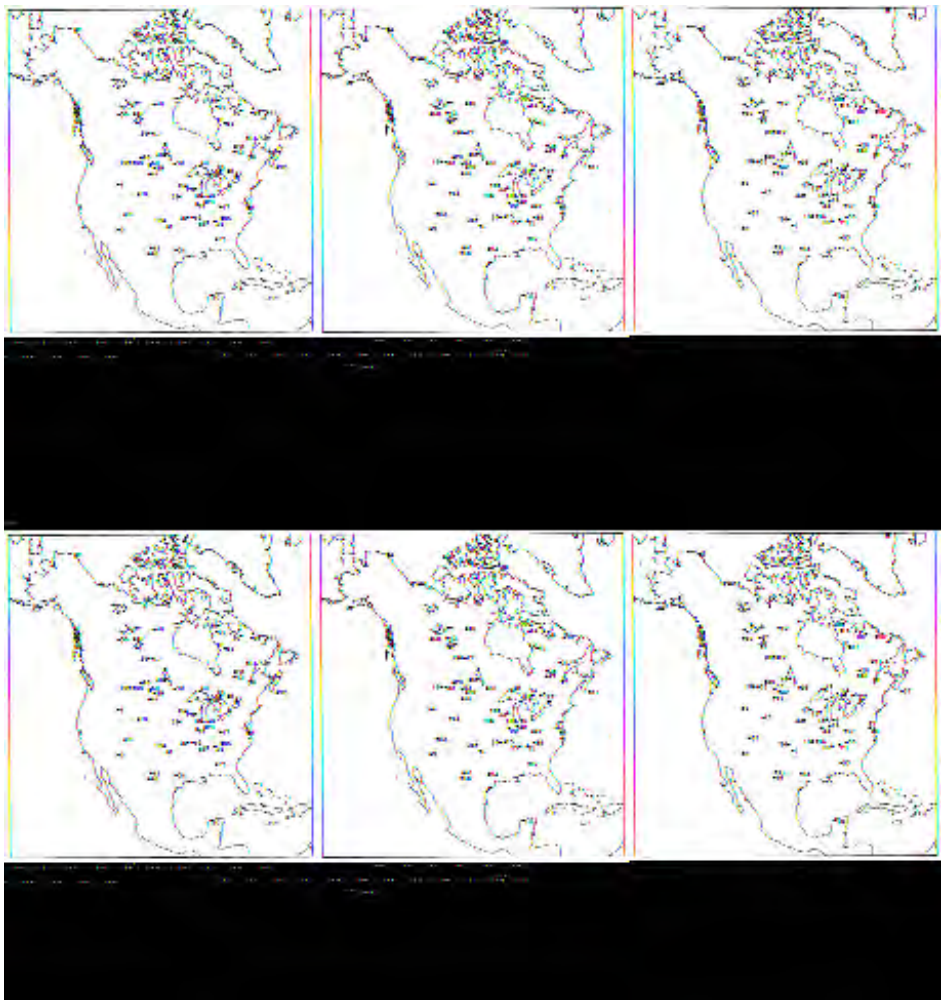
Sometimes "Mother Nature" helps out. Snowfall has emphasized this unidentified crater:



On this and the next page, we will show many satellite-acquired images of previously known, and a few directly discovered in those images, craters. Those in North and South America are considered on this page; those in Europe, Africa, Asia, and Australia are treated on the next page.

We can gain a feel for what to look for at a crater site, especially variations in morphologic expression due to differences in erosional state, by switching again to the Geological Survey of Canada's Web Page on Impact Cratering. Go to the list of individual craters and check these especially interesting sites: North America: Brent/Clearwater East and West/ Deep Bay/ New Quebec; South America: Araguainha; Africa: Aorounga/ Bosumtwi/Rotor Kamm/Vredefort; Europe: Ries; Asia: Bigach/Popigai; and Australia: Henbury/Wolfe Creek. Another collection of air and space imagery focused on terrestrial craters has been compiled by Koeberl and Sharpton as online slide sets

We begin with North America. This map shows the 57 that have been found so far:



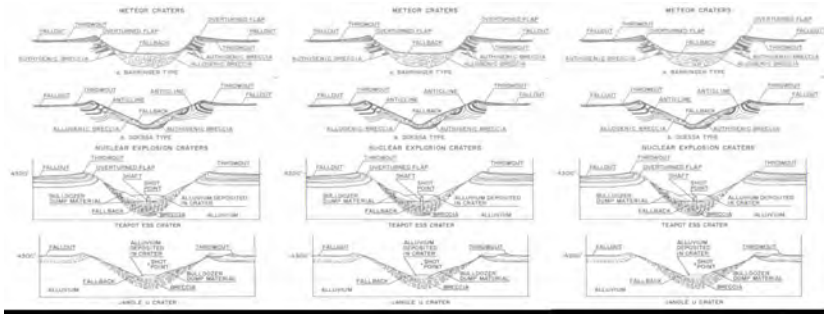
The numbers on the map correspond to a list that makes up this Web Page

Surprisingly, very few of the impact structures in the United States have good surface expressions as craters with rims. Most are astroblemes - eroded morphology but with the rocks still possessing shock effects such as shatter cones or PDFs (the Middlesboro structure in Tennessee is an example of this. Some have been discovered either by geophysics or by drilling (for oil or water). Thus, remote sensing has few U.S. examples to point to. The exception is probably the most famous impact structure in the world - at least to Americans - Meteor Crater (also called Barringer Crater) in Arizona.

This crater exists now as a 50000 year old depression cut into the flat-lying sedimentary layers below the surface of the Colorado Plateau some 73 km (45 miles) east of Flagstaff, Arizona and a lesser distance west of Winslow. An aerial oblique view of this 1230 m (4000 ft) wide crater shows its freshness (pieces of the iron meteorite that caused it can still be found in the ejecta); the road allows thousands of tourists traveling along Interstate 40 to visit its overlook and museum.

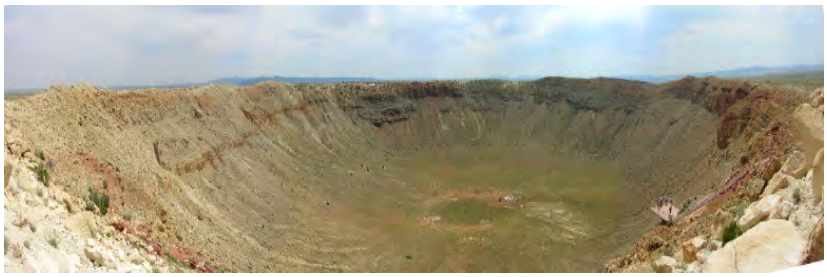
Modern field studies of Meteor Crater in the late 1950s by Eugene Shoemaker and its shocked rocks shortly thereafter by Edward Chao led to the first modern concepts of impact crater mechanics. (The SiO_2 polymorph Coesite was first discovered in impact structures at this crater.)

Shoemaker also was the first to study nuclear explosion craters at the Nevada Test Site (NTS); he and R. Eggleton presented their results at a 1961 Cratering Symposium, with this illustration as their centerpiece:



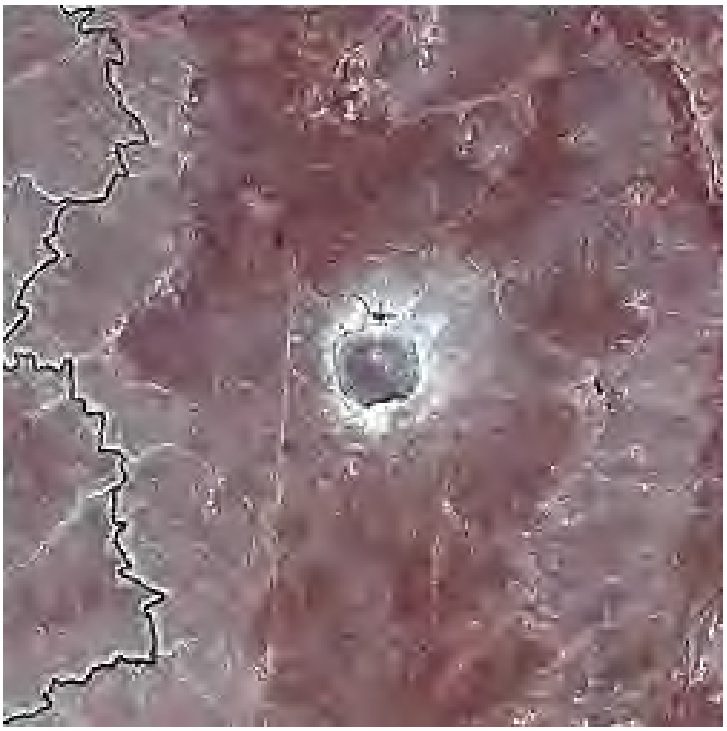
Many craters are now found at NTS. Most of these result from collapse of the alluvium above a short-lived spherical cavity after a nuclear explosion. Here are two such collapse craters at NTS:

The land around Meteor Crater is locally flat. This is how the crater appears as one drives along the paved road leading to the overlook and museum run by the Barringer family, who actually own the crater. Below is a view looking in from the overlook:

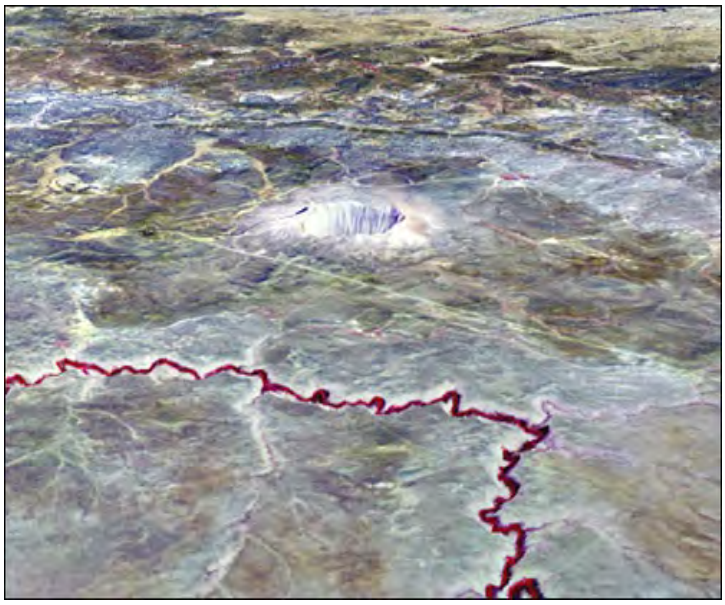


The flat interior floor, without a central peak, is a characteristic of simple craters; Meteor Crater's outline tends towards a square shape - this departure from circularity is controlled by the dominant set of two orthogonal joints (planar fractures) that run through the layers; and the ejecta deposits outside the rim still retain a hummocky (mound-like) topography. Another ground photo from its rim 185 m (600 ft) above the floor gives a sense of its grandiose size; note the displaced (fault-bounded) blocks under the rim in both aerial and ground photos.

The "squareness" of Meteor Crater is obvious in space imagery, such as this subscene made by Landsat; note the white apron of ejecta surrounding the crater (a hint of deposits still farther out is evident in the brown surface beyond the apron):

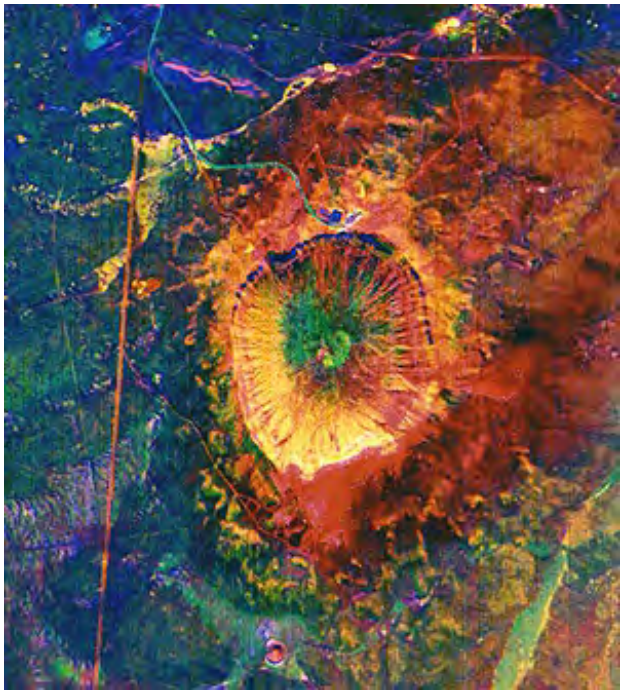


The ejecta apron stands out also in this IR photo:



Meteor Crater has been known for well over a century. G.K. Gilbert at the turn of the 20th Century reported a purely terrestrial origin for this crater, believing that it was probably a blow-out caused by groundwater encountering hot rock below. He largely ignored the presence of iron meteorites found in the apron near the crater.

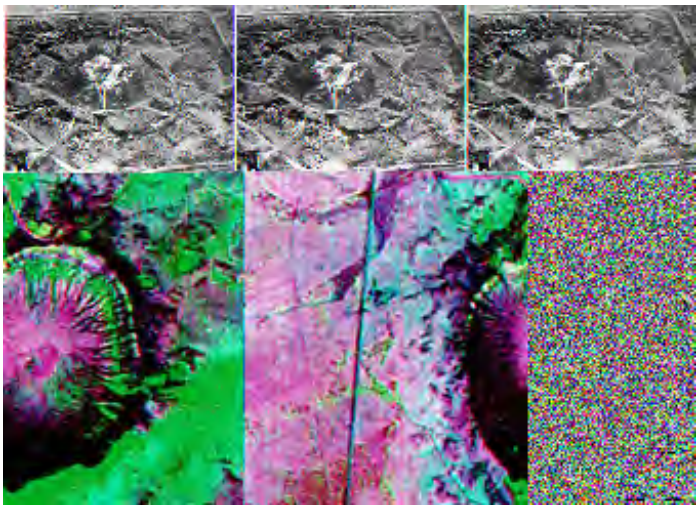
A specially processed image made by the airborne Thematic Mapper Simulator (TMS) shows that the ejecta blanket or apron (in reds and yellows) around Meteor Crater is asymmetrically distributed with maximum extension to the northeast. There is a notable tendency for the ejecta deposits to appear elongated to the northeast; this may be mainly an effect of wind-blown re-working rather than impact angle. The ejecta contain fragments of the iron meteorite which caused Meteor Crater, along with iron melt spherules. The red and blue lines are power lines and roadways.



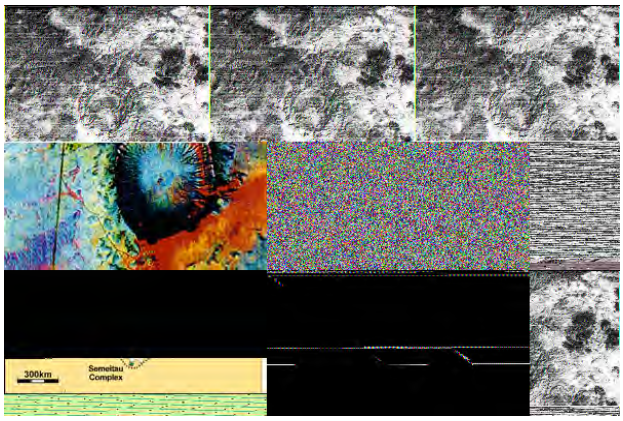
18-12: Assuming the ejecta blanket pattern is not principally a wind phenomenon and instead is the result of ejecta being tossed out preferentially in one general direction owing to the meteorite coming in at a low angle, from what direction did the bolide come? What is peculiar about the crater outline? What might explain the tiny round depression near the left bottom of the image? What could the long straight red line be?
ANSWER

A thermal multiband color image made (courtesy: Dr. James. Garvin) from the airborne TIMS (Thermal Infrared Multispectral Scanner) sensor divulges the expression of this ejecta, with reds and some yellow corresponding largely to Moenkopi Siltstone and Coconino Sandstone (whose spectral properties in the ejecta are influenced by their particulate nature and, possibly, by shock effects) and the blue-greens to the overlying Kaibab Limestone.

Much smaller, but possibly contemporaneous with Meteor Crater is Odessa Crater in West Texas. The only sign that it might be impact or even a crater as such is the disturbed limestones exposed at the surface. Petrographic shock effects are sparse and inconclusive, and shatter cones are absent, but iron meteorites around the site are considered definitive evidence. Here is an aerial view:

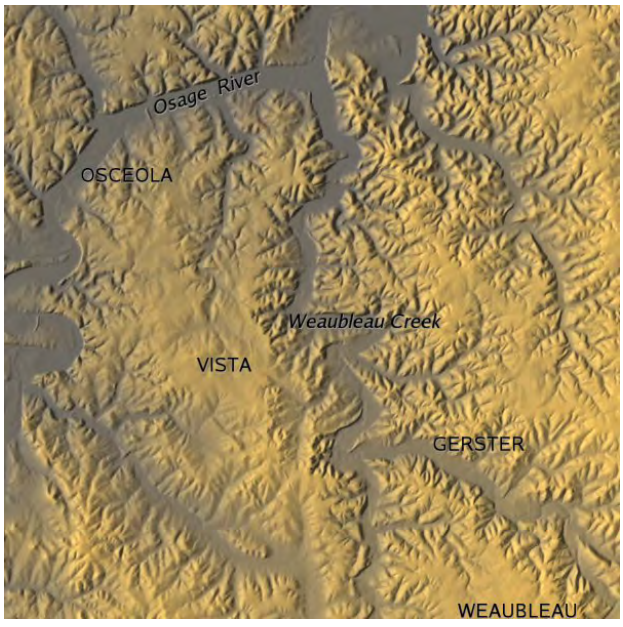


A 13 km wide structure not far from Odessa, Sierra Madera, was first identified as impact in origin from the presence of shatter cones. Here is a space image which clearly shows the crater's outlines and a hint of a central peak:

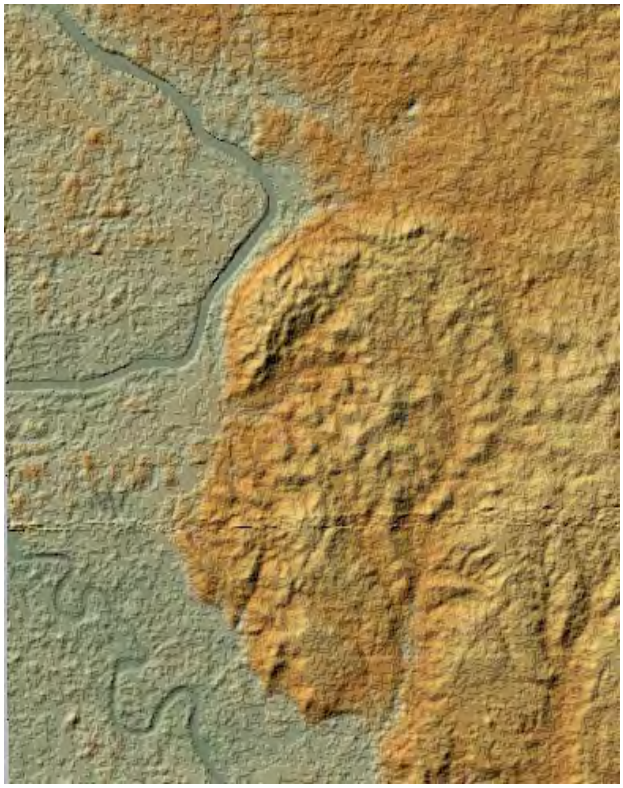


The Crooked Creek structure in central Missouri was first visited by the writer during a 1962 field trip to the so-called "cryptoexplosion" structures in that state. At that time, despite the presence of shatter cones, most geologists held these structures to have formed from volcanic action (ascending lava caused some kind of blow out), to which the name "cryptovolcanic" is given. The impact identity of this 7 km crater, seen in this aerial oblique view, has since been confirmed:

These next images exemplify how craters that don't have good expressions in space imagery can however be visualized in other ways. The Weaubleau structure in northwest Missouri was identified as an impact crater through field studies and recognition of shock metamorphic effects. Its circular outline, although faint, is evident in this shaded relief map;



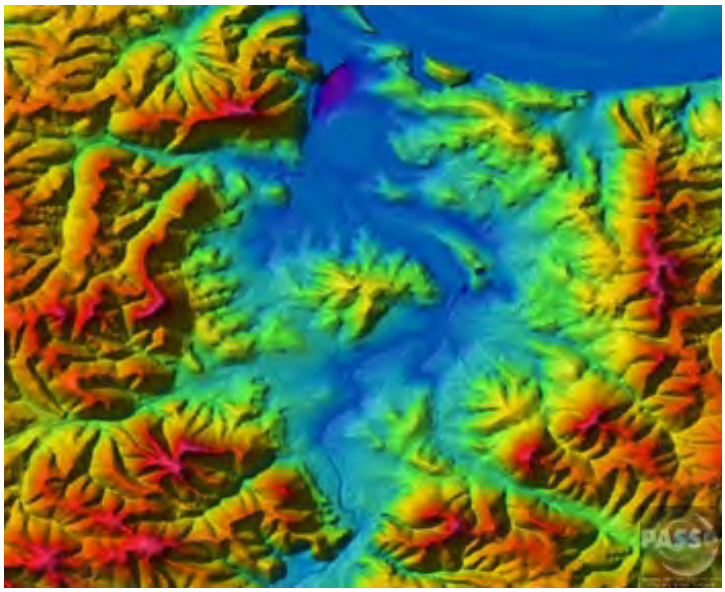
The Wetumpka structure is an eroded 7.6 km diameter structure in Alabama, north of Montgomery. It shows up in both a Landsat image and a DEM product made from SRTM radar data:



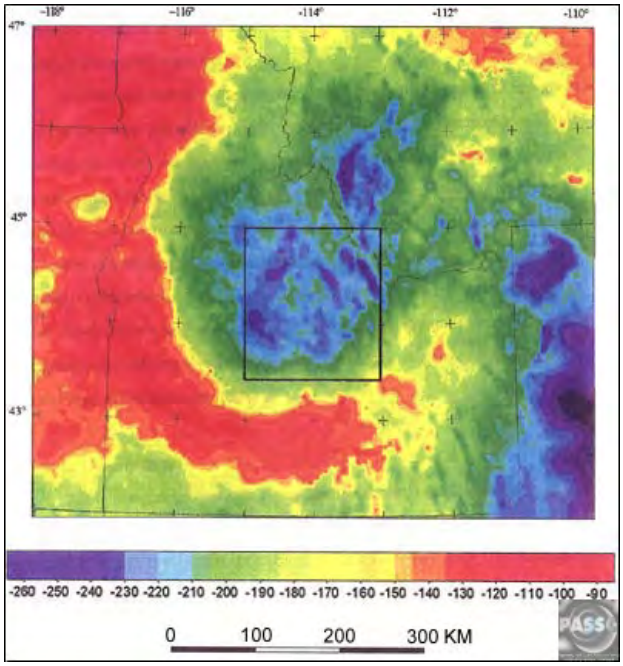
Attention was drawn to disturbed sedimentary rocks in coastal plains units at the Wetumpka site. This led to discovery of shatter cones and other shock features.

A similar subtle expression of crater morphology is evident in this colorized relief map of the 7 km

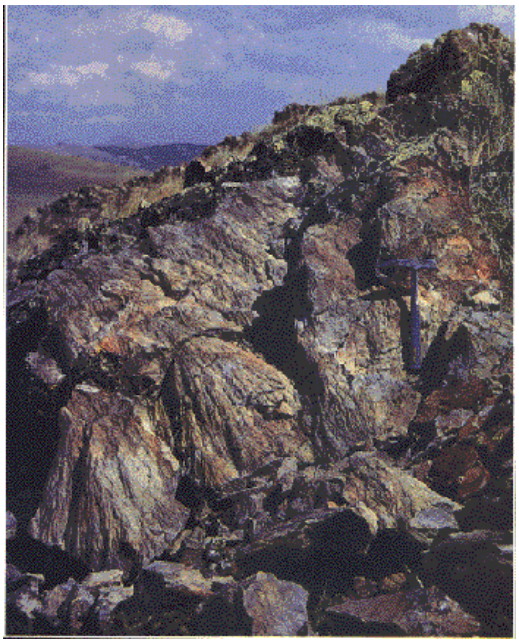
Wells Creek structure in Tennessee, which has a small central peak:

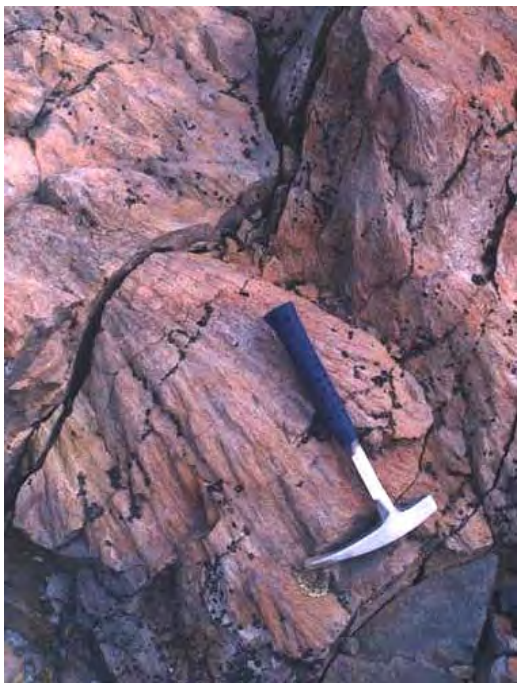


Possibly the largest impact crater in the continental U.S. is the Beaverhead structure, astride the Idaho-Montana border (both states claim it). The structure, which extends across at least 60 km (but may be as wide as 150 km) has no circular outline since it predates the Rocky Mountain orogeny so that tectonic activity has destroyed its morphology. It does have an expression in the Bouguer gravity anomaly map of the region:



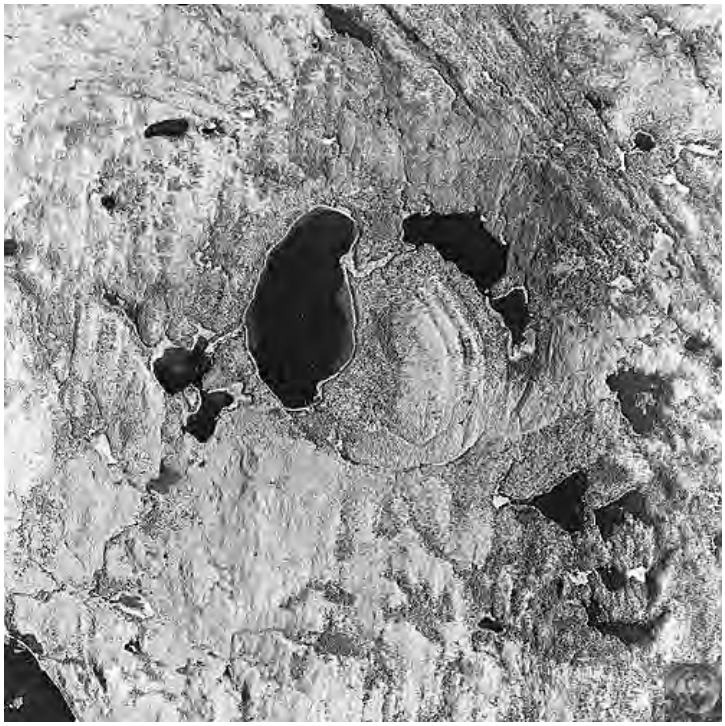
Large shatter cones and shocked rocks have been found in the Beaverhead structure:





Canada has a large number of craters (as indicated on the map at bottom of page 18-1) largely because much of that country is exposed Precambrian Shield - made up mainly of hard, resistant igneous and metamorphic crystalline rocks. Glaciation has carved into some of these craters but glacial debris and boreal forests can mask smaller structures. Typical is the Gow structure - a crater cut into ridges of crystalline rock .

One of the first craters studied in detail (by Michael Dence) in terms of its petrography is Brent in Ontario. Here is an aerial photo of this 3.8 km wide structure, which is severely eroded:



Also in Ontario is the Holleford crater. It was imposed on Precambrian rocks that were later covered by sedimentary rocks. The trace of the Holleford structure is evident in this aerial photo:

East of the Nastapoka Arc are the two Clearwater Lakes structures, formed simultaneously by the breakup of the incoming meteoroid into two chunks. West Clearwater is a complex crater (32 km wide), with a circular ridge as a remnant of the central peak; East Clearwater is a simple crater (20 km diameter).

The 13 km wide Deep Bay crater, in Saskatchewan, appears in this ESA image:

Carswell is a 39 km diameter eroded crater in Saskatchewan; it is faintly visible in this Landsat

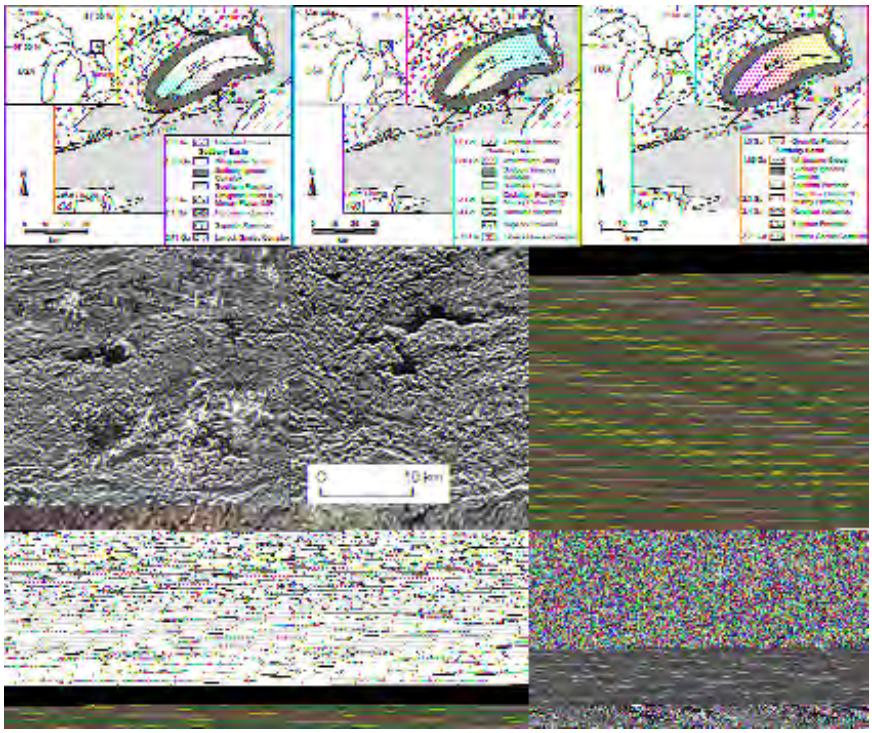
image:

A SIR-B radar image of southern Ontario highlights two juxtaposed but unrelated craters that are very different in age, in size, and in structural state.

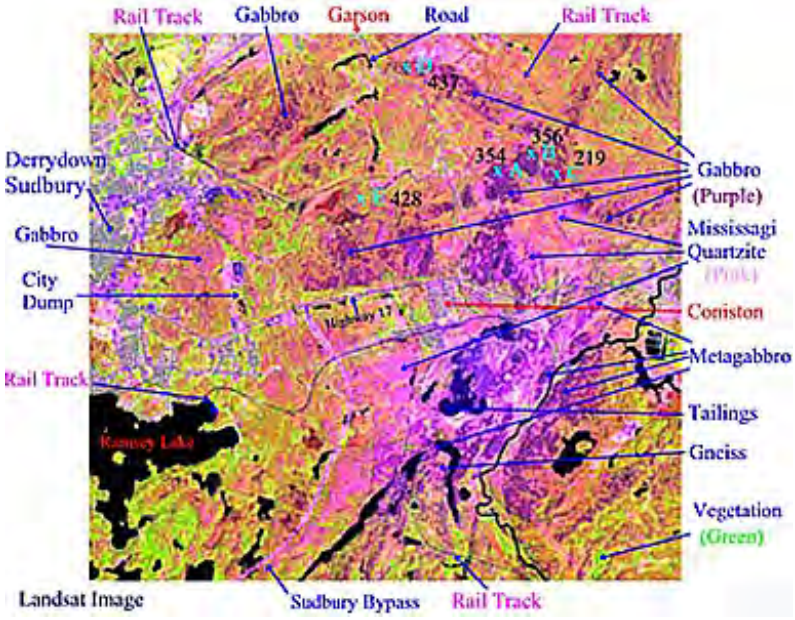
The partially circular lake-filled structure on the right (east) is the 8 km (5 mi) wide Wanapitei crater, estimated to have formed 34 million years (m.y.) ago. The far larger Sudbury structure (second largest on Earth) appears as a pronounced elliptical pattern, more strongly expressed by the low hills to the north. This huge impact crater, with its distinctive outline, was created about 1800 m.y. ago. Some scientists argue that it was at least 245 km (152 mi) across when it was circular. More than 900 m.y. later strong northwestward thrusting of the Grenville Province terrane against the Superior Province (containing Sudbury) subsequently deformed it into its present elliptical shape (geologists will recognize this as a prime example of the "strain ellipsoid" model). After Sudbury was initially excavated, magmas from deep in the crust invaded the breccia filling, mixing with it and forming a boundary layer against its walls. Some investigators think that the resulting norite rocks are actually melted target rocks. This igneous rock (called an "irruptive") is host to vast deposits of nickel and copper, making this impact structure a 5 billion dollar source of ore minerals since mining began in the last century.

18-13: Which part of the elliptical rim of Sudbury seems to have better topographic expression? ANSWER

A geologic map shows the general geologic units present within and around the main structure (just north of the town of Sudbury):

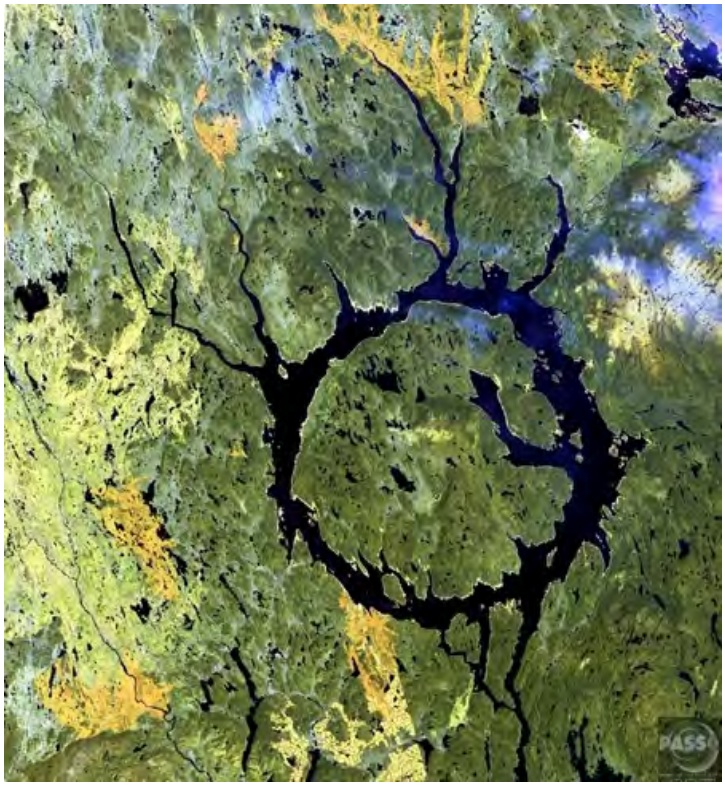


Part of this structure has been displayed in this Landsat image in which classification procedures indicate many of the surface constituents:

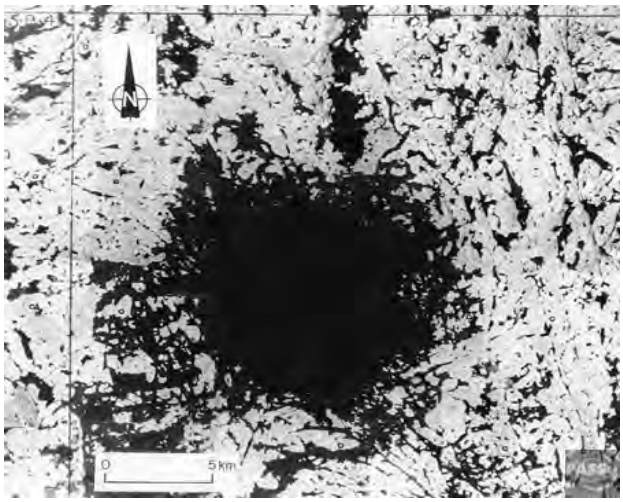


Because the Sudbury event led to invasion of magma that also brought huge deposits of nickel minerals, this structure has been extensively studied by many geologists. Here is a photo of typical Onaping breccia (originally named a "tuff" by Howell Williams) and beneath it of some of the melt rock at Sudbury:

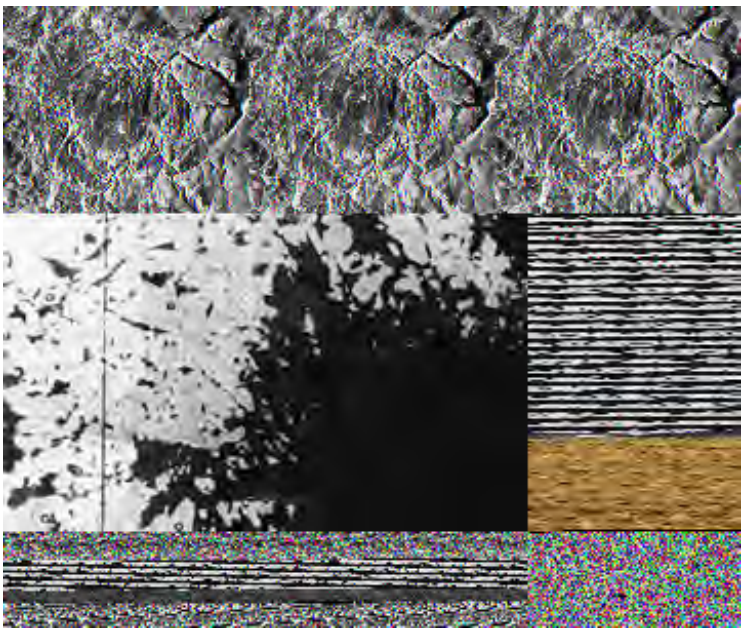
On page 18-2 we first saw Canada's most conspicuous impact structure, Manicouagan, in Quebec. Here is another version, made by Landsat-7, in which the lakes that define the structure appear to be part of an incomplete "8" with the upper half made (coincidentally) by smaller lakes:



Also in the Canadian Shield of Quebec is the 8 km wide Couture crater, seen in this aerial mosaic;



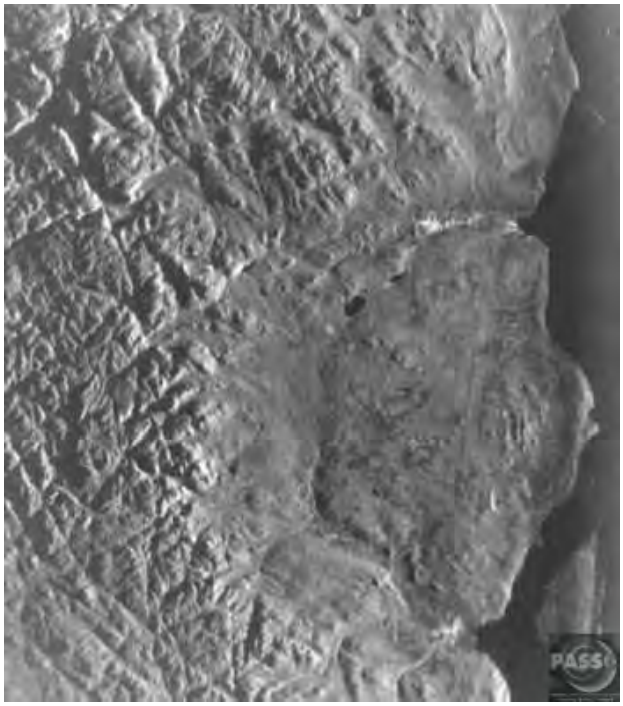
Radar can sharpen the appearance of an impact structure, as demonstrated with this aerial radar image of the Haughton crater (24 km; 15 miles wide) on Devon Island in the Canadian Arctic. Although about 23 m.y. old, much of the crater's morphology has survived erosion.



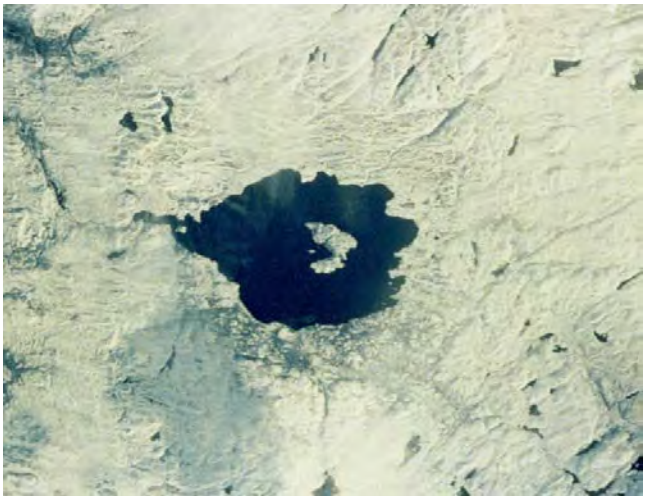
Compare the radar image with this one made by Landsat-7:

18-14: Where does the crater rim of Haughton appear to be? ANSWER

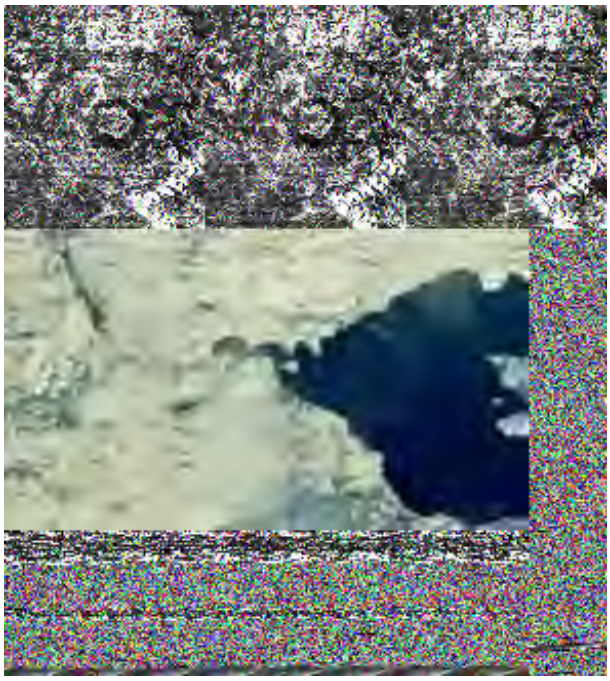
The Charlevoix crater is truncated by the St. Lawrence River in Quebec. This 54 km wide structure is imaged in this aerial radar composite:



Large craters are not always evident morphologically, especially where modified by erosion. The 38-km wide Mistastin impact crater in Labrador, seen in this wintertime astronaut photo, has a lake and central peak but glaciation has obscured its rim boundary:



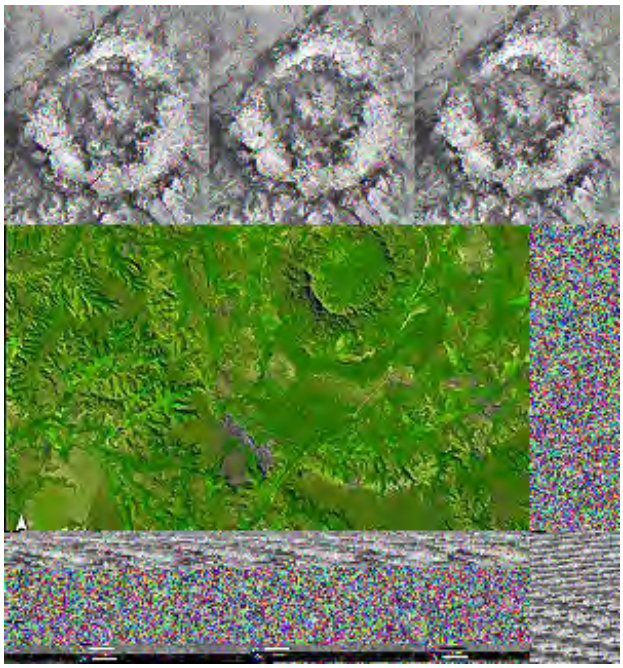
Several South American impact structures have a tie-in with Landsat and other imaging systems. A crater in Brazil named Araguainha had earlier been studied and classified as a dome. But when visited by Dr. Robert S. Dietz - reknown for his ability to find new craters - evidence (shatter cones and breccias) was found that pointed towards an impact origin. Samples were sent to Dr. Bevan M. French, a colleague, to search for shock metamorphic features. On the very same day he confirmed their presence, the writer (NMS) phoned him to say that I had found the following Landsat image, shown here as a subscene:



What we learned from the image was that the crater structure was about twice as wide (40 km; 25 miles) as field studies had suggested.The several tonal bands are due to differences in vegetation in this pampas grass country.

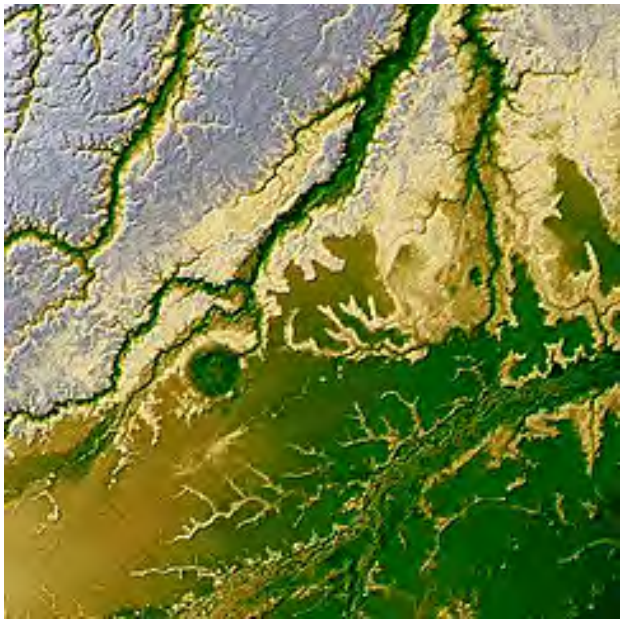
The hunt pressed on to find other craters in vegetated Brazil. Landsat was instrumental in finding this 13 km (8 mile) structure, Serra da Cangalha, with its central rim and inner depression.

Then more looking turned up a smaller crater (4 km), Riachao, about 50 km to the northwest. It is shown as the inset in the above image. An aerial photo taken of it later gives details about its appearance:

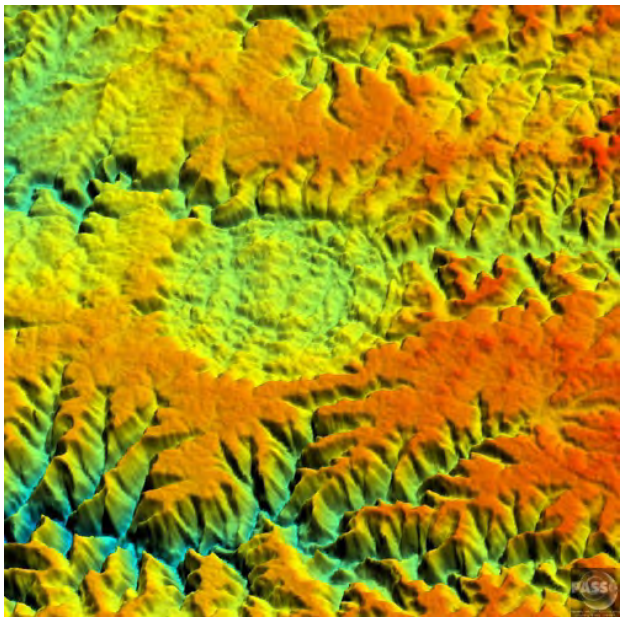


After being visited and sampled, both structures yielded evidence of shock metamorphism, putting them squarely in the impact camp.

Still another impact structure, the Iturralde crater (8 km; 5 mile diameter) was discovered from space just within the rain forest in eastern Bolivia:



Of probably impact origin, the Vargeao Dome, in the Brazilian rain forest, shows up in a topographic image made from SRTM (radar) data:



Lets move on now to impact craters in the eastern hemisphere.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Next to Earth, no other planet has fascinated people as much as Mars, the Red Planet, one and a half times the distance from the Sun as we. Although little was actually known about Mars before the Space Age, the notion that astronomers such as Percival Lowell saw markings on Mars that reminded them of canals led to a ficticious hope that this meant (human) life may be present. Mars figured in several classic science fiction stories, such as H.G. Wells "War of the Worlds". Then, in 1962 a space probe sent back the first pictures of the martian surface and the natural wonders of this planet soon outweighed those of imaginative writers. The next six pages will summarize both what we know and what we are still learning. We start with several good images of the planet in different half sphere positions and then recount the major trips to Mars by various spacecraft.

Mars, The Red Planet

We switch now to the region beyond Earth, as we explore the Mysterious Planet Mars, recently upgraded in interest because of the discovery of possible evidence of single-celled, organic matter (life?) in a meteorite (found in the Antarctic), believed to have been ejected into deep space from an impact on ancient Martian crust.

Mars has always been especially intriguing to the peoples of Earth. It has captured our imagination for many reasons, two of which are the once-believed speculation that it held intelligent life because of so-called "canals" and then the famous book by H.G. Wells, "The War of the Worlds", that had martians invading Earth (this was the basis for the "infamous" radio program by Orson Welles that was so realistic that people panicked during the broadcast). Mars is about half the size of Earth, evident in this picture in which the two planets are shown side by side at the same size scale:

The writer (NMS), a geologist/planetologist, has found (through preparing this subsection) Mars to be, acre for acre, the most intriguing and diverse planetary surface of any of the Solar System planets - including Earth. This is due in part to the 100% exposure of a wide variety of geologic features to observation owing to lack of vegetation. But the variety of landforms and process indicators of tectonic, volcanic, impact, water, and wind action are the chief reason.

After the Earth and its Moon, Mars has become the most intensely studied Solar System body because of its similarity to Earth itself (but as a planet that has lost most of its water) and therefore is a candidate for once or still active organic (biologic?) material. More missions have been sent to Mars than to any other planet; these are summarized on this JPL website. Spacecraft today are orbiting it, surface rovers are exploring it onsite, and NASA - as well as the whole world community - is hoping for enough evidence of its special features to warrant a long-term program culminating in Man's setting foot on its surface. Another reason for the attention given to Mars is that it has many features explainable by extrapolation of geologic studies of Earth (and some features without terrestrial counterparts) -- and these are easily "seeable" because of total absence of oceanic and vegetation cover.

Here are four Web sites to consult if you want a quick overview of Mars or images to download: (1),(2), (3), (4). A splendid synopsis of Mars Exploration is found on this Wikipedia website; it lists 42 missions to the planet of which 20 were successful and 3 achieved partial success.

Also there is a NASA JPL video that describes how geologists use comparative planetology methods (based on familiarity with the Earth) to study, in this case, Mars. The speaker, Dr. James Garvin, is now Project Scientist for Mars Exploration at NASA Headquarters. Access it through

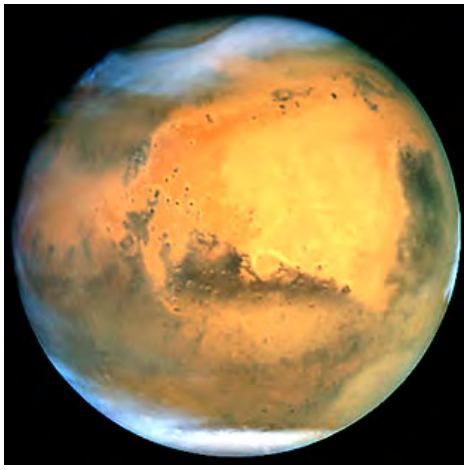
the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes "Finding Mars on Earth", December 2, 2003. To start it, once found, click on the blue RealVideo link.

Mars's mean diameter (6780 km [4875 miles]) is 0.53 that of Earth, but about double that of the Moon, and has about 1/10th its mass; its volume is about 1/8th that of Earth. The planet, fourth from the Sun, has has a mean density of 3.93 g/cm^3 compared with Earth at 5.52 g/cm^3; it may have a much small inner core that would be iron-poor, as borne out by the lack of a distinct magnetic field.

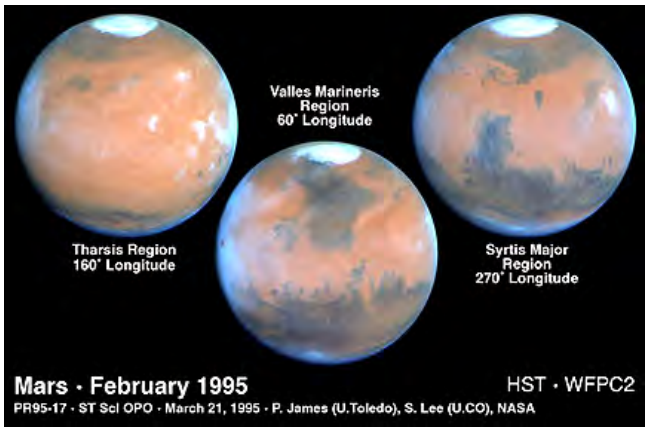
A martian day is almost the same as a terrestrial day - it rotates once every 24 hours and 37 minutes (Earth time). Mars takes 1.86 Earth years (685 days) to complete its orbit around the Sun. Its orbit is distinctly elliptical (1.38 A.U. perihelion to 1.67 A.U. aphelion; average distance of 1.52 A.U.). Its rotational axis is tilted 25°, similar to Earth, which causes Mars to have summer and winter seasons. Mars surface temperatures range from -140° to +20° C.

Look below at a beautiful full-face image of a Martian hemisphere, which should make it obvious why Mars is also known as the Red Planet. This mosaic was constructed from reprocessed Viking images. Two conspicuous features, described later, are the huge gash across the face, known as Valles Marineris, and the three great volcanoes in the Tharsis group, on the left.

Compare this with a single (not mosaicked) image of a face of Mars, taken through the Hubble Space Telescope.

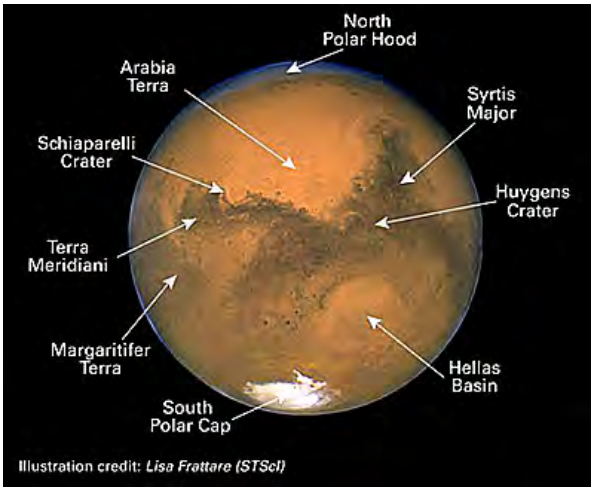


Then, look at these three hemispherical views of Mars taken by the electronic camera system on the Hubble Space Telescope (HST) through filters that allow close approximations to true color. The blues along the limbs are somewhat artificial. Note the presence (white) of material at the poles: these are the polar ice caps (water and carbon dioxide ice).

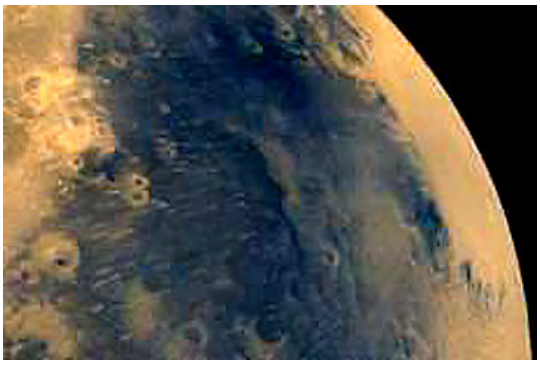


19-33: From these three full face views, what would you deduce about martian rocks?
ANSWER

In late August of 2003, Mars became a "hot item" in TV and Newspaper accounts. Mars' orbit is such that most of the time it lies beyond 160 million km (100 million miles) from Earth. But there are rare times when the two planets positions in their orbits place them much closer together. On August 27, 2003 Mars and Earth were at the closest distance (55,690,000 km or 34,500,000 miles) apart. This had not happened until about 60000 years ago but will happen in about 600 years into the third Millenium. Thousands of amateur telescopes, 100s of Observatory telescopes, and the Hubble Space Telescope all trained on the Red Planet (which was just less than Venus as the brightest object in the Sky) and many photographs and electronic images were taken. The one below was made by the Hubble Space Telescope in orbit :

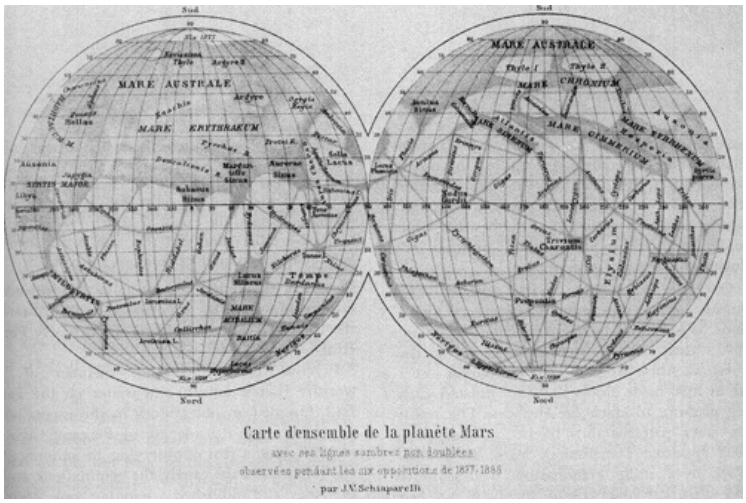


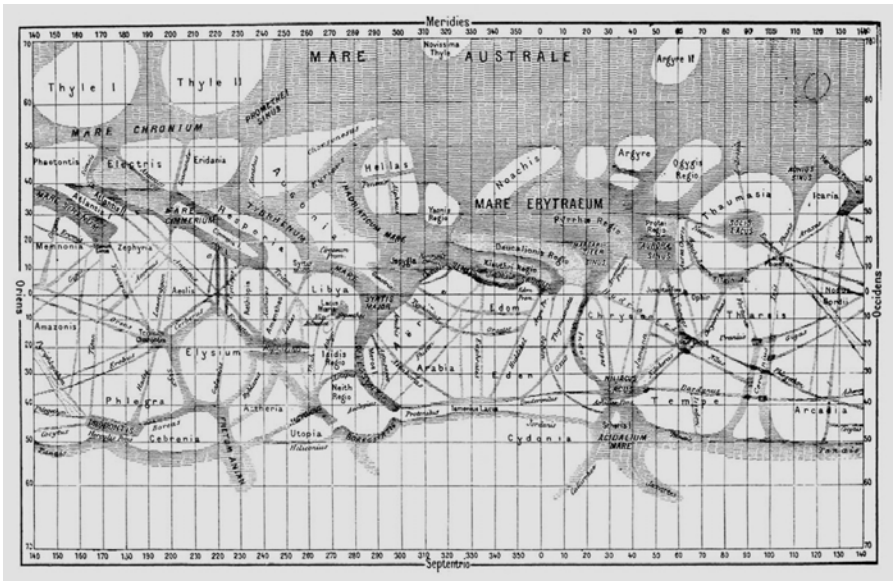
As we shall see on subsequent pages, while volcanic rocks predominate on Mars, there are widespread sedimentary rocks - mostly evaporites and some carbonates. The overall surface state of Mars, judging from these full planet face views, is that it consists of three major color states, each tied to some dominant material: 1) the predominant reddish colored surface, which is now known to be regions in which the rocks, soils and dust are strongly oxidized into phases consisting of hematite, and possibly maghemite (the γ polymorph of Fe_2O_3) and similar minerals (including the group of hydrated iron oxides going under the name of "limonite", if water at and beneath the surface has "weathered" the hematite); 2) the dark bluish to blackish surface, presumably basaltic bedrock with less iron discoloration, and 3) the whitish areas around the poles, identified as a mix of water and carbon dioxide (the outer coating). Color phase 2) implies that the surface is volcanic bedrock with insufficient iron oxide dust cover to significantly alter the color depicted; this suggests much less onsite alteration of the basalt and the transient nature of dust cover as martian winds remove much of previously deposited red dust (but some may be cyclically deposited during strong dust storms [see next page]). As an example of this dark color phase, look at the Syrtis Major physiographic region. It probably is volcanic crust, likely basalt. It size, however, varies from time to time because winds carry red dust back and forth over its boundaries.



The iron rust responsible for the reds and oranges on much of the martian surface is believed to have largely formed early in the history of the planet when its atmosphere may have been different and water in the air was more abundant. In the last billion years or so, the martian atmosphere has been cold and dry. Under these conditions UV radiation bombardment causes a thin top layer in the soil and rock enriched in the oxygen anion ("superoxide", an oxidant) which is capable of further release of iron from mineral by oxidation. However, this is a much slower process and probably doesn't account for most of the iron coloration. One consequence of superoxides is to create an environment that is anathema (destructive) to bacteria and other microbes. This would mean that life right at the surface cannot tolerate these conditions and, if present today, must dwell 10 centimeters or more below the topmost soil layers (life possibilities on Mars is discussed on page 19-13).

Mars was the first extraterrestrial planet mapped in some detail, solely from telescope observations. Thus, in the 1870s, Giovanni Schiaparelli published the map below, in which many of the features he named are still known as such in today's nomenclature. Although some names are illegible in the figure below, we reproduce two versions of this map for its historical importance:



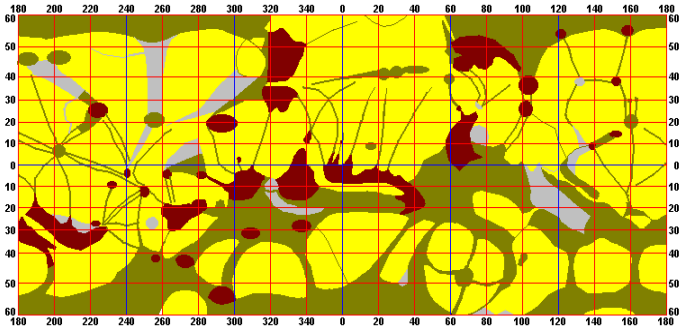


Schiaparelli called attention to long linear features that he "thought" he could see. Schiaparelli referred to them as "canali" to mean "channels" (of non-human origins), which is the normal meaning in Italian but the English world mistranslated the word into "canals" a la Venice. This led to the widespread belief persistent into the second half of the 20th Century that there might be

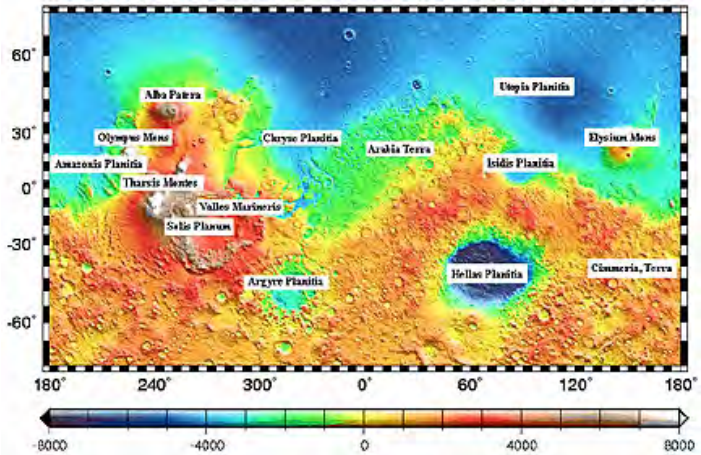
canals on Mars used by its "inhabitants" to transport water. Thus was born the legend of Martians (e.g. H.G. Wells "War of the Worlds" brought these creatures by spaceship to Earth, only to succumb to the adverse biological conditions on our planet). These canals, of course, disappeared when the good views shown above were obtained.

Sir Percival Lowell, using a telescope he built on a hill in Flagstaff, Arizona, extended this mapping and popularized the notion of canals on Mars. His 1909 map depicts a large number of nearly straight lines. Two versions are shown:

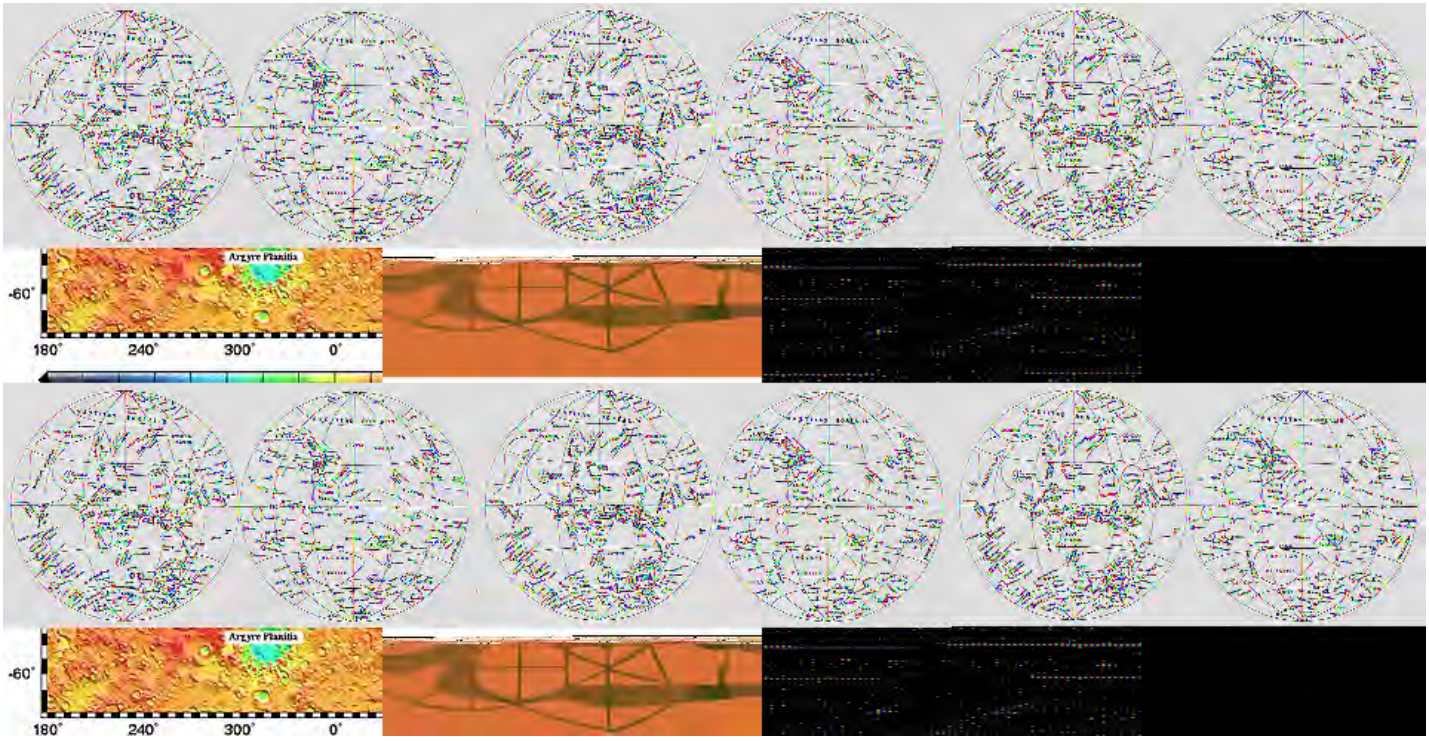
As late as 1940, improved maps of Mars still showed the (fictitious) straight lines on Mars



The probes sent to Mars finally dispelled the martian canal fantasy. It is difficult to reproduce on a Net site a modern map that shows all of Mars with many of its features identified. We will try with this simplified version:



A more detailed version:



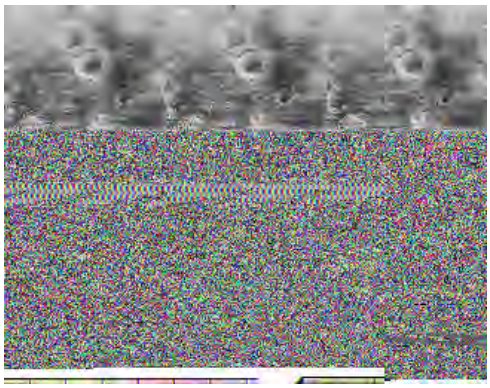
We recognize the difficulty of familiarizing yourself with so many feature names and their locations. However, from this list find the following: Olympus Mons; Arsia Mons; Pavonis Mons; Valles Marineris; Alba Terra; Arabia Terra; Noachis Terra; Argyre Planitia; Chryse Planitia; Hellas Planitia; Elysium Planitia; Syrtis Major; Terra Cimmeria.

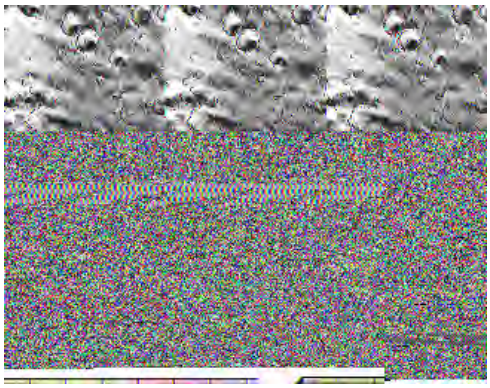
We now know a great deal about Mars from flybys, orbiting surveyors, and landers. The tricky problem of operating in near realtime makes a spacecraft's orbital insertion around or landing on Mars difficult, owing to some extent just to its distance; this has led to a number of mission failures. But, this list documents the successes achieved in the NASA programs for exploration of Mars: **Mariner 4** (1964) --> **Mariners 6 & 7** (1969) --> **Vikings 1 & 2** (1976) --> **Mars Global Surveyor (MGS)** (1996) --> **Mars Pathfinder** (1997) --> **Mars Odyssey** (2001) --> **Mars Exploration Rovers 1 (Spirit) & 2 (Opportunity)** (2003) --> **Mars Reconnaissance Orbiter (MRO)** (2006)--> **Mars Polar Lander** (renamed **Phoenix** (2008). ESA, the Soviets, and Japan has sent missions to Mars. If you want to see a complete chronology of all trips to Mars, check out this List of Mars missions.

At the top of the list of what space probes and landers have discovered about the Red Planet is the evidence that Mars has water now in limited abundance and seems to have had a lot more in its past. Today, most of that water is in the polar ice caps and in a thin permafrost zone in martian soils. In earlier times, Mars had sufficient water to carve out canyons and gullies and perhaps even support an ocean. Sedimentary units (distinct bedding) resulting from sediment deposition have been observed over much of the planet. A more extensive atmosphere than now exists probably held water under warmer conditions than are current. However, over time much of the water has been lost to outer space by the action of the solar wind (Mars has a metal core which seems to have solidified so that the shielding magnetic field it once may have had is no longer operative). Surviving water is largely locked up in ice. But the fact that there still is some water, and even more existed earlier, inspires cautious optimism that some signs of life - most types need water - may yet be found.

On this page we review the result of the first missions to Mars. The first flyby, by Mariner 4, on July 14, 1965, produced 22 TV-like images covering roughly 1% of the planet. The very first of these is:

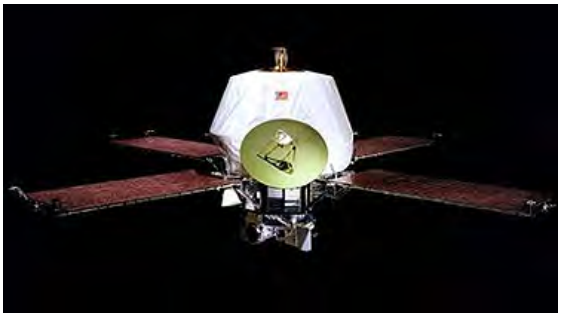
Here is another of those TV views and below it is almost the same area covered by the TV camera aboard Mariner 9 (see below) six years later - a good measure of technological progress:





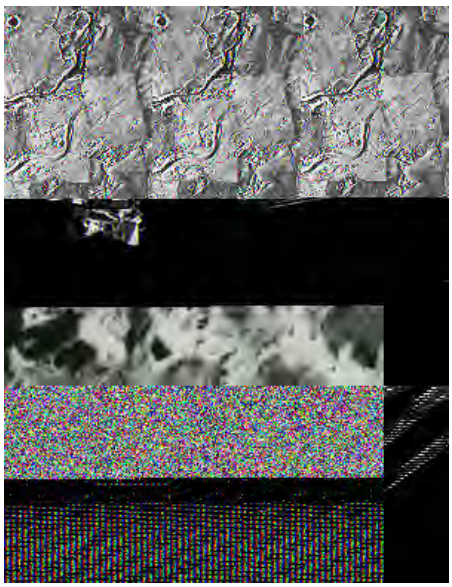
Mariners 6 and 7 passed Mars on July 31 and August 5, 1969, together obtaining 199 images that extended coverage to about 10% of the total surface. A typical Mariner 6 image shows frost-covered parts of the south polar region, with dark craters, furrows, and pits.

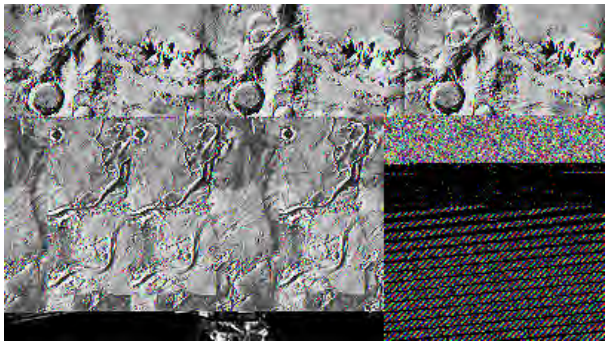
A quantum leap in coverage followed the first successful orbiting of another planet, when Mariner 9 arrived on November 13, 1971. Returning more than 7,300 panchromatic images, a wide-angle TV camera, capable of 1-3 km (0.6-1.9 mi) resolution, mapped almost the entire surface, while a narrow-angle TV camera imaged some areas at 100 m (328 ft) resolution. Mariner 9 also carried an IR radiometer, an IR interferometer spectrometer, and a UV spectrometer. Here is an artist's depiction of this pioneering visitor to the Red Planet:



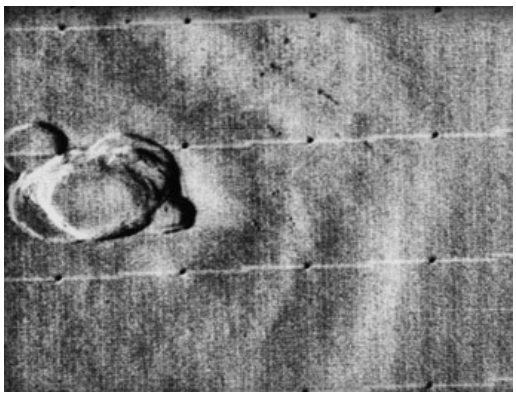
The imaged surface qualifies Mars as one of the most diverse and spectacular planetary bodies in the Solar System. Lacking vegetation and water cover, the easily seen geologic features are often grandiose in scale and sometimes unique. As with most extraterrestrial objects, their interpretation by astrogeologists started by comparing or contrasting features with those known on Earth. Martian features often required innovative explanations.

Two typical Mariner 9 views, as mosaics, show fractures, erosion (stream?) channels, and craters.





One of the grand surprises revealed during Mariner 9's passage was the largest volcano in the Solar System - named Olympus Mons - which previously had only been a suspicious bump in telescope photos. Shown here is one frame, in raw unprocessed format, which displays a part of this huge volcanic structure; a mosaic made from three additional frames is shown on page 19-12.



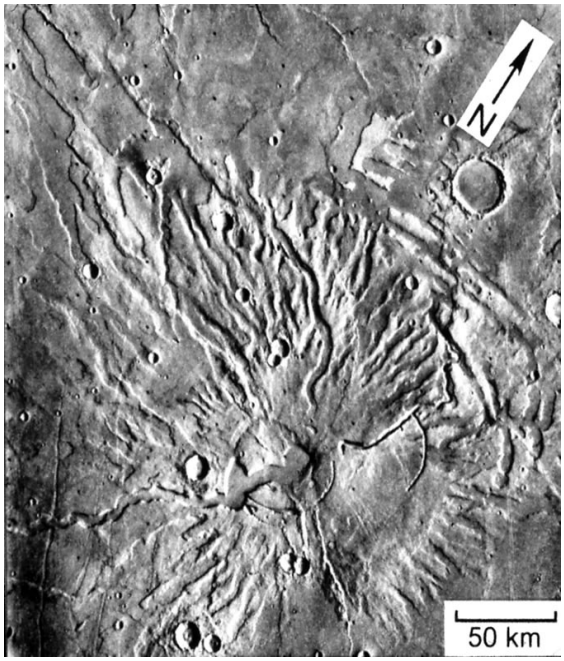
The Russians launched Mars 2 and 3 probes to that planet in 1971, but acquired only limited data, mainly on the magnetic field. In 1973, the former Soviets sent three more spacecraft to the Red Planet. In March of 1974, Mars 4 failed to orbit the planet, but Mars 5 did, long enough to return data, including images during 10 of its 20 orbits. Mars 6 successfully descended through the thin Mrtian atmosphere but failed during the last second before touchdown.

A quantum leap in the U.S. exploration of Mars began on August 20, 1975, and September 9, 1975, with the launches of Vikings 1 and 2 respectively. Their Orbiter components began circling the planet in late June and early August of 1976, coinciding roughly with America's Bicentennial Celebration. Each carried two identical vidicon (TV) cameras, capable of color imaging (from black and white images taken through color filters) and acquiring stereo scenes at 100 m (328 ft) resolution. They also bore a thermal-mapping infrared spectrometer and a second IR spectrometer designed to detect atmospheric water.

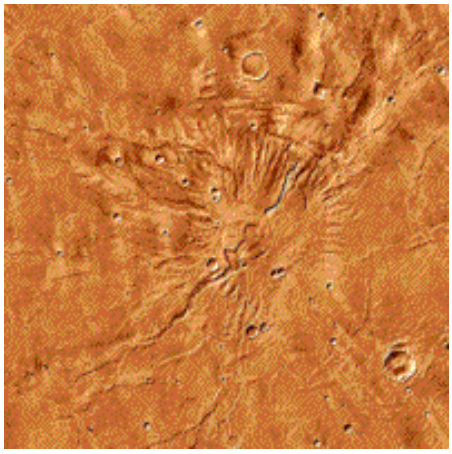
The aforementioned Mars JPL video includes a segment dealing with Viking. Access through the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes

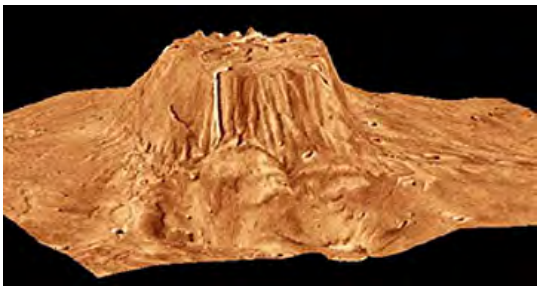
"Viking Legacy and Mars Exploration", July 12, 2001. To start it, once found, click on the blue RealVideo link.

As an example of a Viking image, consider this Viking 1 view of Tyrrhena Patera, a volcano built up from both flows and abundant ash, located near the Hellas Basin in the southern hemisphere.



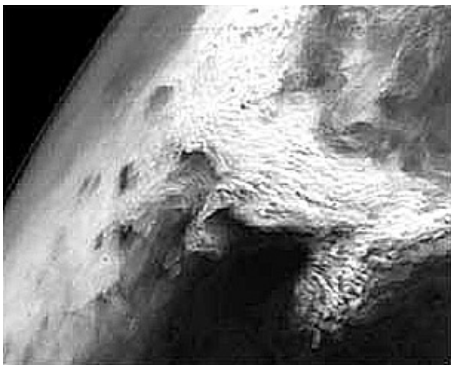
The two Viking images below show color versions of the Volcano Tyrrhena. The first image does not give a sense of relief but when we confound topographic information with this image, a perspective view, marked by strong vertical exaggeration, gives a sense of its height:



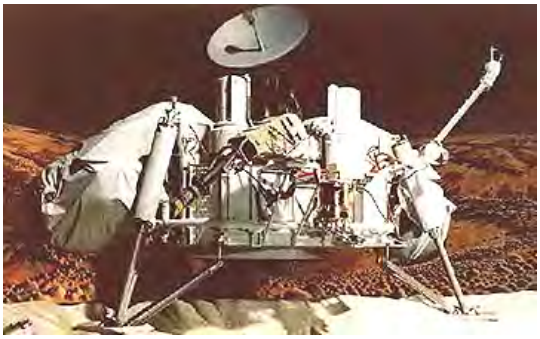


19-34: Critique the above two views of Tyrrhena, specifically in regards to misconceptions they might give about martian topography. ANSWER

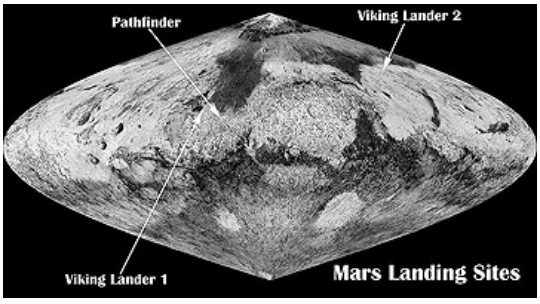
As had been observed earlier in the Mariner 9 mission, the martian surface and atmosphere are actually dynamic. This Viking-2 image shows a large dust storm:



Both Vikings also transported separable landing vehicles that descended by parachute throughout the atmosphere. Here is an image of Viking Lander 1:

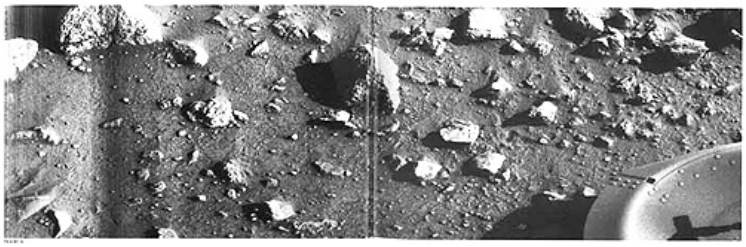


This map shows the two Lander sites and also that of Pathfinder (see below):

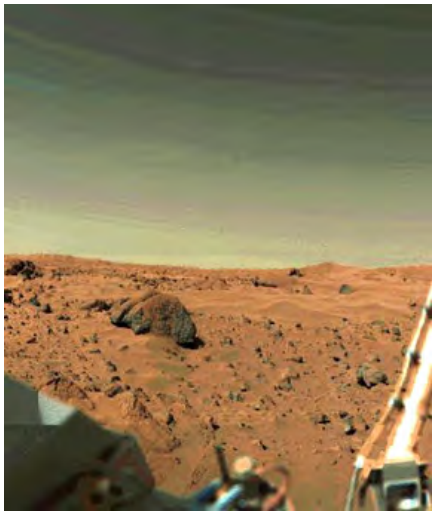


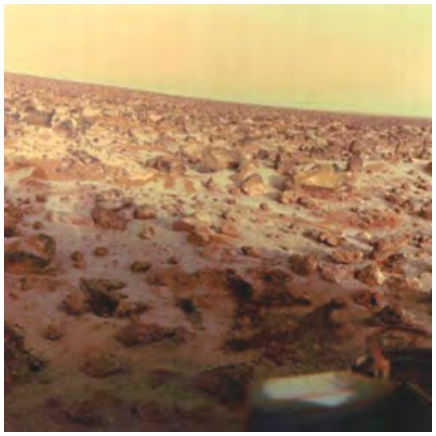
Viking 1 Lander touched down on July 20, 1976, and Viking 2 Lander on September 3. As with the Moon, a chief concern during the touchdowns was the nature of the surface which could be a depression (e.g., a crater wall) or covered by scattered boulders and other debris such as seen here in a view of one of the sites obtained by a later orbiting satellite.:

So far, most landings have avoided these pitfalls (the loss of Beagle2 from the Mars Express mission is unexplained). The first ever image (black and white) of the martian surface from the surface is this:



Both TVs also successfully transmitted surface views in color, of the immediate ground, up to the visible horizon.

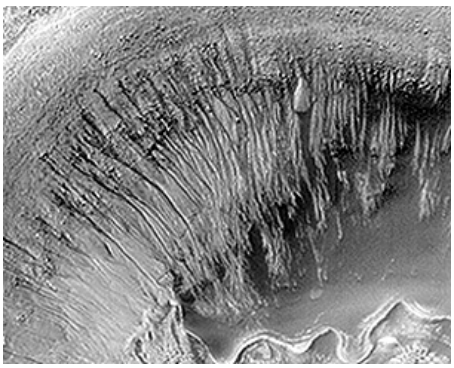




The two Viking Landers were exceptional ground-truth devices. These panoramas from them reveal a red iron-coated surface, consisting of rocks (the largest > 2 m [6.6 ft]) mixed with sand and dust, some piled into incipient dunes. In other views, some rocks look like vesicular basalt. Note the yellow-brown tones in the lower, dusty layers of the sky's rarified gas envelope (mainly carbon dioxide). This stationary observatory was also equipped with magnets, a seismometer, an X-ray fluorescence spectrometer (that receives samples from a movable scoop), a miniature meteorological station, and three experiments seeking signs of metabolic action (mainly, as gases released or absorbed) by biogenic matter (none found).

19-35: Using the left (or upper) of the Viking surface pictures, comment on the conditions above the martian surface. ANSWER

After the Vikings, the resumption of martian exploration was vested in the Mars Observer but this spacecraft failed enroute to Mars in 1993. The next visit was by the Mars Global Surveyor, launched in 1996, which orbits the planet taking pictures up to 3 meters in resolution.Three of its prime instruments are the MOC (Mars Orbiting Camera), the MOLA (Mars Orbiter Laser Altimeter), and TES (Thermal Emission Spectrometer). This spacecraft is described on page 19-13 but as a preview sample here is an image that shows at high resolution the interior wall of a small impact crater within the very large crater Newton; the series of channel strips running down the side has been interpreted as evidence for subsurface water emerging at some depth along the wall and eroding these rills (see page 19-13 for details).



JPL's Pathfinder mission landed a Rover on the surface in 1997 to conduct a series of physical and chemical experiments (page 19-13). Then, in October 2001, the Mars Odyssey mission

reached the planet and is continuing to gather data with THEMIS, its thermal emission spectrometer, and GRS, a Gamma-Ray Spectrometer, and other instruments. In the first years of the 21st Century, several more U.S. and International spacecraft, both orbiters and landers, have been sent to Mars. The successful ones - Mars Express, Mars Reconnaissance Orbiter, the Mars Exploration Rovers Spirit and Opportunity, and Mars Phoenix - are described on pages 19-13ff. Prior to that, selected images from these missions displaying specific features appear in the next two pages.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

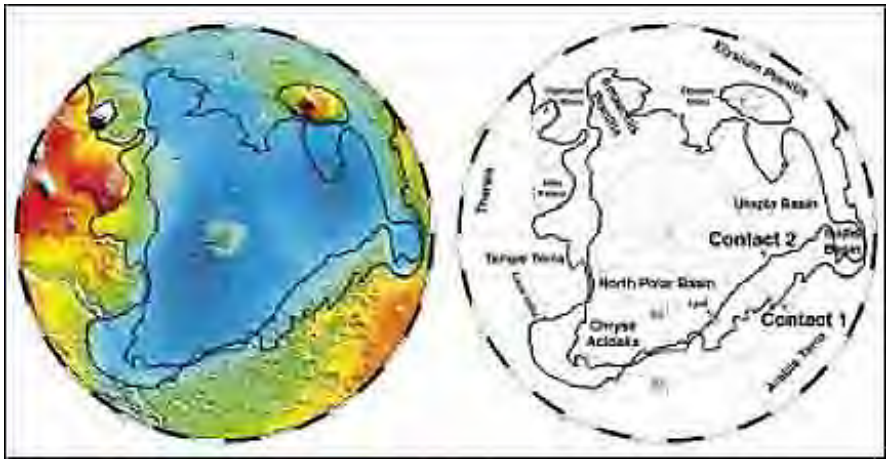
The general geology (stratigraphic units) and physiography (landform types) of Mars are examined. Some types do not have close counterparts on Earth. Although the atmosphere of Mars is thin, high velocity winds stir up loose surficial materials producing dust storms and aeolian deposits such as dunes. The martian polar icecaps have both conventional and unusual features including layering and collapse structures. Features which may represent more widespread glaciation over much of Mars are discussed.

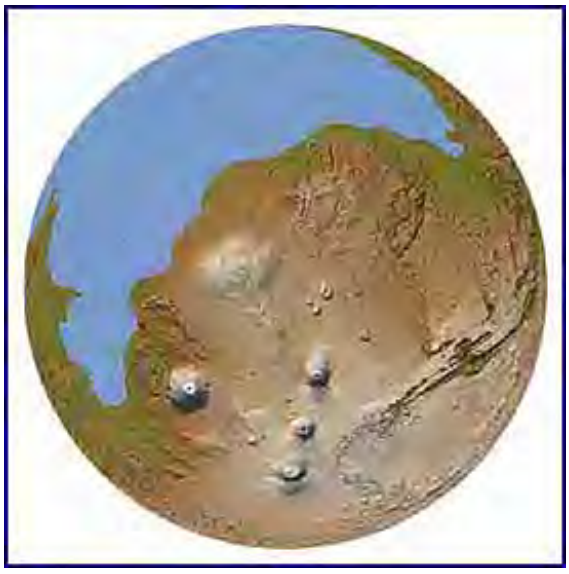
Geology of Mars; The Martian Atmosphere; Ice at the Poles

Stratigraphic Units Maps

As on the Earth, layering - which gives rise to stratification - can be caused by several processes. On Mars, layering from dust deposits laid down in the present and by inference in the past and layering from impact ejecta build-up (probably the dominant process) are established mechanisms that account for perhaps the majority of stratified sequences. As we will see on this page, there is now strong evidence for the action of water from streams and other outpouring mechanisms operating in the martian past.

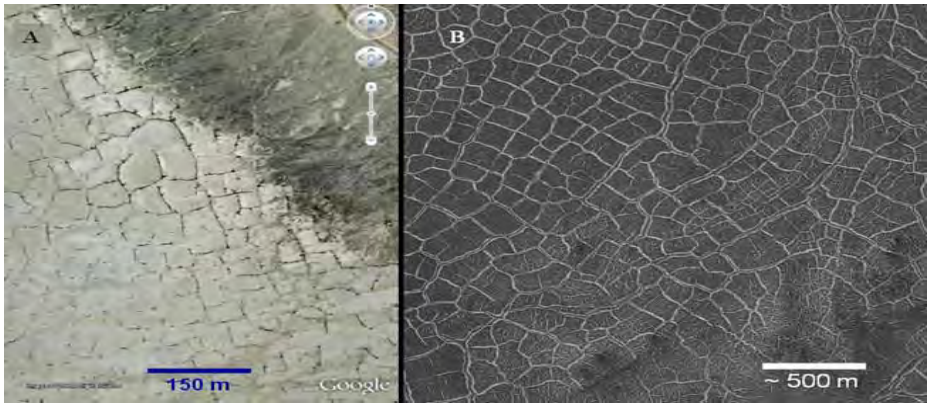
Some students of the martian surface argue that at one or more time(s) in the past, there were lakes and perhaps even oceans on Mars that existed when the martian climate was warmer than today - gone now because most of the water has either dispersed into outer space or become frozen in the upper reaches of deposits near the surface. James Head III and his colleagues at Brown University have interpreted evidence, including low elevations, drainage patterns, and cliffs and terraces that may indicate shoreline erosion, of at least two stages of oceanic concentrations of water in the northern lowlands. (Needless to say, their proposal has stirred up disagreements.) Here is their map and a pictorial showing the emplaced ocean:





Studies reported in 2008 present evidence from terrain analysis around Valles Marineris that the period in which water was present on the martian surface may have lasted until about 3 billion years ago. The water was responsible for producing Valles Marineris and most other channels. But, this claim is moot as to whether the water could (or did) collect in a standing "ocean".

There is greater certainty that ancient water on Mars had collected in lakes. This next pair of images show dessication cracks in terrestrial lake beds on the left and similar features on Mars on the right. The spacing of these cracks is too wide to be accounted for by ice phenomena.



One of the best indicators of the action of water on Earth is depositional layering (note: layering can also be caused by multiple deposits of granular materials by the wind). As will be developed on page 17-13, layering is widespread on Mars. As a preview, note this Mars MOC color image of layers in West Arabia Terra:

The occurrence of stratified rocks is usually the decisive evidence that water had once collected into bodies of standing water. Another example shows a typical set of layers exposed in West Candor Chasm, an auxiliary canyon off Valles Marineris:

Various explanations are put forth as to their nature: Are they marine sedimentary, lake beds, wind deposits, volcanic flows, volcanic ash deposits, or other unknown types, or as seems likely a combination of these modes of layering? The occurrence of *marine* sedimentary layers on Mars is still much debated, with as yet no direct proof. But, like the Earth and its Moon, layered materials found on the martian surface follow the Law of Superposition, which defines the relative ages of overlapping deposits. And, like the Moon and some other planetary bodies, the relative densities of craters have been counted over its surface, which has led to a broad stratigraphic division of the principal geologic units that are recognized on Mars:

Stratigraphy of Mars

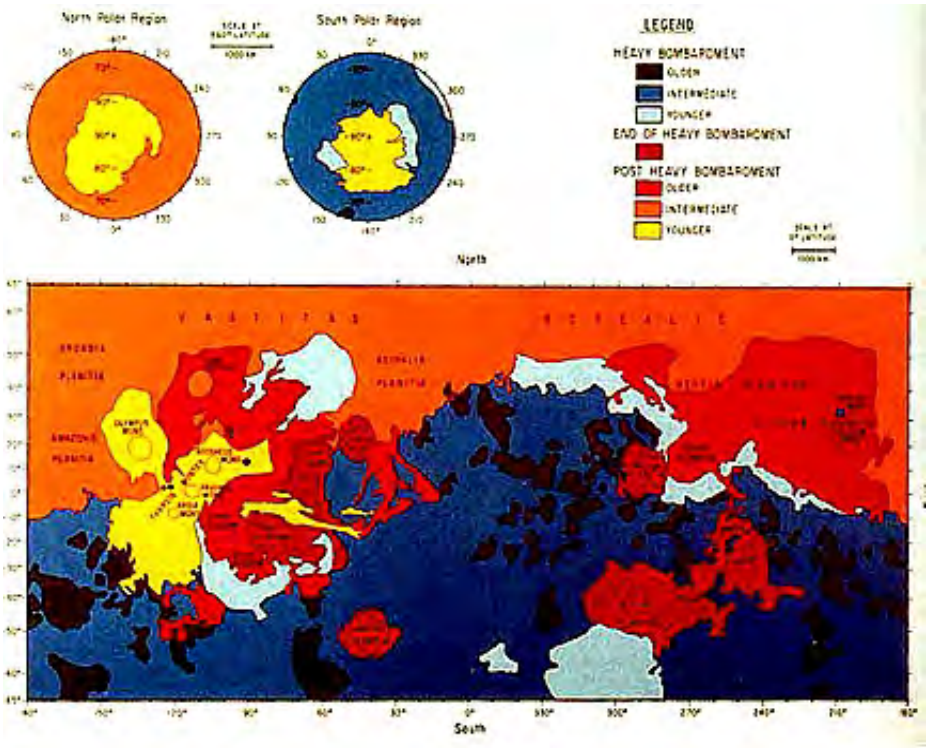
Martian stratigraphy is based on photogeologic units, relative ages derived from cratering density and a partial absolute time scale calculated using comparisons with the Lunar collision record.

Series	Crater Density (N5)	Absolute age Ga
Upper M. Amazonian	< 25	< 2.3-2.9
Lower	25-67	
Upper L. Hesperian	67-125	2.9 - 3.7
Lower	125-200	
Upper M. Noachian	200-400	> 3.7
Lower	> 400	

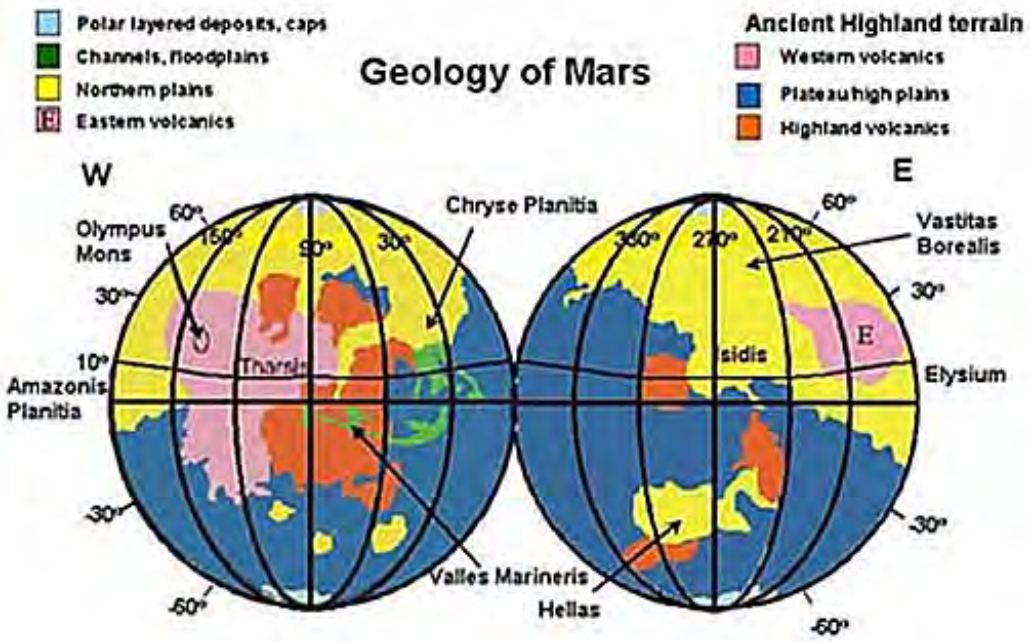
Data from Hartmann and Neukum, 2001.
N5 means crater density > e.g. 25-67 per $10^6/km^2$ for 5km diameter craters

This subdivision yields three broad Eras. The oldest, the Noachian, extends from the formation of the planet to a time estimated to be 3.8-3.5 billion years ago. Crater-forming bombardment was maximum during this time, producing Hellas and other basins. Both extensive lava flows and perhaps a broad regional ocean developed in this time, and the dichotomy of older high plains in the southern hemisphere and lower, more lava-covered plains in the northern hemisphere began in the Noachian. The Hesperian Era lasted from the 3.5 b.y. time period until about 1.8 billion years ago. Crustal fissures allowed widespread volcanic flows to emplace. The Elysium volcanoes formed and those in the Tharsis region began to develop. Again, water may have been active in modifying the landscape. It probably acted upon a large fissure that grew into Valles Marineris. The Amazonian Era continues to the present. Water erosion/deposition may still have played a role but wind now was the principal agent of surface change. During this time Olympus Mons grew as did the Tharsis volcanoes.

The global surface of Mars can be displayed on a map in terms of these relative crater densities, which like the Moon, provide (along with characteristics of volcanic surface) the principal means for distinguishing age units, as shown:



A general "stratigraphic units" map of Mars , produced by standard photogeologic methodology, is shown below along with a units key :

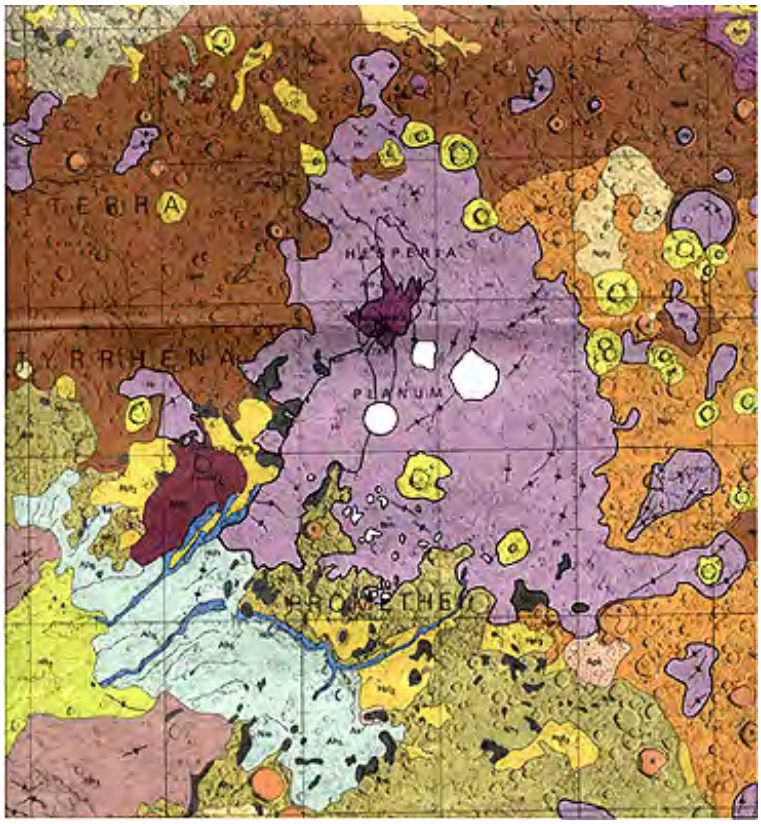


Photogeological units

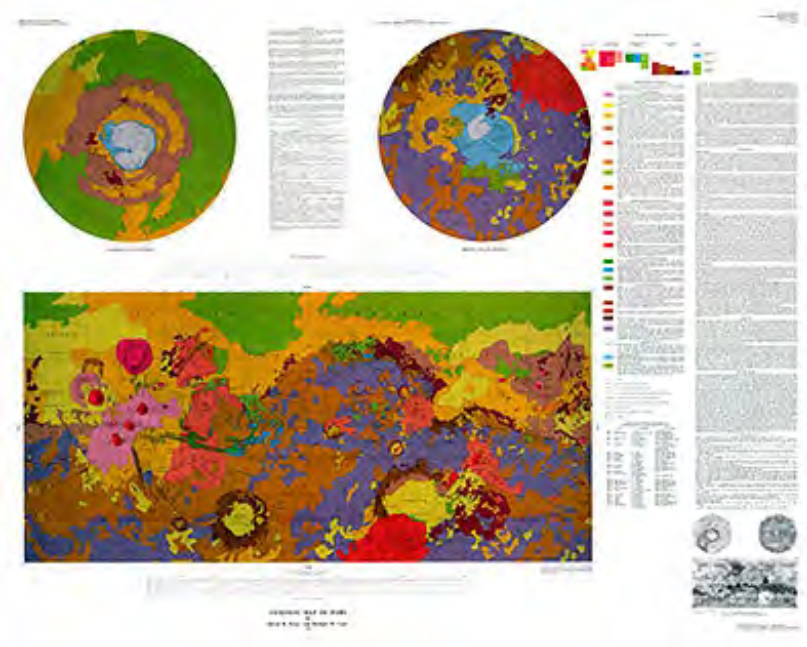
Geological Unit	Series, absolute age
1. Dunes, polar deposits	Amazonian, < 2.9 Ga
2. Channels, floodplains	U.Hesperian-U.Amazonian, < 3.7 Ga
3. Lowland terrain	
(a) Northern plains (lavas, dune fields)	U.Hesperian-U.Amazonian, < 3.7 Ga
(b) Eastern volcanics	
4. Highland terrain	Noachian-Amazonian, < 4.5 Ga
(a) Western volcs. (Arsia, Olympus Mons)	
(b) Hellas assemblage	
(c) Plateau/High plains (large craters)	
(d) Highland volcs, deformed terrain	
5. Other (not shown)(impact crater, ridged plains etc)	Noachian-Amazonian

(from Tanaka et al., 1992, Hartmann and Neukum, 2001), absolute ages based on cratering density

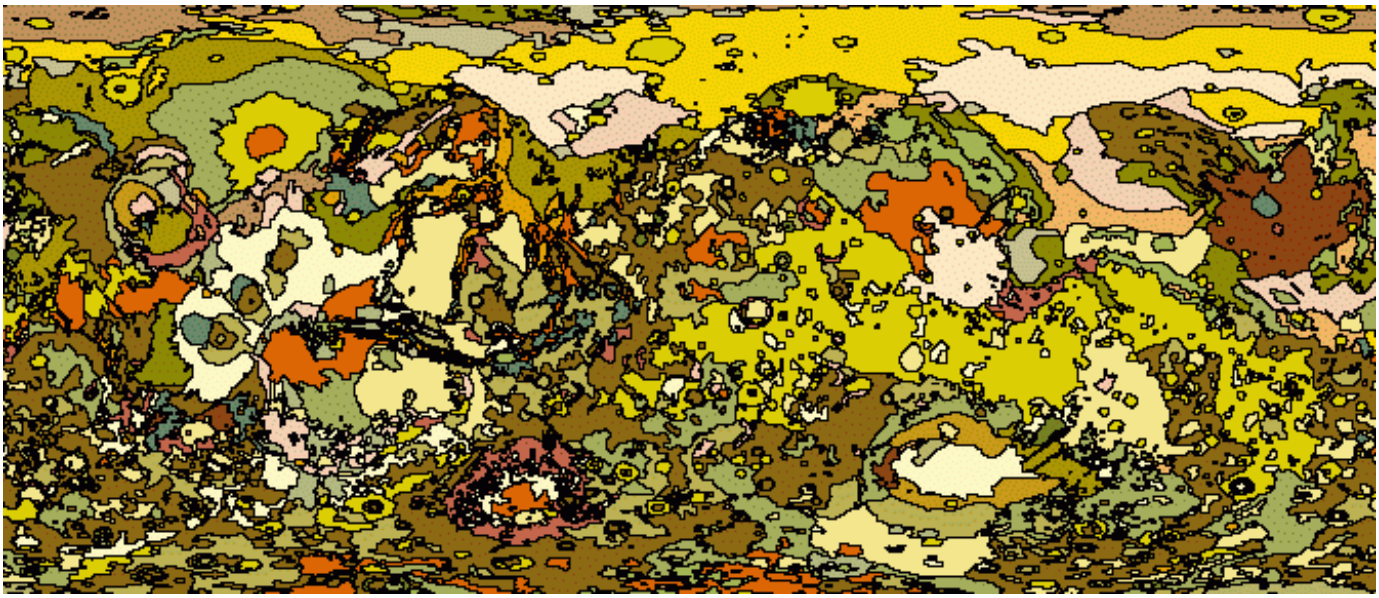
The U.S. Geological Survey has produced quadrangle maps of most of Mars by now, as illustrated by this one which shows the Hesperia Planum and Tyrrhenia Terra regions of Mars.



In addition to the quadrangle maps produced by the U.S.Geological Survey (see above), the USGS has now prepared maps of the entire planet at several scales. The best known of these, at 1:25,000,000, is shown in an overview image below, but none of the text or legend are legible in this version. You can track down a readible version on the Internet by clicking on Mars map kept online by the Lunar and Planetary Institute (note: it is a slow load).



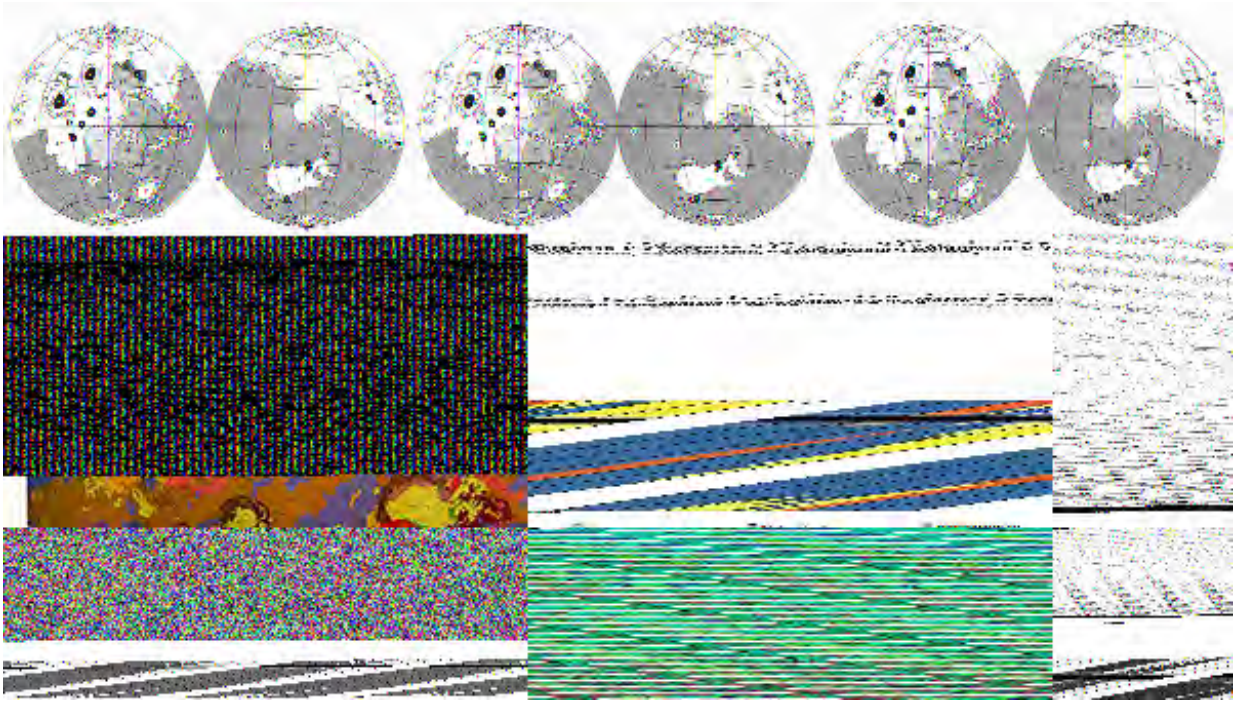
Here is a variant of most of that map but with a different color scheme, shown without a legend or annotation:



It should by now be evident that Mars has a diverse and intriging history, with a wide variety of surface features. These are examined in pages 19-12ff.

Mars Physiography

Before looking at some representative images of the Red Planet, we set up a physiographic framework for Mars: From Earth-based telescopic observations, astronomers suspected Mars had oxidized, iron-rich materials at its surface, was subject to dust storms, and had ice caps at both poles that alternately expanded and contracted over a martian year (687 Earth days). The storms implied at least a thin gaseous envelope (out to about 125 km). Spectroscopic measurements indicated CO_2 and maybe nitrogen. The Mariner/Viking missions greatly modified our concepts of the martian landscapes. Consider the generalized landforms maps of the two martian equatorial hemispheres, drawn on a Lambert equal-area projection from those mission results, as shown here:



From T.A. Mutch et al., The Geology of Mars, © 1976. Reproduced by permission of the Princeton University Press, New Jersey.

Much of the upper parts of the two globes (left and right hemispheres in the diagram) have

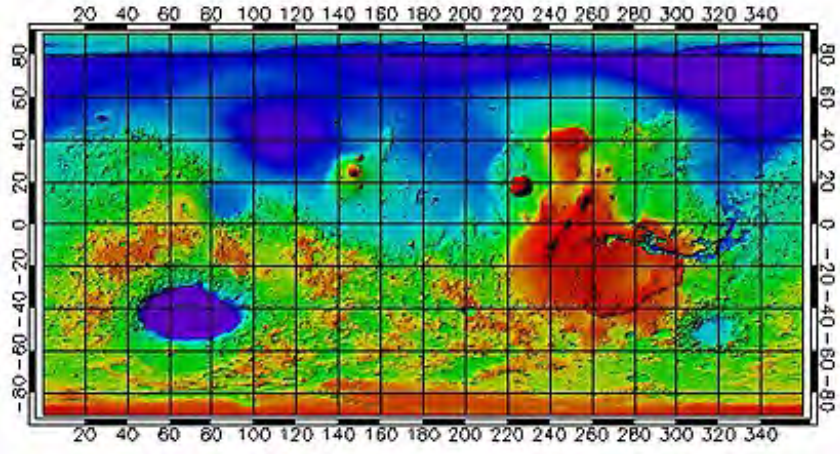
plains units (*p*) (in white), many thought to be volcanic flows (*pv*), and volcanic constructs (*v*) (such as shield volcanoes), some of whose ages could be less than one-half billion years.The bottom polar hemisphere in each globe contains cratered terrain (cu) (dark gray), considered to be several billions of years old (and the likely source of the 3.5 b.y. old martian meteorite containing organics, found in Antarctica), where the bulk of the larger impact structures survive. Other subdivisions include cratered plains (*pc*) and moderately cratered plains (*pm*). Unusual or specialized units consist of channel deposits (*c*), grooved terrain (*g*), knobby hummocky terrain (*hk*), fretted hummocky terrain (*hf*), chaotic hummocky terrain (*hc*), and mountainous terrain (*m*). Units confined to the polar regions are: permanent ice (*pi*), layered deposits (*id*) (thought to be layers of windblown dust interspersed with ice), and etched plains (*ep*). As with Venus, structural signs of plate tectonics are absent.

Compositional data (discussed on pages 19-13 and 19-13a) suggest that the darker shaded southern part on the Mars maps is underlain by basalt whereas parts of the northern hemisphere contain andesites as well as basalts.

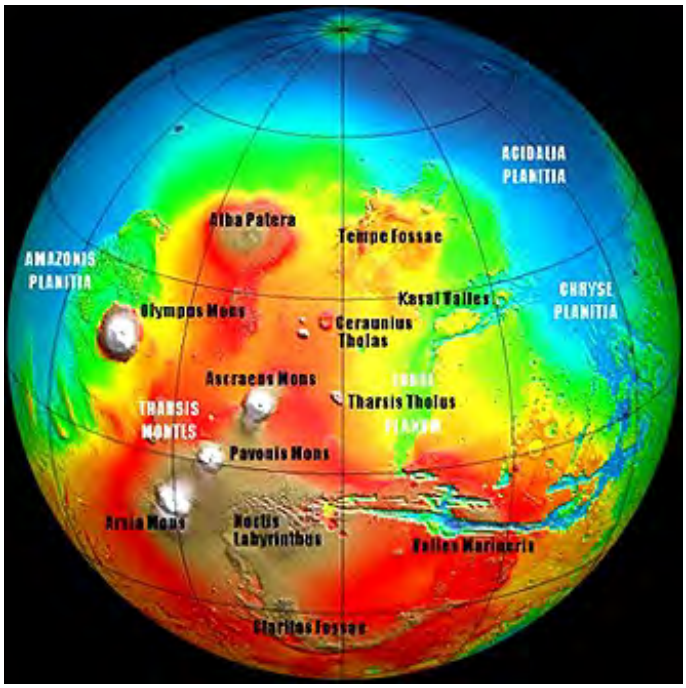
Many features in these categories are also landmarks and physiographic provinces. In the left hemisphere, the three volcanoes in a row (Ascraeus Mons [top], Pavonis Mons [center], and Arsia Mons [bottom]) are in the Tharsis Montes group. The two large ones above and to the left are Olympus Mons (largest volcano in the Solar System) and Alba Patera (also a caldera-capped volcano). The long, curvilinear feature (*c*) near the equator in the left hemisphere is the great Valles Marineris, many times longer and much deeper than the Grand Canyon, with associated chasmas or tributary canyons. In the right hemisphere, the large area (*p*) in the southern half is Hellas Planitia, site of the biggest impact basin on Mars, cut into ancient terrain and backfilled with lava.

19-36: The Tharsis region shield volcanoes are huge by Earth standards. What single volcanic structure on Earth comes closest to these in size and morphology? ANSWER

Further knowledge of martian physiography can be gained from relief (elevation difference) maps of Mars. Several of these for selected regions appear in the next three pages; such maps can be constructed from laser altimeter and radar data, as provided by Mars Global Surveyor and other spacecraft. Shown below is a global shaded relief map of Mars, made from MGS data. Note the higher elevation (red) region around Olympus Mons and the Tharsis volcanoes; the low (blue) depression that constitutes the Hellas Basin; and the generally low state of the northern polar region (James Head's ocean basin). Note also that the southern 60% of Mars is significantly higher (except the Hellas Basin) than the northern part.



The next map, made from MGS MOLA data, focuses in on what is probably the most interesting part of the martian surface - the high terrain that includes Olympus Mons; the other three "Mons", and Noctis Labyrinthus:

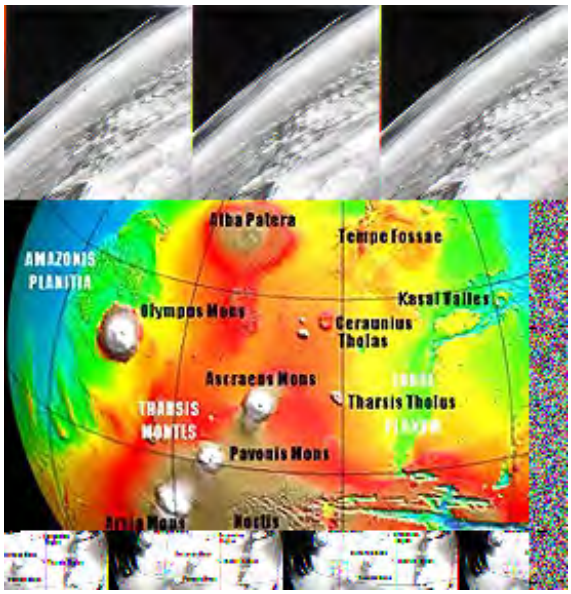


The Martian Atmosphere

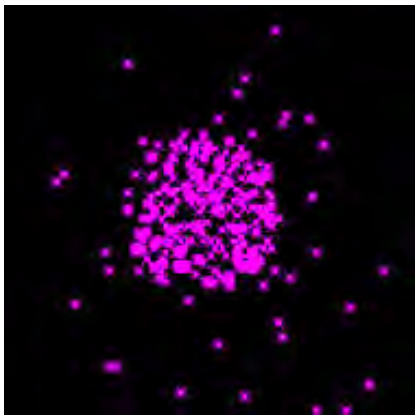
The Voyager and Mariner missions confirmed the thin martian atmosphere consists of 95.3% CO_2, with the remainder being mostly N_2 (2.7%), argon (1.6%) and minor O_2, CO, and traceable water vapor. Atmospheric pressure averages 7 millibars or about 0.7% of Earth's. This thin atmosphere is the survivor of what once may have been a much denser atmosphere. That thicker atmosphere was eventually blown away by the solar winds. For a while the atmosphere may have been protected by a strong magnetic field (which on Earth prevents the solar wind from removing its atmosphere). But the magnetic field eventually shut down (see page 19-13), allowing the solar wind to remove most of the original atmosphere. Evidence for this primitive atmosphere is indirect and inconclusive but an early atmosphere is consistent with some of the signs of water transport responsible for sedimentary layers that are abundant at various locales on Mars (see 20-13ff). The question is important insofar as it bears on possibilities for life.

Despite the meagre present day atmosphere in terms of density, what is there has some dramatic effects on the martian surface. Windstorms in the current atmospheric envelope can reach speeds greater than 200 km/hr (124 mph). These give rise to dust storms and to aeolian deposits.

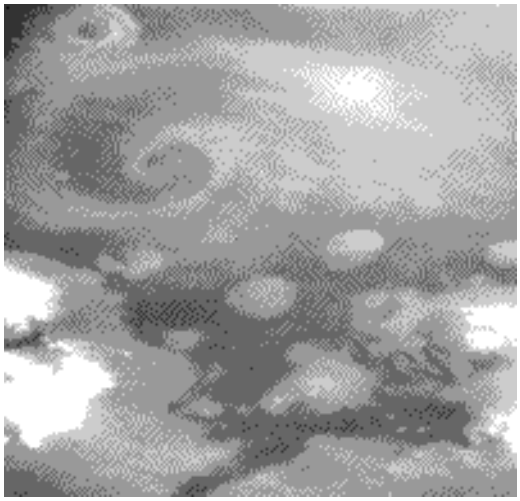
Clouds of water-ice are fairly common in season, especially near the poles. This Viking Orbiter image shows clouds above the martian limb and perhaps clouds over the land.



The presence of small amounts of oxygen has been confirmed by X-ray emissions monitored by the Chandra X-ray telescope. O_2 is excited by X-rays accompanying the solar wind. The amounts seem to vary, as seen in this Chandra image:



Both Mariner 10 and the Viking Orbiters gathered many images showing various aspects of atmospheric circulation, such as this view of spiraling clouds above the martian North Polar region; note also the bright patches of ice.



Clouds reminescent of cumulus types on Earth are seen in the atmosphere. These are evident in this MERS Opportunity (see page 19-13a) image which also show cloud bands formed by atmospheric diversion by the Crater Mie:

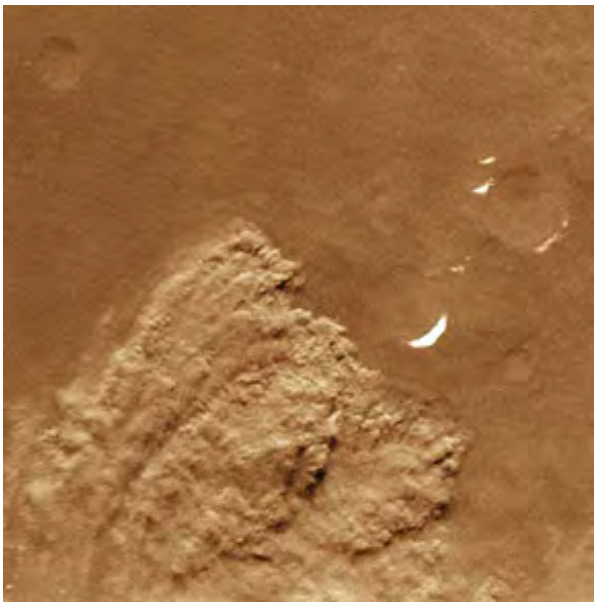
The martian atmosphere is generally hazy, resulting from suspended dust and possibly some water condensate. Instead of the blue sky of Earth, one peering up from the Mars surface would see its sky as grayish-yellow to reddish, depending on the time of the martian day. Although Viking Landers gave some indication of these colors, the next image, from Mars Pathfinder's camera looking well up into the sky, is a good rendition of color near sunrise; the clouds are of the cirrus type as known on Earth.

There are indications that Mars had a much thicker atmosphere one or more times in its past, especially during the first billion years. This is tied into evidence for extensive water on Mars, as will be discussed in subsequent pages. Most of the earlier atmosphere was lost, possibly by escape but more likely by being blown away by the solar wind. Unlike Earth, whose atmosphere is protected from the solar wind by its magnetic field, Mars may never have had a global magnetic field, implying that its core is either not iron or if so, then is not molten now (but possibly in the past).

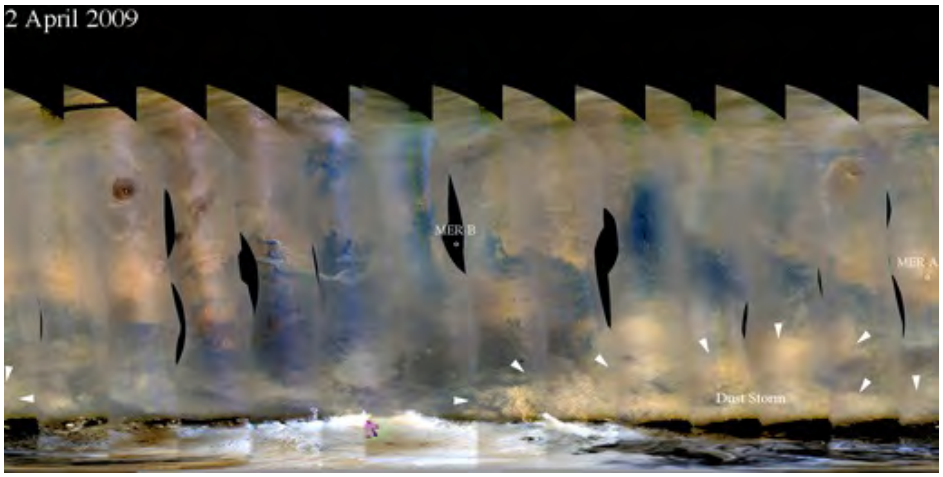
The presence of sedimentary layers on Mars (to be documented later) implies lakes or even seas in martian history. This condition suggests milder climates in the past, during which water could remain unfrozen. There may even be cyclic climate change. Mars wobbles much more than Earth (Earth's wobble is dampened by the gravitational action of its large Moon). The poles may dip as much as 40⪚ from time to time. This could lead to periodic warming.

Thus, Mars is now known to *not be DEAD* as a planet. As we shall see, there may be active water transfer, buildup of ice caps, and possibly volcanic eruptions.

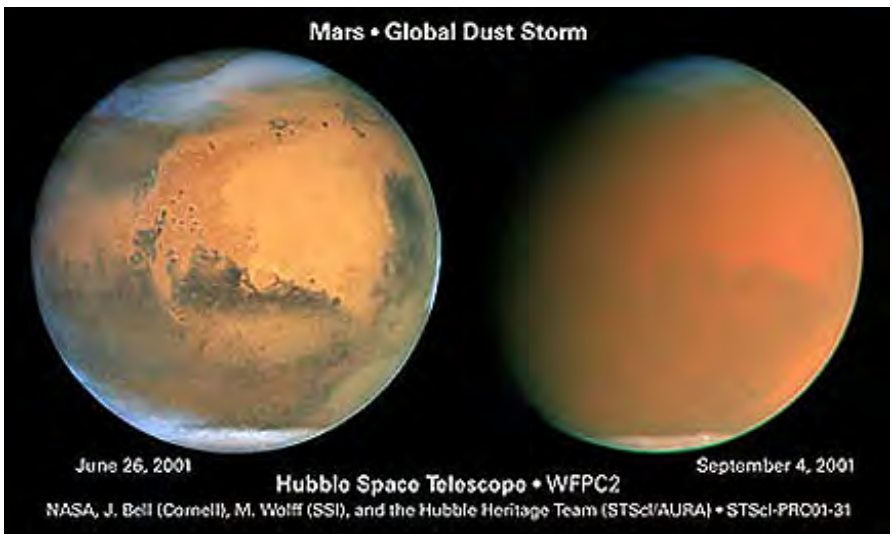
As further proof that Mars is meteorologically quite active despite the thinness of the atmosphere, an extensive dust storm was imaged from an Orbiter in the act of happening:



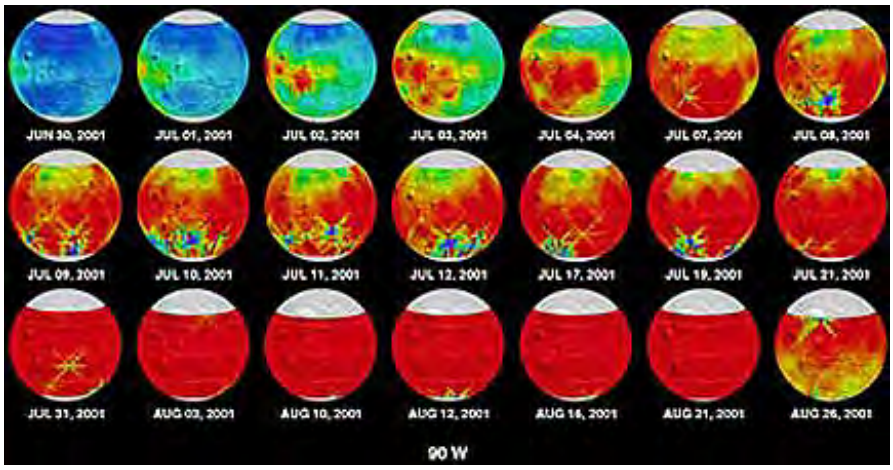
Dust storms commonly begin to form in the martian Spring. Here is a series of images obtained during a single Earth day that show a full-fledged dust storm in the high southern latitudes and dust clouds of lesser intensity over much of the remainder of the planet:



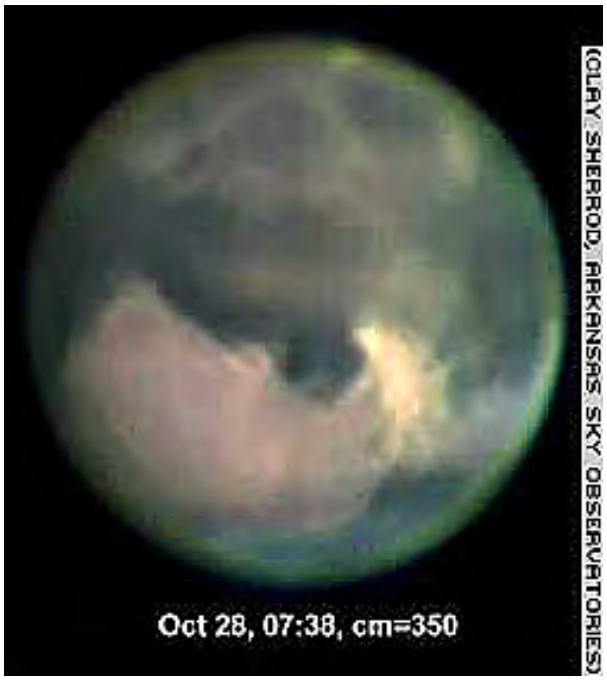
When enough of these regional dust storms form, they coalesce and sometimes cover the entire planet for many (earth) months. More than half the planet was engulfed in a widespread dust storm that began in June of 2001 and started to diminish in September. Compare these "before" and "after" Hubble images of minimal and maximal dust coverage; note how the surface features in the left image are masked by the pervasive, near global dispersal of dust into the atmosphere as seen in the right image.



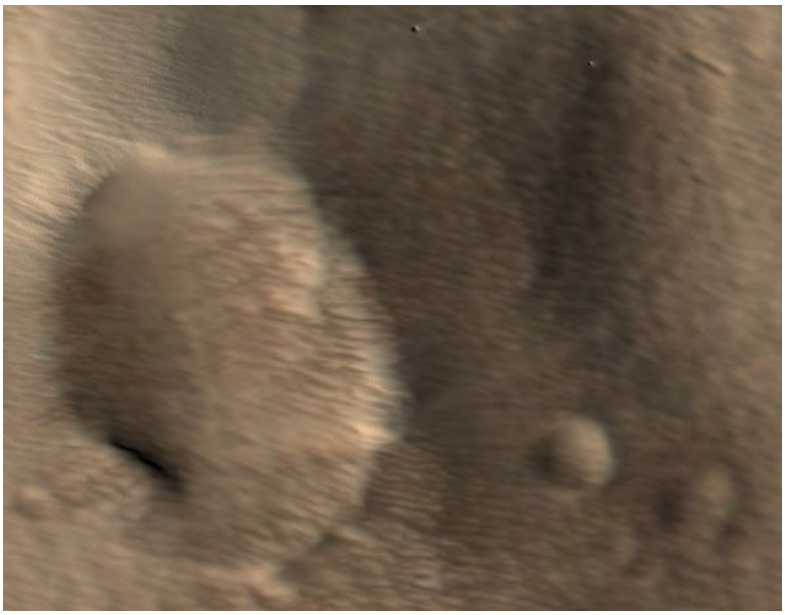
Using the Thermal Emission Spectrograph on Hubble, a sequence of images taken between June and August, 2001 shows the variations in density of dust (shown in red) in the martian atmosphere.



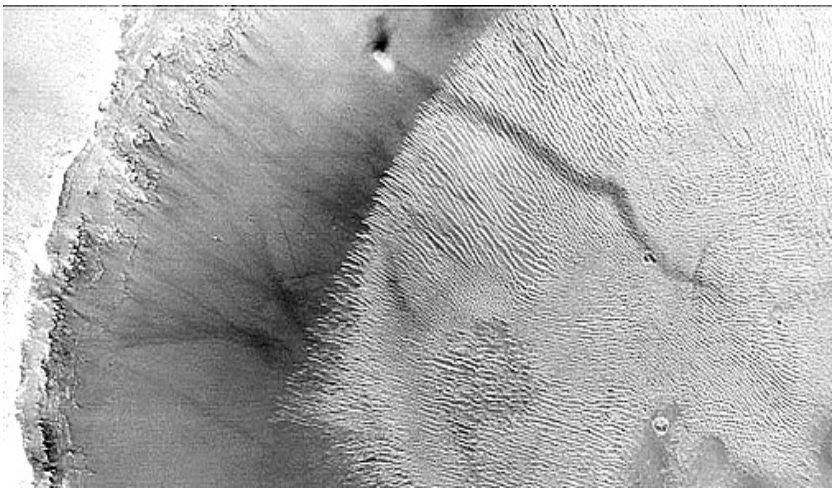
Earth-based telescopes can spot developing dust storms almost as soon as they start. The yellow-brown area in this next full Mars image made when that planet was at its closest distance to Earth (55,000,00 km [34,000,000 miles]) in the last hundred years is an organizing dust storm imaged through a ground-based telescope in Arkansas:



Dust is in the atmosphere most of the time, including periods when dust storms have subsided. This dust can obscure a scene, giving the impression that the image is not quite in focus, as shown here for an MRO image of the Pavlonis Mons region:

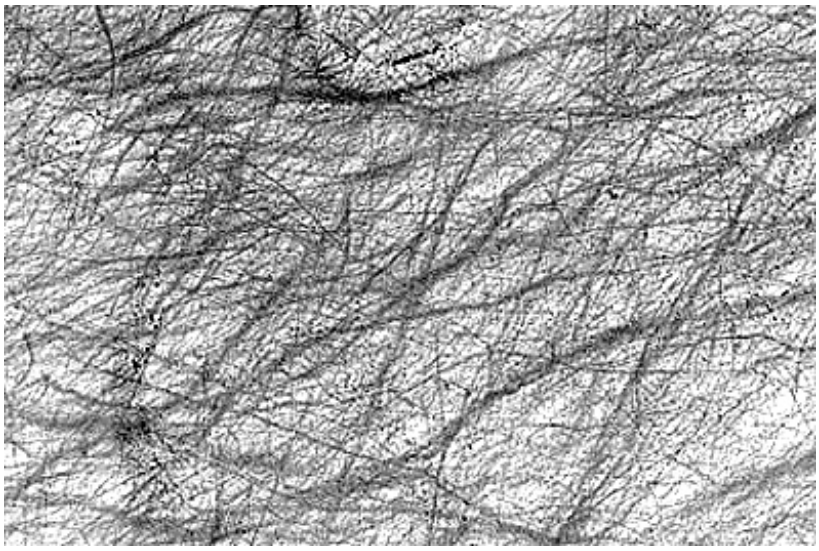


Certainly, as evident from the previous images, its thin atmosphere is remarkably active, with fast-moving winds picking up and redepositing surficial particles in dunes and other aeolian features. Here is a "candid camera" image of what on Earth are called dust devils - tornado-like wind swirls that pick up surface fines to create a tubular or thin funnel-shaped cloud that sweeps across the surface. This Mars Orbiter Camera (MOC; see page 19-13) image shows a thin dark trace along a crater wall made when the dust-devil (actually moving; on top, at the left end of the streak) removed material from the wall:

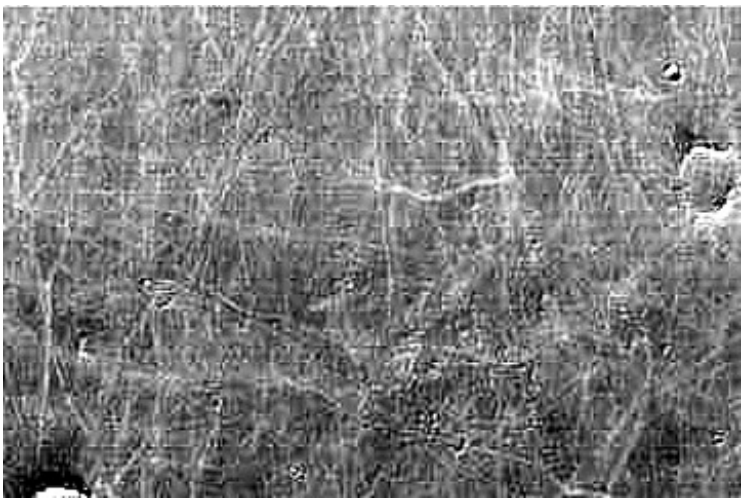


Thin dark markings, often found in swarms, on parts of the martian surface, have now been identified as trails caused by numerous dust devils ("mini-tornadoes") scouring the loose surficial material. Here is a striking example:

Dust devils must be commonplace on Mars. In a single season, the number of trails like those shown below, can be many. The winds seem to destroy earlier ones, then fresh trails are created that expose the dark surficial materials below a new covering of lighter-toned dust:

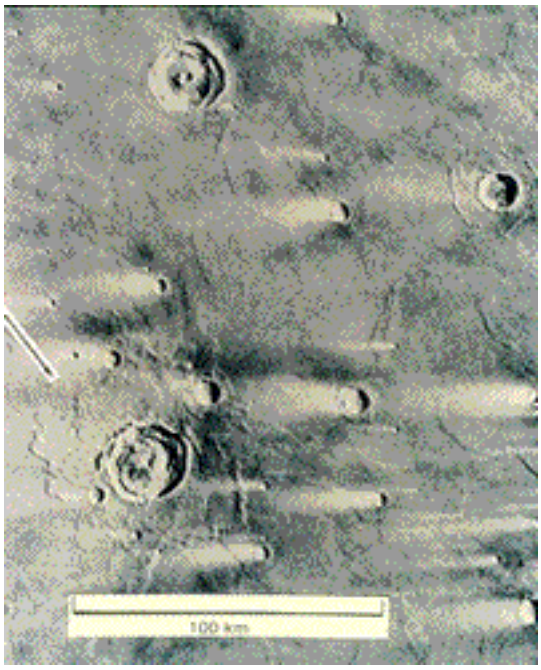


In parts of Mars that have dark surfaces, the dust devil trails appear as lighter streaks, implying the underlying near surface contains lighter materials exposed when the whirlwind picks up the dark fines covering the surface. Thus:

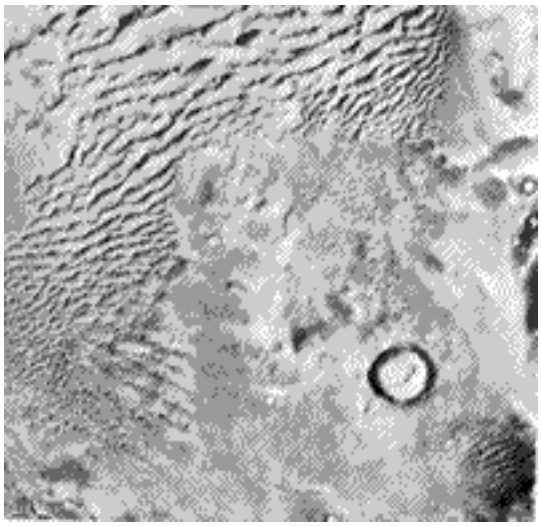


Individual dust devils have been imaged as they actually passed by landers on Mars. The Rover Spirit caught this image of a nearby whirlwind:

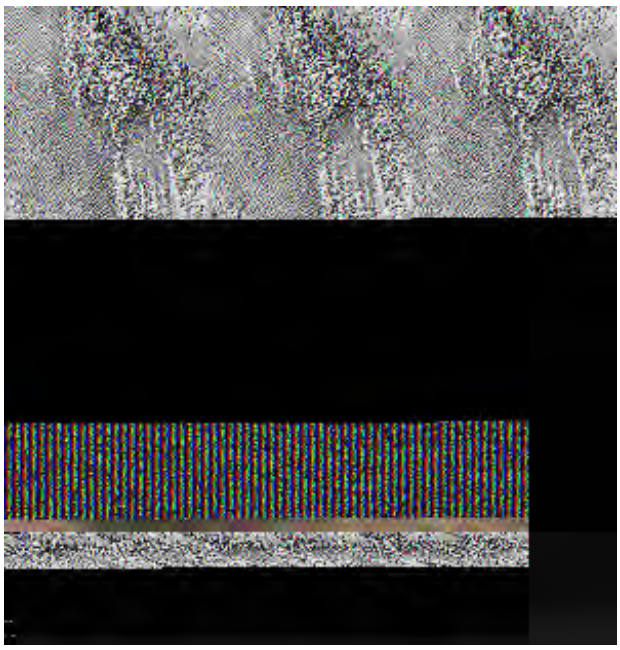
The martian dust and sand is frequently spread out as wind streaks.



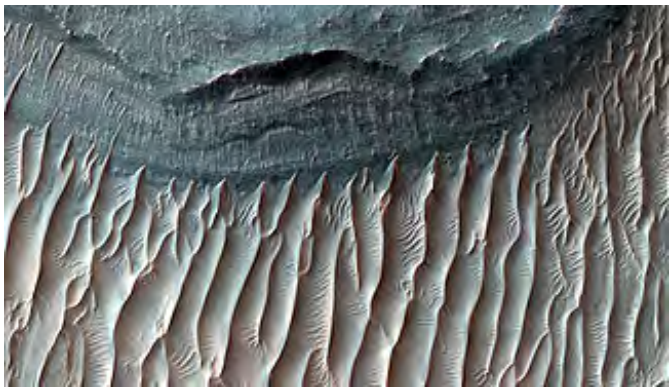
Despite the very low density of today's martian atmosphere, its high speed winds can cause erosive sculpturing, and can deposit uplifted dust and sand into vast dune fields. The Viking Orbiter image below contains transverse (elongated, or parallel) and barchan (crescent) dunes laid down in a plains setting.



A second image shows parallel dunes on the left and barchan dunes on the right.

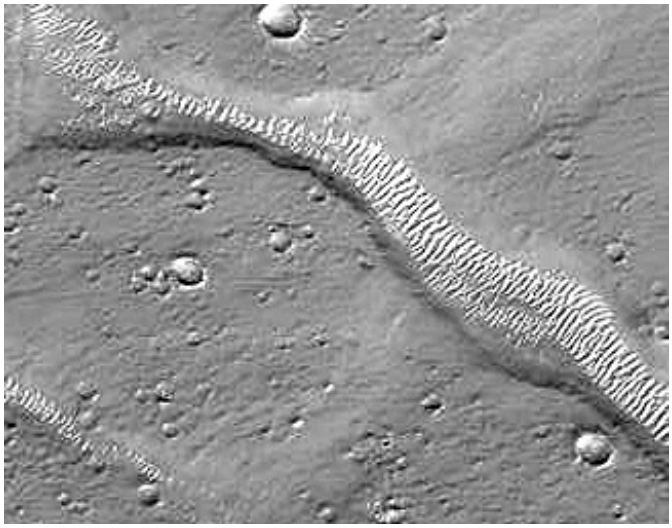


More recent imagery shows the splendor of martian dune fields, such as this one in Ius Chasma:



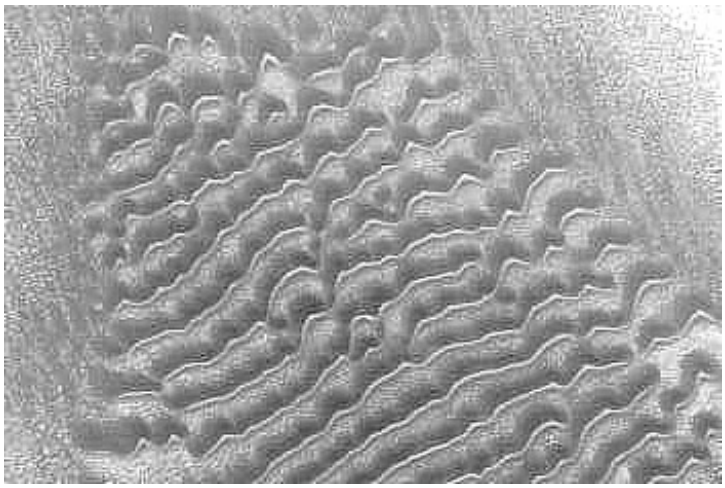
Here are longitudinal dunes adjacent to the Schiaparelli crater:

Distinctive ripple dunes can develop within shallow troughs or linear valleys. This is a good example:

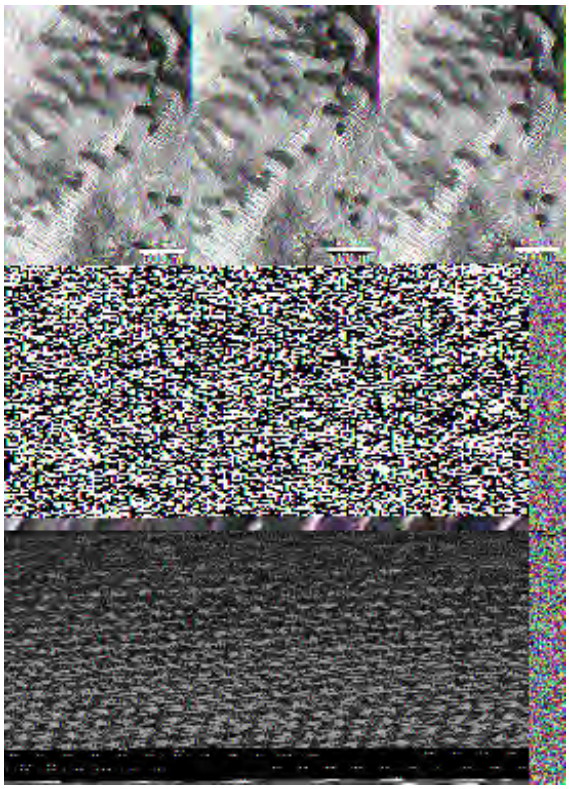


19-37: Why is it rather surprising that dunes with types similar to those on Earth are so widespread on Mars (hint: consider the atmosphere)? ANSWER

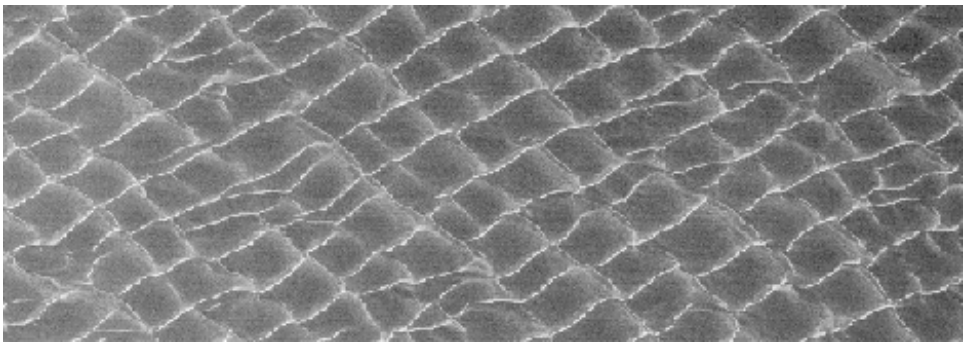
In contrast to dunes on Earth that usually have a light tone because they are made up of mostly clear to cloudy white quartz, dunes on Mars usually are much darker because the grains making them up are either basaltic fragments or darker hematite. This is a typical set of darker dunes:



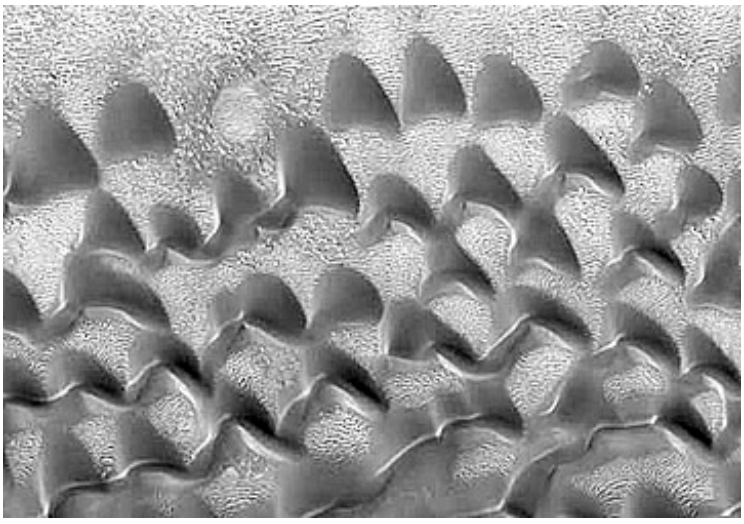
Although uncommon, some places on Mars show two sets of dunes - one dark and the other light in tone. This dual occurrence is hard to explain and suggests two different processes (or time intervals) at work:



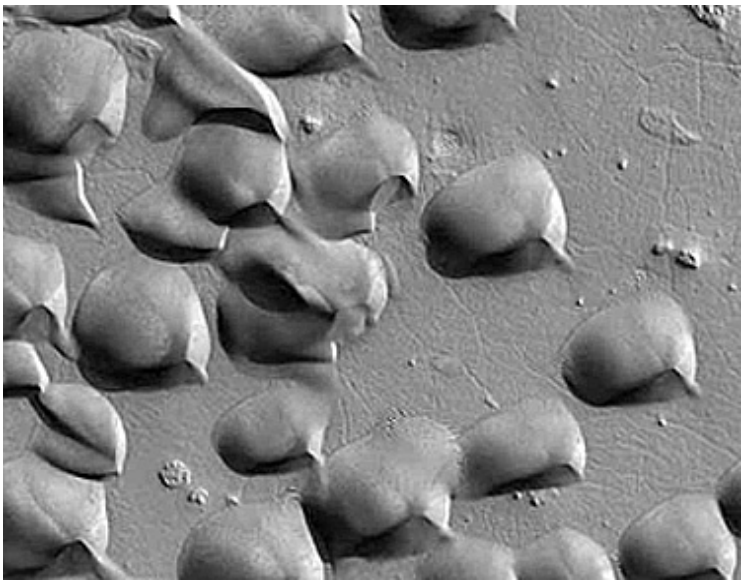
Martian dunes can have a regular, criss-crossing pattern (lattice-like) that is uncommon on Earth where strong prevailing winds tend to favor a single alignment (cross-winds would likely destroy the reticulation). Here is an example from the martian polar region:



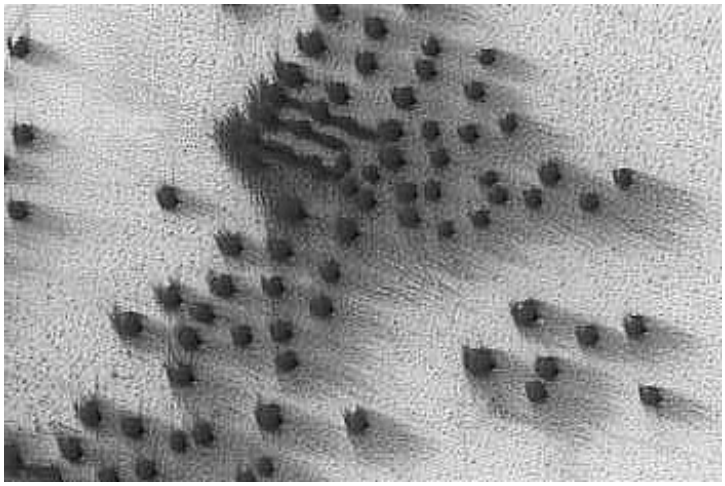
More striking close-up views of dunes have been imaged by the Mars Global Surveyor (MGS; MOC) (see page 19-13). Crescent dunes very similar to those found on Earth also develop on Mars, as seen in this MGS) image:



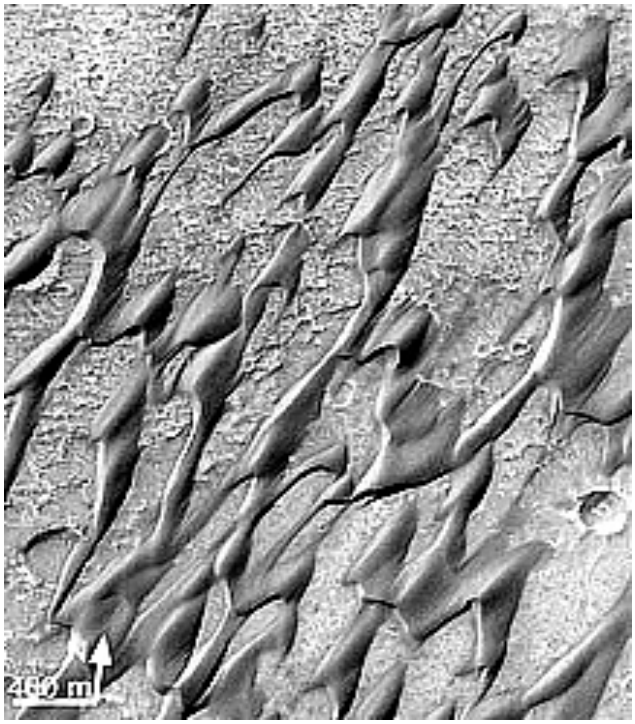
A variant of this type leads to dunes that approach crude circles in shape.



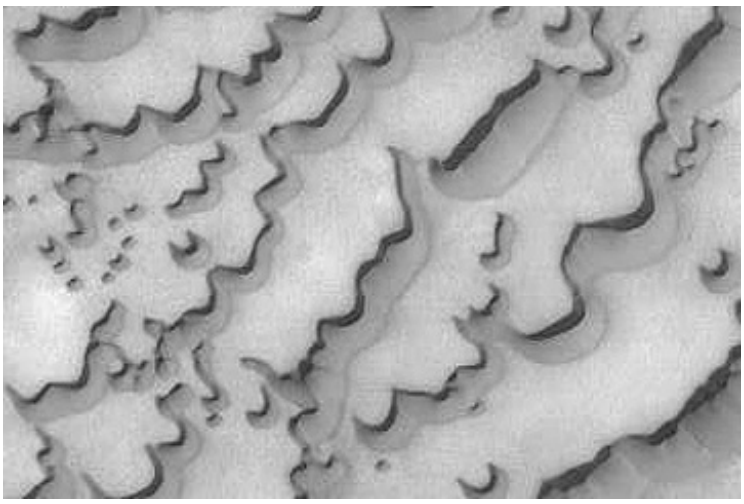
In this next image, dunes having shapes similar to those just above have formed but they are far more widely spaced and tend to be round. Some "wags" have named them "martian cookies". They are most common in the polar regions where their dark materials stand out against a frosty surface.



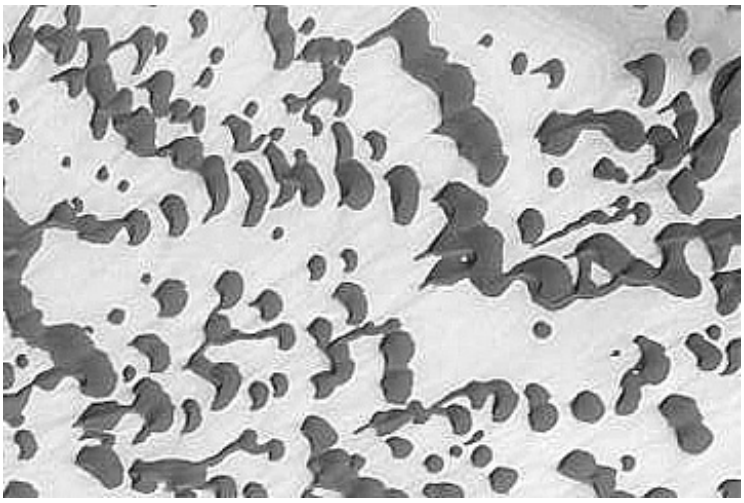
A MOC close up view of barchan dunes appears next:



The martian dunes come in a wide variety of forms. This MOC image shows dunes with a broad, gentle slope and steep, ragged foreface:

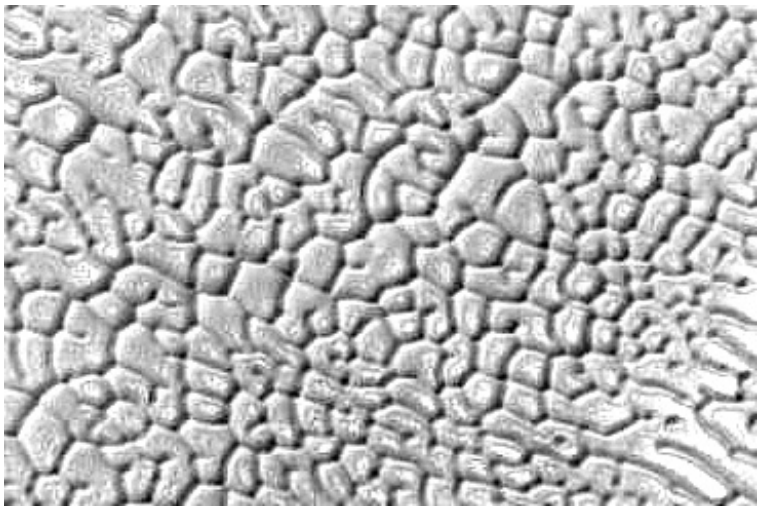


Similar dunes, with odd shapes, are evident in this next image in the north polar region. The dark dunes (which may have a high percentage of basalt-like dust and grains) are now exposed after seasonal frost has evaporated from them (in the martian summer there) but stand out because residual permafrost and/or ice is much lighter (more reflective).



Exposure of basalt dust by frost evaporation is responsible for this strange illusion within a polar dune field. The streaks of basalt dust appear like rows of tree trunks:

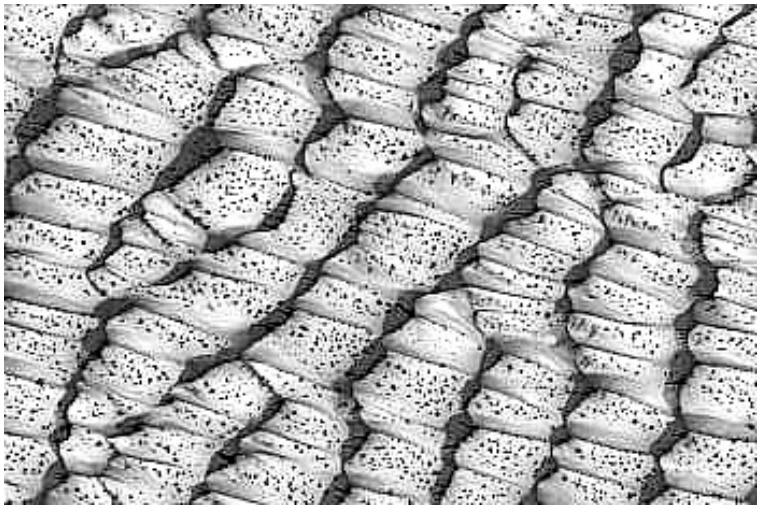
This image shows individual dunes so close-packed that there is no interdune space. Such a density is rare on Earth. The field has developed near the South Pole.



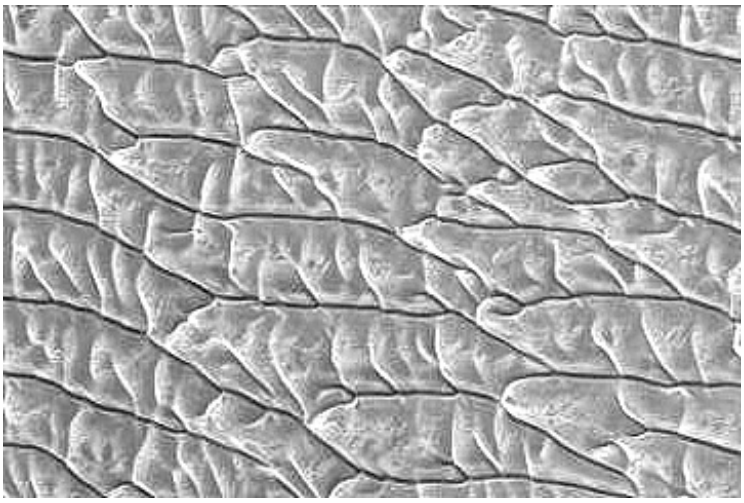
Some dune fields show a series of overlapping dunes, within which thin layers are just visible. Here is an example in the interior of Becquerel crater:

Mars rovers have taken close-up images of individual dunes on the ground. Opportunity (page 19-13a) shows this dune group (longitudinal type) in detail:

Even more unusual, are close-spaced, rectangular dunes in the north polar ice. The dunes themselves are made of dark material but in this view the polar frost has covered them. Melting is just beginning to expose as black spots the underlying ice.

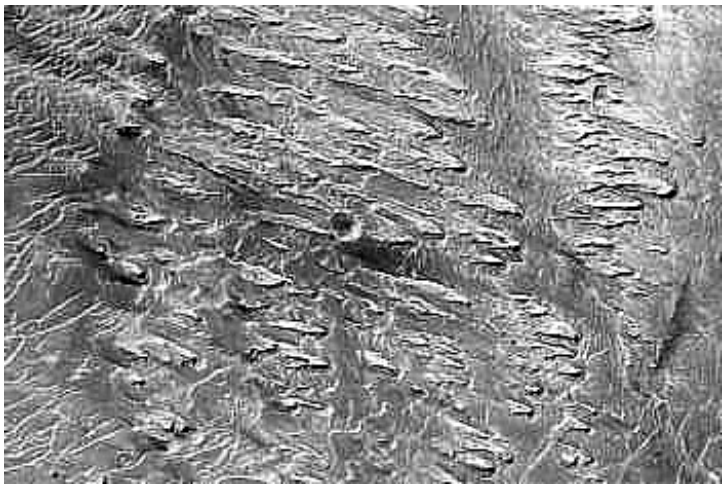


Unusual, close and interfering dunes are found at the North Pole as well. In this image, frost has accentuated the appearance of these lenticular dunes:

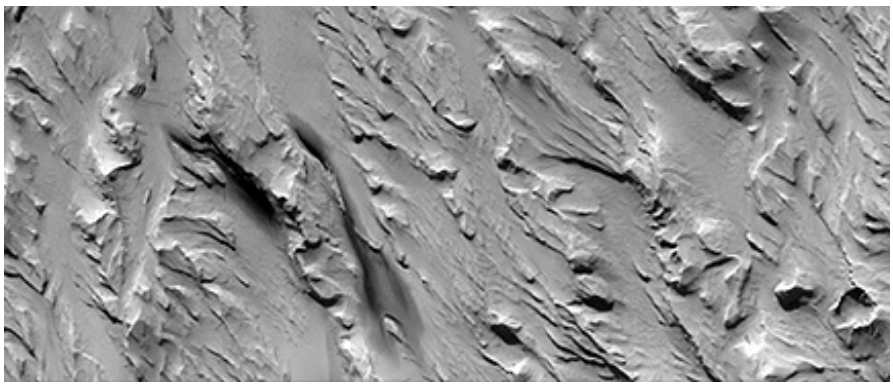


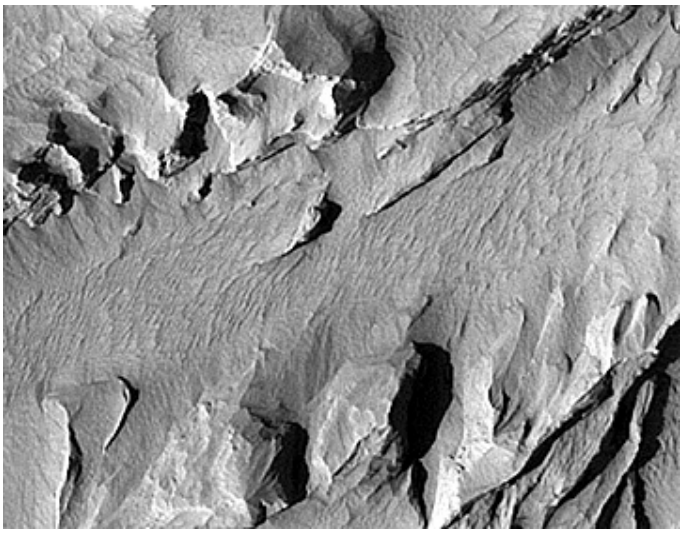
Another feature with a terrestrial counterpart is the yardang (see page 17-5), formed by wind scooping out soft materials and leaving long ridges (usually of the same material) in between. Here is a martian version:

This image shows a field of yardangs, with indications of which way the wind blows in this part of Mars:



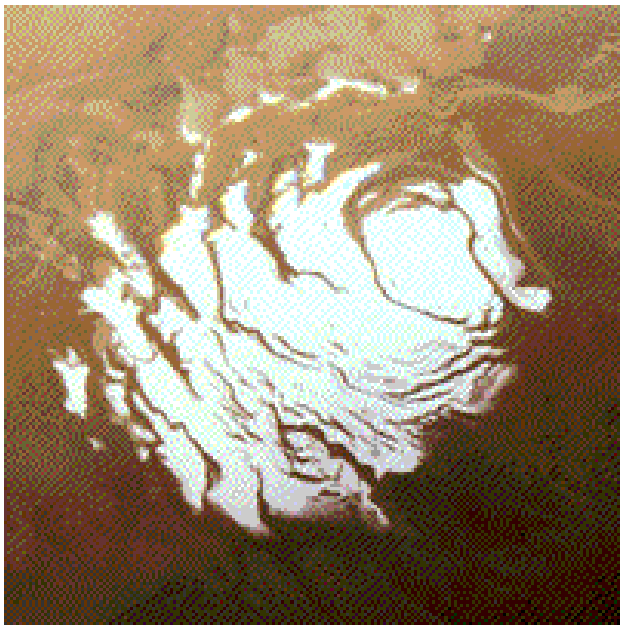
The streamlined shape of the yardangs is a strong indication of the power of martian winds to erode. Wind erosion features of less definitive shape are widespread on Mars, indicating that this type of erosive reshaping is commonplace. The next two images (Medusae Sulci region) show an area of irregular shaped "hills" that are being carved by the wind. The upper image comprises the complete MOC image; the lower image (rotated 90° counterclockwise) is an enlargement of part of the scene in which ripples in the low valleys attest to continuing wind activity that covered the surface with dust:

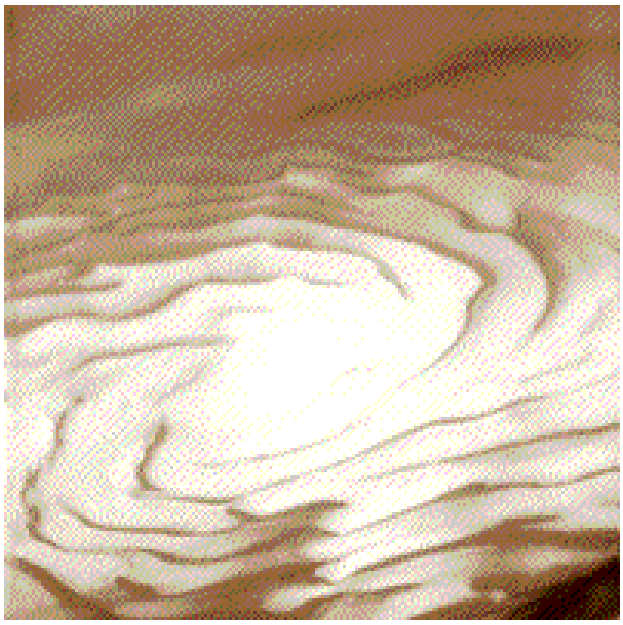




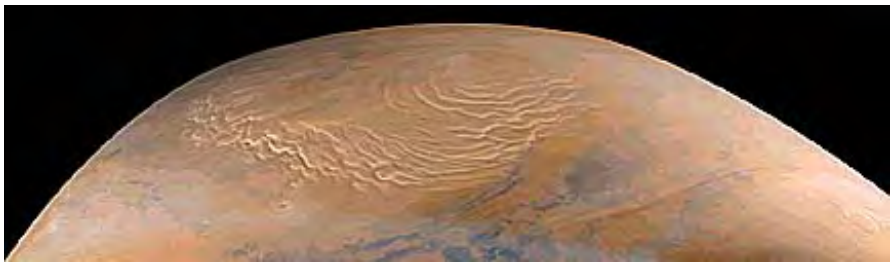
At present almost all of Mars is covered with a thin to thick blanket of dust layers. This often obscures smaller surface features. Here is a MOC view of a smooth dust cover draped over surfaces at Pavonis Mons.

Turning now to the frozen materials that persist or build up/dissipate in the North and South Pole regions. The dominant surface constituent of the polar caps is CO_2, but major amounts of water seem to be locked within the caps themselves as subsurface concentrations (see page 19-13). The next views show the South Polar ice cap (top) and North Polar ice cap (bottom) (which may contain more water) near their maximum growth stage, during a martian winter. This winter recurs about every 685 Earth days at each pole (remember, the rotational pole is tilted about 24°). In about half that time the polar ice at one pole shrinks as summer warming evaporates the frozen gases, and possibly subliming water underneath and the opposing pole experiences ice condensation and growth:



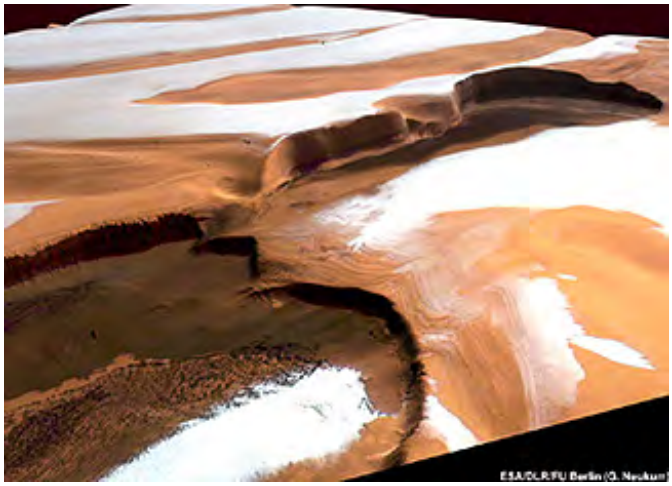


The layering associated with each icecap, probably representing seasonal deposits of dust (see below), is even more obvious in this regional oblique view of the top of the Northern Hemisphere:

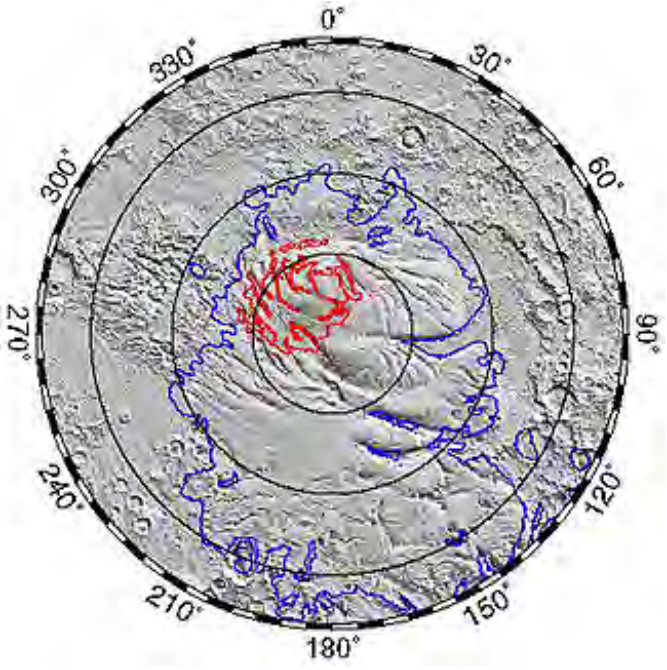


Considerable change in size of both ice caps occur over the Winter to Summer transition. This three-panel set of images made using the Hubble Space Telescope (HST) illustrates variation of areal coverage at the North Pole over a 6 earth month period from late 1996 through early 1997:

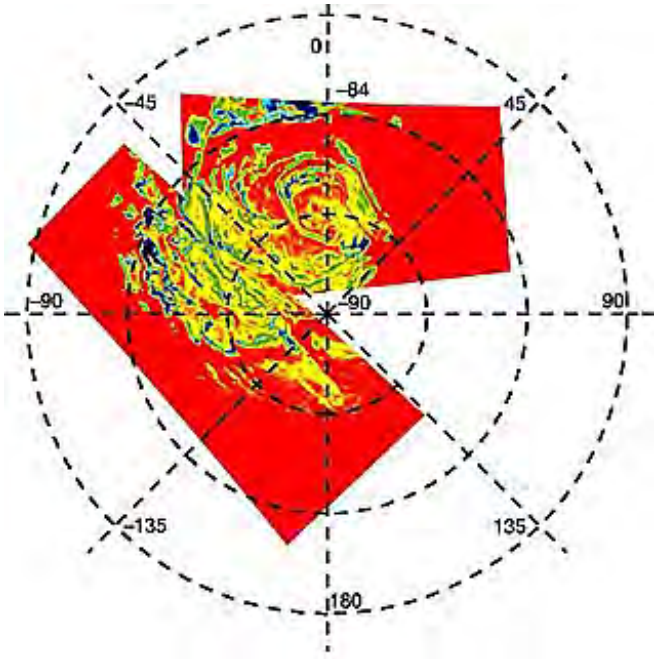
Thus, in the summer the ice cover may have shrunk so that it persists only in patches. That is evident in this Mars Express perspective image of dark material (dust, volcanic deposits, etc.) mixed with a subordinate amount of carbon dioxide/water ice:



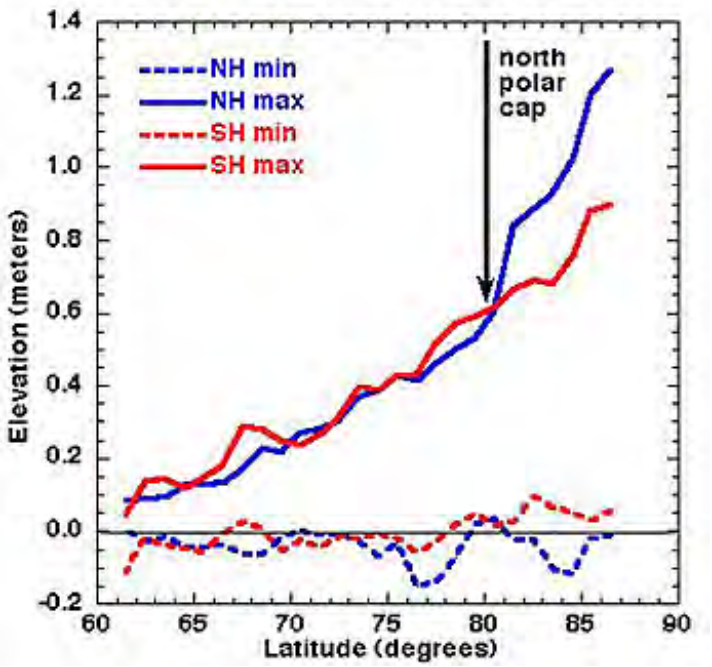
Observations now over several decades have strengthened the fact that the ice is more widespread in the northern polar region, is permanent, but tends to shrink more during the warming season. The higher elevations in the southern hemisphere may account for this difference, being cooler at those heights. Here is a map of the maximum (blue) and minimum (red) extents of the north polar cap during the growth/shrinkage phases:



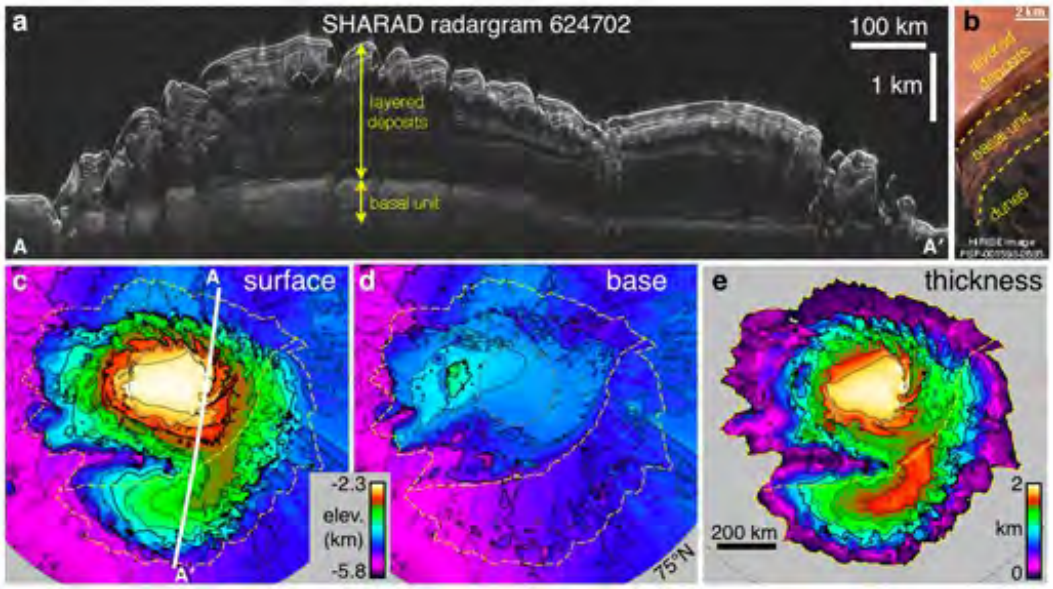
The Mars Express orbiting satellite has now confirmed that there is also permanent ice in the south polar ice cap. Here is a map of water ice (in blue) covering that region:



This change in polar cap thickness over time, determined by elevation differences in meters, has been measured at both pole regions by the laser altimeter (MOLA) on the MGS, with these results:

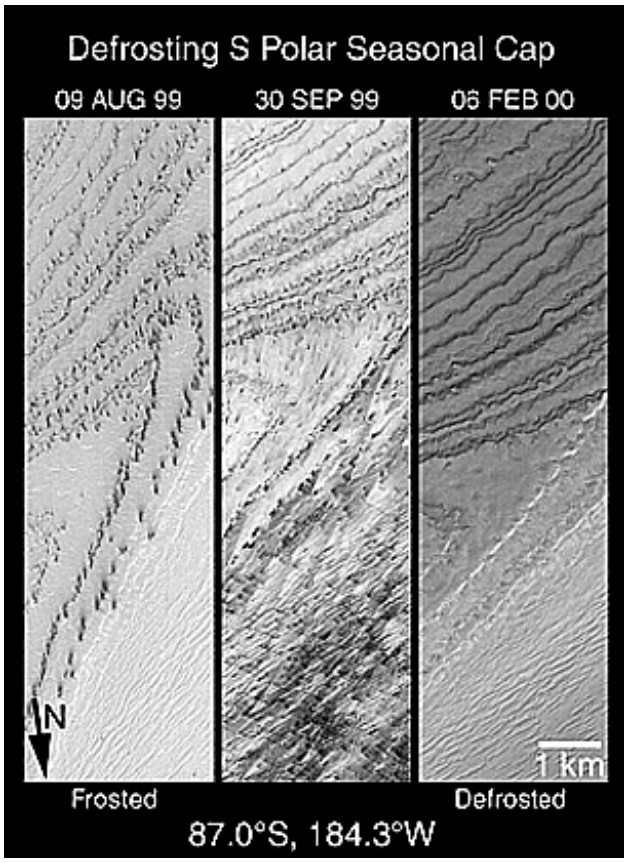


In 2009, NASA released new information on thickness of the North polar ice cap. The Shallow Radar (Sharad) instrument on the Mars Reconnaisance Orbiter (MRO) was used in the determination. The results are summarized in this diagram:

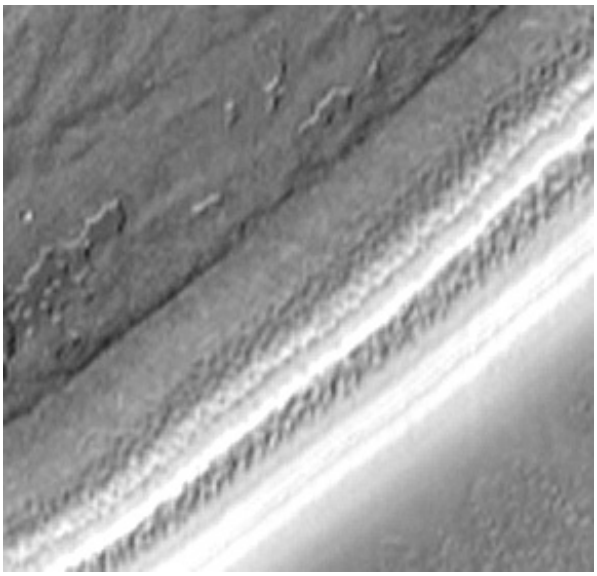


The top diagram (a) shows the radar cross-section of polar layers, many of which contain ice. Diagram (b) is an image looking down at the boundary between the ice-bearing layers and surface materials just outside the cap. The map (c) shows the elevation of the surface of the cap (line A-A' traces the location of the radar cross-section (a)). Map (d) is the elevation of the bottom of the cap (the basal unit) as determined by the radar. The thickness of the cap is plotted in map (e).

The surface of an ice cap shows distinctive changes during sublimation and shrinkage as shown in this MOC image of a part of the south polar ice:

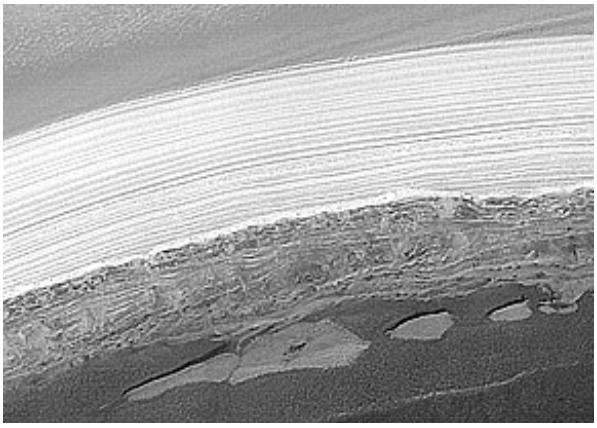


Dust layers in the polar caps have been spotted in pictures from earlier missions that imaged this region on Mars. Details (at 25 m resolution) have been acquired by the Mars Orbiter high resolution camera. Here are layers in the South Polar ice cap:

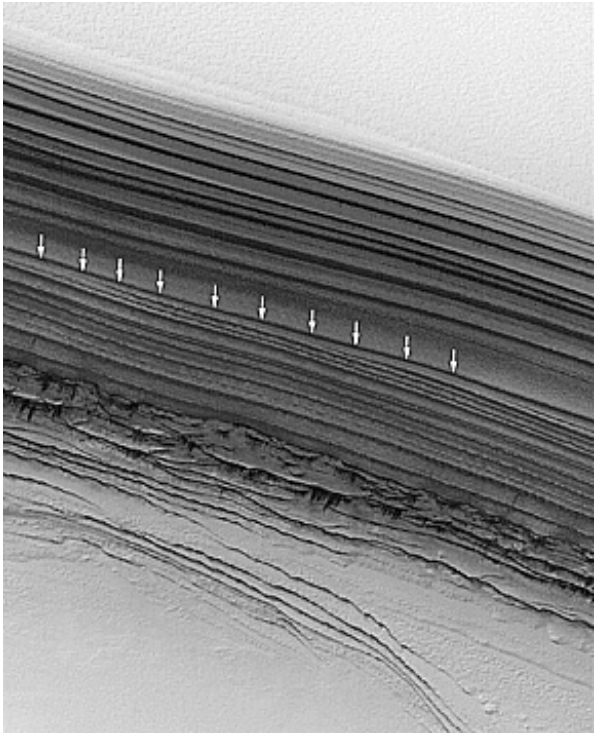


Closer looks at the sides of the polar caps revealed prominent alternating bands of light (the ice) and dark (the dust) materials, as seen here in this view of the south polar cap:

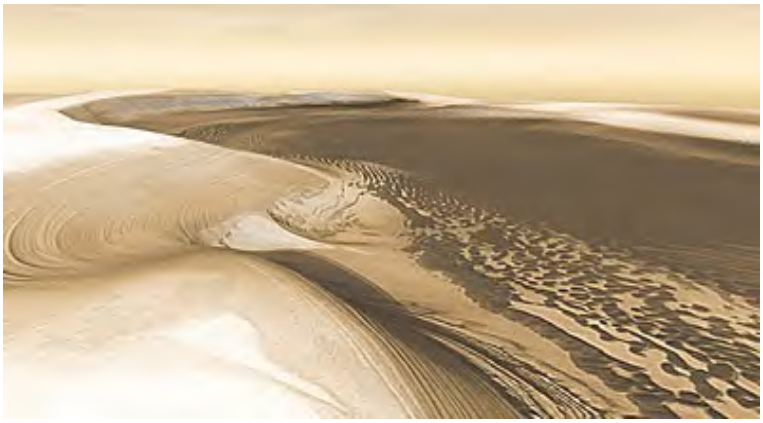
And here is a view of banding in the ice at the martian North Pole:



This MGS MOC image shows banding with different levels of "greyness" in the dust beds; the arrows point to an unconformity (erosional discontinuity in the sequence of layering).



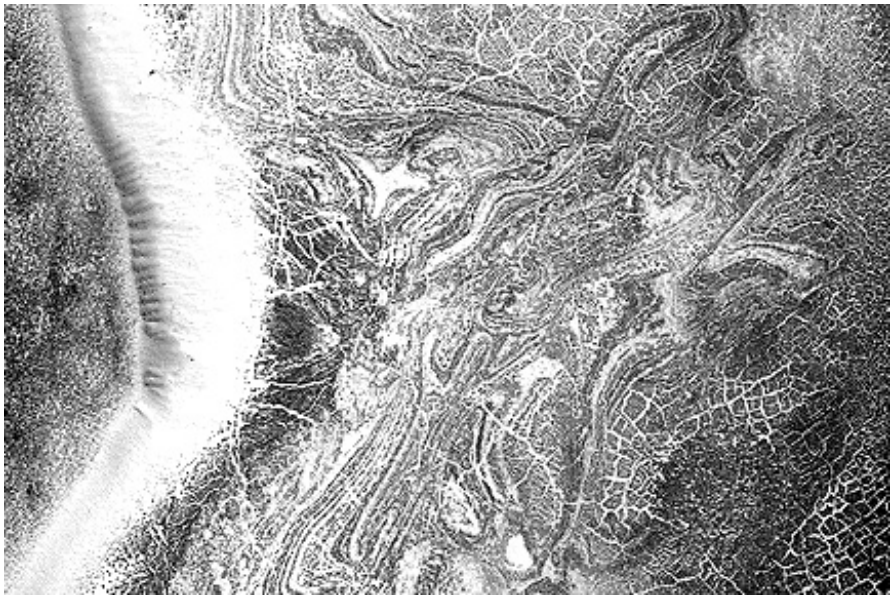
The Mars Reconnaissance Orbiter (MRO) has sent back color images of the polar layering, as exposed in Chasma Boreale (shown below), a 700 km long canyon cut into the ice cap. The redness of the layers suggest windblown dust from lower latitudes but one interpretation favors volcanic ash (not likely):



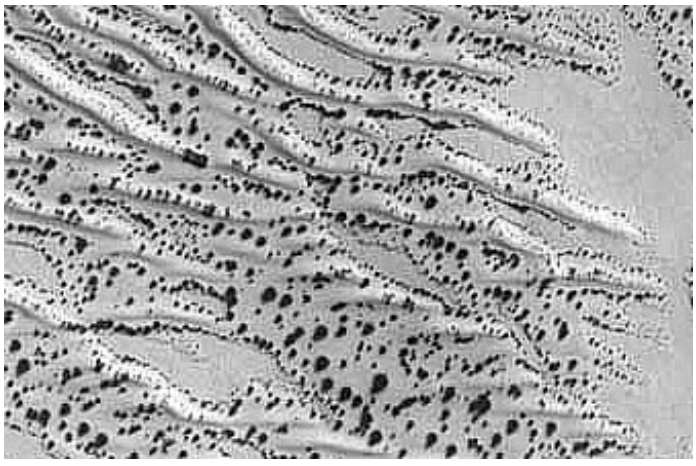
This next image of banding in ice seems to indicate strong light-dark contrasts in an area also undergoing erosional sculpturing; the sharpness may be somewhat illusory.

19-38: Try to devise an explanation for these layers in the image just above. ANSWER

The Mars Global Surveyor has been looking more closely at both poles. As melting proceeds, contorted banding and ice polygons are exposed, and somewhat emphasized in patterns by frost coatings. Here is an example:

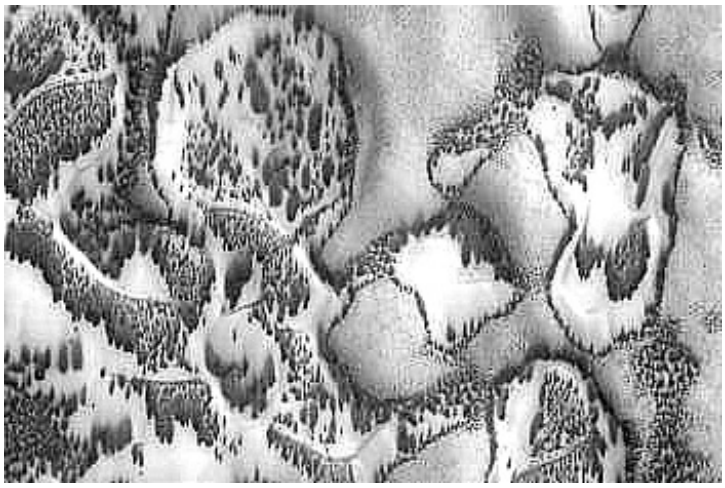


In the fringes around the polar ice caps, dark spots appear and disappear over the course of a martian year. In this image, these spots occur midst a dune field. It is not yet known whether, and how much, water ice and/or carbon dioxide frost coatings are involved in the on-going evaporation that produces the spotting, which may be dune material showing through.

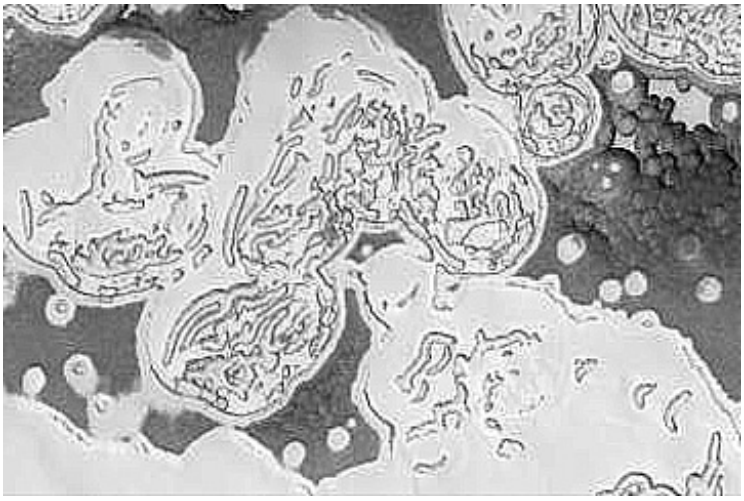


This process of frost evaporation is more advanced in this next image, in which removal of the

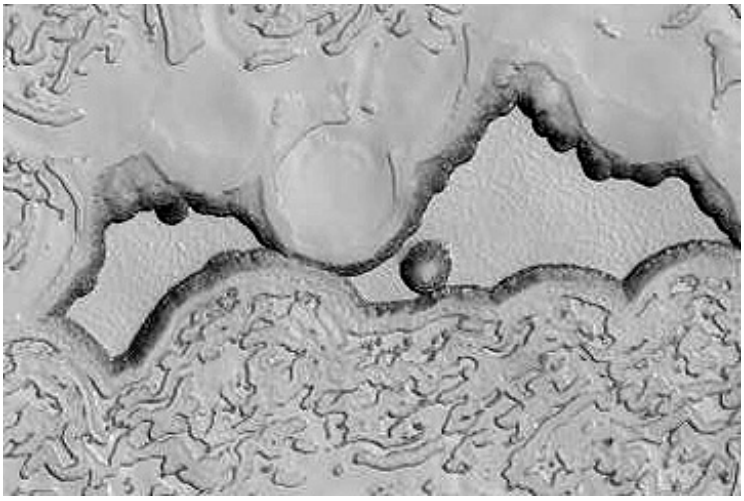
white coating is exposing dark dune sands beneath.



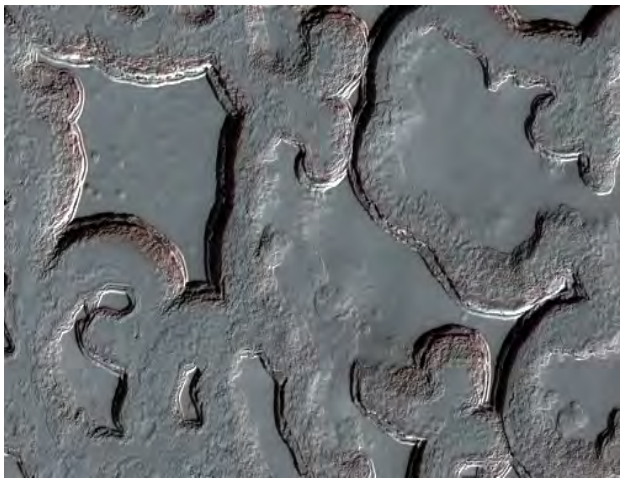
Interesting structures and patterns are developed on the CO_2 ice. This image shows more than one tier of ice, forming plateaus, mesas, and buttes (by analogy with Earth). The darker surface may be ice contaminated with dark windblown dust (either recent or an exhumbed layer):



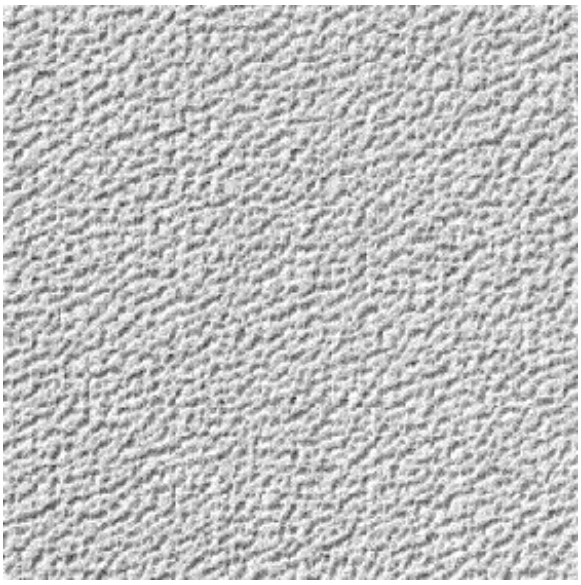
An example of a small mesa apparently composed entirely of CO_2 is shown in this North Polar ice cap. The two mesas are about 1 km and 1.5 km in long dimension and many meters thick. They are very slowly receding by melting and/or ablation at their cliff faces.



The Mars Reconnaissance Orbiter has sent back images that further define the ice cap. These two MRO images show the cap to be permanent enough to have undergone erosion that exposes layers in the ice:

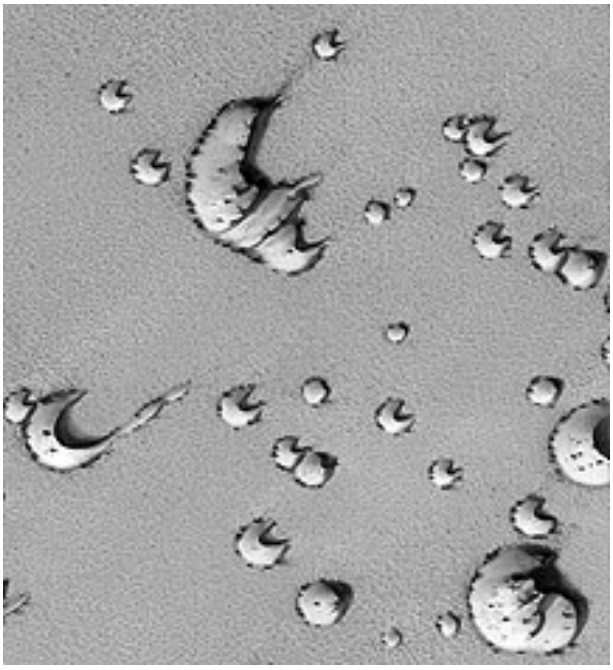


Quaint descriptive names are picked by the investigators to characterize surfaces. In the two images below, both of South Polar ice that is largely CO_2, the top one has reminded them of "swiss cheese" while the bottom looks like a "kitchen sponge.

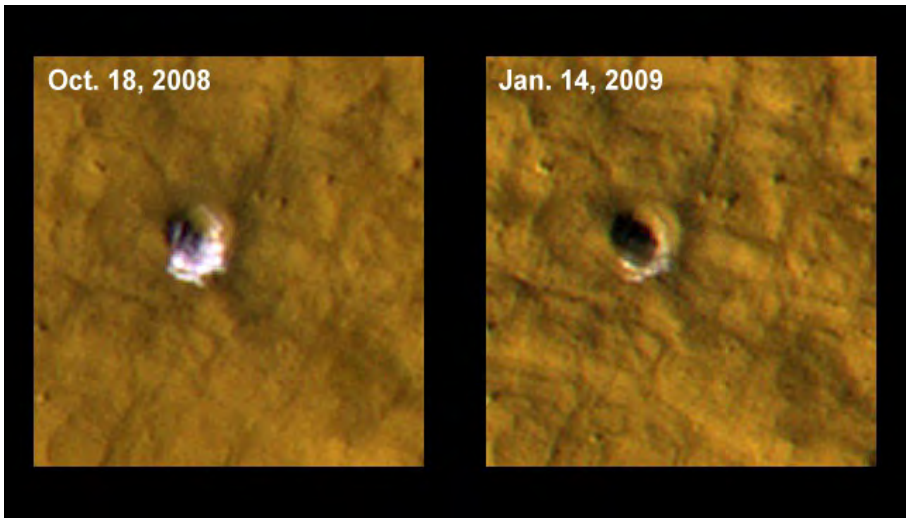


This pattern persists at even higher resolutions; this 1-meter resolution Mars Reconnaissance Orbiter image displays a broken-up surface in martian ice:

Much of the ice at each pole may exist in the form of grains, similar to sand. As the polar ice undergoes changes during the evaporation/deposition cycle that affects the poles, strong winds are capable of moving these grains into barchan-like dunes, as shown in this image of the surface of the North Pole Ice Cap:



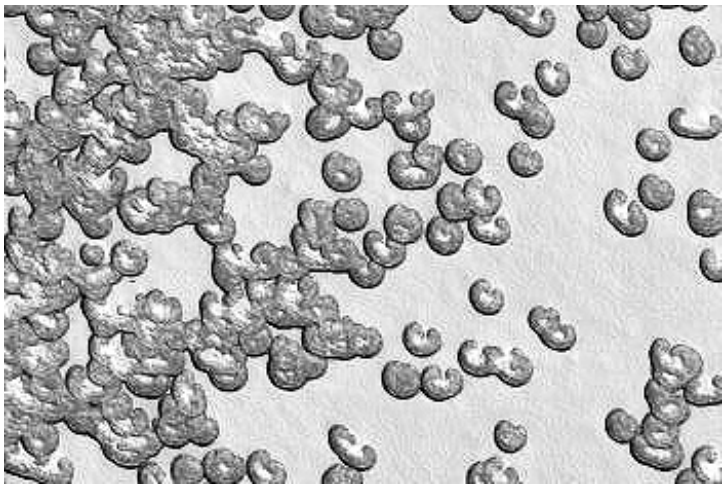
There is growing evidence that ice (water?) is present beyond the poles, at lower latitudes. This pair of MRO images show a bright white fringe around a crater that may have formed recently. The fringe has almost disappeared in 2009, indicating the ice to have sublimed:



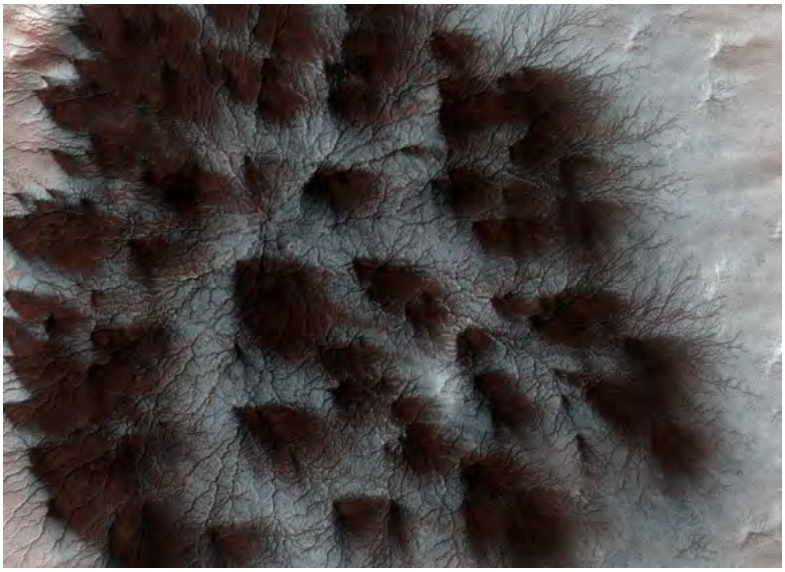
Seen in more detail, the ice around this 3 meter crater is remarkably pure, about 98% water:

A curious feature seen by the Mars Global Surveyor near the South Pole is evident in this next image. Several domelike (nearly circular) structures occur midst what is similar to fretted terrain (exemplified on page 19-13). These domes have steep narrow outer slopes and a wide craterlike interior. Although suggestive of volcanic structures, their origin is unclear.

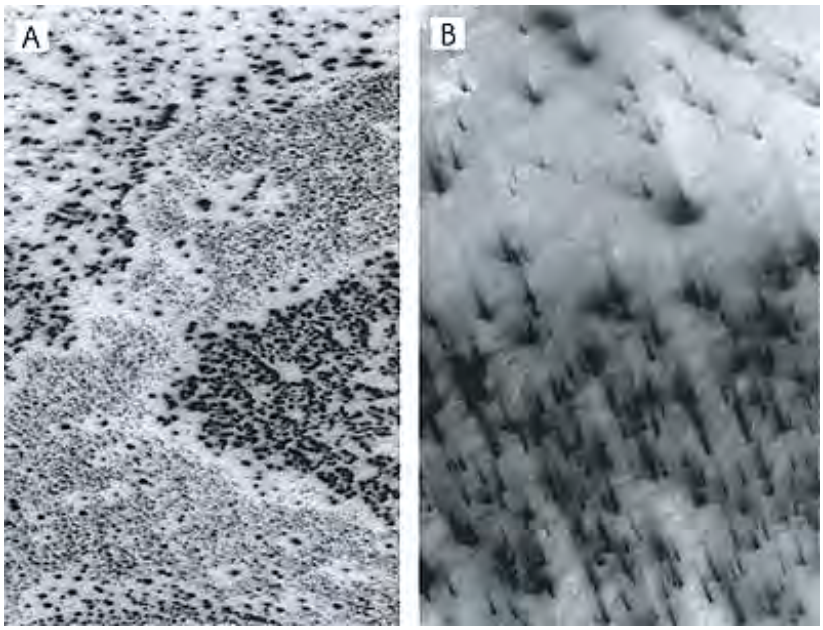
Another unusual feature marked by its roundness is shown in this next image from polar ice. The exact nature is conjectural but the darker clusters of roughly circular objects appear to be caused by evaporation of carbon dioxide during the martian summer warming, leaving behind dust in these "blow holes".



Puzzling features are shown in the next image which displays dark objects in polar ice (they are known colloquially as "starburst spiders"):



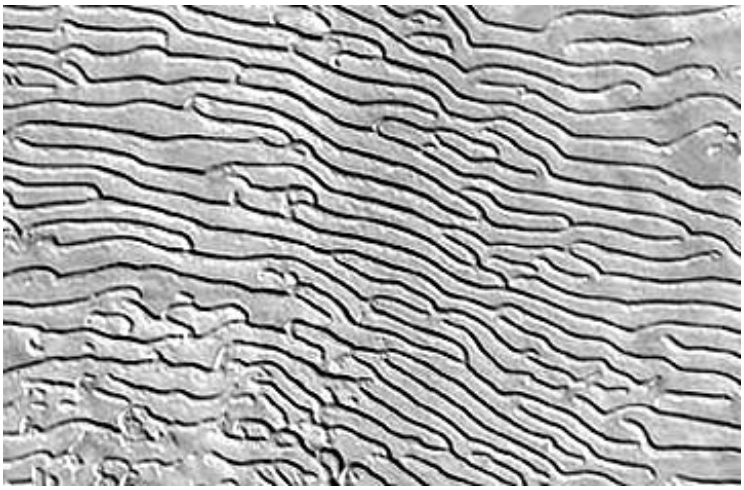
In August 2006 investigators presented evidence citing these features as fountains that during the martian year for several weeks spout CO_2 gas and particles into the thin atmosphere up to heights of 50 meters (160 feet). These illustrations, made by Mars Odyssey's THEMIS (page 19-13a), show two patterns associated with the fountains on the South Polar ice field:



What appears to be happening is this: During the polar summer, surficial water ice evaporates;

carbon dioxide locally sublimates and builds up pressure; the fountains "erupt" and spew gas and particles upwards; the water ice returns in winter and covers the dark spots associated with the vents.

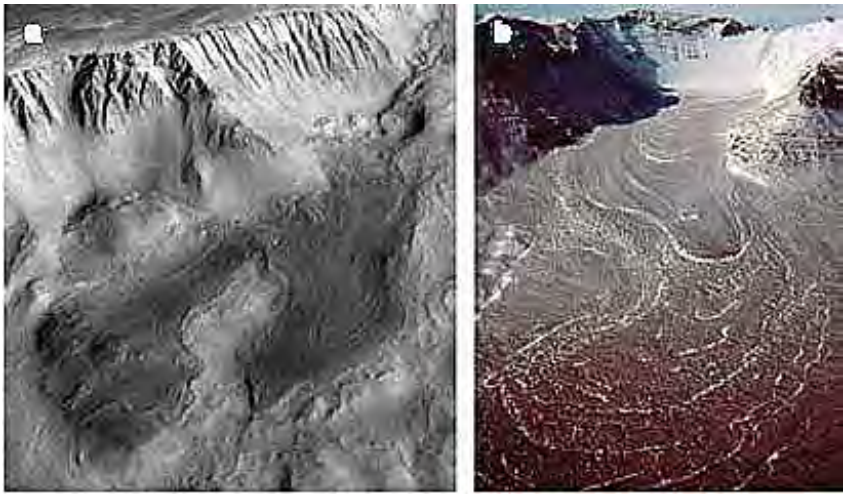
Finally, among other polar landform curiosities is this set of straight, equi-widthed troughs noted in the North Polar cap:

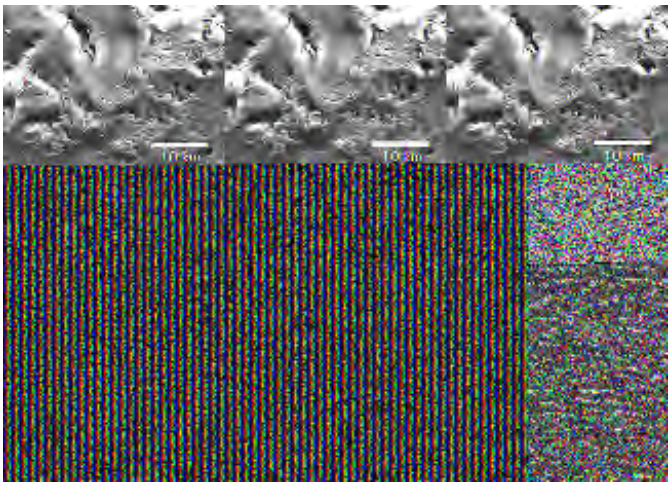


Lest you get an impression that ice is found only at high latitudes on Mars, we state succinctly here (treated in detail on page 19-13) that evidence is strong that water ice is present beneath the martian surface (as permafrost analogous to that found in Alaska and Siberia) over much of the planet. But, being dust and rock fragment covered, it is not visible in images of those regions.

The above review of polar ice features is based on solid interpretations of present-day features associated with frozen water. Much more nebulous is the question of whether in the martian past there was much more widespread ice in forms we associate with glaciers on Earth. This view is not universally held by Mars investigators. But geoscientists like Victor Baker, James Head, and William Hartmann III had favored interpretation of many features at latitudes lower than the polar regions as indicating the presence of ice either in "mountain glaciers" or in sheets (but probably much thinner than Pleistocene glaciation on Earth) or in periglacial and permafrost environments (this last setting is fairly widely accepted now). Several even place the last glacial age on Mars as recently as 10 to 2 million years ago. More signs of glaciation are present in the martian northern hemisphere. Here are some features that have been cited as indications of extensive glaciation; keep in mind that several shown here can have other explanations.

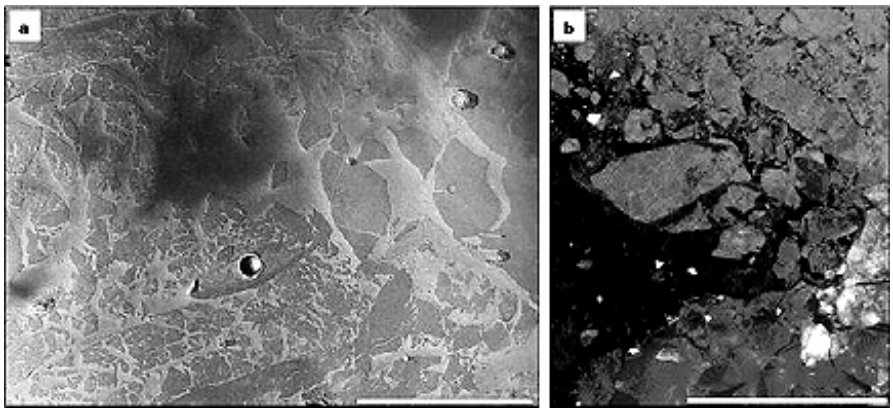
The next three illustrations show deposits of materials in lobe shapes that resemble certain terminal moraines associated with mountain glaciers.



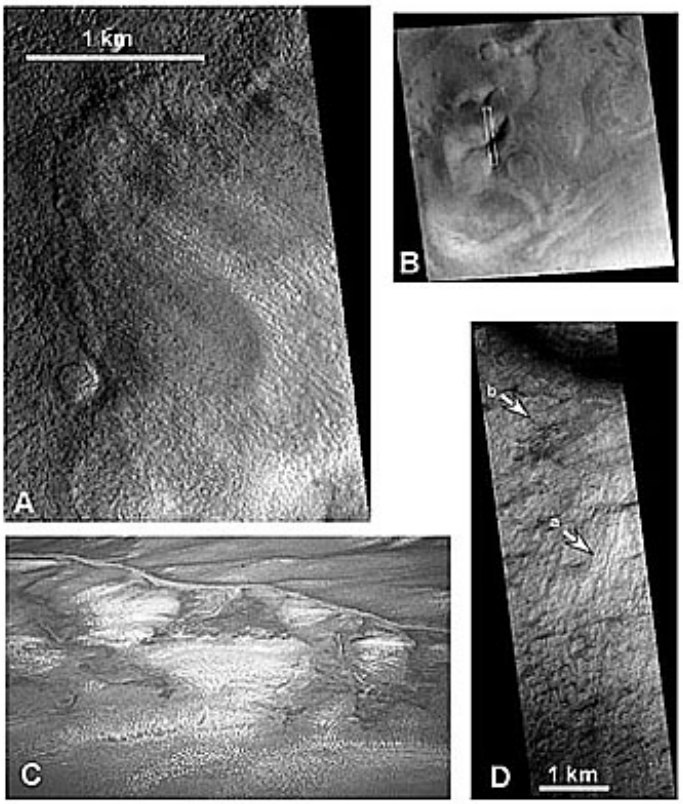


This image shows surface features in the south polar region developed beneath the ice cap which has shrunk to a minimum at the time it was viewed; the landforms have aspects indicative of both glacial and aeolian processes:

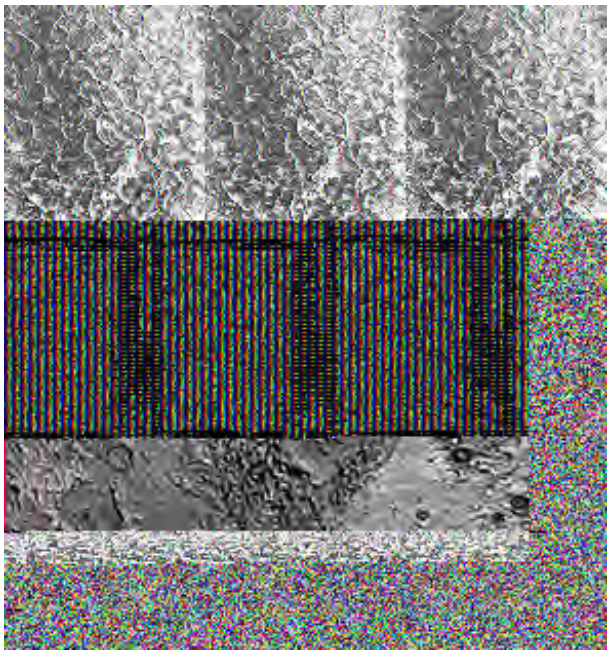
On Earth ice rafts make distinctive patterns at the surface when they are covered by deposits and then melt. This is shown at the right in the next illustration, with a martian counterpart on the left.



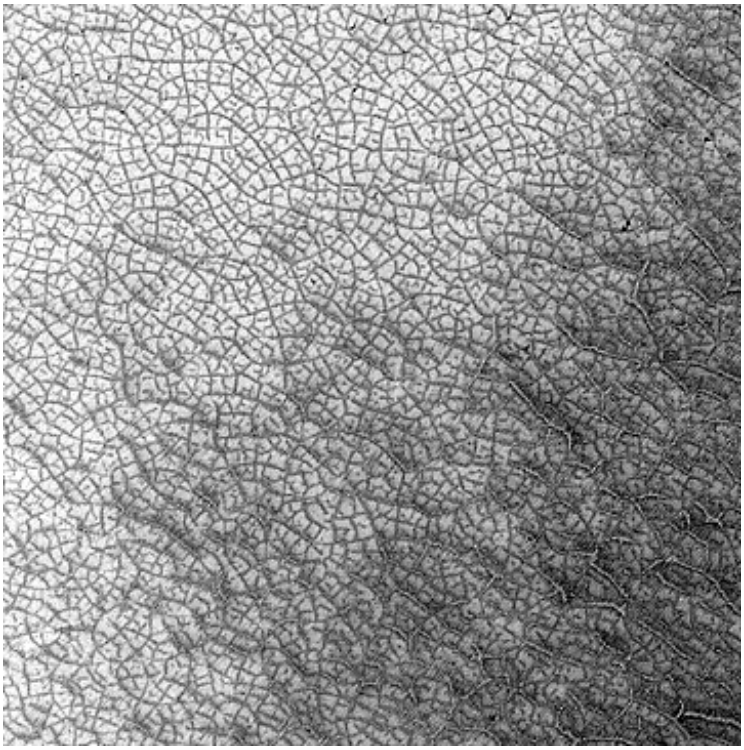
The next figure includes four martian surfaces that have been interpreted as cryoturbation features in a periglacial environment:



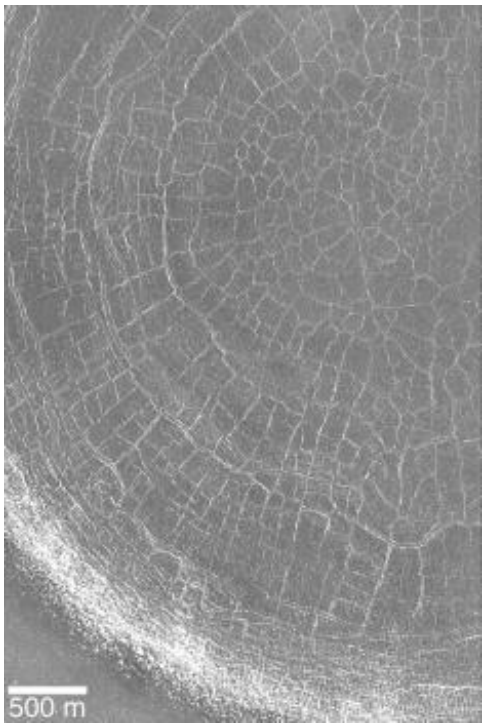
Polygonal structures of varying sizes appear on Mars. Some consider these to be volcanic in origin (see next page); others cite them as related to ice-produced features in deposits related to glaciation.



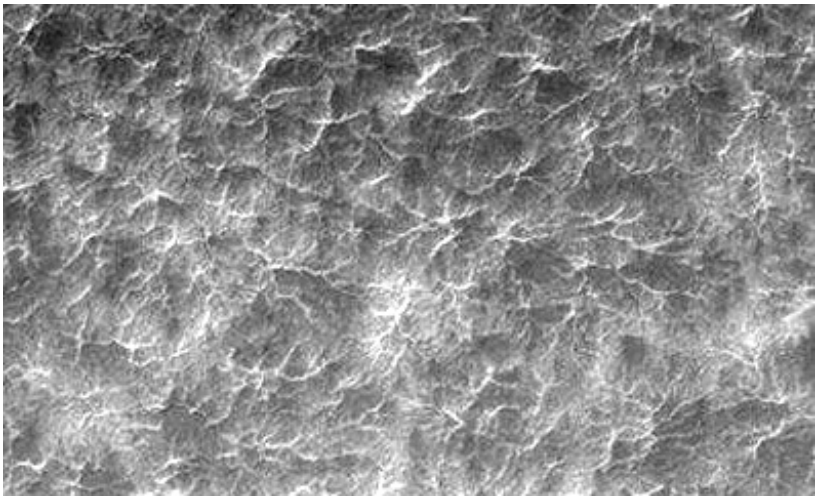
On Earth, one feature that abounds on ice sheets is sets of intersecting cracks that produce what is termed "polygonal" ice. This has been observed also on Mars in the polar regions where is is one type of patterned ground. It usually shows up after the warming of the polar region sublimates the coating of carbon ice. Here is an example from the South Polar ice cap.



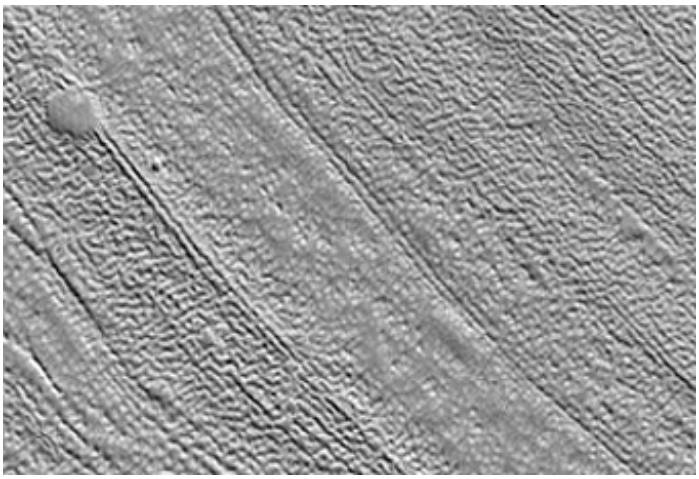
In this example, the polygons are confined to the interior of a large impact crater. If related to ice, one must presume that water filled the crater and then froze.

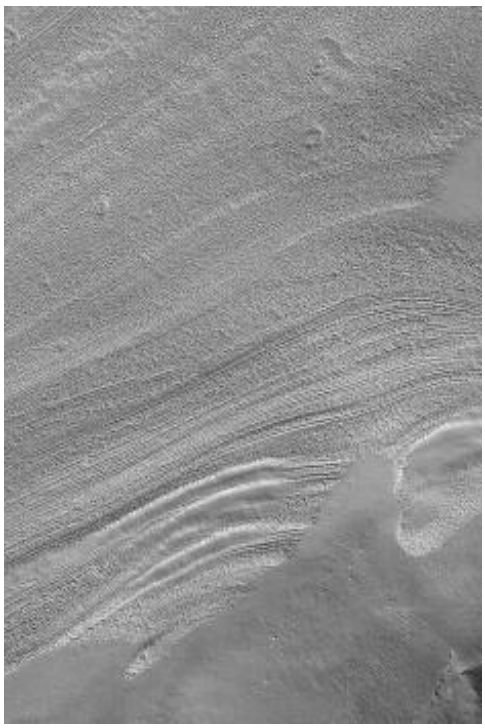


This image shows thin ridge like features over a wide area that seem similar to terrestrial terrains in permafrost regions.



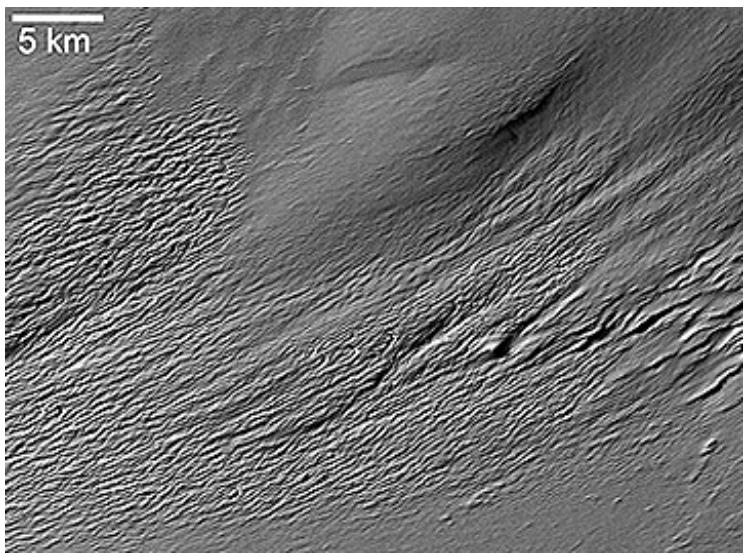
The next three images show terrains marked by linear patterns. These could be glacial, or alternately fluvial or volcanic in origin; see captions for further details:





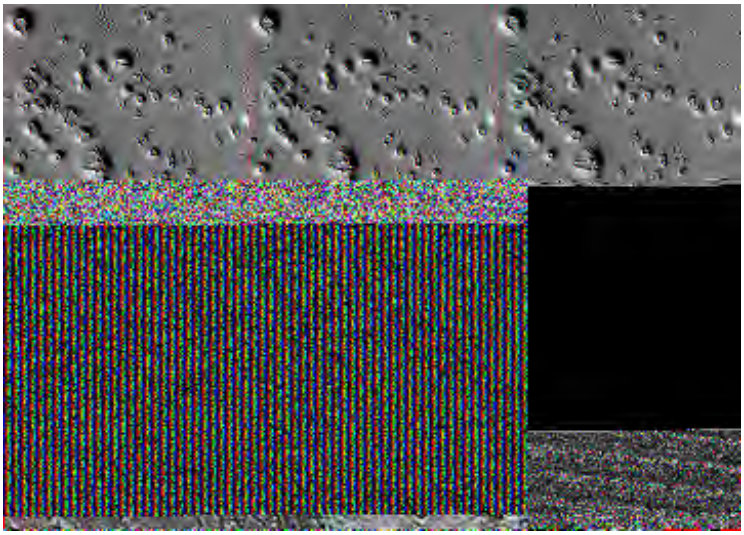
The last two are associated with what is called fretted terrain (see next page). Such a terrain is diverse in its nature, being mostly erosional but with some deposition.

This next scene is ambiguous. It may show aeolian features (close-spaced yardangs; see above), or weird fluvial deposits, or possibly drumlin-like elongate hillocks of glacial origin.



On Earth, when lava outpours beneath an ice sheet it can produce small small protrusions (seen

after the ice leaves) or lava cones. This next image shows what has been interpreted as a lava cone field on Mars, now exposed after glaciation ceased.



We close this topic of martian glaciation by pointing to the opinion of some investigators that many of the thin deposits of sediments discussed on the 19-13 page sequence may have a glacial origin. Most still favor a lacustrine or marine origin.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

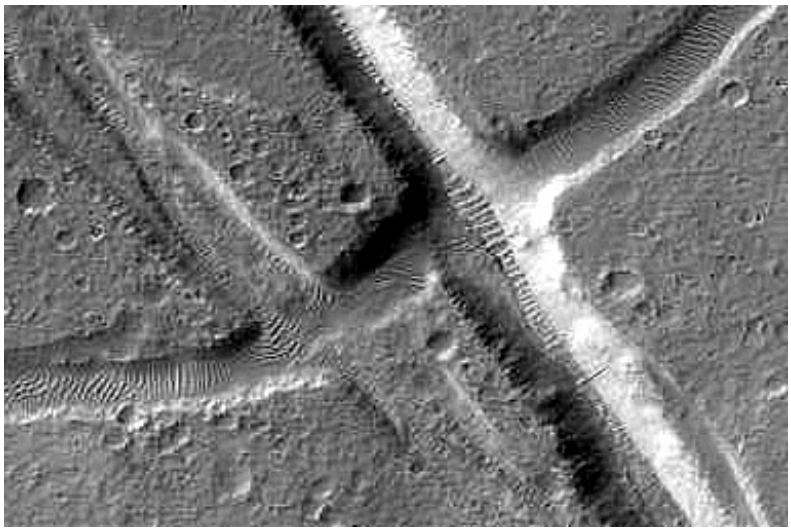
Three prime features of the martian landscape are volcanism and volcanoes, linear structures indicating some tectonism (but not plate tectonic movements of crust), and impact craters. It appears that the bulk of the martian rock materials are volcanic, probably of low silica (basaltic) nature. The martian crust in its broad or first order state can be subdivided into old, heavily cratered terrain in the southern hemisphere and younger, less cratered terrain, with abundant volcanic features in the northern one. The martian surface displays both conventional (i.e., similar to Earth's) and exotic terrains and erosional/depositional processes. One of great importance is sinuous channels that are much like river beds on Earth. This holds big implications: that Mars once had, and may still have, some water, probably for the latter part of its history in subsurface storage. Mars may also have had in the past enough water to develop large lakes or ocean-sized bodies. Like the Moon, Mars has large-scale erosional and depositional units which give it a general stratigraphy that allows geologic maps of its surface to be drawn.

Martian Landscapes: Linear Features, Volcanoes, Impact Craters, Channels; Exotic Terrains

A plethora of spectacular surface images have accrued from the Mariner 9 and Viking missions. Below, we show only a few, to whet your intellectual appetite, but for the curious, consult these two references for many more pictures: *The Geology of Mars*, T.A. Mutch et al., 1976, Princeton University Press and *Viking Orbiter Views of Mars*, C. R. Spitzer, Ed., 1980, NASA SP-441.

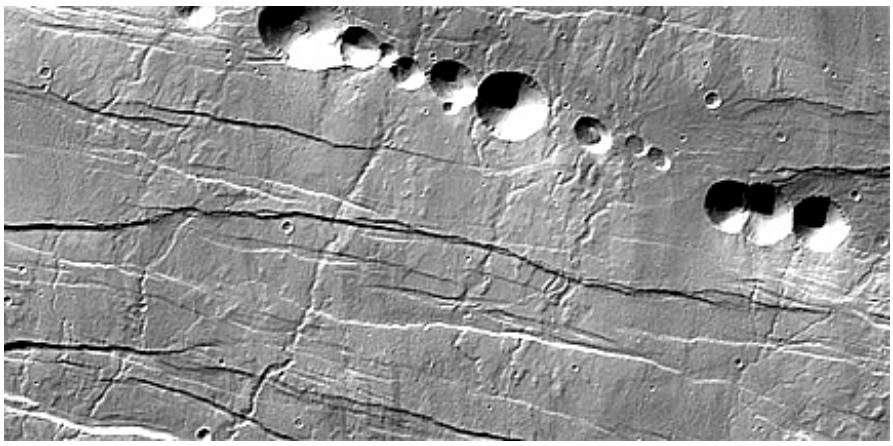
As has been alluded to before in this subsection, Mars seems to consist of two dominant terrains - the northern half is mostly a plains made up by volcanic rocks with some conspicuous volcanic structures, but is only moderately cratered; the southern part of Mars also consists of an igneous (probably volcanic) rock crust but that is much more heavily cratered and has not experienced major surface reworking or lava paving in the last billion years or so. Tectonic features - almost entirely faulting (probably tensional) - are found over the entire planet.

No firm evidence of large scale folding in the Mars rocks, which in many places are layered, has been found to date, although inclined strata have been noted in crater walls. This suggests that compressional force activity is very uncommon on Mars, i.e., plate tectonics as acts on Earth has not taken place. Faults are frequent, however, indicating some extensional forces have pulled the martian crust apart in places. Directions of tension have changed over time. The first image on this page shows three such tensional faults or grabens:



This scene clearly indicates the "Law of Cross-cutting" included in the basic review of Geology in Section 2. The law indicates relative ages. The oldest faulting is indicated by the structure starting in the upper left that slants towards the lower right. It is cut by the fault running upward to the right. The youngest fault, cutting this second fault, appears to the right of the first and appears the freshest (note the sand dunes within it).

Linear faults on Mars are fairly commmon. This image shows a set of parallel faults; along one is a chain of pit craters (surface material sinks into the fractures). Some have interpreted this alignment and association with fresh-looking normal faults as a sign of recent movement in the martian crust, causing "marsquakes" as a consequence:

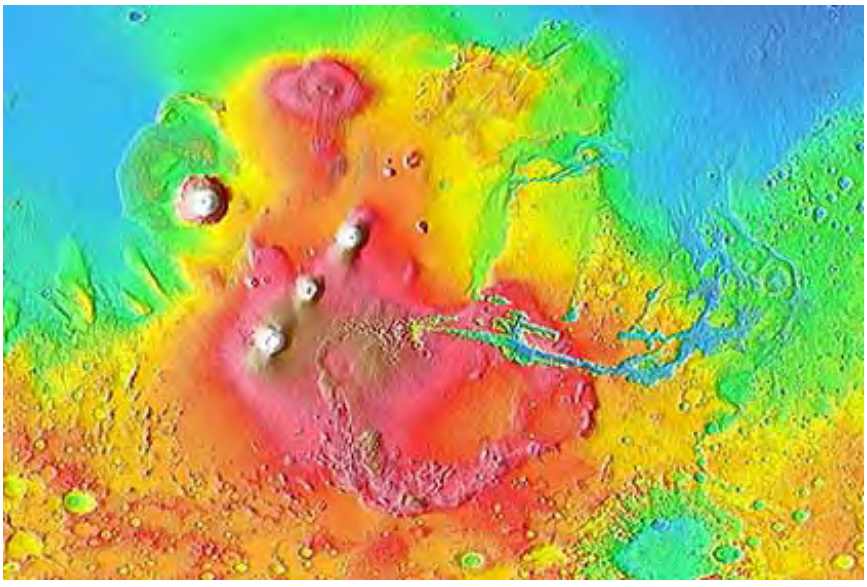


One type of fracture, called "deformation band", is important as a channel for ancient groundwater (a surmise based on analogs found on Earth):

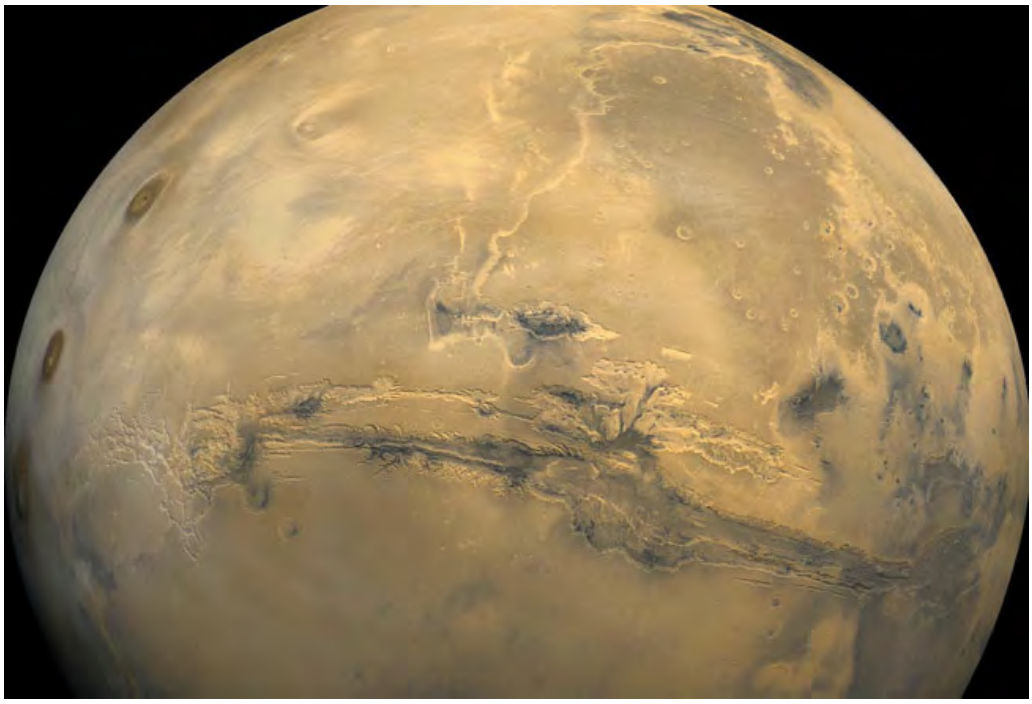
In February, 2007 an image made by the Mars Reconnaissance Observer was released, largely to support a claim that water-related alteration of rocks had been detected in the scene. The image also reveals the best evidence yet of faulting with complications.

In the upper half of this image, there are curved bands which may be strata. A long line separates this block from that in the lower half, where brown lines trend towards the upper left. These relationships may be a sign of a fault, whose plane intersets along the line, and movement (horizontal? vertical? oblique?) of the upper block with its stratalike bands as an offset that brings it against once separate rocks showing a very different assemblage of beds (these are interpreted by others as joints along which alteration is visible).

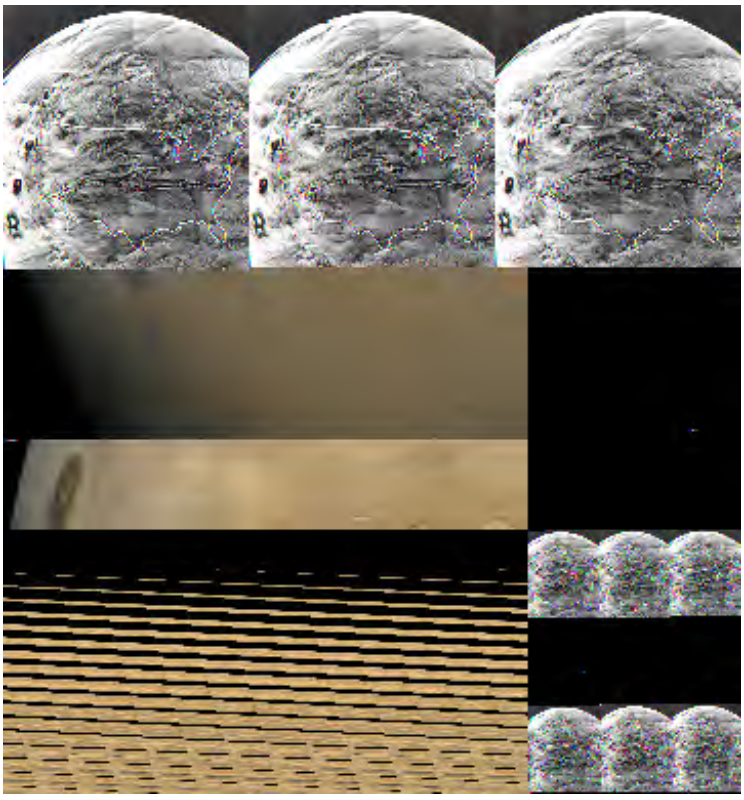
Structural control by faulting may account for the location of some first-order topographic features on Mars. Many of the scenes depicted below that illustrate the topics listed in the page heading occur in and east of the Tharsis region of Mars. To place these features in their physiographic context, here is a MOLA derived topographic map that includes some of the landforms we will visit:



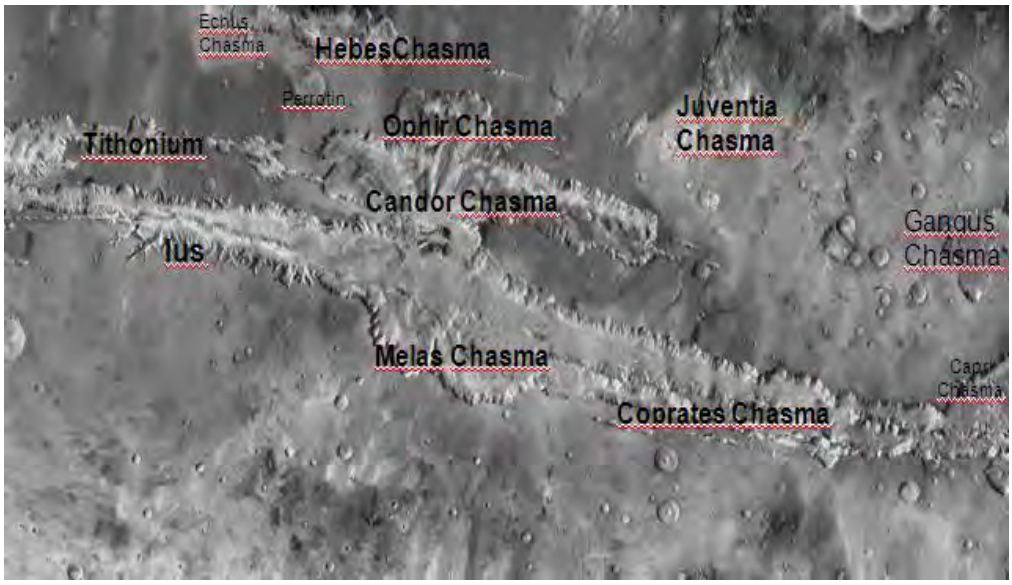
We turn now to the greatest trench system in the ground ever discovered in the Solar System. This is Valles Marineris and some subsidiary trenches. It is a challenge to find an image that shows the entire system but the view made by Viking fits the requirement:



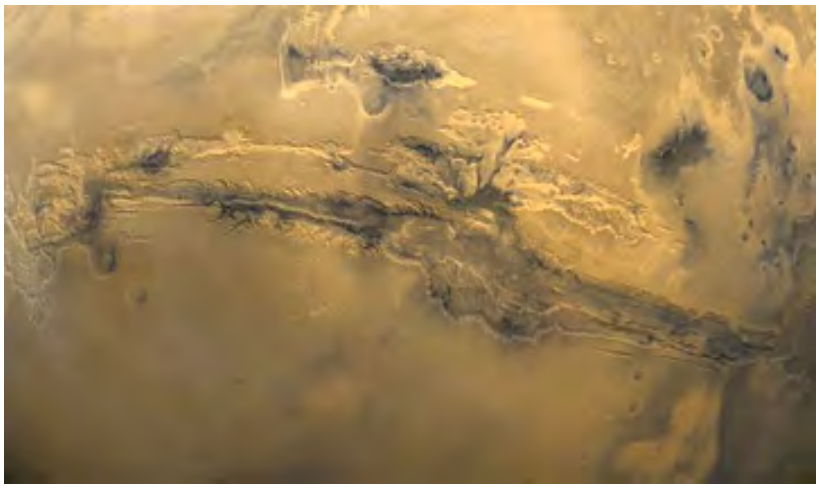
To get a sense of just how big the trench system is look at this Mariner 9 mosaic, extracted from the near hemisphere view, here centered on this most conspicuous feature on Mars. Valles Marineris extends nearly 4,000 km (2,486 mi), can be up to 200 km wide, and attains depths between 2 and 7 km (1.25-4.35 mi). When the outline of the 48 contiguous United States is overlain on mosaic, the eastern edge of Valles Marineris touches the Outer Banks of North Carolina and its western edge reaches to Central California.



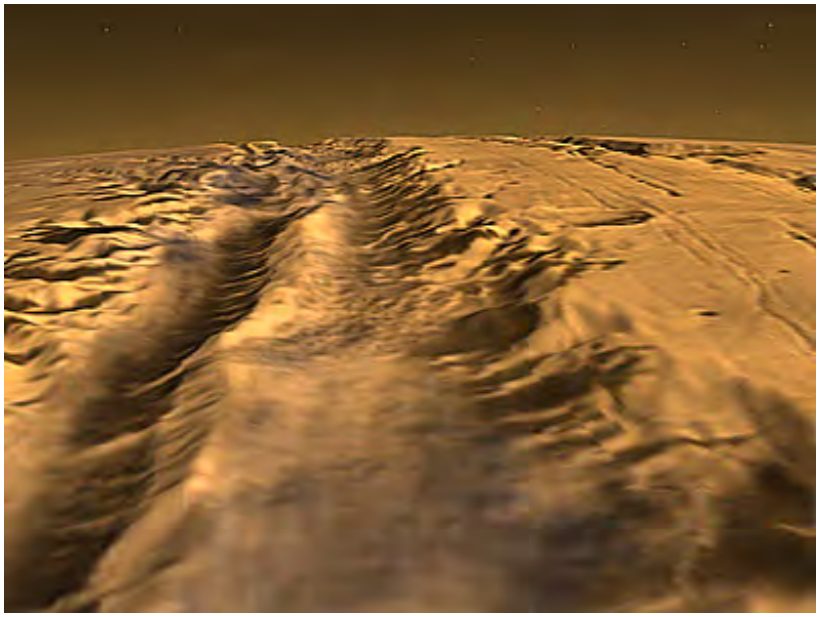
The main features of the Valles Marineris system are labeled in this view:



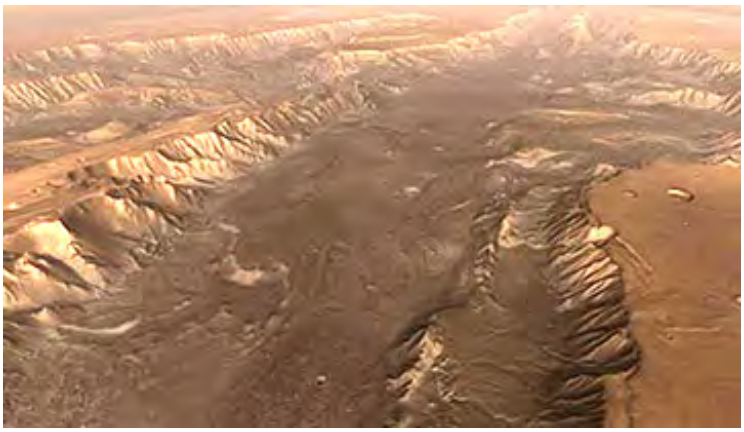
Part of the Valles Marineris around Candor Chasma is shown in color in this Viking image:



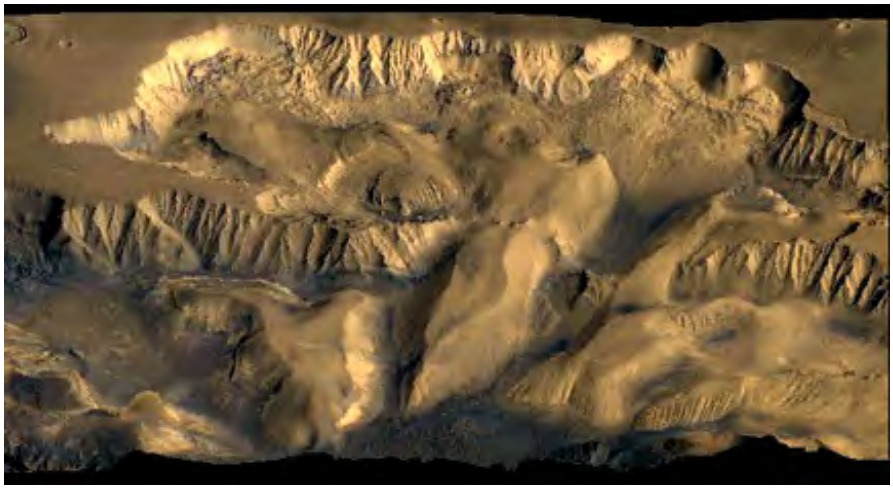
This is what you would see if flying over the Valles (canyon). The image is made by combining a Viking view with MOLA data; there is no vertical exaggeration.



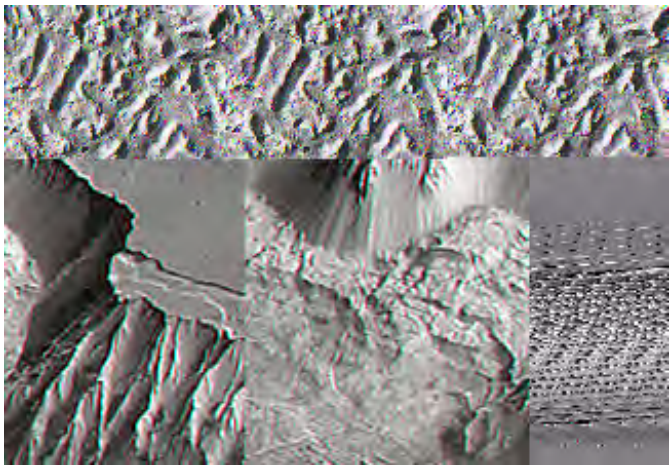
Even more impressive is this low angle view looking at the main canyon, made from Mars Express and MOLA data:



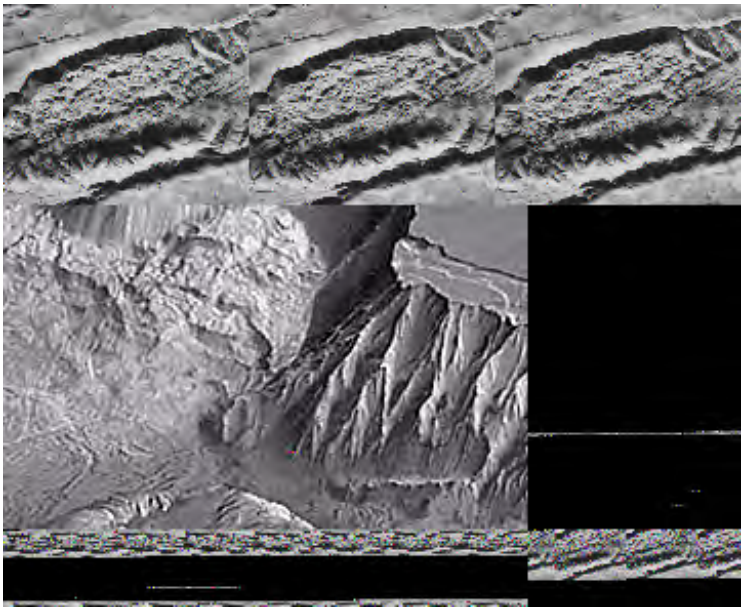
The canyon is actually a series of structural troughs, produced by faulting, radial to the Tharsis bulge to the northwest, which rises some 11 km (6.8 mi) above the surrounding plains, on which are the three dark-shield volcanoes , named on the preceding page. These volcanoes reach about 10 km (6.2 mi) above the bulge. A look inside the canyon wall, along a segment called South Candor Chasma, conveys the sense of steep slopes, perhaps furrowed by water erosion, and basal landslide deposits.



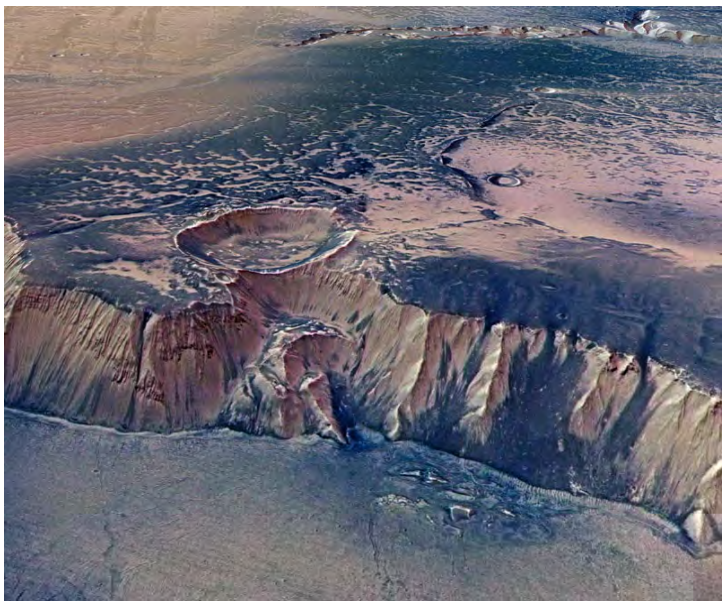
Another landslide into Valles Marineris appears here; below it is hummocky terrain often found in the deposits at the slide's foot.



A close-up of a landslide in Valles Marineris gives details of the massive debris pile-up, as material pulled away from the steep canyon wall.

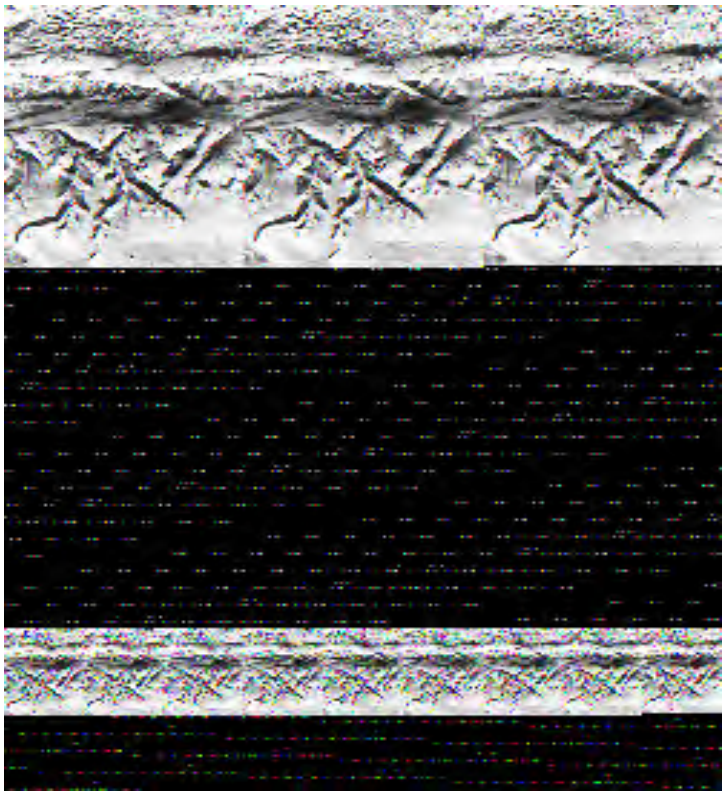


Some of the cliffs in the valleys leading into Valles Marineris are dauntingly high, as for example this 1000 meter scarp face in Echus Chasma:



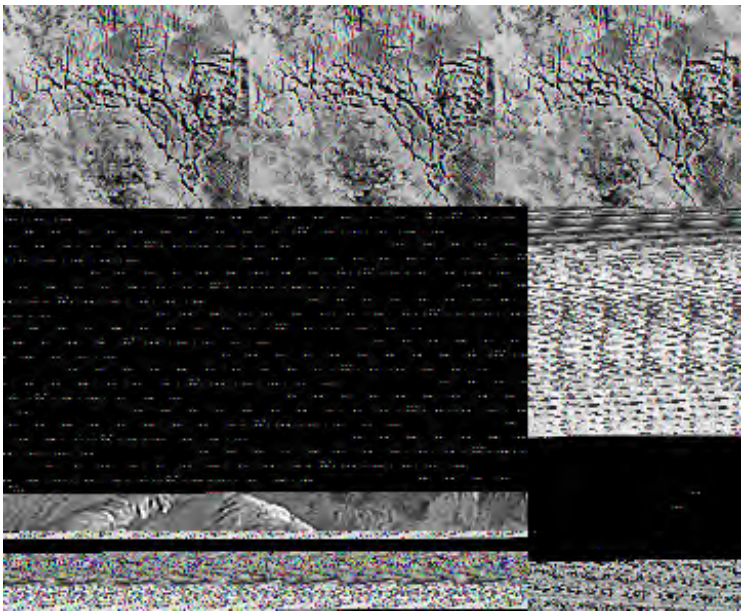
Along some edges of Valles Marineris are what appear at first to be tributary valleys. But they don't enter at levels equivalent to the floor base. They seem to criss-cross in a pattern that suggests tectonic control. One proposed explanation has them as due to subsurface water

sapping.

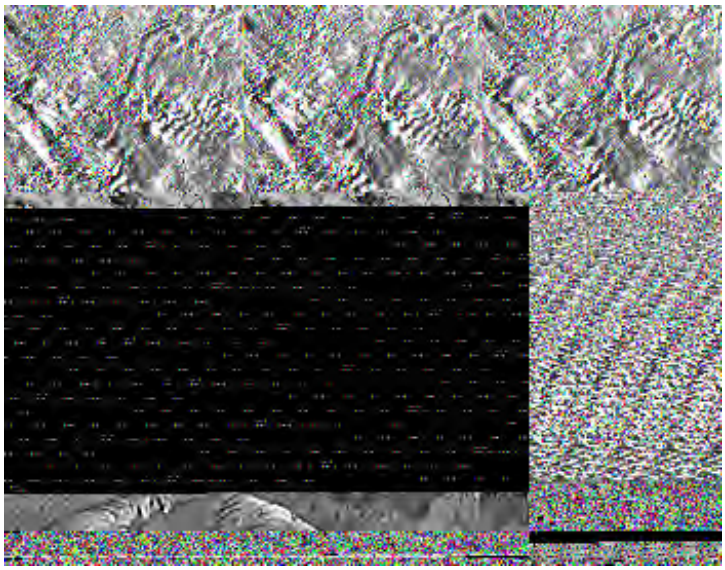


19-39: What general explanation can be given for the formation of Valles Marineris?
ANSWER

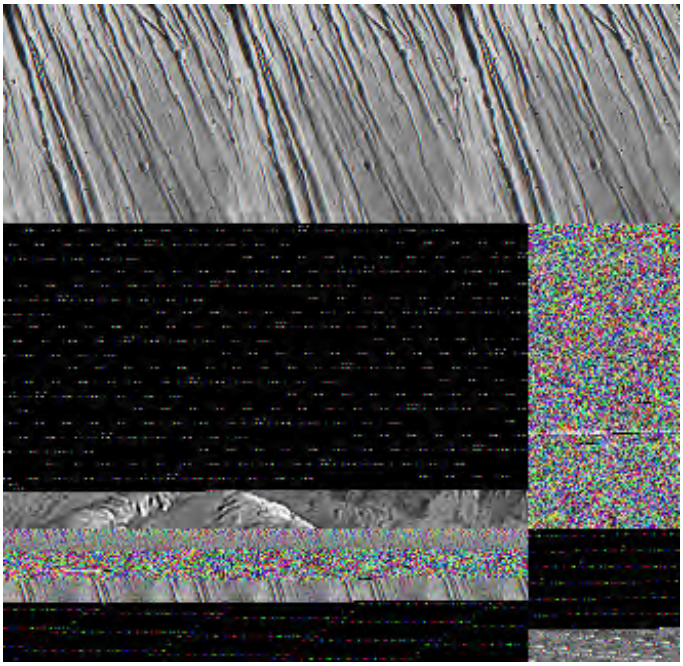
Other tensional grabens are found in various parts of Mars, especially in the newer terrains. These can occur in intersecting networks, such as below which portrays Noctis Labyrinthus in the northern Tharsis region. This tecto-morphological feature is also called *fractured terrain.*



Sets of subparallel fractures cut across the terrain on the flank of the Tharsis region (Tharsis is a huge upbulge of martian crust more than 4000 km across and 10 km higher than surrounding lowlands at its top; Olympus Mons and the Tharsis volcanoes attest to it volcanic nature). In overall pattern, the sets are radial to the Tharsis apex. Here is one such set which cuts across older craters (but several younger craters superpose on the fractures).

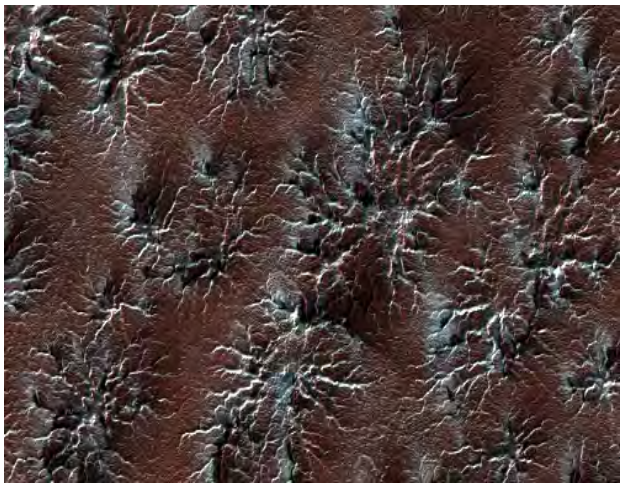


As is often true for volcanic terrains (such as the East African Rift), sets of close-spaced parallel grabens (fault-bounded downdrop blocks of crust) related to tension induced by loss of support after lava withdrawal have been found also on Mars:

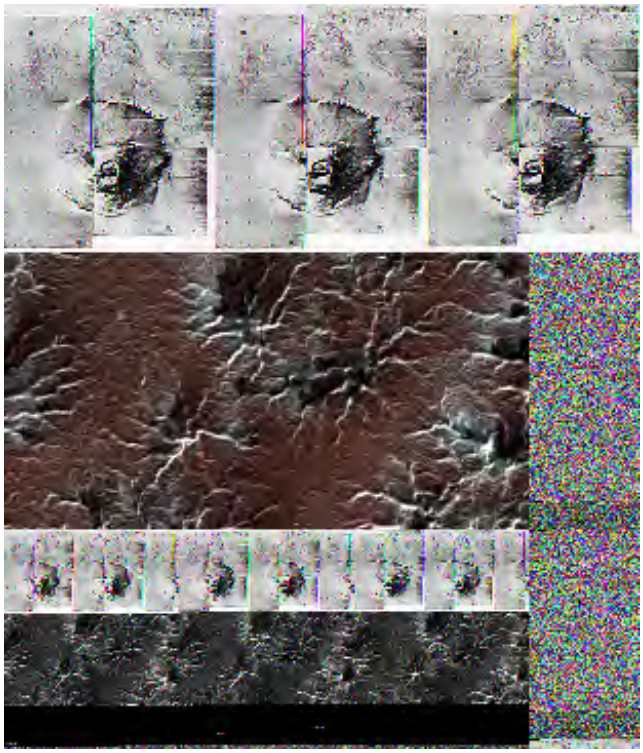


At the other extreme, short fractures may appear as isolated gashes, as seen here. These may be incipient or early stage breaks in a surface undergoing only moderate tensional stress.

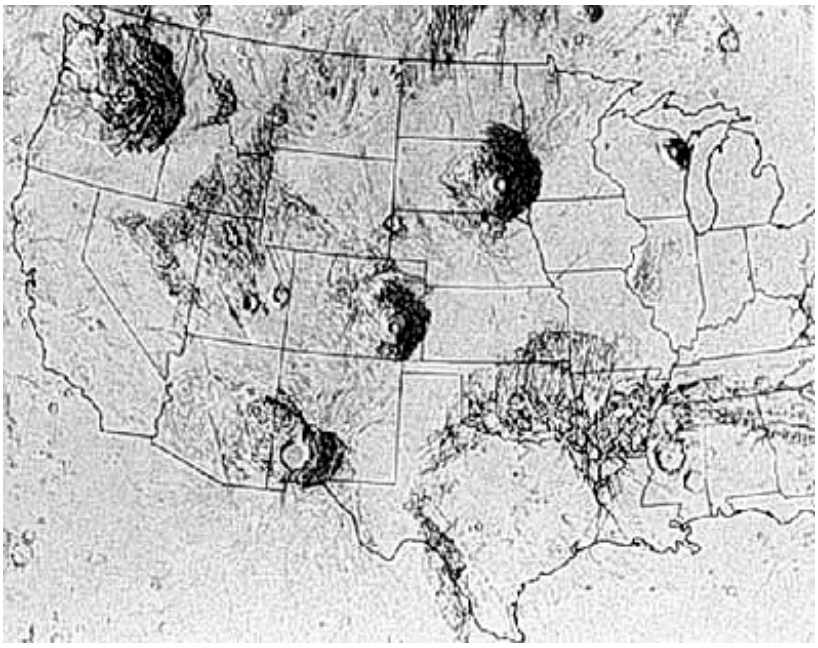
Radiating fractures associated with small domes or swells produce a feature on Martian volcanic lava terrain known colloquially as "spiders". This is a good example:

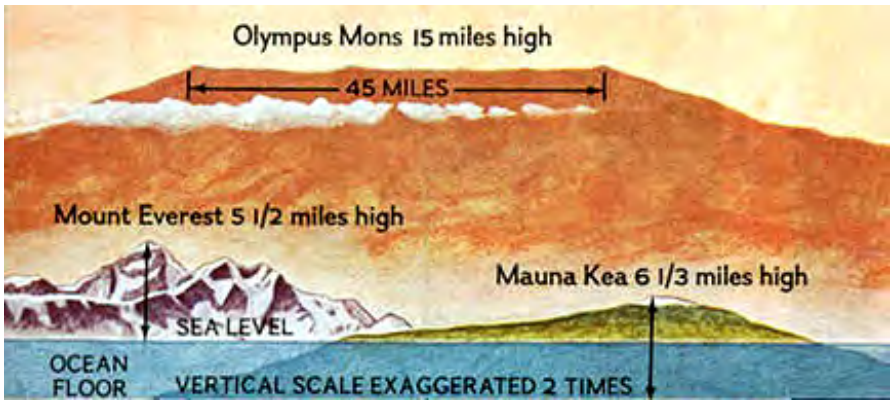


Next, we switch our review to volcanism on Mars. In the Tharsis bulge region, some 4000 km across and 10 km above the mean martian elevation, are four of the biggest volcanoes in the Solar System. The huge structure alone in the western end of the Tharsis region is known as Olympus Mons, which is a broad shield volcano (now dormant), many times the area and volume of the big island of Hawaii, which consists of basaltic outflows from several major vents. Olympus Mons has a median diameter of 625 km (388 mi) and a height of 25 km (16 mi). We show first the famous discovery image from Mariner 9 (top), then a color version from Viking (center).



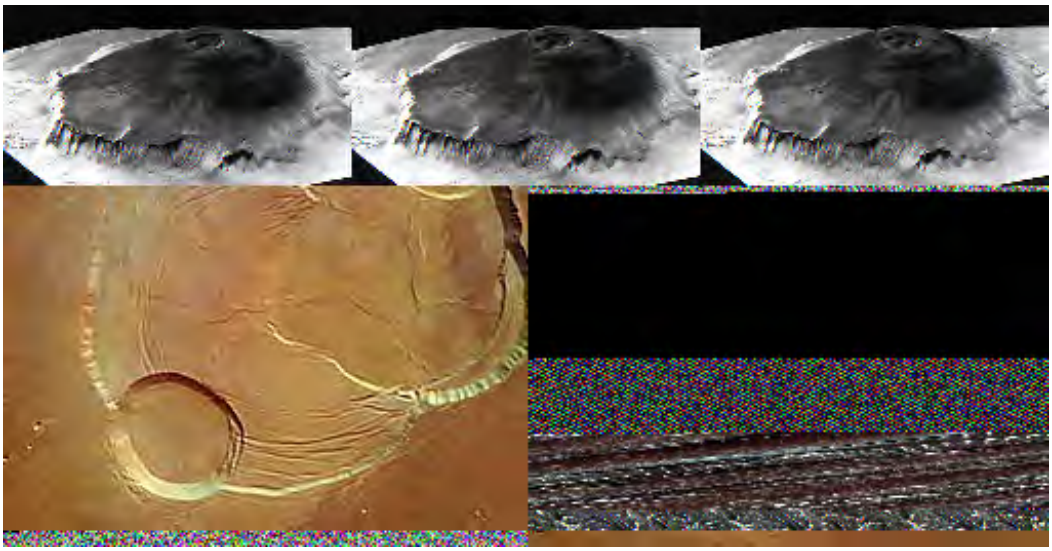
Olympus Mons is the largest volcanic structure known on any of the planets. The major volcanoes of the Tharsis region are all huge by Earth standards. This is self-evident when the next two illustrations are examined. The first plots Olympus Mons and its three Tharsis companions on a map of the United States. Only Mauna Loa on the Island of Hawaii can compete with the three. When plotted as cross-sections the size of Olympus Mons is even more awesome.





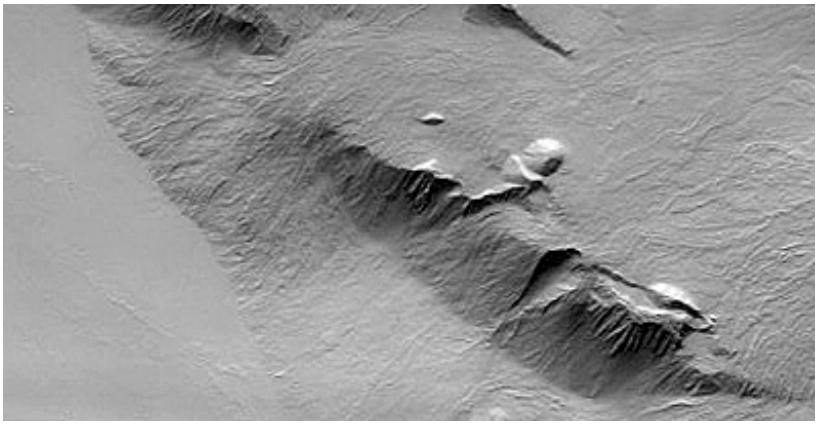
Olympus Mons' elliptical central caldera is 80 km (50 miles) in major axis. Here is a color view looking down from the Mars Express spacecraft:

A steep cliff up to 6 km (3.7 miles) high surrounds Olympus Mons, and stands out in the perspective view (below), derived by combining a Mars Express image with topographic data obtained from laser altimeter data. Scientists still debate the origin of this cliff, but some of them cite it as evidence of an escarpment resulting from wave erosion by an ancient (now vanished) ocean that may have covered at least part of Mars.



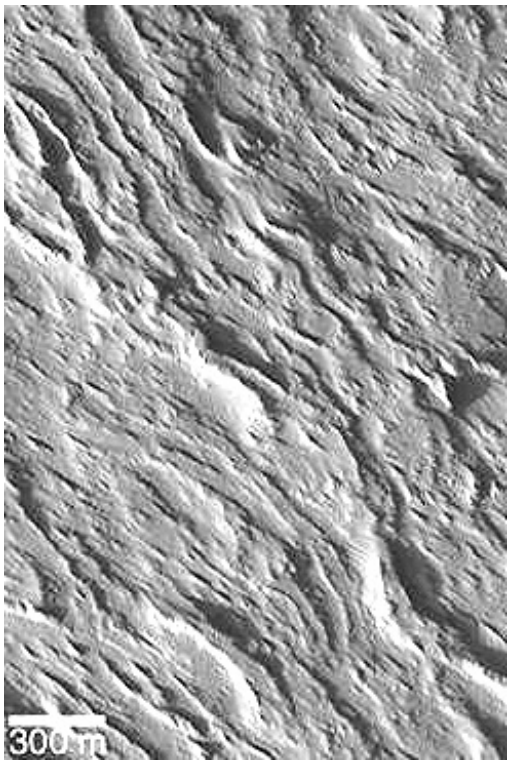
Mars Express imagery has also been manipulated to produce a view of the scarp (cliff) as though seen from the sloping plains beyond it:

Along the scarp these flows have spilled over into a moat-like shallow depression. This indicates that volcanism continued on Olympus Mons after scarp formation.



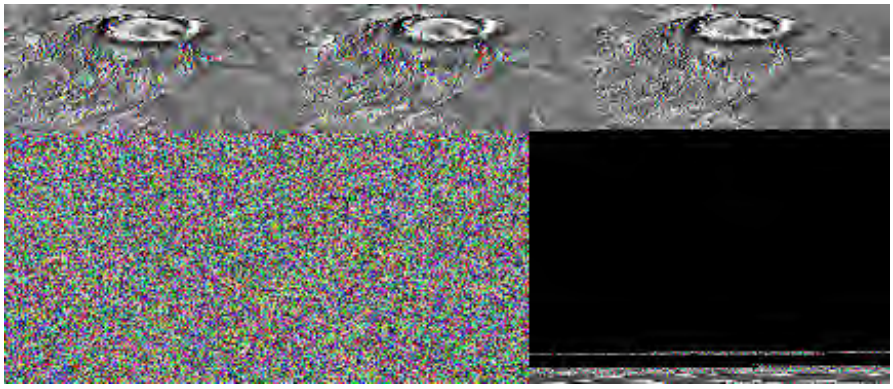
19-40: This steep cliff around Olympus Mons is peculiar and not characteristic of terrestrial shield volcanoes. Speculate on a possible origin. ANSWER

Once above the fringing escarpment, the slope of lower Olympus Mons is gentle - 1 to 3°. However, this incline allows lava flows to move downslope, as shown in this example from Mars Orbiter:

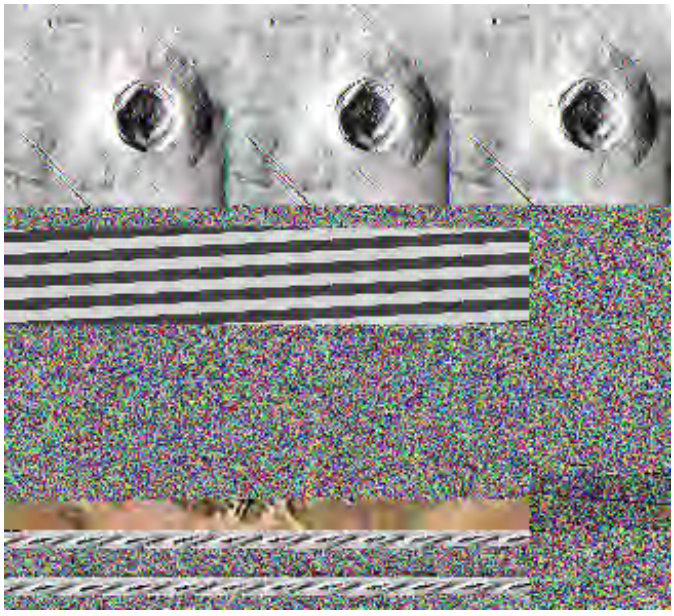


There are three volcanoes lined up in the Tharsis Montes area east of Olympus Mons; from south to north they are Arsia Mons, Pavonis Mons, and Ascraeus Mons. Each is a large shield volcano with a well developed central caldera. Typical is Arsia Mons, about 300 km (200 miles) wide at its base:

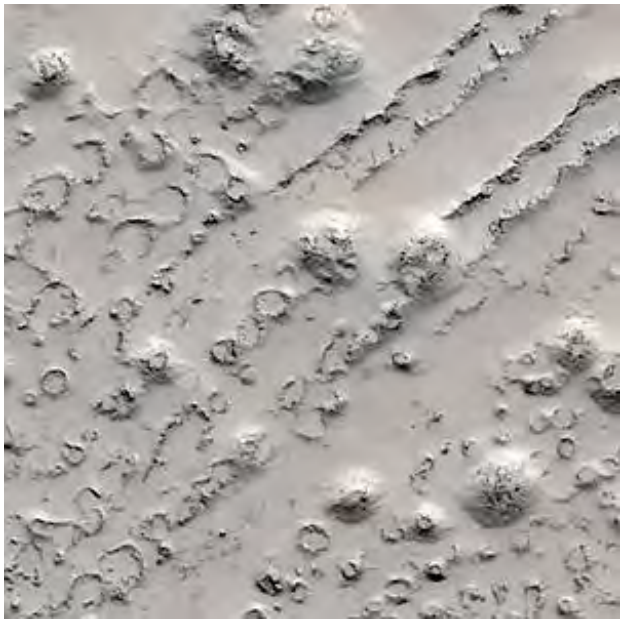
A side view made by combining a Viking image with MOLA elevation data gives this impression of Arsia Mons:



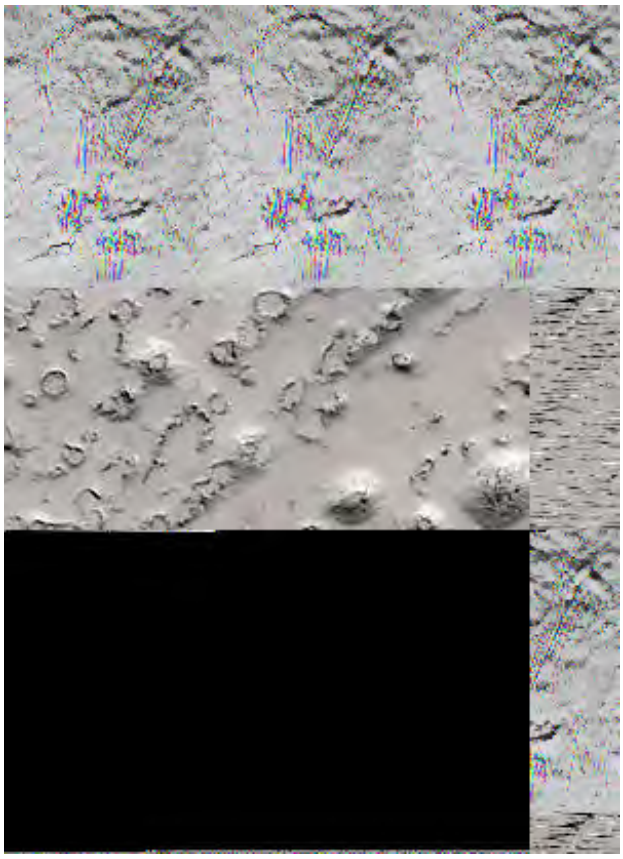
One of the best formed volcanoes in Bilbis Patera, which lies between Olympus Mons and the Tharsis group. As seen here by Viking Orbiter 1, the base of the volcano is about 100 km (62 miles) in diameter. The large caldera is like some volcanoes in the Galapagos Islands off the Ecuador coast.



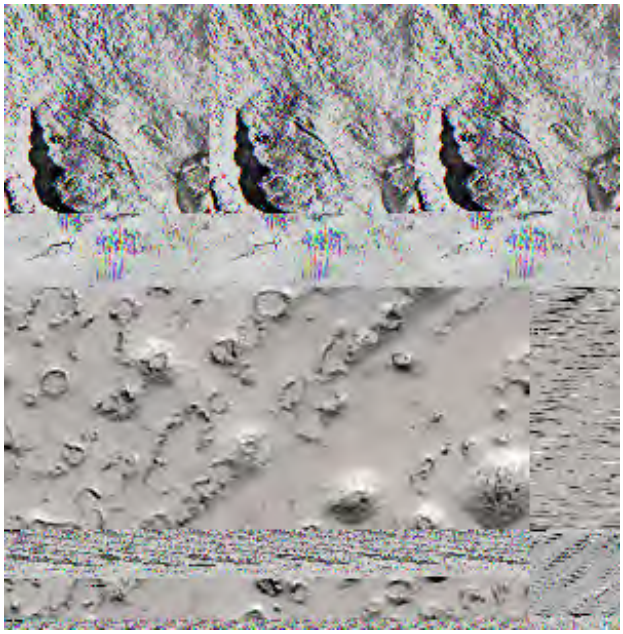
Domes are also found on Mars. Here is three examples:



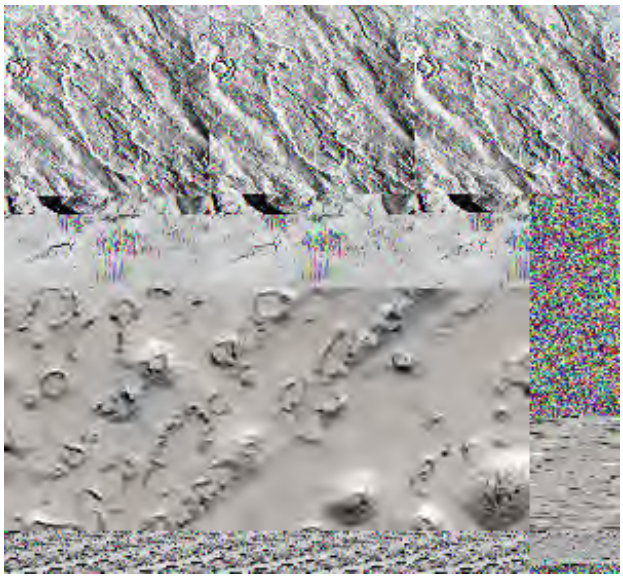
The largest volcanic complex on Mars is the great bulge in the northern Tharsis region known as Alba Patera, only a few kilometers high but 1600 km (1000 miles) in longest dimension. As such, it rates as the largest shield volcano complex in the Solar System. It is broken by a series of concentric fractures and a set of elongate, subparallel fracture grabens, as seen in this Mariner mosaic.



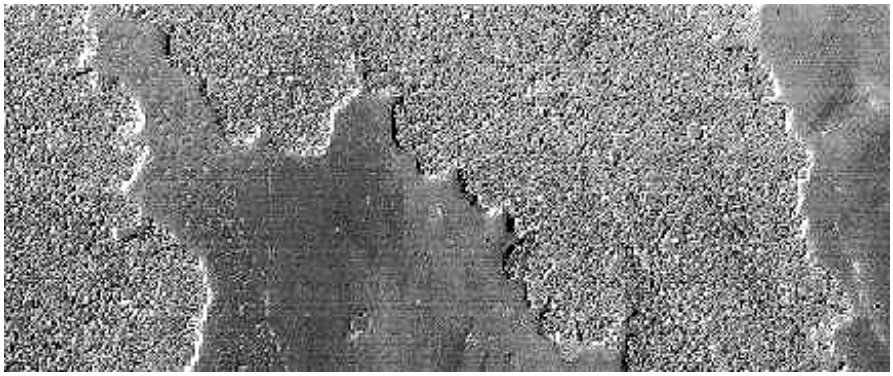
At Alba Patera's top are an older summit caldera (left) and a smaller, more recent one. Note the lava flows descending its gentle slopes.



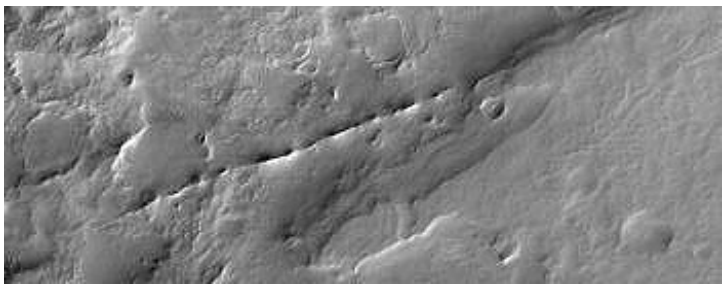
Alba Patera's flanks show numerous overlapping flows, indicating multiple periods of lavas extruded from tube outlets both at the caldera and along the slopes. In this Viking Orbiter image, the flows are flat-topped but steep-sided; volcanic ridges appear in the lower left.



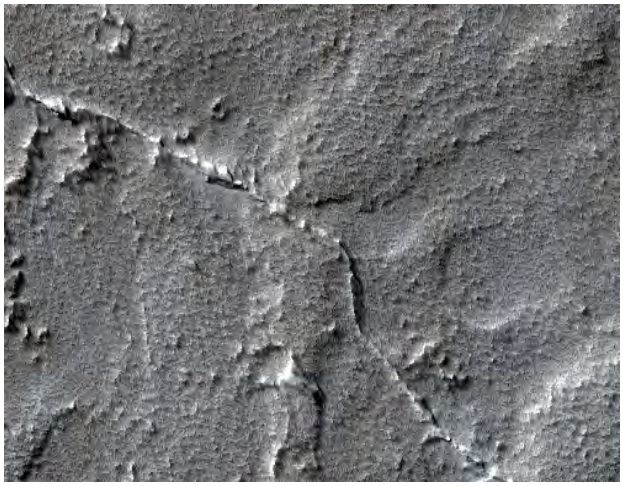
Some lava flows are regular, smooth-to-rough surfaced, and with definite steep fronts, similar to those observed on the island of Hawaii. Here are two examples:



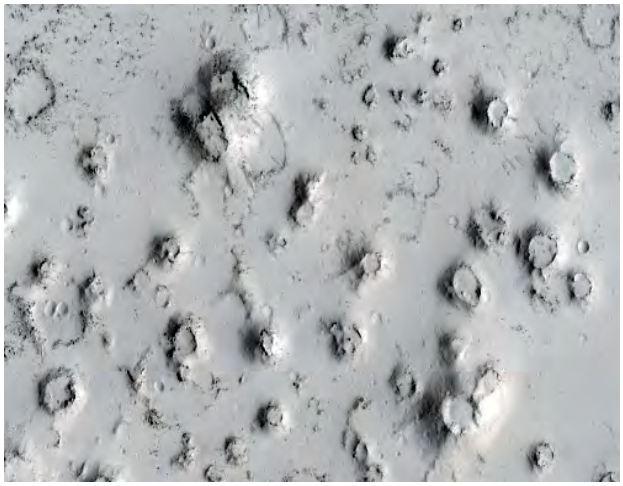
Lava flows issue from volcanoes, vents, and fissures. A narrow fissure can be filled with lava that hardens. As erosion removes its surroundings, the lava mass stands above the surface as a dike, as shown here:



Similar to a dike that has been brought out by erosion is this lava tube which served as a feeder for surrounding flows:

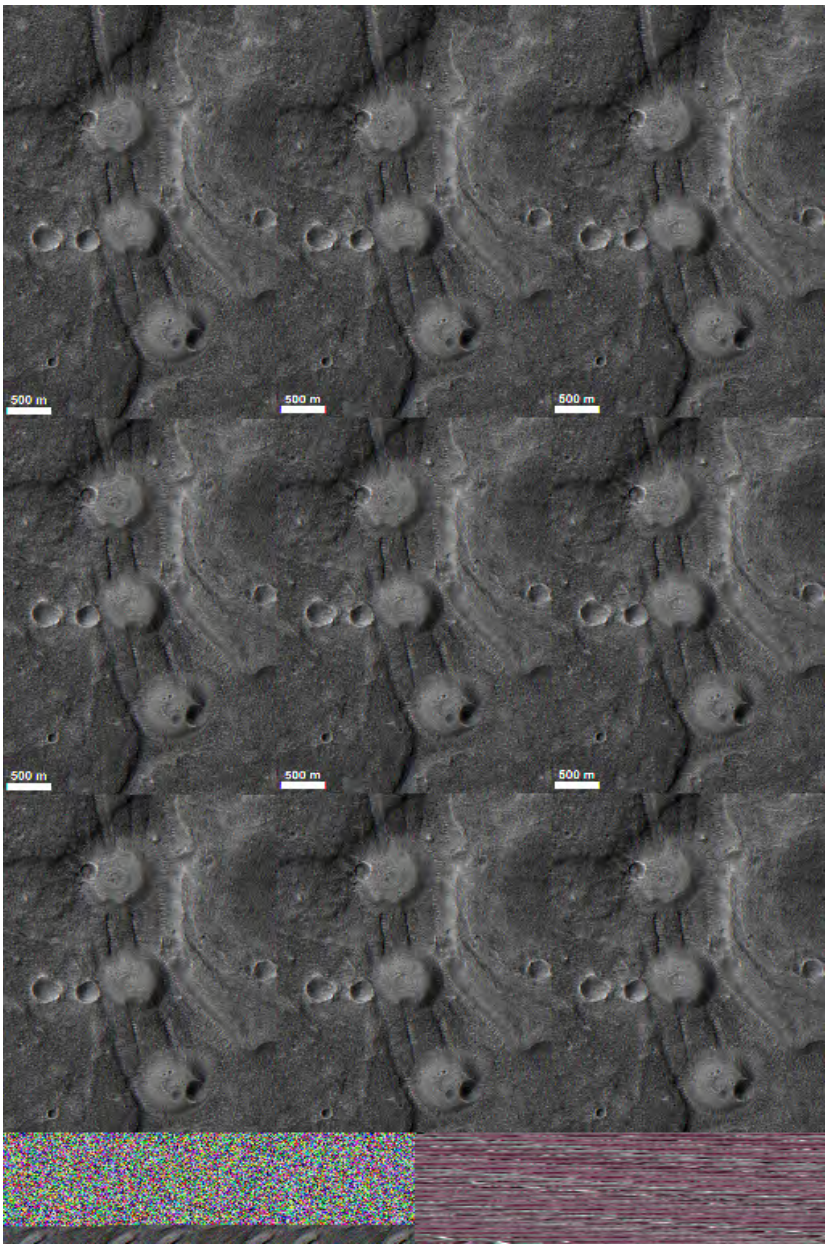


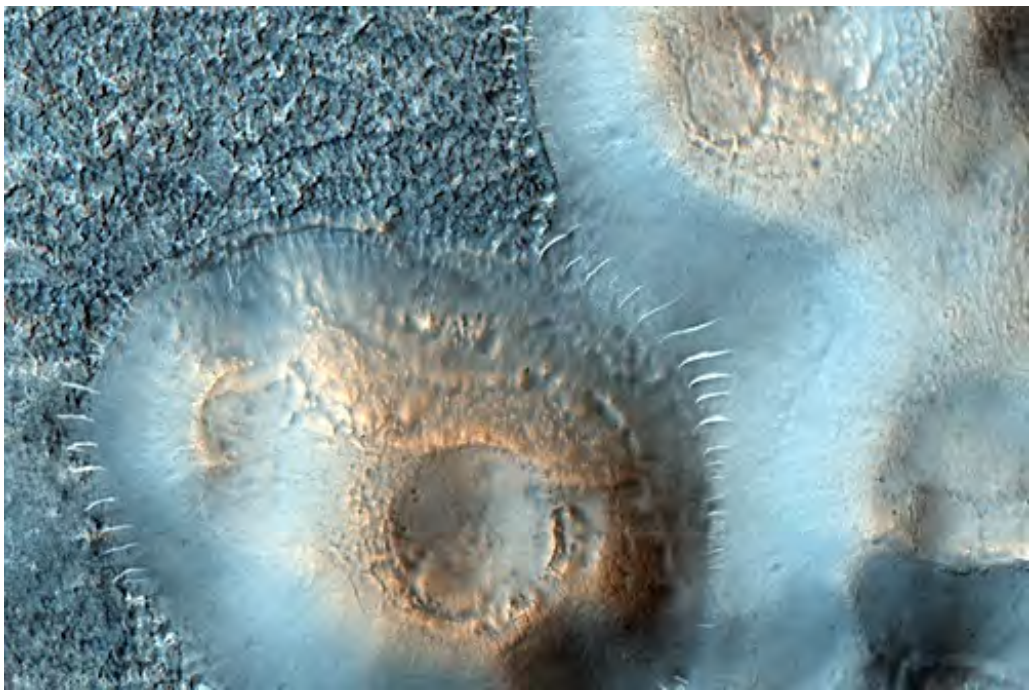
Smaller, more conventional volcanoes on Mars are known as tholii - an example is Ceraunius. Small ones are equivalent to large volcanic cones found on Earth:



Resembling the central caldera of a martian volcano but much smaller is the collapse pit. Here is an example - a string of pits - on the flank of Ascraeus Mons:

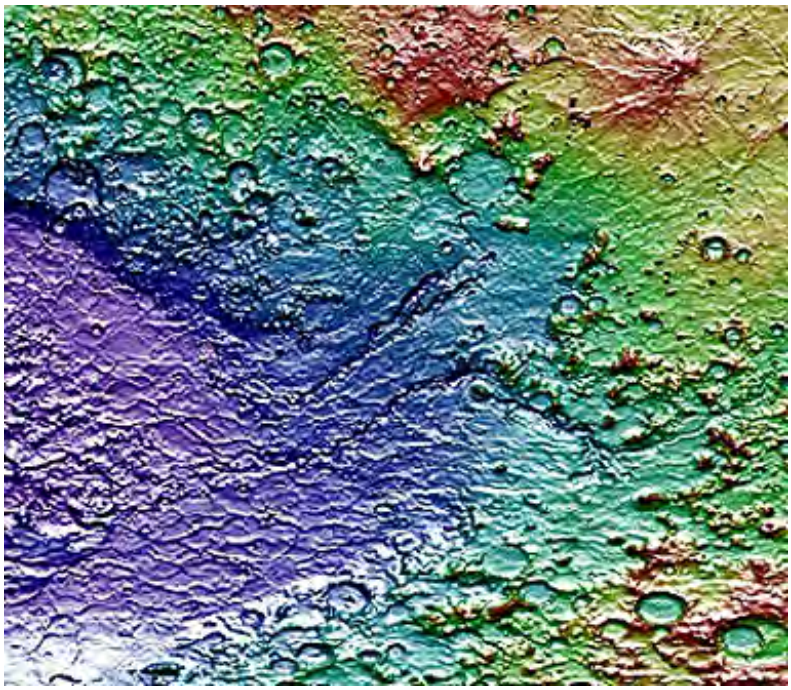
A special type of volcano on Mars has recently been proposed to be caused by escaping gas, fluid, and mud. These 'mud volcanoes' have their counterparts on Earth. On Mars they are observed in the thousands in Acidalia Planitia in the northern Plains. Their importance is that the fluid is likely water and hence such features are more likely to harbor at least single-cell life. Here are two examples:



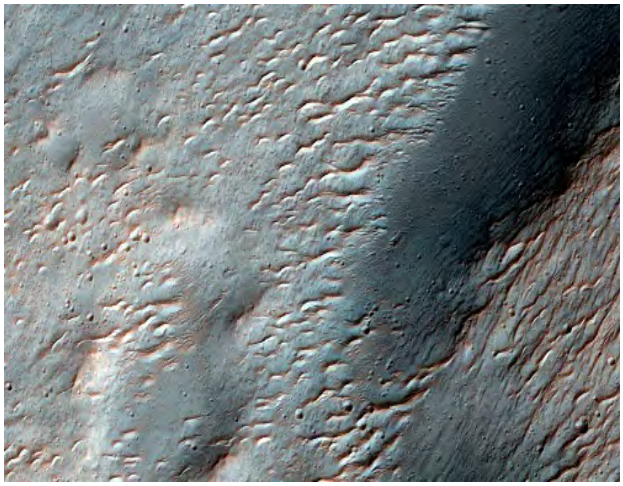


Among martian features believed volcanic in nature are linear ridges similar to the wrinkle ridges found in lunar maria. Here is a topographic map made from MGS MOLA measurements that

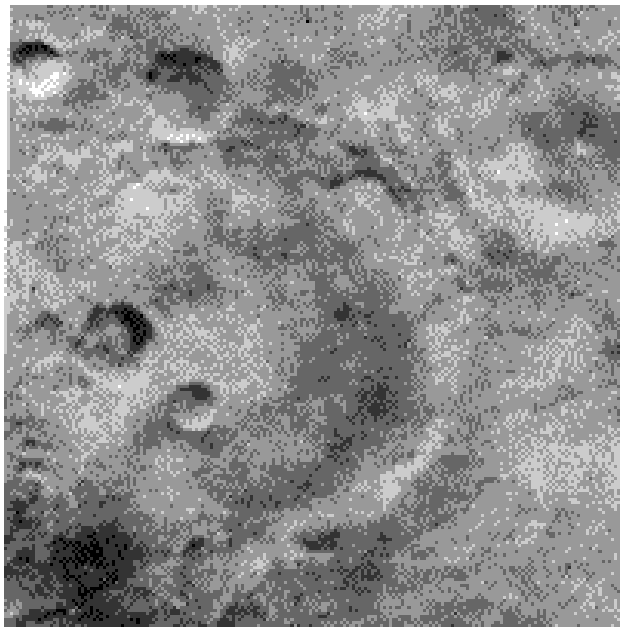
includes (in the purple) these ridges and shows the diversity of other landforms.

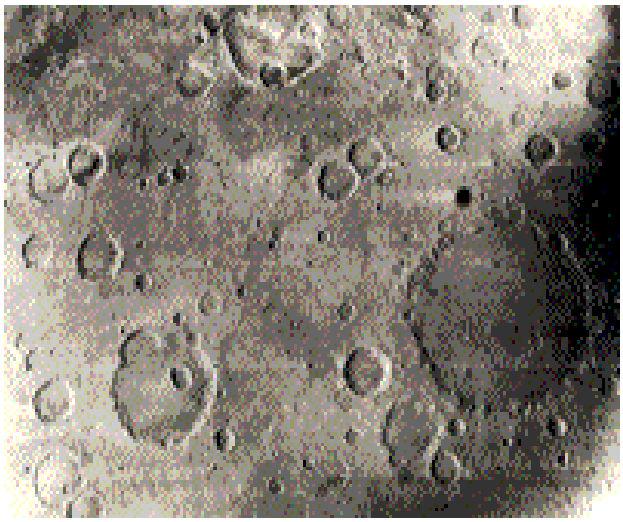


Another example of what is interpreted as wrinkle ridges is this:

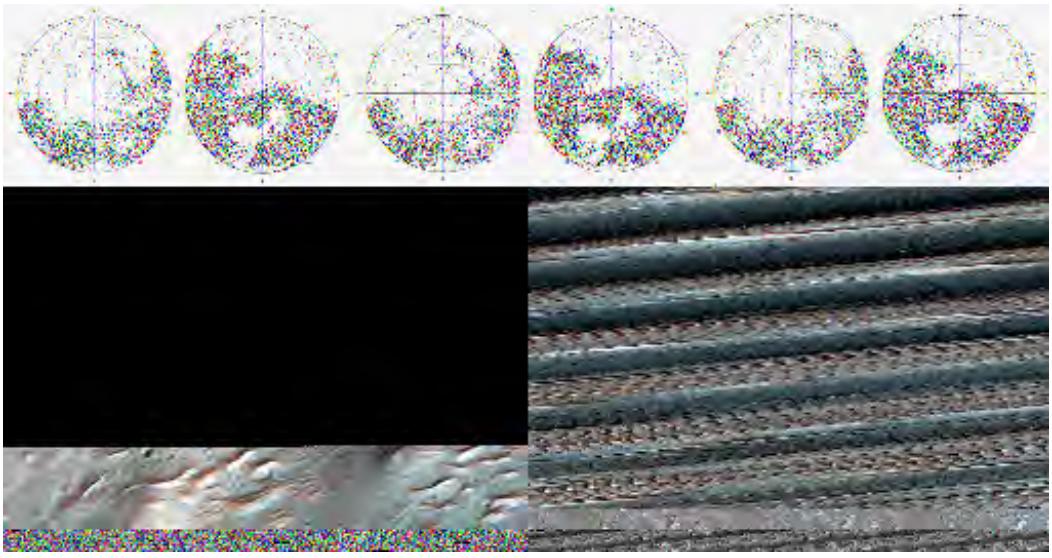


In contrast to the volcanoes described above, which are upward conical prominences, are the downward indentations or craters that can be either volcanic or impact. Both are typical of martian terrains. Extensive impact cratering was observed by Mariner 4, which sent back the first ever images taken of another planet's surface (one of these images is seen below (top) when this probe approached to within 9800 km (6086 miles). As imaged the next year by Mariner 6, the Sinus Sabeus region of the southern highlands (bottom scene) preserves typical impact craters in the ancient terrain that apparently has not been extensively resurfaced by lavas. Note that none of the larger craters in this view have central peaks.

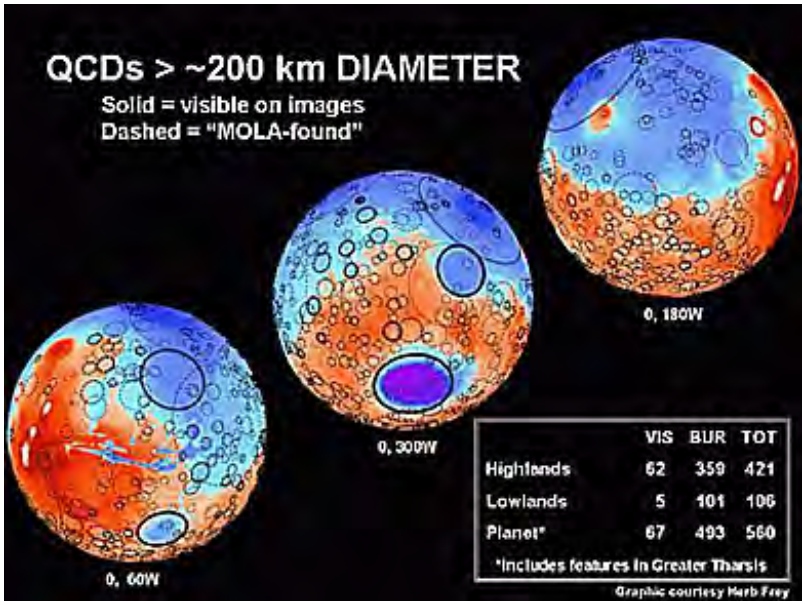




Mariner 9 and the Vikings confirmed that a large fraction of the (older) martian surface, mainly in the southern hemisphere, remains heavily cratered. This is evident in this sketch drawing from Mutch et al., *The Geology of Mars*, 1976 in which all craters larger than 15 km are positioned.

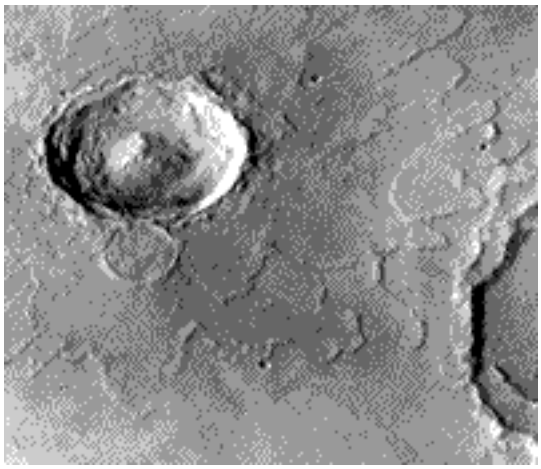


A recent study made by Dr. Herbert Frey of NASA Goddard - assisted by his teen age daughter Erin - has led to a map of the distribution of large surface-visible plus now buried impact structures that nevertheless show circular surface manifestations. The latter have been located using the MOLA laser altimetry data.

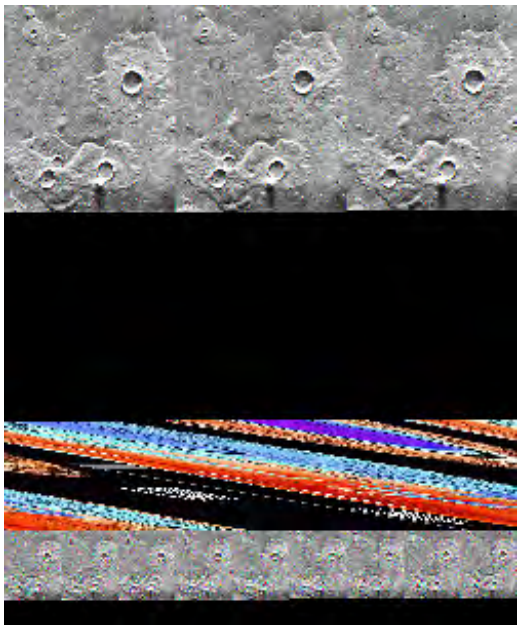


One can argue that this landscape has many similarities to the still cratered Earth in its early stages before extensive water had collected into major oceans. Likewise, buried impact structures can be discerned on the lunar surface. These have since been covered by lunar ejecta. This may mean that the martian Highlands surface is also covered by ejecta deposits that spread over older craters.

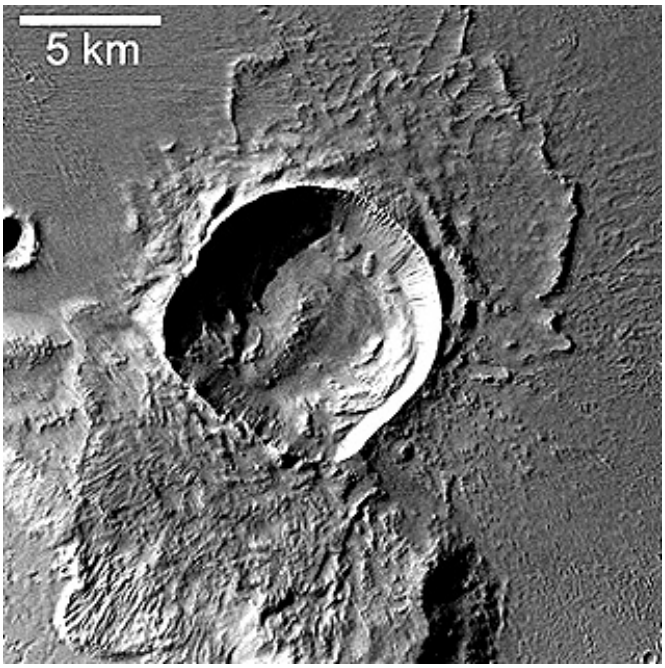
Some of the martian impact structures retain well-preserved ejecta blankets that display prominent lobes, such as seen here around the crater Yuty. The ejecta was probably fluidized by vaporization of carbon dioxide-rich ice lying just beneath the surface.



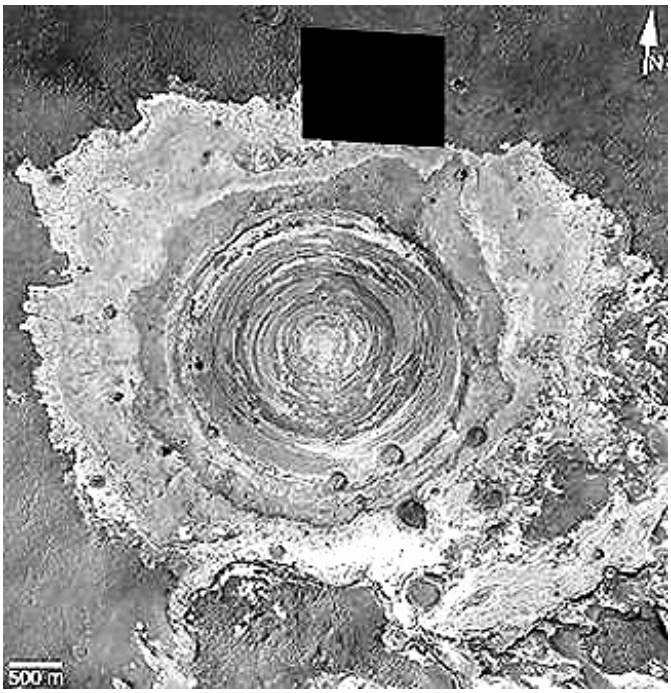
One type of impact crater is different from those on the Moon, Mercury and Venus in that the edge of the ejecta blanket has a steep scarp, evident in the Viking image below, or even a peripheral rise called a rampart. This type is called a pedestal crater.



On Mars many of the younger craters still preserve their ejecta blankets, as exemplified here:



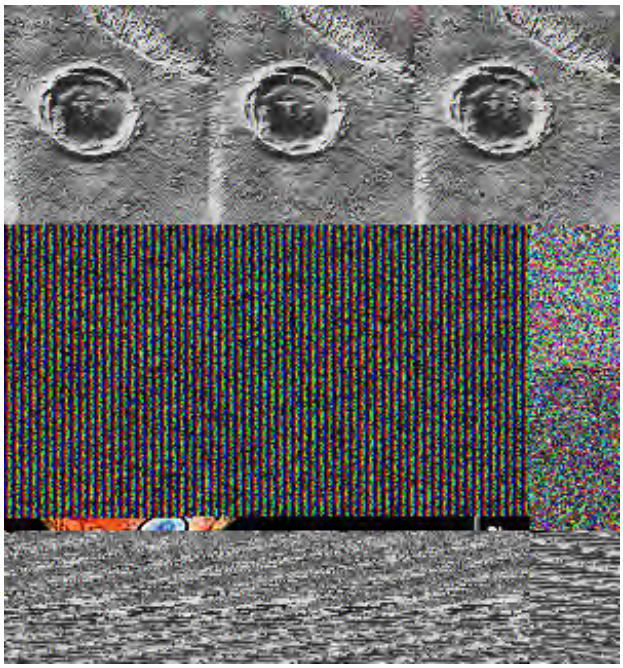
This next crater is small, young, and shows most of the same features as do terrestrial craters. Located in Terra Meridiani, this crater is 2.6 km wide (1.6 miles; rim to rim), has at least 1 nested slump zone in its interior and a distinct exterior ejecta blanket, and has exposed what appears to be internal layering of the martian surface units. The image was made by the Mars Global Surveyor.



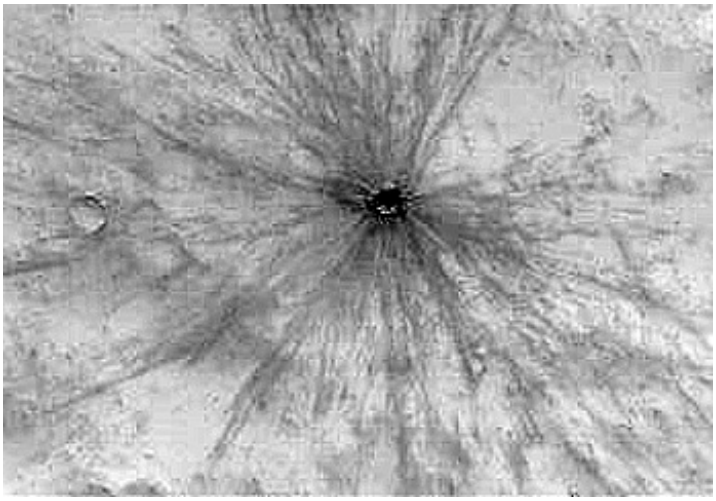
This type of central (interior) layering, almost certainly sedimentary (see pages 19-13a and 19-13b) also appears in the 2.3 km (1.5 mile) wide Schiaparelli crater in the Chrysae Basin, seen below. The layering appears horizontal:

These observations of sedimentary-like crater interior floors and walls (layering is also discussed on the next page) seem rather mysterious. On Earth, craters that still retain their original rims (almost?) never show the bedrock below the final crater excavation wall. Yet this is common in martian craters with initial walls intact. Martian planetologists have suggested removal by erosion (they mean almost certainly wind erosion). There may be an alternate cause: the lower martian gravity allow nearly complete escape during crater formation of the bulk of the ejecta; the floor remains exposed because in the smaller craters slumping has not destroyed the walls.

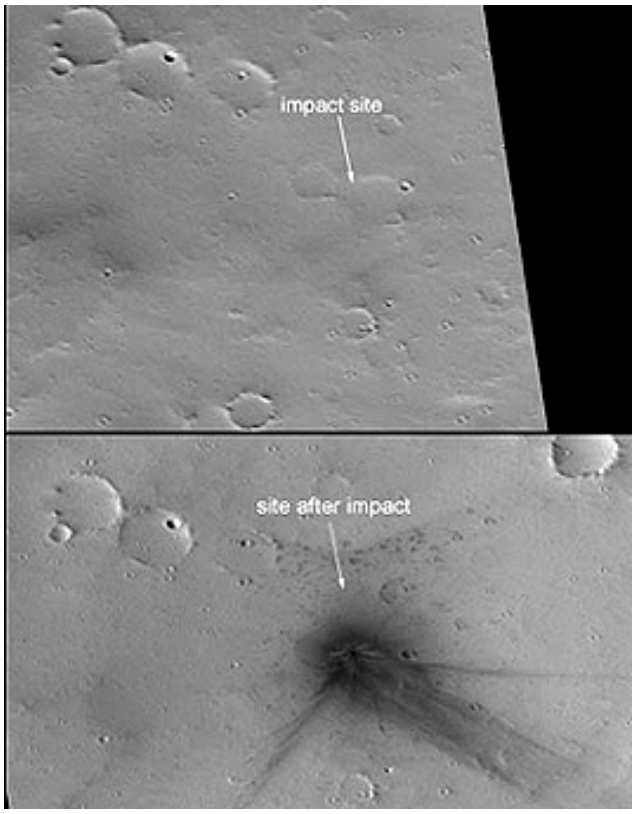
Still another large impact crater, Poona, has a remarkable uniform set of rays, equispaced over the full 360° around the rim:



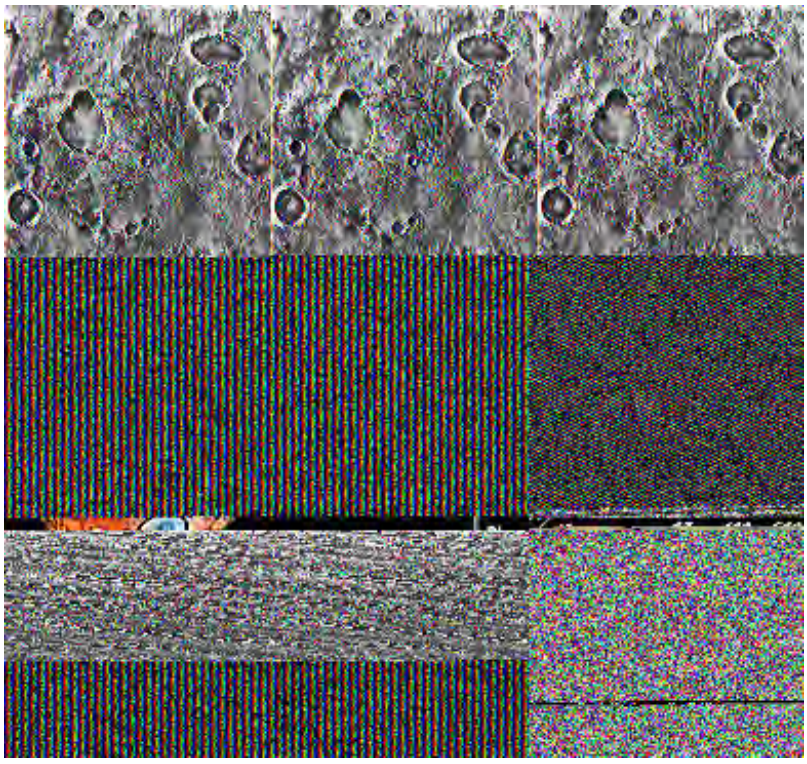
This small crater (below) shows a distinct pattern of dark rays. Because martian winds are continually altering the surface, both removing and covering up debris, the crater (and those above with lighter-toned rays) can be young - age estimates have ranged between a few thousand and a few million years.



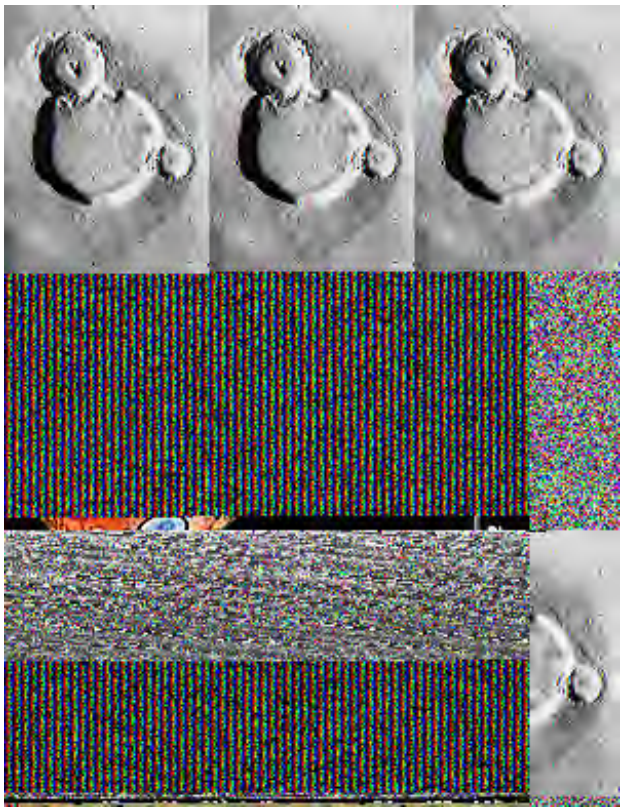
This rayed crater looks fresh. Experience on Earth indicates that impacts occur rather often in terms of a human time frame. A new crater was produced on Mars during the operational period of the Mars Global Surveyor. This before-and-after image pair shows the appearance of dark rays around an area which contains a small hole not there on the earlier date:



This next Viking scene, in the southern Highlands, seems to have both impact and volcanic craters. Some without ejecta beyond their rims, especially the elliptical one, are calderas. Several others have aspects more characteristic of degraded impact structures. This was an active region, with channels (either volcanic or stream) and other types of terrain.

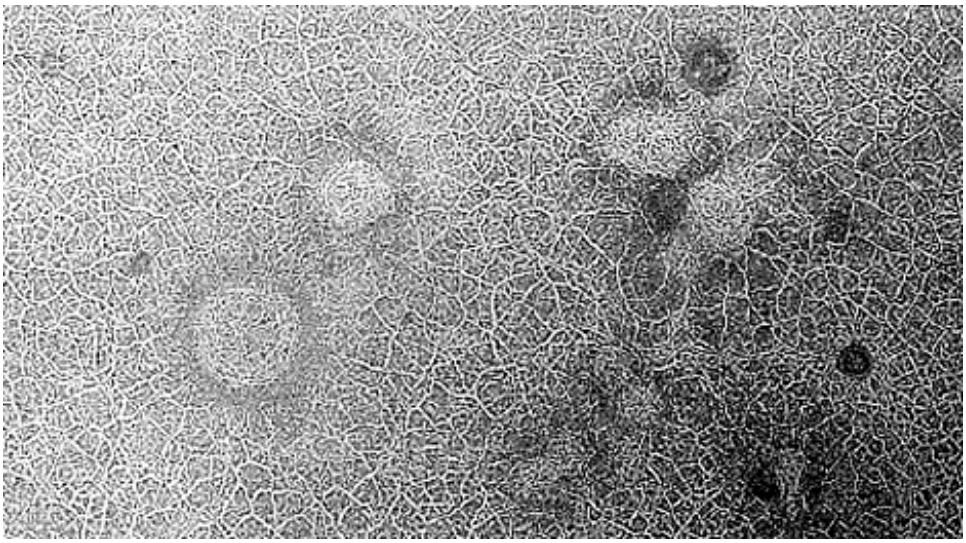


Now look at these three craters (Ulysses Patera):

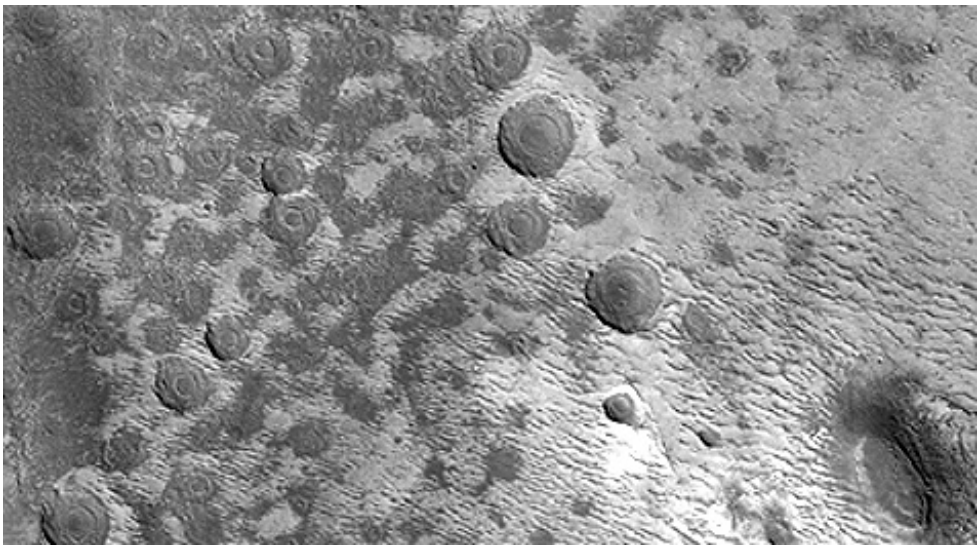


19-41: What type(s) of craters are present in this Viking scene (the largest structure is about 80 km [50 miles] across at its base)? ANSWER

Because of several factors, some martian craters appear as faint rings rather than topographic features raised above the surface. These have been called "ghost" or "stealth" craters. They represent some combination of burial by crater ejecta, wind erosion, dust cover, and ice cover. Here is an example of this last type:



There is evidence that the number of *observed* impact craters on Mars is less than would be expected if the recent activities (dust transport and deposition, ice relocation, etc.) had not buried the smaller ones. The wind, however, is capable of exhuming such craters, as displayed in this image which also shows the exposed craters to contain some signs of filling by sediment, now revealed as faint layers.

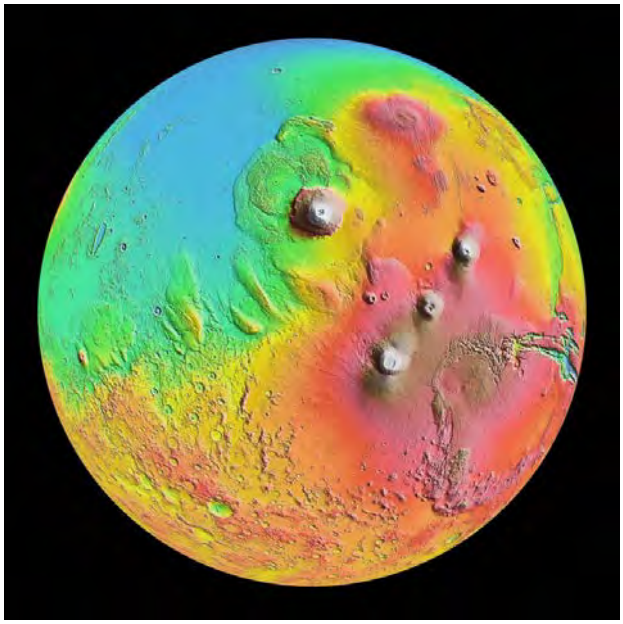


Not all impact craters are circular or slightly elliptical. Strongly elongate craters are found on the Moon. A few such distorted craters are present on Mars, such as the one shown below. The usual explanation is that the impacting body comes onto the surface at a very low or grazing angle, scouring out the surface material as it proceeds forward:

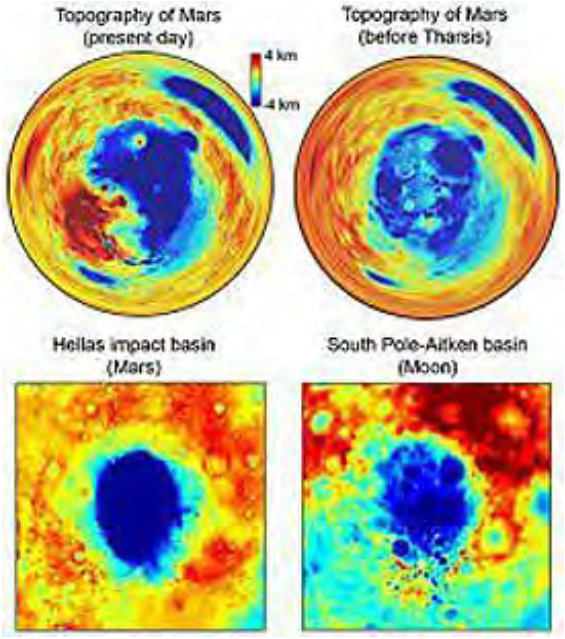
Large, young impact craters are few but conspicuous. Galle Crater is 220 km (138 miles) wide and retains its original rim:

As with the Moon, Mars has a few craters so large that they can be called impact basins. By far the biggest is the Borealis Basin (also known as Vastitus Borealis). Mars geologists (starting with George McGill and Stephen Squyres) have postulated that this feature (which has dimensions of

10,600 by 8500 km) was produced by a glancing collision with an asteroidal body that may have been as much as 1600 km in diameter. This impact, which peeled off at least 3 km of martian surface, may have occurred as early as 4 billion years ago. It accounts for the generally lower topography of the northern half of Mars (see page 19-10), evident in this topographic map (blue low; reds high):

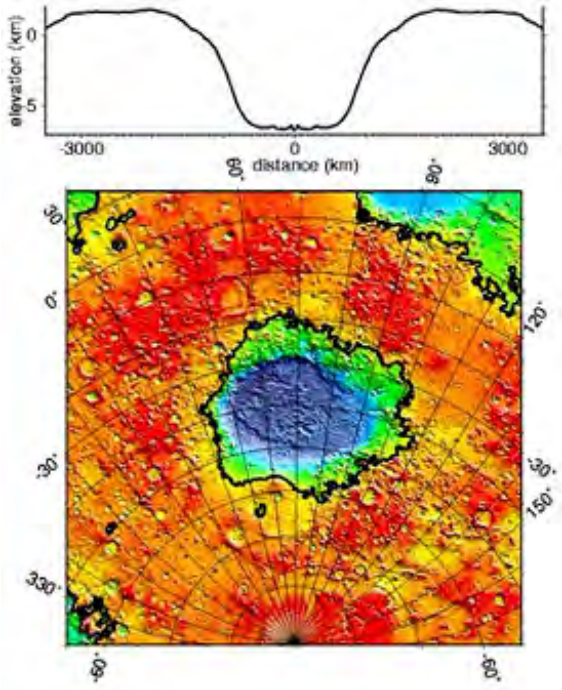


Papers from a group at MIT in a June 2008 issue of the journal *Science* offer strong evidence for the existence of the Borealis Basin; as is often the case, there is a vocal group of doubters. But, the impact hypothesis is a plausible explanation for the dichotomy in martian topography: the distinctly lower top 40% is readily explained as caused by the stripping of crust from an oblique impact. In the next figure, the Borealis Basin (bluish area in upper right panel) is compared with the Hellas Basin on Mars and the Aitken Basin on the Moon:

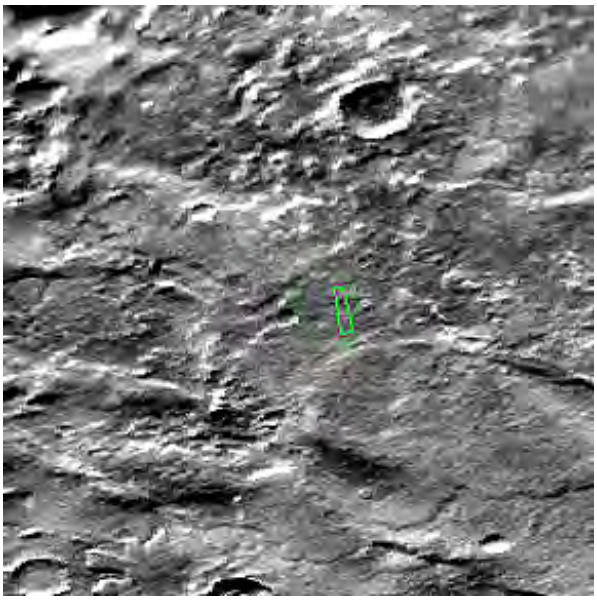


The largest *well-defined* impact basin on Mars, and second in the Solar System only to the Aitken basin on the Moon, is the Hellas Basin in the southern Highlands. Its diameter is about 2100 km (1300 miles), its depth is almost 9 km (6 miles) and its rim exceeds 1.5 km (1 mile). In this view the Basin appears to have no significant landforms within it.

To emphasize the size of this structure: If all material excavated from it were to be spread evenly over the 48 continental United States, a layer of debris some 3.5 km (2 miles) thick would accrue. Below is an enlargement of the map covering this structure.



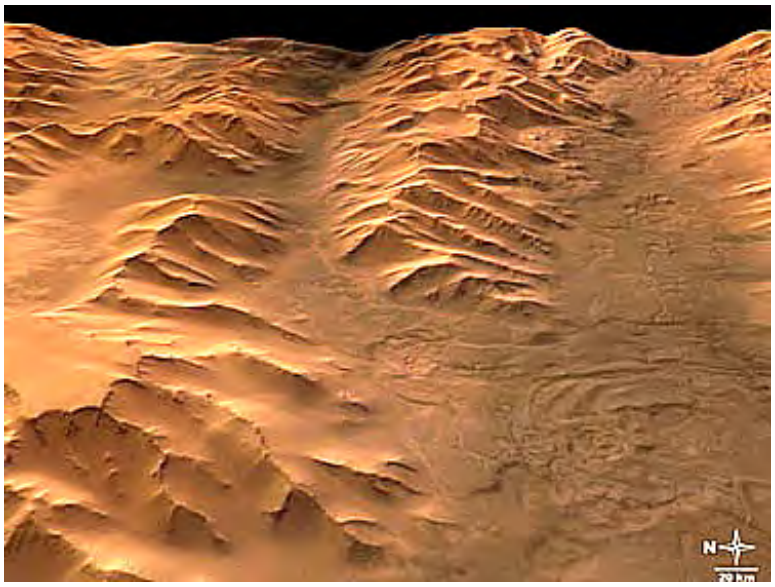
The floor of Hellas actually shows diverse landforms (mostly of low relief), some of which appear volcanic in origin; if so this would imply that the basin filled with melt soon after the impact event, which may have been relatively recent.



Another impact structure is the Argyre Basin (600 km; 390 miles diameter), seen in this Viking view:

Mars investigators have speculated that during the early eons of martian time, when the atmosphere was possibly more abundant (thicker, with greater surface pressure), water released by impacts and other processes could be distributed as rainfall. Some think that shallow lakes filled Hellas, Argyre and other large craters for a time.

In some of the above images, and several on pages 19-13a and 19-13b, features that could be described as mountains are displayed. Of course, volcanoes fall broadly into that category. Rims around large craters also are mountainlike. Here is a series of mostly parallel ridgelike prominences that are considered low mountains, found here in a region called Tithonium Chasma:

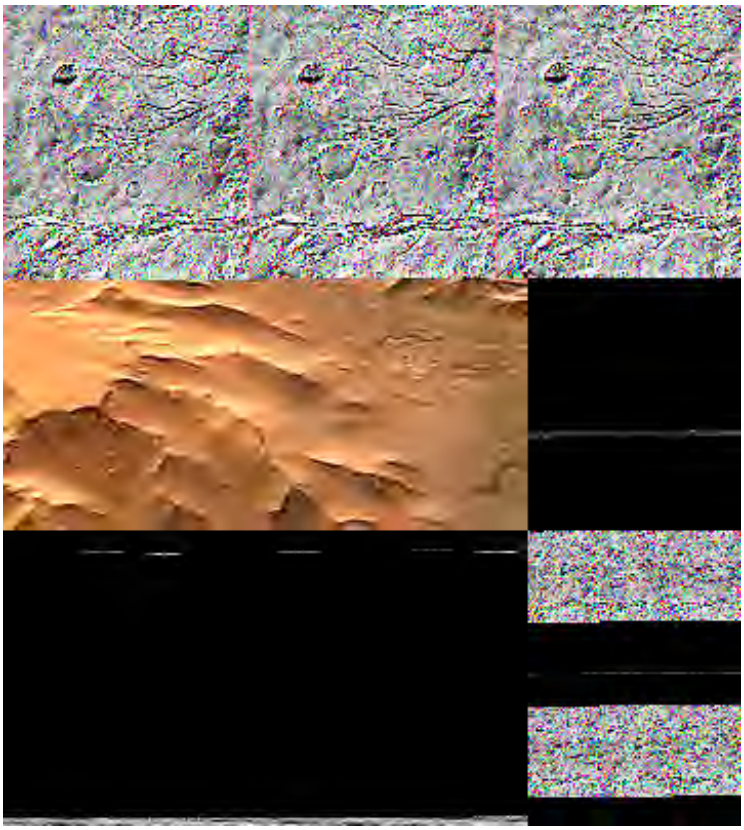


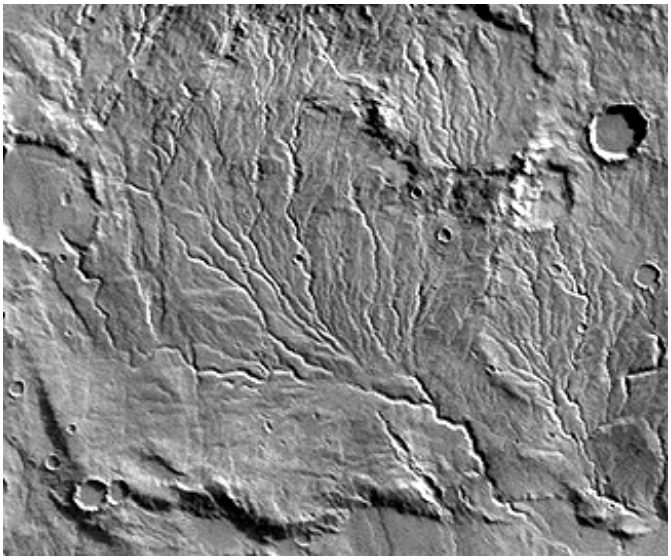
So, once again we see a planetary body with a great variety of landforms, many caused or affected by impact processes. Some of these are unusual (exotic) including those which may reveal water erosion.

Channels and Exotic Terrains

By now, one should be convinced that Mars is a geomorphologist's Paradise. As with the Moon in the earlier days of exploration, landform identification, with educated "guesses" as to modes of formation, has been the prime approach to mapping and interpreting the martian surface. Mars exhibits a great variety of terrains and landforms types. Most are given terms that have a Latin derivation. An excellent summary with numerous examples of these types is found at The Atlas of Mars web site. Click on the terms in the left column which brings up usually many images each displayed by clicking on its entry phrase. It is well worth your time to spend an hour or so looking at the wide range of landforms recorded at this site.

Some of the big surprises were infrequent but distinctive *sinuous channels,* whose morphology is much more similar to river channels that lava channels. One interpretation holds this morphology as evidence of widespread water in the past, in lakes, groundwater or possibly oceans. Expulsion of copious water initiated some sort of hydrologic cycle involving rain storms and runoff. Most of this water has since evaporated into space, although possibly significant quantities may remain frozen as underground ice. Nevertheless, major water activity has recurred as evidenced by the types of dendritic channeling shown in these Viking Orbiter images:





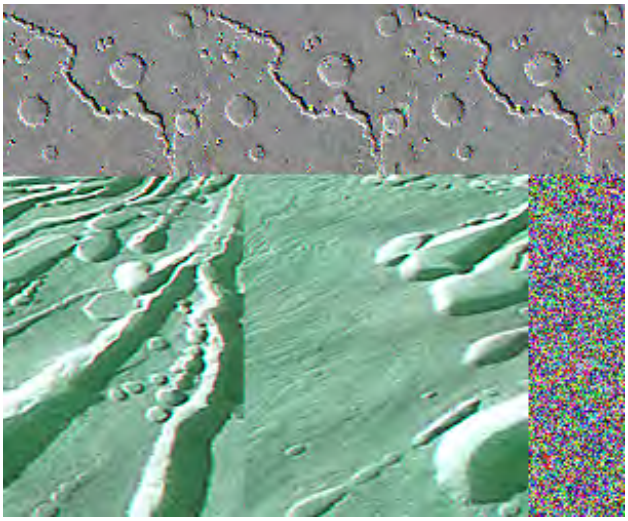
The region depicted in the top image covers the Juvenae Chasma and Vedra Vallis. These are *runoff channels*, a type confined to the ancient landscapes. Stream flow is the favored origin, based on comparing them with terrestrial counterparts. Some channels seem to originate at craters, which could imply that subterranean sources released either water or lava, following impact offloading. The drainage pattern in the bottom image resembles terrestrial patterns found in soft sediments or wind deposits.

19-42: What is the argument that the type of channels shown above is not volcanic sinuous rilles or collapsed lava tubes? ANSWER

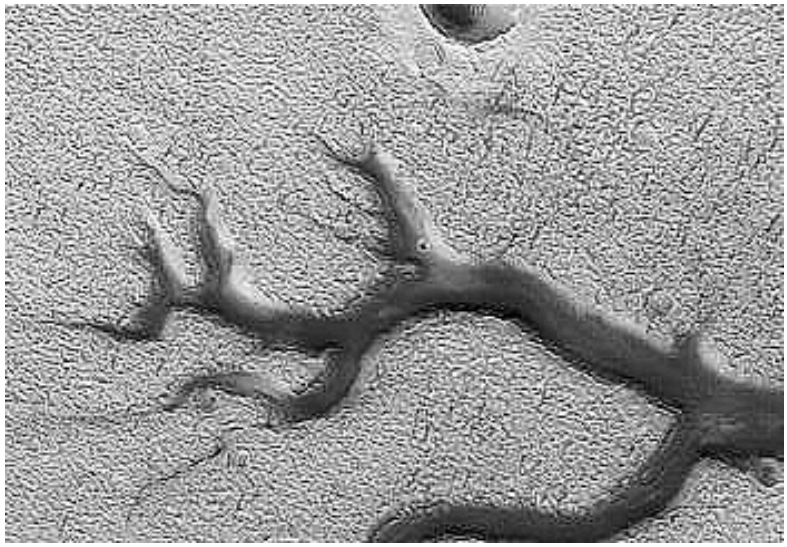
Collapsed lava tubes, which look like some stream channels, have been found in association with martian volcanoes. Here is one example:

Even more striking is this group of lava tubes on basaltic terrain in Pavonis Mons; seen from above this would look like multiple stream channels:

This Viking image shows a channel called Nirgal Valles that looks much like the sinuous rilles described on the lunar surface. Whether this was caused by lava tube collapse or by fluvial action is not obvious at this scale:

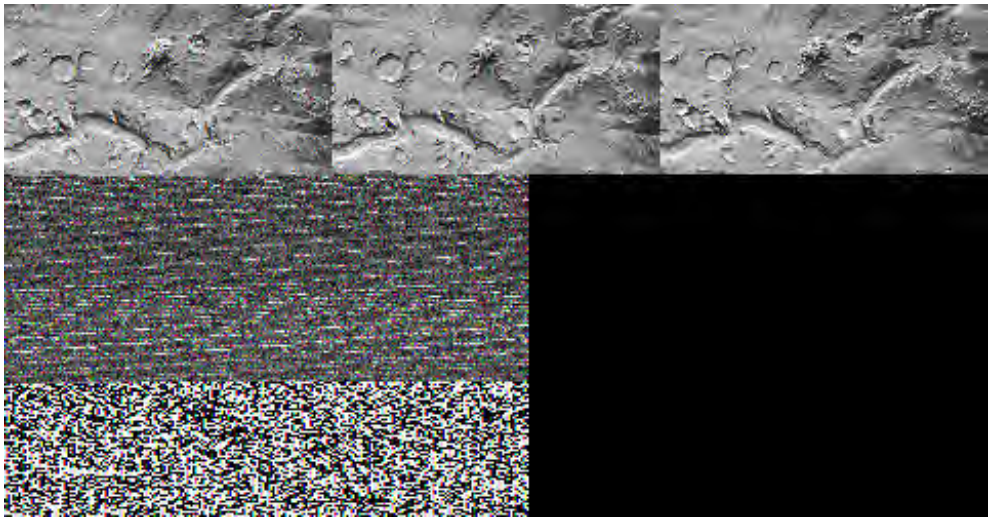


Nevertheless, the resemblance of many of the martian channels to fluvial channels on Earth is particularly evident in the next (MOC) image. Located within the large Newton crater, the dark (windblown sand-filled?), flat-bottomed channels look like some headwater types for streams found on Earth:



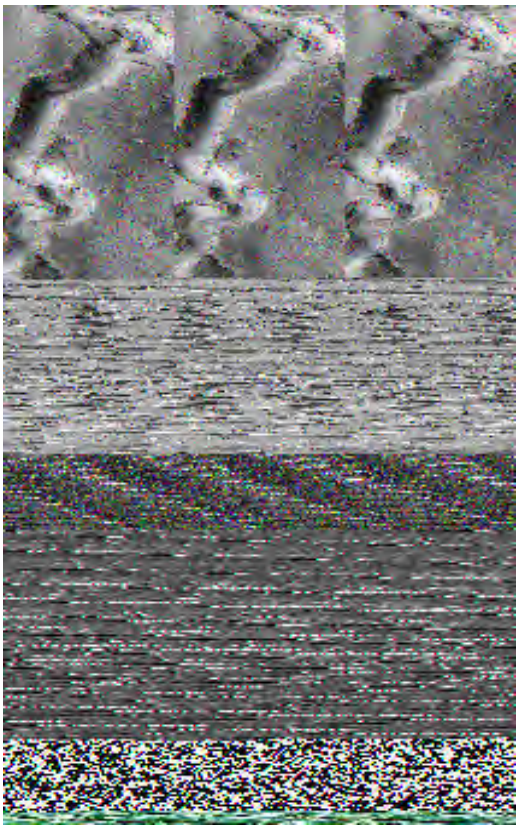
The "Jury is Out" on the exact origin of the narrow channels in this MGS MOC image. What can be said is that over much of the depression light-colored wind deposits have been trapped and shaped into dunes resembling large ripples:

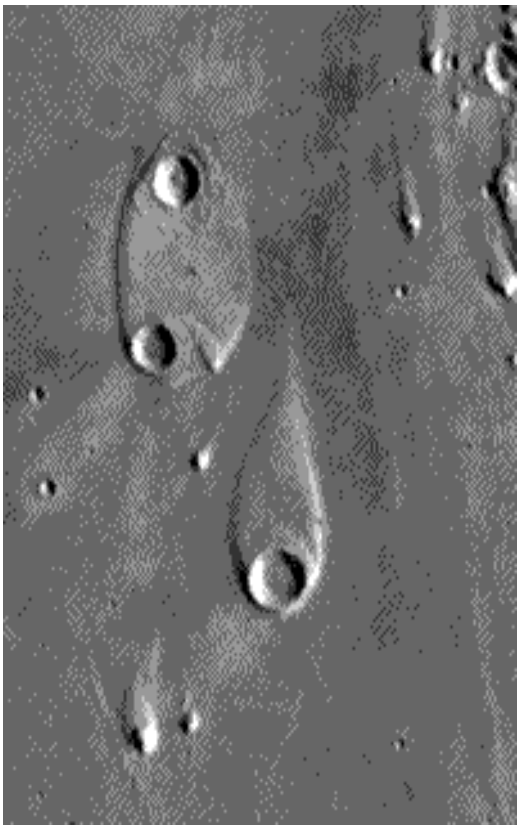
A distinctive type of drainage called *outflow channelling* is typically broad and deep, creating canyon-like depressions. A typical example, seen below, is Ma'adam Valles, some 300 km (185 miles) long, which ends in Gusev Crater (far upper right; see page 19-13a):



This next type of landform (left image) may have been associated with catastrophic scouring during abrupt flooding. A similar example on Earth is the Channeled Scablands of central

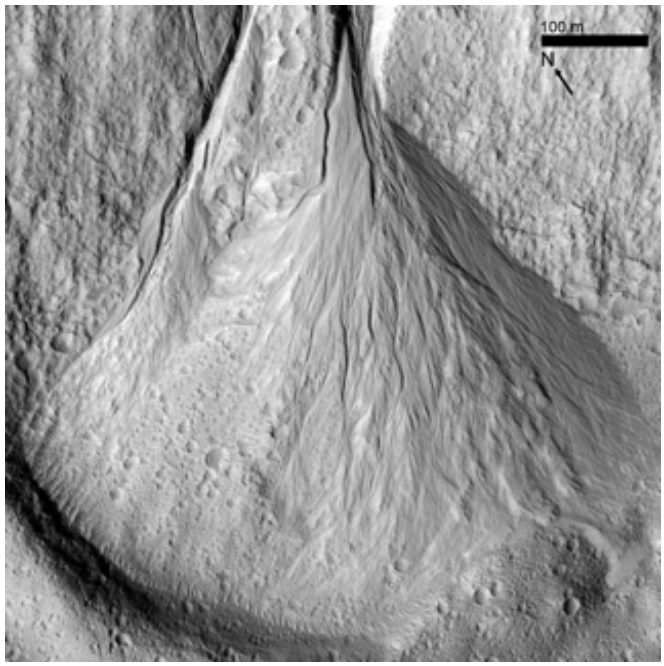
Washington State in the U.S. that developed in just a few weeks from rapid emptying of a huge dammed lake after a natural breakup. Another indication of strong fluid action is a teardrop-shaped landform in Elysium Planitia (right image) a prime example of shaping by streamlining (analogous to aerodynamic sculpturing), in which water flowing from bottom to top has eroded plains material around the rim of a large crater and has terraced and perhaps redeposited debris towards the pointed end.



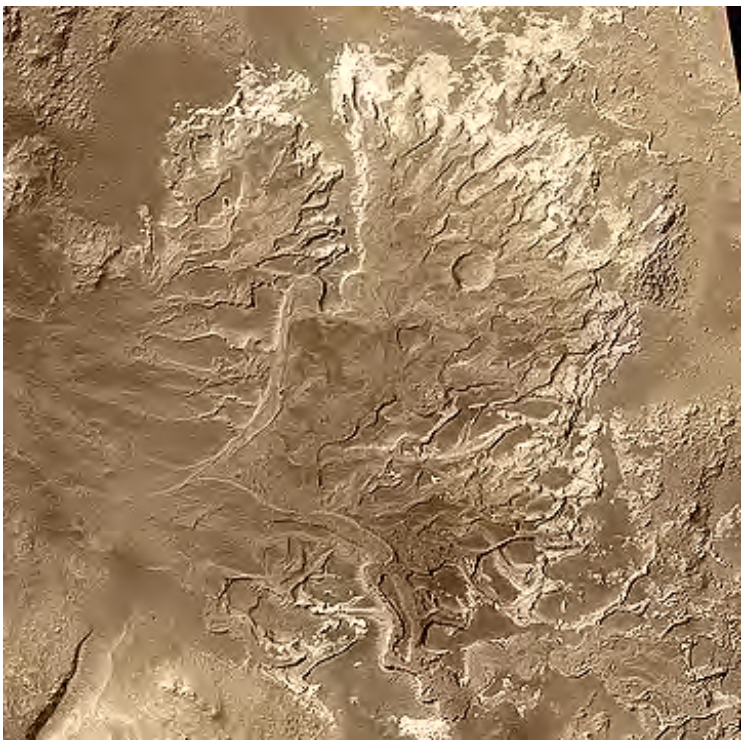


19-43: Present an argument as to why the teardrop landform was caused by water rather than wind. ANSWER

If riverlike channels did once carry water over the martian surface, one landform they should produce is a fan deposit made up of the debris carried by the streams until such streams are slowed such as to cause their sediment loads to be dropped. This fan is located in Prometheus Terra:

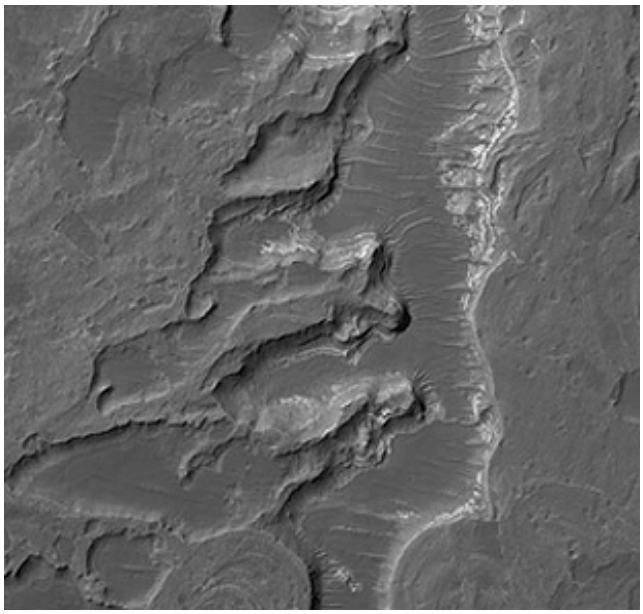


A prominent distributary fan has now been found in the Eberswalde crater in the southern hemisphere. The delta-like fan is 13 km (8 miles) by 11 km (7 miles) in dimension. Here it is in a Mars Global Surveyor MOC image:



MOC close-ups of parts of this fan show several anomalies attributable to fan morphology. In the first image below, one and perhaps two flat-topped ridges emerge above sculpted out surfaces (exposing layers). These ridges may be made of channel deposits that were more resistant than surrounding deposits so that after general erosion of the fan, these remain as topographic highs:

This image indicates the fan's ridges are made up of rock that is consolidated and hardened:



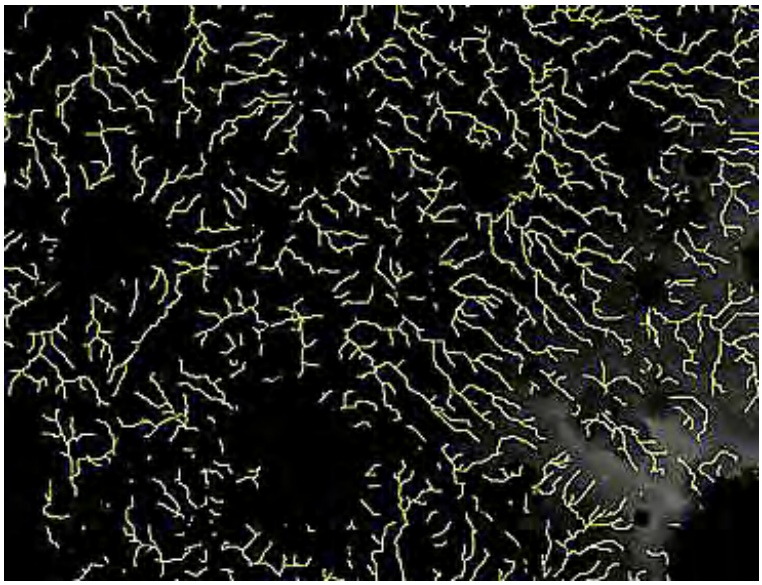
The next two paragraphs are part of a Science release made by JPL/USGS:

The Eberswalde delta provides the first clear, "smoking gun" evidence that some valleys on Mars experienced persistent flow of a liquid with the physical properties of water over an extended period of time, as do rivers on Earth. In addition, because the delta today is lithified -- that is, hardened to form rock -- it provided the first unambiguous evidence that some martian sedimentary rocks were deposited in a liquid (presumably, water) environment. The presence of meandering channels, a cut-off meander, and crisscrossing channels at different elevations (one above the other), provided the clear geologic evidence for these interpretations.

After the sediments were deposited to form the delta, the material was further buried by other materials -- probably sediments -- no longer present. The entire package of buried material became cemented and hardened to form rock. Later, erosive processes such as wind stripped away the overlying rock, re-exposing the delta. Now preserved essentially as a fossil, the former floors of channels in the delta became inverted, to form ridges, by erosion. Channels can be inverted by erosion on both Earth and Mars. Usually this happens when the channel floor, or the material filling the channel, is harder to erode than the surrounding material into which the

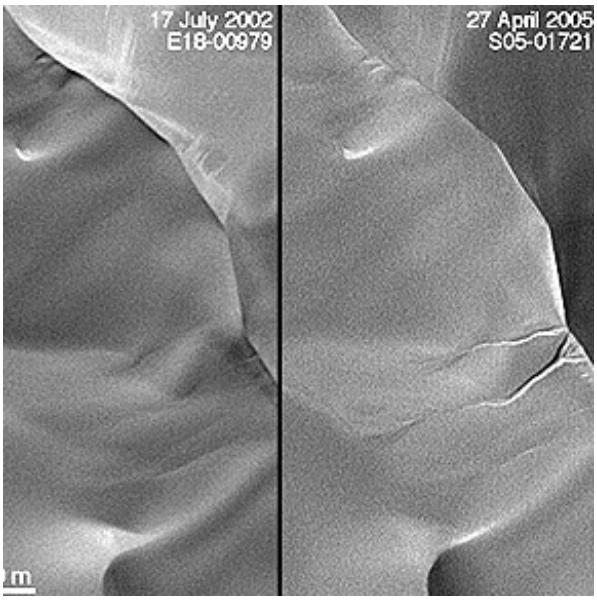
channel was cut. In some cases, the channels on Earth and Mars have been filled by lava to make them more resistant to erosion. In the case of Eberswalde, there are no lava flows; instead, the channel floors may have been rendered resistant to erosion either by being better-cemented than the surrounding material, or composed of coarser-grained sediment (such as sand and gravel as opposed to silt), or both.

The consensus as of 2007 is that many - perhaps most - martian channels were carved by running water at times in martian history when liquid water was much freer to flow in copious amounts over the surface; probably the martian atmosphere was denser in those times. A plot of major channels in the non-polar regions of Mars reveals an interesting pattern:

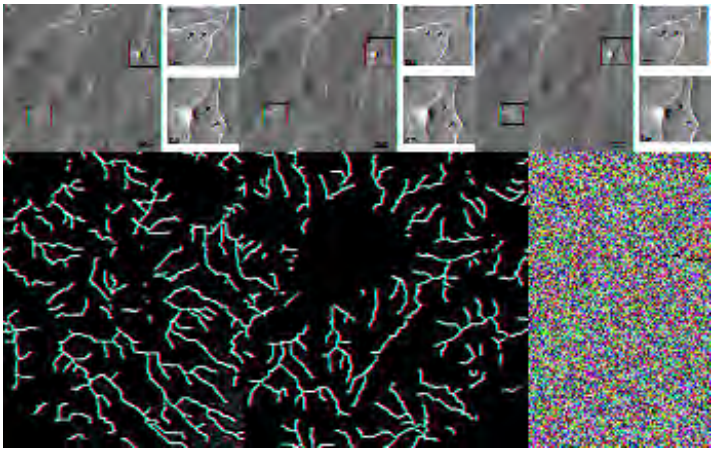


What seems mysterious about this pattern is that most of the channel systems end in a "blank" (black) part of the map. Thus the majority of systems were independent and did not connect with each other. The presumption is that each channel emptied into a relatively small region of Mars. In such a region the water would connect with a standing body of lake-sized proportions. While this in itself does not rule out oceans, those would have existed before channel-cutting. The question that emerges when trying to explain the widespread distribution of sedimentlike layers over much of Mars is whether these were dominantly lake deposits or could at least some represent a more continuous marine stage.

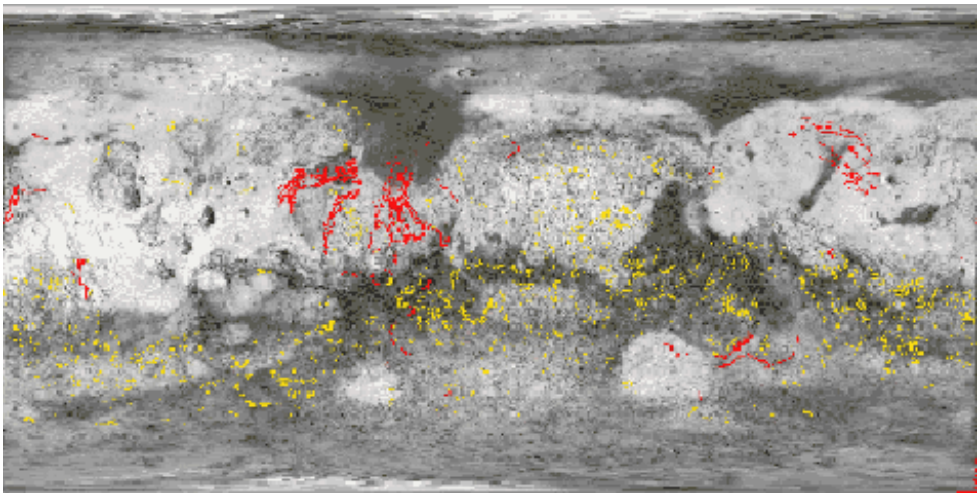
One thing is now sure. Some process(es) is/are still producing new channels. This image below shows the same smooth (sandy?) plains that ends abruptly in a cliff. On the left, the image indicates no channeling; on the right at a later time, a well-defined small channel network has since formed. What caused this remains conjectural; wind erosion is one suggestion.



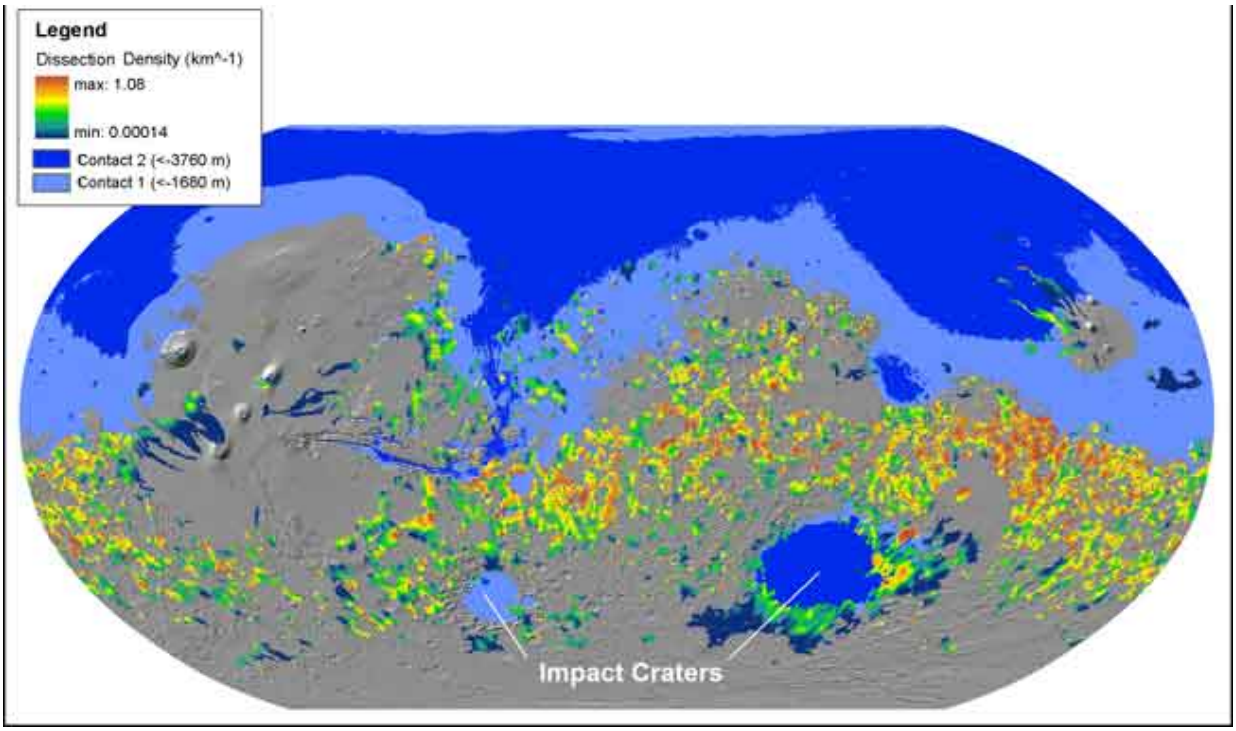
Found both on Earth and on Mars, the "inverted channel" results when a normal channel is filled with sediments that become well-cemented, and later erosion removes surrounding, softer deposits, leaving the channel to stand above the general surface (indicated by shadowing). This is a good example:



Mapping, using largely MGS imagery, has displayed the distribution of channels on Mars. The map below is taken from this website that summarizes the current thinking on martian channels. The type called outflow channels is shown in red and valley networks in yellow.

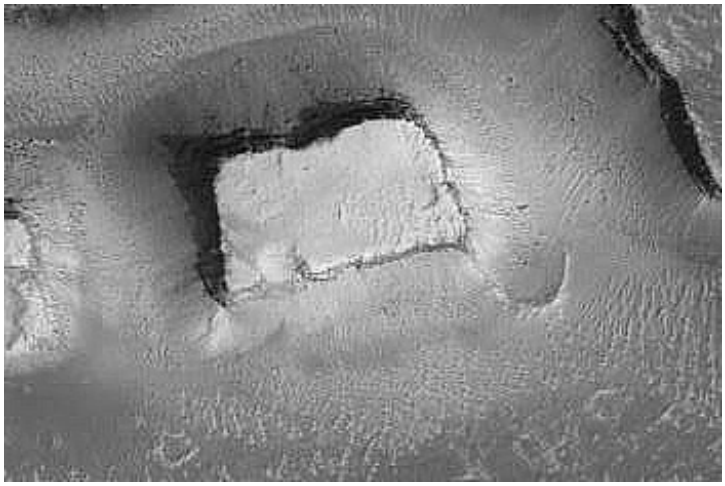


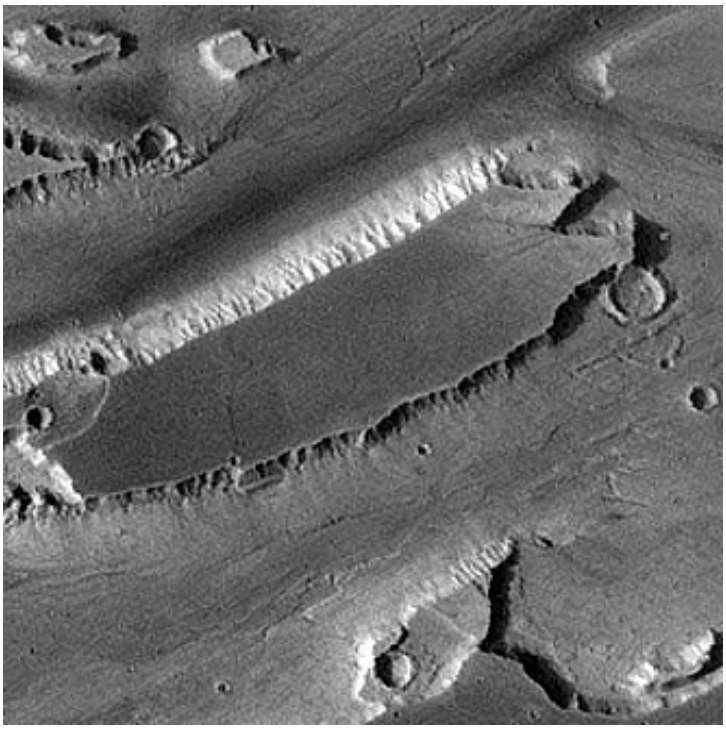
A map - shown below - quantifying the density of channels on Mars was made public in November of 2009:



The greatest number of channels occur in the higher terrain in the central to southern part of Mars, that is, in the Highlands. Note that this band of channels is not far from the vast blue areas that denote the northern Lowlands. Many Mars scientists believe this blue area once held an ancient ocean. Water from this ocean that evaporates would produce rainfall that tended to land on the more northern Highlands and would be much less towards the South Pole. This rain would collect as runoff in streams that produce the channels.

Another landform that, on Earth, is almost exclusively formed by water-involved erosion is the mesa (or its smaller form, the butte); on Mars it is termed "mensa". This terrestrial landform occurs usually when a more resistant layer series is on top of a weaker, more erodable set. Water, usually within migrating streams, attacks the lower layers, gradually exposing the surface, and causing the upper layers to diminish in size and extent as undercutting erodes into those layers. A residual series of higher landforms results as remnants of the original upper layers on the stripped surface. Here are two excellent examples of a Mars mesa, the first found in Granicus Valles, the second is called Lunae Mensa:



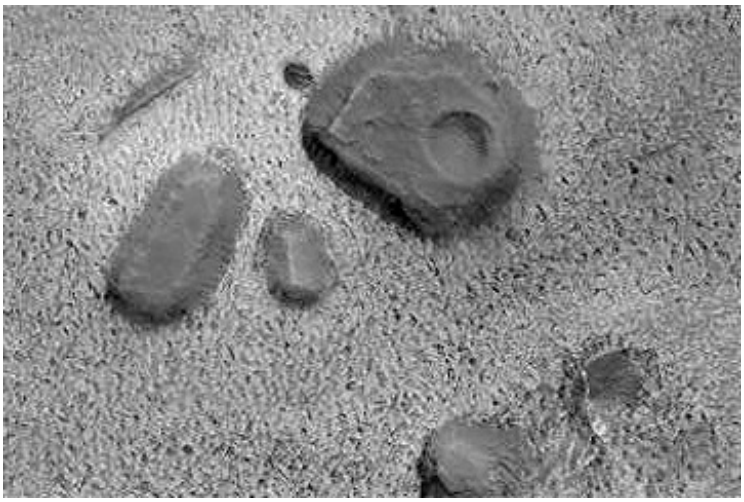


Light-topped mesas are found near Valles Marineris:

Seen closer-up, mesalike blocks of this whitish material shows faint but distinguishable layering. Such units appear to be erosional outliers or remnants of a once continuous sequence of deposits - probably formed by some sedimentary process - found over various regions of Mars.

Well-developed mesas are found at Cerebrus Palin, as shown in this MRO image:

Other variants of the mesa landform include remnants of a thick dark unit found above the lake beds in Aram Chaos (see page 19-13a) and flat-topped CO_2 ice "mesas" separated by flat pits in the south Polar ice sheet.

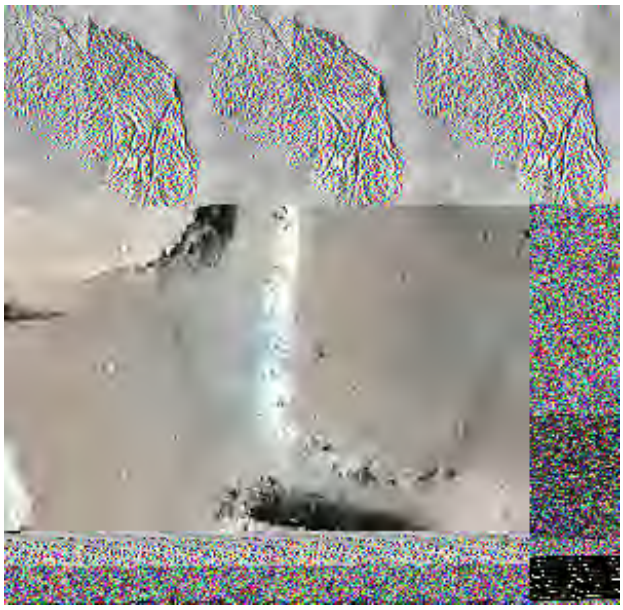


A recent paper has presented an alternative to water as the prime liquid medium responsible for the above channels and streamlike patterns. In this view, CO_2 (carbon dioxide) is proposed to exist in liquid form and in flowing upon expulsion at the surface brings about the erosional features described as fluvial. A variant of this suggests that liquid water at the times in the past when Mars was warmer may have contained a significant amount of dissolved carbon dioxide ("soda water") that increased its ability to erode.

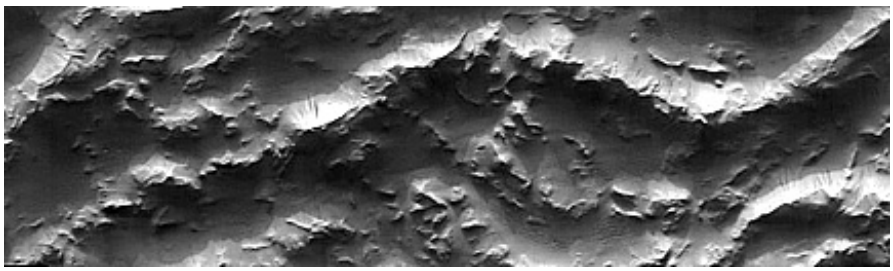
So, as of mid-2006 what can be said that reasonably affirms the presence of water now and in the past in martian history? On the next two pages (19-13a and 19-13b) we will learn of the direct observation of sediments that normally require the involvement of water. The discovery of water in polar ice, probably in subsurface lower latitude materials, and in the thin atmosphere all point to survival of small amounts of water today. This water may just be that released by occasional volcanism or by shock-evaporation of incoming comets. The various riverine landforms shown above on this page seem to point to greater amounts of water in the past. One school of thought concludes that past atmospheres were more dense and had a greater water content. Ancient martian atmospheres have probably been progressively depleted of gaseous and liquid molecules, including water, by thermal activity, by gravitational loss, and especially by the relatively weak but potent solar winds.

The martian surface is amazingly varied, with landforms of diverse genesis, some probably related to water action, as you have just seen, and others to tectonic forces, volcanism, and wind, being given descriptive names. Here are some typical examples:

The mishmash of intersecting linear features, called *grooved terrain*, in this case may be a complex surface of eroded ash deposits or possibly joint enlargement of a now buried remnant of a volcanic lava unit.

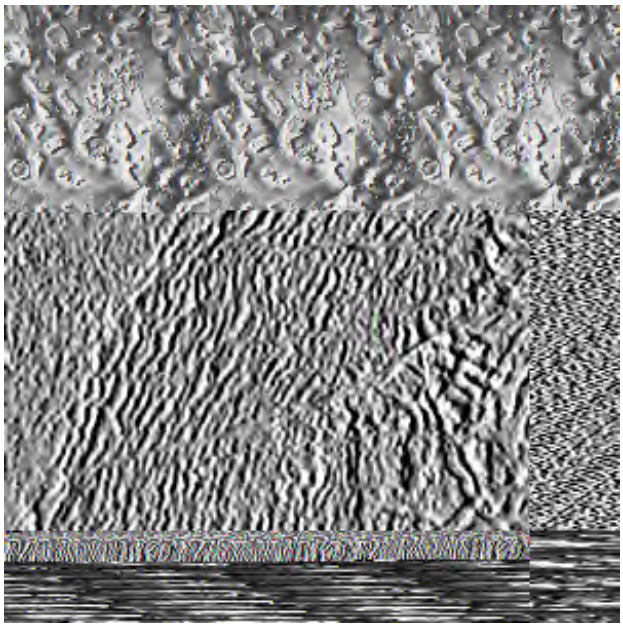


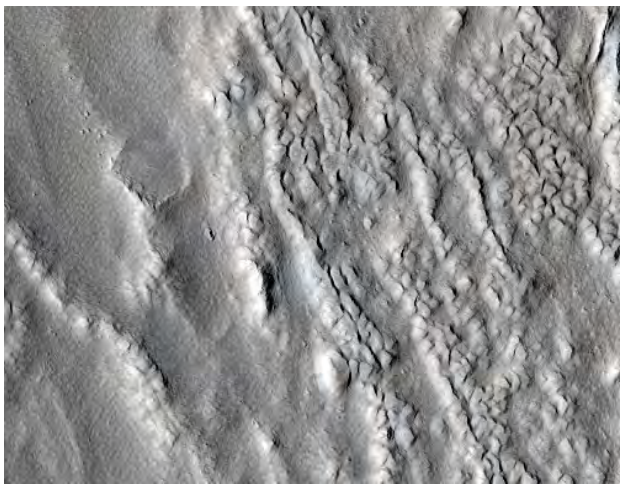
Variants of grooved terrain are known as sulci (singular, sulcus). Here is a closeup example seen in a thermal image made by THEMIS:



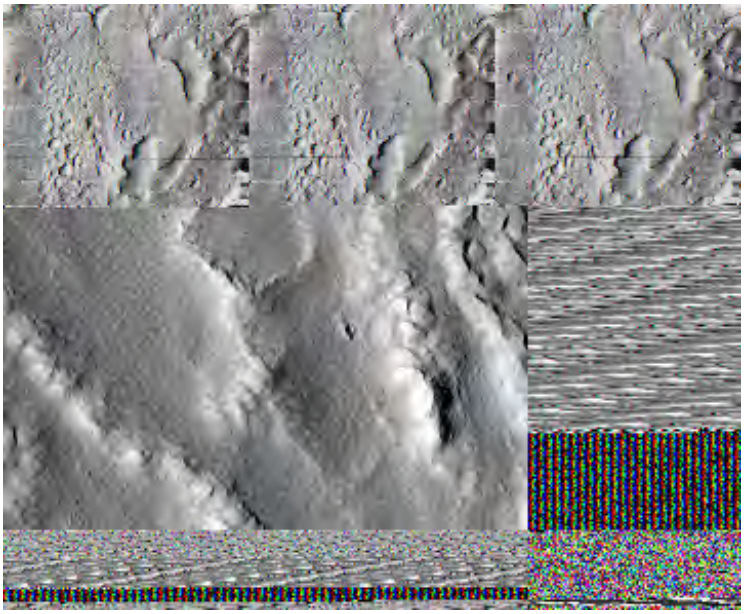
Similar terrain occurs in the slopes beyond Olympus Mons where the features present are called part of this volcano's *aureole*. The criss-crossing grooves and ridges seen here are almost certainly tectonic in nature:

The next pair of images show *fretted terrain*, found usually near cratered terrain, consisting of separated higher mesa-like units, bounded by scarps and set within lower smooth plains. This terrain is usually associated with features suggesting the action of ice (glaciated) and may also be influenced by incomplete dissection of older landforms by water and/or wind. Here are three examples:

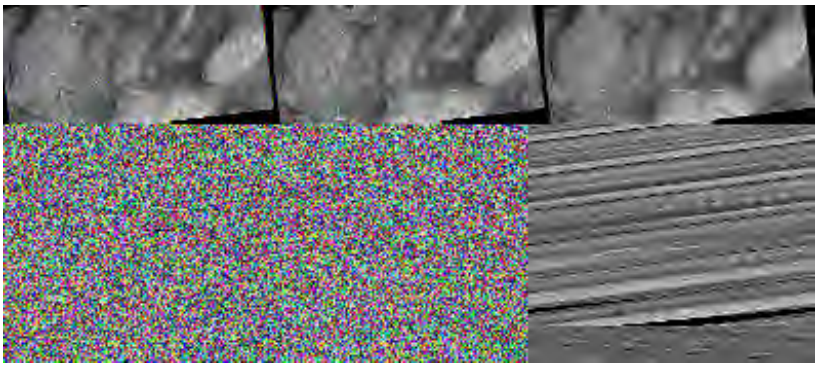




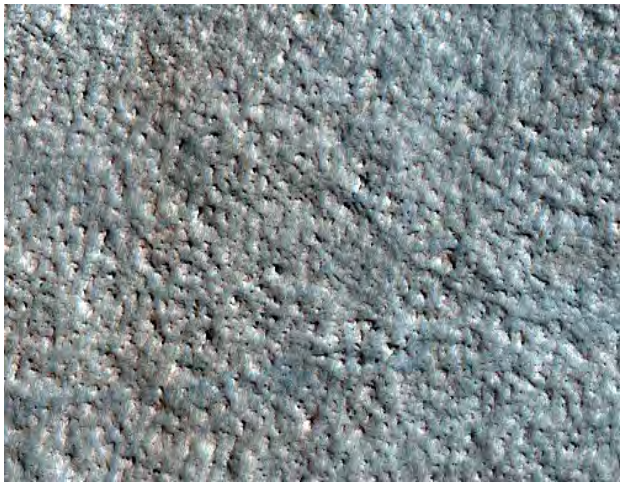
The next image portrays *etched terrain* which consists of shallow depressions likely developed by wind scouring and deflation of easily erodable unconsolidated surface materials.



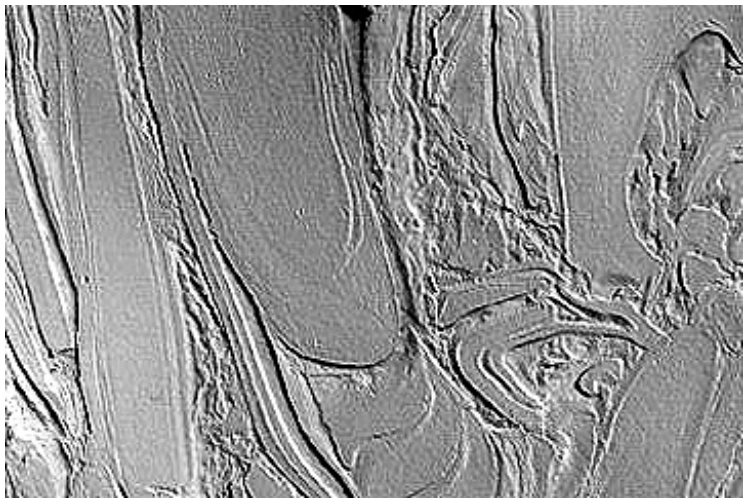
This MRO image shows another variant of etched terrain (also called sculpted terrain), in which the depressions may be caused either by wind deflation or by subsurface sag:



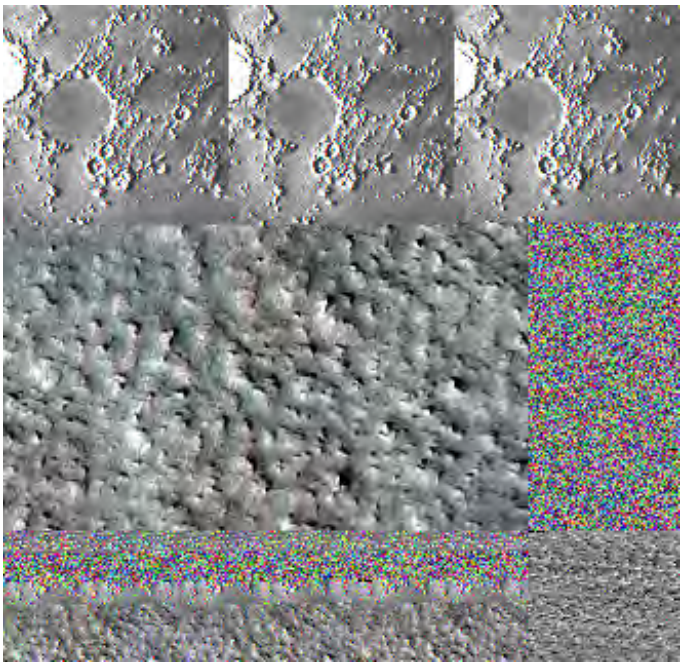
Other landform types given distinctive names (see the map near the top of page 19-11) include: *furrowed terrain*, *knobby terrain, channeled terrain*, and *layered terrain*. Examples of several of these are shown elsewhere in the Mars subsection. Here are two images that show typical knobby terrain:

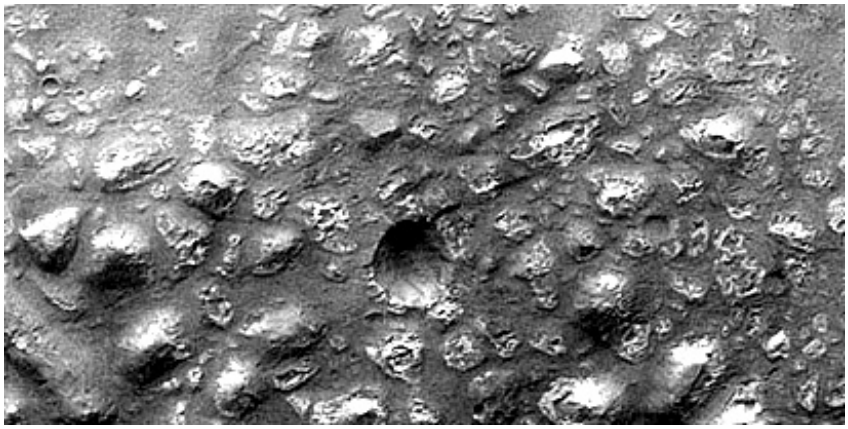


Below is an example of a peculiar terrain found mainly in Hellas Planitia. It is termed colloquially "taffy-pull" terrain. It's formative nature remains uncertain but one interpretation includes the possibility of erosion of hard and soft layers of sediment-like material; this does not quite explain the flow patterns in apparent channels.



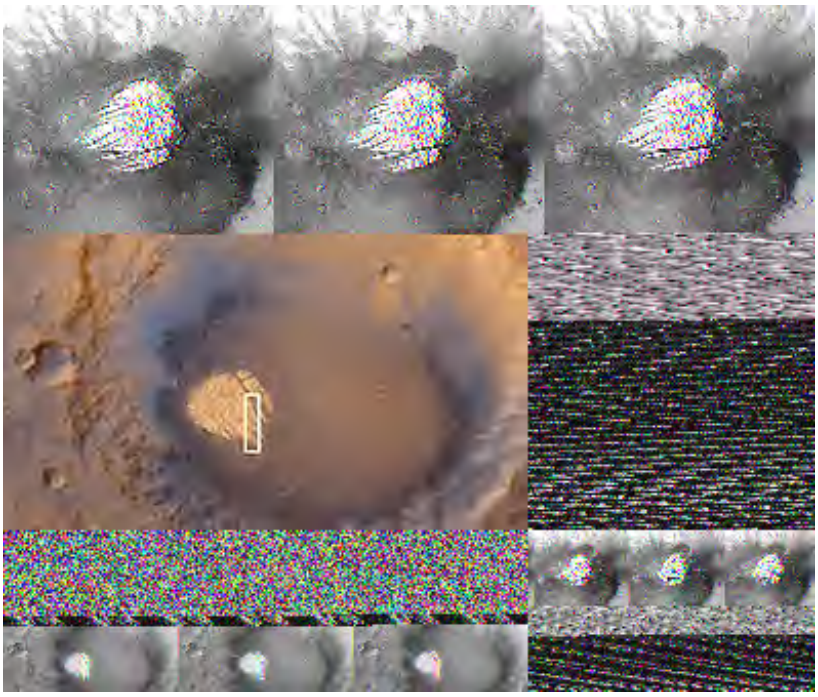
19-44: From the above list in the previous paragraph, decide which name best fits the terrain shown in the two images below; ignore the large craters in the first. ANSWER



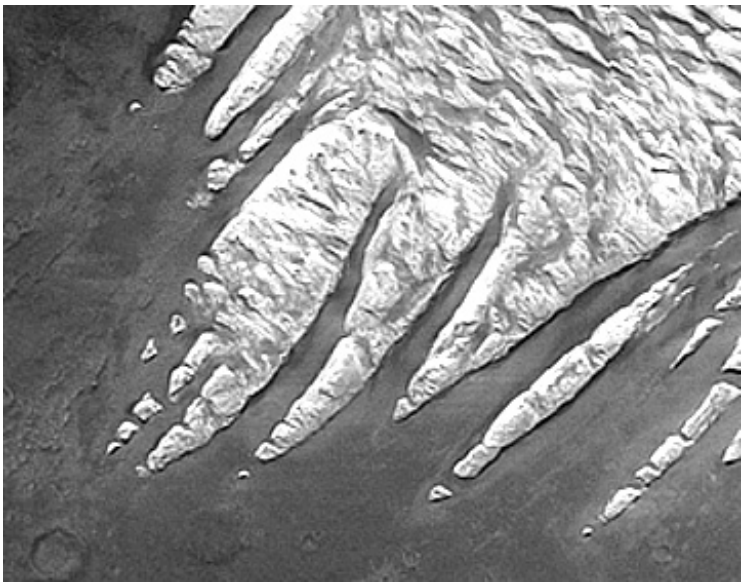


Some of these exotic terrains can also be called *enigmatic*. Lets illustrate this by looking at some images that center on what was called "White Rock" after its discovery in Mariner 9 images. The feature is a light-toned landmass, strongly embayed, that rises above the floor of the crater named Pollack (seen here in a MGS MOC image) in the southern highlands at a low latitude:

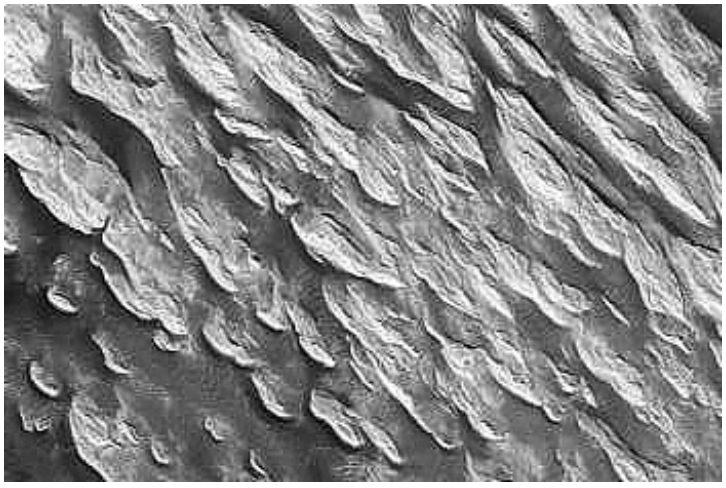
In a Viking black and white image, this feature, which is approximate 12 x 12 km in dimension, indeed has a higher albedo than the crater floor (very dark) and thus stands out as an off-white feature. Seen in a Viking color image (not shown): 1) White Rock has the same reddish surface coating that most of Mars has, and 2) its "whiteness" is largely due to contrast with the floor; its gray tone level is similar to surface materials beyond Pollack's rim.



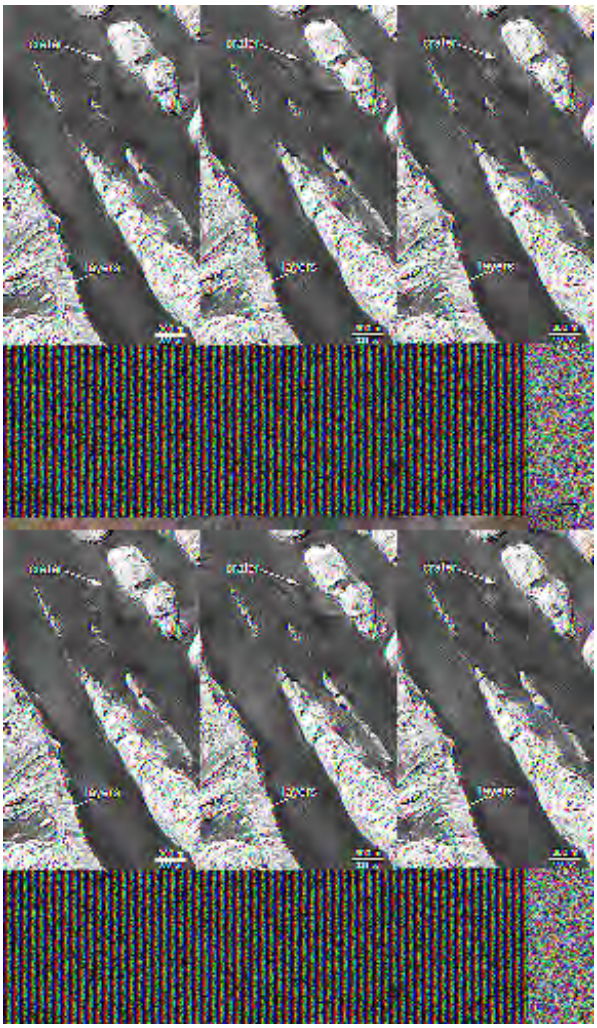
One of the early interpretations considered it to be ice preserved as a patch in the crater. One investigator proposed this feature to have been ice extruded from depth, much like salt forms in domes and may reach the surface. This was discounted by radiometric measurements that indicated too high a temperature and later measurements that ruled out H_2O and CO_2. Interest was renewed in White Rock from MOC images taken onboard the Mars Global Surveyor. Consider this next pair that zero in on the several prongs of the feature:



In the upper right of the above image is a cluster of the teardrop-like features. When enlarged, these seem to show thin layers but their shape may be attributable to yardang sculpturing by wind:



This image concentrates on several white prongs and the terrain in between. Of interest are the series of thin, arcuate equi-spaced lines in the dark areas between the ridges in the lower of the paired images. These seem to be controlled by the ridges. One interpretation is that they are regular dune-like markings that may result from martian winds that are directionally channeled by the ridges - but this is speculation.



Mars scientists now interpret this feature to be dissected lake beds, if Pollack was once filled with water. Others propose volcanic ash deposits that collected inside the crater and are being systematically removed (by wind?); no nearby volcanic vent is evident that would account for this. Still other explanations have been proposed. No consensus explanation has been reached at this time.

There is still much to do to properly categorize and explain the surface features on Mars. The sheer variety of landforms and related phenomena will require continued intensive study. New spacecraft with higher resolution imagers and other instruments are clearly called for. To anyone reading this who has not chosen a profession, the writer (NMS) strongly recommends a hard look at becoming a martian planetologist - probably of a "life's work" scope.

An excellent review article, "The Unearthly Landscapes of Mars" by Arden Albee, appearing in the June 2003 issue of *Scientific American offers further insights into the surface features of the Red Planet.*

As you shall see on the next three pages, the 21st century promises to bear witness to continuing and expanding exploration of Mars, by orbiters, landers, rovers, and, likely, humans.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

In early 2002, another spacecraft, Mars Odyssey, began taking data that provide maps showing the thermal state and composition of selected elements, including hydrogen, on the martian surface. These recent satellites and landers have also greatly improved surface geological and topographic mapping and crustal thickness estimates. The European Space Agency got "into the Mars game" with a strong success achieved by its Mars Express. The latest achievement has been the successful landing and deployment of a pair of Mars Explorer Rovers, a surface-traversing vehicle, Spirit on January 3, 2004 and Opportunity on January 24, 2004. Also on this page is a look at future plans for visiting Mars. The page also briefly treats the two tiny martian moons, both of which were imaged by Mariner 9.

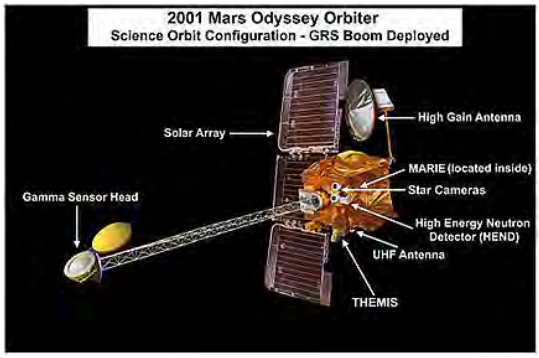
Missions to Mars during the Third Millenium

A series of spacecraft have been or will be sent to Mars in the first decade of the 21st Century. The first, the Mars Climate Orbiter, was launched on December 11, 1998. but failed to reach its planned orbit. This was followed on January 3, 1999 by the Mars Polar Lander which apparently did not achieve a proper setdown, as no signal was ever received. Orbiters and landers that have reached Mars and are returning data are described in this page. There will be a series of Mars probes launched in the next 10 years culminating in a mission which will gather samples for a return to Earth, perhaps by 2014. All of this is connected with the (considered, but not yet approved) U.S. decision to eventually send astronauts to Mars (the year 2030 has been mentioned but is so far just speculative), after the planet has been thoroughly explored with these probes. This decision is embodied in the President's speech given on January 14, 2004 at NASA Headquarters which can be read at this NASA history site; this talk strives to restructure NASA's goals for the next decade, but it has engendered both praise and much opposition (both in its emphasis and its cost).

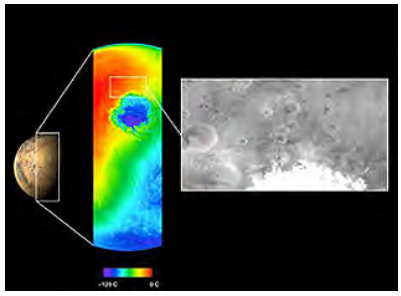
The Mars Odyssey Mission

One successful probe already is Mars Odyssey, launched on April 7, 2001. Access its Home Page at JPL's site. JPL's Webcast series has three broadcasts that cover the Odyssey mission. Access these through the JPL Video Site, then the pathway Subject--> Solar System --> Format -->Webcast --> Search to bring up the list the following: "Mars Odyssey: The Mapping Mission begins", Mar 1, 2002, then for more results "2001 Mars Odyssey Webcast", November 14, 2002, and finally "Live from Mars", March 13, 2002 (this last is oriented towards students and comes from the THEMIS project people at Arizona State University). To start any of these, once found, click on the blue RealVideo link.

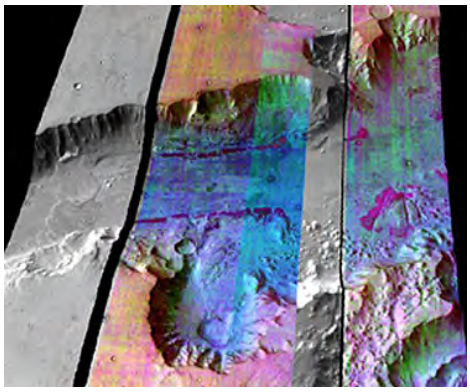
Here is an artist's rendition of the Odyssey spacecraft:



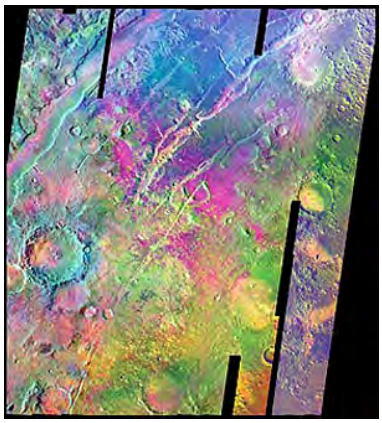
Odyssey was inserted into orbit on October 20, 2001 and has now aerobraked to descend to about 300 km from the planet. One of its sensors, THEMIS, a thermal infrared emission spectrometer, is surveying the gross mineralogy of MARS. A gamma ray spectrometer (GRS) obtains compositional data covering 20 elements, in particular evidence of hydrogen that is tied with oxygen as water. There are actually three instruments in the GRS: the Gamma Ray subsystem, the Neutron Spectrometer, and the High Energy Neutron Detector (built by the Russians). In its fully operational mode, the GRS is located at the end of the boom (see spacecraft picture above) which was deployed in June, 2002, Here is one of the first results: a visible image and an associated thermal map, made from THEMIS data:



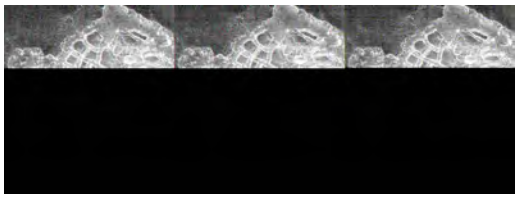
Bands 5, 7, and 8 on THEMIS were used to produce this next color composite of Ganges Chasma in Valles Marineris. The scene is about 150 km (100 miles) on a side. Blues are correlated with basalts; the purples indicate a high olivine content, which is typical of basalts with lower SiO_2 contents.



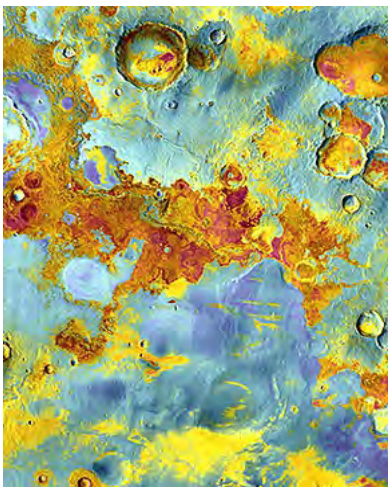
Similar variability in olivine content (magentas and purples) is evident in this THEMIS image covering part of Syrtis Major.

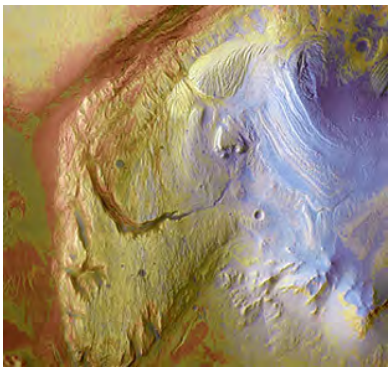


This next image from THEMIS shows a nighttime view of chaotic terrain on Mars. Note its similarity to nighttime thermal images shown in Section 9.

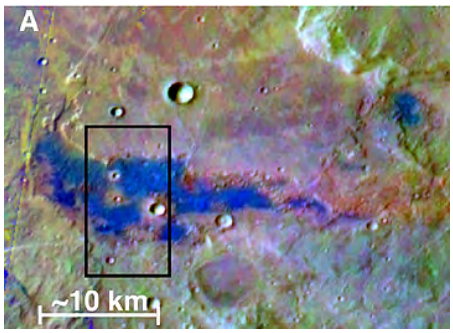


Color composites made by combining THEMIS Day and Night images have revealed additional information. In the color assignment used, reds generally mark exposed bedrock, blues denote loose, often sandy debris, and yellows may indicate coarser particulates derived from bedrock. Here are two image examples:

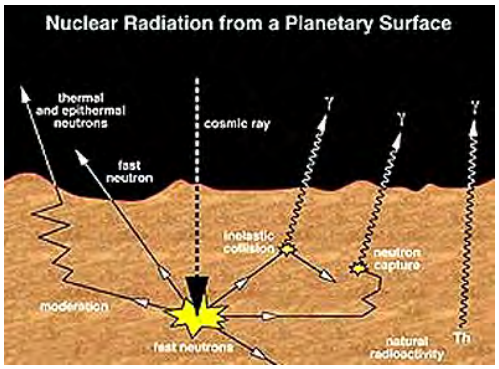




THEMIS is adept at picking out surfaces containing a variety of clay minerals (blue), as shown in this image:

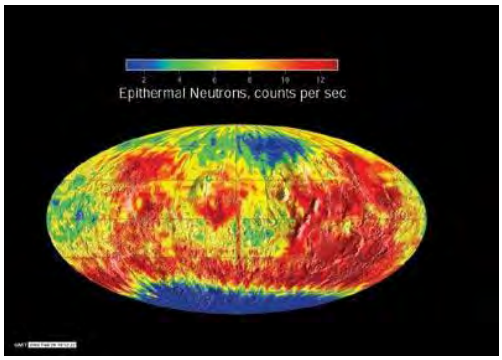


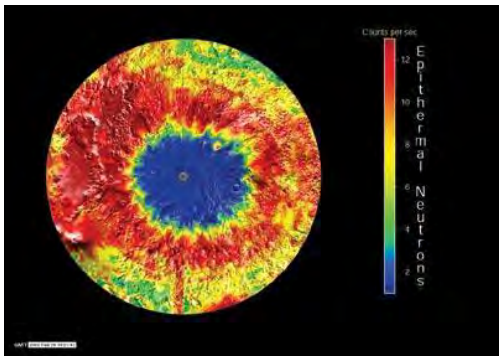
Referring to the GRS instruments, the principle behind production of neutrons from cosmic rays is illustrated here:



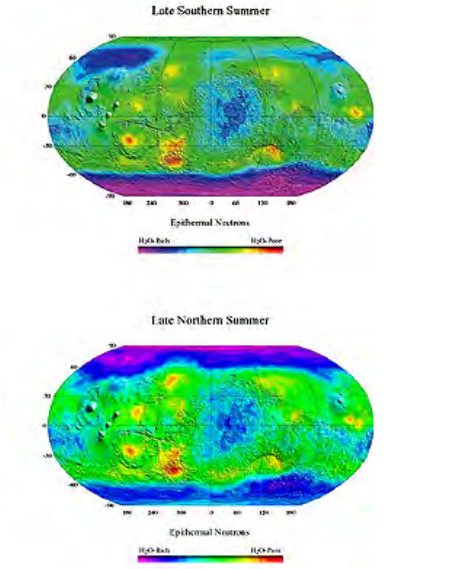
How are gamma rays and neutrons produced by cosmic rays? Incoming cosmic rays--some of the highest-energy particles--collide with atoms in the soil. When atoms are hit with such energy, neutrons are released, which scatter and collide with other atoms. The atoms get " excited" in the process, and emit gamma rays to release the extra energy so they can return to their normal rest state. Some elements like potassium, uranium, and thorium are naturally radioactive and give off gamma rays as they decay, but all elements can be excited by collisions with cosmic rays to produce gamma rays. The HEND and Neutron Spectrometers on GRS directly detect scattered neutrons, and the Gamma Sensor detects the gamma rays.

A preliminary map (below, top) of the global distribution of epithermal neutrons and a closer look at the South Polar region (bottom) appear below. Blue indicates conditions explainable by high hydrogen content; hydrogen moderates (absorbs) neutrons better than the heavier elements. While several explanations could account for this, the most probable is hydrogen bound with oxygen as water. More observations will be needed to confirm this.

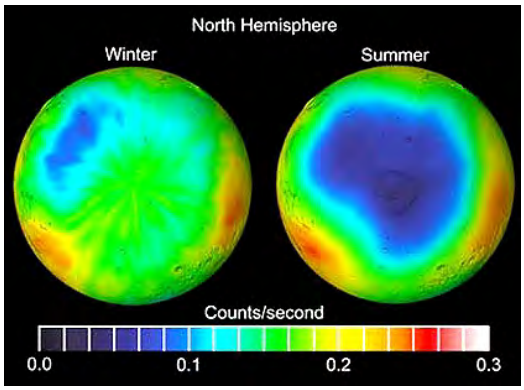




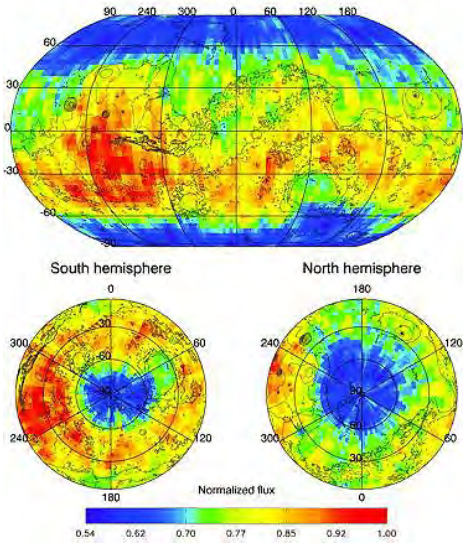
Maps made by the Neutron Spectrometer, released in December 2002, indicate even more potential water (higher hydrogen content in purple) than earlier announced. This pair of global martian maps shows the hydrogen distribution first when the south polar cap was largely free of CO_2 and later when the CO_2 had sublimated of the north polar cap and to some extent resolidified at the south cap.



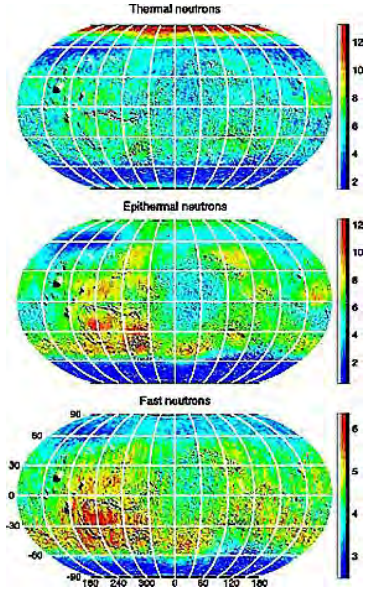
A 2003 map based on neutron data from Mars Odyssey emphasizes the wide variation in water ice at the northern pole beteen martian summer and winter. This is displayed best when first the CO_2 has sublimed off the cap leaving a dominance of the hydrogen-rich (low neutron count) substance whose identity is almost certainly water:



Russian co-investigators have produced a map of high energy fast neutron results which also show similar distributions of water:

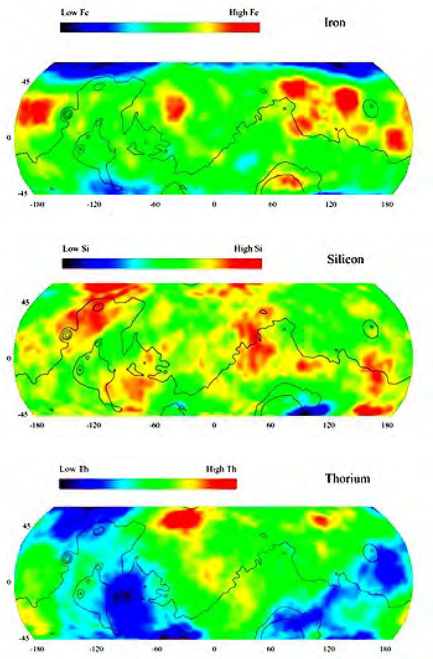


For close comparison, this diagram shows the general distribution of all three types of neutrons:

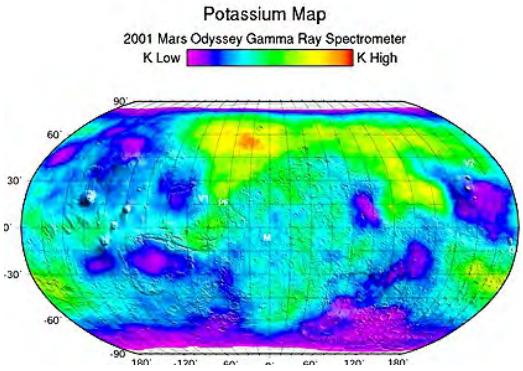


Calculations indicate there is enough water present in rocks and soil near the surface so, if released, water could cover the entire planet by up to 10 cm (4 inches) in thickness.

Although up to 20 elements can be detected by GRS, efforts were concentrated on these six: H, Fe, Si, K, Cl, and Th. Here are the results for Iron, Silicon, and Thorium:

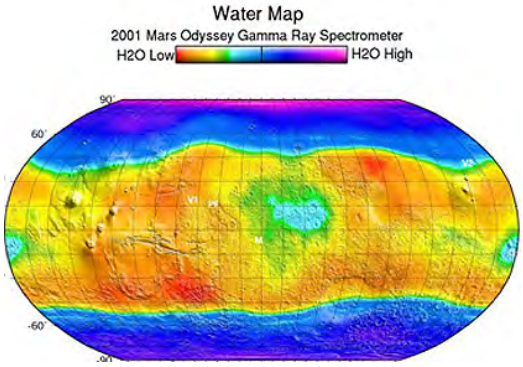


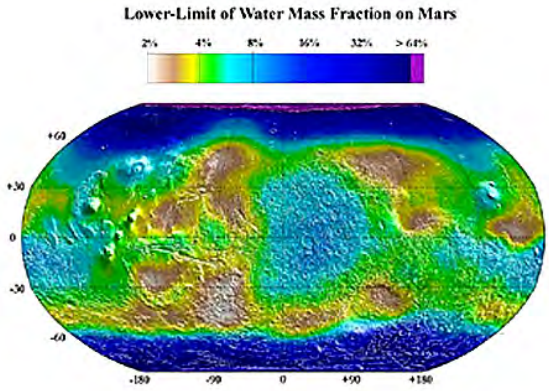
Another of the elements detected by analyzing the produced neutron flux is potassium. This global map of potassium distribution on Mars shows higher concentrations than initially predicted:



But water continues to be the focus of intense investigations. Improved water maps (based on hydrogen absorption of neutrons) continue to be released to the scientific community. Both MOLA and MGS data are used in making these maps. Hydrogen data come from the neutron spectrometer on Odyssey. In the blue areas, the amount of water ice is between 2 and 10% (assuming that all hydrogen is present only in water; hydrous minerals may contain both water and OH). Intermingled water increases to around 50% (yellows and oranges) to 70% (dark red to black) at the surface mainly around the poles; estimates approach 90% near the North Pole. But water highs are found in Arabia Terra and other areas at lower latitudes.

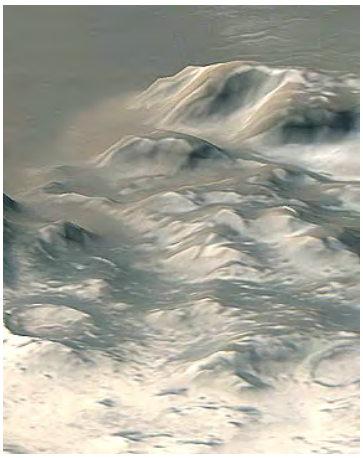
Compare the next two global maps, both made by Mars Odyssey. The top one, dating from 2001, shows a qualitative distribution determined without actual percentages of water. A global map released in August 2004 casts the immediate subsurface water content (dominantly as ice) in the fractional percentage (lower limits) of water present in the top meter:





Thus, the current conclusion from these neutron data sets is that there is much more water at and just under the martian surface than had been anticipated from all previous hints from earlier missions. In the polar and high latitude parts of Mars, it appears that water ice is quite abundant in the top meter or so of the debris that seems to cover most of the planet. The layer may be similar to permafrost in terrestrial rock and soil materials in Alaska, Siberia, and other high latitude land cover. Some of this martian layer can be considered as an ice bed, with subordinate amounts of rock fragments. This may grade into a predominance of fragments cemented by subordinate water.

As indicated above, some ice seems to extend into the low latitudes of Mars. There may be both a permanent subsurface layer and transient coatings. This perspective view of the Charitum Montes area (near Argyre Planitium), made by combining red and blue band MOC images with MOLA altimeter elevation data, shows a coating of (water?) frost on the hilly surface.



Frost had actually been seen closeup from the Viking 2 lander, as shown here:

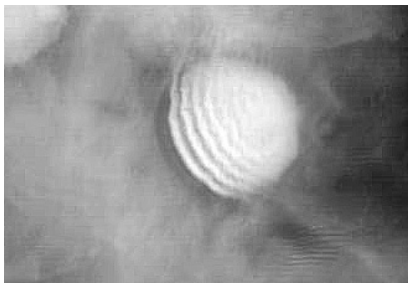
Frozen water has also been observed in shadowed portions of martian craters:

During each martian year a fraction of the water ice evaporates and is transported in the thin atmosphere to other locations. The transportation mechanism may be condensation of sublimed water and CO_2 as coatings on dust stirred up by the strong martian winds. Estimates of the amount of water thus bound to the surface environment have ranged from the total present in Lake Superior to much higher.

Such observations have led Philip Christensen and associates, planetary geologists at Arizona State University, to postulate that for periods of thousands of years conditions on Mars have permitted snow, frost, and ice to accumulate over wide expanses between the polar ice caps. This happens during certains phases of the change in Mars' rotation axis from 15° to 35° from the vertical relative to the orbital ecliptic plane. (They postulate from their model that early in martian history the shift ranged from 0 to 60 degrees.) The change proceeds at a rate of approximately 1 degree per 100000 years. This shift is oscillatory, currently going from 15 to 35° and back to 15°. During the process the climate changes owing to the different angles at which solar radiation hits the martian atmosphere and surface. At the higher angles, the poles receive more radiation, thus causing more water and carbon dioxide to evaporate and redeposit as snow and frost in lower latitudes. As the pole returns to lower angles (presently, now at 23°), this frost and snow resublimes and returns to the polar regions.

Besides the frost observations, another line of evidence they cite is the presence of the local gullies, such as those in the crater above. They believe that the water may be covered during low latitude accumulation and perhaps converted to ice. But climatic conditions at some stage bring about melting that carries the water just below the surface to crater walls and cliffs where the outflow of this groundwater gives rise to the numerous small gully channels.

MOC images may be monitoring transient "fogs" in the circulating atmosphere that are made up of water vapor and perhaps ice crystals. Look at this image:



The uniform gray is thought to be the "fog". As it passes across a crater (66°S), the topography causes a disturbance that produces ripples in the water-bearing thin atmosphere.

In a widely held model, many martian specialists believe the planet began with **an active magnetic core, had a much thicker atmosphere than present, was warmer, probably had water, and may have had primitive life**. Mars probably changed dramatically in its early years to become more passive ("dead") but the possibility of **significant amounts of water ice**, when/if confirmed by future landings (preferably, polar latitudes), indicates that somehow life of some kind may have survived since its glory days in the first billion years.

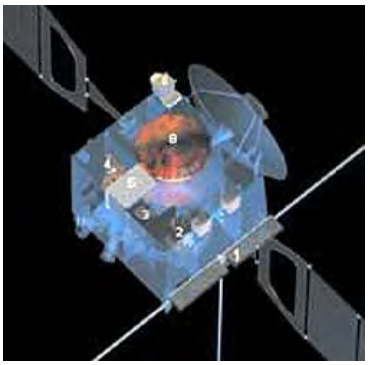
This seemingly valid discovery of surface and atmospheric water needs verification and more specific new data. Needless to say, the likelihood of abundant water has galvanized that segment of the scientific community that promotes an ultimate manned mission to Mars. Presence of extractable water not only supplies drinking needs but processes exist to break

down the water into hydrogen and oxygen gases which can be made part of a fuel system to power vehicles returning to Earth. Oxygen thus released could be used for breathing in any base established on Mars for continuing manned exploration. A number of missions to continue exploration have either been approved or are being carefully planned and considered. Among these that already have happened are ESA's Mars Express and two NASA Mars Explorer Rovers (MER) both of which started for Mars in 2003 when the planet was in an orbital position that brings it as close as 35 million miles from Earth.

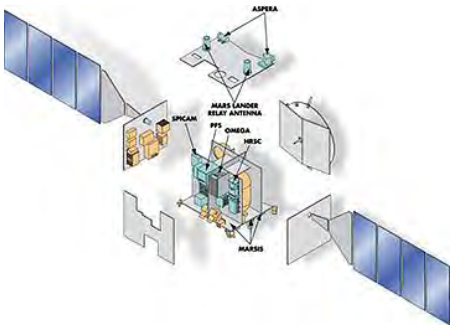
(The Japanese Space Agency sent a Mars probe, called Nozoma, in July, 1998. Because it failed to get enough "kick" to its thrust towards Mars it had to receive the needed boost by repeated orbits past the Sun. This meant that the total trip was 4 1/2 years long. As it approached Mars, onboard systems failed and the spacecraft was lost without achieving any of its goals).

The Mars Express Mission

The European Space Agency's Mars Express was successfully sent towards Mars on June 2, 2003, and arrived in December of that year. It is pictured in a sketch below.

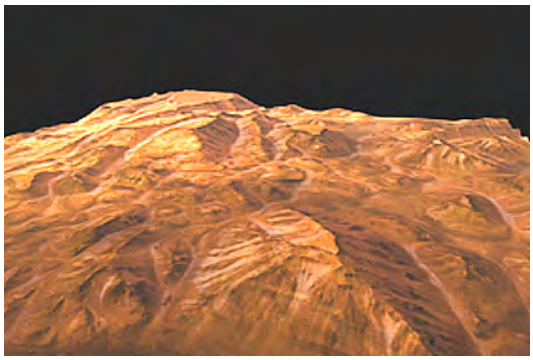


Its instruments are described at this Mars Express site. Among its 9 instruments is MARSIS, a radar capable of penetrating many meters into the Mars surficial deposits. It also carried a lander, named "Beagle 2. This view is an "exploded" diagram of the prime instrument on this spacecraft:

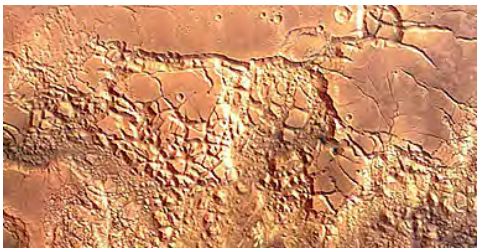


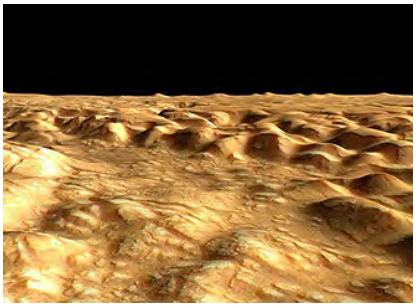
Mars Express orbited Mars successfully and on December 23, 2003 released and propelled the Beagle 2 towards its landing site. After initiation, nothing was radioed to Earth that would validate a successful landing. Meagre evidence indicates that either the probe was destroyed enroute or did not deploy correctly (even if intact, its radio antennae may be pointed such that no signal is leaving the surface). Thus, this part of the mission failed but the Express has sent abundant data back, thus rescuing the effort from total disaster.

So, the Express orbiter is working well, as indicated by the next two images covering parts of Valles Marineris. Its imaging device produces stereo, so that the resulting views have a more 3-dimensional character than normal (but, of course, previous orbiters have yielded 3-D imagery after processing). The ground-penetrating radar deployed its three booms (holding antennae) and became operational in May, 2005. Examples of a vertical and a perspective image of a region are given by these views of part of Valles Marineris.



Another pair of examples are the chaotic terrain in Aureum Chaos near Valles Marineris. The nature of that terrain is evident in the diversity of forms in the vertical view. When part of the image is converted to a perspective view, the nature of the hills displayed leads to questions about their origin. One view considers them to be the result of differential settling as underlying ice was melted.

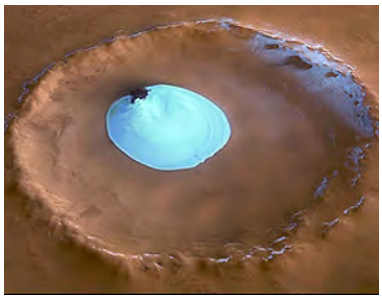




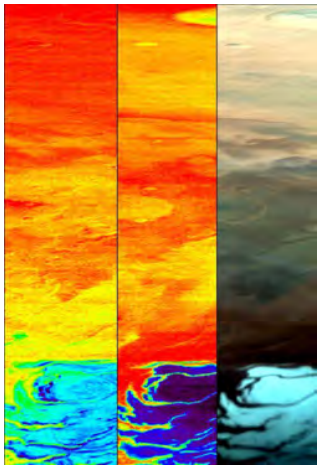
Another image shows curved grabens (uplifted) and horsts (downdropped) blocks caused by faulting in the Tharsis region, in which the martian crust has been generally raised and stretched causing tensional stresses that result in this faulting.

This Mars Express image shows the Reull Vallis, interpreted by ESA scientists as a valley that almost certainly was formed by low viscosity fluid erosion, namely water.

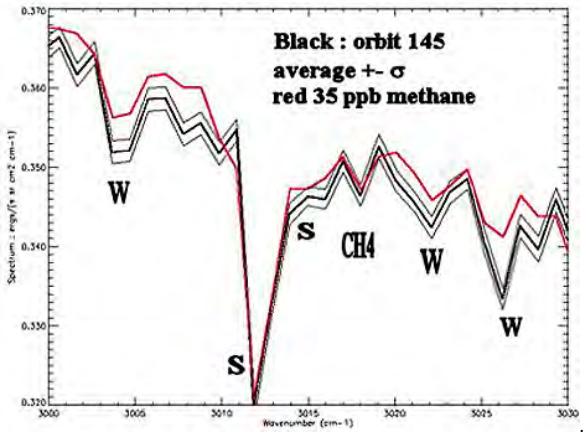
One surprise Express find: a standing body of frozen water in a large "pond" within a crater some 35 km (21 miles) diameter in the Vastitas Borealis region at a high latitude south of the Northern Ice Cap:

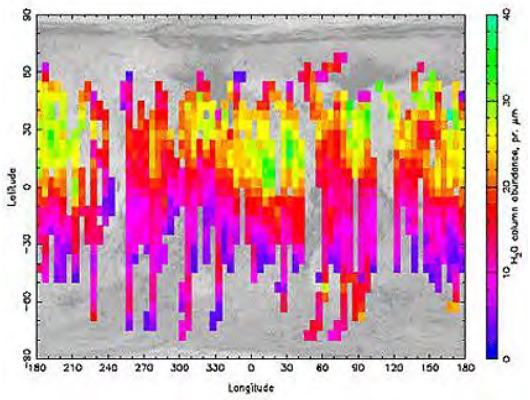


Mars Express has also confirmed the presence of water in the South Polar Cap ice. It has an instrument, Omega, that has a Visible-IR spectrometer. Below, the right image is a visible color view of part of that Cap; in the center, the blue pattern demarcates the frozen carbon dioxide; in the left image, the blue indicates a frozen water response:



Among significant findings by Mars Express so far are the detection, using the Planetary Fourier Spectrometer [PFS]) of trace amounts (10 parts per billion[ppb]) of methane (CH_4) and water vapor in the martian atmosphere. These seem to be more concentrated in certain regions:



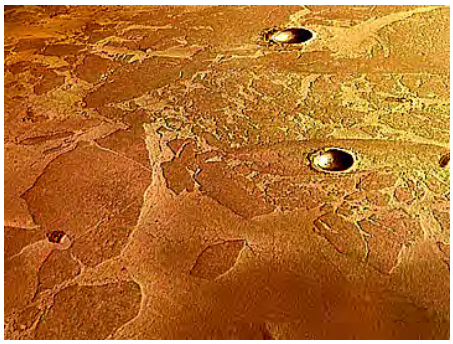


The methane may be residual from early days in the evolution of the atmosphere but its presence may indicate some youthful source(s). One could be from vaporized comet(s) that impacted the surface, but no large young craters have been found. A second possibility is from volcanic release, but again no visual evidence of recent volcanism has been detected. The third option is both provocative and conjectural: release from the decay of organic matter in surficial deposits. The co-association of methane and water vapor could point to either of the second and third options. But the presence of small volcanoes over the "highs" in Arabia Terra place this last option as presently the most favored Further observations are needed to determine if the CH_4 is uniformly distributed or is concentrated geographically. Ultimately, to prove an organic

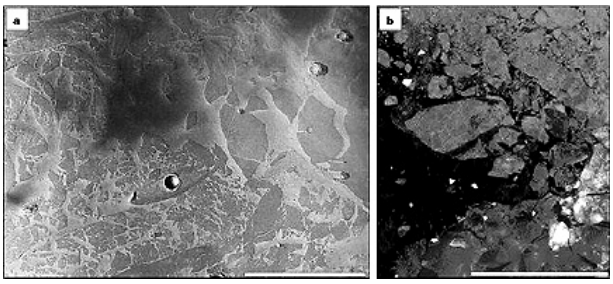
nature for the methane, a future probe or lander hosting a mass spectrometer will be needed, since the isotopic proportions of C_{12} to C_{14} can clearly distinguish between an inorganic and an organic (biologic - bacterial decay) origin of the methane.

A recent idea from martian scientists based on Mars Express imagery claims to see evidence that water has been expelled from larger volcanoes over hundreds of millions of years. This conclusion is based on two arguments: 1) there are numerous small channels on the gentle slopes of these volcanoes that are possibly from water (but some could be lava channels); and 2) using crater counts and other age dating methods, the various overlapping flows on the volcanoes can be relatively dated (and from comparison with crater data on martian plains, rough actual ages can be estimated).

On February 23, 2005, at a European conference on Mars Express results, researchers presented information that interpretation of topography in Elysium Planitia can be evidence for a vast field (800 by 900 km [450 x 560 miles]) of what could be frozen water ice covered by a protective layer of volcanic ash. The flat plates, the fracture patterns, and the curved rims of inpact craters (typical of craters that form in ice) are the basis for their claim. They believe the ice to be a once liquid "lake" that quickly froze, perhaps as recently as 2 to 5 million years ago. This observation is a high priority for confirmation by the MARSIS radar.

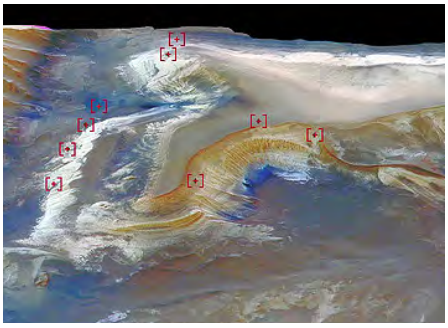


In one locality, the Mars Express image seems to indicate the ice is at or very near the surface, over a wide area, such that one interpretation holds it to be pack ice. In the image below, pack ice from the Arctic is shown next to the martian ice for comparison.



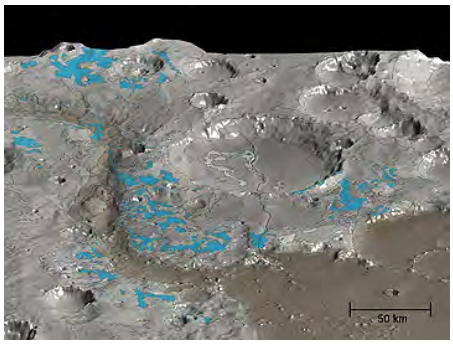
In another region, the Mars Express may have detected a flowing glacier presumably of ice. It is shown here compared with a typical small terrestrial glacier, on the right:

In December 2005 various groups involved in Mars Express data analysis revealed important new information at a press conference. First, was confirmation that parts of Mars have phyllosilicates (flaky clay minerals) that resulted from weathering of ancient volcanic rocks. These clays contain water introduced when it was plentiful enough to produce these minerals during a hydrous phase of surface activity, probably early in martian history. Here is one map showing the clay mineral distriution, as mapped by OMEGA, shown in brown (the perspective view was created using Express's HRSC imagery):

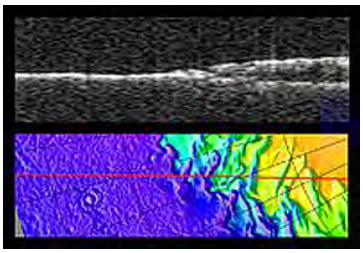


OMEGA also can detect certain sulfate minerals that, as will be discussed in the Mars

Exploration Rover segment on the next page, have been found elsewhere besides the MER sites. This illustration portrays detectable Bieberite (a magnesium sulfate mineral) depicted in blue that forms almost exclusively through the action of water. The area shown is the Marwth Valley; other areas where this mineral has been found include Arabia Terra, Terra Meridiani (within which the two MER spacecraft landed), and Syria Major.

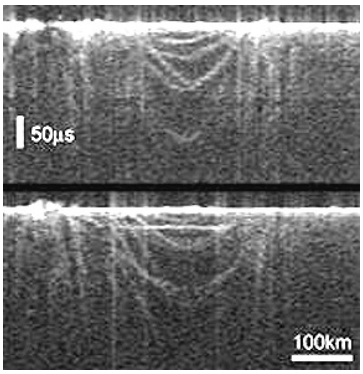


That press briefing also included several of the first images created by MARSIS (described at this ESA website). The top image shows surface radar reflections from an area that includes part of the South Polar ice cap.

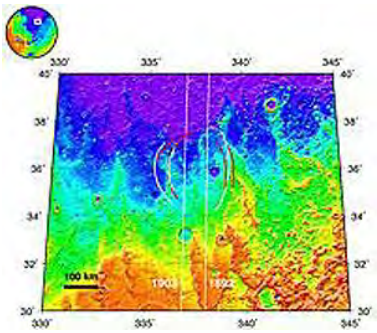


Note that the radargram (an echo signal) splits under the ice cap. The top bright echo is a surface reflection from the ice; the bottom echo marks the base of the ice. The thickness is up to 1.1 km (0.65 miles).

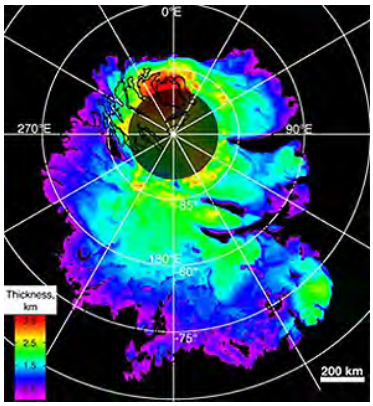
In this region, deep radar penetration has detected reflections from a large (250 km; 156 mile diameter) bowl-shaped feature which is being interpreted as a buried impact basin.



The outline of this buried basin is plotted on this map of Chrysae Planitia.



The most significant finding by Mars Express was released in March of 2007. The MARSIS has completed enough of its scheduled survey of the southern polar ice cap to produce this map of the thickness of the frozen water below a surface of mixed water ice and frozen carbon dioxide:



The maximum thickness is 3.7 km (about two and a half miles). The ice seems to have a purity of > 90% H_20. If all of this ice were melted, there is enough to cover the entire martian surface with an "ocean" 11 meters (about 35 feet) deep.

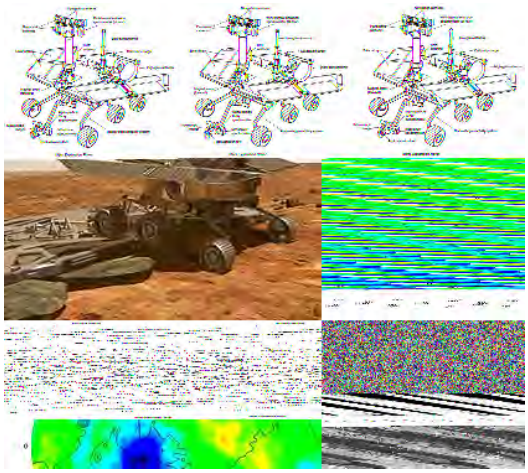
A large number of images are being made by the M.E. spacecraft and some are released periodically. Seek these on the Mars Express home page site.

The Mars Exploration Rover (MER) Missions

The American Rovers are another story with happy results. The place to start is the JPL Rover Home Page; after looking at its options choose Overview. Also, visit the JPL movies that set the stage for the Rover program. Access through the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes "Rough Guide to Mars", February 4, 2003 and "Rover Mission to Mars", June 6, 2003. To start either one, once found, click on the blue RealVideo link. A third webcast in the series also concentrates on the Rovers but is brought online through a different pathway, Access it through the JPL Video Site, then the pathway Subject-->Von Karman Series 2003 --> Format -->Webcast --> Search to bring up the list that includes "The Mars Exploration Rovers", August 21, 2003. To start it, once found, click on the blue RealVideo link.

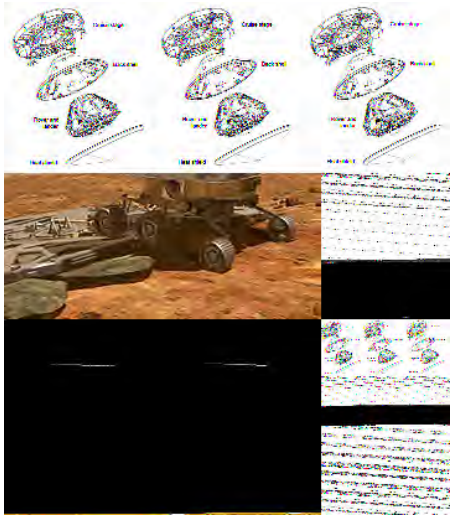
The two MERs are both designed to search for water using a variety of instruments, including a mini-Thermal Emission Spectrometer, a Mossbauer Spectrometer, and Alpha Particle X-ray Spectrometer. Each MER looks like this:

This schematic diagram shows the main instruments on a MER.

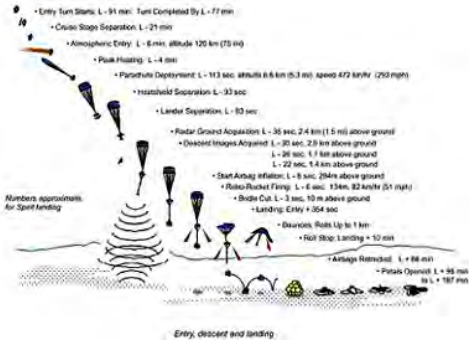


The Mars people at the Jet Propulsion Lab have produced a video of both landing and then deployment that can be accessed at this website. I have downloaded Microsoft MediaPlayer (MPEG links) to monitor the animation (QuickView also works). This overview is highly recommended, especially since it highlights the technical achievements involved (shown on a PBS NOVA program prior to the landing).

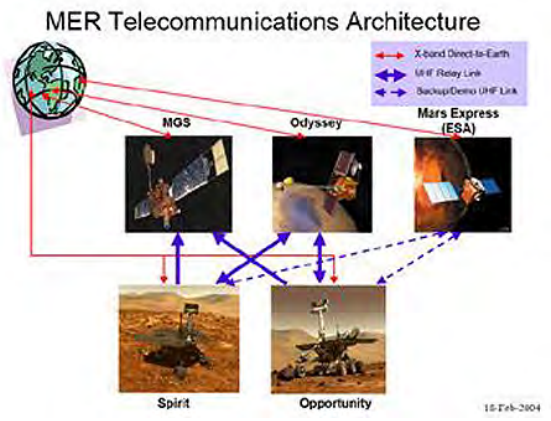
The four major components of the MER spacecraft are sketched and labeled in this schematic (see caption for more information):



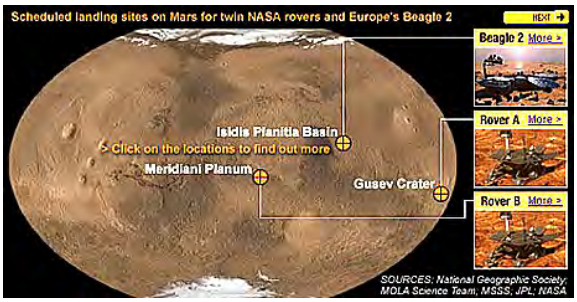
The manner of its landing (much like the earlier Pathfinder) is spectacular: consider that the vehicle needed instructions (some were preprogrammed) sent over a distance of more than 100 million miles) and each step had to be perfect in execution. Using both parachutes and complex airbags, the landing sequence is depicted here:



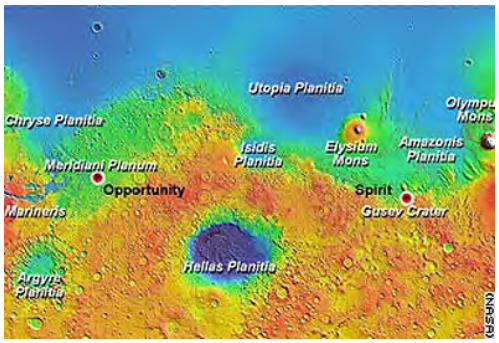
General communications with, and data from, the two rovers are transmitted either directly back to Earth (about 20% of the total) or are related to Odyssey, MGS, or Mars Express, according to this scheme:



MER-1 (also referred to as MERA), nicknamed "Spirit", was launched towards Mars on June 10, 2003. MER-2 (MERB), called "Opportunity", was launched on July 7, 2003. Here is a plot of the landing sites for Express's Beagle 2 (failed) and the two MERs set on a backdrop of the Nars surface in a global projection:

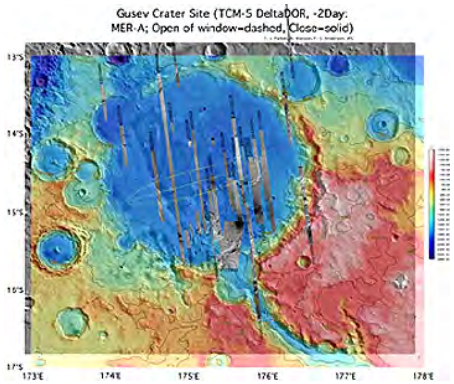


Spirit arrived at the Red Planet on time and sent its Rover to the surface with a **successful** landing at 8:35 PM (PST) on January 3, 2004. Spirit set down in the expansive Gusev Crater, which was postulated to contain frozen water inside. Opportunity (discussed on the next page) went to Meridiani Planum and landed among rocks rich in hematite (iron oxide) which can be associated with water-rich conditions. Here are both landing sites as mapped using Mars Orbiter data:



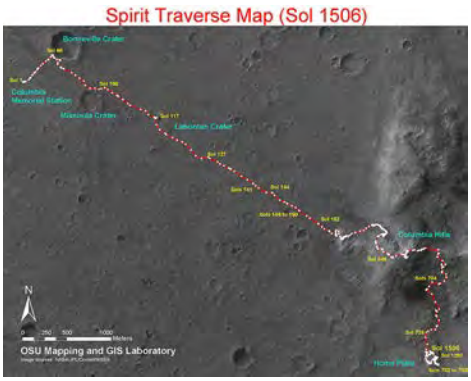
The Gusev crater, which is about 200 km wide (125 miles), appears as follows in a Mars Orbiter image:

A map of elevations in and around the crater appears here:

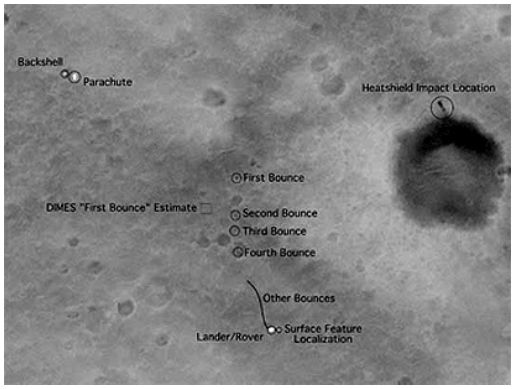


The rationale behind choice of Gusev as one of the MER sites was based on several criteria: 1) the apparently smooth, flat surface is typical of lake beds (Gusev was postulated to have been filled with water at some time); 2) a long 800 km [500 mile], sinuous channel, Na'adim Vallis, seems to have breached the crater walls; 3) faint layering is visible in high resolution images of the crater walls. Another constraint on site selection was to have it about 180° from the second rover site, so that one site faced Earth during an approximately 12 hour period (allowing signal transmission) and the second site could broadcast during the other 12 hour time frame (Mars' day is a little longer than 24 hours).

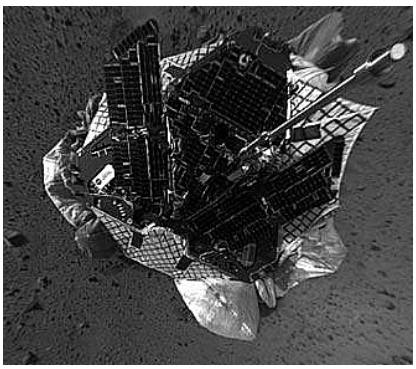
This MOC image shows the initial touchdown point at Bonneville Crater and the subsequent traverse path followed by Spirit as it eventually reached and climbed the hills to the southeast.



Soon after touchdown, Spirit was located by the Mars Global Surveyor, in this MOC image, which also shows what is interpreted to be surface disturbances caused by the bouncing (an estimated 28 times) of the airbag-cushioned rover as it moved up and down after first impact:



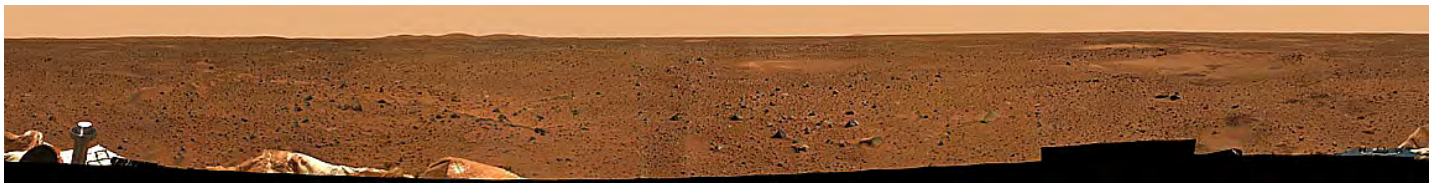
The Spirit Rover took a picture of itself using the Navigation camera mounted on a pole at the top. Here it is sitting on the surface:



Pre-landing interpretation of the Gusev floor's geology led to the assumption that this crater;s floor was once covered with water, i.e., it was an ancient lake. Thus rocks within, and perhaps the soil, might show some signs of life, since water normally is a prerequisite. The first color image produced by Spirit does not show any direct evidence of lake deposits; in fact, it closely resembles surfaces imaged by the Vikings and Pathfinder:

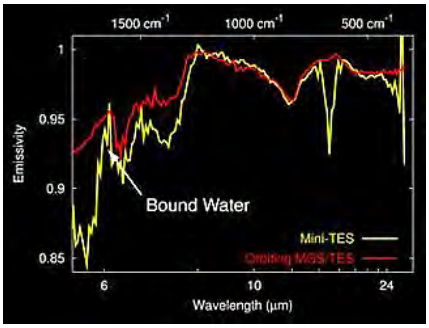
None of the rocks looked at during the early traverses (examples just below) seem to show lacustrine layering (lake bed deposits on Earth often show narrow bands of alternating composition - light/dark/light/dark - indicating seasonal changes). First impressions are that these rocks are similar to basalts, which may be the dominant rock type on Mars. Even if Gusev does contain lake deposits, many (most?) of the surface rocks at the site could be ejecta from impacts beyond the crater that penetrated into basaltic bedrock. The hope for confirming lake beds rests mainly with the presence of their characteristic rocks as ejecta from local smaller craters within Gusev Crater. These rocks may be lighter in color/tone; several lighter rocks have been spotted so far

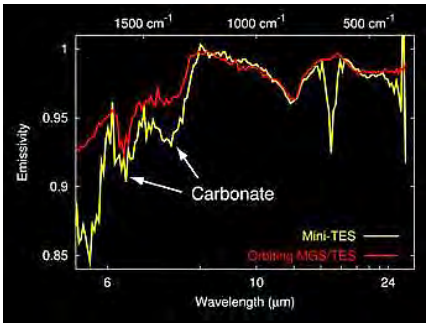
While waiting to deploy the Rover on its first excursion, sensors on Spirit acquired various imagery and data. Look first at a 360° panorama around the touchdown site, made using more than 200 images (scroll horizontally to see the full extent):



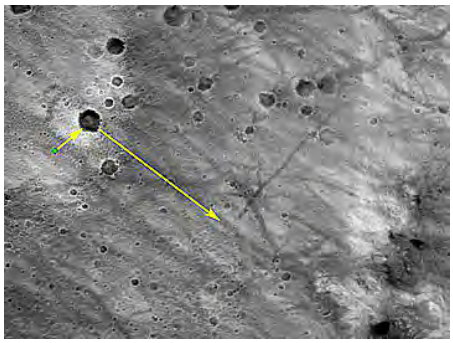
The surface within the crater is not all flat. Low hills are seen in the background. These initially were called the East Hills but since have been renamed the Columbia Hills. Chances are the hills themselves are upthrown floor or interior peak rocks made during Gusev's impact formation and would likely be older martian crust (basaltic) rather than lake bed deposits. The seven crests along the hills were remaned after the 7 astronauts who died in the Columbia Shuttle disaster. Here they are, seen in a panoramic view from Spirit:

The onboard MTES (Mini Thermal Emission Spectrometer) has produced spectra for the area close to the instrument. Both silicates and CO_2 were detected; the latter is not in the rock/soil but in the martian atmosphere. Here are the plots.

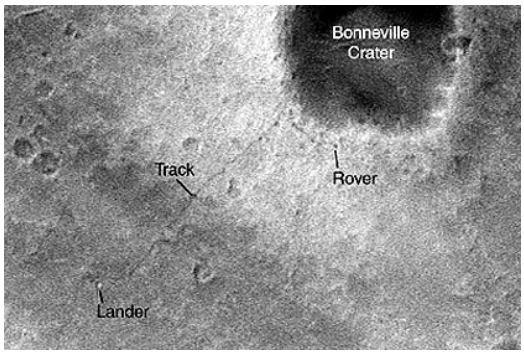




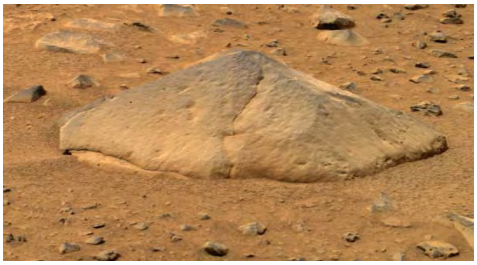
Spirit took off on its first excursion on January 14, 2004. The planned traverse is shown below. The first major objective was the impact crater named Bonneville which has excavated the possible lake beds to a depth of 30 m. Rocks enroute were examined to see if they are compatible with the types formed in lakes. The dark bands in the image probably represent the tracks of dust devils that remove surface fine particles exposing darker rock below.



Using a special technique, the Mars Global Surveyor was able to image the Bonneville scene at 0.5 m resolution, sufficient to pick out the landing site platform, the travel track to Bonneville, and the rover itself on the edge of the crater ejecta slope (note that the surrounding blanket is lighter in tone than the general martian surface traversed; this is not an indication of a compositional change but probably results from the higher reflectivity of the ejecta associated with the crater).

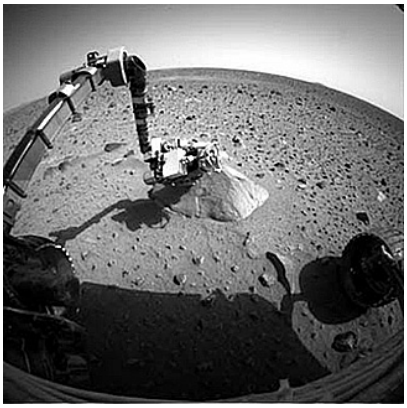


The trip to this crater (whose dark surface suggests basaltic rock) took time, as experience was gained in navigating the Rover from millions of kilometers away. The first stop was an analysis of a football-sized rock nicknamed "Adirondack" (an American Indian name for "They of the Great Rocks"). It is believed to be volcanic (likely, basalt) in nature. Here is a natural color view:



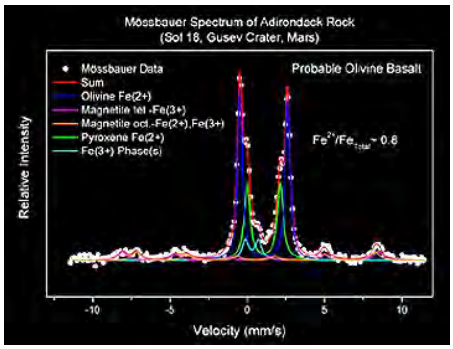
Enroute to Adirondack, Spirit's analysis of nearby soil showed the presence of olivine, a greenish magnesium-iron silicate that is a constituent of basalt. The mineral probably is present as loose grains blown to the site by winds.

This view shows the extended robotic arm (Instrument Deployment Device: IDD) containing several of the analytic instruments up against the Adirondack rock:

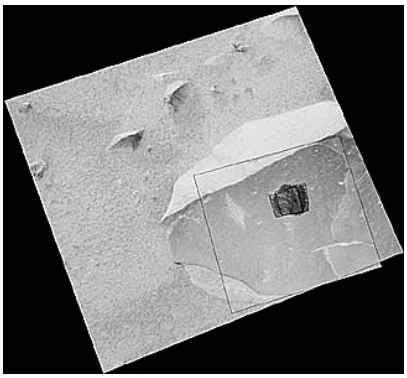


On January 21, 2004 Spirit went temporarily into a "repair mode", as it ceased to send back data although ground controllers at JPL could still communicate with it. Thereafter, several small streams of data were received but were largely unusable. The problem has been diagnosed as a software misfunction in which the storage device simply ran out of memory when commands were misinterpreted, and caused file corruption. The "flash memory" has since been purged and debugged, and functions reset. Mission Control is cautiously optimistic that the Spirit Rover can continue its excursions and collection of scientific data indefinitely and that the same problem in Opportunity can be avoided by modifying its software.

On January 30, Spirit returned a spectral analysis of the Adirondack rock, confirming it to be basaltic in composition, as seen in this spectral curve:



Then, on February 4th Spirit began a visual inspection of the Adirondack surface. On a "hunch" the operational team suspected there might be a thin coating of dust on that surface. Attached to the RAT (Rock Abrasion Tool) is a wire brush. The rover was instructed to brush off a small patch on the surface, with this result:



The dust coating (almost certainly hematite) was less than a millimeter thick; it may be adhering by electrostatic forces. This dust appears to be widespread across Mars' surfaces and may be compromising some of the compositional data obtained from the Viking sites and from orbit (e.g., Odyssey and MGS). When the dark patch was examined by the Microscope Imager (MI), this image was obtained - it appears consistent with a rock having a texture typical of basalt.

Adirondack is the first rock to be drilled into, using the RAT (Rock Abrasion Tool). Here is the rock showing the circular cutting outline:

As Spirit moved on, it spied one unusual rock lying as float on the surface. This rock (below) is splitting into thin flaky layers or chips. It resembles a fissile shale on Earth but is richer in iron than most shales (but there are red shales). Such layering is usually a consequence of continuing deposition in a fluid medium, usually water.

Spirit's Microscope Imager looked at tiny dunes (sand ripples) as it traversed towards Bonneville crater. The resulting grains on a ripple crest range from angular to round (and may relate to the spherules found at the Opportunity site?):

On March 10th, Spirit reached the rim of Bonneville crater (about 200 meters in diameter) and looked in to see the interior:

At this resolution, no obvious layers of rock appear within the crater (some blocks in the far slope may be loose or may be outcrop in otherwise loose fill).

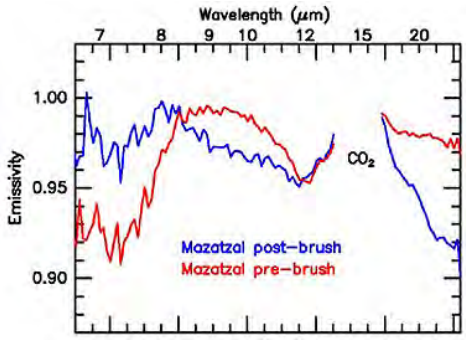
While in the rim, Spirit took a pan camera image of a scuffed (scraped) surface. It shows that darker material lies just underneath, which may be basaltic sand covered by hematitic dust:

Spirit's Microscope Imager has also examined a sand deposit on the Bonneville Crater rim. After it was scraped, this image shows a crust of lighter-tone granular material overlying a darker substrate:

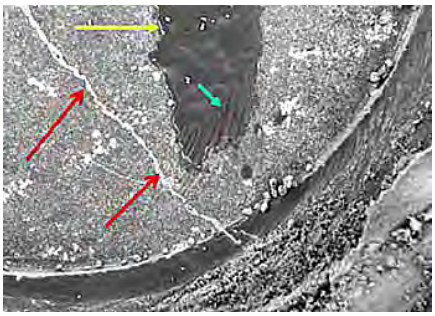
The lower zone contains rounded particles up to about 0.5 cm in diameters mixed with much smaller grains (~50µm).

As it moved from Bonneville, Spirit spied a large rock which was named Mazatzal. It was partially covered by windblown sandy soil

Several RAT grinds penetrated this rock. It revealed at least two very thin layers which have different emissivities, as indicated in this plot:

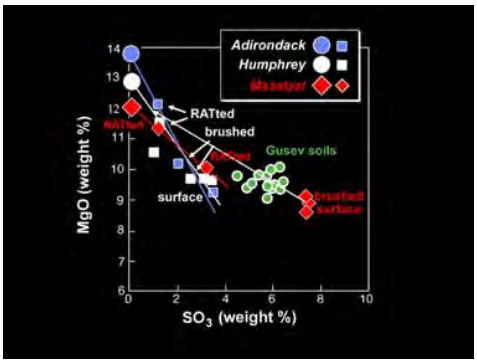


The pre-grind surface is darker in the visible; after grinding the rock underneath is brighter. Thus, this basalt block has experienced some process that darkens its outermost layer (or deposits it). An analogy found in Earth rocks is the formation of "desert varnish" which is blacker because of MnO_2; but this composition has not been confirmed for Mazatzal. Part of that dark surface remains in the RAT hole seen in the next image:

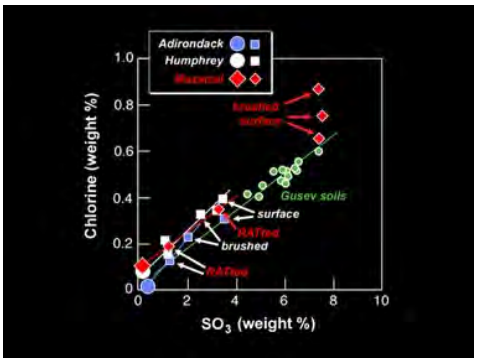


The presence of a cross-cutting vein filled with light material has caused excitement among the Spirit scientists. Although its composition is as yet unknown, the filling is postulated to be the result of deposition of mineral(s) of secondary origin that precipitated out of a hydrous solution that came from beneath the basalt before it was released (probably by impact) from its bedrock. This is believed to be an indicator of hydrothermal activity after lava emplacement and cooling. (Groundwater circulation is another suggestion.) This veining in this and another Gusev crater rock is the best evidence so far for some water having affected the bedrock present in the crater.

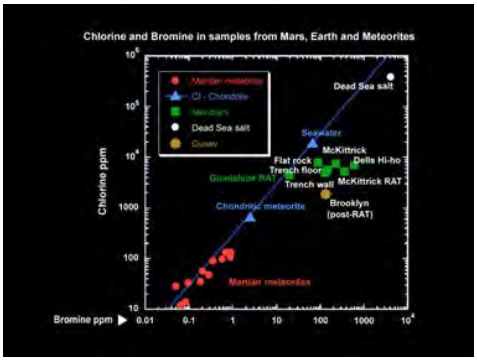
However, APXS data have confounded interpretations of whether a water body at least briefly covered the Gusev surface after the basalt blocks were brought to the surface. Examine this plot of Magnesium oxide (MgO) content versus SO_3 content of the very thin veneer covering several Spirit rocks including Mazatzal:



The three large symbols on the ordinate are from the rock materials beneath the veneer. No free sulphur is present, indicating its absence in the basalt. However, fresh, brushed, and slightly RATed surface all have evidence of sulphur, shown as a trioxide but without any specific mineral type being identified (still unknown). Speculation has it that this is some type of coating that may have been associated with precipitation of sulphur-bearing minerals on the surface and in the soil. A similar pattern is found for the data that indicate Chlorine (Cl) to be present in the veneer and to increase in proportion to the rise in SO_3.



Both Chlorine and Bromine are found in coatings on rocks from both the Spirit and Opportunity sites. On Earth Chlorine and Bromine tend to increase linearly in evaporites. The situation is somewhat different in rock coatings from both MER sites, as shown in this general plot of the concentrations of these two co-varying elements.

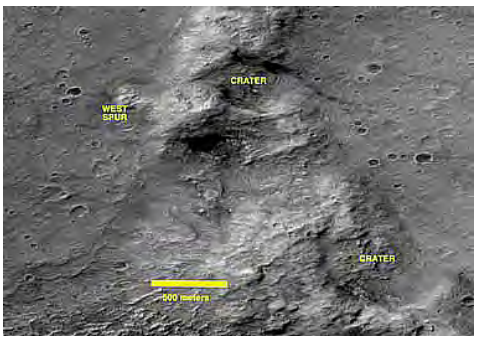


Summing up the first year, the story at the Gusev Spirit site was this: The only rock type firmly identified is basalt; hematite is of the red variety; although several rocks may contain hydrous mineral(s), no other signs of lake water or water-related sediments, including strata of any kind, had as yet been detected; however, hydrothermal or groundwater activity from subsurface sources has been verified. Coatings containing Cl, Br, and SO_3 remain enigmatic. One hypothesis is now more favored: the coating(s) may be dust that has been cemented as a crust onto the surface with water from an atmospheric source. In retrospect, the Gusev site does not now seem to have had a lake or extensive sediments, still, water activity is probable but not to the extent favoring life, and the sinuous channel may not be fluvial in nature. The possibility that Gusev contains sediments below a basalt layer as established from the composition of the rocks scattered about the surface, the latter thus being a later emplacement of covering lava, cannot yet be discounted but later cratering into the basalt has not penetrated to the depths of these possible deposits.

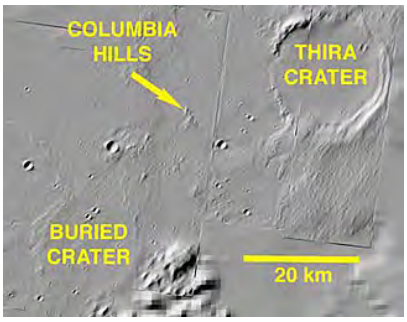
Spirit set out in late 2004 for the Columbia Hills and reached them in early 2005. In this view of these Hills hints of layering can be seen through its pan camera. Enroute, Spirit has examined other surface rocks, especially those lighter in tone, as seen in these views, to test the surface coating hypothesis or find alternate explanations:

As the Columbia Hills were approached, possible outcrops were detected:

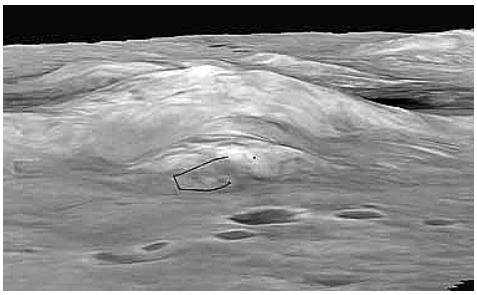
As seen in this MOC image from MGS, the Columbia Hills appear as below.



As Spirit closed in, the rounded nature of these hills becomes more apparent, as shown in the above images. From these images, one might get the impression that they are rather high. In actuality, the highest point is just 90 meters above the edge of the float block apron. This MOC image shows that these Hills are just a "blip" on the surface inside Gusev Crater. Nevertheless, they may reveal the nature of materials in the pre-impact surface.



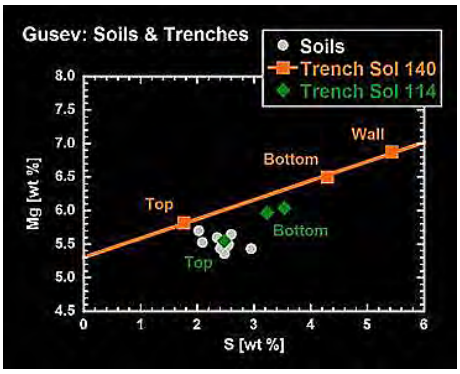
This next image is a perspective of the Columbia Hills made from the same MOC data. It shows the path being followed at West Spur by Spirit during its explorations.



West Spur was examined in early June of 2005. Spirit then moved to Lookout Point, and then ran over a low pass into the Columbia Hills Inner Basin,

As these hills were neared, Spirit stopped to trench the outer slope sediments and found

suggestive proof of water action. The trench showed varying amounts of Mg and SO_4 (expressed as sulphur), with signatures suggesting possible Kieserite, a magnesium sulphate mineral present in abundance at the Opportunity site. Note this plot:

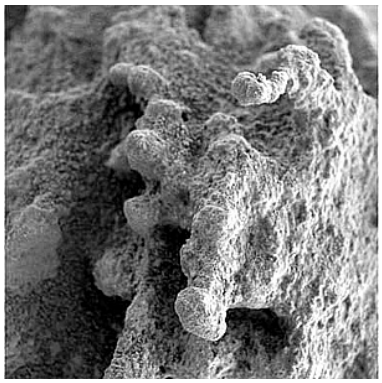


Spirit spent considerable time on the Columbia Hills, visiting a number of outcrops. This view shows inconspicuous, non-definitive evidence of layers at West Spur.

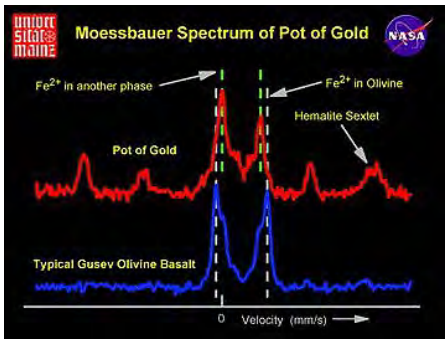
However, the loose float around this spot was indeed bedrock of a blocky nature. Some rocks were more slablike, possibly of a sedimentaary nature. Instrumental measurements will zero in on the question of their origin. Here are two scenes containing rocks.

">

Pot-of-Gold, the flat rock in the image above, may also be layered. Inspection with the Microscope Imager has revealed rather peculiar nodules, some at the end of tiny columns.



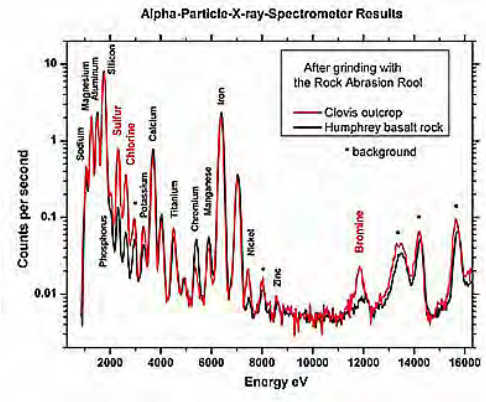
Mössbauer (spelled in label as Moessbauer to obviate use of 'o umlaut' in German) spectra for Pot-of-Gold shows the presence of Fe^{+2} peaks attributable to olivine and pyroxene, and four extra peaks representing Fe^{+3} which is best explained as hematite. The rock seems to be a basalt that has been altered to some extent to hematitic iron oxide.



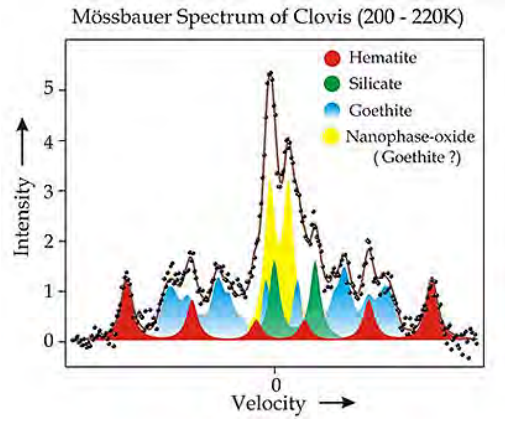
While on West Spur, Spirit got its first good look at thin dark layers outcropping midst the regolith, rock fragments and small boulders. This is likely to be older basalt flow rock brought up from the crater floor during the Gusev cratering event.

A lighter-toned outcrop, dubbed "Clovis" rock, was found on West Spur. It is currently being analyzed.

APXS plots for a basaltic rock (Humphrey) and the Clovis rock show one major difference: At least some of the Clovis rock contains Magnesium, Sulpher, and Bromine - indicators of either a sedimentary unit or alteration that includes Magnesium Sulphate (Epsom salts).



Results from a processing of Mossbauer data from Clovis has brought to light one of the most rewarding discoveries yet at the Meridiani site. The mineral Goethite, hydrogen iron oxide, $HFeO_2$, containing 10% water by weight has been identified, as shown here:

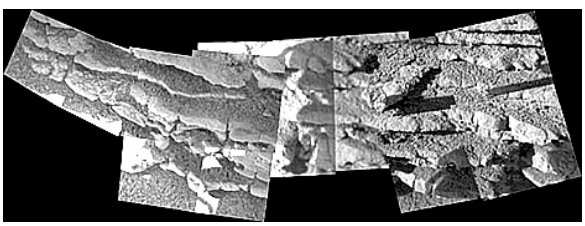


On Earth, Goethite - a very common mineral associate with Limonite - is found as an alteration product or as a direct precipitate in the so-called "bog iron" deposits, which result from a reducing, water-rich swampy environment. That form of Goethite is usually produced with the aid of bacteria but can also form inorganically. The mode of origin of the Clovis Goethite is still "up for grabs" but the presence of this mineral suggests a significant role for water in Mars' past.

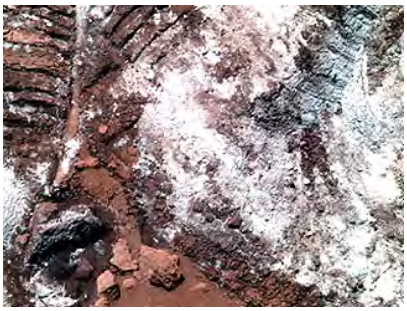
Of renewed interest in the possibilities of significant layered rock units is this view looking further from the West Spur position.

The first close-up view of layered rocks is the detached block named TetI, shown here:

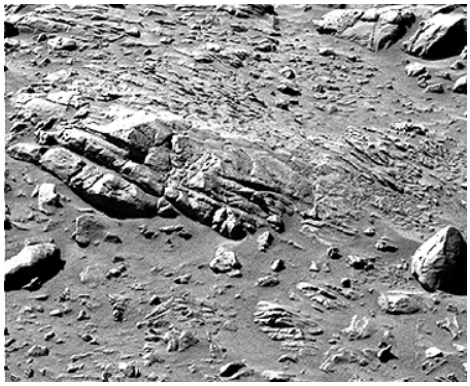
TetI comes from an outcrop area that consists of distinct layered units, seen in this composite image.



As Spirit drove off from the Columbia Hills it encountered in the soil and rock fragments a pronounced white encrustation that looked much like a crystalline coating. Chemical analysis show this to be mostly magnesium sulphate. Whether formed by surface water evaporation or by groundwater seepage containing dissolved salts, the discovery is the best evidence yet for the onetime activity of water at the Gusev crater site.

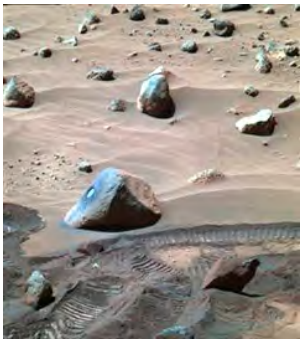


Despite "limping" along with two of its six wheels having problems, Spirit has visited more individual layering features - outcrops and loose rock. Uchben is an example where there is some evidence that the layering is caused by volcanic ash deposition. Microscope Imager views of surfaces close-up show angular flaky particles similar to those occurring in terrestrial ash deposits. Water has been found in its composition: the water may have been in the volcanic material derived from the interior, or may have been present in later altering hydrothermal solutions, or could have been incorporated from a standing body of water once at the site of the Columbia Hills (probably before their present uplift). So far, deposition by direct ashfall, by wind action, or in water all remain possibilities:

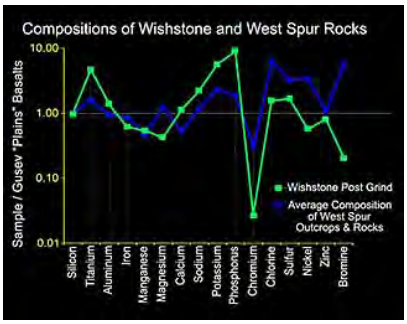


Because of wheel problems Spirit has been more cautiously guided. By disabling its front right wheel, and proceeding slowly, Spirit has covered much of the first ridge in the Columbia Hills, has moved over West Spur, and has moved into the flats:

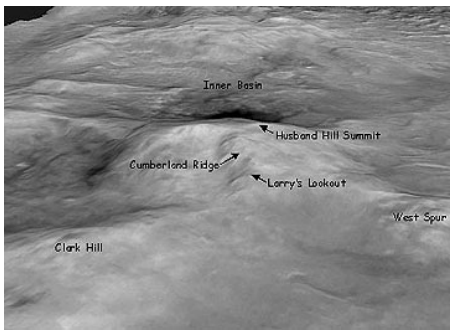
As Spirit traversed the West Spur, it encountered a rock nicknamed "Wishstone" whose appearance enticed the JPL team to decide to investigate it in some detail. The RAT was used to cut into its surface. The next two images show Wishstone (near center) after the abrasion hole was produced and a close-up of what was exposed:



The exposed interior shows that Wishstone has a granular texture. A few dark fragments are noted. The texture is consistent with either a volcanic tephra deposit or water-deposited fine particles; the interpretation remains ambiguous. The APXS instrument produced this plot of elements detected. The most significant feature is a higher than average (using West Spur results as reference) amount of phosphorus. This points to some mode of aqueous activity affecting the rock before it left its initial location (by impact ejection or by rolling over the West Spur slope).



Spirit has continued to move through the Columbia Hills visiting Rover team-named landmarks as seen in this perspective from orbit:

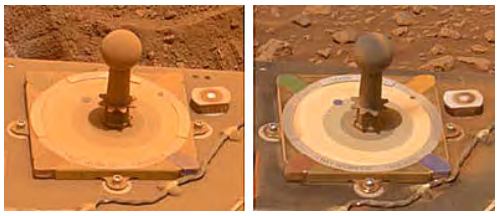


These hills seem to be underlain by layered rocks. Most evidence so far points to volcanic deposits, perhaps of several types such as flows and ash falls. At an outcrop named Methuselah, a rock called Keystone shows thin layers:

With one exception (shown on next page), trenches in and beyond the Columbia Hills did not disclose a light-colored bedrock unit comparable to that we shall see at the Opporutnity site. The layers seen in the Columbia Hills appear to be made up of grains of basalt loosened and transported into thin units as local deposits. The interpretation now favored has subsurface water leaching out chemicals from bedrock and soil, rising towards the surface, and depositing the sulphate salt near the surface as a cement for the granular units. This is the best evidence to date for the action of water at the Gusev Crater site; this water probably came from beneath the surface rather than from any standing surface water.

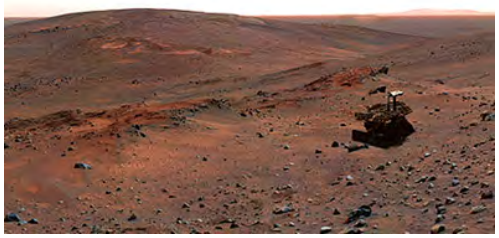
From time to time, both Spirit and Opportunity's cameras have captured the quick passage of atmospheric micro-cyclones or "dust devils". One was shown on page 19-11; here is another example:

In a remarkably serendipitous event, Spirit has been given a new rebirth that makes its already exceeded design lifetime even longer. Over the year plus of operation both Spirit and Opportunity have been gradually coated with martian dust. Their solar panels were becoming sufficiently plastered by dust to cut the efficiency that the panels should have for generating electricity from solar photon input. This cuts down operational time and threatens to fall to minima in which instruments cannot work. Then, sometime between March 5 and 15, 2005, the power output of Spirit dramatically spiked upwards. By looking at its surroundings, the JPL team concluded that a dust devil, the atmospheric whirlwind that occurs in large numbers of mini-tornados, must have swept over Spirit, sucking up (vacuuming) the dust coating, and exposing the solar plates to near optimum efficiency. TV images of a part of the spacecraft taken before and after this lucky encounter show conclusively that Spirit has been cleaned:

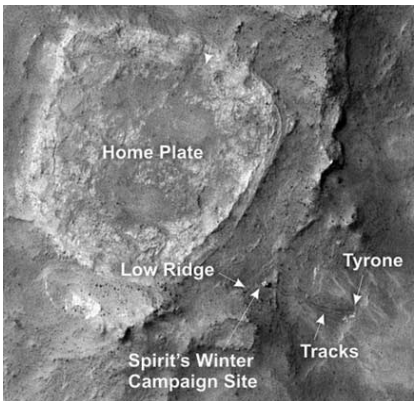


(The Opportunity team is hoping intensely for a similar quirk of fate to rejuvenate its power supply.)

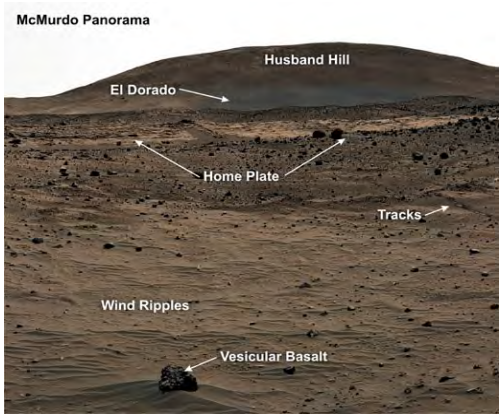
By late August, 2005 Spirit had climbed towards Husband Hill, a peak in the Columbia Hills (see map above), and spent several months gathering data. Here is a panoramic view made as Spirit (put in the scene by trick photography) in November started its descent from the top; in the scene are some layered rocks which will likely be visited in the future:



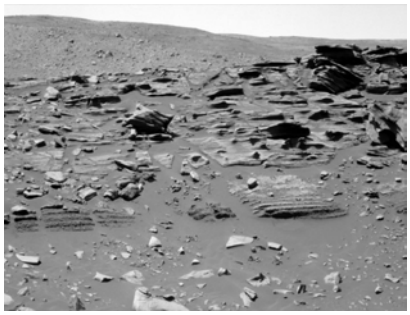
Spirit came to an area in which a low embankment contains outcroppings of slablike rocks. These are nicknamed the "Home Plate" site. Here is a Mars Reconnaissance Orbiter image of Home Plate (named from its shape) - the site consists of volcanic rocks which may get this platy structure from jointing or the individual plates are parts of a succession of thin deposits:



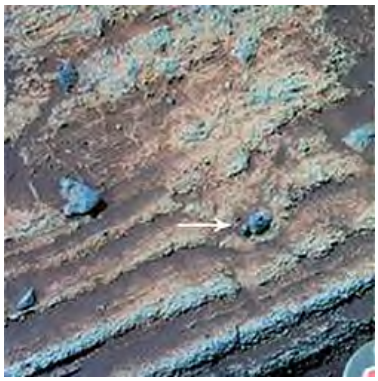
This view shows how the Home Plate rocks are interspersed with darker rocks:

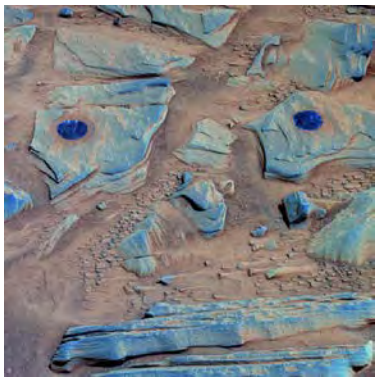


From the ground, the Home Plate layered rocks looks like this:



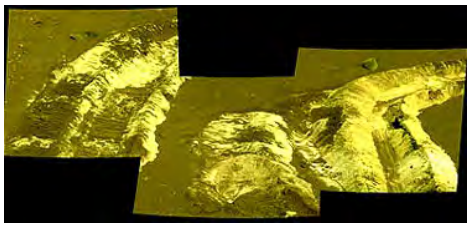
Here are two close-ups of several of the thin layers at Home Plate. These have been interpreted as possible volcanic ash deposits (pyroclastics).





In May, 2006 Spirit, with one wheel broken, was placed in a "hibernation" mode, with few activities that require power and with its solar panels oriented towards the Sun while Mars passes through an 8-months winter. The vehicle was parked facing northward on a gentle slope named Low Ridge. Spirit did resume exploration in 2007. While spending the martian winter at one place, when conditions are favorable Spirit has gathered data during this period. For example: Here is a close-up of the immediate surface on which Spirit sat, showing several light yellowish-brown exposed patches of altered bedrock determined to be rich in $CaSO_4$ with perhaps some $FeSO_4$.

In early 2007, during renewed exploration Spirit's wheels exposed a thin continuous layer of a yellow deposit rich in sulphates:



In May of 2007 an intriguing discovery was made when a light-colored layer was uncovered during Spirit's traverse; it came to light serendipitously because a stuck wheel on the Rover dug deeper than normal as it spinned in place:

A mini-TES analysis of the layer revealed it to be nearly 90% silica. This was not quartz sand but a fine crystalline mass or perhaps amorphous. Its specific nature has not yet been determined but speculation centers on alteration of the material by acidic vapors or fluids, leaving the silica as a residue. The silica deposit is similar to that found in thermal hot springs associated with volcanism on Earth, and as such is an indicator of aqueous activity. This next image, with a different color shading, shows the white material to better advantage:

As martian spring began in 2009, Spirit has been reactivated (its was parked in 2008 during the martian winter), and has backed away from Home Plate to explore elsewhere. It has started towards the Columbia Hills but as of August 2009 has been stuck for a while in loose sand. Efforts to get it unstuck have so far failed. Although efforts continue to pry it free, the consensus in 2010 is that Spirit is now permanently trapped in its present location.

Thus, the history of exploring the Gusev crater site and beyond was at first suggestive of the absence of any strong evidence for water activity but once the Columbia Hills were reached, possible evidence was acquired in several settings. However, the water, if verified, seems associated with volcanic activity rather than any marine or lacustrine deposition. The results from the MER Opportunity rover, which discovered minerals almost certainly of hydrous origin and has found indicators of deposition from discrete water bodies, are covered on the next page (19-13b).

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

In the past ten years renewed interest in Mars exploration, in antipation (or hope) that it will become the next big planetary goal for NASA and the worldwide space community, has shed much new light on the Red Planet. Many images of distinct layering, attributable to (a likely) deposition from fluid, have now been acquired. The possibilities of significant amounts of water in the martian past, and some still extant even now, excite planetologists because H_2O is a necessity for life. Some evidence of microscopic life forms in meteorites known to come from Mars is claimed but this still engenders strong debate. Further exploration of Mars has become almost a crusade for the NASA milieu and other martian devotees. This is underway with the successful Mars Global Surveyor and Mars Pathfinder programs, reviewed here.

Life on Mars?

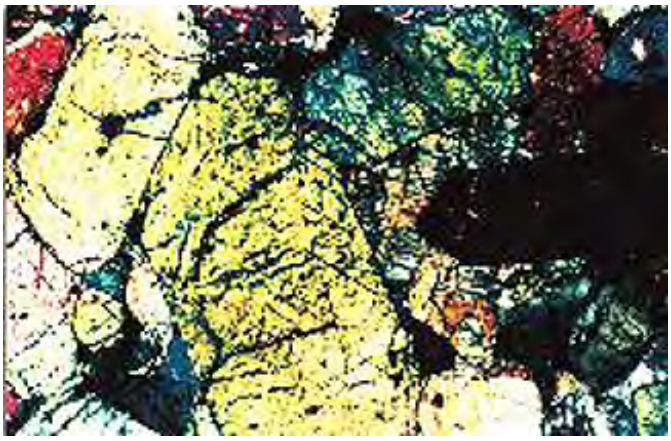
Starting with Percival Lowell's popularization of "Canali on Mars", that planet has continued to fascinate both scientists and the general public. Readers unfamiliar with what is known about ancient life on Earth may want to work through the paragraphs on that subject in the middle of page 20-11 to familiarize with characteristic of primitive life forms.

The importance of those landforms with fluvial signs looked at on the previous page is that they suggest a probability that water did (does?) exist in sufficient volume and concentration on Mars as an essential ingredient for the inception of organic molecules. Carbon, the other vital constituent, was certainly present as indicated by today's atmosphere. Whether these elements came from the interior or from meteoritic matter, during the accretion phase or later, is still uncertain.

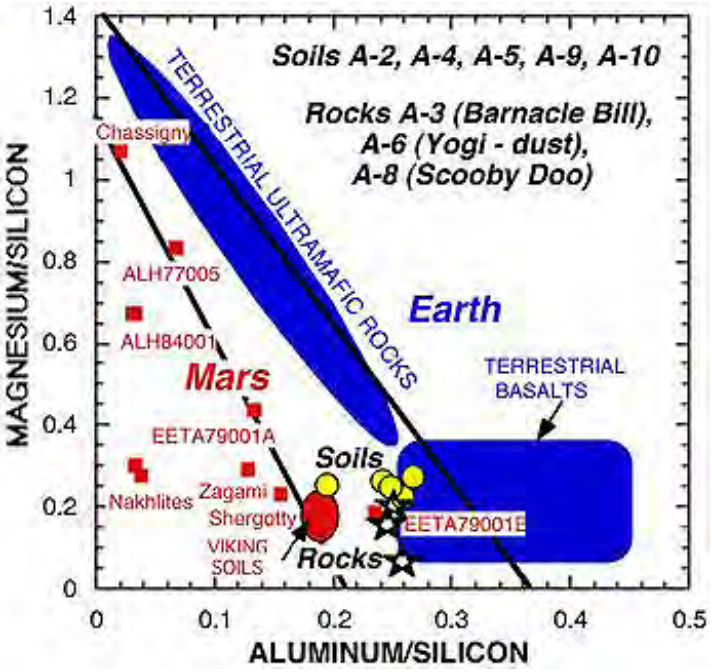
The absence of any viable organics in samples analyzed at the Viking sites does not disprove the possibility of biogenic forms in modern Mars. The sampling may not have encountered organics at these two isolated sites but they dwell elsewhere. Both sites were in younger terrains, so the nonoccurrence could mean that any earlier life or unorganized organic matter had perished by then. However, one of the Viking investigators, Dr. Gilbert Levin, has always contended that there is indirect evidence of organic matter extractable from the measurements made onsite and transmitted to Earth.

Before (or after) reading through this subsection, you may wish to look over these three Web sites:(1), (2), and (3).

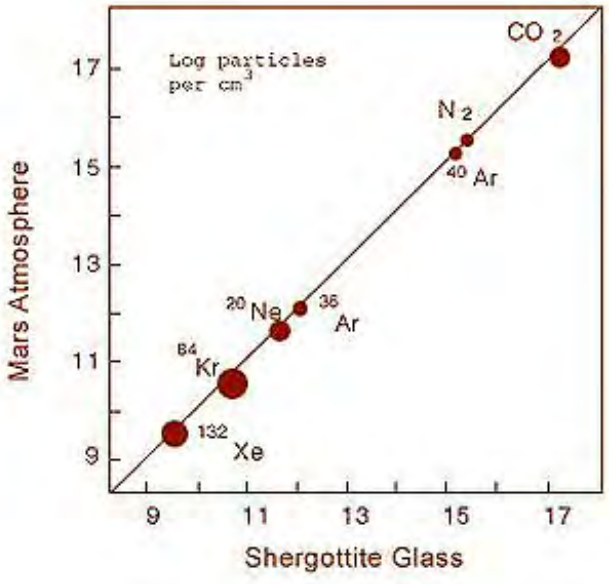
For the time being, an alternative approach would be to find meteorites on Earth that can contain strong proof that they came from Mars, being ejected through impacts into Earth-crossing orbits until a few pass onto the Earth. Such meteorites have certainly now been found. These form a group known as the SNC meteorites (consisting of the Shergottite, Nakhlite, and Chassignite types). A good summary of the SNC meteorites is found at this Internet site. That they come from Mars is supported by their compositional similarities to rocks examined during the Viking and the Mars Exploration Rover missions (see below). They compositionally are similar to dunites, lherzolites, clinopyroxenites, and basalts. Some show several shock features noted in basic igneous minerals, indicating that these were close to the impact points on Mars; others without shock effects represent samples derived further from these points. At least 12 such meteorites had been found prior to the 1980s and more recent searches place the number of probably martian meteorites now being studied at more than 40. This is a thin section that shows the characteristic minerals in a Chassignite:



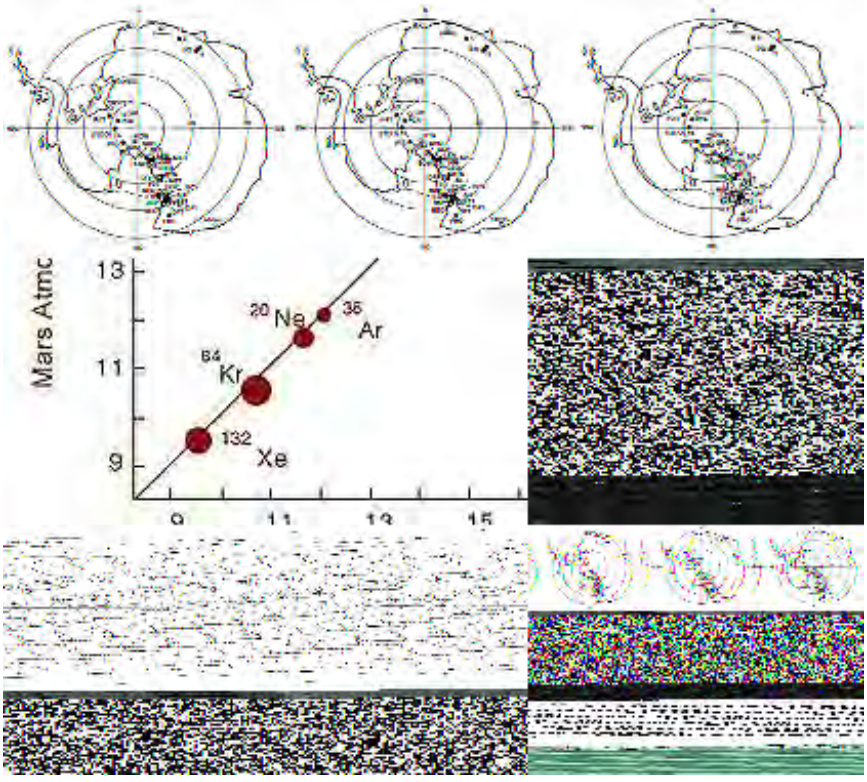
Several lines of evidence support the martian origin postulate: One is the similarity of composition between the SNC group and inferred rock types at the two landing sites on Mars described below. Examine this plot, which shows several SNC meteorites (in orange letters) falling within a compositional field notably different from the two common terrestrial basic igneous rocks:



Other evidence includes isotope compositions that do not fit terrestrial rocks or Asteroid Belt meteorite rock types. A very strong proof is the almost identical composition of gases found in a Shergotty meteorite with gases sampled on Mars, as shown in this plot:

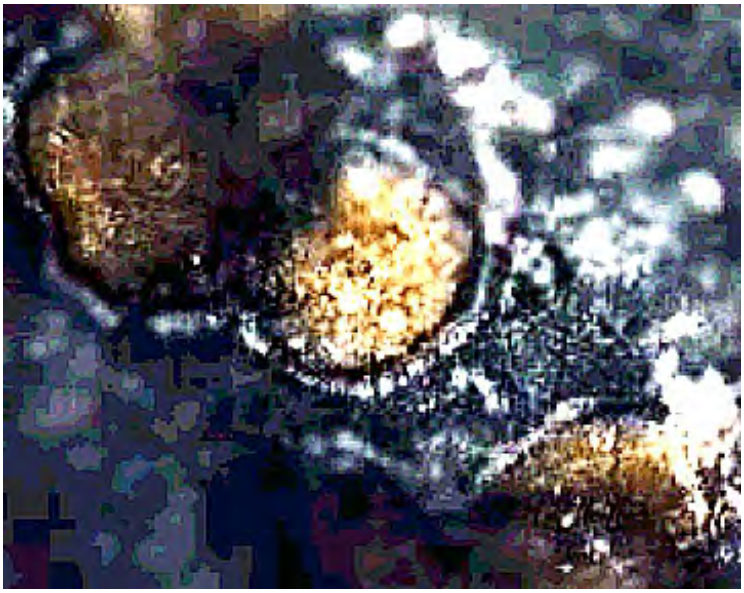


As mentioned elsewhere in this Tutorial, the number of meteorites of all types and sources being found annually went up very significantly when by the late 1970s Dr. William Cassidy of the University of Pittsburgh, and then scientists from other institutions, determined that these "stones from heaven" could remain for long periods of time on the ice surface of the Antarctic. Many such rocks have since been collected. Included in these are some 35 SNC types, found at the locations shown in this map. The majority of these martian meteorites date as younger than 1.3 billion years, implying that martian volcanism has continued at least til then.

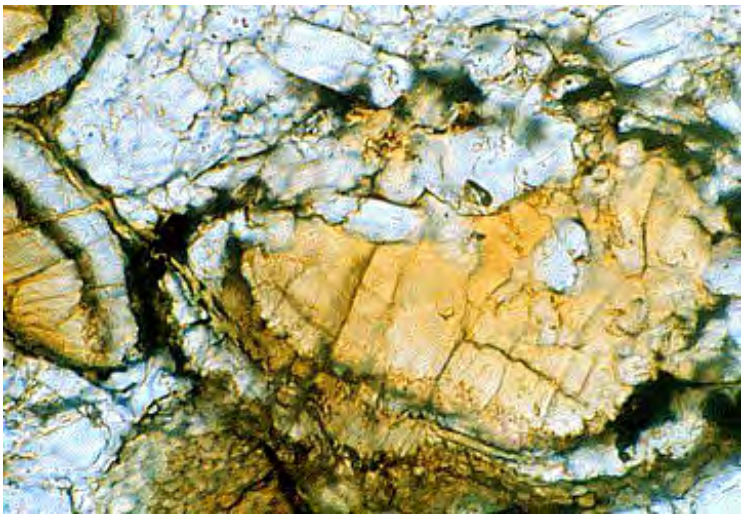


A short-lived sensation in the scientific world during 1996 was the claim that a meteorite from Mars, found in the Antarctic, contained evidence of life. It is known as the Allen Hills meteorite (ALH84001), which is characterized by containing orthopyroxene and has several features interpreted as consistent with organic structures found in some ancient terrestrial rocks. So far, only this meteorite contains enticing (but not conclusive) signs of organisms but several others contain tantallizing objects of possible biogenic origin. Below is this Allen Hills meteorite (one other has also been collected there) first as it appears before being sawed into (the black coating is glass [fusion crust) and then after the rock was sawed in half, exposing its inner materials.

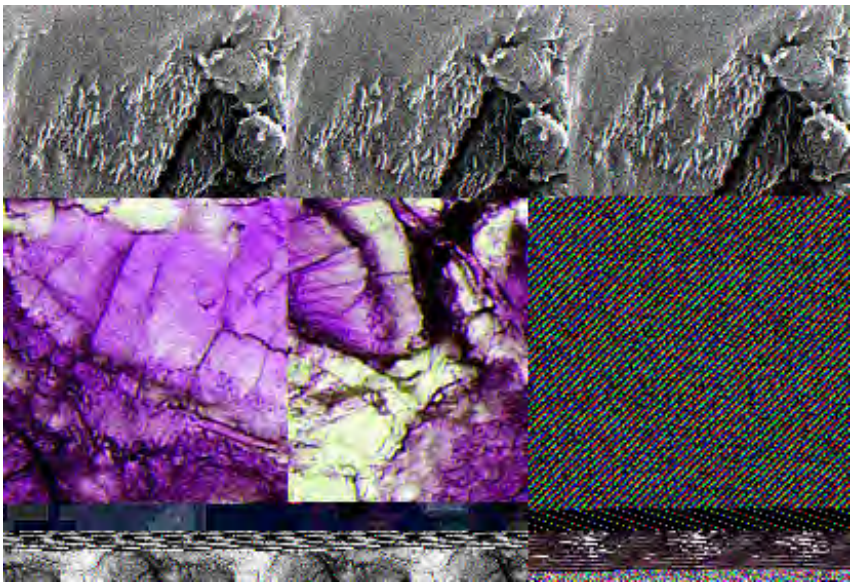
This fragmental rock contains dark materials, thought to be old martian crust, that date from 4.5 billion years ago. Golden-orange carbonate globules, dated at about 3.6 b.y., are considered evidence of a primitive ocean or in another interpretation weathering of surficial rock through reaction with the CO_2 in the atmosphere. Here is an electron micrscope image of two Mg-Fe carbonate balls found in this meteorite:

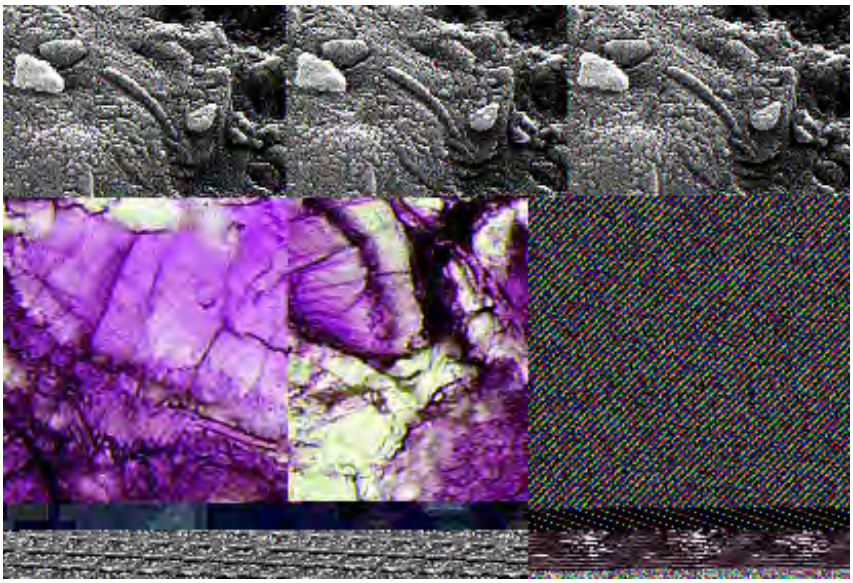


A thin section of ALH84001 also shows the orange (iron-enriched) carbonates and orthopyroxes (clear):



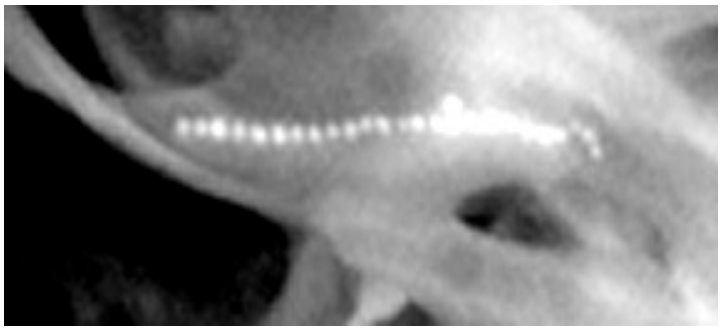
Fossil-like bodies consisting of the iron oxide magnetite and iron sulphides are present in the globules, as seen at high magnification under the electron microscope. Thus, the size of the elongate tubes falls between 1/100th and 1/1000th of a millimeter. If these are truly fossils, they are about a tenth the size of nanofossils found on Earth.





Keep in mind the subroundish features seen in these images. Similar structures are seen in rocks at the Meridiani site by the Mars Rover Opportunity. They are all likely to be inorganic in origin but some that approach spheres in shape are suggestive or tiny microfossils.

Even more lifelike in appearance is the elongated chain of tapering magnetite crystals seen below, an arrangement known to be produced on Earth solely by certain bacteria. Some scientists hail this as convincing proof of martian life, formed very early in its history. David McKay and his cohorts at JSC have studied magnetite organisms on Earth and cite 6 criteria that associated with an organic mode of origin. They claim that all six are involved with the Allen Hills magnetite bodies.



Associated with the globules are small amounts of polycyclic aromatic hydrocarbons (PAHs) that, while not necessarily biogenic, are interpreted by some as indigenous to Mars rather than contamination after Earth-arrival.

Still, it is hard to consider this rock itself as the primary (initial) host of tiny life forms that originated within it during its formation. The rock is a igneous cumulate - enriched in pyroxene by crystal settling during magma differentiation - and thus is a totally alien matrix for life to develop within. The carbonate mineral is likely secondary in origin, formed perhaps by subsurface weathering as water percolated downward. If these strange features are true life forms, they almost certainly were introduced somehow from an external source - perhaps when a sea, lake, or other water body covered the rock before it was heaved off Mars during an impact.

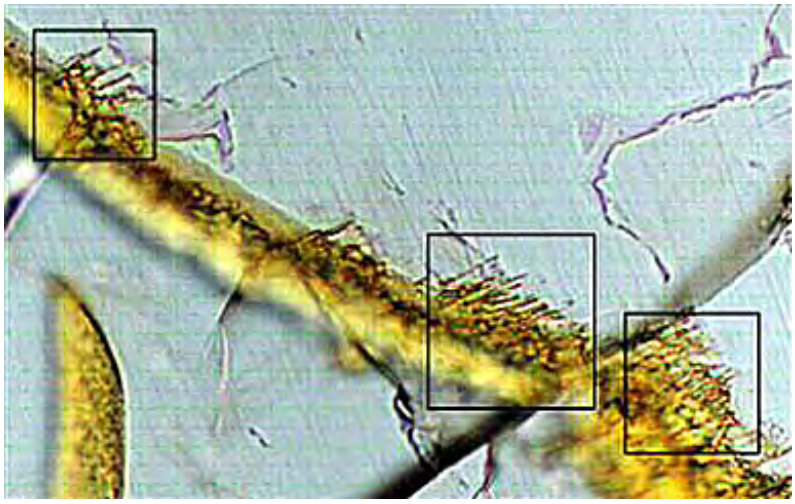
The great value of ALH84001 is that it contains provocative features that offer an incentive to look at other Mars rocks, either as meteorites or by trips to Mars, for similar features. Whether it really contains evidence of once-living biogenic matter, it has served as a major inducement to expand the exploration of Mars.

Another Antarctic martian meteorite, EETA79001, contains carbonate (probably an alteration product but possibly evidence of martian sediments) and organic matter.

Needless to say, many skeptics have argued that this evidence is not persuasive; some similar features found in Earth rocks have been shown to be inorganic. A recent report now claims that these tiny forms are indeed primitive bacteria but very similar to types still thriving on Earth; the implication is that this is just terrestrial contamination. Another report claims that the researcher

has found very similar features in terrestrial rocks that fail to show any signs of active life during their formation.

A research group at Oregon State University, led by Prof. Martin Fisk, has presented evidence of microscopic tubes or tunnels in the Nakhla meteorite that are very much like similar features in terrestrial rocks that also contained bacteria. These features have been generally accepted as produced by rock-feeding bacteria. The Nakhla meteorite is a volcanic rock - one of the defining types in the SNC group. It fell to Earth in Egypt in 1911, and was quickly collected. Here are several sets of tunnels, all emanating from a fracture.



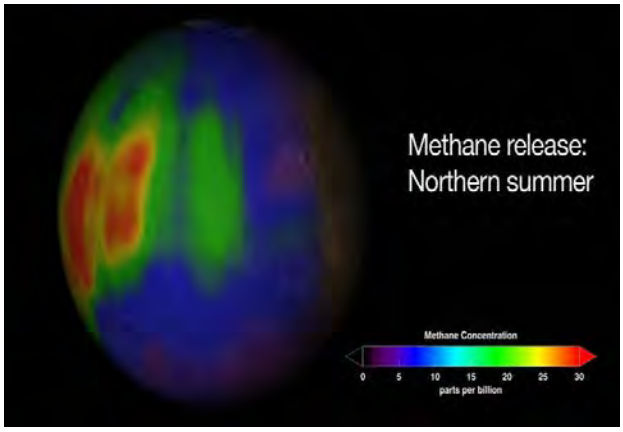
David McKay of JSC has reported finding Iddingsite, a clay mineral derived from olivine, in the Nakhla meteorite. Associated with the mineral is a dark brown to black substance rich in carbon. Although the meaning of these Nakhla discoveries is being hotly debated, their presence encourages the Mars life devotees to redouble their efforts for convincing proof.

However, the possibility that they are genuine persists: if scientists eventually certify life on Mars, the Earth would no longer retain its unique status as the living center of the Universe; although it is a huge leap from microscopic primitive organisms to the intelligence that then understands them. Suffice to comment that some scientists are touting these presumptive meteorite "life forms" as the most pressing reason to formulate and accelerate a major space effort to return to Mars for more detailed exploration.

19-45: What would be the best proof of life on Mars? ANSWER

As the search for any hints of life on Mars continues with the 2004-6 exploration by the MER Rovers, described on the next two pages, attention back on Earth is being fixed around environments where levels of life (see page 20-12 for an overview of this topic as applies to planets in general) may be controlled by environments that may have counterparts on Mars. Extremophile environments like oceanic "black smokers" and the Antarctic both support life. The only place on Earth where almost no life whatsoever (including microbial) has been discovered is the central region of the Atacama Desert in northern Chile. This place will likely become a testing ground for life-searching robotic equipment and life form-identifying software (NASA ASTEP program) that will apply techniques for pattern recognition of morphological objects in the martian rocks of a suspicious nature as conceivable evidence of organisms. Here is a view of hills in the Atacama which, except for the dark skys, certainly reminds one of some Mars scenes:

In January of 2009, NASA confirmed and augmented evidence of methane (CH_4) in the atmosphere over several martian regions. This gas was detected by the NASA Infrared Telescope on Mauna Kea in Hawaii. Methane could indicate organisms that provide it as a by-product of their metabolic action. But methane is also released by volcanoes. Thus, this result for the moment is ambiguous. Here is the methane map that shows the distribution of this gas:

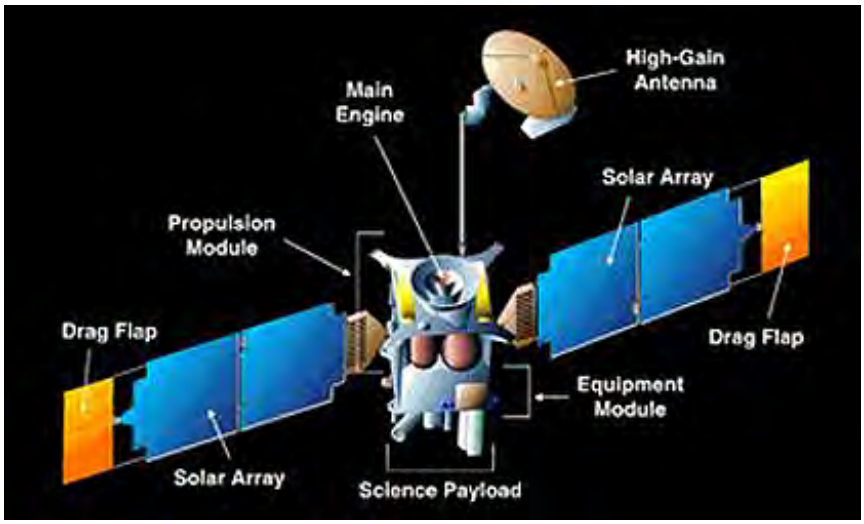


As we shall see in the next pages on Mars, evidence is mounting that water existed and still exists on Mars. Thus one of the essential needs for life seems to have been present. But a study by CalTech/MIT professors of the Argon in the above meteorites indicates that the high retention rate of this radiogenic end product implies low temperatures at the surface of Mars, persisting for most of its lifetime. This would preclude any long term buildup of large bodies of water, in streams, lakes, or shallow seas. Another argument against the presumption that life may have once existed on Mars is the discovery of parts of Mars containing layers of evaporites that formed from very acid waters - some scientists contend this acidity is too extreme for life to have even staarted. But parts of the martian surface (specifically, around the north polar region) have alkaline soils (page 20-13b). For now, the question remains open, as the more recent exploration of Mars has revealed conditions that could indicate standing and flowing water at some time in the past.

Resumption of Martian Exploration

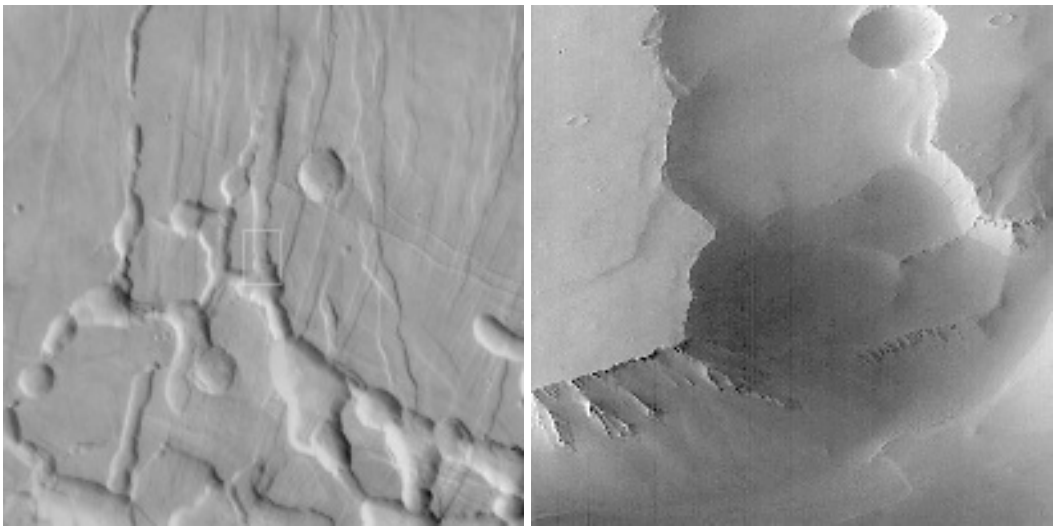
After a 20-year hiatus in Mars exploration, the quest for onsite information about our red neighbor has resumed. Several of these are described at JPL's Missions sites (Current comes up, check also Past and Future). First up was a launch in 1993, of the Mars Observer, a billion-dollar spacecraft, which, regrettably, failed enroute. The Mars Global Surveyor (MGS) followed on November 7, 1996, with its operation directed by JPL. The spacecraft arrived for orbital insertion on September 12, 1997. In December of 1996 the Pathfinder mission was underway, to land a roving vehicle capable of viewing the local terrain and making sophisticated measurements. An orbiting spacecraft, whose task is mainly climatic measurements, was launched in mid-December of 1998.

A brief synopsis of the Mars Global Surveyor mission is given at JPL's movie site. Access through the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes "Mars Global Surveyor Across the Centuries", April 17, 2003 (note: this is an hour-plus long lecture). To start it, once found, click on the blue RealVideo link. If you don't visit this video site but are curious as to what the MGS looks like, here is an artist's sketch of the deployed spacecraft:

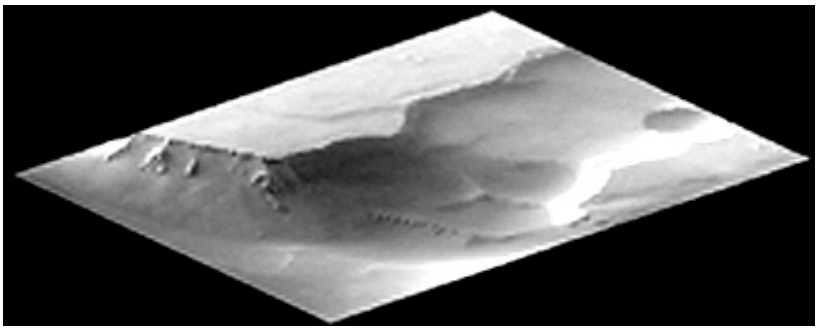


From an initial orbital altitude of 400 km (249 mi), the Global Surveyor used the aerobraking technique, which involved progressive slowing through air friction (drag) with the thin atmosphere, until, after six months, the spacecraft lowered to 110 km (68 mi). From there, using thrusters to maintain this altitude, a mapping mission began in late-March of 1998 to last a minimum of 687 Earth days. A second lowering was executed in 1999.

The Mars Global Surveyor has a wide- and narrow-angle camera system (MOC or Mars Orbiting Camera; see next page), a Thermal Emission Spectrometer (TES), a Laser Altimeter (MOLA), a Magnetometer/Electron Reflectometer, and two other instruments. Among the first surface images, taken from the higher altitude in early October of 1997, are these two examples of the many views from the mission. The left image (175 km; 108 miles) is of a section of Labyrinthus Noctus, a maze of interlocking grabens. The right image is 12 by 12 km; it shows a canyon wall in an unidentified area of Mars.



This next image is not a computer-generated perspective of the bottom (or right) image above, but was taken from MGS when its MOC was tilted 25° towards the horizon so that the view of the canyon walls is an actual scene with a 12 meter resolution, even though the target view is 1600 km (1000 miles) away.

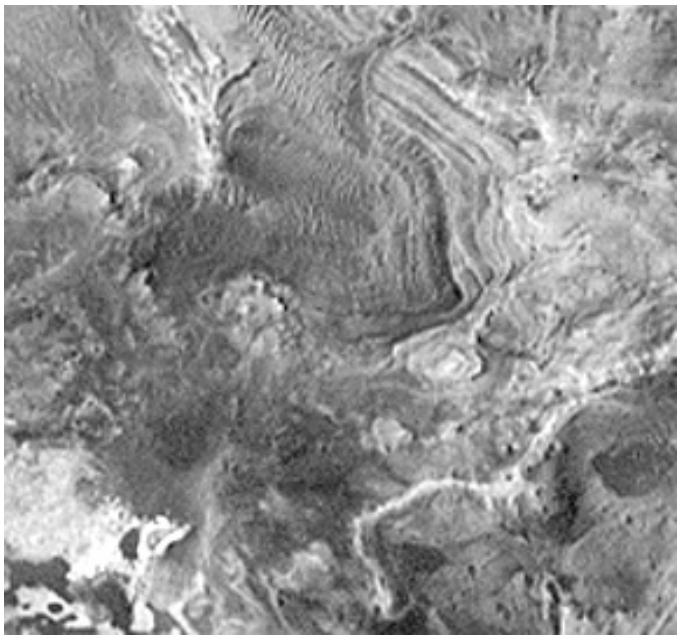


The Mars Orbiter Camera (MOC) can take pictures with resolutions as good as 3 m (9.8 ft). This resolution will improve slightly over time, as the orbital height lessens, because of the aerobraking effect. A very large number of these images can now be viewed online at the Malin Space Sciences System site - this is certainly worth visiting now that they are putting up a new image each day. A number of the images shown on this and the two succeeding pages are credited to NASA/JPL/Malin Space Sciences Systems (referenced by MSSS).

The next view (supplied at the MSSS web site), taken after the orbit was lowered to the altitude chosen for the first major scientific measurements, shows details of the terrain surrounding an elongated volcanic caldera, 2 km in length in the Tempe-Marcotis Fossae region. Almost touching it is a round crater that may be impact in origin. The surface surrounding these two features is marked by small dunelike features probably built up by wind action

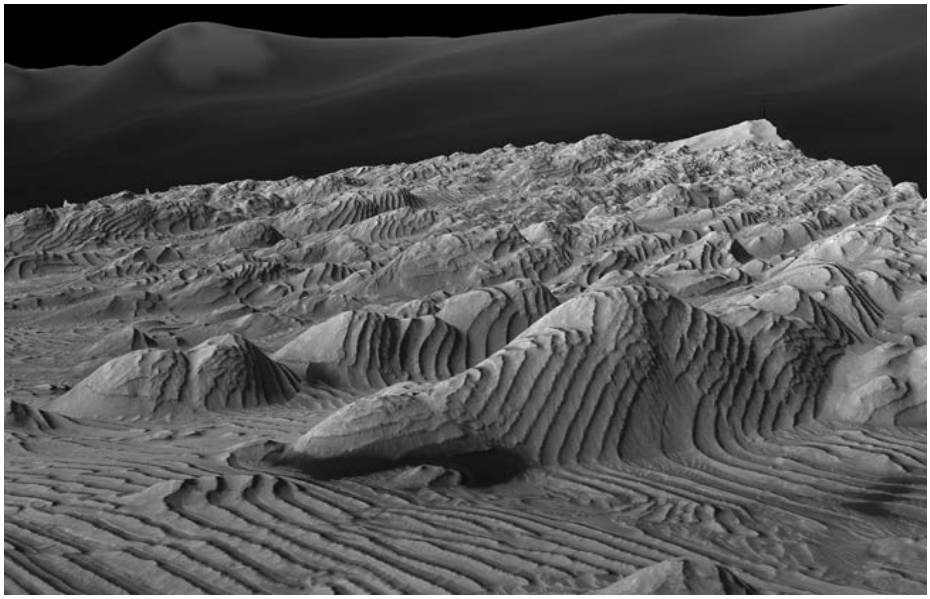
For life to have ever existed on Mars, the best environment to expect evidence or trace of primitive life forms would be sedimentary rocks, primarily in layers, which would suggest some type of water action (although layering from volcanism or from wind deposition of loose surficial grains could also produce depositional layers). MGS's MOC has been especially valuable in imaging layering of the bedrock of Mars (presumably either lake sediments or dust beds or volcanic flows) that has been found at many martian locations.

Here is an MOC view at 7 meters resolution of the wall of Candor Chasma in Valles Marineris that shows distinct layers.

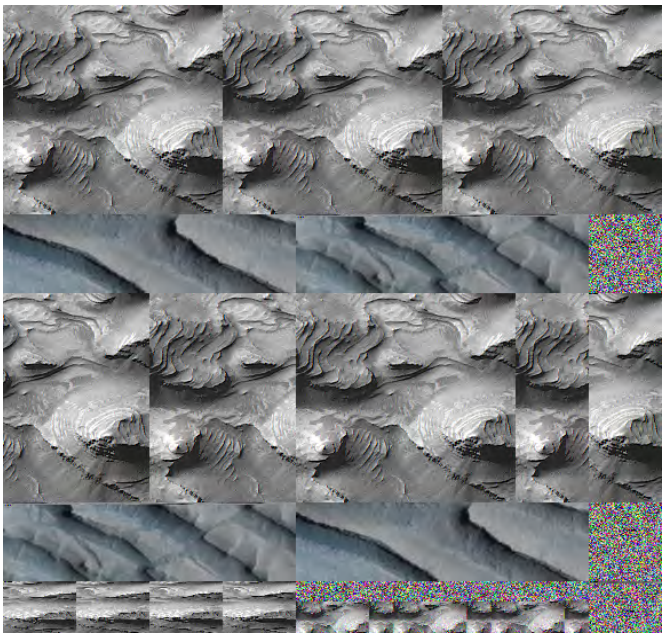


19-46: What can you say about the layering seen along the left facing cliff in the right center of the image? ANSWER

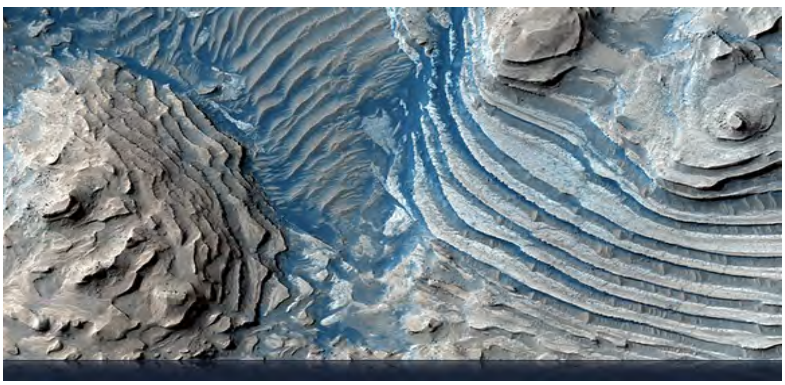
The next pair of images present solid visual evidence for what appears to be significant stratification of bedrock units. The scene just below is so striking - almost artificial - that one might believe it to be a hoax. But cross-checking on the Internet verified its authenticity; it was imaged by the HiRISE camera on the Mars Reconnaissance Orbiter. The hills shown with inclined layers are in the Arabia Terra region of Mars. A close-up of the layers appears below the perspective view:



Many Mars investigators, including Dr. Michael Malin, assert views such as above as unequivocal indications for depositional layering from an aqueous medium such as water. Examine this view (MSSS) of another part of Candor Chasma:

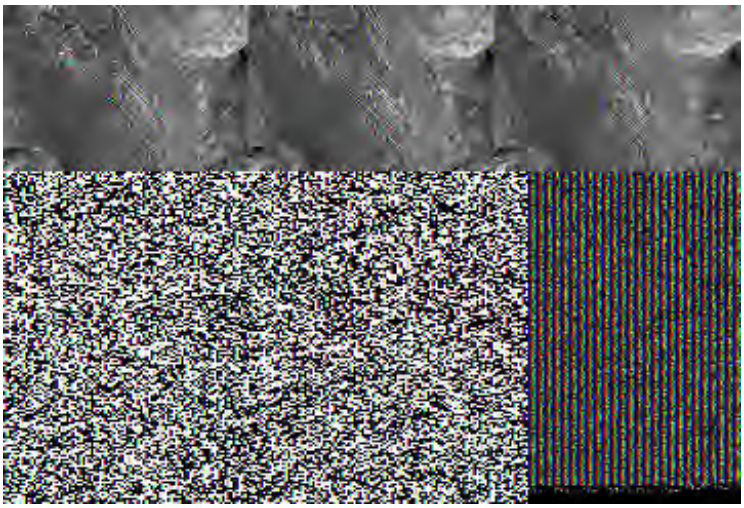


The layers can be expressed in offset ledges, referred to as "stairs-and-mounts", as exemplified here:

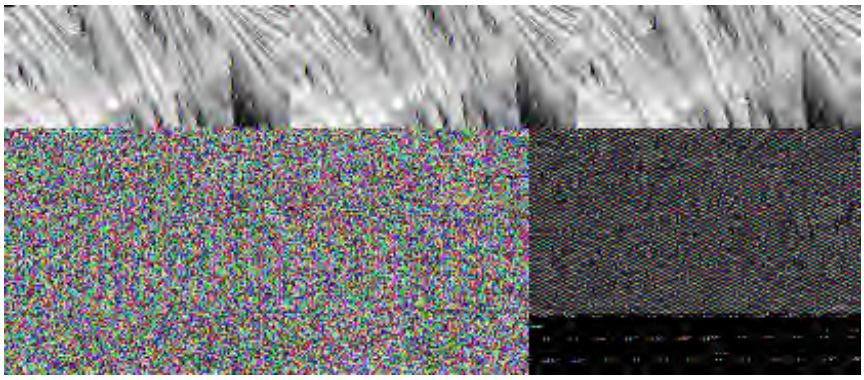


The Valles Marineris canyon and trough system is very well suited to the recognition and characterization of thick sequences of martian layering because, like the Grand Canyon of Arizona, it has cut so deeply into the upper martian crust. Thick layers are especially well-exposed in the tributary canyons of Valles Marineris:

Layering can show cross-bedding even when viewed from above, as evident in this example from a Mars Reconnaissance Orbiter (MRO) image of a Valles Marineris tributary valley:



High resolution MRO images are providing new details about martian layering:

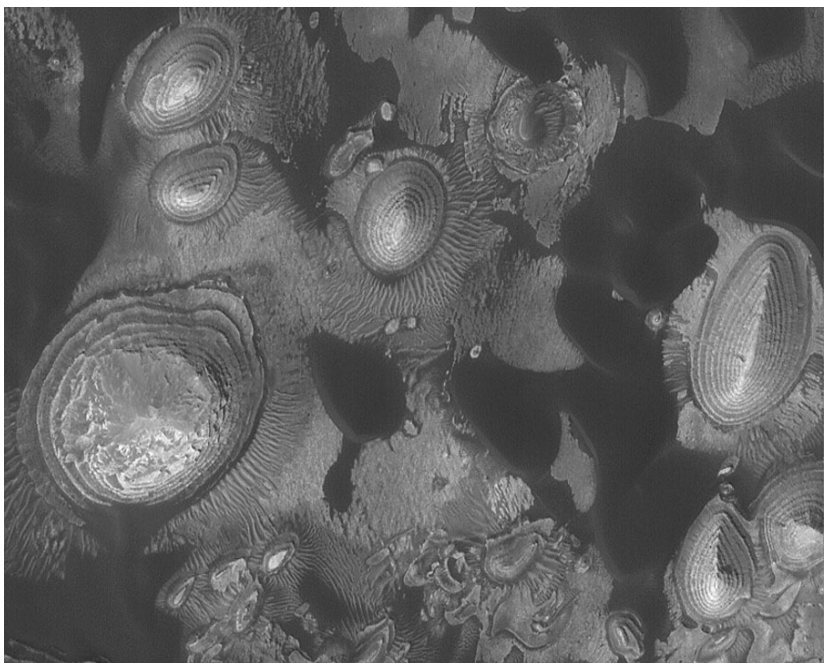


Here is another example from MGS MOC imagery, showing both (presumed sedimentary) layering and linear features that are likely dune deposits, here in the walls of Melas Chasma, a tributary canyon to Valles Marineris:

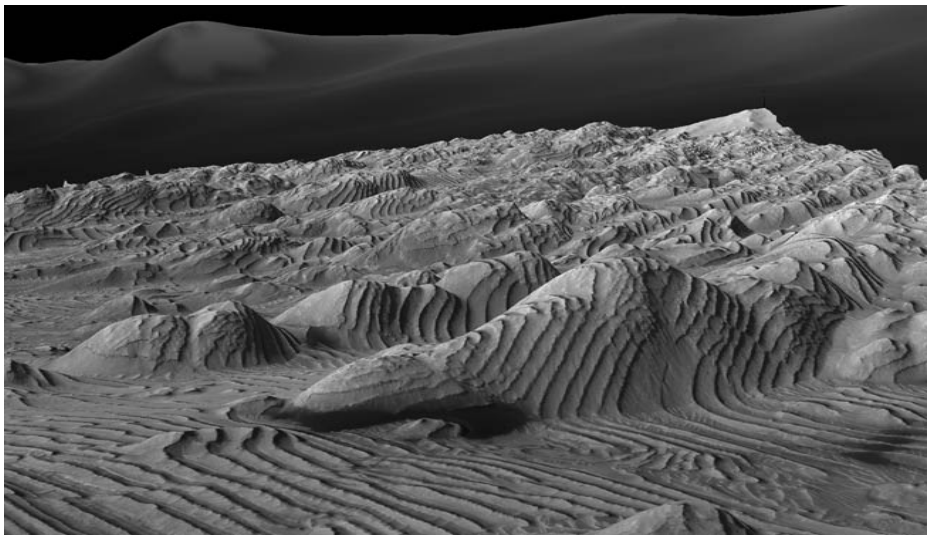
Note that the so-called layers are arranged stepwise in terrace or benchlike patterns in part of the image. Such offset "layering", if this it be, is sometimes found associated with mesas in the horizontal rocks of the Colorado Plateau. In the butte-like prominence at the center right of the image, the layers are well exposed along the steep sides. Two more MOC examples further support the presumption that this layering is sedimentary in nature; the precise mode of deposition is still being debated, with two favored choices being either lake beds or volcanic ash deposits from cyclic eruptions. They show layering in the wall and the floor of Becquerel Crater in West Terra Arabia:

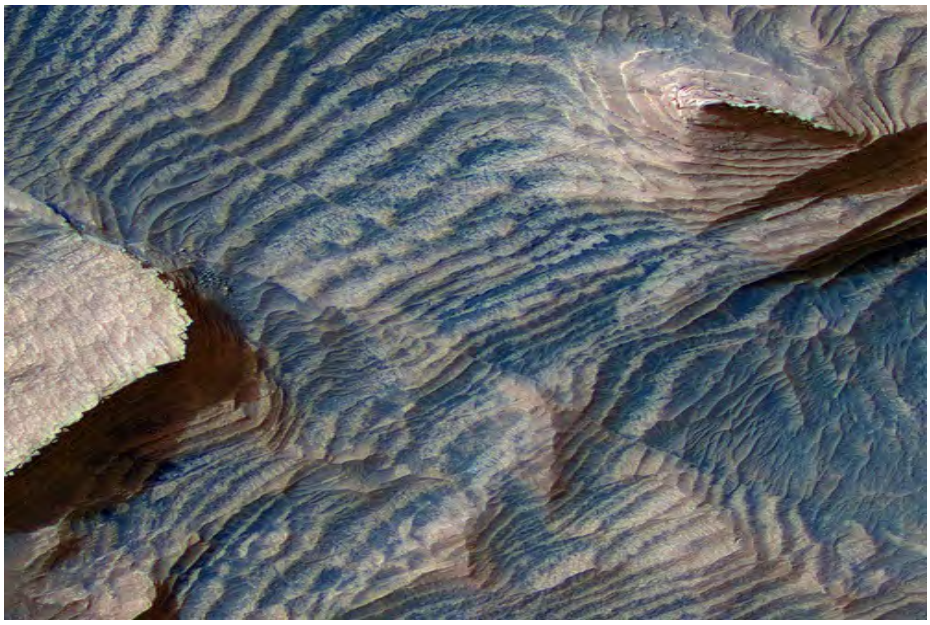
Units that are almost certainly layers have been exposed in craters located in the West Arabia Terra as seen in these MOC images. The top image shows layers that appear to have filled an earlier-formed crater. The layers have since been partially sculpted out as an erosional depression has formed.

Another impressive image of layering in Arabia Terra is seen next, in which the exposure is evident in the slopes of isolated hills:



These next two images provide additional examples of layering in Arabia Terra.

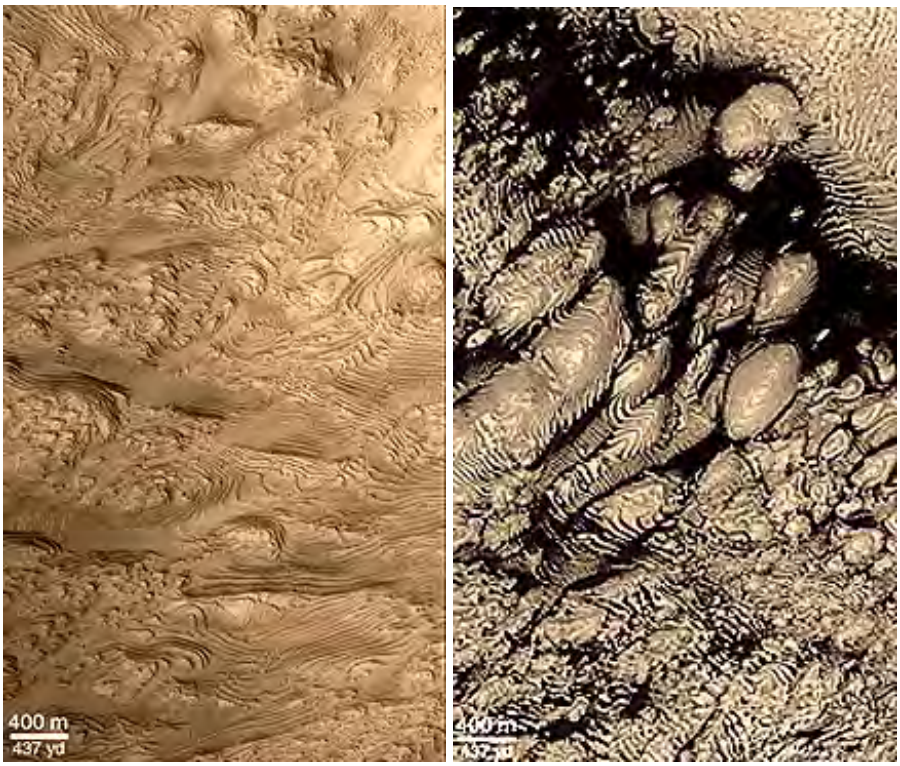




In rare instances, the layers can show visible offsets, indicating faulting occurs within them. This image of layers in an impact crater in Arabia Terra reveals subtle but detectable fault lines:

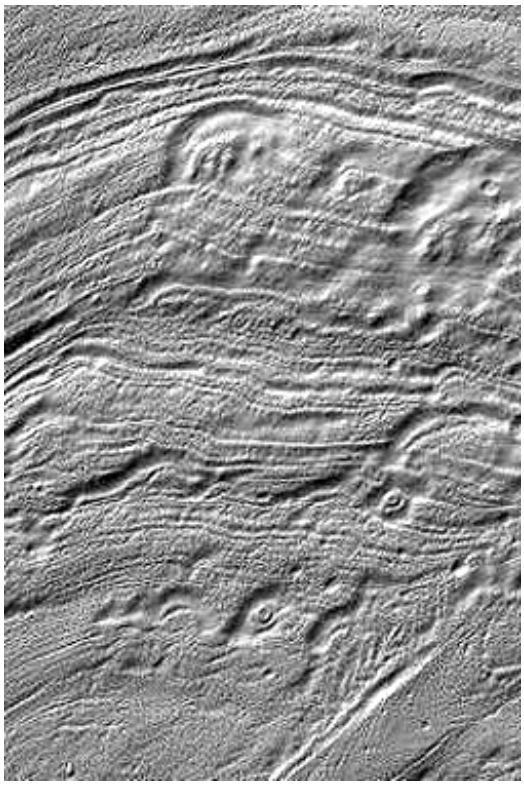
In the color images below showing part of a larger crater in West Arabia Terra, the layers appear inclined or dipping; but in an image from the crater center they look more horizontal, as expected if an impact crater hits on a sequence of non-inclined beds, with the dips then caused by the cratering action (on Earth craters [e.g., Meteor Crater in Arizona] in horizontal bedrock will show strongly inclined and even overturned layers outward towards the rim). The left image is near the rim top. The right image is further down the crater walls and contains black deposits which could

be shock-melted rock or black basalt sand brought in by the wind.



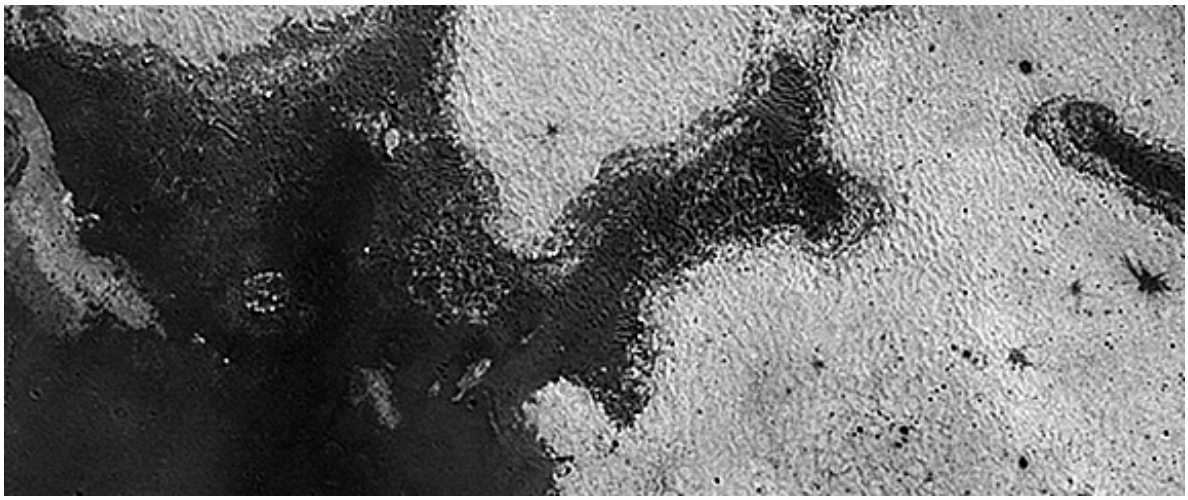
Eroded terrain, with horizontal beds, very similar to the WTA crater has been imaged by MOC in the floor of West Candor Chasma, as seen in this black and white image:

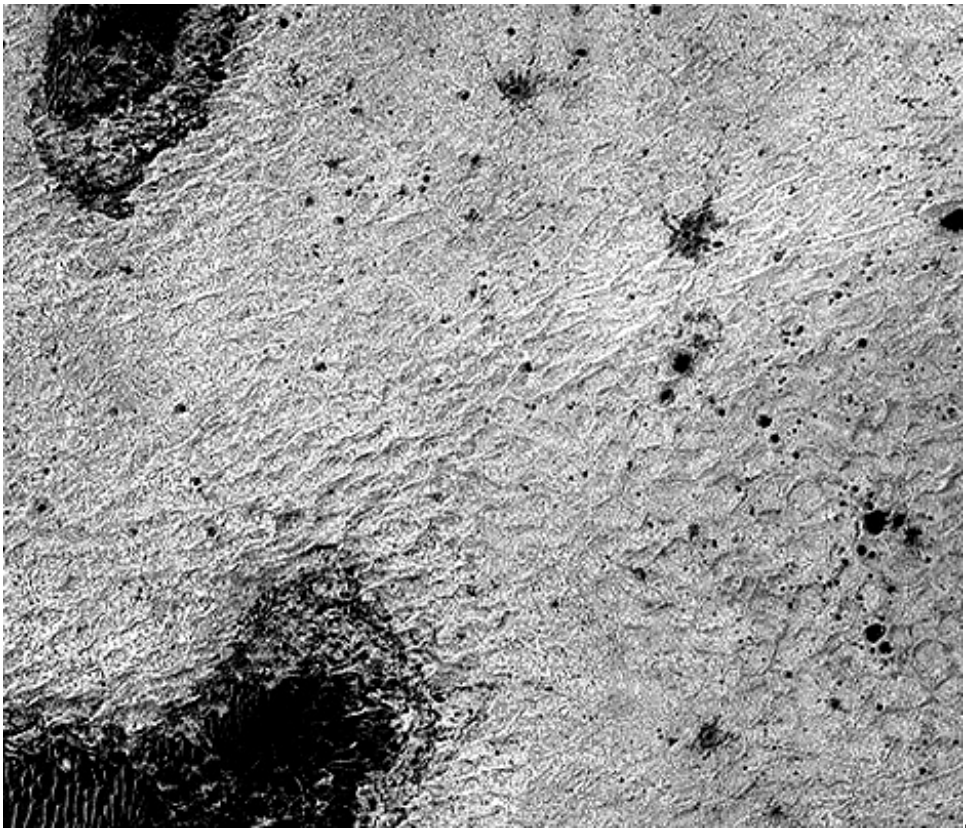
One of the largest features on Mars is the Hellas Basin. It too contains areas where prominent layering is observed, as at Aeull Valley, shown here:



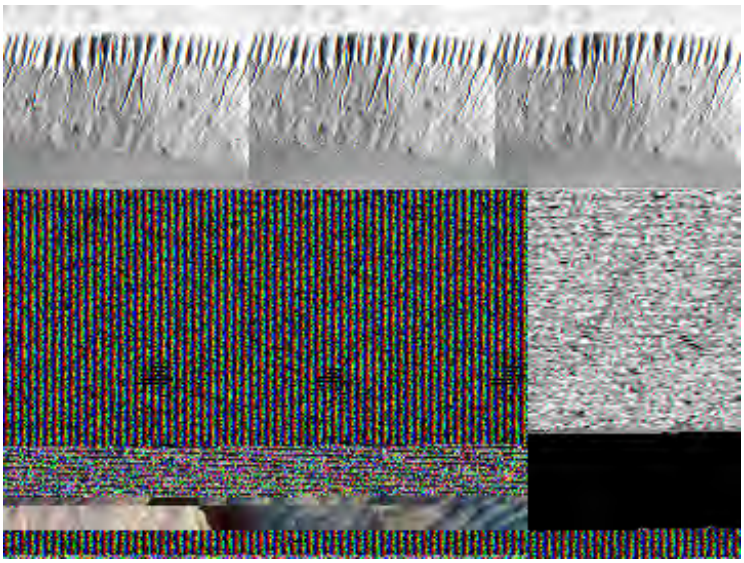
As seen from the above images, direct visual signs of layering are widespread on Mars. Some is caused by deposition of dust/fine sand layers in annual ice deposits in the polar regions, and is therefore presumed young. However, much older layers of varied nature - mainly basalt and sedimentary lake-or-ocean deposits - occur over wide areas of the martian terrains. In the MOC image below, layers are standing on end - nearly vertical - within the Oudemans crater near Valles Marineris. In appearance these look similar to those found at the MERS Opportunity Meridiani site (page 19-13b) and are probably sedimentary units that were shoved up in the central peak that developed during the actual cratering.

An area in the Meridiani region is interesting because the light sedimentary units are topped by a dark unit (basalt?) which is now partially removed. The general setting is shown first; below it the light units are displayed in an enlargement to show the ripple and polygon features that may be primary sedimentary structures:

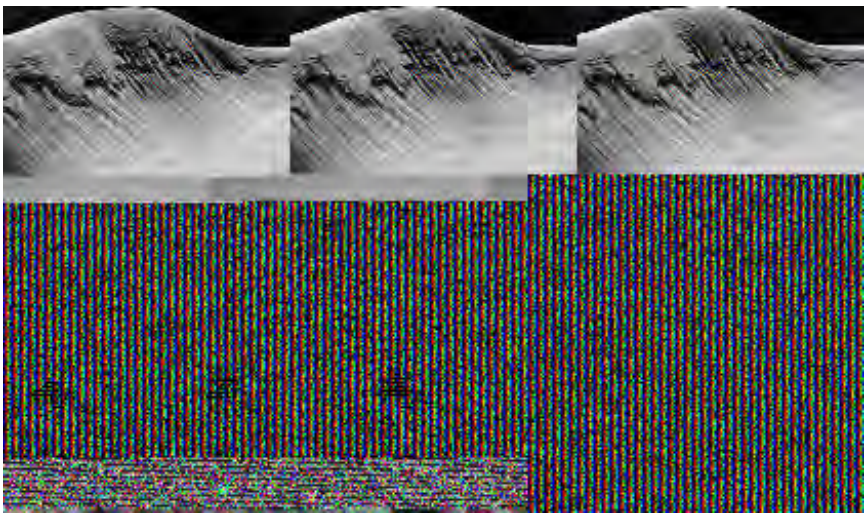


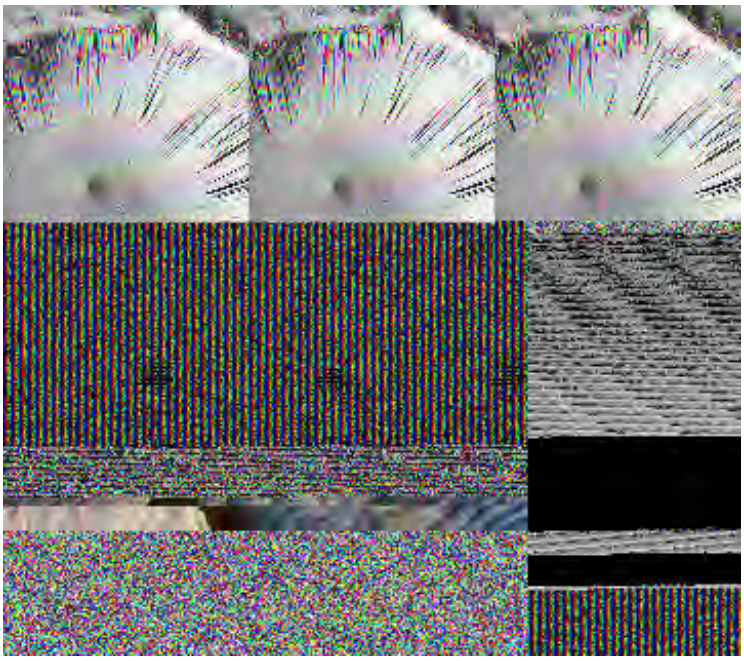

">

The MOC onboard the Mars Global Surveyor has been returning some provocative high resolution images that, after much interpretive debating, many geoscientists now believe offers supporting evidence of either water escape in the recent past or even indications of water still emerging. This MOC view below (5 m resolution) indicates channels coming from material in or below the edge of the southern ice cap. The stream outflow, if that is a valid identification, in each rill ends up with a small fan of deposits at its terminus - again, a feature consistent with fluid transport.



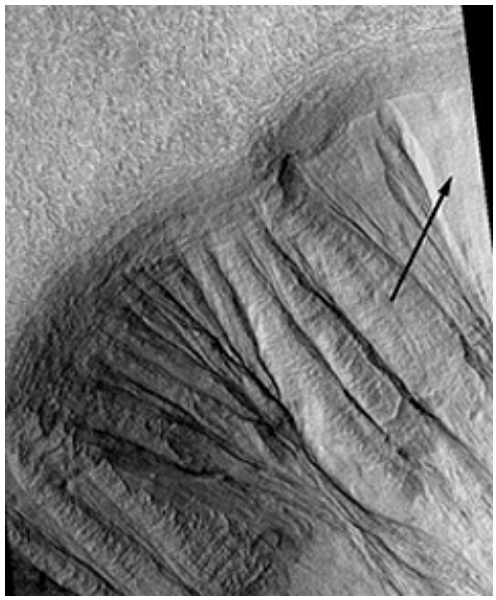
Gullies, or rills, also are noted in the walls of large craters, such as seen below. These may be caused by extended outflow of warmed subsurface ice which upon melting flows out at the walls. Other examples (second image) are considered by some to be streaks of dust avalanches, in which surface dust is mobilized, perhaps aided by water flow:



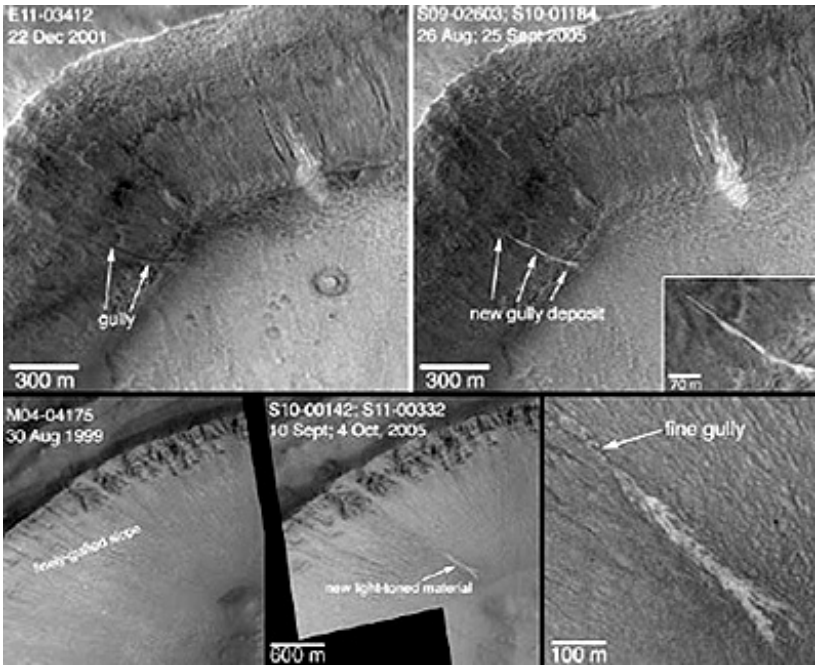


This Mars Global Surveyor image shows something unusual - what appears on the right to be a wedge of frozen ice cascading downslope. This has been interpreted to indicate that water

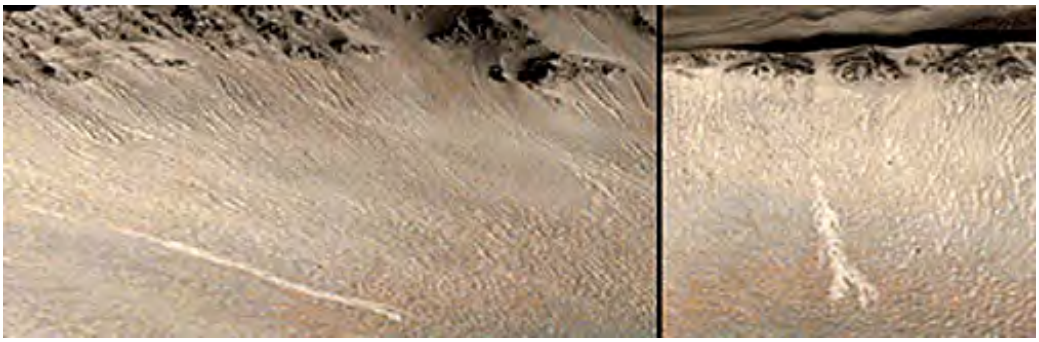
moves at shallow subsurface depths and here extrudes at a canyon wall where is freezes (much like the ice sheets seen in roadcuts during winter). If so, this further supports the argument that the rills or gullies also seen in this image are water-generated.



A MGS MOC image analysis team has accumulated visual evidence that some active process - the best "guess" favors running water - is taking place today on Mars. Before-and-after images show light-toned deposits along gullied crater walls that were definitely not there on the earlier date:



A color view of the Centauri Montes crater deposit indicates that it is much like similar deposits caused by water on Earth:

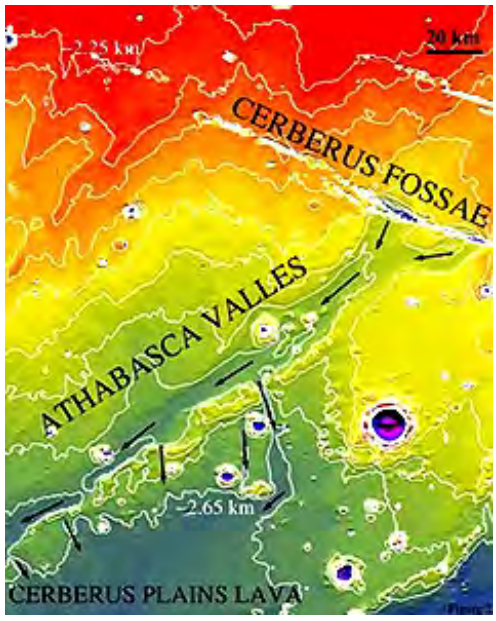


Early surmise held these deposits to have formed by release of liquid water. However, modeling has shown a much likelier explanation is formation by landslides that release dry powdered soil whose tone is less dark than the thin surface layer.

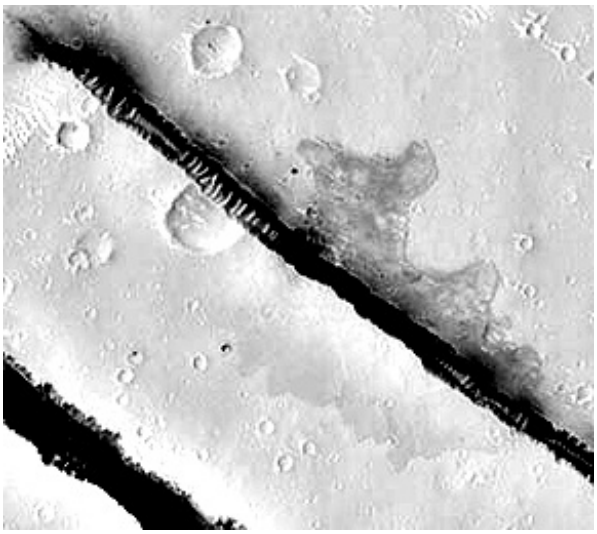
Water seems to exist on Mars in four plausible settings: 1) in small amounts in the atmosphere; 2) in polar ice caps; 3) in subsurface ice; and 4) as yet unproved, liquid water probably derived from melting of that ice along crater walls; even if so released, this water does not survive long in the fluid state but evaporates and/or freezes.

This next pair of images offers solid proof that new small and local markings develop even today on Mars. A single black streak appears on the left (1998) image of part of a crater wall. On the right arrows point to new black streaks that appear between then and 2001. These are interpreted as related to water outflow.

The next three images (NASA JPL and Malin Space Science Systems [MSSS]) demonstrate how planetary geologists go about interpreting and reasoning to conclusions using imagery returned from martian satellites and probes. The argument to be devised uses Martian Global Surveyor data to examine the region illustrated in this topographic map:



The Cerberus Fossa (channel) includes this 100 m wide, 10 m deep straight feature that, by comparison to similar areas on Mars, is almost certainly a tectonic graben, bounded by faults on either side expressed as steep walls. Emanating from the upper wall is what appears to be lava flows spread over both the flat land beyond the fault and into the graben floor.

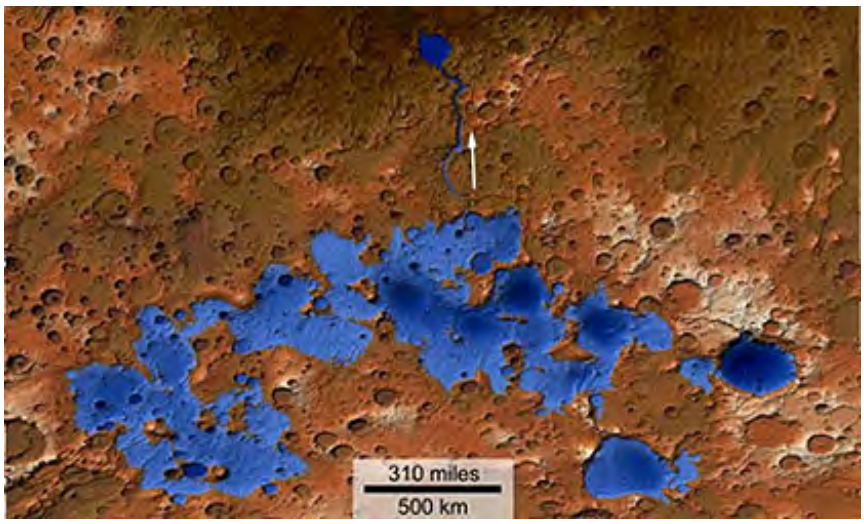


Water is presumed to have excaped from the Cerberus Fossa during, or at different times from, periods of volcanic activity. The amount of water needed to develop the valley is estimated to be about that found in Lake Erie today. Based on crater counts and other data, geologists believe that the principal valley-forming events may be as recent as 10 million years ago.

19-48: What causes the numerous indentations found in several of the dark layers?
ANSWER

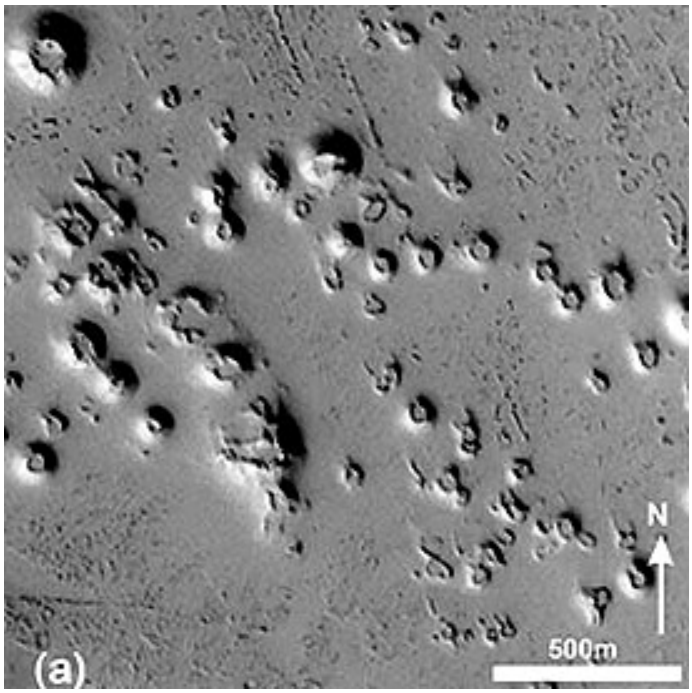
The Athabasca Valles extends southwest of the Fossa. It resembles a stream-cut valley similar to some on Earth. Within it is strong evidence of stream flow in the form of the streamlined, tear-drop shaped flatlands within it (see above on this page), such as shown here:

The Ma'adim Vallis (white arrow) is a 2.1 km (6900 ft) deep cut into the Southern Highlands that runs from a topographically lower area to the Gusev crater (see illustration on previous page). H. Irwin III, G. Franz of the National Air and Space Museum (NASM) and others have published evidence that the low area was once filled with water in several lakes whose extent, if on Earth, would extend across Texas and New Mexico. Their starting point is the low dark areas present in the MGS MOC image of this region. Using MOLA data to determine elevations, they have reconstructed these lakes (not now existing) by coloring the topographic lows blue against a color image of the region:



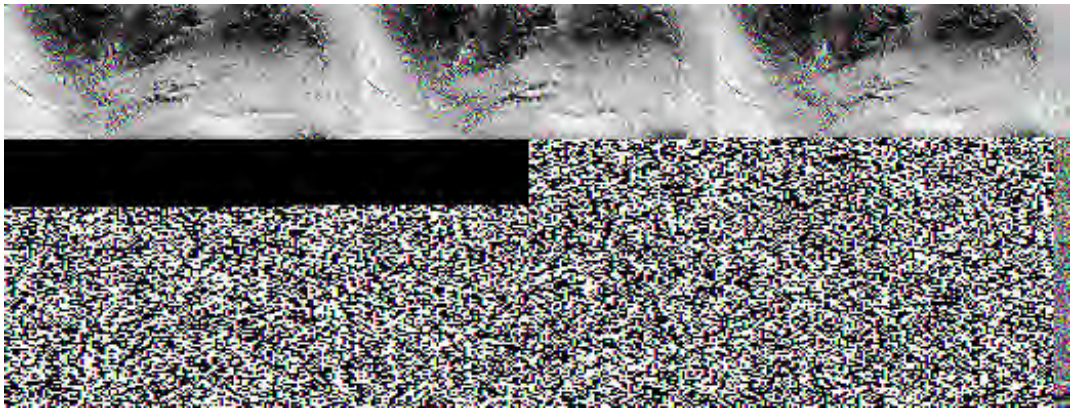
In their model, at some stage in Mars' past, the water from the central lake burst from its confines and rushed through the already forming valley, thus deepening it rapidly (analogous to the Scablands of the State of Washington; see page 17-4).

A new line of evidence for water and/or ice on Mars has come from MGS images of what is considered to be analogs to phreatomagmatic "rootless" volcanoes on Earth (especially in Iceland). Consider this image (MSSS) of a swarm of cones in the Cerberus Plains near the martian equator:



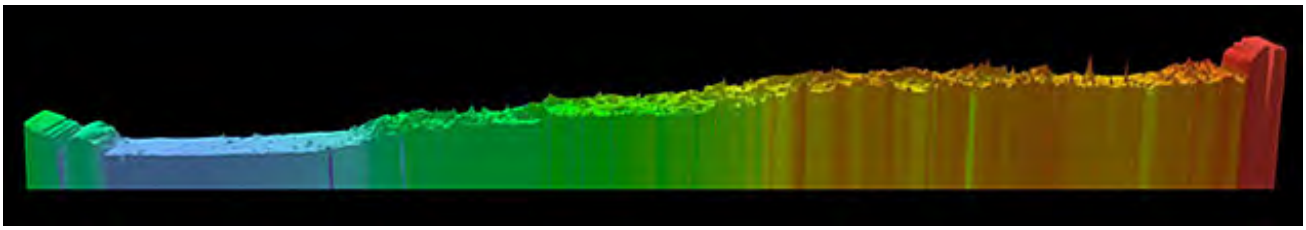
The origin of these miniature cones (typically, about 50 meters at their base) requires water to have been released at earlier times that becomes trapped near the surface (perhaps at depths of 3-5 meters). At the cold martian temperatures, this water is preserved as ice. Then, later eruption(s) of volcanic flows heat the water to steam, causing a pressure buildup that ejects material locally to build up the cones. The implication of the presence of these cones on what may be a young surface is that ice entrapment may occur within suitable martian terrains and could thus be available as a source of water during human exploration of this planet.

Some Mars scientists are arguing for the presence of water as ice and/or "snow" deposits that form both varying and persistent deposits in areas of lower latitudes. This is related to the migration of water vapor and CO_2 during the changing martian seasons. Perhaps the most provocative indicator of subsurface water that outflows in non-polar regions of Mars is shown in this next illustration:



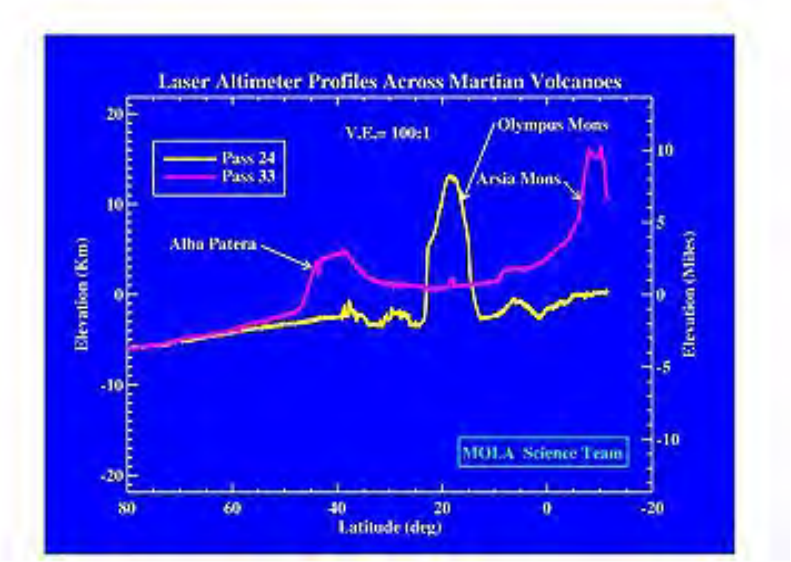
This feature has been given two interpretations: 1) it is an ice glacier (note the parallel ridges); or 2) it is a rock avalanche. Its geomorphic expression as seen from above seems better fitted to the first explanation.

The Laser Altimeter (MOLA) on the Mars Global Surveyor yields maps and profiles that show martian relief. The southern hemisphere is, on average, 10 km (6 miles) higher than the younger northern hemisphere. This is brought home by this block diagram that shows variations in topography alone the 0° latitude running from the North Pole (left) to the South Pole:



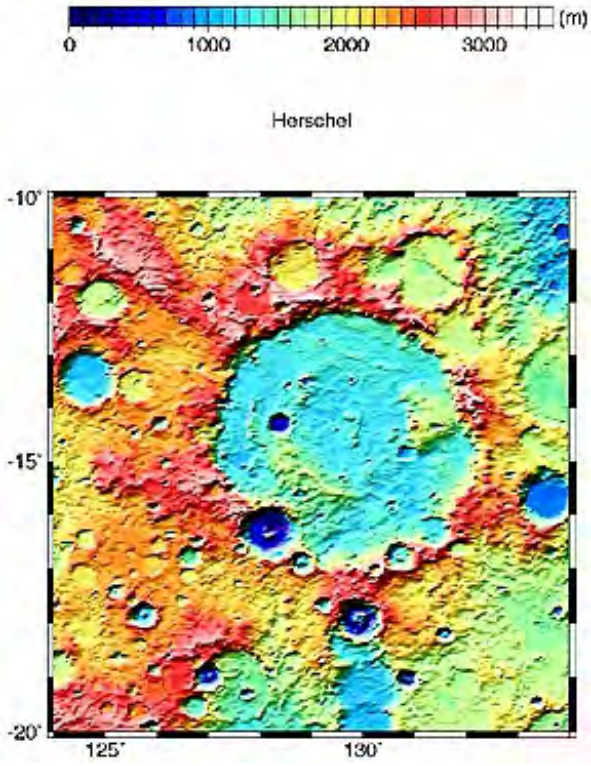
The maximum relief on Mars ranges to 30 km (19 miles), determined by Valles Marineris as the high point and the Hellas Basin the low point, about 1.5 times greater than that on Earth. The three white round areas are the Tharsis volcanoes and the offset white one is to the left. The brown rounded area above these is Alba Patera. Valles Marineris is obvious. The blue-green circular feature below is Argyre Planitia. The large elliptical blue region in the red southern hemisphere is a huge topographic depression about 2100 kilometers (1300 miles) wide known as the Hellas Basin.

One of the most dramatic of these MOLA data plots presents an exaggerated profile across Olympus Mons and Arsia Mons/Alba Patera (which is actually higher than Olympus).

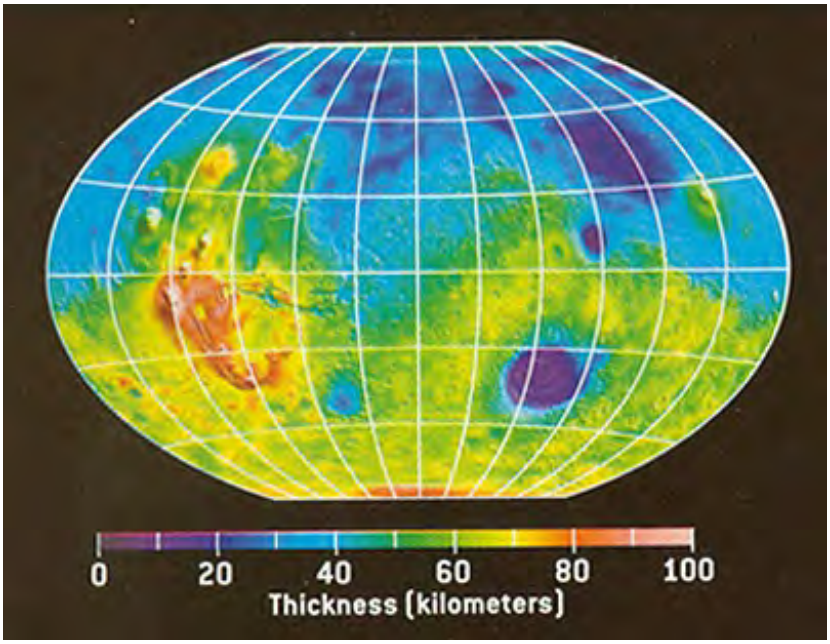


19-47: Compare the relief (difference in elevation) between Alba Patera and Olympus Mons. ANSWER

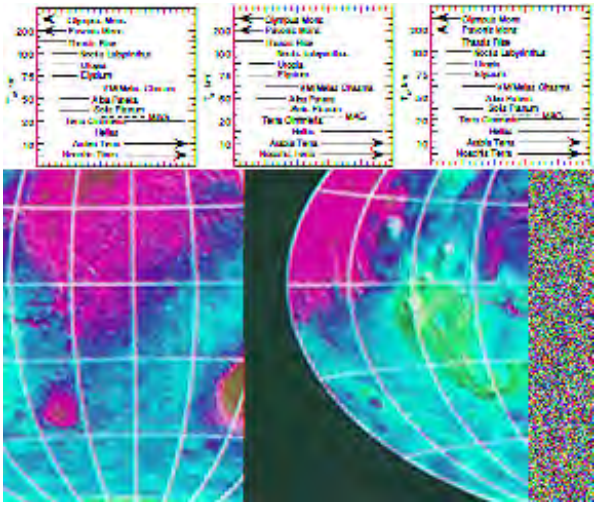
Still another example of a MOLA-based topographic map is that centered on the Herschel crater east of Hesperia Planum:



The topographic map of Mars, combined with gravity and other data, allow calculation of the general thickness of the martian crust, as shown in this global map:

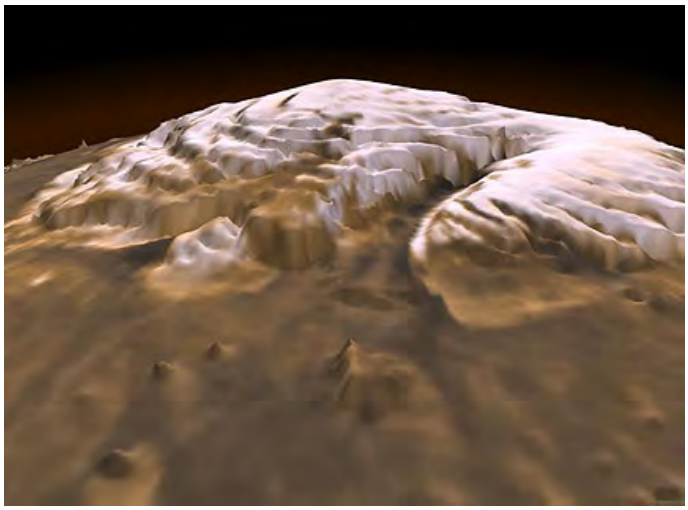


Thicknesses of the crust at major locations are estimated as shown in this Table:



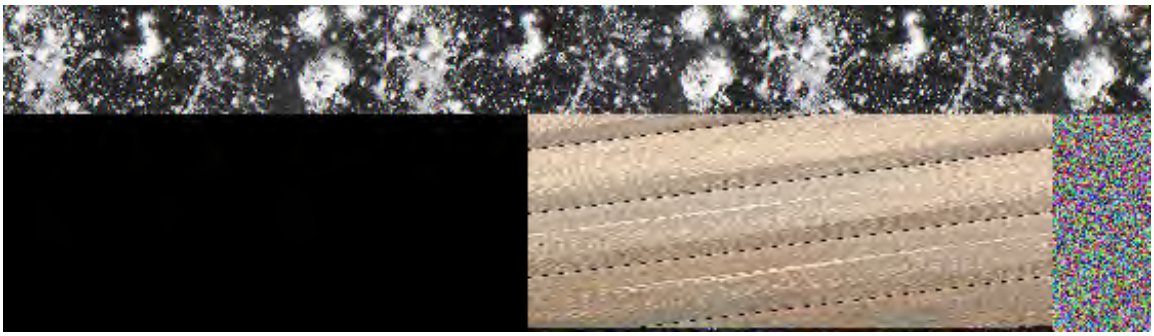
Note the fairly strong correlation between topography and thickness, emphasized here by using the same scheme of reds/orange for higher elevation/greater thickness and blues for lowest elevations/minimal thickness in the two relevant maps.

Another exciting and informative scene taken by the Global Surveyor combines a color view of the North Pole Ice Cap with topographic data obtained from a series of Laser Altimeter passes to produce this three-dimensional view:

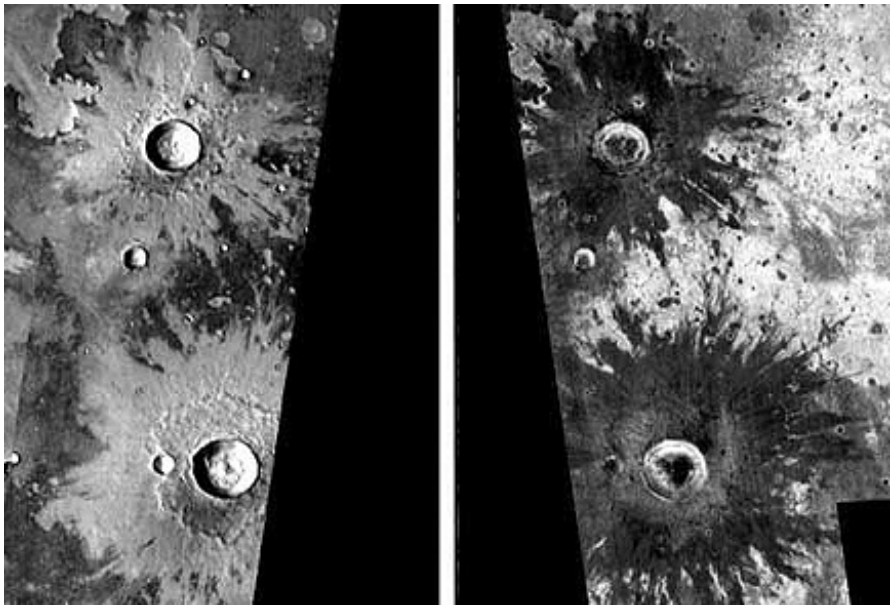


The size and amount of ice (which the spectrometer confirmed was largely water ice) was less than anticipated from pre-Surveyor observations. The cap is roughly 1,200 km (746 mi) across (crudely circular) and as much as 3 km (1.9 mi) thick. With an average thickness of 1 km (0.62 miles), the volume of ice is about 1.2 million cubic kilometers (288,000 cubic miles), roughly half that of the Greenland Ice Cap on Earth. The cap's general surface is quite smooth, but the ice cap is cut by large, deep (up to 1 km; 3,280 ft) steep-walled canyons and troughs, which may have come from cracks enlarged by wind and possibly meltwater. The amount of water located today at the North Pole, and lesser quantities now at the South Pole, does not seem sufficient to have once made an ocean over parts of Mars. It instead may be the remnants of any seawater that could have existed in the Mars past. Any relationship of this surviving water to the river-like channels on Mars is still speculative. Water may be widespread even now but would be subsurface. Detecting that water is one of the objectives of future Mars missions.

MGS's Thermal Emission Spectrometer (TES) can produce images similar to those made using visible light (as an example, check the image of Ma'adim Vallis on the previous page). Other THEMIS (longer) IR wavelengths form images which clearly differentiate hotter (whitish) from cooler (dark) surface materials. This image shows a thermal pattern in which the blackish part indicates a smooth, dust-covered surface and the light blotches represent blocks and debris in some cases associated with craters:

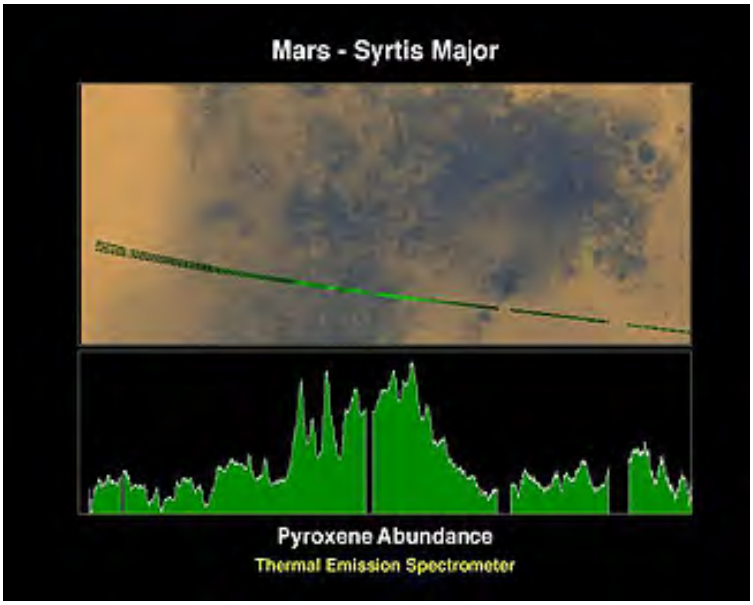


When TES surveys the same scene during martian daylight and again at martian nighttime, the differences in tone reveal information about the thermal inertia of the materials involved. Here are two views of a martian crater.



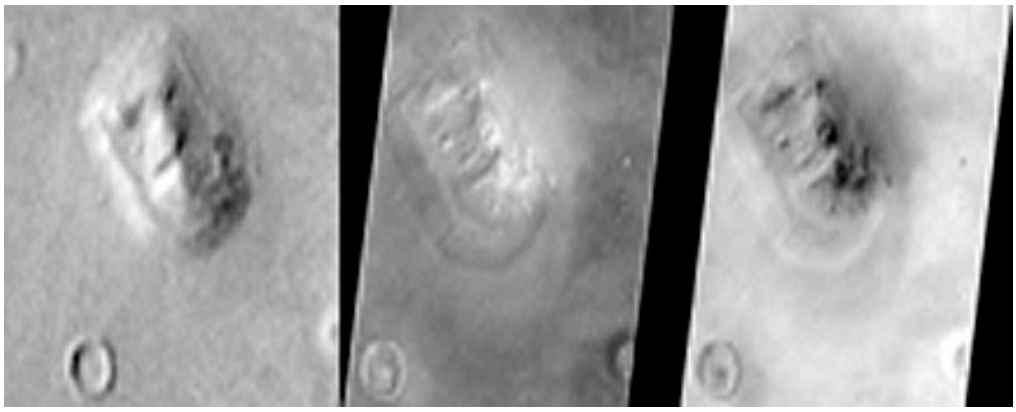
The ejecta blanket holdss more heat during the day and loses more during the night, owing to its porous nature. The surrounding rock/soil remains warmer during the night.

TES can identify some individual minerals and determine in a semi-quantitative way their proportions in the surface rocks (believed to be primarily basalt and andesites). The mineral Pyroxene (an iron-magnesium silicate) occurs in rocks found in Syrtis Major. Note its variable distribution, with the largest amounts present in the darker areas of this region.

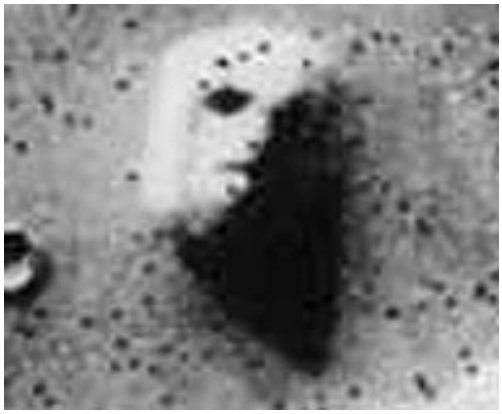


19-49: Account for the area that shows the highest pyroxene content; why? ANSWER

MGS sheds new light on the famous "face" on Mars first noted in a Viking image and seized upon by zealots favoring a lost civilization on the Red Planet as an artifact carved (like the Sphinx) into bedrock. The left view is the Viking image; the center (MGS) view shows the feature to have several mini-mountain peaks partly responsible for the fortuitous shadowing that highlights the face; the right view is a negative of the center view that recreates the shadowing.

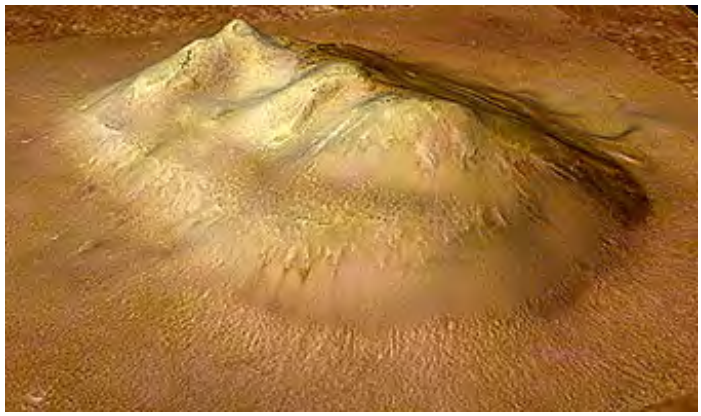


A more recent image of the "face", taken by Mars Odyssey (see page 19-13a) reveals an even more uncanny likeness to a face:



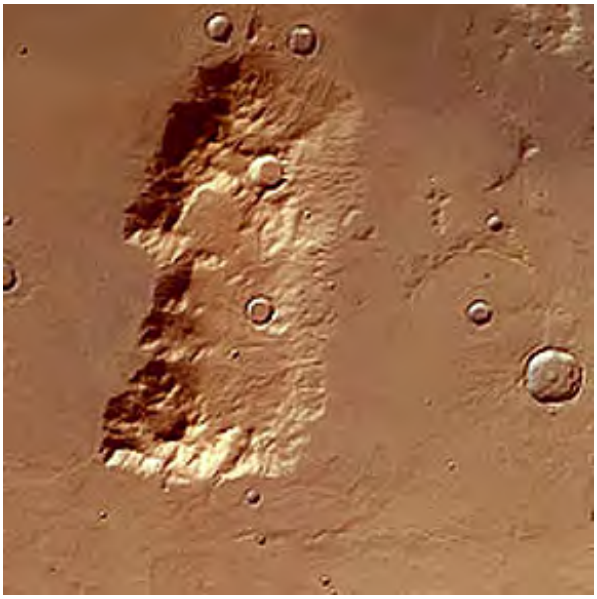
In this image, the rock promontory that looks like a nose even has a "nostril". Surrounding the "face" is a blanket of material that resembles hair. Judge for yourself what this feature might really be.

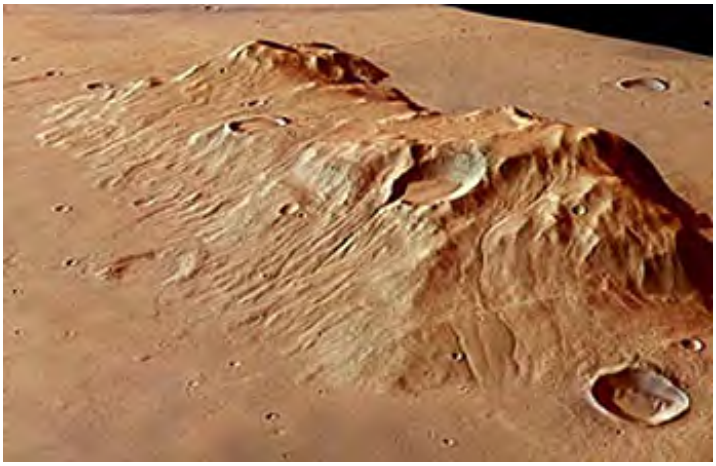
Perhaps another view will help in your surmise. This is a perspective image made from Mars Express data:



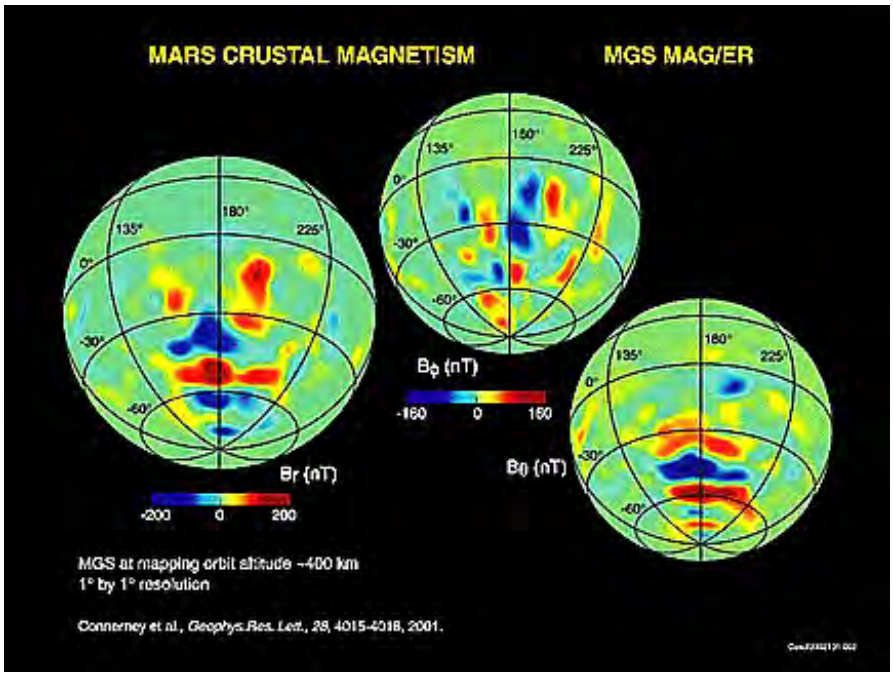
This view indicates the face to be an eroded mountain. The term "massif" applies to this and

similar features. When seen as a vertical image looking straight down, the height of such a massif is not particularly impressive. But when seen in perspective, this topographic type comes across as a dominating landmark. Consider these two Mars Express images of Ausonia Massif; the first looking straight down; the second as a perspective view:



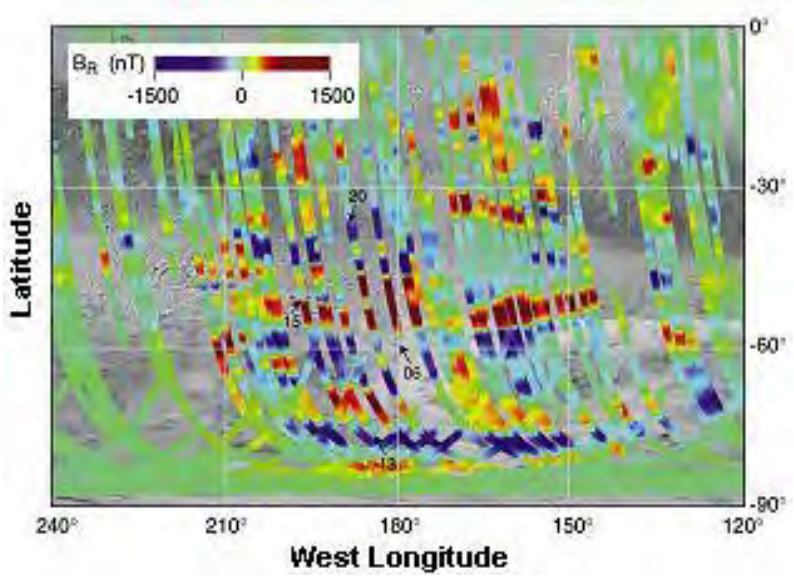


The MGS on Odyssey had an instrument capable of detecting weak magnetic fields. At the present time no active magnetic field exists. But earlier spacecraft suggested that one once existed. MGS provided this data set shown as three maps of the field strengths in different directions:

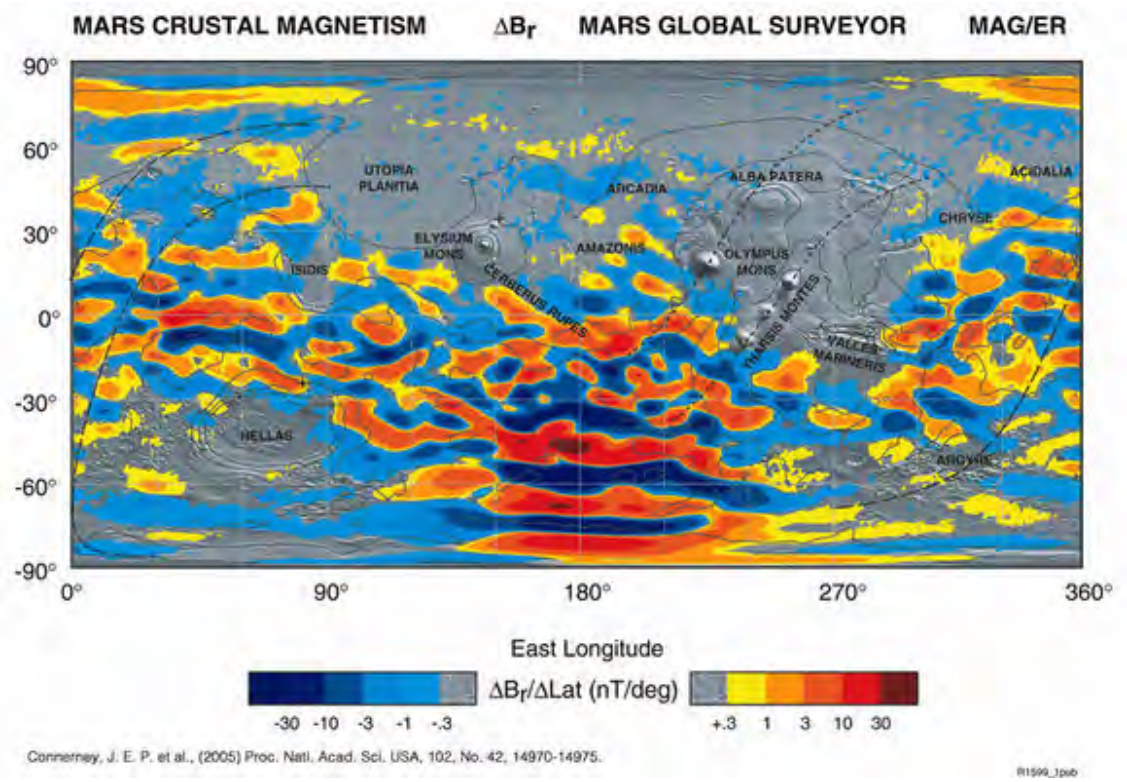


It is still somewhat uncertain whether this is an active magnetic field (implying some surviving liquidity of the core, within which magnetic fields are generated) or is a fossil (remanent) field. Opinion currently favors the latter; there are strong negative fields associated with the Argyre and Hellas Basins, which suggests an already frozen field lost strength from the impacts that caused these two structures. If true, then the protective cover afforded by a magnetic field that captures solar wind and other radiation did not last long. But prior to that an active field would help to maintain a thicker atmosphere which in turn might have supported water vapor that could produce rain, streams, and primitive life.

An April 1999 report by Mars Global Surveyor investigators has created a flurry of excitement about an aspect of the martian crust which may be a counterpart to the mobile segments of terrestrial crust that are involved in the concept of Plate Tectonics - a cornerstone in our understanding of the operation of the Earth's outer layers embodied in the idea of "continental drift". The Mars Global Surveyor magnetometer has picked out a series of magnetic stripes on Mars that systematically reverse their polarities. Thus:

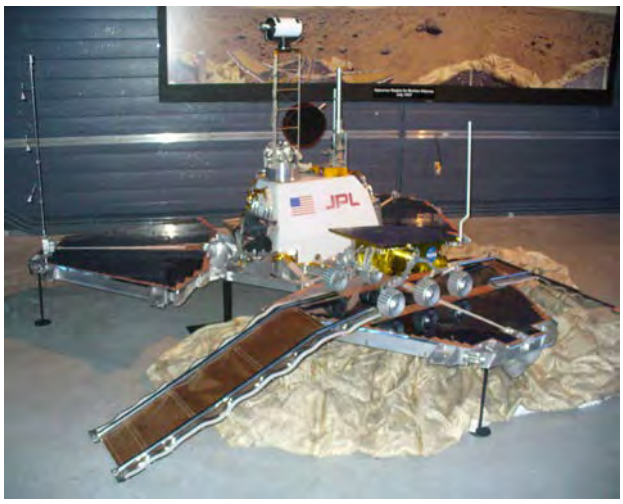


As more data accumulated, a nearly complete map of magnetic strengths has been made. Again, the stripes are the most obvious feature. One interpretation is that they are related to a time in the past when Mars had some mode of plate tectonics, with crust forming, rising, and subducting.



These martian strips are found mainly in the older terrains of the Southern Hemisphere. They trend at high angles to the polar axis. The implication is that this crust was formed by outpourings of lava from analogs to the terrestrial oceanic ridges that push the plates towards subduction zones (no such features have yet been recognized on Mars nor are there changes of mountains that result above these zones). This also suggests a much stronger magnetic field during early martian times; this would indicate a (partially) molten core. The full meaning of this new discovery is yet to be assessed.

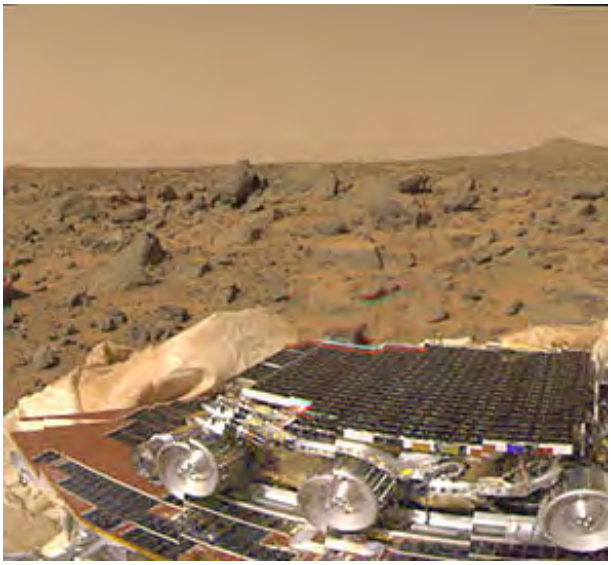
Martian exploration reached a new plateau when landers allowed on site close-up inspection of martian terrain and rocks. Launched on December 4, 1996, JPL's Pathfinder landed on July 4, 1997, in the Ares Valle near the earlier Viking 1 site . This shorter transit time (seven months) resulted from a better alignment of Mars and Earth in their orbits. Pathfinder used a cluster of small inflated bags which, together with a parachute, cushioned the descent and deployment onto the martian surface. Here is a mock-up version of the equipment as exhibited at the National Air and Space Museum. Note the small Sojourner rover.

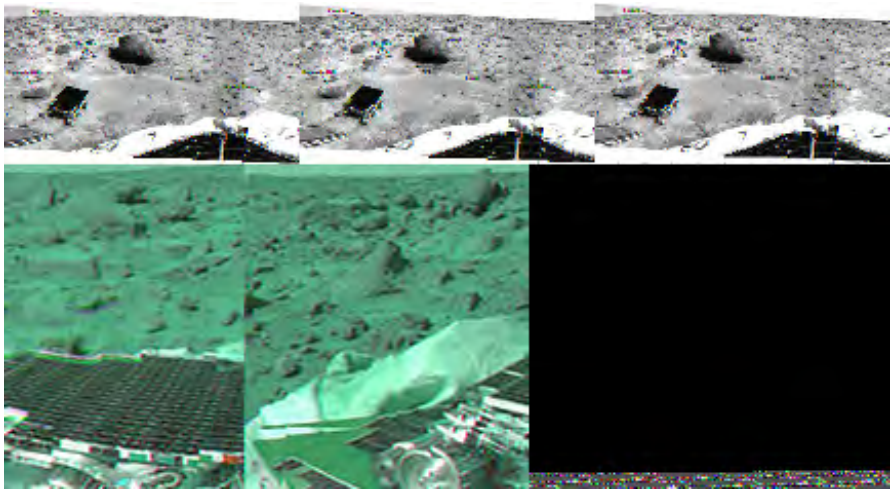


The primary scientific purpose of Pathfinder was to release the small Sojourner Rover, guided from Earth through the lander's communication system. The Rover moved independently (after departing the lander down a ramp) up to 20 m away on its six, flexible, rocker wheels, to take pictures and bring its spectrometer against rock surfaces for analysis.

This microrover, which weighed 11.5 kg (25.4 lbs), measured the following: length = 630 mm (24.8 in), width = 480 mm (18.9 in), and height = 280 mm (11.0 in). Using power from a solar panel, Sojourner moved at a maximum speed of 40 mm/min (0.13 ft/min) during the martian day. Its front and back wheels moved independently to control its steering. The Rover used a stereo camera, mounted at its front, which imaged objects in its path. Each excursion was intermittent, because operators on Earth determined direction changes to reach target rocks and avoid obstacles. This took extended time periods because of the long distances radio and video signals traveled back and forth. When it encountered small rocks, the wheels raised up to glide over the obstruction.

The stereo camera system on the Pathfinder took images of the surface around the lander, obtaining black and white views and colored ones with color filters. Sojourner's camera also collected images (mainly rock close-ups). In the three images below, the first shows part of the scene viewed by the camera, the second is a montage of part of the surrounding site, in which we identify (using friendly names that speak a personal touch) some of the rocks that it visited, and the third is a panorama of the 360° view around the Pathfinder (after Sojourner had departed) made by the camera mounted on the mast as it looked downward.

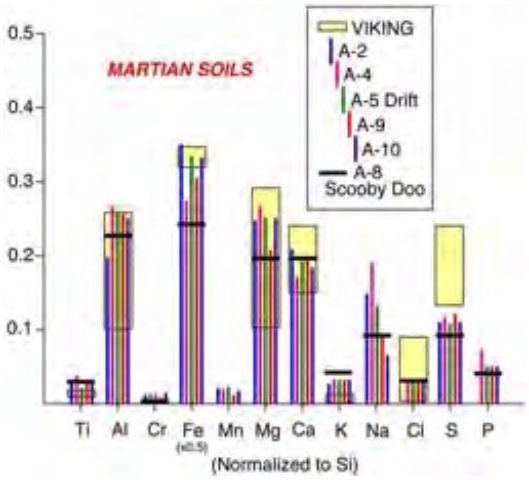




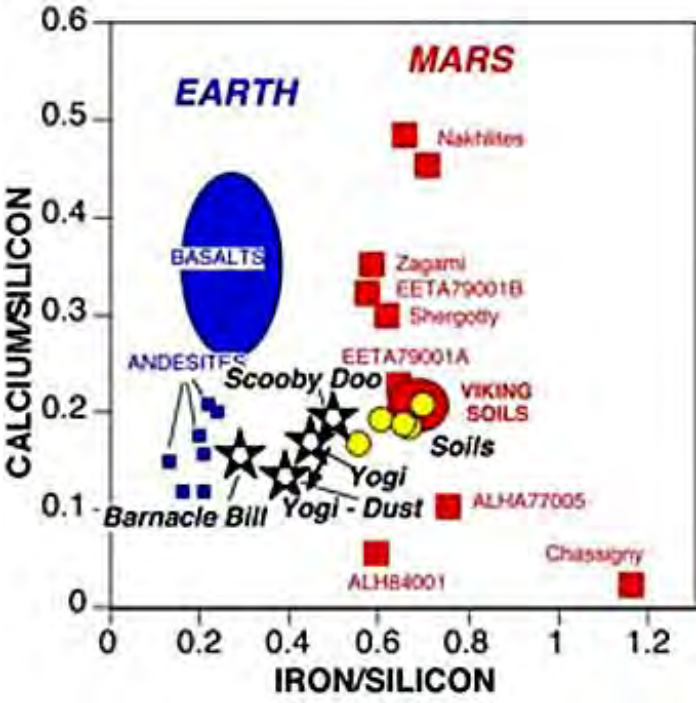
Many of the rocks are grayish on fresher surfaces. These surfaces appear weathered and pitted, and often are, at least partly, dust-covered. A typical large rock is Yogi (shown below), which appears to have a reddish dust blanket on one side and a sharp boundary with fresher rock (right side).

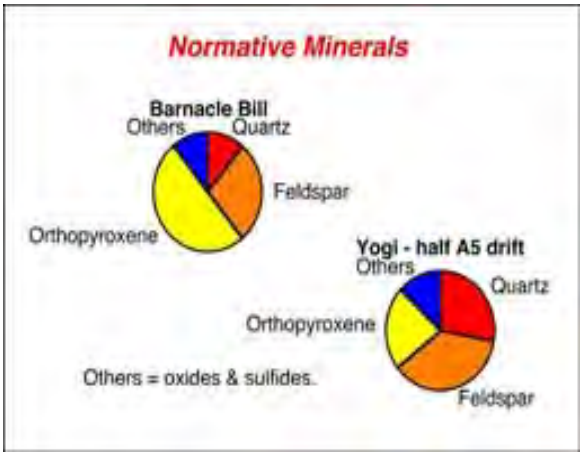
The primary instrument on the rover is the alpha-proton-x-ray spectrometer (APXS). This instrument uses high speed alpha particles to bombard soil or rock surfaces. These particles generate a back scattering of alpha particles, alpha-proton particles, and x-rays, whose energies as determined at the detectors are indicative of a wide number of elements. It thus detects, quantitatively, elements such as those in rocks, i.e., Si, Al, Ca, Fe, Mg, K, Na, Ti, Mn, Cl, S, P, O, and C. It can report these elements in their elemental form or as oxides. The diagram below is a histogram of many of these elements from Barnacle Bill, Scooby Doo, Yogi, and some soil sites.

Barnacle Bill's composition is close to that determined earlier at the Viking site but the other rocks and soils are different.



The next graph (top), is a plot of values of elemental ratios of Fe/Si and Ca/Si from the Pathfinder site, compared with certain Earth rocks. The Mars rocks at this site are definitely not basalt. They plot closer to but not within the compositional field of the terrestrial volcanic rocks known as Andesites. When we recalculate the elemental composition of the rocks into what petrologists term normative rock compositions, we see that the mineralogy is similar to andesites (bottom diagram).

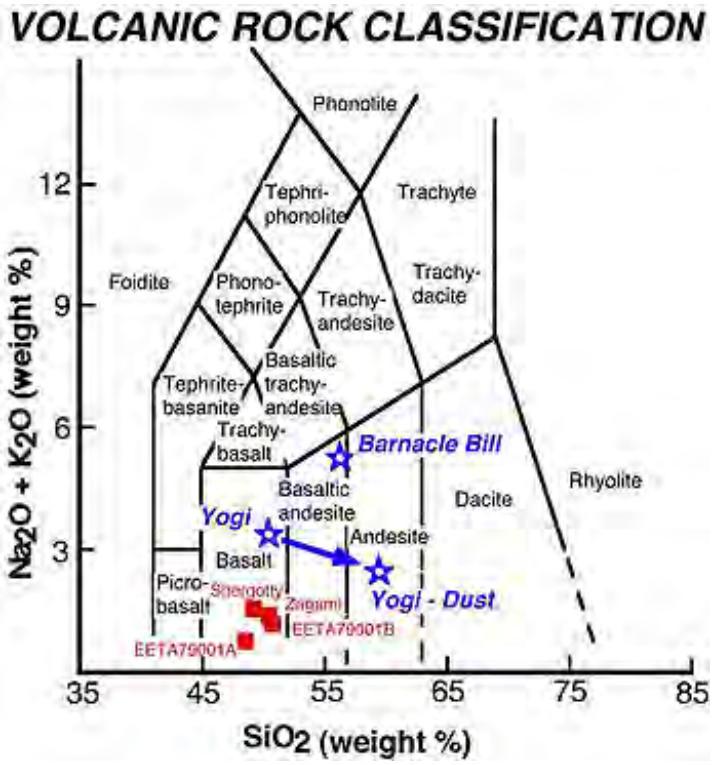




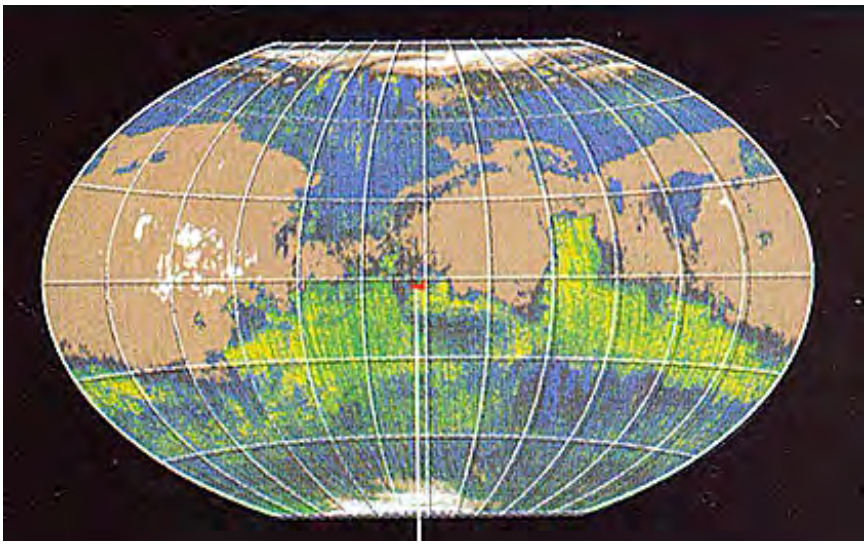
But the presence of normative quartz (leftover silica, after we assign all other SiO_2 to the other minerals), technically makes this rock type a dacite. Andesites and dacites are the common volcanic rocks emitted from continental volcanoes, such as those along the Ring of Fire (e.g., the Pacific Coast Cascades). The presence of these rock types at this martian site, if typical of other sites, suggests that the Red Planet had melted and differentiated, so that at least part of its crust is andesitic. Paul Lowman (author of Section 12 of this Tutorial) had postulated more than a decade ago that an andesitic primordial crust probably developed early in Earth history and may be the norm for other inner planets.

19-50: Which rock should be darker: Barnacle Bill or Yogi? ANSWER

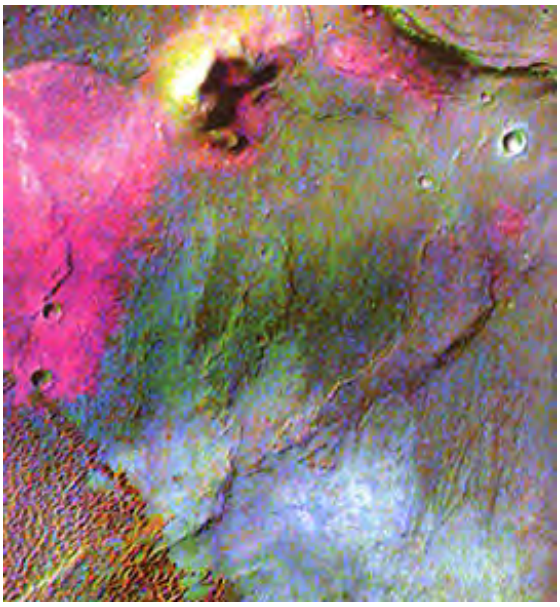
Most of the rocks plot in the basaltic and andesitic fields of a diagram which shows the classification of basic to silicic igneous (volcanic) rocks found on Earth in terms of $Na_sO + K_2O$ versus Sio_2:



Interpretation of MGS and Pathfinder data, supplemented by that acquired later by Mars Odyssey (see below), has led to this general map of the surface rocks and sediment cover on Mars. Basalts, overall the most common of the martian surface volcanic rocks, are shown in green and yellow; andesites in blue, and regions where dust cover is so thick that data on underlying bedrock could not be recorded appear in medium tan:



Some individual volcanoes appear to produce lava outpourings of different compositions from time to time, suggesting differentiation in the magma chamber. A good example is Nili Patera, in Syrtis Major, which produces both basaltic (blue) and andesitic (red) lavas as determined from MGS data; some of the latter may actually be dacites (andesitic but with quartz); one or more small peaks within the scene may have been further enriched in silica to gain a granitic composition:



The occurrence of andesites leads to several postulates: 1) these rocks are evidence of some differentiation in the martian mantle; 2) they may represent rocks involved in subcrustal release of water; or 3) they are anomalous local thin veneers on the dominant basalts caused by surface altering processes. No definitive conclusion has been agreed upon.

So far, martian Landers have gathered compositional data at five widely-separated martian sites (Pathfinder; the the two Vikings; the two MERS rovers.). In general, bedrock at the sites are similar in their chemical makeup. This may simply mean that the martian crust tends to be homogeneous, but even more likely is the explanation that martian wind circulation has scattered and homogenized compositional differences across the planet. The distribution of rocks and boulders at the Pathfinder site further supports earlier ideas that at one time in the Mars' past there was considerable water cover and resultant flow currents, causing a pattern sometimes observed on Earth as flood-related.

The last Pathfinder image we show below is a classic sunset picture, taken through the thin, dusty, martian atmosphere that produced a fan-shaped glow around the distant Sun.

The martian exploration program since the opening of the next millenium (we accept that time start as January 1, 2000) is presented on the next pages (19-13a and 19-13b).

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

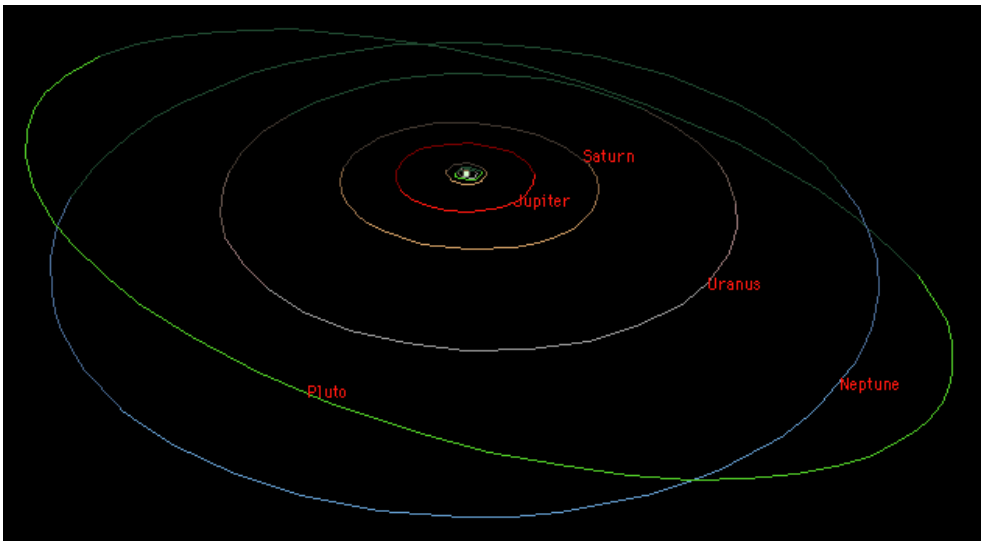
So far we have surveyed the general characteristics of four Rock planets (usually called the Terrestrial Planets because of gross similarities to earth). Now we proceed to look at the Gas planets (normally referred to as the Giant Planets - Jupiter, Saturn, Uranus, and Neptune - because all four are (except for the fifth, Pluto, a maverick) very much larger than the inner rock ones. All of these gas balls are made up principally of hydrogen, some helium, and other gases arranged as thick "mantles" with small central cores of rock with possibly some metal. All have distinct atmospheres with circulation driven in part by planet rotation. The internal temperatures are far too small to favor nuclear reactions, so these planets are in no danger of converting to stars. The general characteristics of the Giant Planets are given on this page. The history of their exploration, first by the Pioneer and Voyager spacecraft, and more recently for Jupiter, Galileo, and for Saturn, Cassini, is outlined. There is a brief preview of a special feature of each - the presence of both spherical and irregularly shaped satellites made up of rock and ice, and for several sulphur- or organics-rich surfaces.

The Giant (Outer) Planets; Pioneer and Voyager

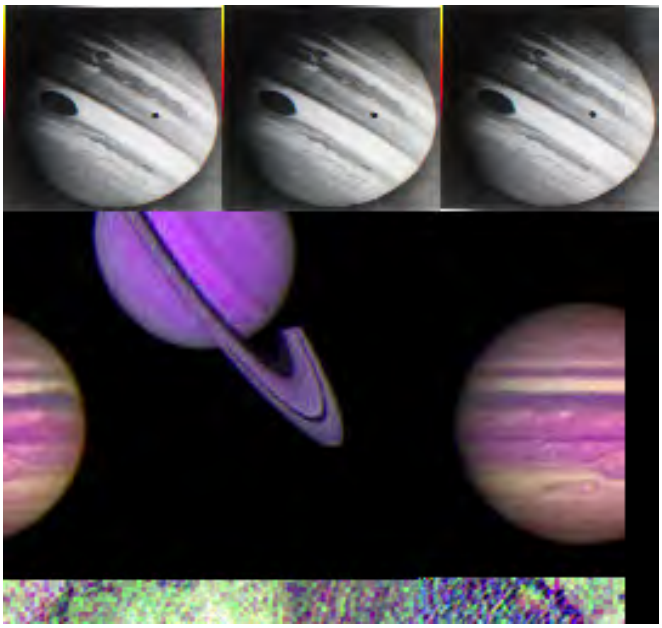
Overview of the Outer Planets

Our attention turns now the the four Giant outer planets (plus a preview of Pluto, a much smaller rocky body that is no longer a legitimate ninth planet) that consist of huge amounts of hydrogen and helium gas surrounding small rocky (and possibly with some metal) cores. Here is a composite showing a full disc space image of each Giant planet, shown in their relative sizes.

It also helps to place these four and Pluto in their orbits at their relative distances from the Sun, as shown here (the innermost brownish-orange ring is the Earth's position).



The first trips to the outer planets were the Pioneer 10 and 11 missions, launched in March 1972, and April 1973, respectively. Pioneer 10 went past Jupiter in December 1973, enroute to becoming the first man-made object to leave the Solar System. (After 25 years of continuing contact, communication with this spacecraft was finally terminated at the end of March, 1997, as Pioneer 10 reached a distance of more than 11 billion kilometers [6.8 billion miles] from Earth.) Pioneer 10's principal finding was to measure a magnetic field around Jupiter that is about 2,000 times stronger than Earth's and to detect intense radiation belts and a Jovian ionosphere. It sent back excellent close-up, black and white images of the planet's surface from its TV camera. Here is one of the classic (historic) images, taken through a blue filter on the Photopolarimeter lens, showing a full face view of Jupiter, with its conspicuous dark ellipse known as the Great Red Spot (an atmospheric cyclone). The small black spot to the right is the shadow cast by one of Jupiter's moons, Io. At the time, Pioneer 10 was about 2,500,000 km on the Earth side of the planet.



Pioneer 11 passed Jupiter in December 1974, enroute to a flyby of Saturn in September of 1979. This spacecraft traversed Saturn's rings without being destroyed (as some predicted) by their widely-spaced particles. Pioneer 11 also found evidence that Jupiter's magnetosphere extends beyond Saturn's orbit in the direction away from the Sun.

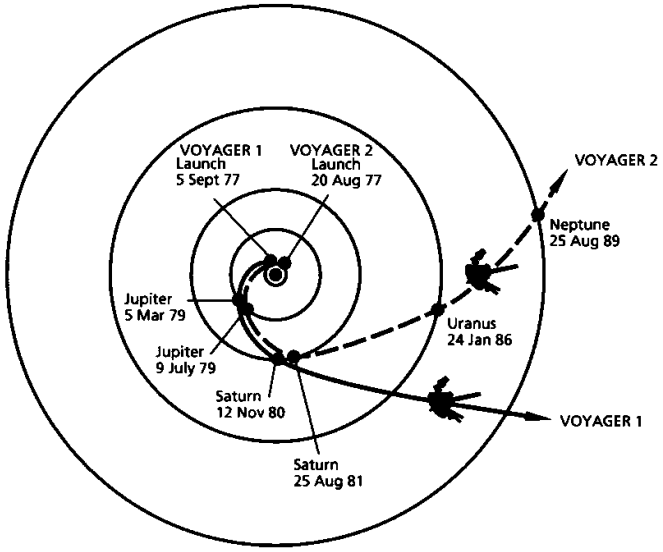
We come now to what is indisputably one of the most outstanding planetary space missions ever sent forth by mankind: the Voyager 1 and 2 probes to the outer planets. At a cost of under one billion dollars - a true bargain even then; two remarkable spacecraft visited all of the giant

planets, Jupiter, Saturn, Uranus, and Neptune. They started their grand tours in 1977 and are still going strong, as these vehicles leave the Solar System. They imaged all four planets and 48 satellites, some of which they discovered. A general summary of the Voyager missions is offered on the Jet Propulsion Lab's (JPL) Voyager home page, and its links. It will lead to a 15 minute video on the Voyager missions. Or you can access this summary by clicking on Voyager videos, that gets you to the Multimedia menu, and select "Voyager: The Grand Tour" by clicking on its blue background logo.

One impetus for the missions was the favorable alignment of the giant planets that occurs once every 175 years, such that initial trajectories plus corrections might allow at least one spacecraft to fly past Jupiter, all the way to Neptune. With each planetary flyby, the spacecraft received a gravitational "kick" to speed the vehicle along, so as to favorably alter its course and shorten its travel time significantly (without this boost, the trip to Neptune, using the initial launch velocities, would have taken 30 years).

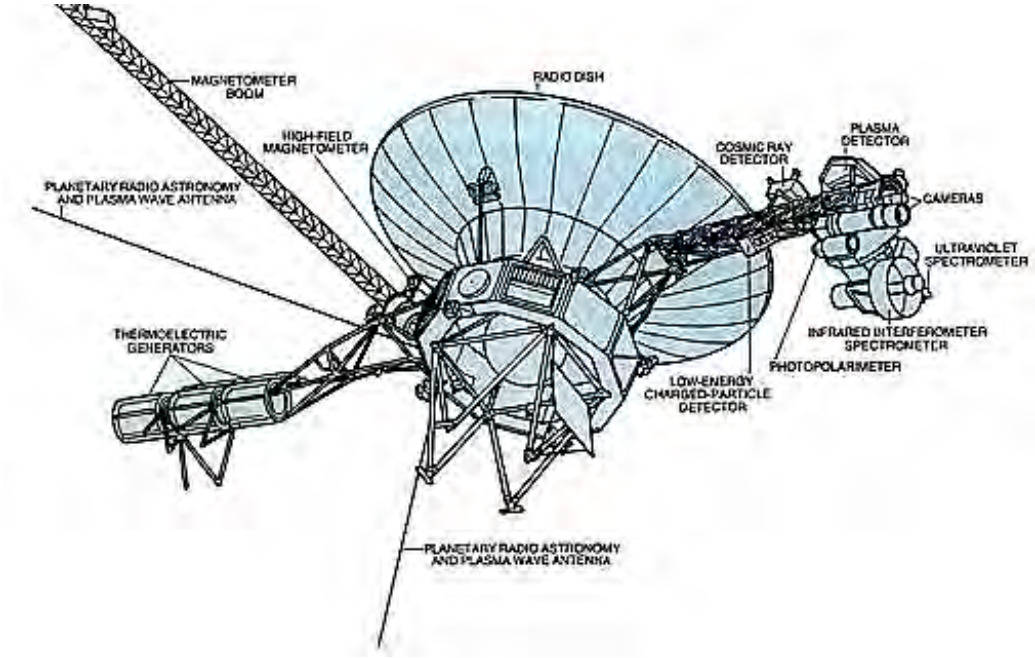
Voyager 2 launched from Earth on August 20, 1977, and arrived at Jupiter on July 9, 1979. Voyager 1 took off on September 5, 1977, reaching Jupiter on March 5, 1979. Even though leaving later, Voyager 1's shorter path reduced its trip time. Both spacecraft then went by Saturn, Voyager 1 during November 1980 and Voyager 2 in August, 1981, receiving another gravity boost. At their Saturn encounters, the Voyagers had met the initial requirement of five years of operation, but Voyager 1's chosen trajectory (past Saturn's moon, Titan) did not favor it continuing to Uranus. So JPL officials allowed it to depart upward (towards celestial north) out of the ecliptic plane. Voyager 2's "all systems go" status qualified it, with appropriate course corrections, to head to Uranus, encountering it on January 24, 1986. Further course adjustments sent it on to Neptune, with closest approach on August 25, 1989. At each new rendezvous, the spacecraft performed superbly, providing detailed data on each giant planet. Both spacecraft have now left the planetary region of the solar system.

This sketch indicates the trajectories each Voyager followed:



An artist's rendition of the identical Voyager spacecraft, traveling through space, appears below. The scientific payload appears at its upper left and also on the large boom extending right to left.

The Voyagers carried instruments designed to support 10 experiments. In addition to TV cameras, others included UV and IR sensors, and cosmic-ray, plasma and charged particle detectors, and magnetometers. These are shown in this schematic diagram:



Because of their extreme distance from the Sun, they could not use solar panels for power, but, instead, they use radioisotope thermoelectric generators (RTGs) that convert heat from decaying plutonium into electricity. Because of this wise power choice, both spacecraft, even today, beam radio signals full of new science data from well beyond 100 A.U. (1 A.U. is the distance from the Sun to Earth: 150 million kilometers [93,000,000 miles]), having passed the heliopause, the outermost boundary of measurable solar wind.

As an introduction to the superb images sent back from the Voyagers, consider the next two views, each showing a segment of its planet, along with two visible moons. At the top is Jupiter, seen by Voyager 2, on February 3, 1979, from a distance of 20,000,000 km (12,400,000 mi), with its moons in the foreground: Io on the left and Europa to its right. At the bottom is an image from Voyager 1 that displays part of the saturnian surface, its magnificent rings (innermost = A, then the Cassini Gap, then B, the Encke Division, and C and the outer rings), and the moons Dione and Tethys.

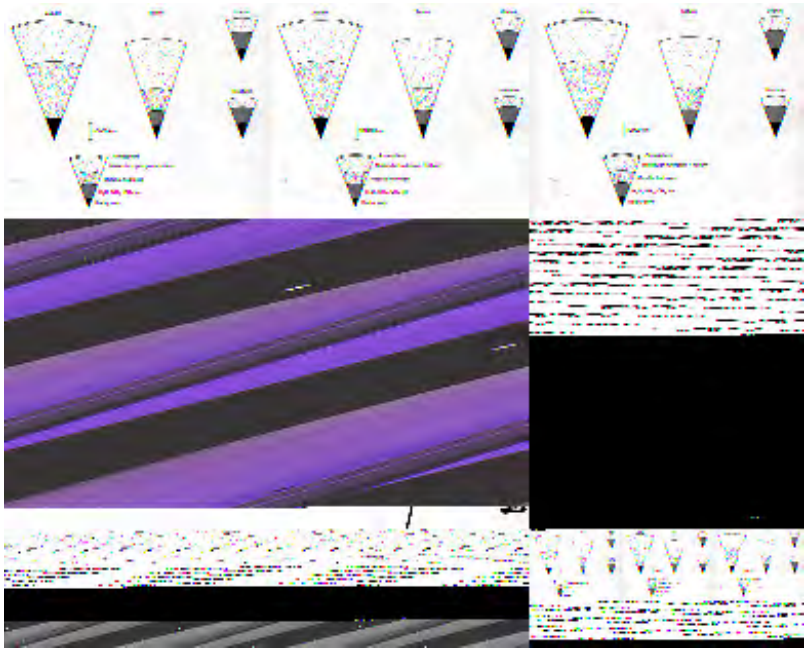
19-52: What are the three dots (glitches?) in the Saturn picture? ANSWER

Galileo was a major mission that went to Jupiter alone, arriving in 1995. It studied both the planet itself and its major moons. We will treat these in detail in the next two pages As a preview, JPL has a webcast from its von Karman Series that looks to the Galileo mission in detail. Access it through the JPL Video Site, then the pathway Subject-->Von Karman Series 2003 --> Format -->Webcast --> Search to bring up the list that includes "Galileo's Odyssey to the Worlds of Jupiter", September 18, 2003. To start it, once found, click on the blue RealVideo link. We will cite another lecture concerned with the Cassini-Huygens mission on the first Saturn page.

The Giant planets are just that: vastly larger, near-spherical bodies, dominated by atmospheres that comprise the bulk of their volumes. Orbiting at 5.2 A.U., Jupiter is the largest, with an equatorial diameter of 142,984 km (88,850 mi, measured from the center to an outer level, where we estimate the atmospheric pressure is 1.0 atm.) and a mass of 1.9×10^{27} kg (4.19×10^{27} lbs) (greater than twice that of all other planets combined, and 318 times greater than Earth's). Moving further from the Sun, the planets are progressively smaller, as follows: Saturn (9.54 A.U.) - 120,530 km (74,897 mi)(about 90 Earth masses); Uranus (19.22 A.U.) - 51,116 km (31,763 mi); Neptune (30.1 A.U.) - 49,528 km (30,777 mi). They experience polar flattening of about 2% to 10%. Each planet rotates very rapidly for its size; for Jupiter and Saturn their rotational periods are about 10 Earth hours; Uranus and Neptune have rotational periods less than one Earth day.

The Giant planets consist mainly of hydrogen and helium (for Jupiter and Saturn, the ratio of H/He is about 9:1 by number of atoms and 7:3 by atomic weight); minor amounts of methane and ammonia; and traces of nitrogen and water. Jupiter and Saturn are massive enough to convert the inner layers of hydrogen into a metallic liquid state. For Jupiter, this region is about 40,000 km (24,856 mi) thick. It is less than 15,000 km (9,321 mi) thick for Saturn. For both planets, the region

grades outward into liquid molecular hydrogen (Jupiter's outer region is about 18,000 km [11,185 mi] thick), and then comparatively thin atmospheres. Metallic hydrogen is absent in Uranus and Neptune, so, their interiors are mostly molecular hydrogen. Based on density calculations, all four planets may have a rocky (silicate) core, subjected to high pressures and temperatures (Jupiter: about 20,000° K). The radius of Jupiter's core is about 5% and Saturn's about 10% of the planetary radius. The cores of Uranus and Neptune extend to about 1/4th their planetary radius.



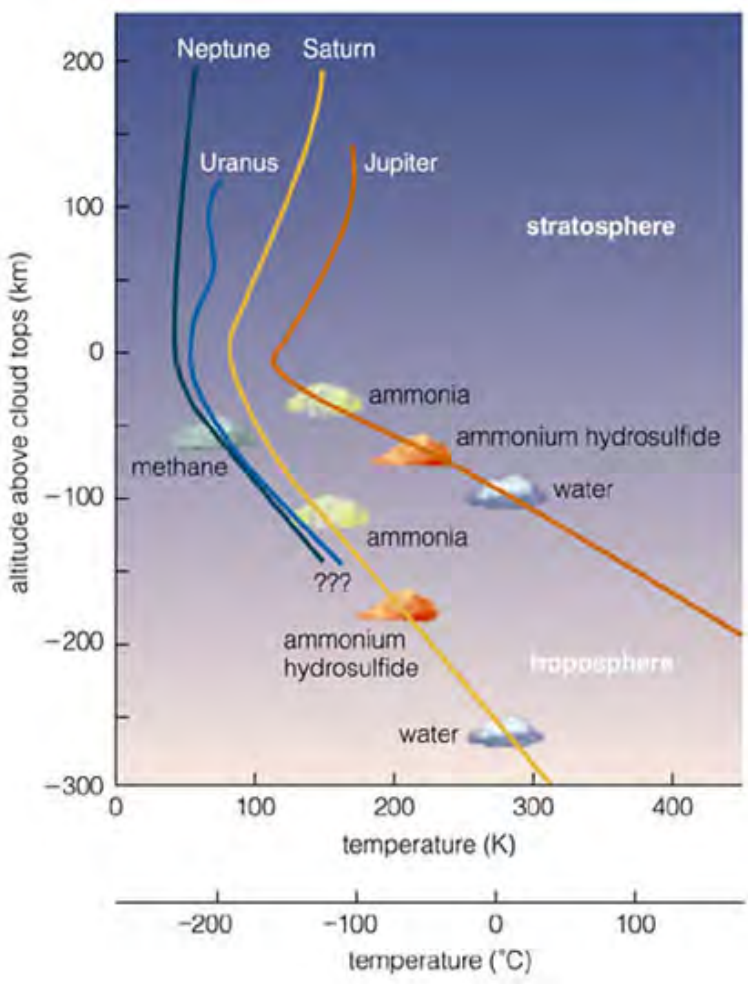
Reinterpretation of earlier data (to be refined from future Cassini measurements) has cast doubt on any solid core at all inside Jupiter. The argument is that the intense pressures in the core region, coupled with higher temperatures, prevented rocky core formation, so that the elements beyond helium are still dispersed in the gas/liquid phases.

Thermal energy released from compression is largely responsible for the internal heat that makes each planet warmer at the surface (reverse thermal gradient) than expected from solar heating alone. For Jupiter, heat from the interior provides 70% more thermal energy than received from the Sun. This energy promotes the fluid dynamics that drive the atmospheres into intense motions in belts or zones, roughly parallel to their equators, with wind speeds far in excess of those on Earth. Thermal conditions at great depth cause electron stripping and other mechanisms involving conductive motions that produce magnetic fields around each planet. The fields of Jupiter and Saturn are about 4,000 and 1,000 times, respectively, stronger than that of Earth's magnetic field. The fields of Uranus and Neptune are much weaker. However, the magnetospheres of Uranus and Neptune have unusual shapes and distributions, with the magnetic pole axis of each strongly inclined to the rotational axis. Uranus's spin axis tilts 98° , so it almost lies in the solar ecliptic plane.

Prior to the Voyagers' visits, we only knew Saturn had rings of material moving in near-circular paths at varying distances beyond the planet's (gas-pressure-defined) surface. But, as the spacecraft imaged the three other planets, we found positive visual evidence of generally more subdued rings at each one, so that we now consider rings the norm for this type of gaseous body. The rings consist of minute particles of ice and rock. At Saturn, some are as large as a small house, and concentrate in orbits only a few kilometers thick. Within the rings, the particles are widely separated. The origin of such rings is still uncertain, but opinion favors disruption of earlier satellites. However the rings may have collected debris that eroded from existing satellites, or that the planet captured from interplanetary space, or from "leftovers" during the planetary accretion.

The Voyager and Cassini visits have sent back data that allow some specification of the atmospheres (composition and temperature profile) of each of the Giant planets. This is

summarized in this diagram;



Each of the outer planets has a number of orbiting moons. Many of these have been discovered by visiting probes such as Voyager and Galileo. The planets may have captured some from asteroid-like belts or from passing comets and other space debris. Most of the satellites consist of mixtures of rock and ice (water, carbon dioxide, or other frozen gases). Ice is often the dominant material at the surface. The surfaces of many satellites show varying, often heavy, cratering that may be preserved states from soon after they formed. Some satellites have unusual, patchwork-like multi-typed terrains that suggest they were reassembled from pieces torn apart from a collision. Indications of later, partial melting are present in a few. One satellite has an atmosphere; several may contain liquids at or below their surfaces; and at least one has ongoing volcanic activity. All in all, the satellites have a lot in common, yet each seems to retain its own distinctive "character".

Using both ground-based telescopes and space probes, a large number of Giant Planet moons (satellites) have been discovered, as summarized below. These fall into two categories as defined by their orbital parameters: 1) regular moons - those with almost circular orbits that generally follow nearly equatorial paths; and 2) irregular moons - those with highly elliptical, large, and tilted orbits; many with retrograde motion, i.e., moving in a direction opposite to the rotation of their parent planet. The bulk of these satellites are small (many just a few kilometers in longest dimension) and irregular in shape. Most are likely captured asteroids and perhaps comets.

The current inventory of Giant Planet moons is as follows (information extracted from Jewitt, Sheppard, and Kleyna, *The Strangest Satellites in the Solar System*, Scientific American, August 2006, pp. 41-47):

* Jupiter: 8 regular moons; 55 irregular.

* Saturn: 21 regular; 26 irregular

* Uranus: 18 regular; 9 irregular

* Neptune: 6 regular; 7 irregular

* Pluto: 1 regular; 2 irregular

This grand total of 153 moons will likely increase as Uranus, Neptune, and Pluto are eventually visited by spacecraft that can detect small objects not yet visible from ground telescopes. This number of moons is in stark contrast to the inner planets which so far have only 3 satellites orbiting around Earth and Mars.

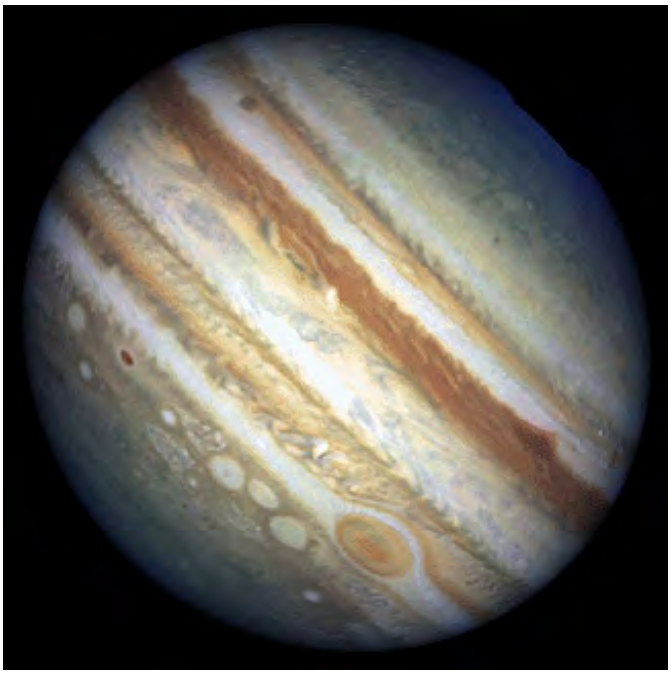
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Jupiter is the largest, and closest to the Sun, of the Gas Giants. It has been visited by two Pioneer missions, Voyagers 1 and 2, and Galileo which has been orbiting the planet since 1995 until it was directed to plunge into the jovian atmosphere in 2003 (to avoid possible crashing on one of the satellites, contaminating it with radioactive material onboard to provide power). Cassini is the latest to pass near Jupiter, at the end of 2000. All gave better looks at this massive planet than we had before with telescopes. A better understanding of atmospheric composition and circulation came from each. The Voyagers and Galileo have also provided tantalizing views of the jovian satellites. This page considers Jupiter's atmospheric properties, including some of the measurements made of its composition, and the discoveries of rings (inconspicuous relative to Saturn's), lightning, storms, and auroral effects near the poles.

Jupiter: An Overview

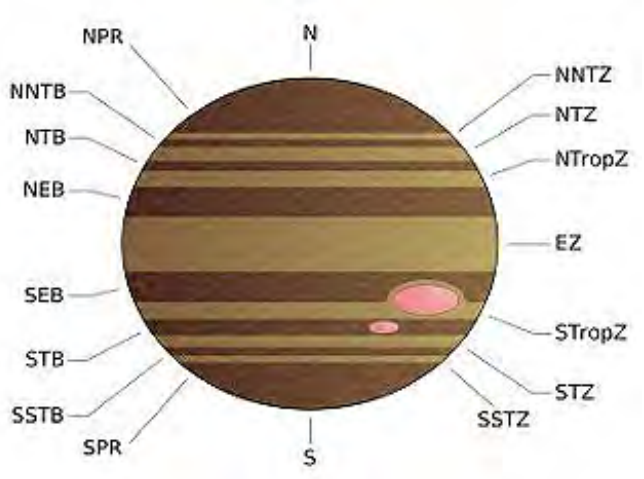
Our tour begins with Jupiter, shown here in full disk images taken by Voyager 1 (top), enhanced to bring out various color tones and (bottom) taken from the Galileo spacecraft that reached the planet 18 years later which affords a closer approximation to the true color of this planet's atmospheric exterior. This, and other imagery, are shown in the Wikepedia website overview of Jupiter.



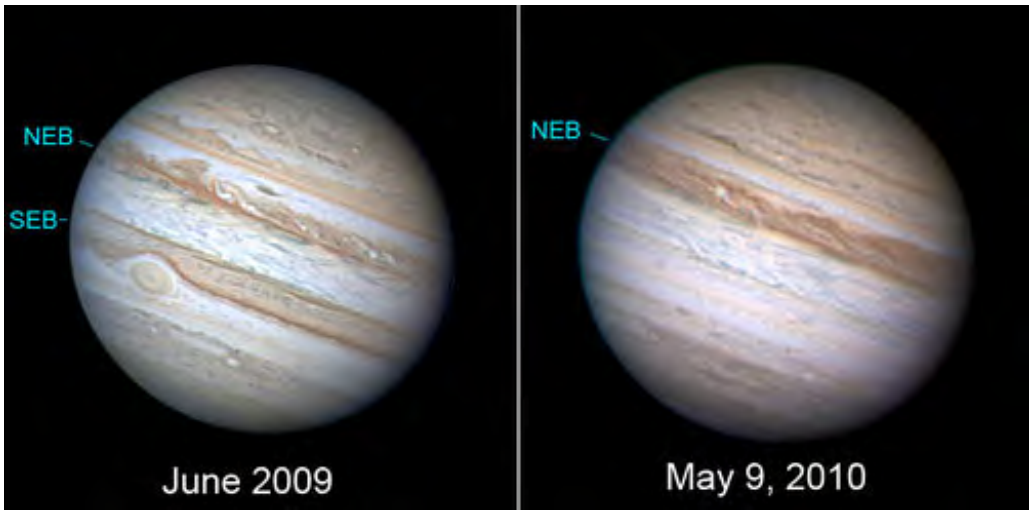
In one sense, Jupiter is the most important of the solar planets in that its size and mass have played a major role in the development and survival of the inner (terrestrial) planets. It does two vital things: its gravitational action on these planets have slowed their inward progression towards the Sun and it helps to keep most large asteroids from continually bombarding the inner planets.

The jovian atmosphere is approximately 75% hydrogen and 24% helium by mass, with the remaining 1% of the mass consisting of other elements. The interior contains denser materials such that the distribution is roughly 71% hydrogen, 24% helium and 5% other elements by mass. The atmosphere contains trace amounts of methane, water vapor, ammonia, and silicon-based compounds. There are also traces of carbon, ethane, hydrogen sulfide, neon, oxygen, phosphine, and sulfur. The outermost layer of the atmosphere contains crystals of frozen ammonia. Through infrared and ultraviolet measurements, trace amounts of benzene and other hydrocarbons have also been found.

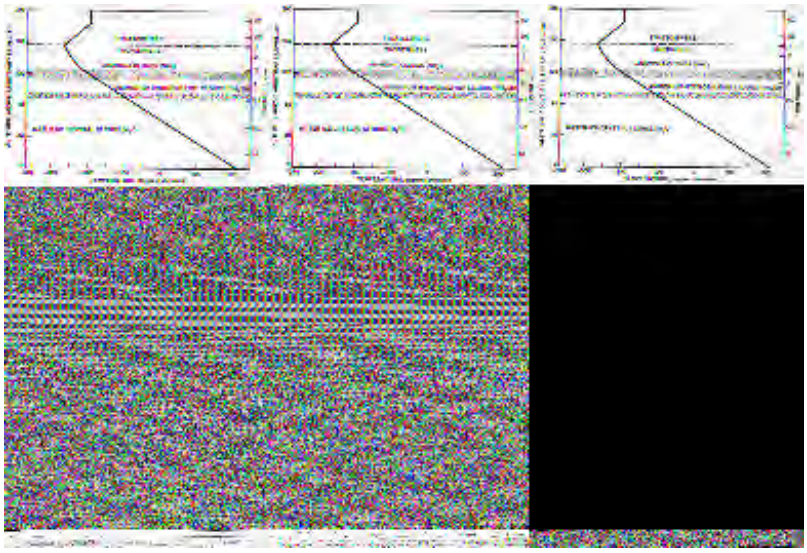
The most obvious features of Jupiter's gaseous surface are the colored bands (from chemical reactions), in which the darker reddish-browns are called *belts* (low pressure atmospheric motion) (temperatures around -147° C) and the lighter, more bluish are *zones* (higher pressure) (temperatures about -154° C). Belts (atmosphere rises) and zones (atmosphere sinks) also move in opposite directions. The orange, red, and brown coloration in the clouds of Jupiter are caused by upwelling compounds that change color when they are exposed to ultraviolet light from the Sun. The exact makeup remains uncertain, but the substances are believed to be phosphorus, sulfur or possibly hydrocarbons. Nomenclature for the belts and zones are given in this illustration (from Wikipedia).



These belts and zones are not stable and permanent - they can change rather abruptly but usually are restored over time. In May of 2010, the large Southern Equatorial Belt almost vanished, as seen in the comparison images below. The cause is uncertain but one hypothesis is that whitish clouds form from time to time, obscuring the dark belt below. Another is that the color fades owing to chemical changes.

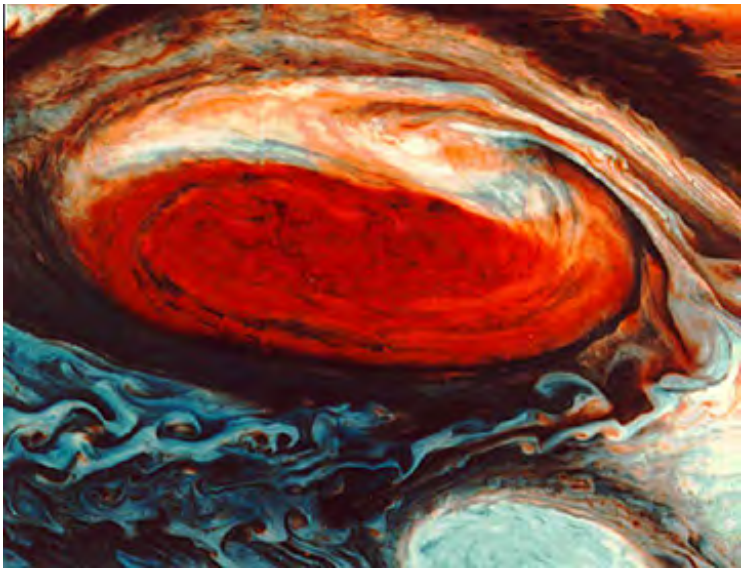


In Jupiter's northern hemisphere, jet winds at velocities up to 600 km/hr (420 mph) blow westward south of zones and north of belts and eastward north of zones and south of belts; this is reversed in the southern hemisphere. Global cloud decks include an outer layer of ammonia ice crystals, above a layer containing small amounts of ammonium hydrosulphide (NH_4SH), and possibly a water ice crystal layer beneath. Upwelling in the belts may convert the NH_4SH (perhaps including also phosphorus compounds) into the darker reds by photochemical reactions with sunlight. This profile diagram describes the pressure-temperature-composition relations in the jovian atmosphere:

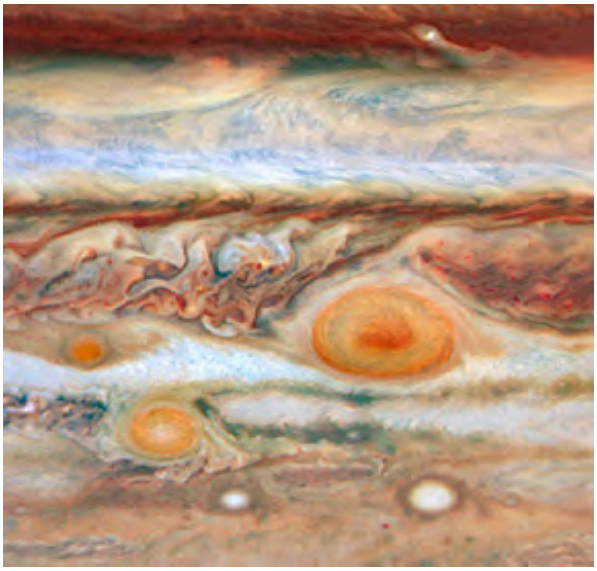


A hallmark of Jupiter's atmosphere is the prominent Great Red Spot (GRS), an oval feature (25000 x 12000 km [15500 x 750 miles]) that has persisted at least since its discovery by telescope in the 17th century. The GRS is a high pressure disturbance caused by thermal convection that carries gases up to 8 km (5 miles) higher than its surroundings, thus cooling them; its anticyclonic winds move counterclockwise (in the southern hemisphere) as it drifts through the bands. The GRS experiences a complete rotation in about 6 days. The false color view below is a Voyager close-up of the GRS in which images made by the red and blue filters are enhanced relative to green which also helps to emphasize some of the swirls and plumes peripheral to it. Red areas are higher than blue above the level of the jovian atmosphere that appears as a quasi-surface. Theories supported by laboratory experiments suppose that the

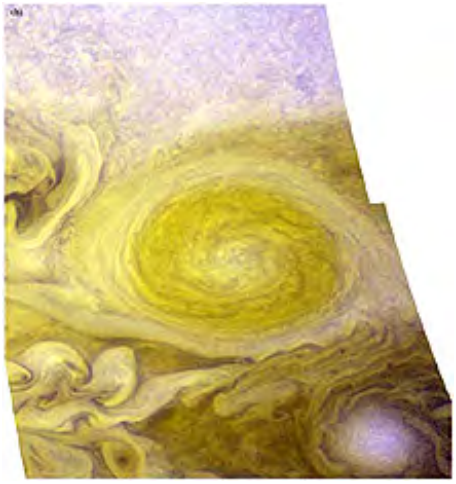
GRS's color may be caused by any of "complex organic molecules, red phosphorus, or yet another sulfur compound", but a consensus has not been reached.



The GRS has existed much as is since first observed by Galileo. More recent are two more Red Spots as seen here (the one at the same latitude as the GRS first appeared in the spring of 2008):

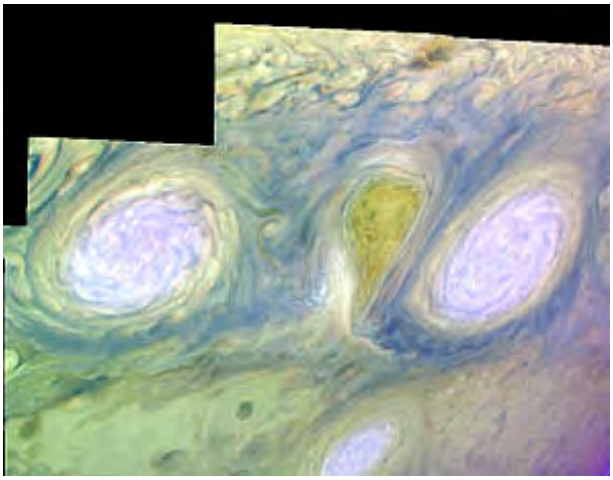


The Red Spot at a more southern latitude has been around since the 1940s (it developed from the merger of several storms). Named the Little Red Spot (LRS)(it does not appear red in the image below) because different wavelengths are involved), the swirling patterns of individual curve cloud streaks indicate fast motion of the gases; a maximum wind speed of 524 kph (385 mph) has been measured:

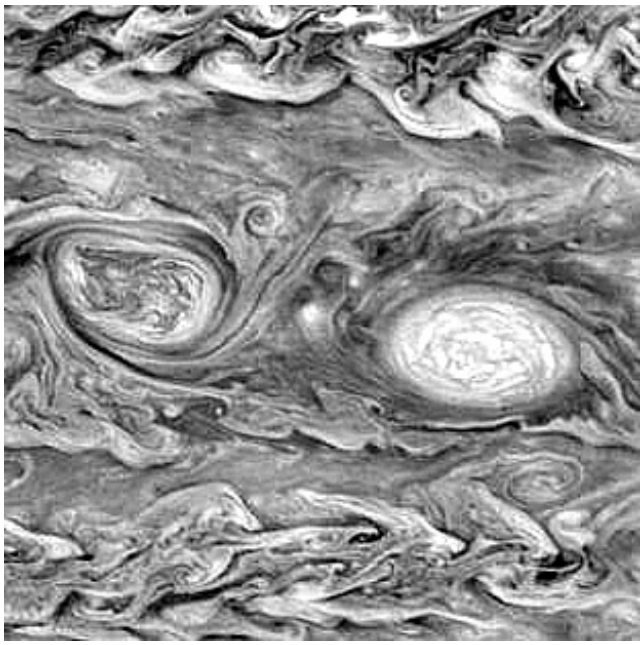


But other cyclonic storms can have much shorter durations. Sometimes two storm spots within a band will close the gap between them and even merge, as shown in this sequence:

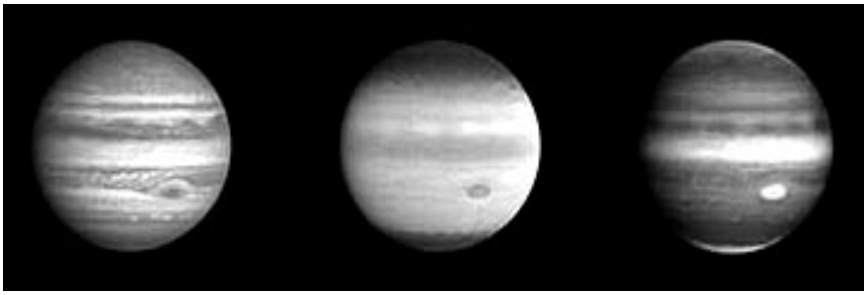
Here is a color version of two of the circulating features known as White Ovals:



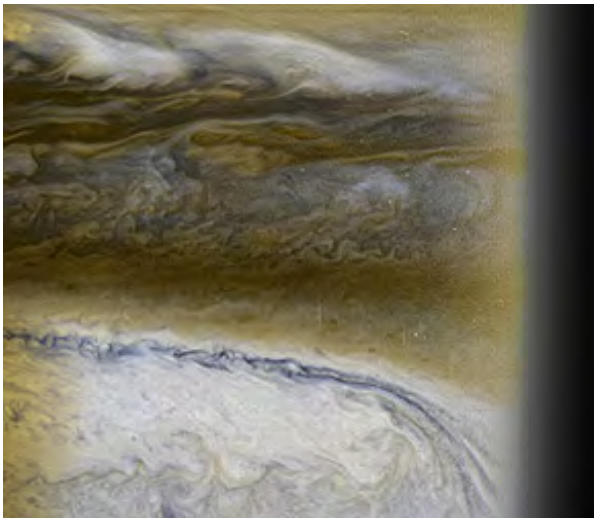
The Galileo spacecraft has returned numerous images showing details in the circulation patterns of individual bands. Here is one typical example:



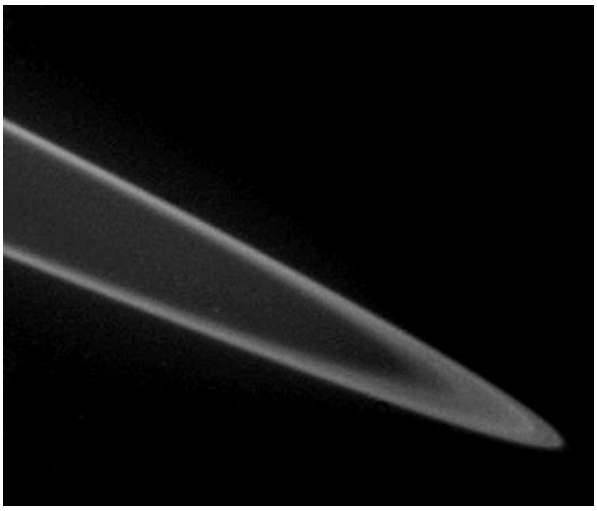
Multispectral imaging of Jupiter helps to reveal variations in the compositional banding and the appearance of cyclonic spots. In this next view, the three images were made in the Visible Blue (left), the Ultraviolet, and the Near IR (right)



In February, 2007, the New Horizons spacecraft passed Jupiter enroute to Pluto (see page 19-21). It proved the quality of its instrumentation with excellent looks at the planet and its moons. Here is a closeup of the atmosphere; note that a thin visible cross-section of the outer atmosphere shows at the limb.



One of Voyager's discoveries is that Jupiter has a ring system. It is more nebulous than Saturn's rings, as evident in this image.

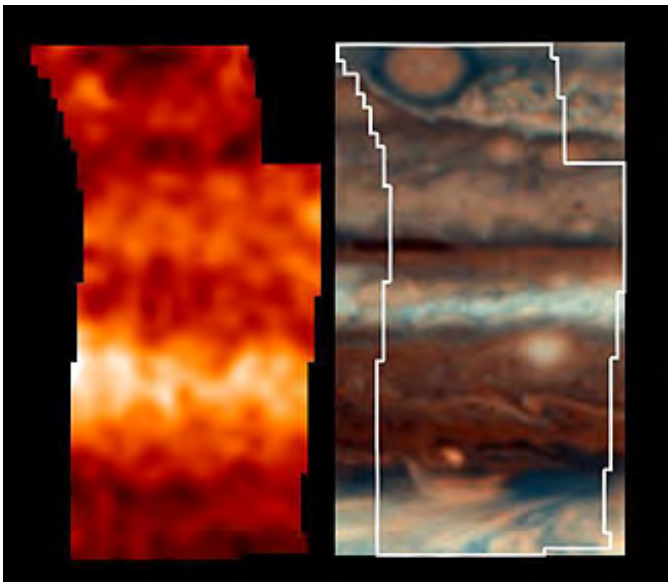


The ring's inner edge begins at about 123,000 km (76,432 mi) from the planet's center. Its fine particles, still effective as sunlight reflectors, form a continuous band about 5,200 km (3,231 mi) wide and 30 km (18.6 mi) thick that shows no conspicuous gaps.

After the Voyagers, Jupiter has been revisited in the '90s. The primary objective for the Galileo mission was to penetrate the jovian atmosphere. This spacecraft, launched on October 13, 1989, separated its probe from the Orbiter on December 7, 1995. The probe's initial velocity on entering was about 106,000 km/hr (65,868 mph), but it then slowed greatly, due to atmospheric friction aided by a deployed parachute. It transmitted data for 57 minutes, as it descended in excess of 150 km (93.2 mi), where the pressure exceeded 24 bars (23.7 atm) at a temperature around 150° C. It encountered no thick, dense clouds, and it detected less sulphur, oxygen, and neon than expected, but found no water. The helium levels were above what was expected. The data overall suggested the probe may have entered at an anomalous hot spot, with abnormal cloud conditions. As it observed Jupiter, the Orbiter saw evidence of large thunderstorms, with rain, and some lightning. The Orbiter also monitored the Galilean satellites, taking close-ups of their surfaces that surpass those of Voyager.

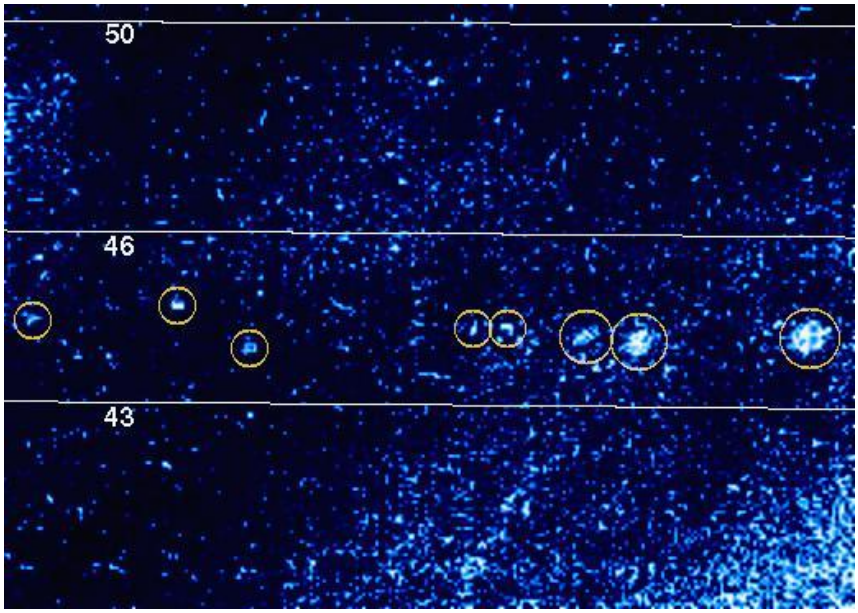
JPL has three brief videos that describe the Galileo mission. Access through the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes "Galileo's Legacy at Jupiter", October 16, 2003, "The Big Spacecraft that Could", February 4, 2003, and "Galileo at Jupiter", May 8, 2002. To start each, once found, click on the blue RealVideo link.

Galileo spent part of its observing time monitoring Jupiter's atmosphere. The bands and convection plumes show temperature differences that broadly correlate with the patterns these atmospheric features assume. This correlation is evident in the Galileo Orbiter view, below, of a segment of bands and spots, in which the temperatures derived from the photopolarimeter divide into warmer (yellow) and somewhat cooler (orange-red to black) zones.

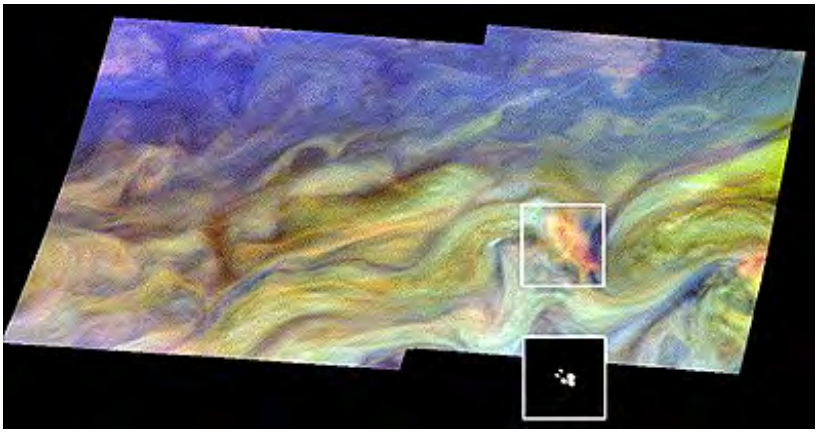


19-53: How does the Great Red Spot differ from hurricanes on Earth? ANSWER

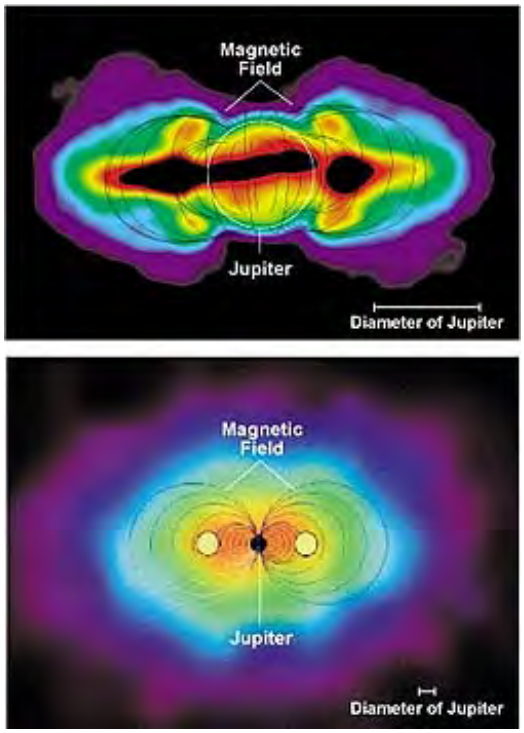
The jovian atmosphere is not only dynamic and turbulent but is electrically charged, as is evident from this next view that shows lightning-like discharges, imaged when the Orbiter is looking at the nightside of Jupiter's outer atmosphere. These are the first visible lightning flashes seen on another planet.



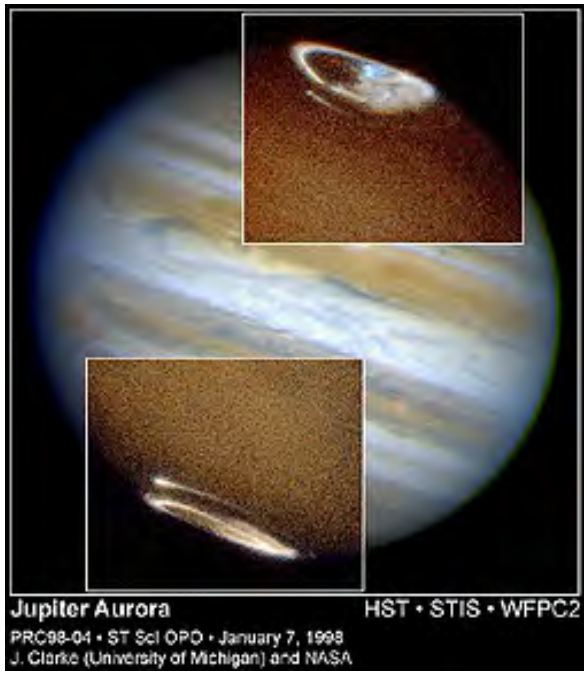
This next image associates lightning strikes (bright marks in lower insert) with a specific, localized storm in the atmosphere.



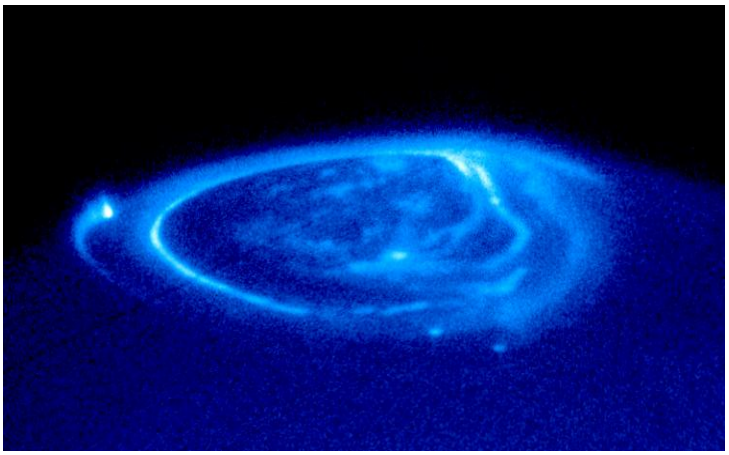
Jupiter has both the strongest (10x that of Earth) and the largest magnetosphere of any solar planet. The magnetic field extends out to at least 100 jovian radii from the planet (reaching to Saturn). If seen in real size from the Earth, in which Jupiter appears at its normal size, the field would cover space on the celestial sphere equal to almost three times the full Moon. Here are two plots of the magnetosphere: one using radiowaves to visualize the field; the lower one constructed from data gathered by a sensor aboard the Cassini spacecraft:



The Orbiter also observed aurora effects as rings at the north and south magnetic poles. The view below shows a Galileo visible image of the planet, on which are two superimposed insets made with the UV camera on the spacecraft.

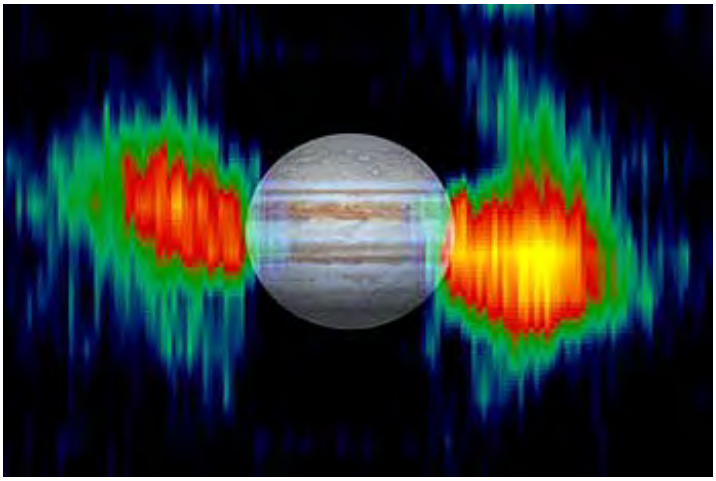


This is a closer view of the northern aurora in UV light:



Ultraviolet auroral light comes from emissions resulting from excited ions and electrons emanating from the innermost satellite, Io, released during volcanic activity (see the next page) that Jupiter's magnetic field captured and concentrated in the polar regions.

Jupiter is an emitter of strong radio waves (peaking at 10 cm), caused by both thermal and non-thermal processes. These waves come from the Van Allen-like radiation belts around the planet. The belts trap electrons, producing a field of 400000 rad strength (enough to kill any astronaut caught within it). Here is a plot of the detected radiation belt around Jupiter:



Jupiter continues to be a target of additional imaging and data gathering. The Cassini spacecraft, enroute to Saturn for a rendezvous in 2004, passed by Jupiter on December 30, 2000 and continues to acquire information. Although the images are not much different from those obtained by Galileo, they do extend the monitoring and add more instrumental analysis. Here is a view of part of the jovian surface and the satellite Io; this helps to realize how small are these "moons" relative to the huge size of the planet.

Now, on to the four fascinating major satellites of Jupiter, plus a look at some smaller moons.

PREVIOUS NEXT

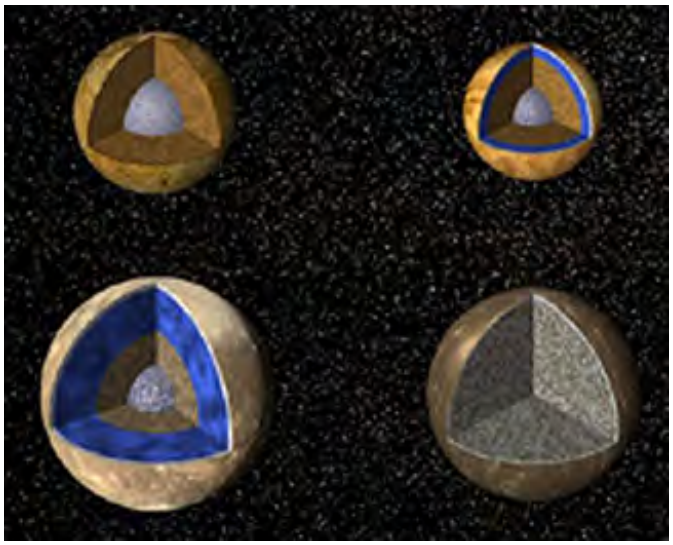
Primary Author: Nicholas M. Short, Sr.

The two inner jovian satellites - Io and Europa - are among the most interesting planetlike bodies in the Solar System. Io has a hot mantle that continues even today to send volcanic materials to the surface; its closeness to parent Jupiter leads to strong tidal forces that help to account for this activity. For its size, Io has more active volcanoes, some spewing sulphur compounds, than even Earth, the only other spherical body in the S.S. on which current volcanism can be monitored. Europa is quite different: it has a crust largely composed of ice, mixed with rock. Below that cracked crust there is mounting evidence of a fluid shell - probably of water - of moderate thickness which has an analog in the ice-covered Arctic of our Earth where ocean water lies beneath. This suggests a possibility, which NASA hopes to test, of a medium in which some forms of marine life may have developed.

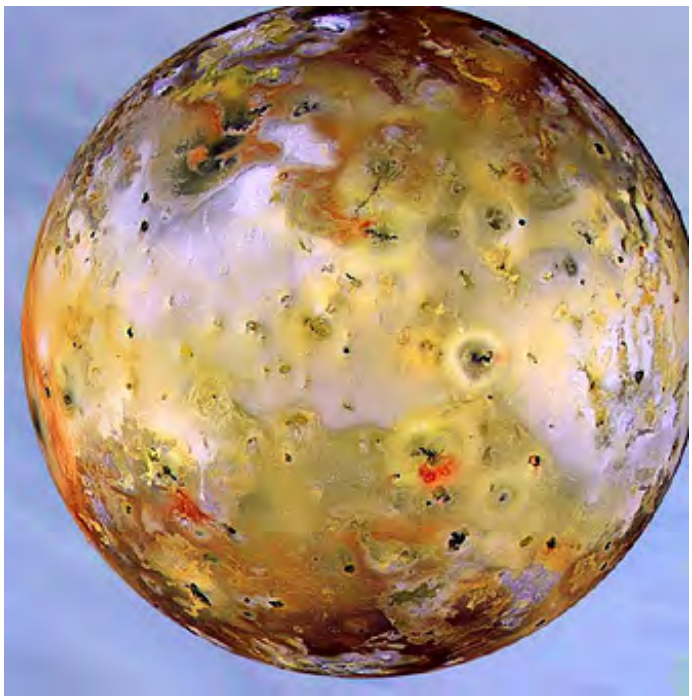
Jupiter's Galilean Satellites: Io and Europa

Jupiter has 47 named satellites (we use the term "moon" to describe spherical satellites while "satellite" may apply to an irregularly-shaped body). All but four are small and irregular, and most of those were discovered by telescope from Earth, but Voyager found three. We generally consider the four large ones, known as the Galilean Moons, in honor of the Italian scientist, Galileo, as among the most extraordinary, as a group, in the Solar System. We introduce these now in the Voyager montage below, created from individual global (full) views scaled, so that each retains its size, relative to the others (see caption for identification):

Data gathered both by the Voyager and Galileo spacecraft on the densities and nature of the crusts of these four Moons have resulted in general models of the interior of each. Io (upper left), Europa (upper right), Ganymede (lower left) all have cores (denser rock; possibly some solid metal); Callisto may have a small core but its interior is thought to be mainly water ice plus rock (largely undifferentiated).

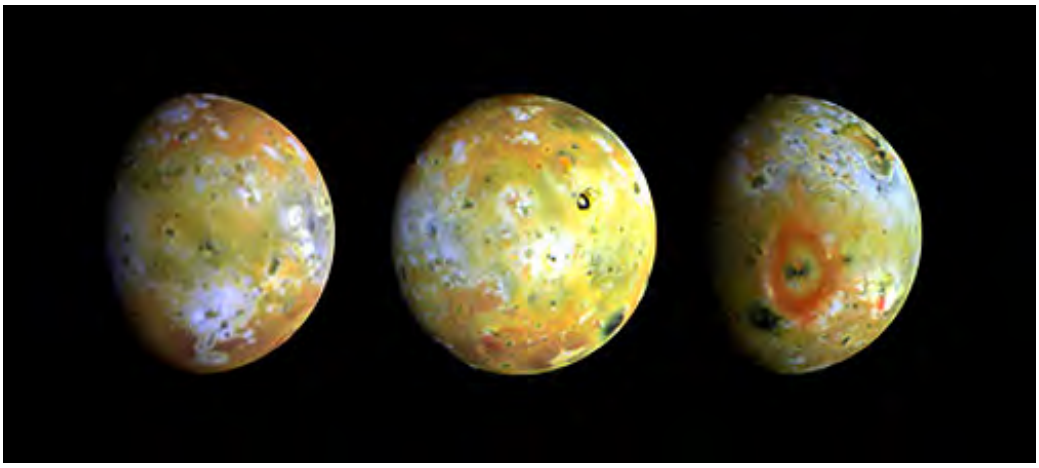


The innermost moon, Io, lies some 422,000 km (262,231 mi) from Jupiter's center. It's about 3,630 km (2256 mi) in diameter, making it just slightly larger than Earth's Moon. (This particular image has reminded some planetologists of a "pizza".)

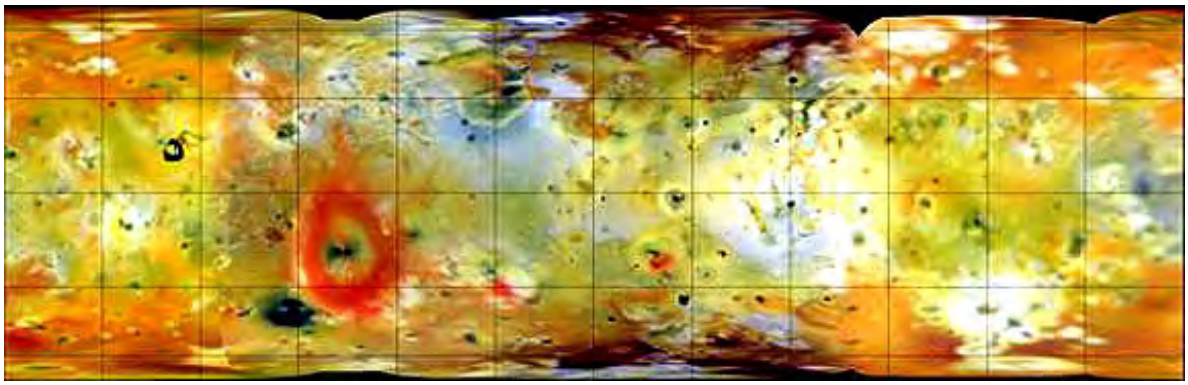


For a quick summary of what space observations by these probes have told us about Io, check the JPL movies listed here. Access through the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes "Galileo's Farewell to Io", January 15, 2002, "Io Flyby", August 5, 2002, and "Io: A Volcanic Inferno", August 2, 2001. To start any one (they tend to be redundant), once found, click on the blue RealVideo link.

Different parts of Io are portrayed in these three views taken by the Galileo TV camera:



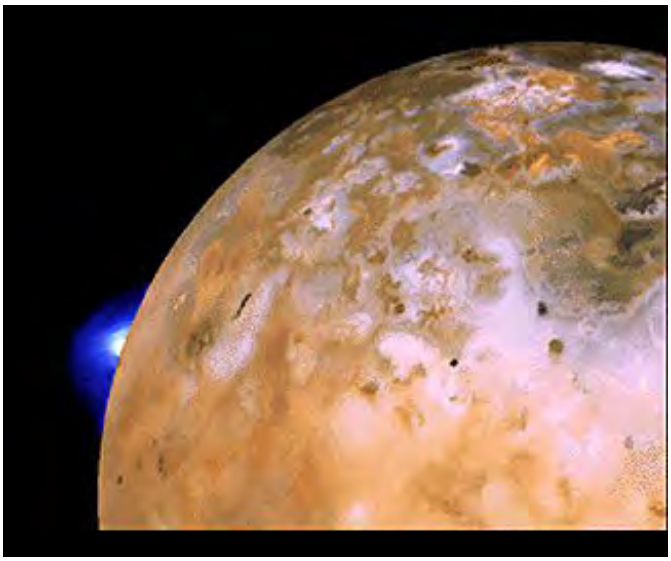
The ionian surface can also be displayed in a flat projection (which compresses and distorts polar regions):



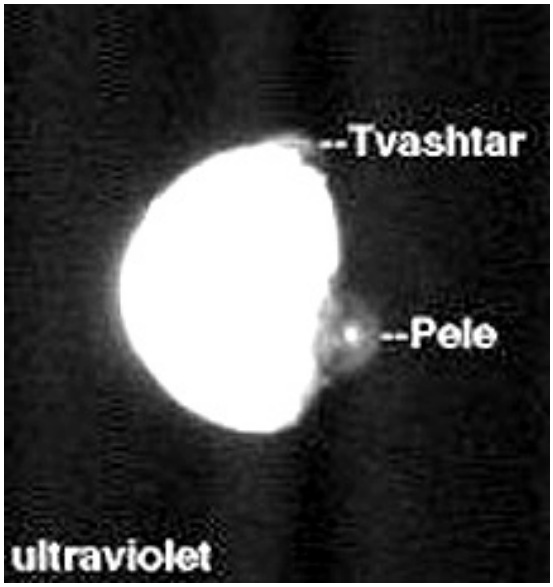
19-54: What common planetary feature seems to be absent on Io? Venture a guess as to what may be the dominant process acting on this satellite. ANSWER

Io displays prominent orange colors, with blackish mottled areas (that has reminded some of a giant pizza with olives). It may well be the most remarkable of all satellites, in that the surface is undergoing constant dynamic change, largely through volcanic processes. Astrogeologists have rated Io as the "most geologically active planetary body in the Solar System", roughly 100x that of Earth on a surficial area basis. What they mean is that it is undergoing more volcanic eruptions, for its size, than any other planetary body (at least 16 volcanoes that are active have been detected on Io); besides Earth, it is the only other object with live volcanoes spewing hot lavas. Although the ionian surface generally has a temperature of -140° C, hot spots with temperatures up to 1600° C (2900° F) attest to volcanically active areas. Io's surface is geologically young (no impact craters have been found on it), as it is constantly undergoing "resurfacing" by volcanic processes.

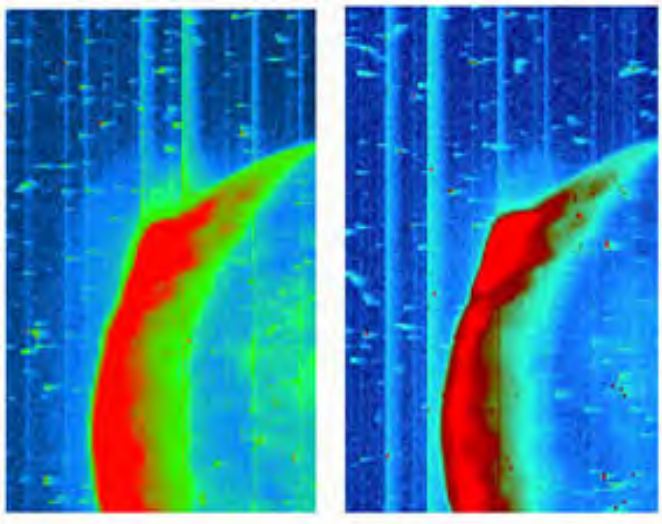
One Voyager 1 image startled JPL scientists, because it captured a major plume from the eruption of an ionian volcano that was then tossing material up to 300 km (186 miles) above its surface. This was the first clear indication that Io was a body with (very) active volcanism.



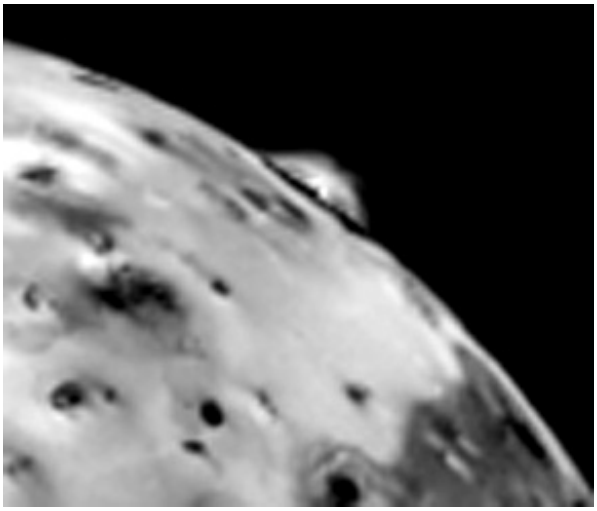
The plumes are particularly likely to be detected when their source volcanoes lie during rotation on or near the limb of the illuminated/dark parts of the satellite. In this UV image made by the Cassini spacecraft, two plumes from separated volcanoes are visible against the dark background just past the limb terminator. These eruptive plumes are nearly symmetric hemispheres because of lack of strong winds to disperse the ejected materials directionally.

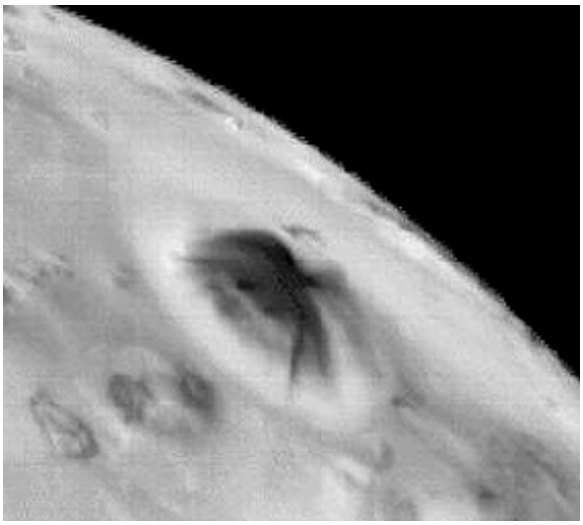


In August 2001, Galileo had a stroke of good fortune: An unknown volcanic source was caught in eruption while the spacecraft was actually looking at a different volcanic site, Tvashtar, whose activity had been persistent earlier. This time Tvashtar was apparently quiescent but a plume, imaged in the infrared, was observed along the illuminated crescent as it rose to 300 km (200 miles) above Io's surface at a point that had not earlier disclosed any discrete volcanic features. This is how this eruption appears in this observational mode; most of the excited particles causing the red color are SO_2.



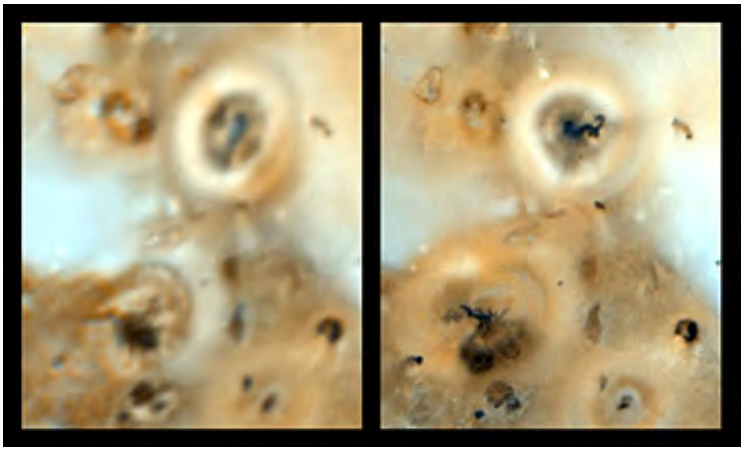
Other eruptions were "caught in the act" during the fly-bys meaning that, along with signs of young surface deposits, for its size Io is more volcanically active than Earth. A major eruption in progress at a volcano named Prometheus was captured by the Galileo TV camera, as depicted in these two views:





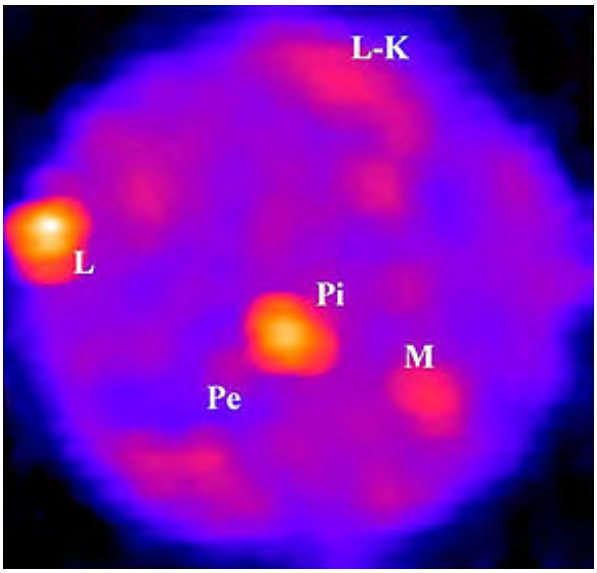
19-55: Look at the Prometheus images. How does this eruption differ from the usual big ones (typically from stratovolcanoes) on Earth? ANSWER

By comparing images of Prometheus acquired by Voyager in 1979 (left) and Galileo in 1997 (right), several distinguishable major changes are evident:



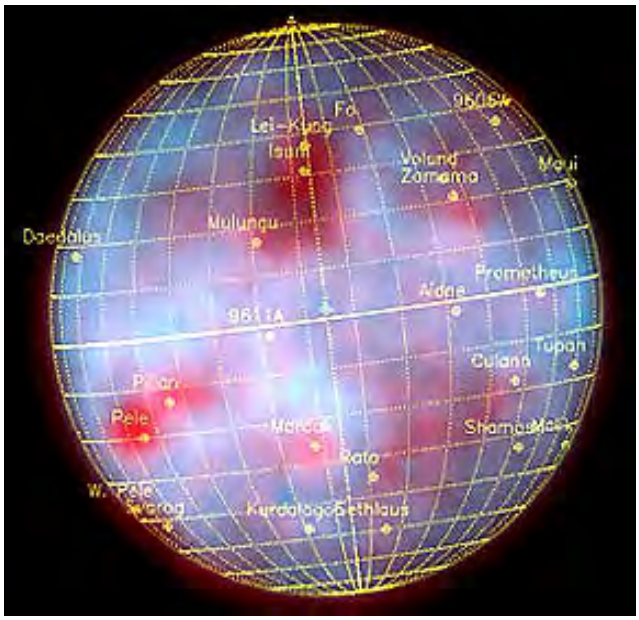
The caldera central to the Pele eruption site was imaged after it had discharged sulphurous lava, ash and gases out several hundred kilometers in all directions, forming a heart-shaped halo deposit (SO_2 frost), named the Matuike Patera (volcanic field).

A thermal sensor on Galileo has taken measurements showing temperatures associated with specific volcanic hot spots (see caption for identities of individual volcanic centers):



The actual temperatures are mostly low, by Earth volcano standards. The blue denotes temperatures of 90° K (-297° F) and the yellow denotes 170° K (-153° F). But some hot spots have local temperatures up to 1500° K (2240° F).

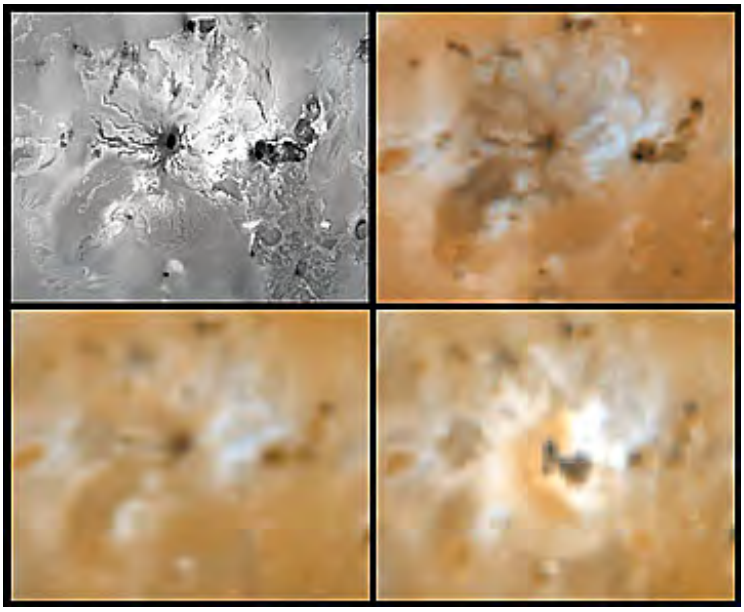
Another part of Io, overlapping onto the Pele area, has been examined in the infrared by the European Southern Observatory (an Earth-based telescope). The image below was made by using these bands: 1) Blue = 1.6 µm; 2) Green = 2.2 µm; 3) Red = 3.8 µm. The last band singles out areas on Io that are notably warmer than their surroundings.



Galileo has imaged an active or recent lava flow, still hot as suggested by its orange color. This is the first time a "living" flow has been seen on any other planetary body:

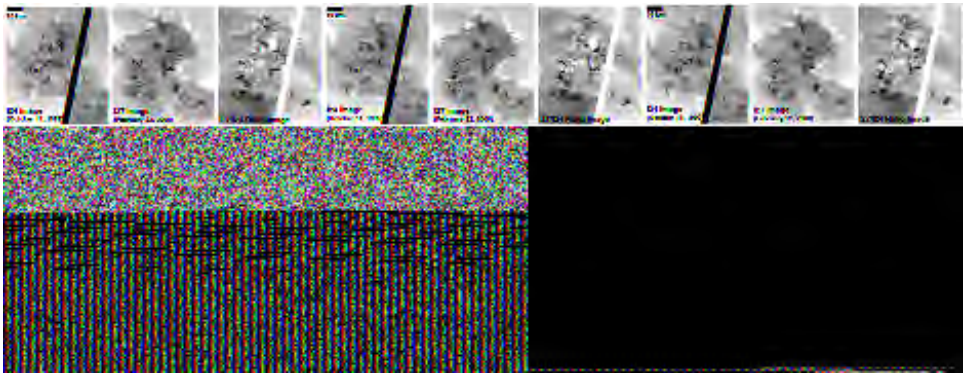
Some of the Galileo images, when converted to approximations of natural color but with some artistic license, are quite attractive to behold, as is this one of Culann Patera:

At least 100 eruption sites have been identified, but many may be inactive. More than 30 appeared to undergo one or more events during the observation periods. Volcanic deposits dominate the entire ionian surface, some being flows and others, probably fragmental. We can trace many of these to calderas. Volcanic sites can undergo significant modifications in just a few years, as demonstrated by this series of images taken over the last 18 years from Voyager 1 (upper right) through Galileo (lower right) around the volcanic complex known as Ra Patera (upper left):

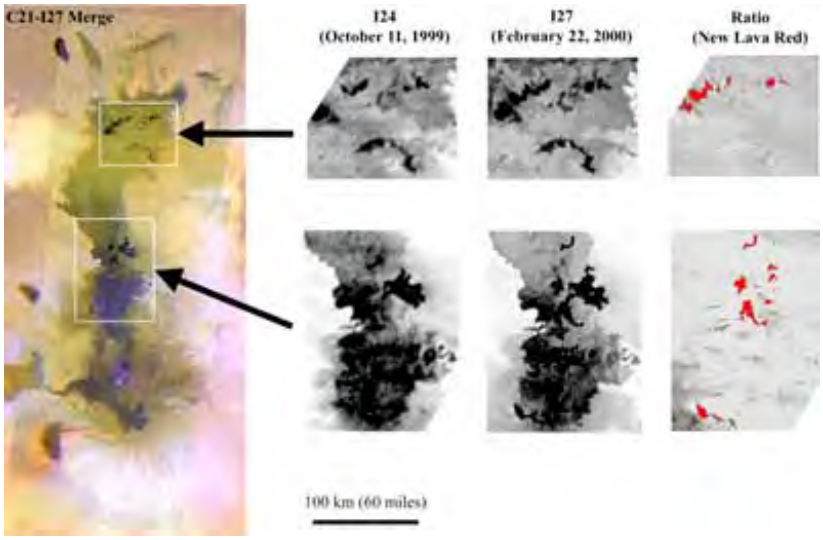


19-56: In general terms, describe the changes at Ra Patera. ANSWER

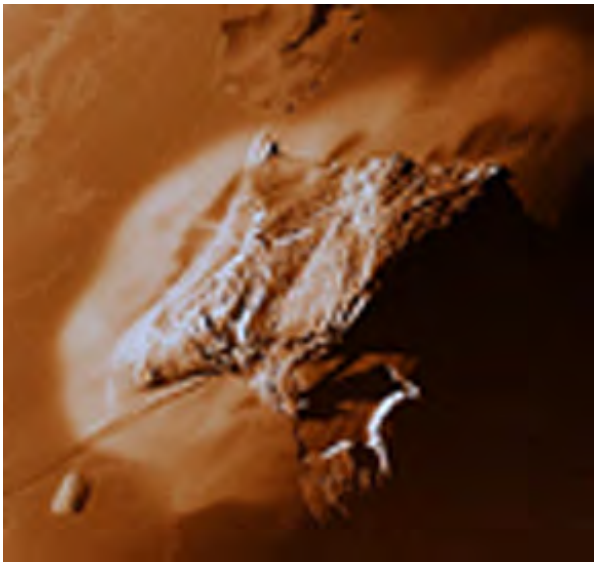
Standard image processing for enhancement or change detection that we've used in this Tutorial for terrestrial images work just as well for other planetary bodies. Here are three Galileo images of an active volcanic terrain on Io. The left and center were taken four months apart. The right image is a ratio product of the first divided by the second date; the new tonal patterns indicate changes most probably of volcanic deposits over this short interval.



An even more recent observation has pinpointed the volcanic complex called Amirani-Maui where new lava is pouring out at a rate of 100 cubic meters/hr. This is producing a lava flow continuum that is at least 350 km (220 miles) long (longer than any equivalent on Earth). The images below, on the right, show areas of new lava cover (red patches, determined by image differencing) built up in just 4 1/2 months.



Most of the ionian volcanoes do not appear to be high stratocones such as are displayed as the majestic volcanic peaks observed in the Ring of Fire on Earth. Where some relief is evident, many Ionian volcanoes are more like terrestrial shield volcanoes, often with wide calderas. Haemus Mons is an example:

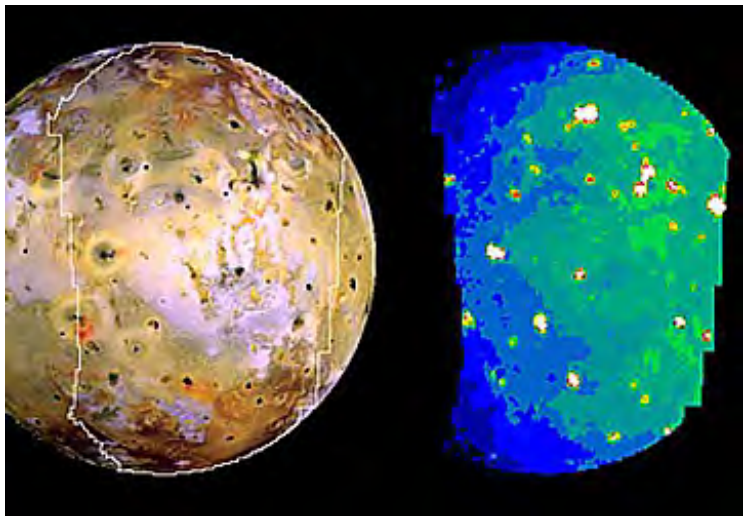


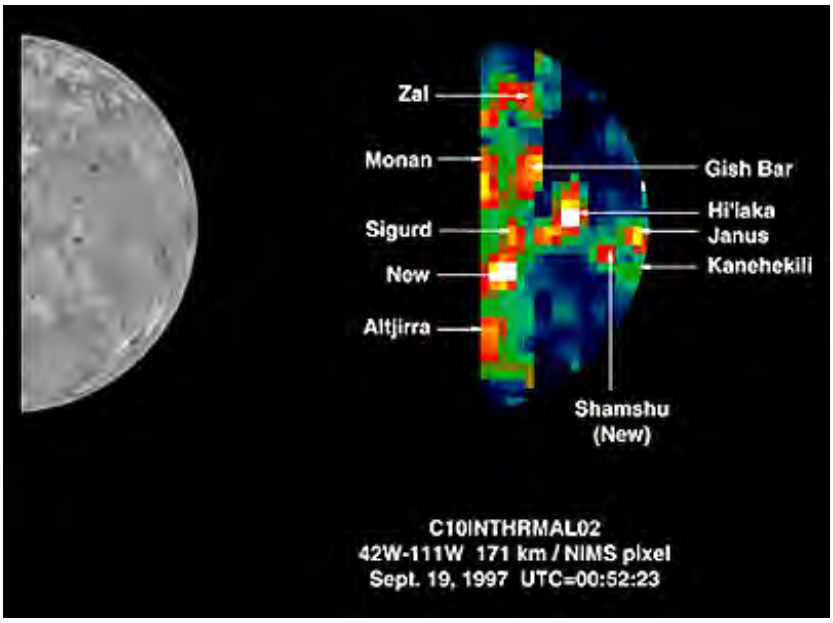
The volcanic materials are likely silicates, mixed with large amounts of (native or pure?) sulphur and sulphur compounds. These compounds can account for Io's orange, yellow, and black colors

that are typical of this element. Sulphur-rich lakes have been observed. Sulphur dioxide makes up a tenuous atmosphere.

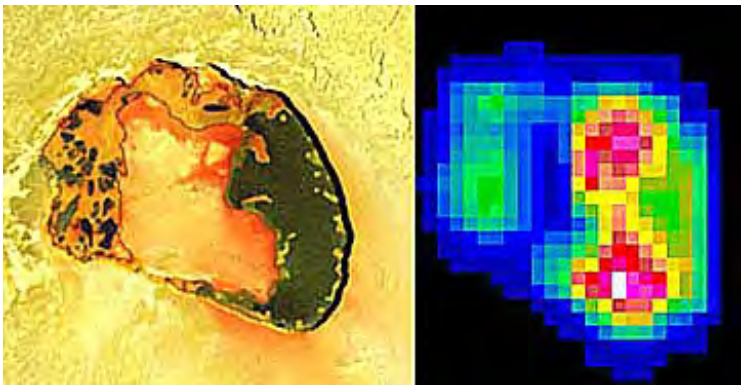
The cause of this (totally unexpected) activity on Io is a process termed "tidal pumping," in which the outer part of this satellite flexes and stretches as much as 100 m (328 ft) as Jupiter and the other moons tug on it. The tidal energies acting on Io induce heating, and the stretching causes it to wobble. This conversion of tidal energy to thermal energy is also cited as the principal energy source for activities that may be continuing today on Europa and Ganymede, since both of these satellite display signs of surficial changes that appear young.

Io has the *highest degree of volcanic activity* of any spherical (planet or planetary-like) body in the Solar System. Some of this activity was missed during the earlier imaging of Io in the visible light range. But thermal infrared images have found at a large number of hot spots (at least some being volcanoes) representing new locations; later inspection of visible imagery then found evidence for most of these being associated with volcanism. Two examples illustrate this:





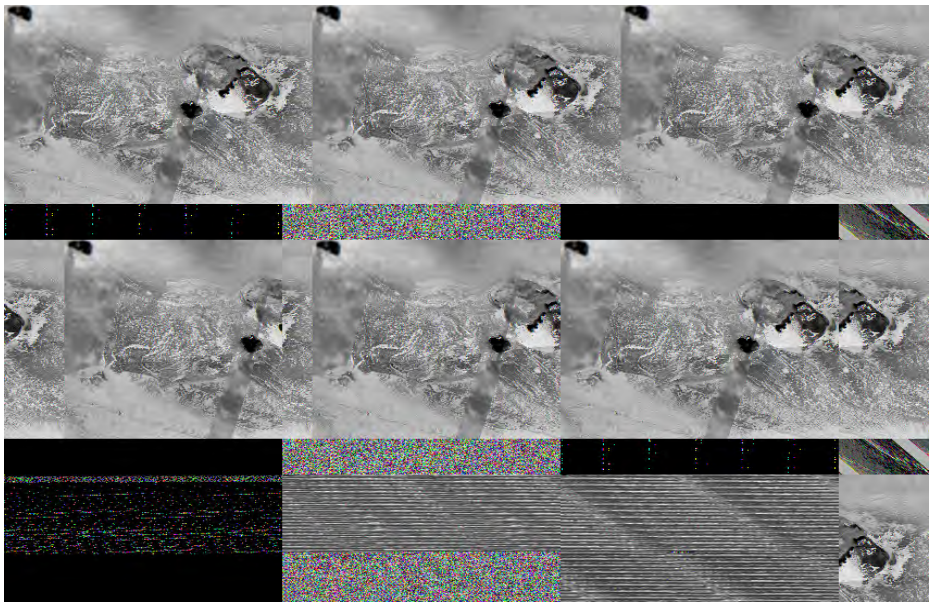
An image made during an October, 2001 Galileo pass by Io shows the Tupan caldera in a false color rendition and in a thermal plot using data from the 4.7 µm thermal channel:



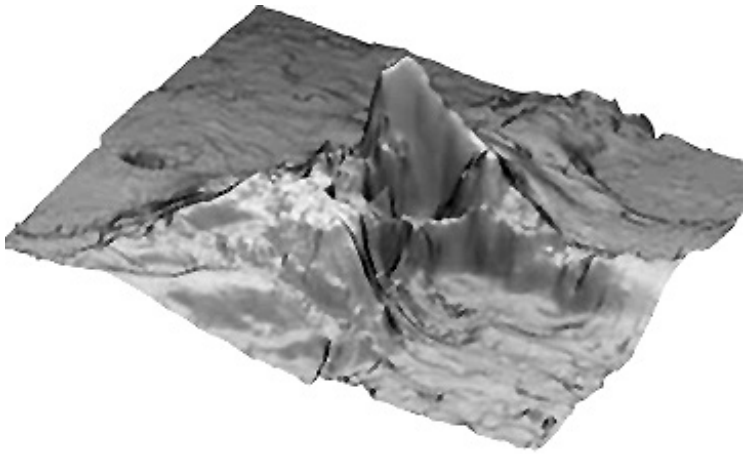
The present ionian surface is now believed to be quite young since lava flows like those shown above are commonplace. With such activity, the surface is continually being replated with lava cover. Some planetary geologists contend that Io's surface today is similar to that of the early Earth, in its first billion years, and is thus a "living" model of terrestrial conditions not long after its formative years (involving intense planetesimal bombardment and cratering) came to a close and a continuous crust developed. However, the Earth of about 800,000 years of age probably didn't have a sulphur-rich crust.

Amazingly, Io also has the highest mountains on any of the rocky or ice-covered satellites, although these are uncommon. Using both Voyager and Galileo images, stereo viewing has allowed estimates of elevations of various peaks rising from the general surface. Some exceed 10 km (6 miles) in height (relief). One of the more prominent is Tohil Mons, a volcano with a caldera and associated lava flows, that resembles similar volcanoes on Earth:

This mosaic shows the region around Tohil Mons, with its sets of what appear to be individual lava flows:



This next image is a perspective re-creation (with added vertical exaggeration) from the above imagery of Tohil Mons; note that the viewing direction is different. The peak is more than 6 km (20000 ft) above the plains.

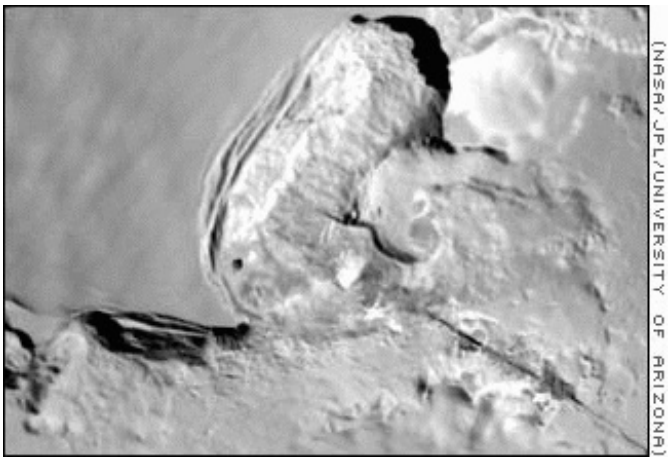


Viewed close up, this mountain (~6150 m or 18500 ft high) is seen to have a subsidiary rimmed crater (to its northeast in the above view) with a smooth central filling that appears to be a (solidified?) lava lake.

Recently, Galileo sent back evidence that Io has an iron-rich core whose radius is about half (~900 km; 610 miles) that of the complete satellite. Io is generating a small magnetic field that interacts with (and may be largely caused by) the intense field caused by Jupiter. Io also seems to have an ionosphere. It is surrounded by a vapor cloud made up of hydrogen, sodium, potassium, and ionized sulfur (derived by sputtering off the ionian surface) which has been

stretched out to form a torus around its parent planet. A flux tube between Io and Jupiter allows a current of 10 million amps. to flow.

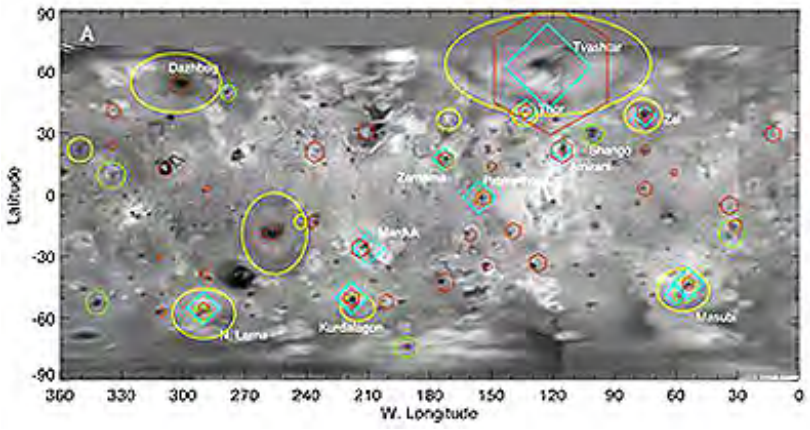
As would be expected on such an active, dynamic planet-like surface, other processes besides volcanism are taking place. This Galileo image shows landslides against the steep front of the Telegonus Plateau:



Io also has an atmosphere made up of gases and dust ejected from the volcanoes (most eventually escaping to space) and ionized particles. This next Galileo image shows the atmosphere (reddish band near top) and volcanoes in eruption.

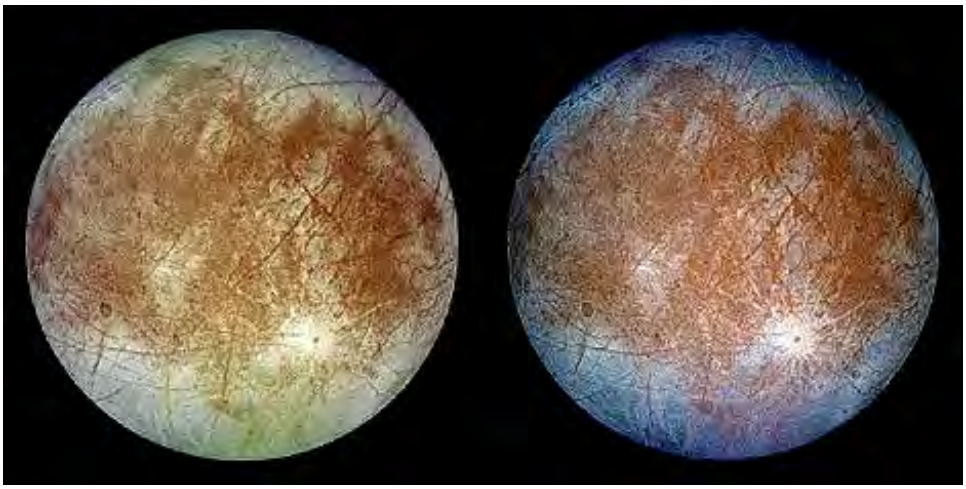
Galileo has made a remarkable image of the presence of ionized sodium in its ionosphere. Sodium ions when excited give off visible light (remember sodium vapor lamps) at 5890 angstroms (or 589 nanometers or 0.589 µm). Using a yellow-green filter, its camera captured this view of Io and surrounding space. The yellow is an approximation of the sodium light. In the Io sphere, the Sun has lit the far side, as seen here, but there is a bright white light on the dark side which is sunlight illuminating the plume of ash coming from a Prometheus eruption in progress.

As mentioned on the previous page, the Pluto-bound New Horizons spacecraft passed near Jupiter in February, 2007. It obtained good views of Io; comparison of surface features in the new imagery with those obtained during Galileo's activity has shown that there are many subtle to sharp changes detected since the earlier passages. Some of these are shown in the image:



The yellow ovals refer to new plume deposits; the green circles to new lava flows; the cyan diamonds to active volcanic plumes; and the red hexagons to volcanic hot spots.

The next moon is Europa (diameter = 3,140 km (1951 miles), 671,000 km (416959 miles) from Jupiter. Below are images (using different filter combinations), as it appears from Galileo:

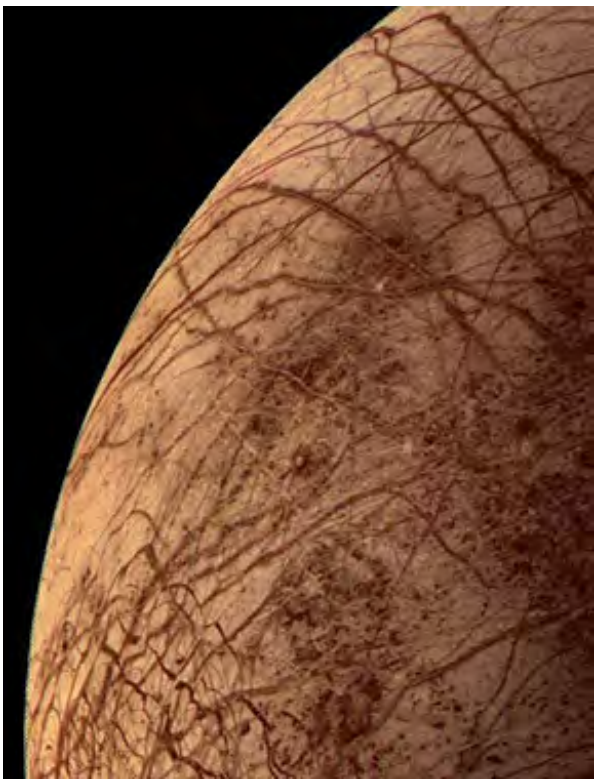


The left view is a natural color image. On the right is a false color image, made by projecting the violet band through blue, the green band through green, and the IR through red.

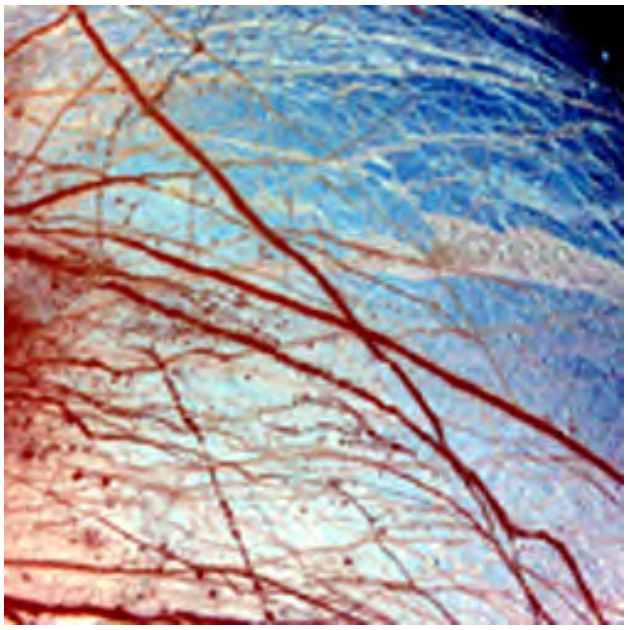
19-57: What Earthly object with which you may be familiar does Europa remind you of (hint: think of a game you play indoors)? ANSWER

Seen by Voyager close up, the surface is extremely smooth, with few markings (minimal craters). Much of the ice making up the european surface is blue, similar to fresh ice found in polar regions or in glaciers on Earth. This is typical:

Other regions on Europa have a brown coloration. These contain pronounced sets of darker brown "fractures" that divide the crust into polygonal segments. These cracks are only slightly higher or lower (<300 m [984 ft]) than the main surface that appears to be water ice, which may top an ice-water slush or a pure water "ocean" at some depth before grading again into ice and then rocky material.

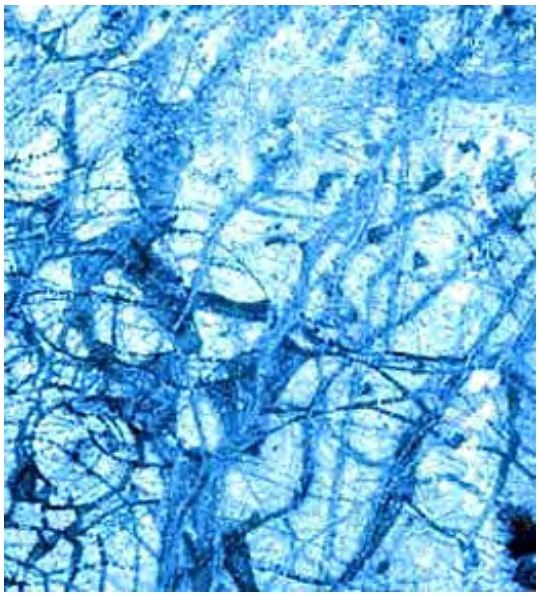


Another scene in the Minos Linea region of this surface viewed from Galileo has been reprocessed to give a color version in which the ice is blue and the crack-filling material is dark red - a color very close to that of hematite, the non-hydrated form of iron oxide (suggestive, but not proof it is that mineral).



Still another eye-appealing color version of a small part of Europa's surface:

At higher resolution, we can see several orders (size discontinuities) of fracturing. Some are distinctly curvilinear (a common pattern in rapidly cooled liquids, such as those used to make silicate glass). Slivers of dark material are scattered within. On Europa, some interior rocky material may mix with the ice.

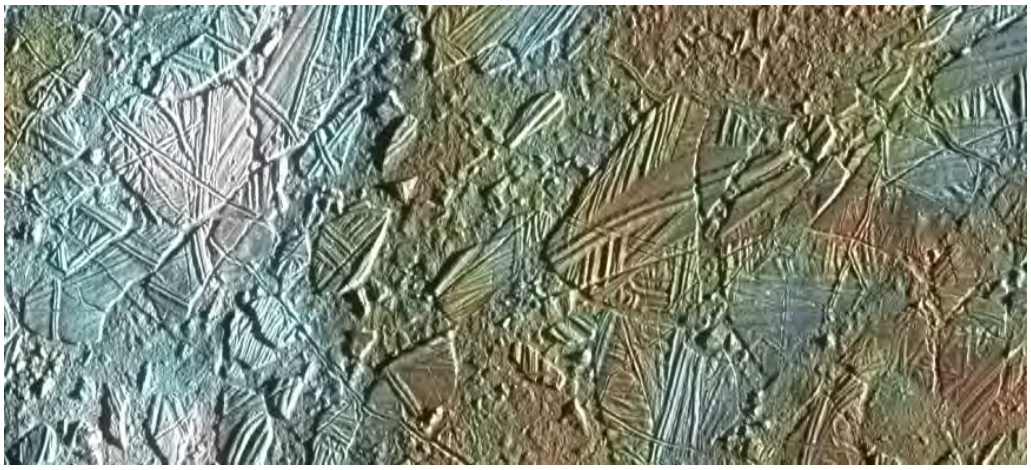


The origin of these linear breaks is uncertain. Scientists compare these breaks with polar sea ice on Earth, which experience cracks or "leads" by tension, as water currents cause them to flex. These cracks then reheal. Ridges, formed by compression, are an alternative.

19-58: On what basis can you conclude that Europa's surface is relatively young?
ANSWER

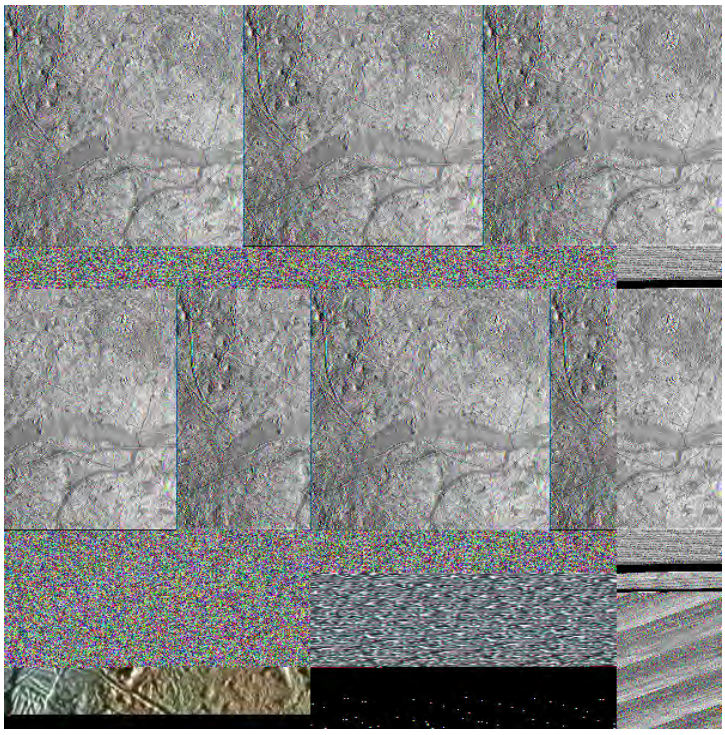
Galileo shed more light on this, when it made a close approach (586 km [364 miles]),on February 20, 1997. The picture below shows closely-spaced parallel cracks, with different orientations, in two blocks separated by a high ridge, furrowed in its center.

This image contains, in addition to cracks, a break-up of the crust that produces what have been called "ice rafts" which rise above the surface. Individual rafts are grooved, but the grooves vary from raft to raft indicating that they floated as breakoffs from grooved terrain or alternatively were reoriented by pushing around from a disrupted terrain.

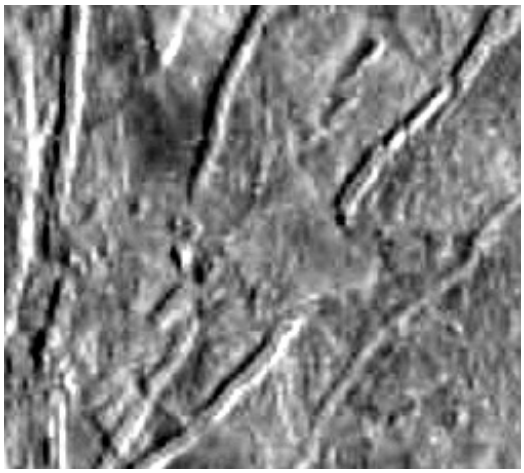


These appear to be break-ups of a thin ice cover (with its leads), in which the rafts are free to move about, with additional evidence favoring a liquid water body below. This is now being referred to as the "buried ocean" - some estimates envision this as a continuous subshell of water that may be as much as 100 km in depth that extends over the entire subsurface of Europa. The implication of a liquid "ocean" topped by a frozen crust (analogous to the polar Arctic on Earth) has led some scientists to speculate that conditions may be favorable for primitive life in this liquid. This supposition has fueled intense interest in getting even closer approaches to Europa, which planners scheduled in the continuing Galileo observations. Officials are considering a new mission in the early 21st century, specifically to search for more signs of life. The mission would attempt to reach the subsurface water (one way: a probe with radioactive material that generates heat, allowing the probe to melt its way through the ice crust).

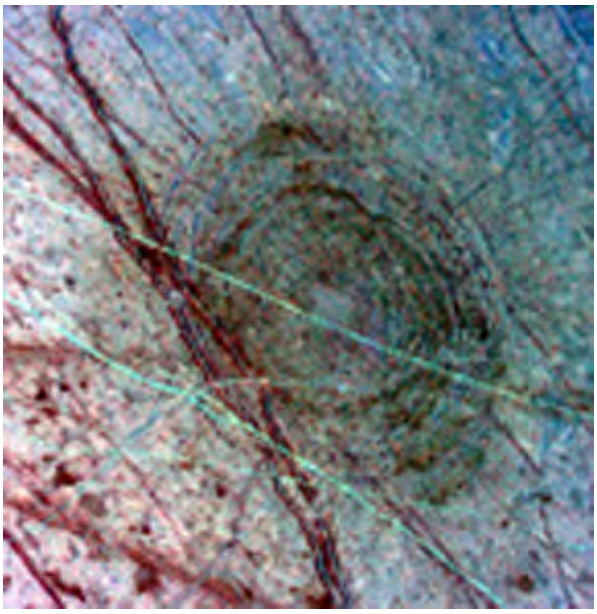
More recent observations shed further light on the variety of terrains on Europa. This image shows what appears to be an ice flow or glacierlike ribbon cutting across the surface:



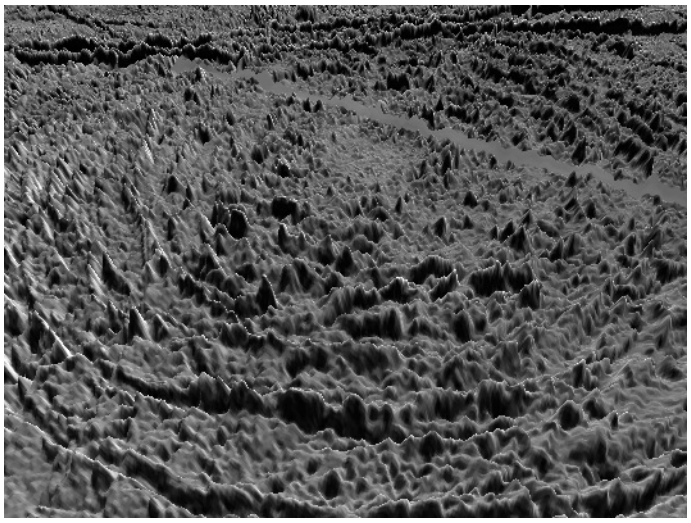
Recently, Europa investigators have proposed the possibility of "ice volcanoes" in which viscous ice (perhaps with some water) erupts from vents and flows like a glacier over the surface. Galileo has imaged such flows, as seen in this image in which a flow (in the center) is extending over large cracks.



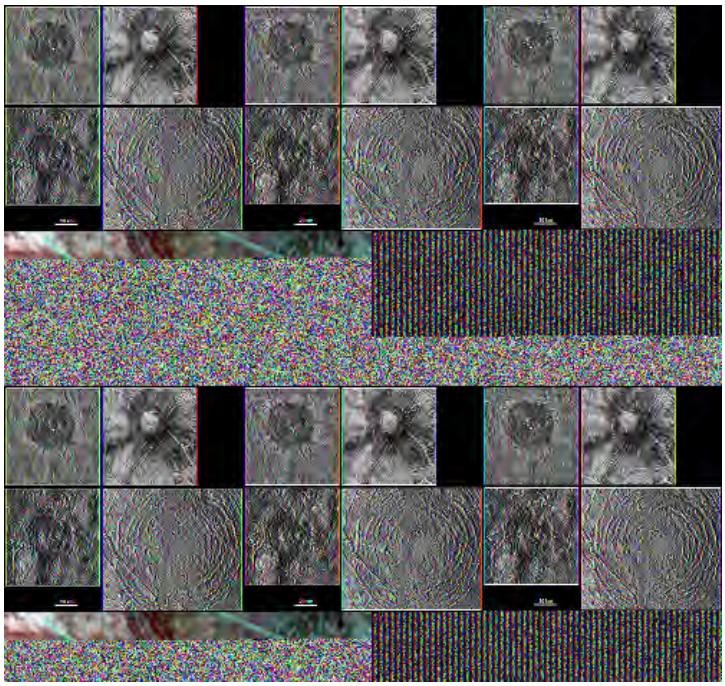
Although impact craters, especially larger ones, are quite uncommon on Europa, a few exist, such as this one, not yet destroyed by ice movements, known as Tyre Crater. This sparsity indicates that Europa is an active planet, with continual or periodic replating of its surface in a process affected by the presumed liquid ocean beneath its ice cover.



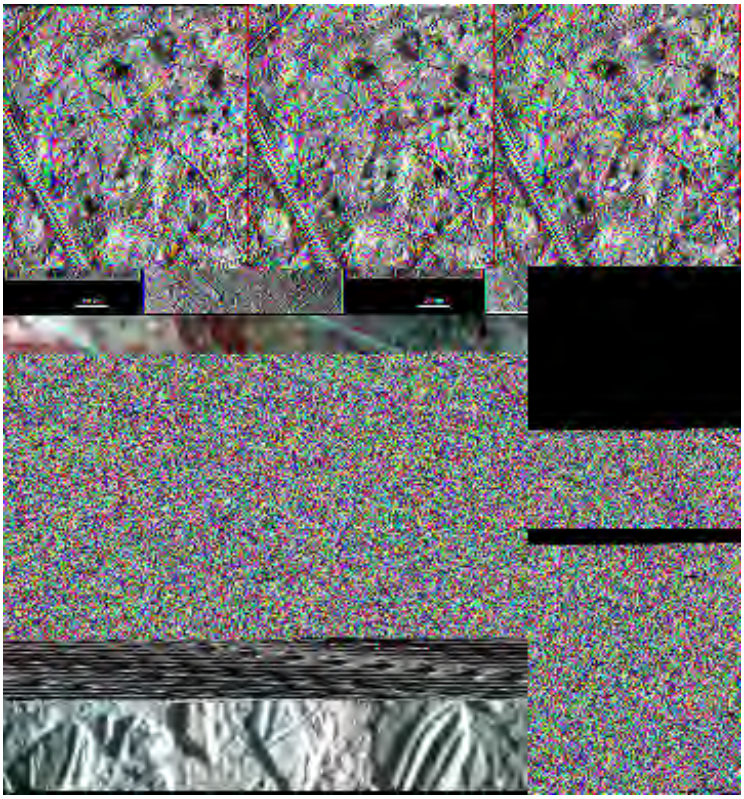
Tyre Crater has numerous rings. These are made more distinct by exaggerating their relief, using a computerized image-processing program



Here are four impact craters on Europa:



This last view (Galileo) is again typical of the european surface, with fractures and grooves, ridges and small domes (up to 8 km [5 miles] in diameter) that are believed to represent salty liquid pushing upward into the crust which may be as much as 13 km (8 miles) thick.



Now on to the remaining two galilean moons of Jupiter.

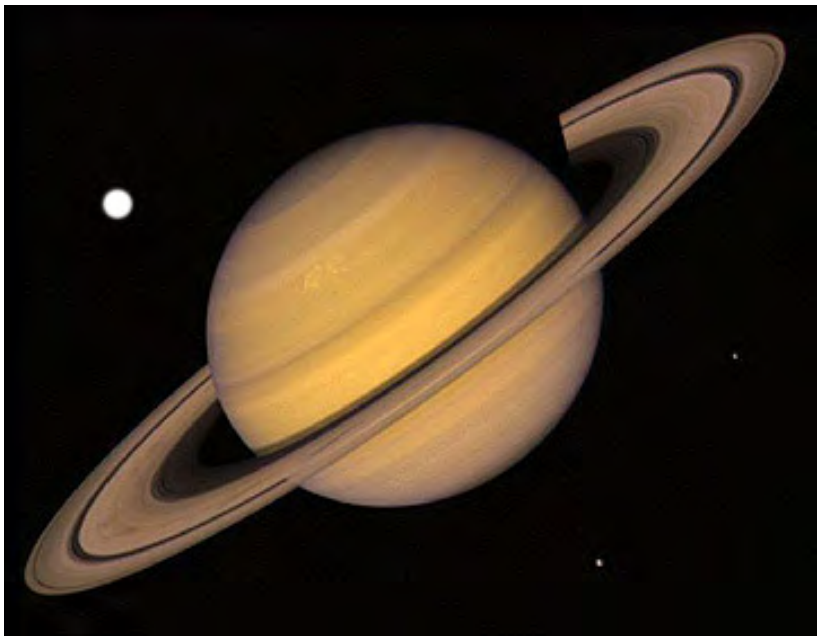
PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Saturn, somewhat smaller than Jupiter, is notorious for its resplendent rings, which can even be seen in the telescopes of amateur astronomers. These rings consist of rock and/or ice debris, a few pieces of which are as large as a house, and most smaller. Image processing of the Voyager flyby images of the rings can make these stand out in beautiful (though artificial) colors which facilitate their separation into individual ring units (identified by letters). Saturn has since been visited by the Cassini spacecraft. Saturn also has a large number of satellites (34) of which seven are big enough to have evolved into spherical shapes. While sharing an icy composition, there are notable differences among the seven: Mimas, Enceladus, Tethys, Dione, Rhea, Titan, and Iapetus.

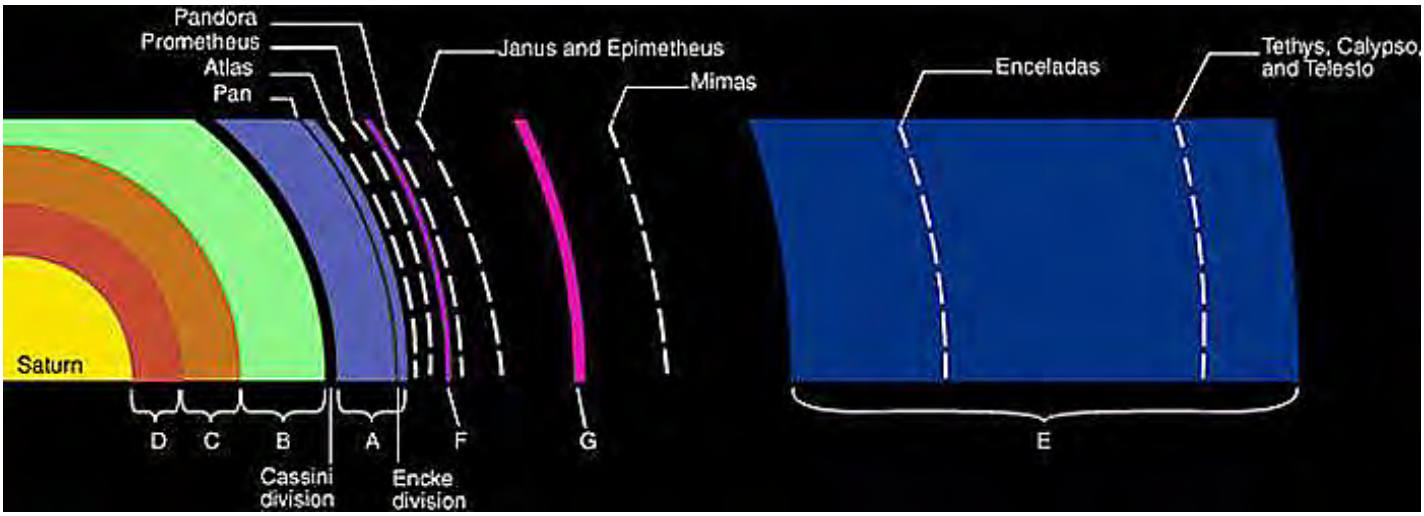
Saturn and its Moons

Now, on with the Tour! Both Voyager spacecraft observed the second largest planet, Saturn (diameter 84.3% that of Jupiter and 10% shorter at the poles). Voyager 1 flew past Saturn in November 1980; Voyager 2 passed by in August of 1981. Much of what we now know is summarized at this Saturn web site. When the first spacecraft was slightly above the rings, on the sunlit side, this planet presented a stunning image:



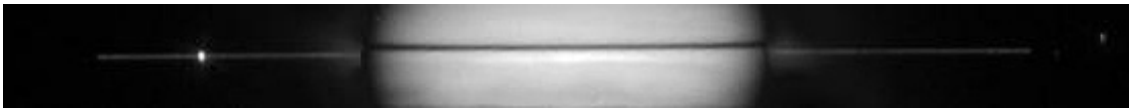
Cassini, the largest spacecraft ever to be sent to planet, was placed in orbit around Saturn on July 1, 2004.

The seven major rings begin about 7,000 km (4,350 miles) beyond the planet's discernible atmospheric edge and continue for another 74,000 km (45,984 miles), to the faint F ring (discovered by Voyager), a diameter of greater than 260,000 km (161,564 miles). But, the entire system is only about 1.5 km (about 1 mile) thick. The rings have been named using alphabet letters, arranged in the order of their sequence from closest to farthest from Saturn's gaseous surface (but the letters relate to the order of discovery): D, C, B, A, F, G, E. This can be shown diagramatically, in which the specific rings and the major satellites are shown in their relative positions:

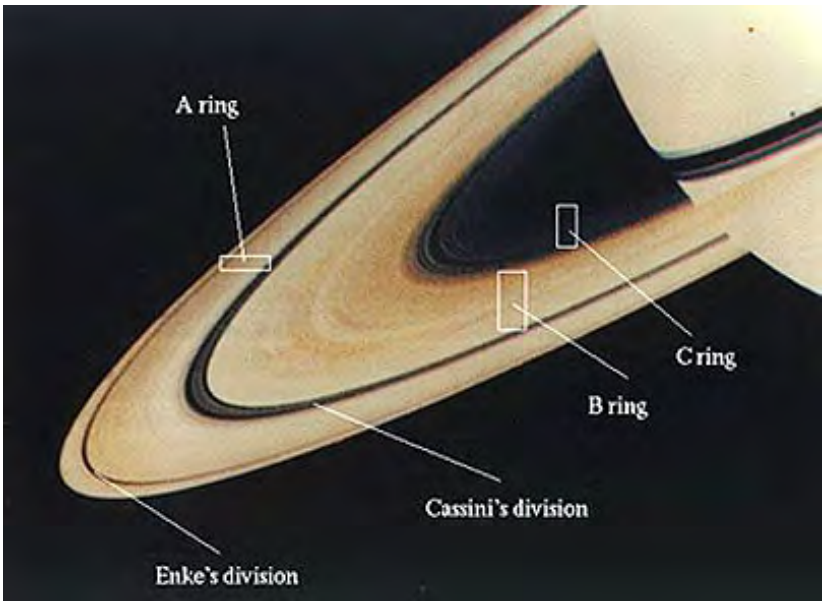


The E and G rings are usually not imaged. This composite mosaic made by the Cassini spacecraft when it was 3 million kilometers away shows these two and their relation to the inner rings:

The rings are in fact very thin (less than a kilometer). This is suggested by this Cassini image, which also shows one of the saturnian satellites, which indicates how small these are relative to Saturn's size:



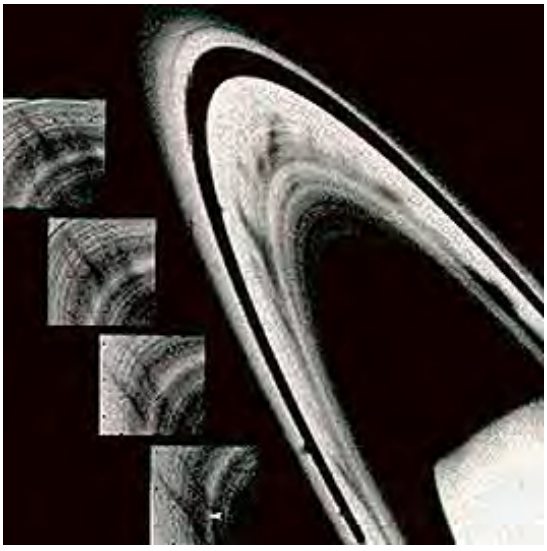
Let's reexamine Saturn's inner rings. The Cassini Division (also referred to as a Gap) separates rings A and B, and the Encke Division lies within the outer part of A.



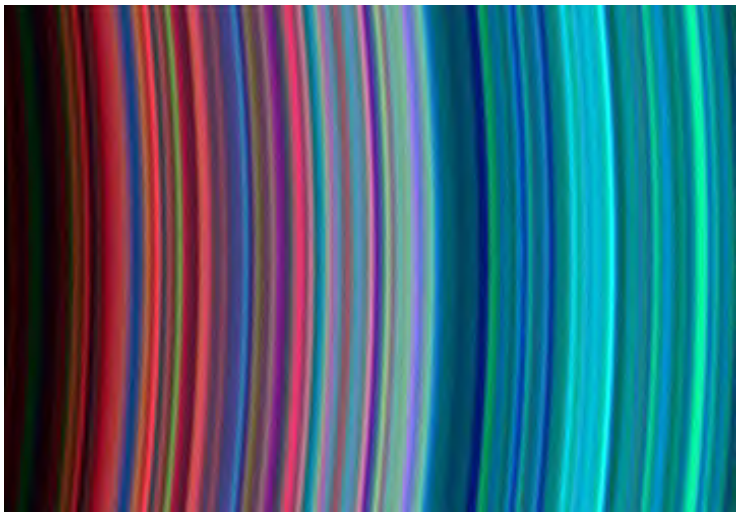
This image made by the Cassini spacecraft is the closest to the true color of the rings:

Note the details of the ring structure in this color-enhanced version, made by assigning red, green, and blue to clear, orange, and ultraviolet image frames, respectively. We chose it for its esthetic quality and to help further to distinguish its ring divisions:

The inner blue region marks the C ring. Beyond it is the B ring, whose inner part is orange and outer is greenish-blue. There is a large dark area known as the Cassini division (a gap) that separates the B ring from the outer (purplish) A ring. A smaller gap, named the Encke gap, subdivides the A ring. Although not visible in this image, parts of rings (mainly in B) contain "spokes", which are narrow darker linear "shadows" that lie perpendicular to the bands. These may be density discontinuities or clusters of small particles levitated above the residual larger particles in the rings by electrostatic charging. Note the many dark bands, which are gaps with lower densities of particles. This Cassini image shows examples of the spokes:

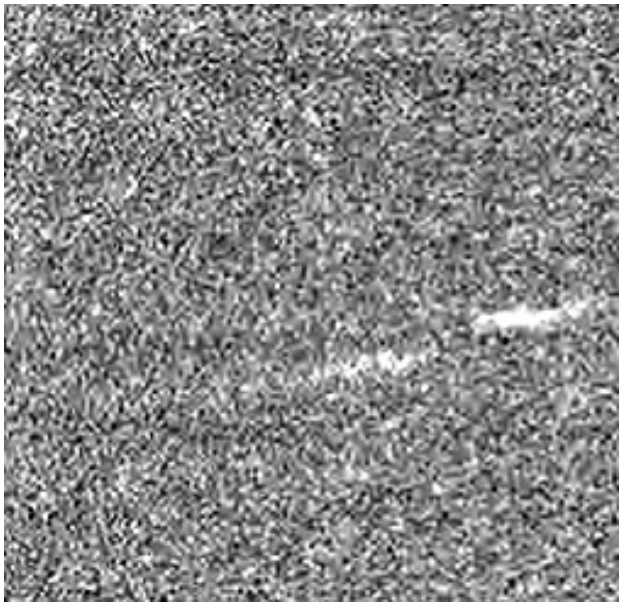


The Ultraviolet Spectrograph on Cassini has produced images whose colors indicate compositional differences. In this next image, of the B ring, the red bands have a higher proportion of "dirt" (rock/dust) whereas the blue bands are mostly just water ice. One model for the origin of the rings considers the band components with more dirt to be older than the blue bands.



The visible rings consist of mostly ice and ice/rock fragments that (although spread apart) reflect sunlight. Most particles fall within the millimeter to several meter-size ranges. A few chunks attain sizes up to 100 meters. Here is an artist's conception of the objects within one of the rings:

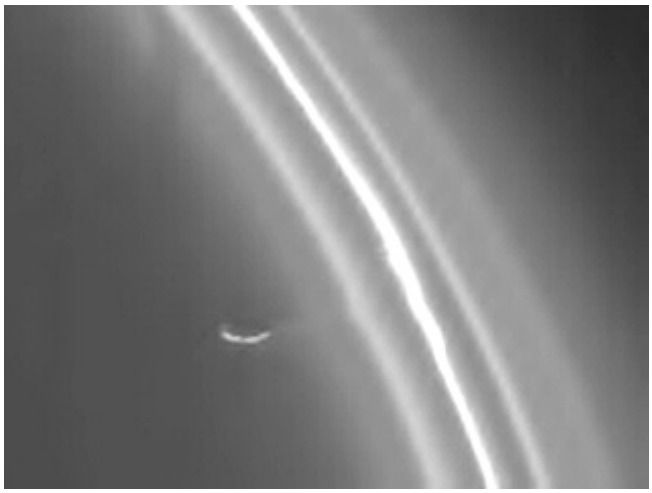
The actual ring particles were imaged just once, when the Cassini spacecraft actually passed through the gap between the F and G rings as it was about to enter its orbit around Saturn. This image shows that the particles in the F ring are mostly around meter-sized:



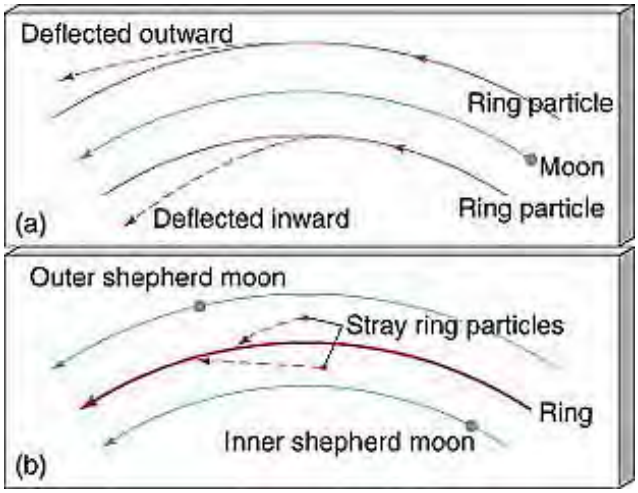
The origin of Saturn's rings is still unsettled. The materials may be even now in the process of organizing into a satellite or could be residual materials around Saturn that never succeeded in

building into a single large body. Or they might be debris from a large impact that shattered an earlier satellite. Or they could be the residue of a pre-existing satellite that moved so close to Saturn that the planet's tidal forces disrupted it (the distance at which such break up must occur is called the *Roche Limit*; for Saturn this is 1.5 times the planetary radius; for the Earth-Moon system it is 2.44 times Earth's radius). Recent discovery of small moonlets (typically 100 - 200 meters) has supported the idea of the rings being residuals from the destruction of a larger satellite moon. One generality: only the Giant planets seem capable of maintaining rings once formed, owing to the ability of their strong gravitational attraction to keep the rings from dissipating over time. One line of evidence considers the saturnian rings to be no older than 100 million years.

At least the outer rings, A and F, are kept from drifting apart by the action of three small satellites orbiting close to them. Called "Shepherd Satellites", their gravitational focusing acts to push straying particles back in line. Their action also may cause the braiding observed in the A ring. The F ring is controlled by the small satellites Prometheus and Pandora. Prometheus appears in this Cassini image:



As each satellite moves along, it deflects particles inward or outward depending on its position inside or beyond the ring. This diagram is relevant:

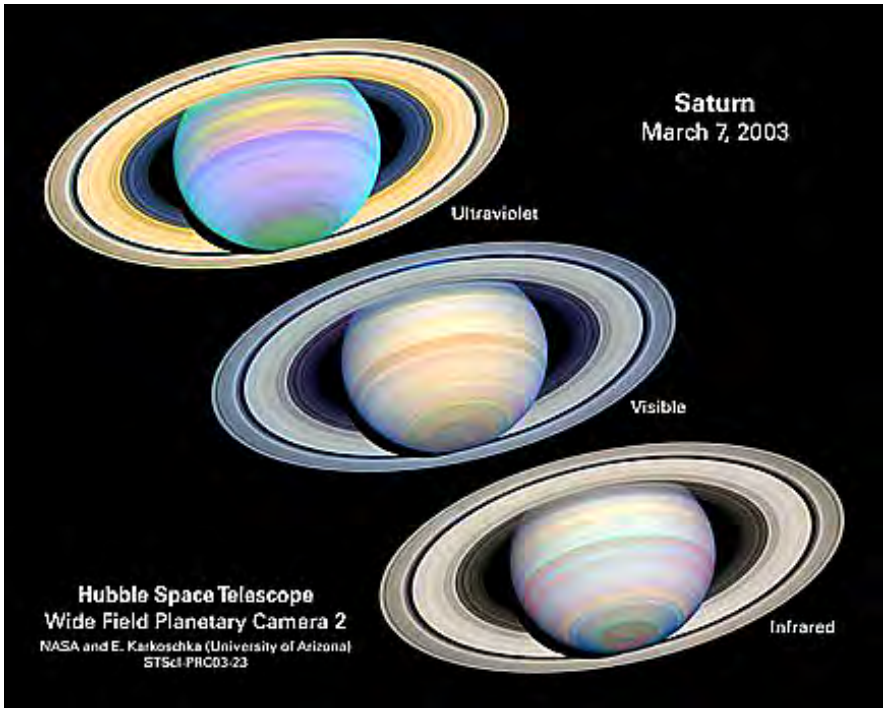


19-63: Speculate on possible fates of the Saturnian rings. ANSWER

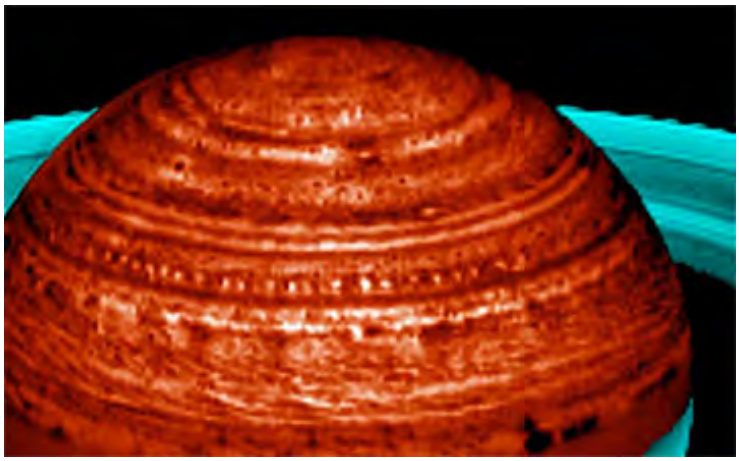
Saturn has about the same composition as Jupiter, namely Hydrogen and some Helium, but lower internal pressures and temperatures (up to about 12,000° C) may limit the extent of the metallic Hydrogen zone. Helium makes up about 7% of Saturn's atmosphere (on Jupiter He was 11%). Saturn's density, ~ 0.7 gm/cm^3, is the lowest of any planet. Saturn's thin surficial atmosphere, which is responsible for the visualization of an "apparent" surface, although moderately banded, lacks the reddish coloration typifying Jupiter's belts, but its yellowish color is similar to the Jovian zones. Circulation patterns, driven by jet streams, are also similar, including

oval spots, cyclonic storms, eddies, and swirls. Winds can exceed 300 km/hr. The depth of atmospheric circulation may be as much as 2000 km (1200 miles). The atmospheric bands differ somewhat in composition (which has been investigated further when the spacecraft Cassini arrived there in 2002).

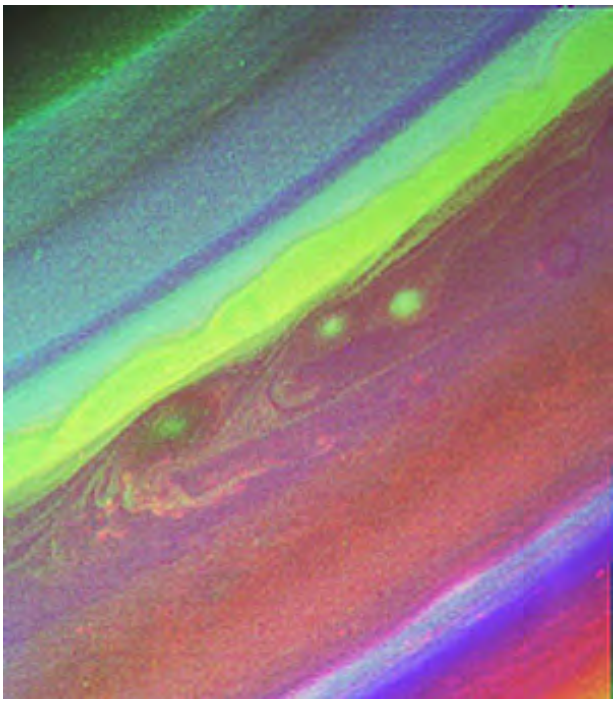
The NICMOS camera on the Hubble Space Telescope (Section 20) has taken an interesting multispectral image of Saturn (including IR bands) in which the belts are assigned contrasting colors (like what was done on the rings above) based on spectral variations to help emphasize the significant differences that do not stand out in true color images. This trio of images compares Saturn's gaseous surface and rings as seen in three spectral regions: UV, Visible, and near Infrared. Again, color assignments are somewhat arbitrary:

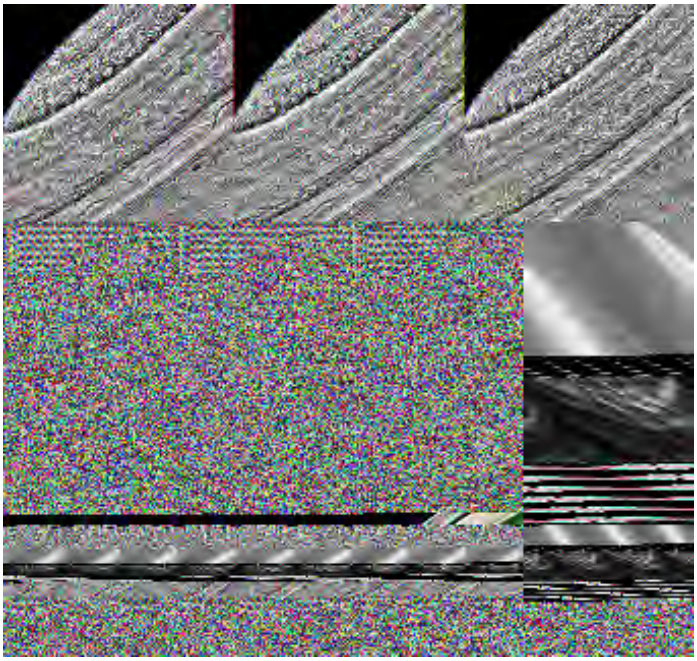


The various cloud bands and decks on Saturn leap out when viewed in the Infrared by the Cassini spacecraft. The white dots in this image are warm clouds arranged in a pattern described as a "string of pearls".

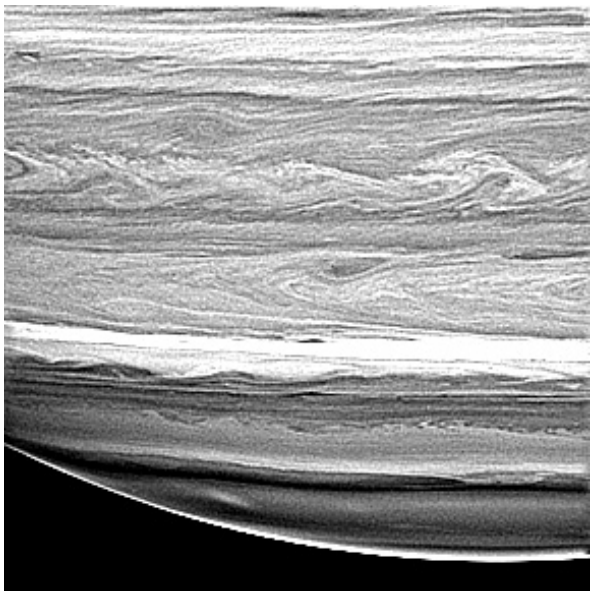


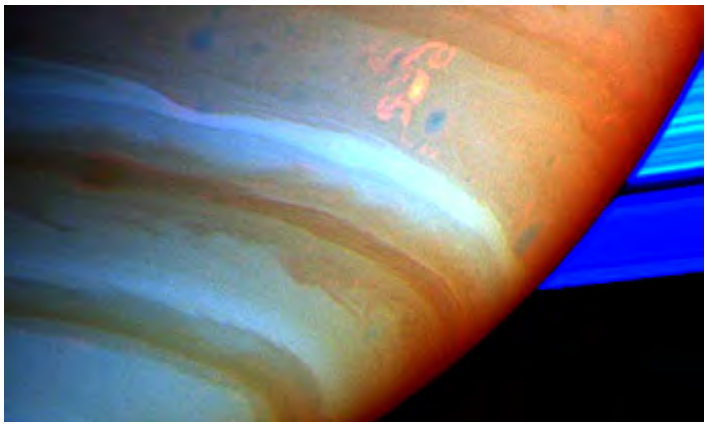
When Voyager looked at parts of the cloud surface in detail, some interesting fine structures were evident:



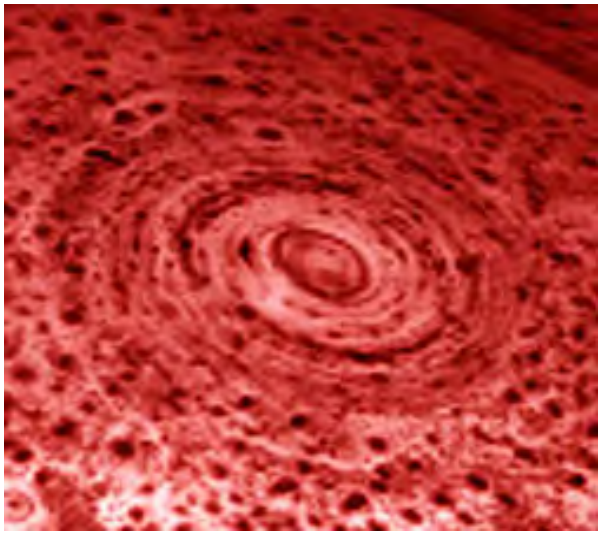


Compare these with the next two images of saturnian atmospheric bands obtained during the Cassini mission (next page):

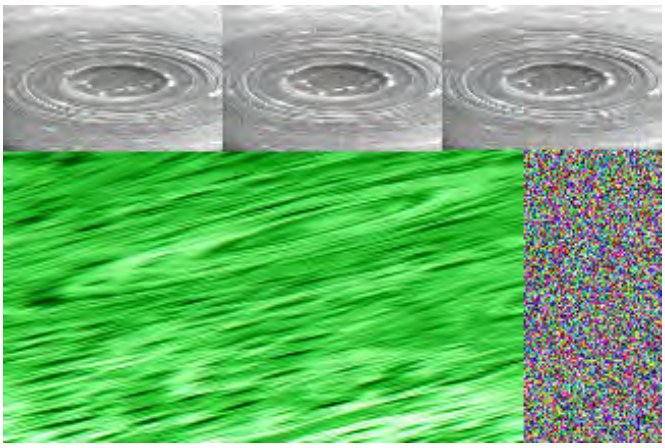




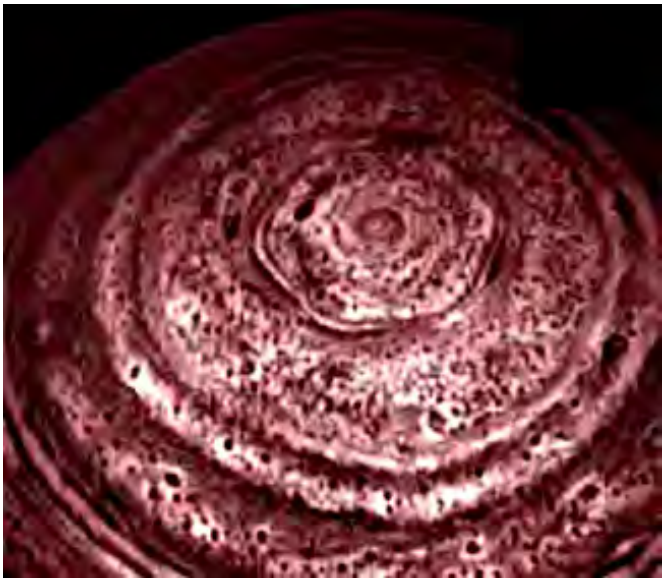
Similar to Venus, Saturn has a well developed vortex of swirling clouds at each pole. Here is the one formed over the south polar region:



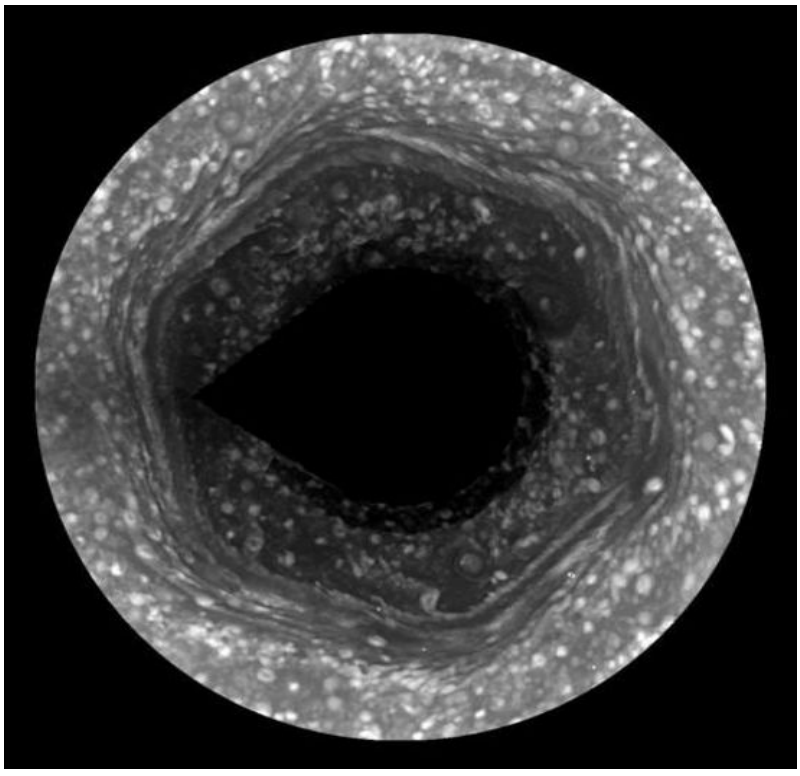
These clouds reach different heights, producing a topographic-like effect:



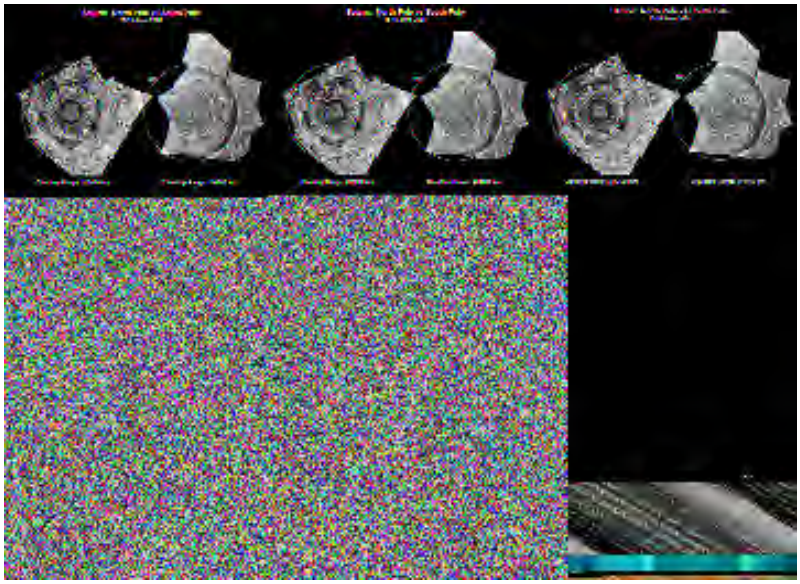
Cassini's sensors caught remarkable images of clouds above the North Pole. In the image below, note the general circular pattern, but the inner clouds have formed into a hexagon shape; this feature is more than 40000 km across and 200 km high. Why this distinct shape has developed is without any explanation but it seems related to actions caused by jet winds.



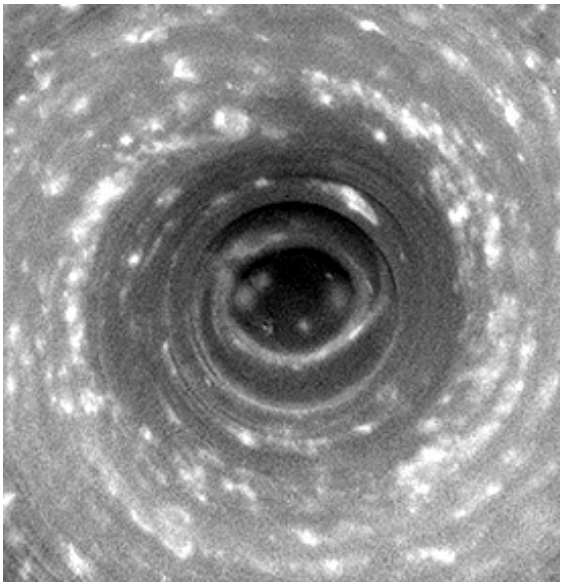
In November of 2009 Cassini got an even better look at the hexagon:



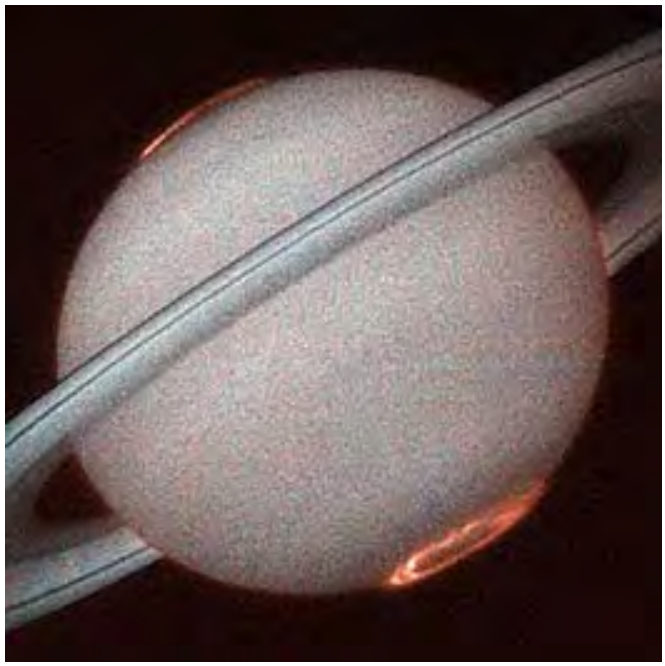
This next image pair compares the two polar cyclonic systems side by side (North is on the left):



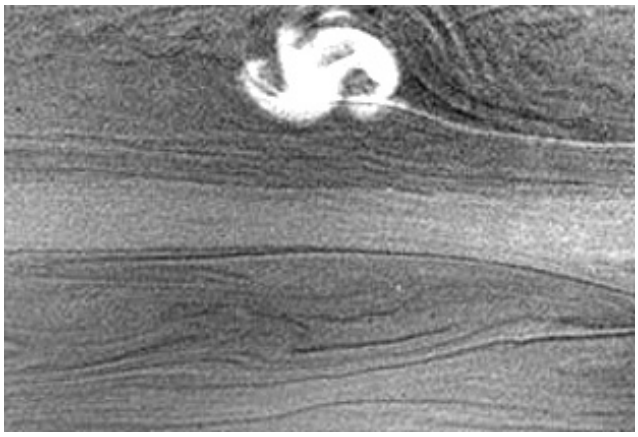
Like Jupiter, Saturn can have huge storms in its atmosphere. This next image shows what may be the biggest "hurricane" (8000 km; 5000 miles across) ever found in the Solar System, as it developed on November 7, 2006 in Saturn's southern hemsphere. The white cloud heights reach to about 60 km (40 miles); note the conspicuous central "eye":



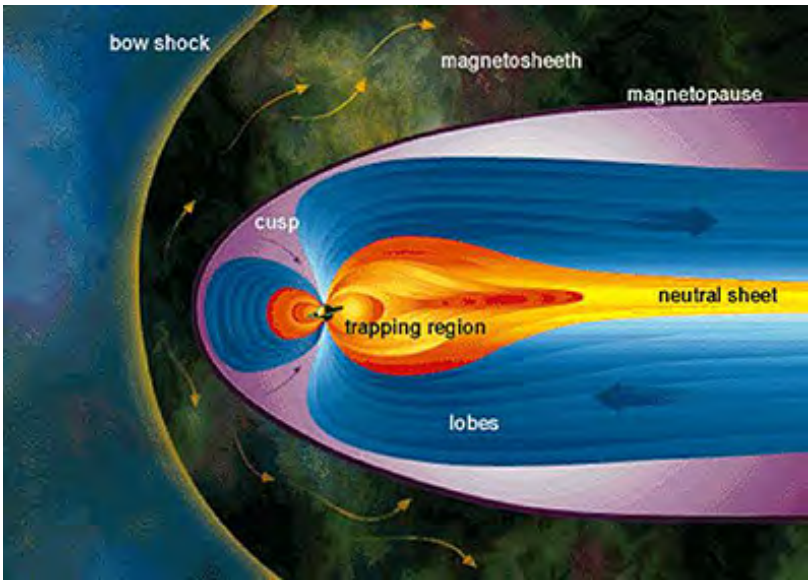
Voyager observations also indicated the possibility that, like Jupiter, there are auroras at each of Saturn's poles. This was confirmed in the '90s by the ultraviolet measuring capability of the Space Telescope Imaging Spectrometer (STIG) on the Hubble Space Telescope. In the STIG image below the auroral torus is in orange tones, representing atomic hydrogen emission as the solar wind interacts with the saturnian magnetic field.



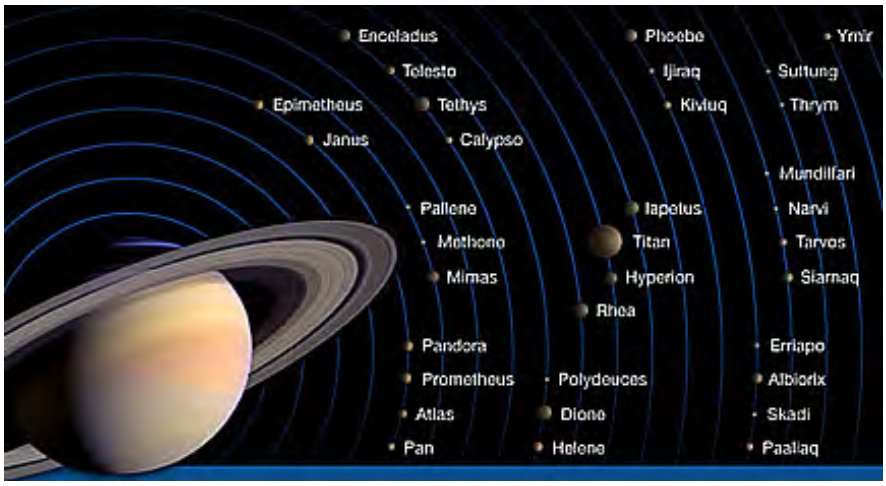
As might be expected, active atmospheres such as on Saturn can have severe electrical storms. This next image captures a cascade of lightning as large in area as the United States:



Saturn has a strong, albeit conventional magnetosphere, as shown in this diagram:

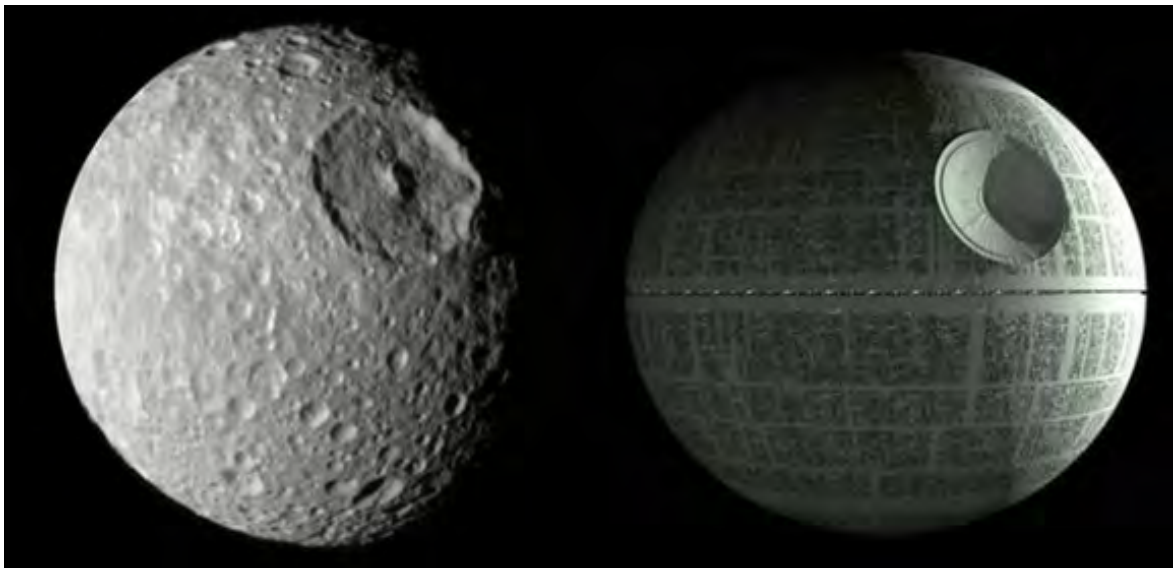


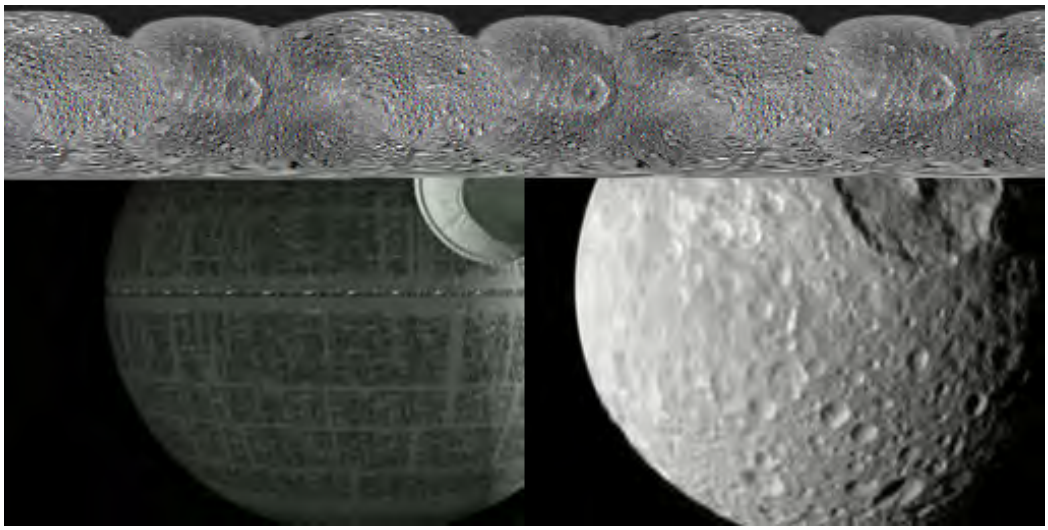
Saturn has 34 known satellites. Eighteen of these (all small) are located beyond the rings. The two diagrams below name these but their location on the charts is not really their actual positions in terms of distance:



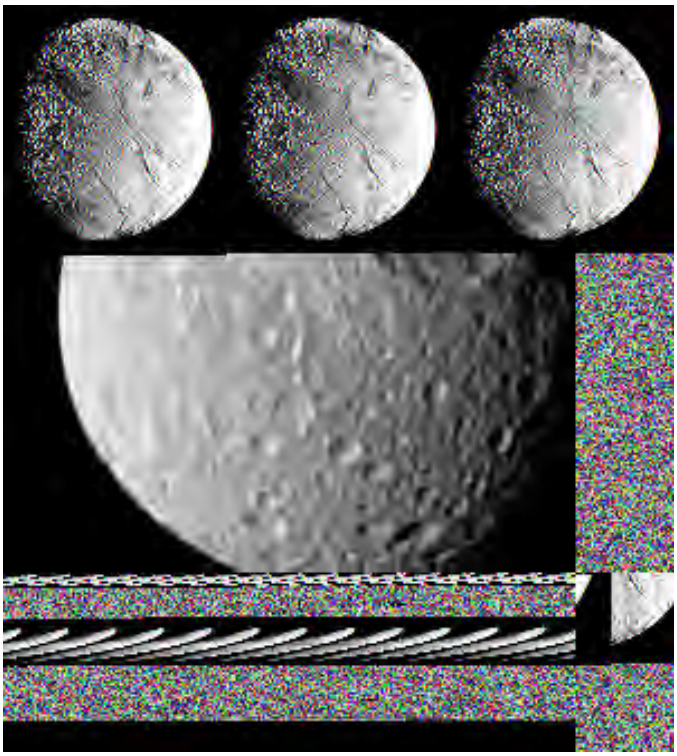
All but one (Iapetus), orbit close to the planet's equatorial plane. Seven of the interior moons are spherical, ranging in diameter from about 400 km to 5,100 km (249 mi—3,169 mi). Their names, in orbital-position order, starting at the innermost, are Mimas, Enceladus, Tethys, Dione, Rhea, Titan, and Iapetus. The remainder are smaller and have irregular shapes (although Phoebe is nearly spherical), and are mixtures of ice and water. Of the larger ones, six appear to consist almost entirely of water ice, but interior compositions are unknown. The seventh and largest, Titan, seems completely different, as we describe below. Most of the satellites have synchronous, prograde (orbiting in the direction of the planet’s rotation) orbits, locked by Saturn's gravity that causes each satellite's rotational period and revolution (about Saturn) to coincide. That means the same hemisphere always faces the inner planet (as the Moon faces the Earth). We will look at one or two representative images of each of the big seven, describing the satellite briefly.

Mimas is only 392 km (244 mi) in diameter. Its surface holds innumerable small craters. Because of the 136 km (84.5 mi)-wide crater, Herschel (approaching a size where disruption might have occurred), with its cyclopean eye-like visage, it has gained a special mystique because of its resemblance to the Death Star ship in the Star Wars movie series IV.





The larger (501 km [311 mi] diameter) Enceledus is the brightest object, in terms of albedo, in the Solar System. This brightness suggests regions of fresh or uncontaminated ice. This Voyager 1 image has been contrast-stretched to enhance major features which do not show well in the high albedo versions.



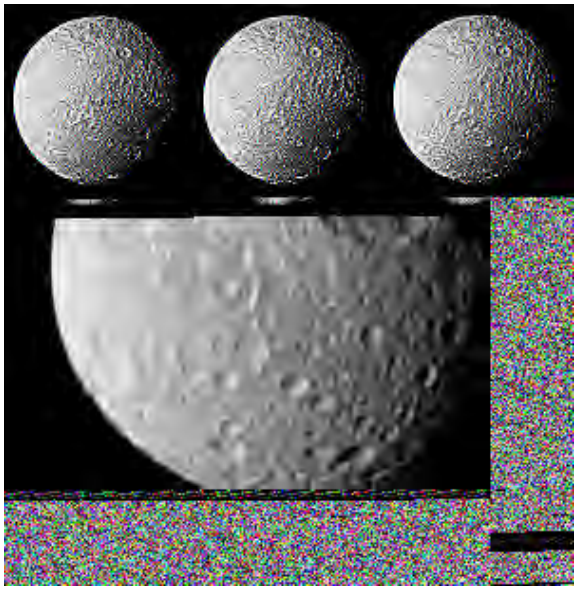
When examined from the multiple images returned, five distinctive terrains are visible on this satellite: 1) Smooth Plains (craters sparse) 2) Cratered Plains (relatively few craters); 3) Cratered Terrain (more craters but still less than the other ice satellites, implying a rather young surface); 4) Ridged Plains (subparallel ridges); 5) Grooved Terrain (including graben-like, straight to curvilinear depressions). Some of these features are indicative of tectonic stressing of a brittle crust underlain by liquid, with the energy to fracture it coming from bulging in response to tidal interaction with Dione, which orbits in resonance with Enceladus. Several linear scarps are faults that show offset of lines crossed by them.

19-64: Locate the five Tethys terrains on the image above. ANSWER

Tethys is a heavily cratered satellite (diameter = 1,060 km [659 mi]), on which appears a very large (about 400 km [249 mi] diameter) impact structure called Odysseus. A close look at Odysseus indicates both its rim and central peak have diminished in height under viscous flow. Because of this, its depth to diameter ratio is lower than normal for craters of this size.

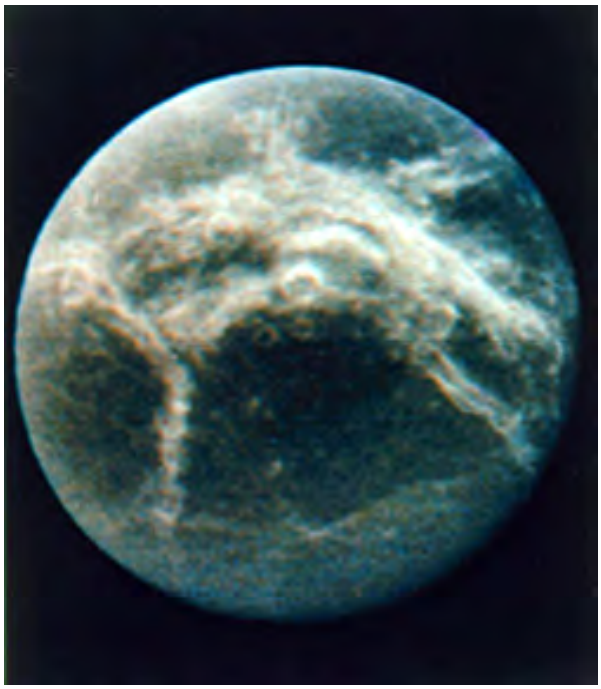
A huge valley, 2,000 km (1,245 mi) long, 100 km (62 mi) wide, and up to 5 km (3 mi) deep, known as Ithaca Chasma, suggests that the ice cracked at some stage, perhaps when rifting occurred, as water converted to contracting ice that redistributed tension. It is hard to see in this next

image but look for the long, broad "slash" in the upper left quadrant.



The slightly larger (1,120 km [696 mi]) Dione has two dissimilar hemispheres: on one side, the leading hemisphere (because of synchronous locking, where the same side always faces the direction of the advancing orbit), the surface is heavily cratered and fractured, with a uniformly medium bright tone.

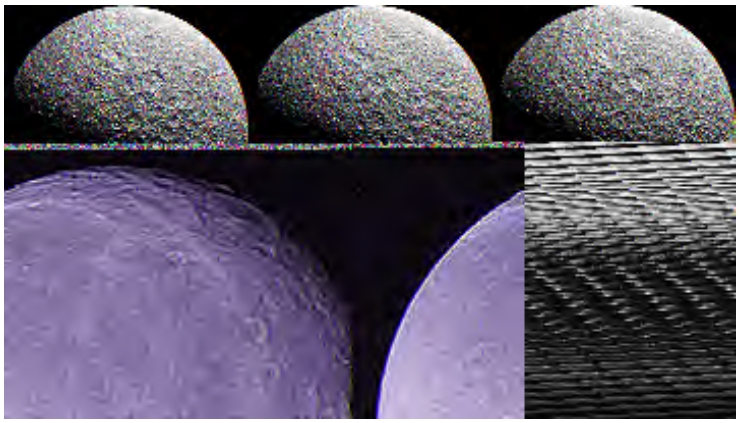
The other side, the trailing hemisphere, is darker, has a much lower crater density, and contains broad, very bright streaks in a diffuse network. Planetologists believe this strange pattern is frost and ice, expelled or extruded from fissures, and possibly from ice volcanoes, in which the water is the "lava" in an otherwise frozen structure.



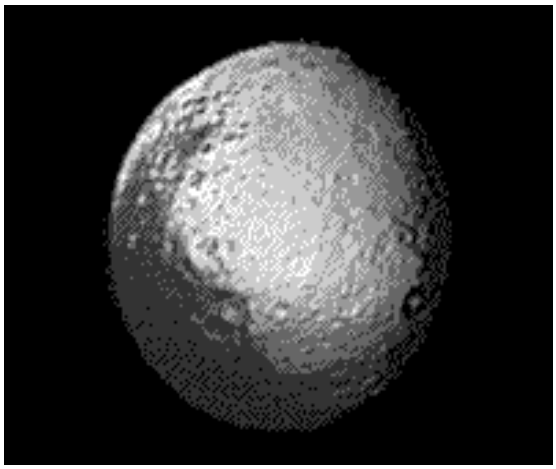
19-65: Just why is one hemisphere peppered with so many more craters than the other?
ANSWER

Rhea (1,530 km [951 miles] in diameter) is very similar to Dione, in having a leading hemisphere that has crater densities comparable to the Moon and Mercury, as is hinted at in this full disk view, and a trailing hemisphere with few craters, streaks (ejecta rays?), and fractures.

This is the heavily cratered side of Rhea:



This same theme repeats at Iapetus (1,460 km [907 miles] in diameter), which is further from Saturn. The leading hemisphere has a very low albedo (0.03-0.06), in sharp contrast to previously described satellites, and with abrupt boundaries, against the trailing edge terrain (albedo about 0.5), which is heavily cratered. The reason for this dichotomy within this satellite is still uncertain. This trailing-edge material may have emerged from within Iapetus during crustal foundering. But, the dark terrain is consistent with silicates that have high carbon content, such as those that comprise carbonaceous chondrite meteorites. Thus, this part of Iapetus could have formed as ejecta, which came from another satellite or more likely from infall of a similar, asteroid-like body onto Iapetus.



Typical of the 23 smaller satellites is Phoebe (200 km]134 miles] diameter), orbiting nearly 4x as far out as Iapetus. The first look at this tiny moon was taken by Voyager; later, much better views were obtained by Cassini.

Slightly larger (286 km; 179 miles) than Phoebe is the more angular Hyperion:

As a preview of the extent of improvement that Cassini has provided for the saturnian satellites, compare the above view with this view of Hyperion made by Cassini:

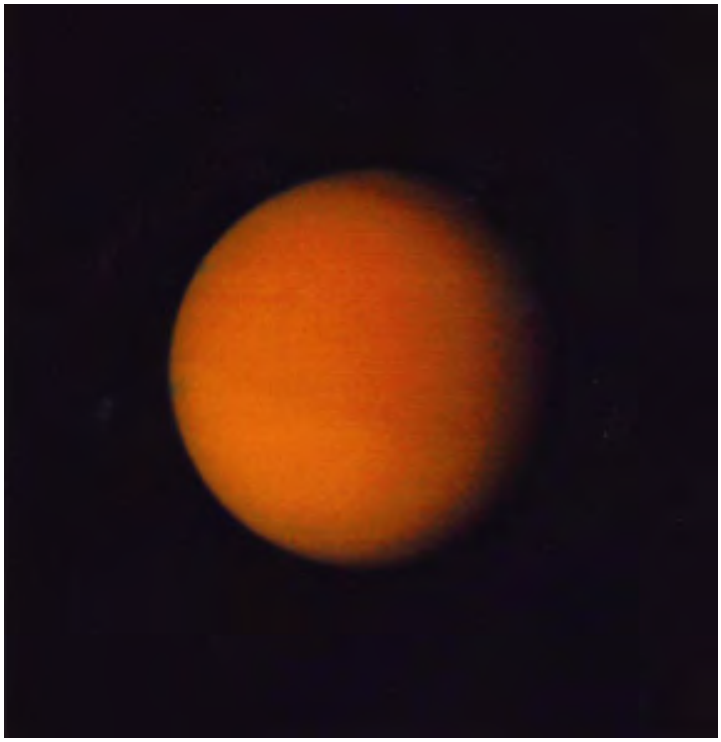
Now move to the next page which covers Titan and the still incoming views of Saturn and its satellites being obtained by the Cassini-Huygens mission.

PREVIOUS RST NEXT

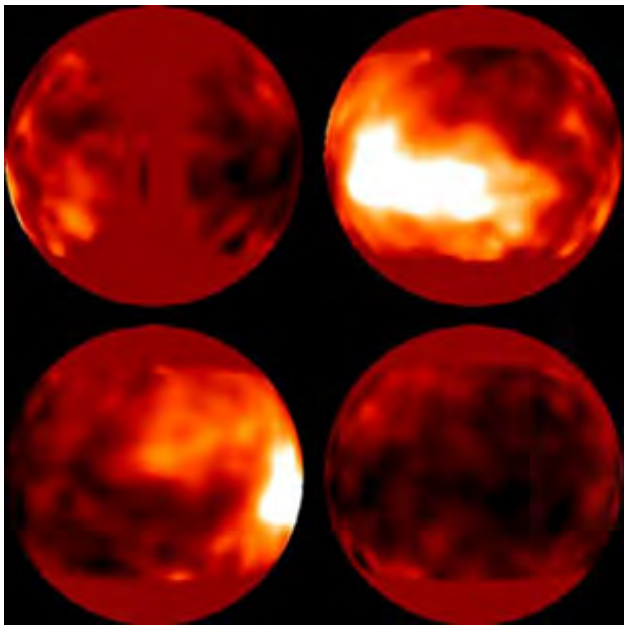
Primary Author: Nicholas M. Short, Sr.

We continue with our exploration of saturnian satellites by concentrating possibly the most intriguing of the group, and certainly one of the most interesting in the Solar System - Titan is a "maverick", being surrounded by a nitrogen atmosphere and some organic molecules that give it a distinctive reddish-orange color. Brief treatments of several of the irregular-shaped satellites are then considered. The important Cassini-Huygens mission in 2004 to Saturn is presented on this page and the next, with images added from time to time as they are sent back after that spacecraft achieved orbit and began its 4 year study of Saturn and its moons.

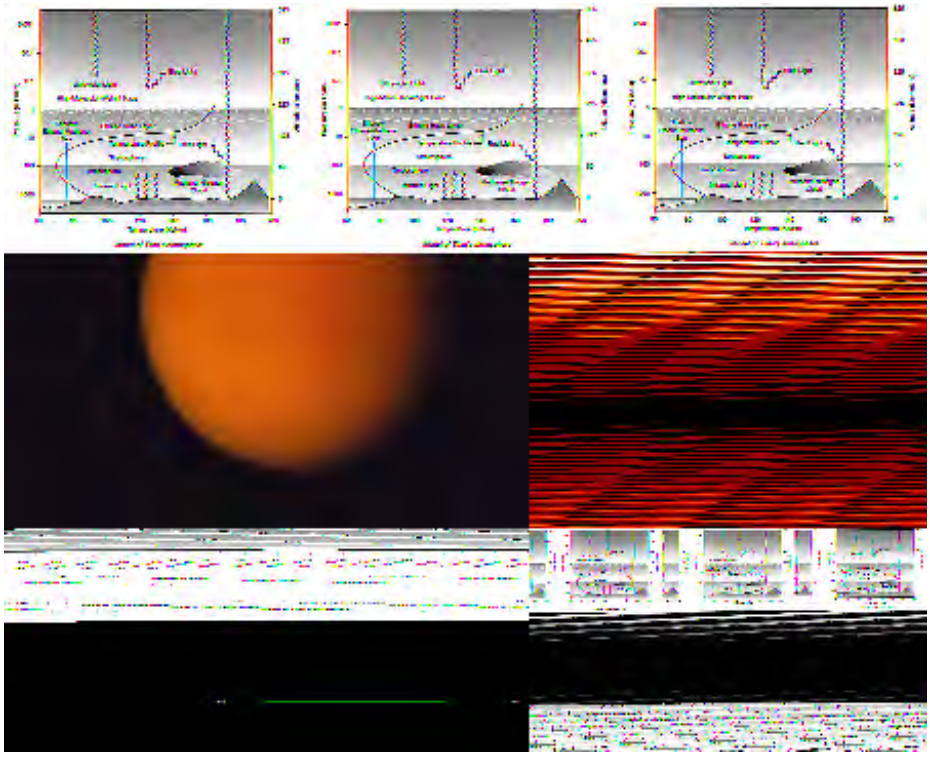
Between Rhea and Iapetus is the large (5,150 km [3,200 mi]) satellite Titan, which is a "maverick" among the icy group, in that it is quite different in its appearance owing to the presence of an atmosphere with a distinctive color described as a brownish-orange. In a sense it is as odd amongst its peers as Io is within the Galilean satellites.



Some of the variations in the atmosphere can be discerned as Titan rotates. Here are four views obtained through the Hubble Space Telescope, using selected bands in the infrared that penetrate the atmosphere.:



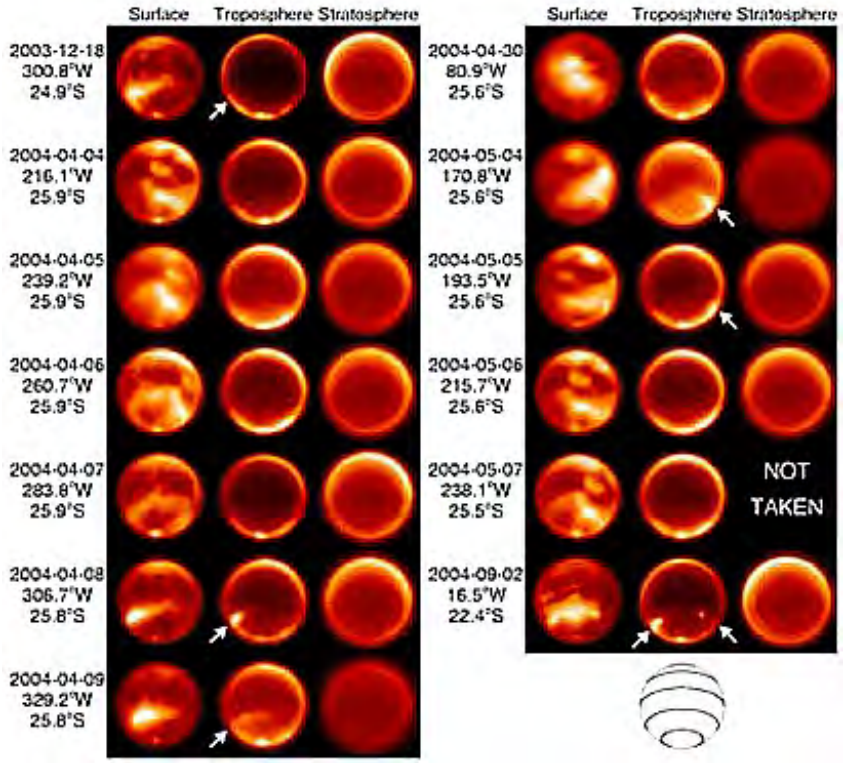
A diagramatic summary of what is known about Titan's atmosphere, prior to the Cassini mission, appears here:



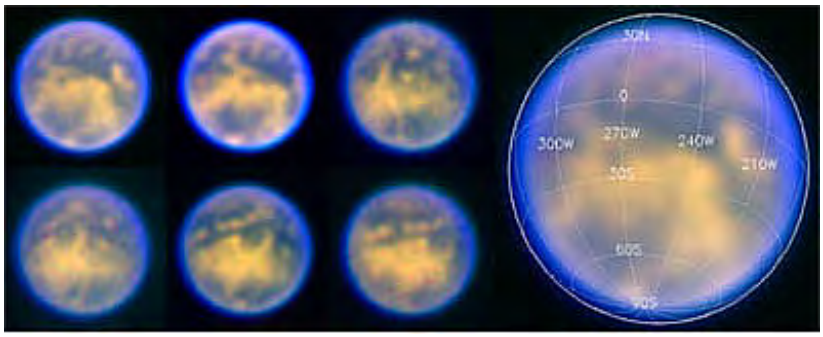
Estimates of the relative amounts of atmospheric constituents, in percentages and parts per million, are given in this table; nitrogen is the principal constituent:

Nitrogen	N_2	90–97%
Argon	Ar	0–6%
Methane	CH_4	2–5%
Hydrogen	H_2	0.2%
Ethane	C_2H_6	10 p.p.m.
Carbon monoxide	CO	~10 p.p.m.
Acetylene	C_2H_2	2 p.p.m.
Propane	C_3H_8	500 p.p.b.
Hydrogen cyanide	HCN	170 p.p.b.
Ethylene	C_2H_4	100 p.p.b.
Acetonitrile	CH_3CN	5 p.p.b.
Carbon dioxide	CO_2	10 p.p.b.
Cyanoacetylene	HC_3N	10 p.p.b.
Methylacetylene	CH_3C_2H	5 p.p.b.
Cyanogen	C_2N_2	5 p.p.b.
Water vapour	H_2O	8 p.p.b.
Diacetylene	C_4H_2	1 p.p.b.

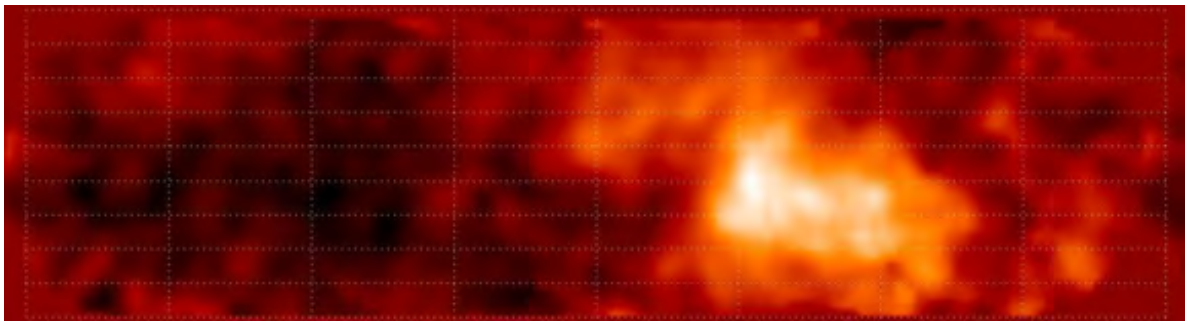
Titan has been studied from Earth. Photos taken through ground telescopes over several years show variations in lighter-toned areas which are assumed to be some type(s) of clouds that come and go; these are mainly just above the surface:



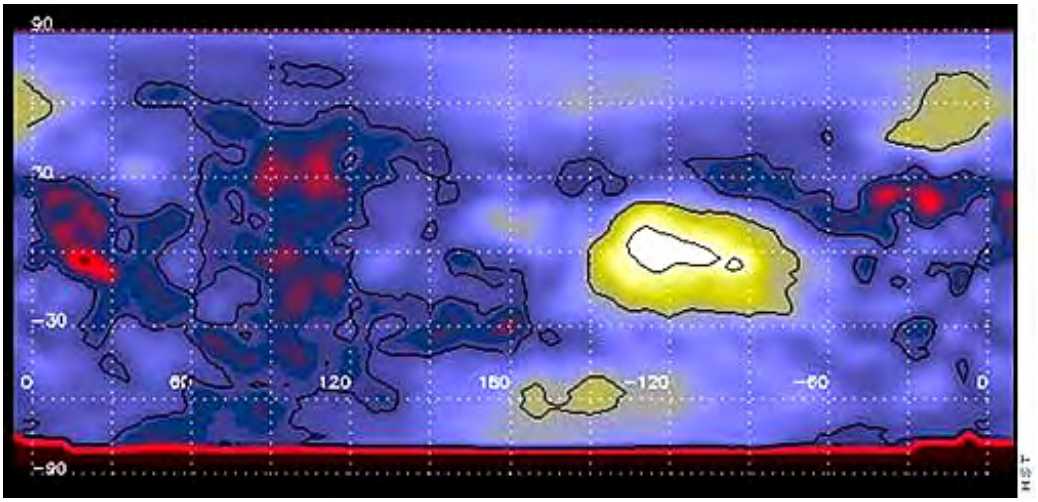
The European Southern Observatory (ESO) Very Large Telescope (VLT) in Chile has obtained sharp images using an instrument called the Simultaneous Differential Imager (SDI) camera, which obtains Near-IR images at 1.575 and 1.600 µm (surface sources) and 1.625 µm (atmosphere). SDI data taken on different nights can be reprocessed by differencing to give this group of images representing changes which may be due to the surface; the actual surface itself is likely to be partially obscured by the organic haze in the atmosphere:



The color version can be enhanced to bring out as many details as possible - rendered in reds and yellows - rather than in the natural color that is a medium brownish-yellow. Titan appears unlike the others because its dense atmosphere (with surface pressures of 1.5 atm, about 50% higher than Earth's) hides its surface. Nitrogen is the main constituent of the atmosphere but the 2% of methane and associated organics cause the colors noted. The continuous clouds are somewhat analogous to the haze or smog found above urban areas such as Los Angeles. We show some of this atmospheric structure here:



What is below the clouds has been revealed somewhat by the Cassini mission, described on this page. At Titan's atmospheric temperatures (typical value: -179° C), the ethane (and probably methane and nitrogen) may have condensed into a liquid "ocean" that is up to a kilometer deep. Below this, Titan probably consists of a layered mix of silicates and water ice. An infrared image, acquired by the Hubble Space Telescope, suggests that there may be several upwellings of (solid?) material, similar to "continents" (see yellow area in the image above). Recent studies now point to some of the variations in the atmosphere imagery are related to topographic irregularities on Titan's surface. In fact, prominences as high as 2000 meters have been postulated to account for some of the patterns in the image. A map has been prepared that purports to indicate the height variations in that surface:



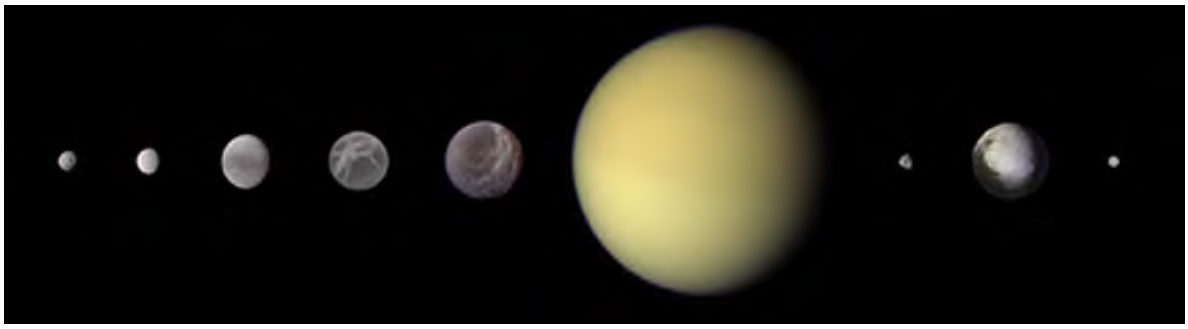
This atmosphere consists of about 90% molecular nitrogen, 6% argon, with the remainder being methane and organic derivatives, such as hydrogen cyanide, ethane, acetylene, and CO_2. Solar radiation breaks some of these constituents into colored compounds, analogous to urban smogs on Earth. Ammonia, now a trace constituent, may once have been more abundant, until

dissociating into carbon-nitrogen compounds.

This mysterious satellite (larger than Mercury and Pluto) - totally disparate, compared to the other saturnian moons - may harbor still other organic molecules of unusual nature. This surmise just cries out for follow-up study, since no other planet besides Earth or planetary satellite in the Solar System seems so promising for finding and identifying the more complex molecules that may be present.

19-66: What is the main implication of this unusual subsurface liquid and the chemical compounds found at Titan? ANSWER

19-67: Now that you have been introduced to the saturnian satellites, let's see how well you can recall and identify each of the larger ones. Examine this montage gallery of 9 satellites and name each one. ANSWER

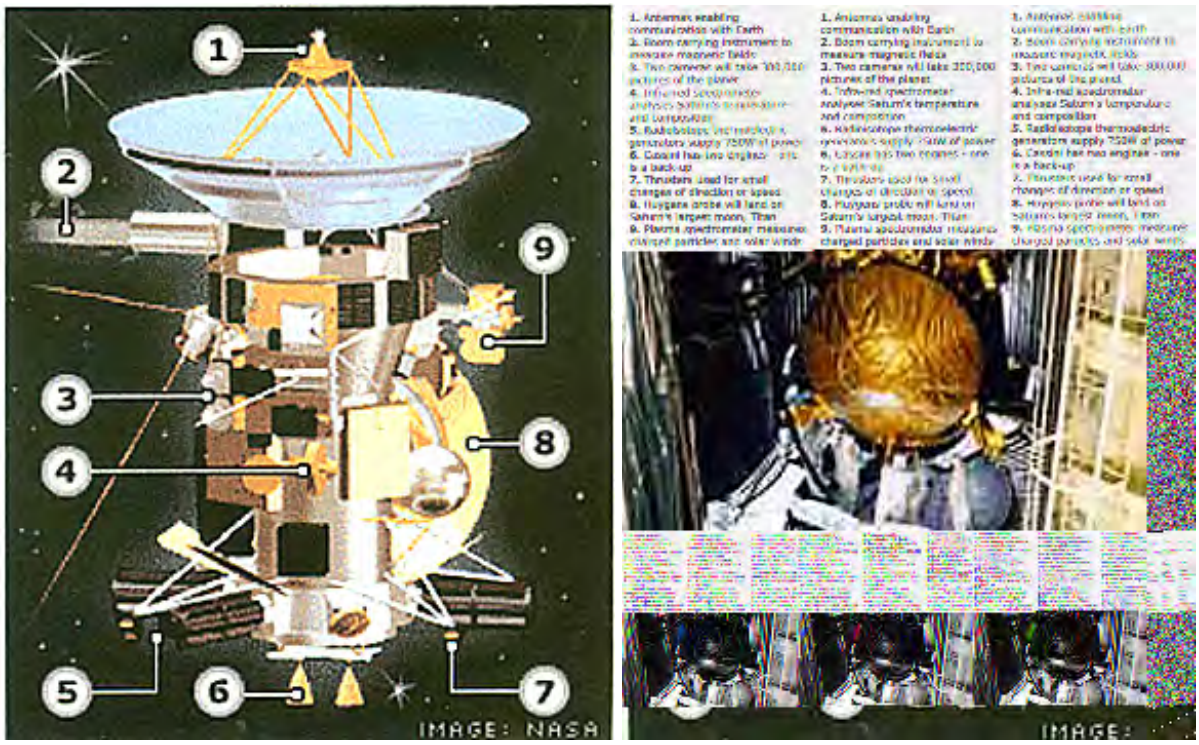


The Cassini-Huygens Mission

Just as the Galileo spacecraft has added significant knowledge of Jupiter, a sophisticated mission to Saturn is underway. That probe is part of the Cassini-Huygens mission, named for the 17th Century French-Italian astronomer who pioneered telescopic observations of Saturn. The Cassini spacecraft successfully launched from Cape Canaveral at 4 AM, on October 15, 1997, to begin a 6.7 year journey to the Ringed Planet. This vehicle has taken a round-about trip to gain velocity by gravity assisted "kicks". By passing Venus twice, then Earth again, and Jupiter later, the spacecraft used their gravitational pulls to increase its speed enough to reach Saturn. JPL manages the mission, with experiments from NASA, the European Space Agency, the Italian Space Agency, and other participants. A good overview of the Cassini mission is found on JPL's Cassini website. The Cassini Orbiter's task is to explore Saturn's magnetosphere, its rings, its icy satellites, its atmosphere, the atmosphere of the moon Titan and, with luck, its surface as well.

Built at a cost of $3.27 billion dollars, the complete Cassini spacecraft, the size of a school bus (5.7 m [22 ft]; 5.83 tons), is shown here in its assembly room before being taken to its launching rocket:

The Orbiter contains a variety of instruments, as listed below. Power for the spacecraft is a RTG (Radioisotope Thermoelectric Generator). The general layout of its major components is portrayed in this schematic:



* **Imaging Science Subsystem (ISS**) - observes particle properties, vertical distributions (~6 km/px, 0.6 Mbit/frame with 2x2 summing). ISS will also examine wind/cloud motions; (3-12 km/px, 3 images/timestep in CB1 filter to increase SNR), and search for and monitor lightning/aurora (High-resolution imaging, 50-200 m/px, special targets, emission angles < 45° prefer IR-polarizer (phase 45-110)).

* **Cassini Plasma Spectrometer (CAPS)** - investigates large-scale and distant aspects of the Titan interaction with Saturn's Magnetosphere by observing during the entire period around an encounter from 10 to 25 RS.

* **Composite Infrared Spectrometer (CIRS)** - obtains information on trace constituents in Titan's stratosphere. Integrate on limb at two positions. Obtain information on CO, HCN, CH_4.

Integrate on disk at air mass 1.5-2.0. Pointing: -Y to Titan, X away from Sun.

* **Ultraviolet Imaging Spectrometer (UVIS)** - observes the star Beta Ori as it becomes occulted by Titan's atmosphere.

* **Visible and Infrared Mapping Spectrometer (VIMS)** - provides new high resolution images that will help understand Titan's geology and the fate of CH_4. VIMS will also search for and study the evolution of mid-latitude clouds, and search for lightning and hot spots.

* **Magnetometer (MAG)** - examines large-scale and distant aspects of the Titan interaction by observing during the entire period around an encounter from 10 to 25 RS. T13 is an equatorial wake flyby under plasma conditions near Saturnian local midnight with 1852 km altitude at closest approach. Thus it is very similar to T13 even according to local time.

* **Magnetospheric Imaging Instrument (MIMI)** - investigates micro-scale and near aspects of the Titan interaction by observing during about one hour period around an encounter. With -Y pointed toward Titan, when within 30 minutes of the targeted flyby, optimise secondary axis for co-rotation flow as close to the S/C -X, ± Z plane as works with the other constraints on Pointing. Also, measure Titan exosphere/magnetosphere interaction by imaging in ENA with INCA (when Sun is not in INCA FOV).

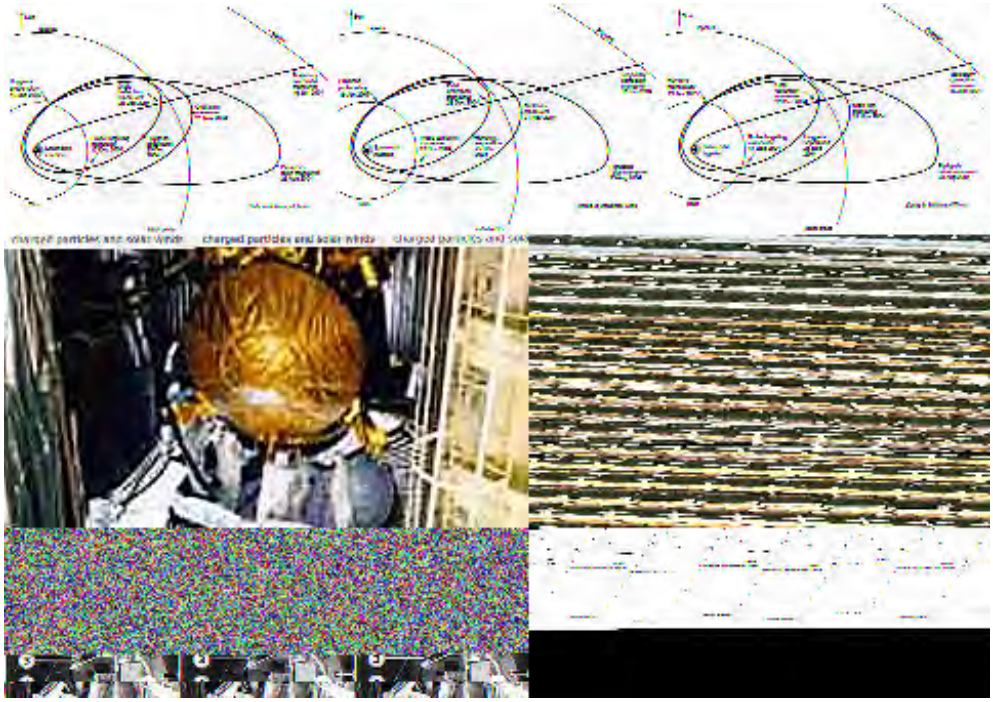
* **Ion and Neutral Mass Spectrometer (INMS)** - procures data regarding Titan's atmospheric and ionospheric composition and thermal structure. INMS will also observe the magnetospheric/ionospheric interaction.

* **Radio and Plasma Wave Spectrometer (RPWS)** - performs observations in the immediate vicinity of Titan, including thermal plasma density and temperature measurements with the Langmuir probe, search for lightning and other radio emissions, characterization of plasma wave spectrum, search for evidence of pickup ions. Langmuir probe within 90° of spacecraft ram at closest approach, co-rotational ram outside of ± 15 minutes.

* **RADAR** - conducts low and high resolution SAR (Synthetic Aperture RADAR) imaging of Titan's surface. Additionally, RADAR will collect Altimetry, Radiometry, and Scatterometry Data. SAR swath cuts right across Xanadu, as well as some areas where there exists good ISS/VIMS hires coverage, enabling useful comparative studies.

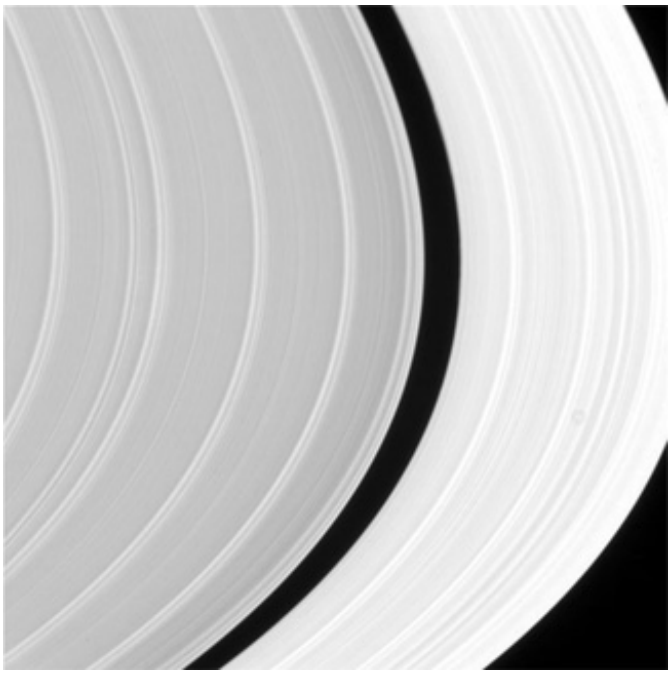
The nearly 7 years needed for Cassini-Huygens to travel the 3.5 billion kilometers (about 2.1 billion miles) involved a series of 4 gravity "kicks" (velocity boosts) by flying the spacecraft relativelly close to planets, as shown in this diagram (from the Cassini-Huygens Home Page):

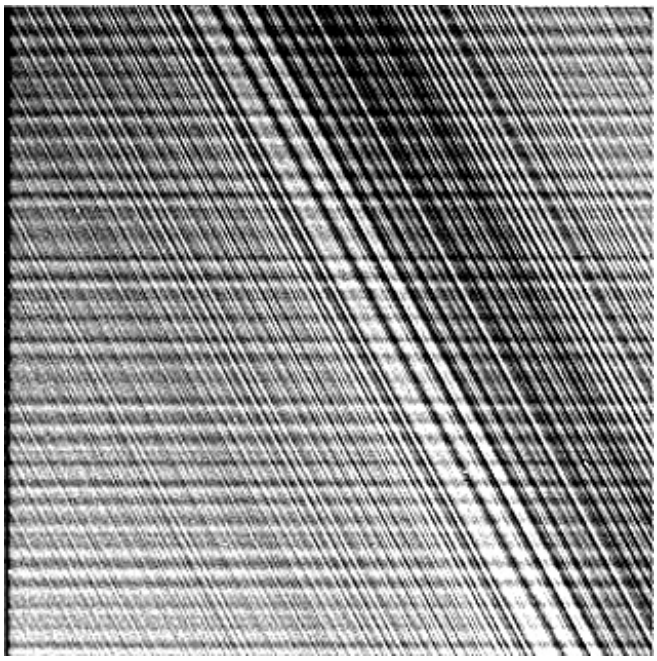
After arrival on late June 30, 2004 Cassini underwent a series of fuel burns to progressively insert it into orbits of different radii and ellipticity around Saturn. This diagram gives the shifts during the first 7 months:



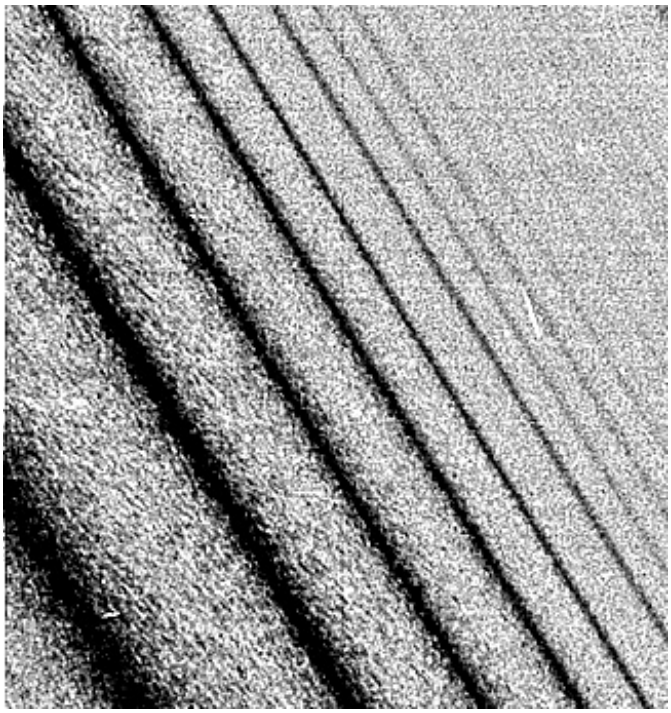
Cassini, which cost in excess of 3 billion dollars (and raised protests from some environmentalists fearful of its 77 grams of plutonium that will serve as a power source), is the most sophisticated and ambitious probe yet sent to explore the planets.

Cassini began its process of insertion into orbit around Jupiter at 6:30 PDT on June 30, 2004 by traveling through a gap in its rings. One of its antenna was deployed to act as a shield against microparticles of ice in the ring complex. A burn sequence of nearly 90 minutes was required to achieve a proper orbit for the first phase of exploration. Total success during this phase was achieved. A series of hundreds of spectacular images of the rings was obtained in this timeframe. Since they show imagery similar (but at better resolution) to the Voyagers, we will present only a few of these here:

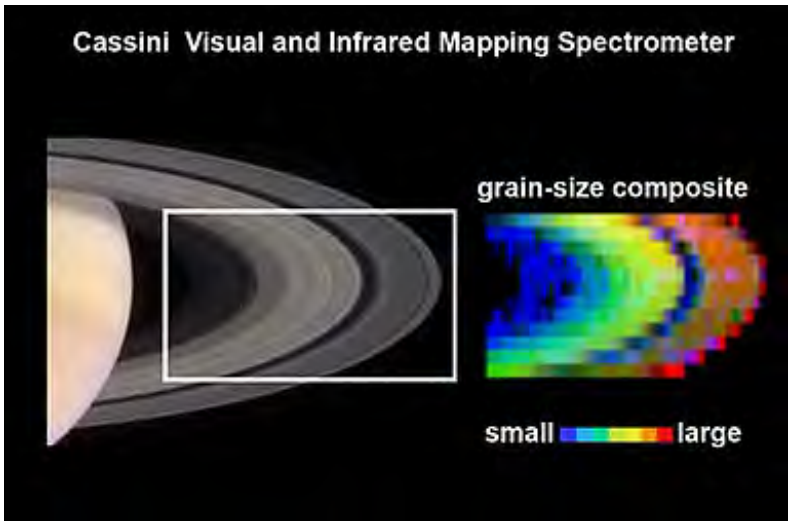




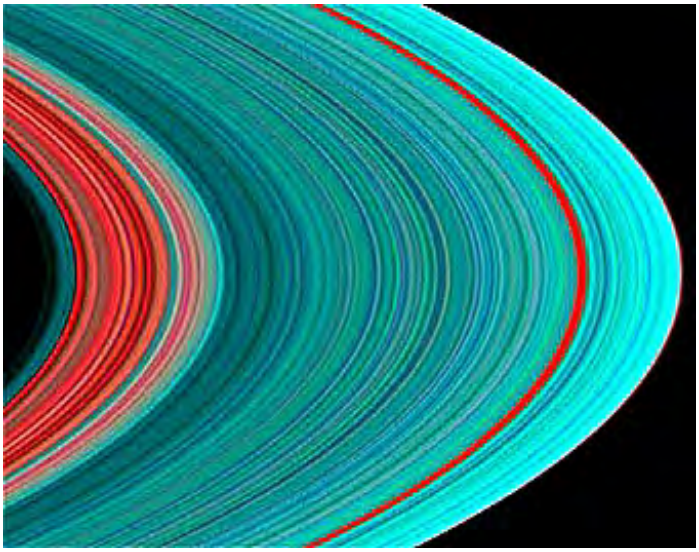
In this next image of Ring A, the closeup detail shows the individual ice block nature of the larger particles making up the ring itself.



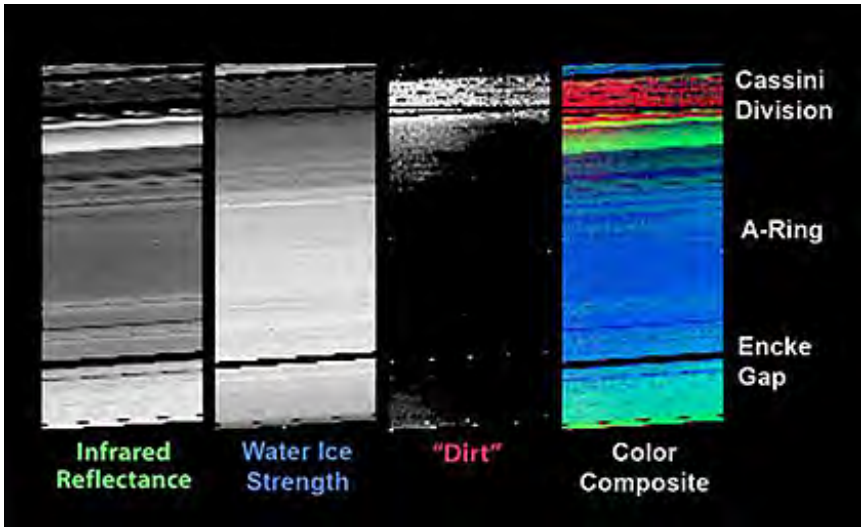
Cassini has determined that there is a gradual increase in size of the ice particles going from the inner main ring to the outer main ring:



This variation in particle size and in possible non-ice "dirt" are factors in these striking color renditions of the A and the B-C ring characteristics, as determined from UV images interpreted by a group at the University of Colorado-Boulder:

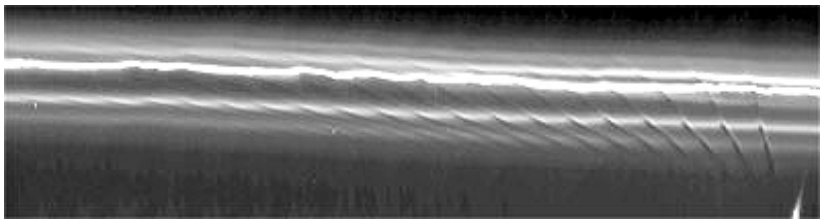


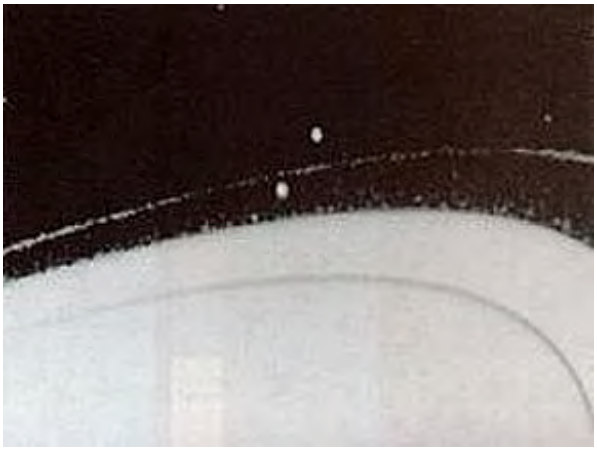
Another discovery, using the Visual and IR Imaging Spectrometer, is that the Cassini Gap, and by inference the Encke Gap, is not a region of minimal particles but instead contains "dirt", particles of non-ice composition. These may be remnants of the rocky portion of a satellitic fragment composed of low reflectance basaltic material.



The F ring is narrow and seemingly isolated, as seen in the first image below. The F ring has always merited special attention because of its twisted appearance. Cassini shows that in this

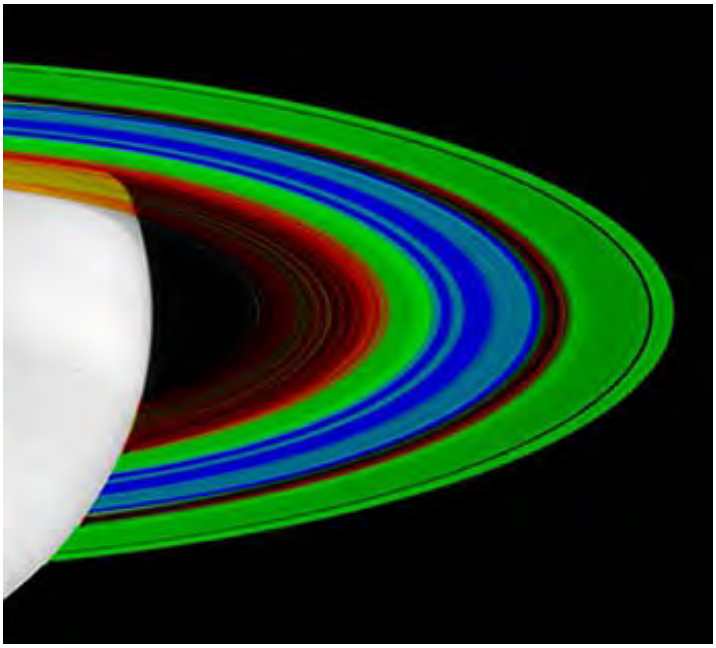
mosaic image. The contortions are caused by gravitational interactions with the small satellites Prometheus (145 x 85 x 62 km) and Pandora (90 x 53 x 39 km), spoken of as "shepherd" moons because they help to guide and maintain the ring material:



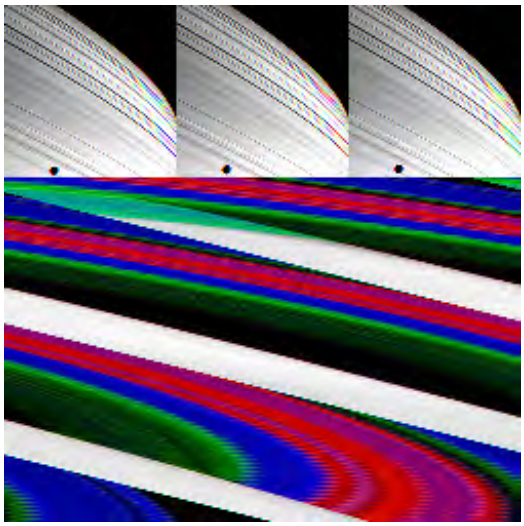


Here are Cassini views of the shepherd moons Prometheus and Pandora:

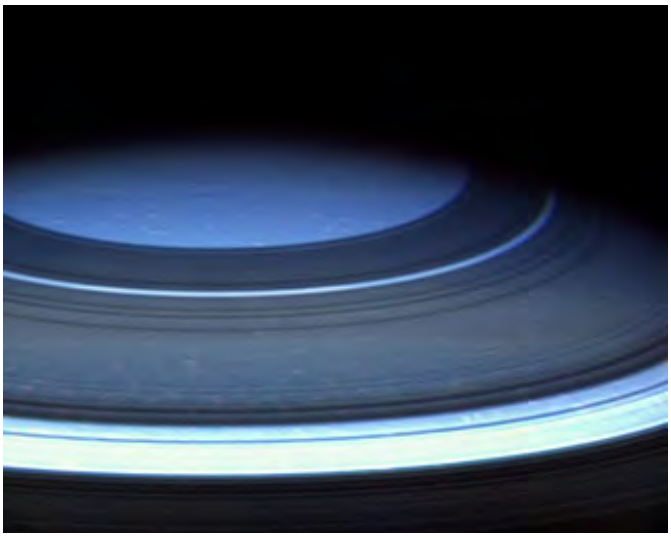
Athough some data relating to the mean temperatures in the individual rings had been obtained prior to Cassini's visit, those data were refined to give more exact values. A temperature profile was obtained during passage and used to assign colors to the rings by extrapolating the strip profile to the entire rings. In the image below, blue represents a temperature range clustered around 70° C (-333 ° F), green 90° C (-298° F), and red 110° C) (- 261 ° F):



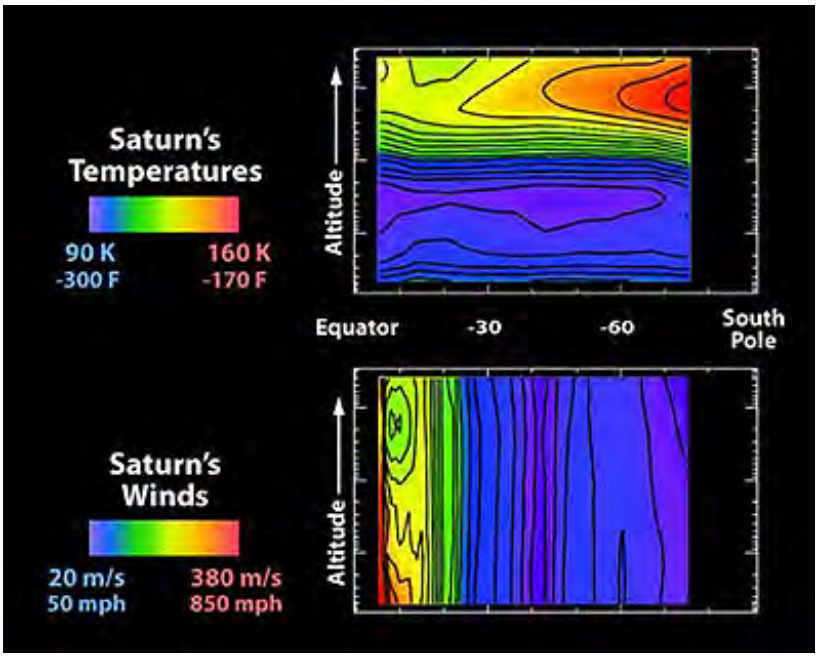
This next Cassini image shows a novelty - yet not unexpected. Under certain solar lighting conditions, sunlight is variably absorbed by Saturn's rings. The result is a series of thin light to dark bands on the saturnian surface, representing shadows. B and C ring shadows appear in this image; the black dot at the bottom is the shadow image of the moon Mimas.



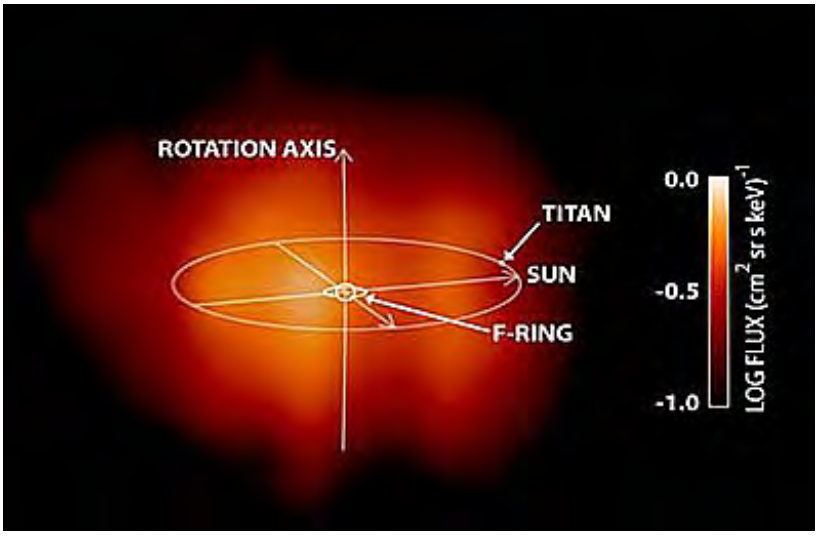
Cassini has also produced near-true color images of Saturn's surface. This next image shows shades of blue in the northern hemisphere. This color is due to scattering by gases, much like that observed in the Earth's sky during a clear day. The variable shading is caused by ring shadows.



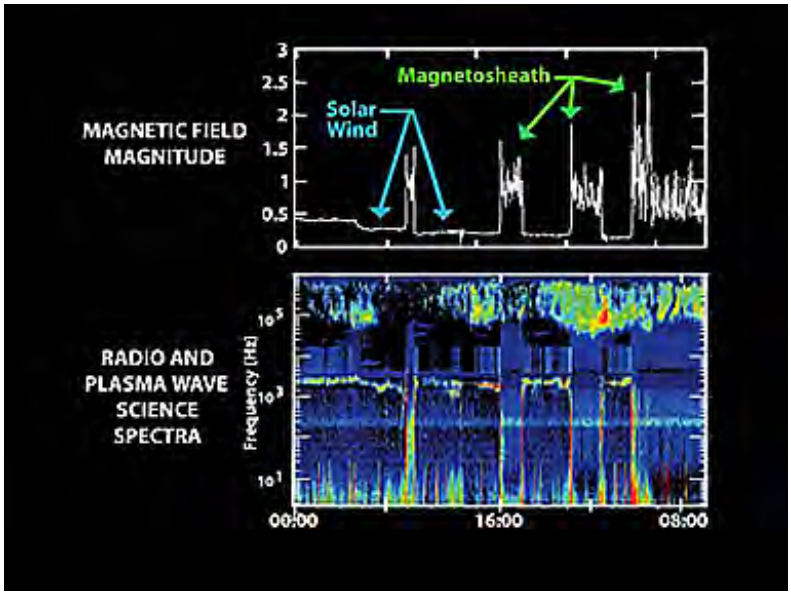
Temperature and wind speed variations in the atmosphere above the saturnian clouds are shown in this next diagram. Highest temperatures are towards the upper atmosphere; strongest winds at this time in the saturnian year are near the equator.



As Cassini approached Saturn, its Ion and Neutron Camera was able to take data that lead to a determination of the Magnetosphere and Magnetopause. In the image below, excitation of hydrogen ions gives rise to the orange glow that indicates the extent of the magnetosheath.

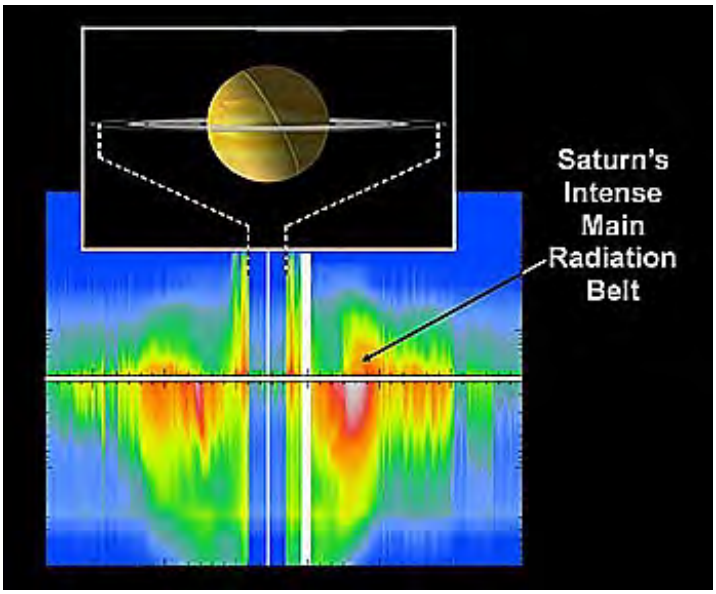


During the approach to Saturn Cassini's magnetometer encountered several magnetic spikes that are equivalent to magnetopause bow shocks within Saturn's magnetosphere.



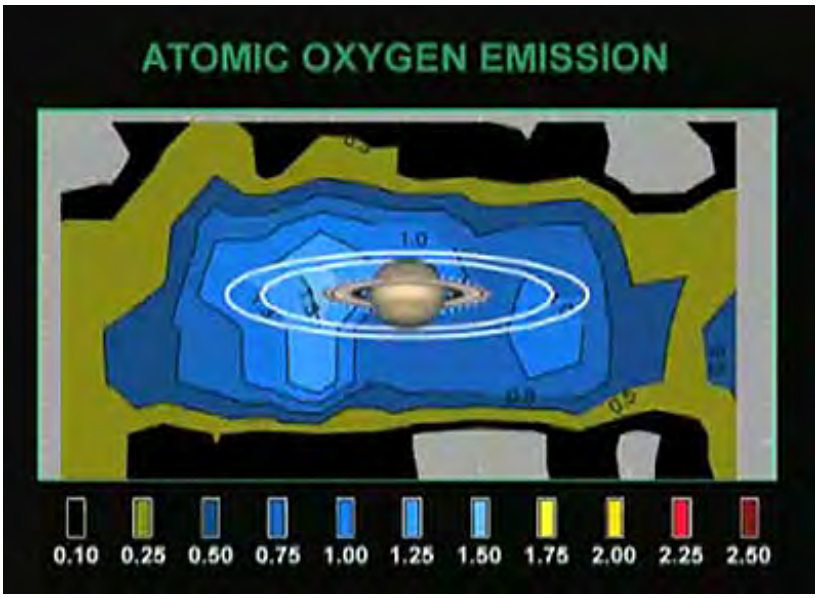
The magnetometer detected a very small secondary magnetic field around Titan, induced by interaction of its interior with Saturn's field.

From a far distance as it approached during its first major swing around Saturn, Cassini's Magnetosphere Imager has measured the distribution of trapped radiation in the Main Magnetic Belt around the planet, as shown below



This belt begins (from the surface) out at 70000 km (48000 miles) and reaches to about 783000 km (489 miles). Cassini also detected a previously unknown inner belt near the surface that is about 6000 km (4000 miles) thick.

One of the discoveries, made by the UV spectrometer, is that there is a small but detectable concentration of oxygen over much of Saturn's environment beyond the rings. One tentative hypothesis: Ionic bombardment of the E ring caused a sequence of reactions in the ice particles that led to free molecules of oxygen that dispersed in the pattern shown here that covered much of December 2003. Readings taken in 2004 showed an increase in oxygen for several months, but the amounts may now be abating.



There are now so many informative images from the Cassini mission that it is helpful to add another page. So proceed to page 19-19a by clicking on Next below.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Astronomers from ancient times until Galileo knew almost nothing scientific about stars and planets except that the latter moved in a regular fashion through the skies. One of the first examples of planetary remote sensing was Galileo's use of a primitive telescope to discover the Moon's craters and the moons of Jupiter. In the early 20th century, Hubble learned that most stars were actually galaxies - clusters of billions of stars. But it took the space program, with its probes to and orbiters around the planets, to open up the other planetary bodies in the Solar System to a systematic examination. The wealth of knowledge this has brought, largely through the remote sensing devices carried on the spacecraft, has given astronomers today remarkable insights into the nature and history of the planets. This Section will convincingly underwrite that statement. The first page reviews the different sensors and parts of the spectrum used in these great advances.

THE SOLAR SYSTEM AND PLANETARY EXPLORATION

The Role of Remote Sensing Techniques

The main thesis of this Section on the Exploration of the Solar System is simply that most of it is carried out using the tools and instruments associated with Remote Sensing. The primary sensors have been the camera, scanners, or comparable imaging devices. Also, geophysical instruments were emplaced on the surface by the Apollo astronauts. But much has been learned from the sample returns brought about by the astronauts who went to the Moon and the study of rare extraterrestrial meteorites found on Earth.

Remote sensing by imaging, as applied to Earth, goes back to the middle of the 19th Century, when balloonists took the first photos. As applied to the rest of the solar system, we must look to the first observations (documented by sketches) made by Galileo in 1610, when he turned a telescope to the heavens and caught a glimpse of the surface complexities on our nearest neighbor, the Moon. Later, he confirmed the Copernican theories with his discoveries of moons, or orbiting satellites, around Jupiter. Since then, we have many observations of our Solar System neighbors, first with telescopes and, after the opening of the Space Age, with orbiting spacecraft, flyby, probe, and lander missions. Nowhere else in the diversified and imaginative programs of NASA and other space agencies from different nations has there been such a plethora of observational and scientific triumphs as the exploration of the planets and the Cosmos beyond.

Most of the same instruments that survey the electromagnetic spectrum (EM) making up the radiation emanating from the Earth have been the principal tools for exploring our planetary associates and beyond; searching well into outer space at stars and other members of the Universe. Here is a list of remote sensing methods using EM spectral measurements that have provided exceptional information about planetary surfaces, atmospheres, and, indirectly, interiors: *

METHOD	EM SPECTRUM	INFORMATION	INTERPRETATION	MISSION

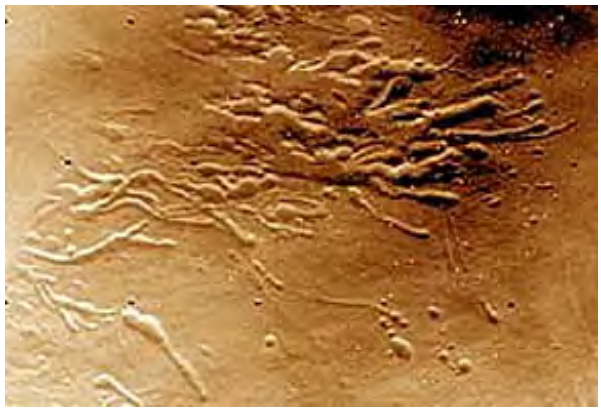
Gamma-Ray Spectroscopy	Gamma rays	Gamma spectrum	K, U, Th Abundances	Apollo 15, 16: Venera
X-ray Fluorescence spectrometry	X-rays	Characteristic Wavelengths	Surface mineral/ chemical comp.	Apollo; Viking Landers
Ultraviolet Spectrometry	UV	Spectrum of Reflected sunlight	Atmospheric Composition: H,He,CO_2	Mariner; Pioneer; voyager
Photometry	UV, Visible	Albedo	Nature of Surface; Composition	Earth Telescopes; Pioneer
Multispectral Imagers	UV, Visible, IR	Spectral and Spatial	Surface Features; Composition	On most missions
Reflectance Spectrometers	Visible, IR	Spectral intensities of reflected solar radiation	Surface Chemistry; mineralogy; processes	Telescopes; Apollo
Laser Altimeter	Visible	Time delay between emitted and reflected pulses	Surface Relief	Apollo 15,16,17
Polarimeter	Visible	Surface Polarization	Surface Texture; Composition	Pioneer; Voyager
Infrared Radiometer (includes scanners)	Infrared	Thermal radiant intensities	Surface and atmospheric temperatures; compos.	Apollo; Mariner; Viking; Voyager
Microwave Radiometer	Microwave	Passive microwave emission	Atmosphere/Surface temperatures; structure	Mariner; Pioneer Venus
Bistatic Radar	Microwave	Surface reflection profiles	Surface Heights; roughness	Apollo 14,15,16; Viking
Imaging Radar	Microwave	Reflections from swath	Topography and roughness	Magellan; Earth systems
Lunar Sounder	Radar	Multifrequency Doppler Shifts	Surface Profiling and imaging; conductivity	Apollo 17
S-Band Transponder	Radio	Doppler shift single frequency	Gravity data	Apollo
Radio Occultation	Radio	Frequency and intensity change	Atmospheric density and pressure	Flybys and Orbiters

* Adapted from Billy P. Glass, *Introduction to Planetary Geology*, 1982, Cambridge University, Press

This list is incomplete but is still highly representative. The exploration of the planets, while dominated by remote sensing devices, is also supported by some non-remote sensing methods. Chief among these is landing astronauts on the Moon to observe first hand, deploy instruments, and collect samples. Landers have set down on other planetary bodies as well. So far in the study of stars and galaxies, the methods used have been entirely remote sensing, as will be evident in the Section 20 review of Cosmology.

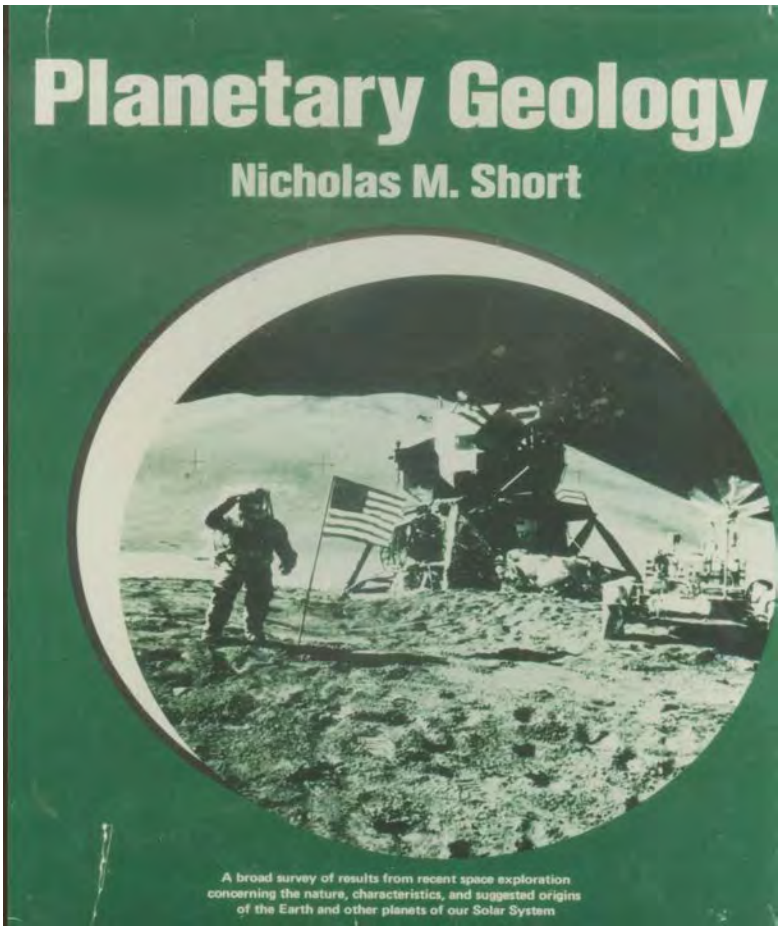
The Command and Service Module on the Apollo lunar missions carried a complement of remote sensors and other instruments including alpha-particle spectrometers, mass spectrometers, magnetometers, far UV spectrometers, scintillometers, and others designed to measure geochemical and geophysical properties. The astronauts also deployed, on the surface, instruments for specific studies. Among these were seismometers, magnetometers, gravimeters, solar wind gauges, cosmic-ray detectors, heat flow probes, and laser ranging retroreflectors. However, in retrospect, sensors that produce images, especially photographs and similar items, have provided the most direct and readily interpretible sets of data, and will continue to be a mainstay of future missions.

While remote sensing, especially in the optical or visible segment of the EM spectrum, is a mainstay in planetary studies, the resulting data still need to be interpreted. The new observations of a planet's or moon's surface tend to reveal exotic features which at first seem alien to those who live on Earth. Yet the very familiarity of the Earth to these observers is often the key element in explaining extra-terrestrial features, since Earth's surface has been well explored and documented visually. The Earth then is the "frame of reference" that commonly provides features resembling those on other planetary bodies; and, much is normally known about the mode of origin and development of these features. This approach has been termed "**Comparative Planetology**". As an example of how one proceeds in identifying and describing a geological feature on another planet in terms of a terrestrial counterpart, consider these two views of channels on Mars and then a similar set of channels on the Earth (in Africa):



The Chad volcano has been studied in the field, so that the role of running water in carving out the channels shown (they tend to follow fractures) is well documented. Note the similarity in morphology to the two martian sets of channels. This close resemblance illustrates the type of argument planetologists use to explain martian channels: those channels look like terrestrial channels - they probably have similar origins (this still is inference rather than firm proof).

Comparative Planetology, along with Remote Sensing, really began to take off with the space program. Satellites and flyby probes, and then landers, could visit the planets closeup. No longer was it necessary to rely on poorly resolved images made through earthbound telescopes to learn important details about the planets in our Solar System. New subject fields started to emerge in the 1960s. They were given various names, such as Planetology (which was slanted from an astronomical perspective), Astrogeology (which actually has little "astro"), and Planetary Geology (which emphasizes the geological aspects of the "comparative"). To the best of my knowledge, I, the writer (NMS) wrote the first English language textbook on Planetary Geology, which was published in 1975.



Before proceeding, it may be helpful to you to visit and browse a website that deals with (mostly NASA's) Solar System programs - past, present, and future. This website displays many (but not all) relevant missions, including those dealing with earth-observations and with astronomy, by displaying each mission as a panel logo; click on any of these to access a mission description. Check, too, the Nine Planets and Solar View websites that list most of the spacecraft sent to other planets and solar system objects. To see a large collection of images of the nine planets, go to JPL's Photojournal website, and click on the planet of interest. Then check out one of JPL's movies. Access through the JPL Video Site, then follow the pathway Format-->Video -->Search to bring up the list that includes "Interplanetary Superhighway", July 17, 2002. To start it, once found, click on the blue RealVideo link.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

The last two Gas Planets, Uranus and Neptune, are less than half the diameter of Jupiter and Saturn. Both appear a near-sky blue in colored versions constructed from Voyager 2 images. Uranus has five large satellites (and 17 small ones) and Neptune two, plus 6 small ones discovered by Voyager. Both have ring systems, so that all four Giant Planets possess this feature. On this page all of the five large uranian satellites, each with an outer shell of ice, are pictured. Neptune's satellites are examined on the next continuing page.

Uranus and Neptune, and their Satellites

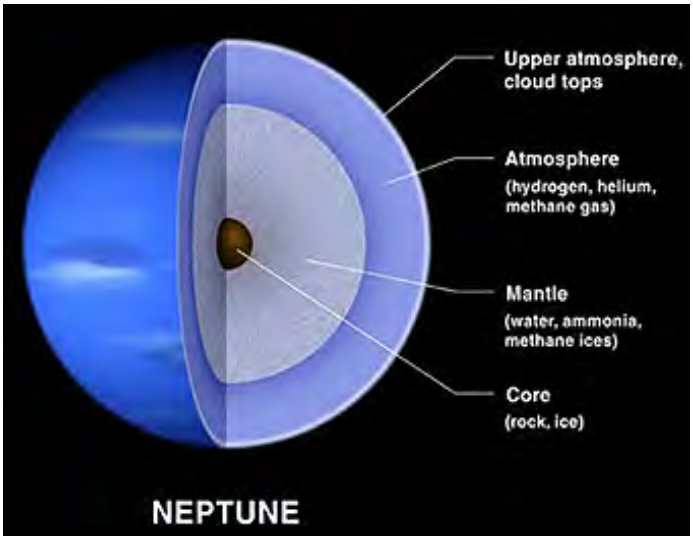
From Saturn, Voyager 2 moved onward, to pass Uranus and Neptune. As with Saturn, we suggest you take a quick look at the U. of Arizona LPL summaries (by Bill Arnett) of Uranus and Neptune.

As we noted before, both of these outer Giants are about 40% the diameter of Jupiter. Uranus and Neptune have densities of 1.28 and 1.63 g/cm^3, respectively. Both have thick cores composed of rock mixed with water, ammonia, and methane ices, lack metallic hydrogen mantles, have thick gaseous envelopes (about 15% of the planet's mass) consisting of about 83% H, 15% He, 2% CH_4 (methane) and traces of other gases. They also have rather faint, thin rings, and strong but distorted magnetic fields. Uranus has 22 known satellites, of which 5 are greater than 450 km (280 mi) in diameter. (Interestingly, many of the satellites discovered by Voyager 2 are named after well-known Shakespearean characters from his plays.) Neptune has 8, of which 2 are greater than 400 km (249 mi) wide. In this discussion, we describe only the larger ones.

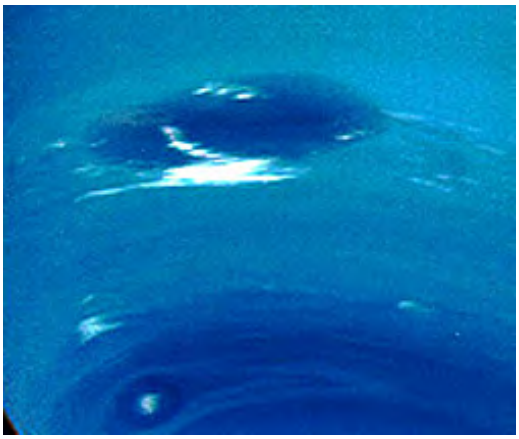
Seen in color by Voyager cameras, both planets appear similar, as shown here (Uranus on left; Neptune right):

Both Uranus and Neptune have small rocky cores, a thick mantle made up of a mix of water, methane, and ammonia ice, and a thick atmosphere of hydrogen, some helium, and methane

clouds. This cross-section of Neptune applies similarly to Uranus:

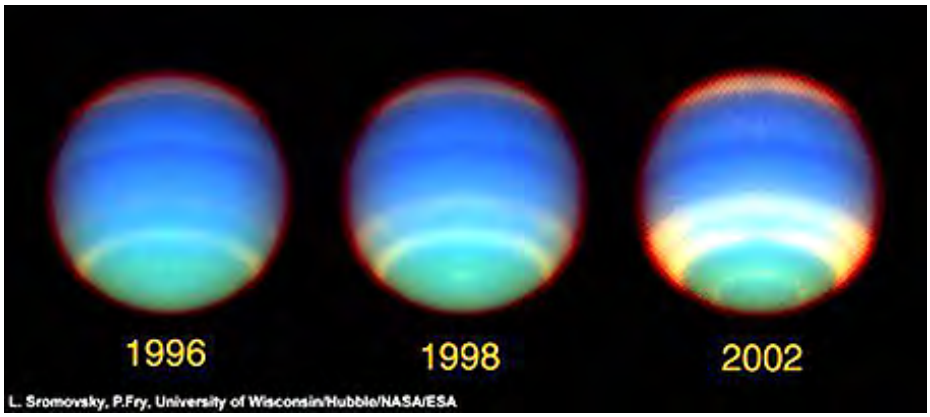


Each has a notably bluish color, because methane gas absorbs longer wavelengths (reds and yellows) and reflects shorter wavelengths (blues). The atmospheres possess weak banding and Neptune displays a large Dark Spot and fast moving cloud spots that represent rising plumes and turbulence. Neptune's zonal winds can move as fast as 900 km/hr–the fastest measured, so far, in the Solar System. Neptune's Great Dark Spot, shown in detail below, is accompanied by an elongated bright patch (nicknamed "Scooter"). It is evidence of a storm at the time of the Voyager flyby. In the late '90s, observations by the Hubble Space Telescope revealed the Spot to be gone (unlike the Great Red Spot on Jupiter).



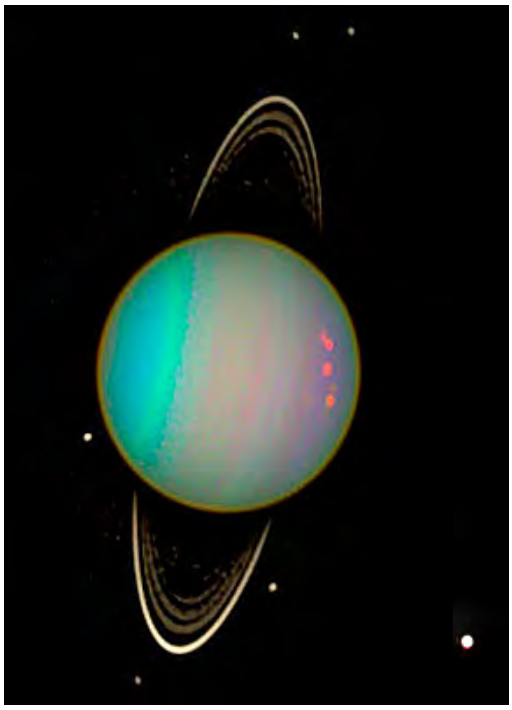
19-68: Why are Uranus and Neptune different in color from Jupiter and Saturn?
ANSWER

Neptune takes 165 years to fully orbit the Sun. Its axis is tilted 29° from the ecliptic. For this reason, like Earth, scientists postulated seasonal changes in the weather patterns exhibited by the bands. If these seasons are analogous to the four on Earth, about 41 years is needed to complete a neptunian season. After 7 years of observations by HST, these changes - brightening and broadening of the bands first noted - seem to confirm that such changes are happening.



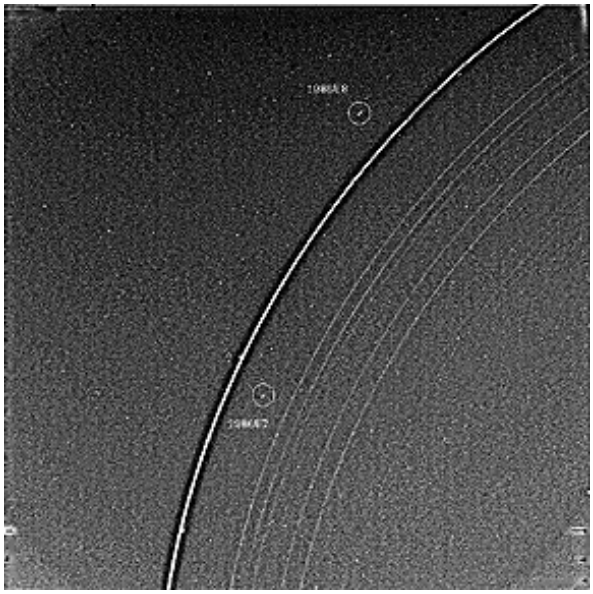
Both Uranus and Neptune have rings, which are comparatively faint and thin. A ground-based telescope image made by a European Space Agency telescope seems to indicate Uranus' rings are as well-developed as Saturn's.

The rings are actually much more tenuous than those of Saturn. This enhanced Hubble Space Telescope of Uranus highlights both the variations within the atmosphere and its most prominent ring. Note too that this image shows the planet in its proper orientaton, with its rotation axis "flipped" about 90 ° so that it is almost in the plane of the ecliptic along which it orbits.

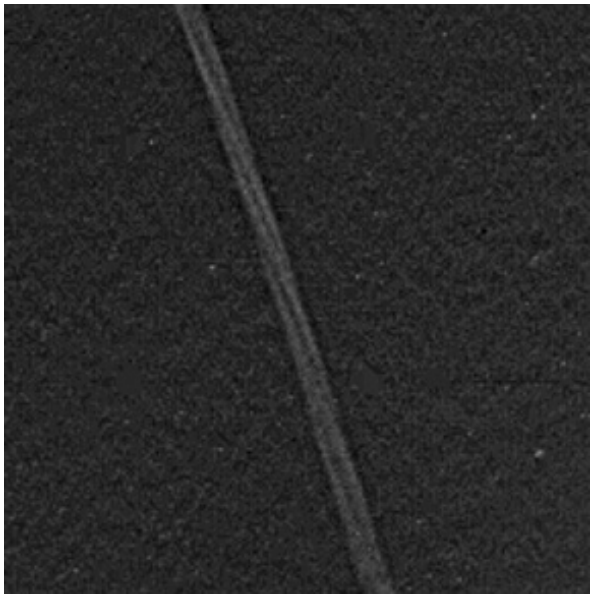


This view shows details of the ring system:

All told, Uranus has 11 rings, with the outermost (Epsilon) being much brighter than the others, as shown above. Two small satellites are visible on either side of the Epsilon ring.



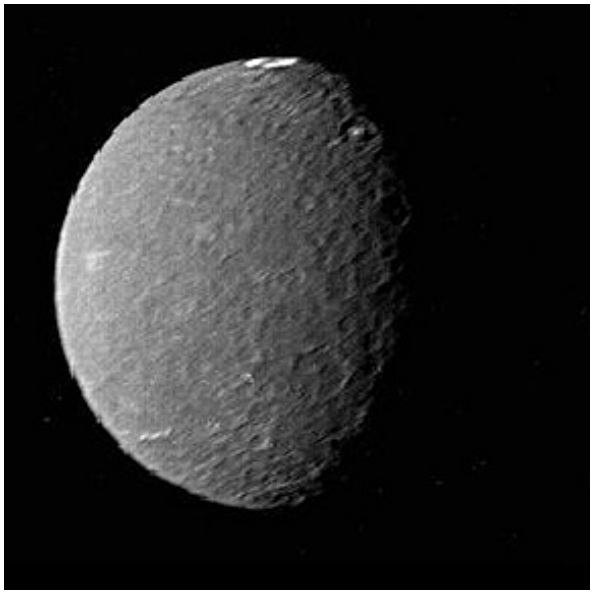
Scientists call these satellites, "shepherds," because they seem responsible for keeping that ring intact, by pushing stray particles back into line and possibly gathering and inserting particles from beyond. One of the rings around Neptune has a distinctive "braided" (twisted) structure:



The five outermost uranian moons are also that planet's largest. They occur in this sequence (from the left): Miranda (472 km [293 mi] in diameter), Ariel (1,158 km [720 mi] in diameter), Umbriel (1,172 km [728 mi] in diameter), Titania (1,580 km [982 mi]in diameter), and Oberon

(1,524 km [947 mi] in diameter). Oberon is the farthest from Uranus at 583,300 km (362,463 mi). Each has an icy surface, extending to unknown depths, making up about 40-50% of the volume but ending in a rocky interior.

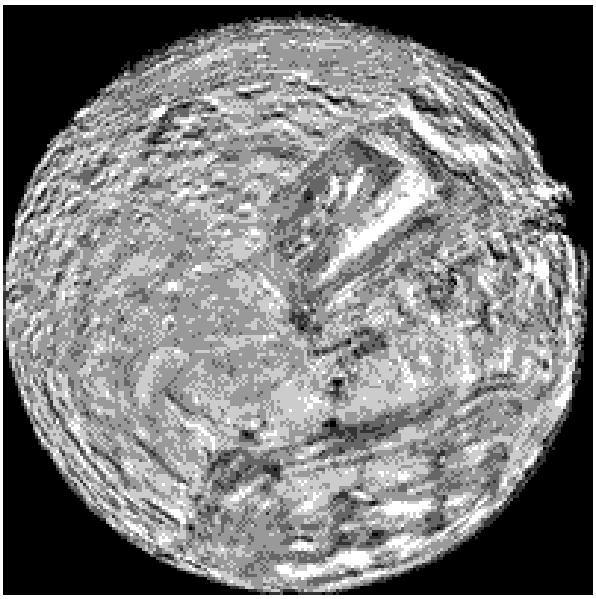
The last four comprise a group, in which all show strong similarities, with their dominant features being craters (especially Umbriel and Oberon) and deep (some greater than 10 km [6.2 miles]), interconnected, gash-like valleys (particularly Titania and Ariel).

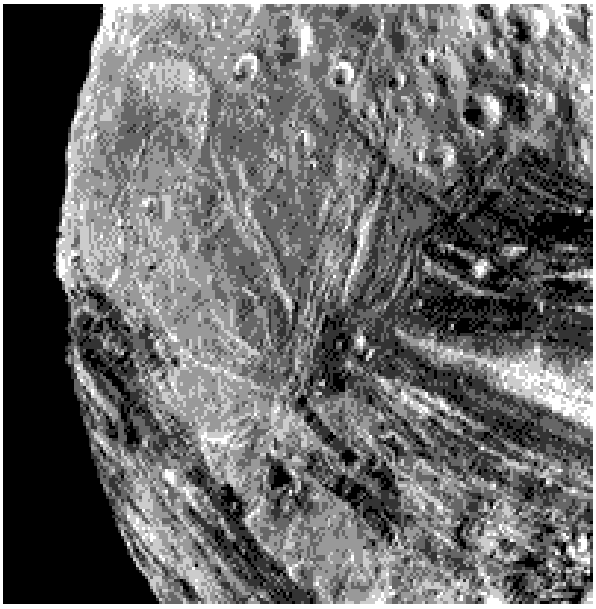


One hypothesis for the origin of these cracks contends that the satellites once had enough internal heat to melt the exteriors into liquid water, which induced tensional fracturing, when it later froze and expanded the ice.

19-69: Which of the four above satellites appears to be older? ANSWER

Miranda is unique among the uranian satellites but, excepting its much smaller size, strongly resembles Ganymede in the Jovian group. By serendipity, Voyager 2 made the closest approach to Miranda of any of the satellites, so it could image the peculiar terrains in detail. Investigators based this flyby decision on the need to use this satellite to give the spacecraft a gravitational boost towards Neptune. First, look at the full view (left), a mosaic of one face. Then (right), view a close-up of the surface.

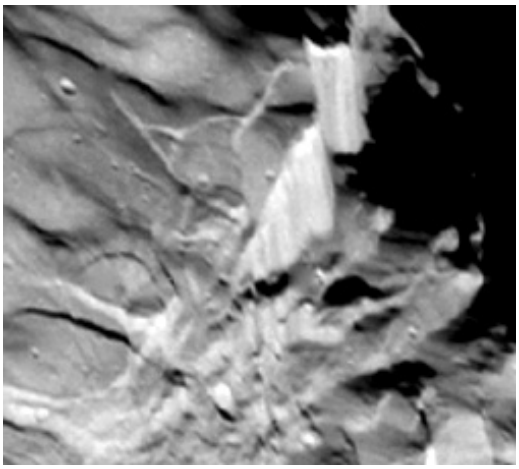




Two sharply contrasting terrains are evident. Much of Miranda's surface consists of rather bright, heavily cratered terrains, similar to the above four satellite exteriors. But there are three large oval-to-rectangular terrains (called coronae), whose physiographies are quite different (to each other and to the dominant terrain). They may contain subparallel grooves, several as deep as 2 km (1.2 miles), or alternating light-dark bands, or ridges and stripes that take sharp bends. A closer view (at bottom, above) of ridged terrain shows the generally abrupt boundaries of these terrains with the encompassing older terrain. Scientists still debate the origin of these patches. An early interpretation considered them to be fragments of a shattered proto-Miranda or some other disrupted satellite, incorporated in a re-assembling Miranda. The paucity of craters in these isolated terrains argues against this. Another opinion holds them to be a mix of rock and ice that formed a crust, which broke up and was invaded by upwelled ice masses from the interior.

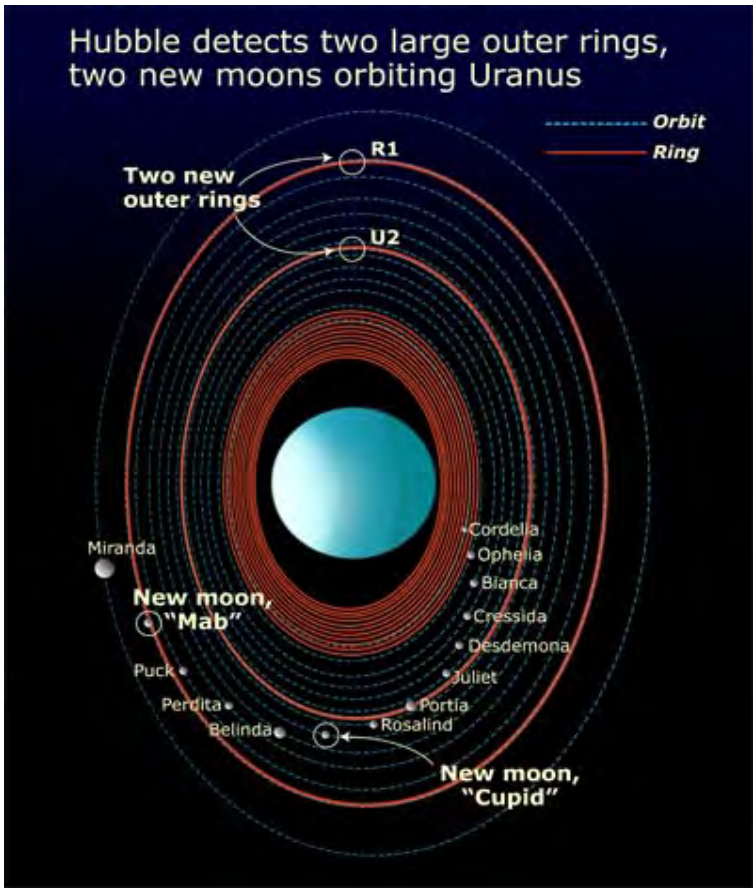
19-70: Develop an argument against the impact disruption hypothesis. ANSWER

Still another close view of the Mirandan surface shows high fault scarps and grabens in jumbled terrains.

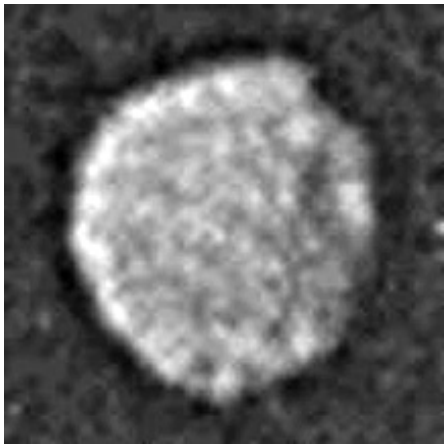


The Voyagers were unable to get many close views of the larger uranian satellites. Here is a view of Ariel's surface:

There are 13 small uranian moons inside Miranda's orbit. They are identified in this schematic:



Most of the seven small satellites discovered by Voyager (six more have since been found in photographic plates exposed through Earth-based telescopes) were not sharply imaged. An exception is Puck, about a 200 km diameter object, shown below. At least some of these bodies are probably captured asteroids.



PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Just in our Solar System alone there are billions of moving bodies besides the planets and their satellites. One class consists of rocky bodies of varying compositions: carbonaceous chondrites, chondrites and metallic iron objects (similar to the meteorites that fall on Earth). Most of the meteorites come from the asteroids whose sizes range from less than a kilometer to hundreds of kilometers. Nearly all have irregular shapes. Most lie as a swarm of unconnected bodies in the Main Asteroid Belt between Mars and Jupiter. Others occupy different orbital regions, some of which periodically cross Earth's orbital path. The other class consists of comets which are relatively small "dirty ice balls" or a nucleus of ice and rock. As those with certain orbits approach the Sun, the solar wind and other forces cause them to ablate, developing a small local atmosphere or coma and eventually a tail of particles which along with the coma glow brightly from sunlight. Comets have a wide range of orbits, some near the solar ecliptic, others at high angles, but all still under the influence of the Sun's gravity. The Oort Cloud and the Kuiper Belt are two sources that lie well beyond the orbit of Neptune. Long-Period comets come from great distances such as these sources or elsewhere; short period comets tend to have orbits much closer to the distances defined by planet paths. Both comets and asteroids occasionally strike the Earth. Programs to detect possible bodies that could threaten Earth are now activated. Several missions to comets and asteroids are described.

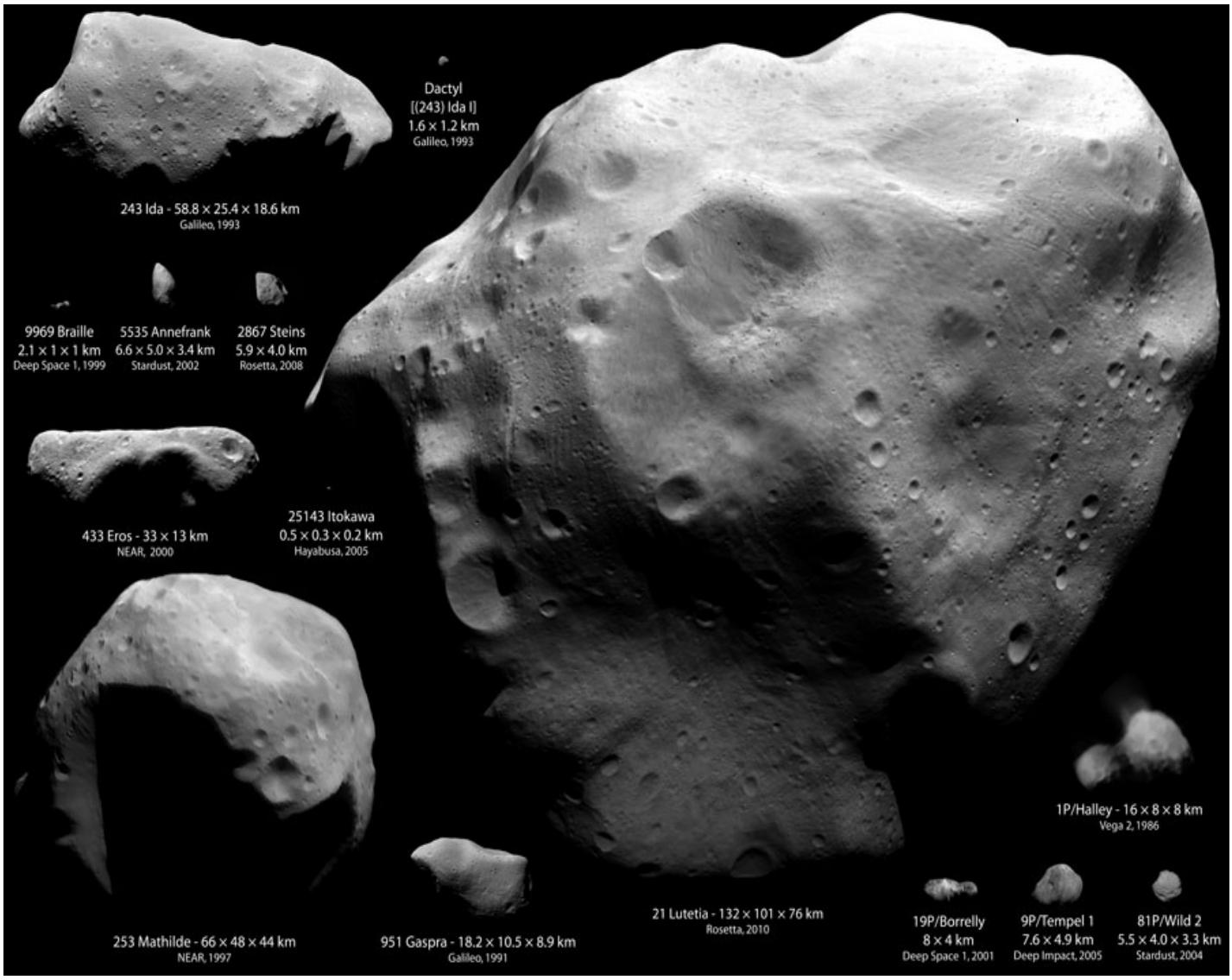
Asteroids and Comets

This page will be a thorough review of the basic knowledge about asteroids and comets. But you can supplement this information by clicking on the Wikipedia sites that cover Asteroids and Comets.

A very large number of small (less than 1500 km maximum dimension) bodies composed of rock/dust, ice and condensed gases orbit around the Sun as *asteroids* and *comets*. The asteroids are composed mainly of rock, with some ice, and are detectable by solar light reflected from their low albedo surfaces. Asteroids are mostly found in discrete, nearly circular belts that orbit at about the same distance from the Sun over long periods of time; some asteroids escape their belt and wander through the Solar System - thus, a few have hit the Earth and others remain a threat in the future. The nucleus of a comet generally is dominantly ice with some rocky material that ablates off the surface into one or more "tails" that become visible because solar light reflects from an envelope of gases and rocky particles that stream from the nucleus. Comets have more elliptical orbits that extend well out into the Solar System but may approach and pass around the Sun - a few have struck Earth in the past and more may strike in the future

Three good overviews of asteroids and comets are found at these sites: (1), (2), and (3).

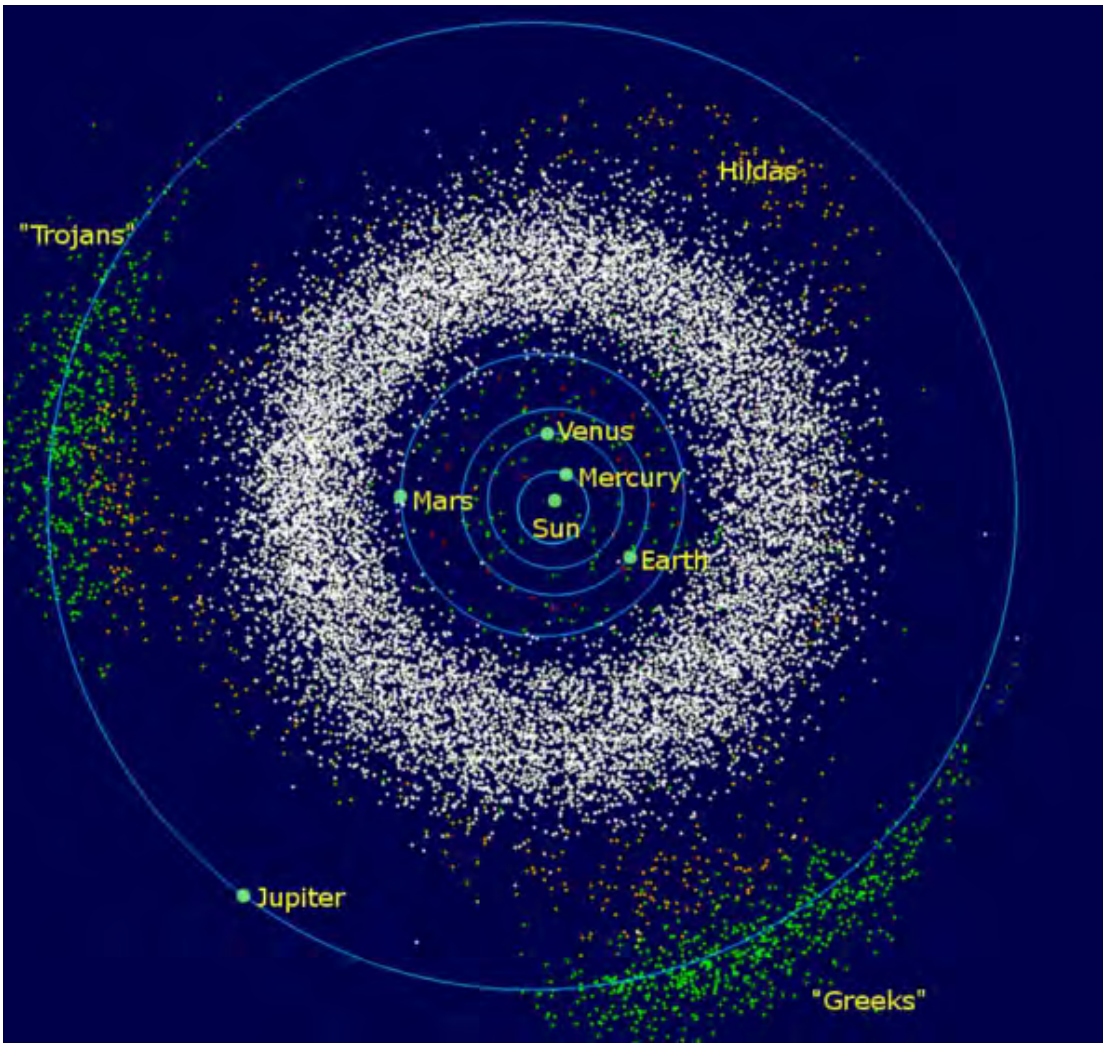
Several missions from Earth to the outer planets have, either as their primary goal or serendipitously, imaged named asteroids. This montage shows some of these:



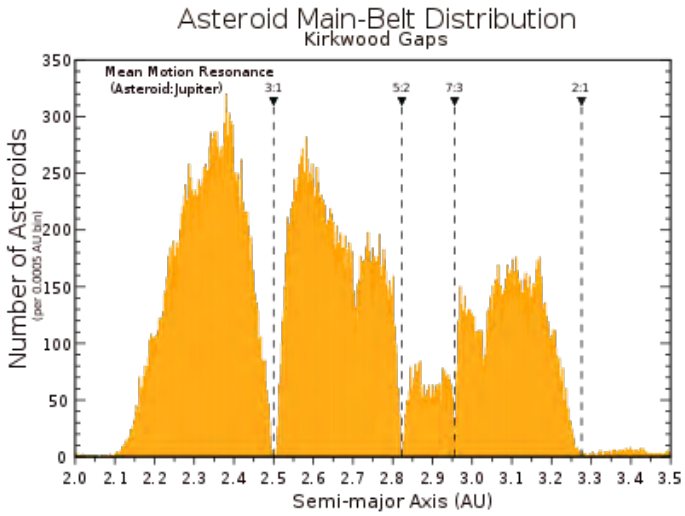
About 10000 **Asteroids** have been discovered as individuals so far. Most are located either within the "Main Belt" or the "Kuiper Belt". Some astronomers have estimated that at least 400,000 bodies larger than 1 km exist within the Solar System; the total number, including small ones is likely well in excess of a million. More are being found each year owing to a stepped-up search program (see below). The number of known comets is smaller.

In the 20th Century, the *asteroids*, solid fragments of varying size that orbit in several regions of the inner Solar System, have had dramatic increases in interest, in part because they are the more frequent bodies that have collided with the Earth and the Moon, and most other larger solar bodies, producing impact craters which on our planet could have (and have had) diastrous consequences - castastrophic in terms of effect on life - of magnitudes only now being properly appreciated. We will start this review with the nature and distribution of asteroids.

Tens of thousands of asteroids occupy the Main Belt, an interval of mean distance 2.4 A.U., between Mars and Jupiter. This diagram shows the belt in relation to the planets, along with several other asteroid clusters closer to Jupiter.



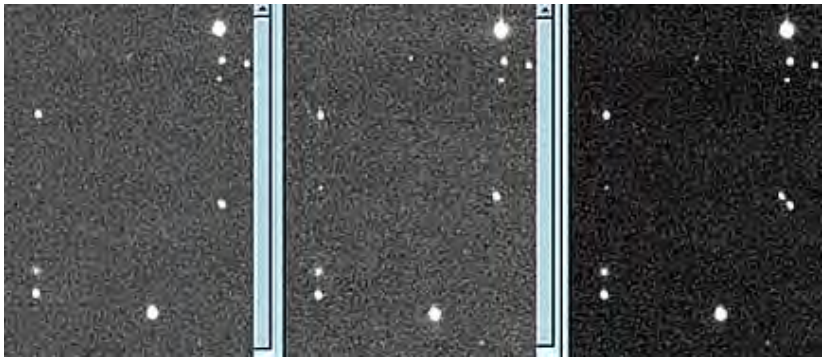
Contrary to the above illustration, the asteroids are not uniformly dispersed. They form several belt swarms, with reduced numbers in between. These lower regions are called Kirkwood Gaps:



The average size of a Main Belt asteroid is about 100 km. Typically, the separation between individual asteroids in the belt is about 1600 km (1000 miles). Telescopic observations have indicated that about 75% of these asteroids have very low albedos, which is consistent with compositions that are similar to carbonaceous chondrite meteorites, generally considered to be the most primitive Solar System material. Other (about 17%) asteroid groups occur in different locations, such as the Apollo belt (Earth-crossing) and Trojan belt (Jupiter-crossing). Many of the Trojan's are darker and redder. As they orbit the Sun, they also rotate, some with periods of a few days to a week or so and others in a matter of hours ("tumble"). Mars and Neptune also have some associated asteroids that cross into their orbital zones.

Comets are easier to detect than smaller asteroids. Comets have higher albedos, and thus are brighter, and as they enter the region of the Solar System within the outer planets they develop luminous tails. Asteroids, on the other hand, present greater difficulties in finding them and determining their orbits. Most asteroids have such low albedos that they don't begin to reflect enough sunlight until relatively close to Earth. Only the larger ones in the Main Asteroid Belt are fairly easy to spot and measure. To illustrate detection techniques, let us describe how the Near Earth Asteroids (NEA), the Apollo belt, are found and monitored to determine their size and orbits.

Optical telescopes, with mirrors around a meter in diameter, are the instruments most commonly used. A portion of the sky (usually around 1 to 10 Moon diameters in width) is viewed over a short time interval (hours to a few days). Successive views are recorded on film or electronically using CCDs. Stars and galaxies will retain their fixed positions relative to each other during the interval. An asteroid will look much like these fixed distant stellar/galactic bodies if it is close enough to "shine". But being very close with respect to the distant background, it can move a measurable distance in that time. Consider this trio of telescope photos:

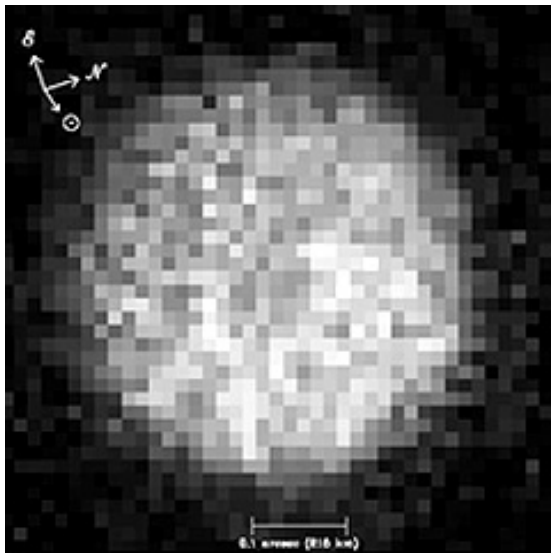


In the left and center photos, all the starlike objects appear to be in the same fixed positions. But one such body, in right center, actually moves to a slightly new position, towards the upper left, in the middle photo. When the two photos are superimposed, as shown in the right image, two "stars" appear instead of one. The left "star" in the pair is just the same body that has moved in this time interval (30 minutes). This is explained by assuming the body is so close to the Earth that its motion is discernible over short time periods. The presumption is that it is an asteroid. Working with that idea, the actual distance of the object from Earth is determined by parallax (using two telescopes). Once that has been calculated, the size of the body can be estimated (taking its albedo into account) and its velocity is likewise determined from the distance/time relation. Repeated observations allow its orbital parameters to be specified. The painstaking search efforts involved (if done manually) has been eased by computer programs that automatically seek out diagnostic displacements.

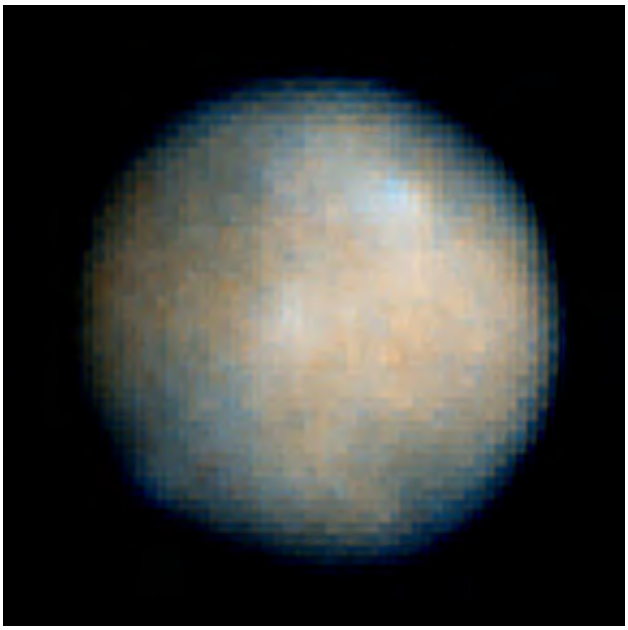
The Earth-crossing asteroids all are less than 10 km in maximum dimension. From actual counts, plus extrapolations, the number of Earth-crossing asteroids greater than 1 km in diameter is now placed between 700 and 1000; a collision with Earth for one of that size is estimated to be around 1 per 1-2 million years and for a 10 km asteroid once every 100 million years. The majority of all asteroids have irregular shapes.

Among the Main Belt asteroids the biggest is Ceres (933 km [583 miles]); second in size is Pallas (530 km [331 miles]) and Vesta is slightly smaller; these all reside in the Main Belt. The larger asteroids tend to approach spherical in shape; Ceres is the only one presently known to be almost spherical. At least some of these have melted and differentiated, so that they have developed iron-rich cores, and possibly a mantle with dispersed metallic iron (the pallasite meteorites may derive from this). One school of planetologists considers these to be "small planets".

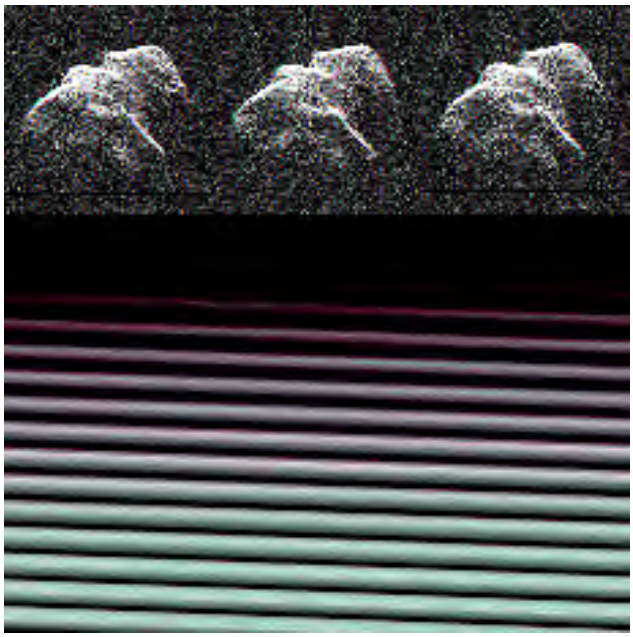
Asteroids have been imaged by conventional telescopes, by radar, and by passing or orbiting space probes. This first image shows that ground telescopes do not normally produce good images, even of the larger asteroids, such as Ceres.



A recent Hubble Space Telescope (HST) image has been reprocessed to give the best look yet at Ceres. This is important because this asteroid has now been tentatively proposed as a dwarf planet within the Main Belt asteroids.

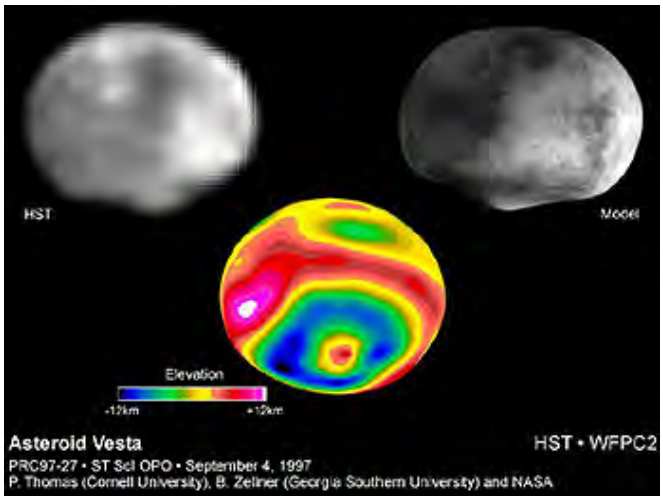


As implied on page 18-3, the dinosaur "final solution" from a huge asteroidal impact, followed by several movies that deal with world-threatening (and destroying) collisions with outer space debris, have led to a more organized search and inventory (and orbit-determination) for asteroids and comets. Optical telescope methods look for tiny light blips in photographs that move with significant displacements relative to fixed star backgrounds. This is turning up new asteroidal bodies - many quite small - with notable numbers added each year. Another approach uses radar: this image of Toutalis (4.6 x 2.4 x1.9 km [2.9 x 1.5 x 1.2 miles]) is a mosaic of several radar returns:



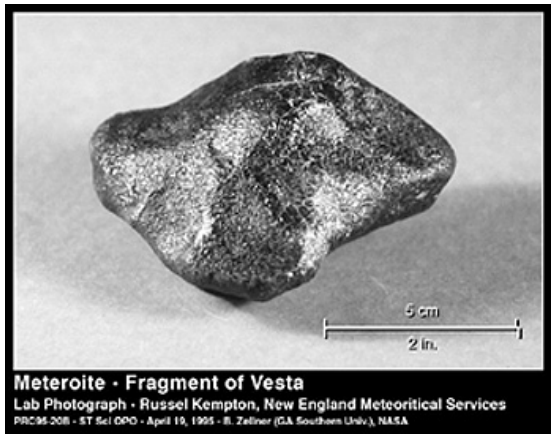
Radio astronomy is also adept at imaging larger asteroids. This image of asteroid 1999JM8 was made using the large dish at Arecibo in Puerto Rico:

Vesta, at 525 km (325 miles) in long dimension, has been imaged by the Hubble Space Telescope, as seen here:



Although not shown clearly, there appears to be a large impact crater with a central peak on Vesta. Spectral analysis of light from Vesta produces values that are very close to the mineral group of pyroxenes. In 1960, a meteorite was found in Western Australia that is entirely pyroxenite (making this a rare type). Many asteroid specialists believe this came from Vesta; if

so, it is our first sample of an asteroid that we can analyze.



Both ground telescope and even HST images of small objects such as asteroids can be blurry. Thus, to effectively image in detail an asteroid a space probe needs to visit its vicinity. The Galileo spacecraft was programmed to transit close to two asteroids. It passed the first, Gaspra, in 1991. This asteroid, 19 x 12 x 11 km (~12 x 7.5 x 7 miles) in size, consists of iron-nickel and iron-magnesium-rich silicates. This is how it appears in color:

Galileo approached and imaged the second asteroid, Ida, on August 28, 1993, as shown here (at about 33 m [108 ft] resolution):

This asteroid, about 58 x 23 km (36 x 14 miles) in dimensions, is chondritic and, like Ida, pockmarked with craters. Totally unexpected was the presence of a small orbiting body, about 1.5 by 1.2 km (0.93 x 0.75 miles), named Dactyl–, making this pair the first known binary asteroids. Close-up, Dactyl has a small crater, causing it to look a bit like Saturn's Mimas in miniature:

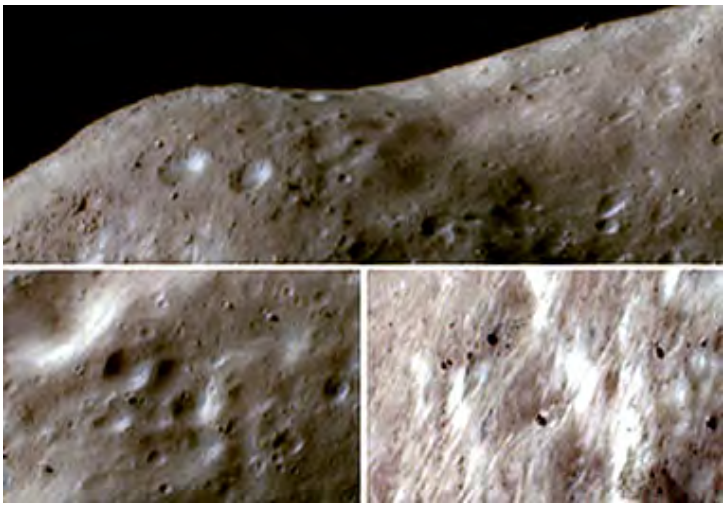
The first visit exclusively to the asteroid belt has been made by NEAR, for Near Earth Asteroid Rendezvous mission, launched in June 1997 to reach the large asteroid Eros in January, 1999. After launch, the spacecraft had the name Shoemaker added to it, in memory of the Dean of Astrogeologists, Dr. Eugene M. Shoemaker (see bottom of next page). Enroute NEAR-Shoemaker took a close look at Mathilde (59 x 47 km; 37 x 29 miles) on June 22, 1999:

Mathilde has one of the lowest albedos, less than 4%, of any asteroidal object. It also has a low density, 1.4 g/cc, indicating that much of its interior is either ice and/or voids - thus it has been described as a "fluffy" ball. Its solid component has probably a carbonaceous chondrite composition.

In 2000, NEAR-Shoemaker successfully orbited Eros and has taken data on its composition. Below is a view of Eros, whose dimensions of 33 x 13 x 13 km give it a peanut-like shape. Eros is rapidly rotating, as evident from the different positions over a short time span on a December 1998 date.

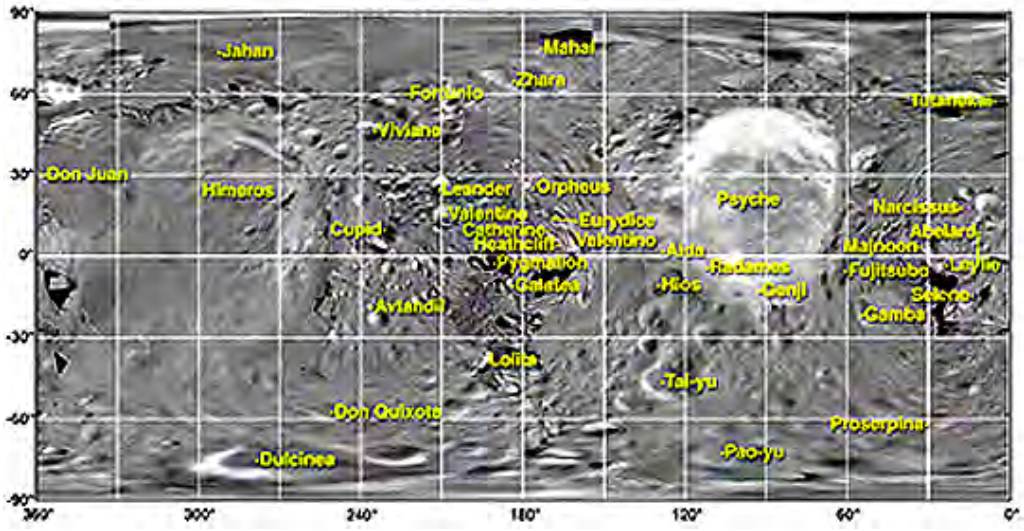
NEAR-Shoemaker has operated successfully around Eros for several years, taking thousands of images at various resolutions from different orbital heights. This next image is one of the best, taken in a false color mode in which a green and two NIR bands are used.

Here are three views of parts of EROS:

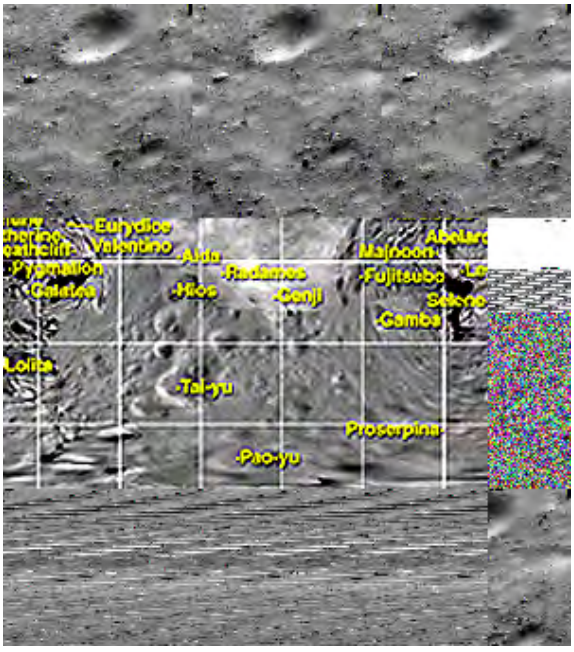


Here is a close-up (resolution: just a few meters) of a crater on Eros:

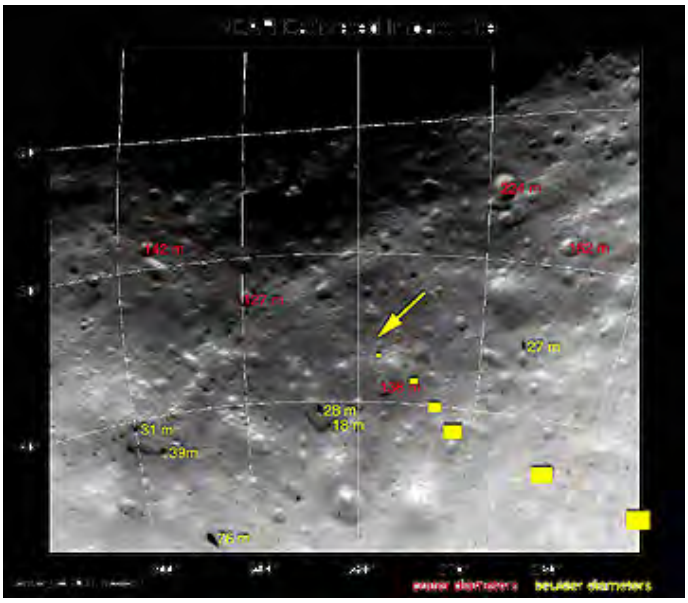
After completion of the NEAR-Shoemaker mission, the mappers gave names to the major features on a flattened mosaic of its surface, as follows:



NEAR-Shoemaker made a close approach to Eros, reaching an altitude just 6 km (4 miles) above its surface, on October 27, 2000. Here is part of a mosaic of images made during that pass. This view shows a crater and rock debris as small as 1 meter across:

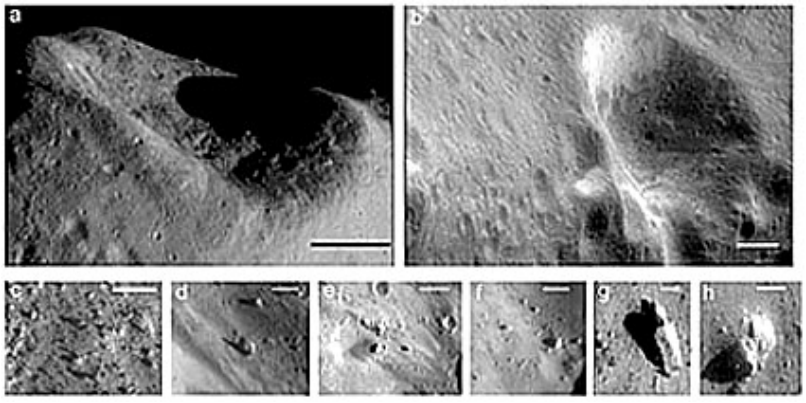


In early 2001, it was evident that the spacecraft's fuel was nearly depleted. The NEAR-Shoemaker scientists and managers pondered how to finish the mission and decided on a bold course - to land the spacecraft on EROS's surface. The asteroid was 196,000 miles from Earth; if successfull this would be the farthest any manmade object had set down on a solid body in the Solar System. Good operational maneuvering and some real luck were essential to the endeavor. Here is the target area chosen:

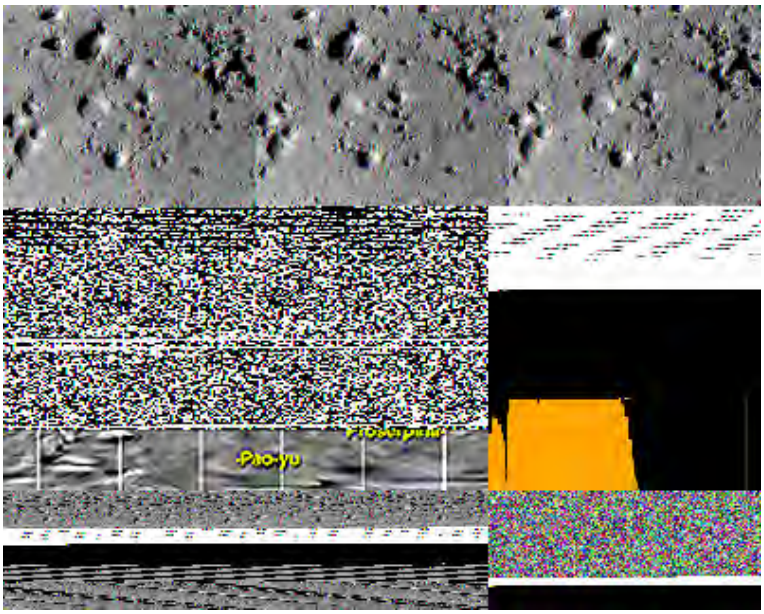


The landing attempt began in the afternoon (EST) of February 12. Five burns (fuel blasts) were involved over a 50 minute interval, each slowing the spacecraft and thus causing it to drop ever closer to the surface. The last burn dropped the descent speed to 6 km/hr (4 mph). All the while, images were taken and radioed to Earth. The spacecraft not only reached the surface but survived the touchdown and continued to send back signals.

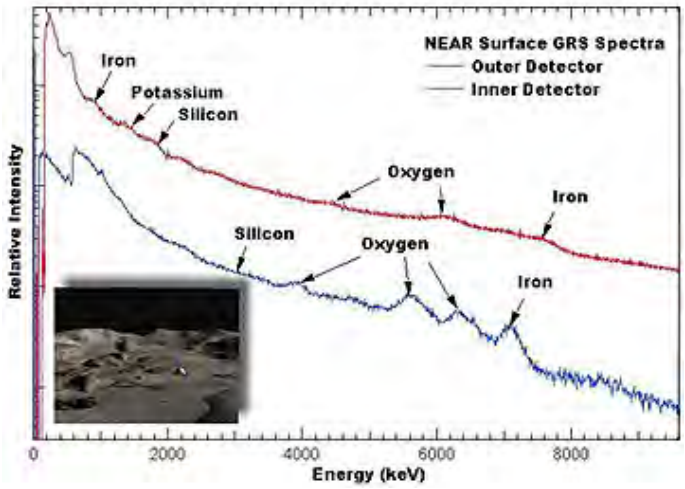
Pictures obtained as the spacecraft moved ever closer to its landing resemble in their increasing detail those returned from Ranger impacts on the Moon. Here is a sequence obtained as NEAR-Shoemaker approached a large crater (again, named Shoemaker). The caption gives the spatial resolution achieved for each image.



Here is a NEAR-Shoemaker descent image taken when the spacecraft was just 250 m (820 ft) from the surface. Rocks less than a meter in dimension are scattered on the surface, as had been expected.



Astonishingly, the gamma ray spectrometer on the spacecraft was not damaged. When its solar panels were arrayed so as to pick up sunlight and hence provide power, the spectrometer was turned on and data were collected from two depths. Here is one of the plots:

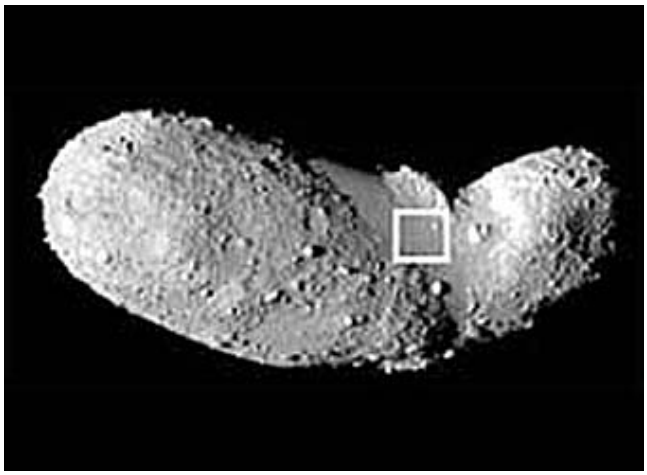


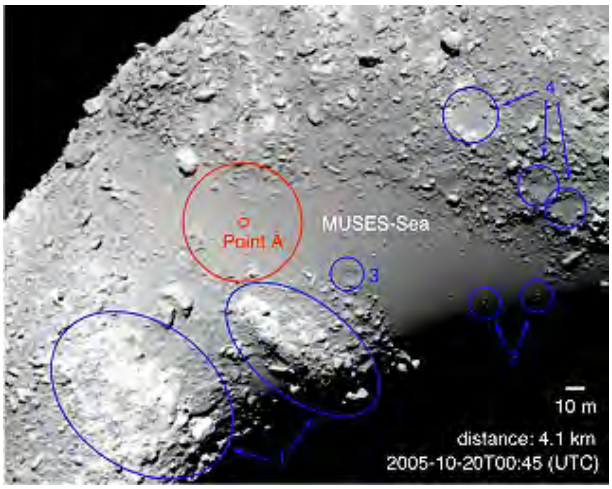
This triumph, with its serendipitous landing results, augers well for future missions to asteroid surfaces; sampling these (either by instrumental analysis or by sampling collecting and return to Earth) will give scientists their first solid data on the composition of these asteroids which have the most primitive material in the condensed planetary rock system. This should thus confirm whether any of the meteorites fallen on Earth are truly samples of the planetesimals which, for Earth and the inner planets, were the primary constituents that eventually melted.

The first try at landing on an asteroid, collecting samples, and then returning those to Earth is being carried out by the Japanese space program. The small (535 x 294 x 209 meters) Apollo Belt asteroid Itokawa was reached by the Hayakawa spacecraft and then placed in near orbit around this irregular object - the asteroid Itokawa.

19-73: What have you seen (as images) before in Section 19 that remind you of these asteroids? ANSWER

The initial attempt at landing failed but success occurred from a second attempt in November 2005. The spacecraft may have collected samples (although probably not more than small amounts of dust and then headed enroute back to Earth, with arrival in 2010. Here is the landing site:





Analysis of instrumental data on Hayabusa and other sources have now led scientists to conclude that Itokawa is an assemblage of rock particles ranging from dust to boulders, held together by weak gravity and electrostatic attraction.

The return trip of Hayabawa to Earth was fraught with difficulties, which were largely overcome by the daring and skills of the Japanese technicians. On June 12, 2010 the spacecraft reached Earth, entered its atmosphere, and released a small capsule which parachuted into the desert of central Australia and was quickly found and recovered.

NASAs Dawn mission to asteroids was launched in September of 2007. The probe will reach 4 Vesta in 2011 and 1 Ceres in 2015. An overview of the mission is given at this Wikipedia website. JPL is directing the mission; check its Dawn homepage at JPL index.

The redirected long-term space program announced by the Obama administration includes a possible manned mission to land on an asteroid. This would allow sampling similar to the Moon visits during the Apollo program. It is very important to fix the composition of a typical asteroid, as this would in principle determine the nature of the starting materials believed to dominate the rocky planets and probably the cores of the Giant planets.

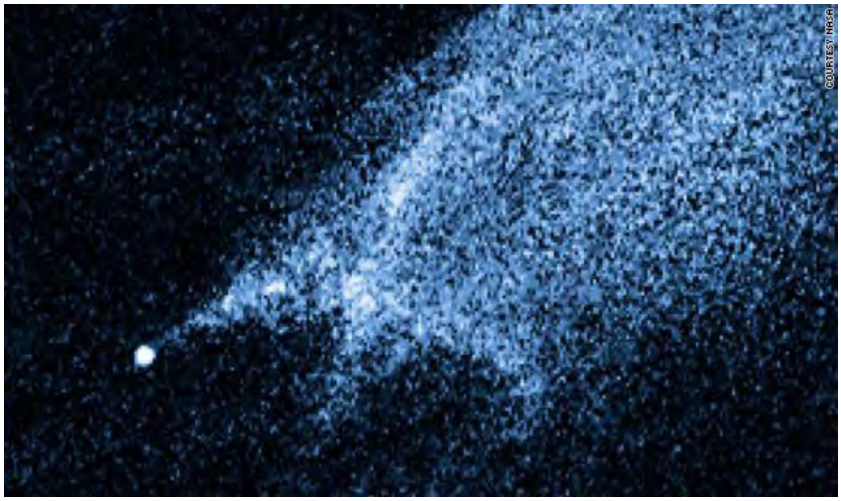
Planetologists generally agree that they already know some basic facts about asteroid composition. Most meteorites are considered to be parts of asteroids or pieces broken off of asteroids. But until and unless the Hiyakawa mission is successful, no meteoritic material tied to a specific individual asteroid has been acquired - until 2008. That year a tiny asteroidal fragment (about 3 meters), named TC3, was discovered hurtling towards Earth. It was predicted to hit the atmosphere in northern Africa in September. This indeed happened, with the asteroidal body exploding at a height of about 32 km (20 miles). The event was observed by nomads in the Namib desert of Sudan. Scientists journeyed into the area below the fireball and found many pieces of the meteoritic material surviving the explosion. You can read details of this event at the Astronomy Now website.

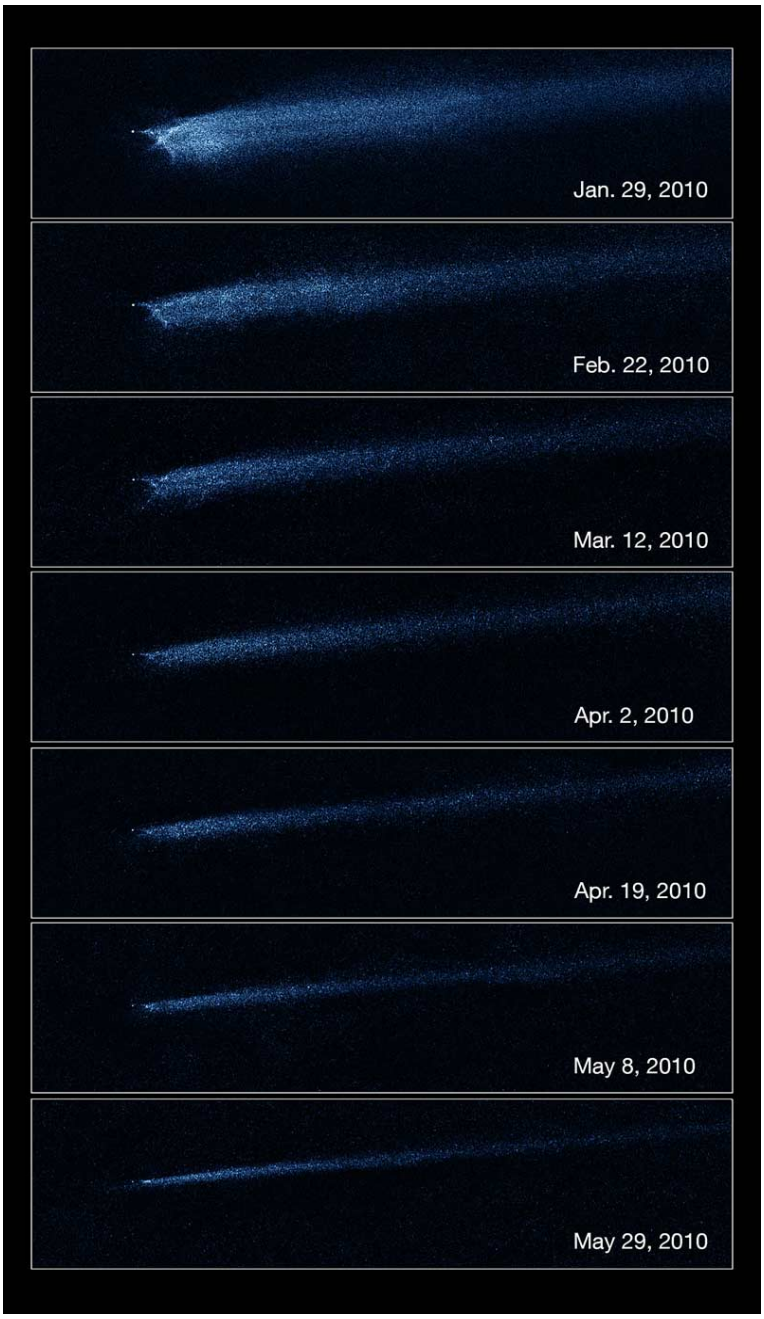
Many more asteroidal bodies populate the Kuiper Belt. There are now several large objects that orbit the Sun from within or near that Belt. Their status as large asteroids or mini-planets

remains unsettled. Here they are as depicted in an artist's conception:

The object named 2003 EL61 is peculiar. Its shape is ellipsoidal - almost like a football with rounded tips. It is the fastest rotating object in the Belt, tumbling in a full rotation every 3.9 hours.

With many thousands of asteroids in the Main Belt, it is not surprising to learn that occasionally two such bodies will collide and the smaller one may then disintegrate. This is the best explanation for the observation by the Hubble Space Telescope in late January of 2010 of what appears below as debris in the Belt. This rubble trails out from the moving asteroid much like the tail of a comet. Because the debris stream diminishes so rapidly, the collision is interpreted as having been very recent:





The question of the origin of asteroids is still debated. All agree that they represent primitive materials and are at least as old as the planets. A few planetologists still argue that some, perhaps all, of those in the Main Belt represent a disrupted planet. But most subscribe to the concept that these are really planetesimals (accretionary bodies built up of fragments of gravity-attracted small solids that condensed and organized during the first stages of solar history) that never succeeded in building up by accretion into (a) planet(s). The reason for this failure to reach planet-size is ascribed to the perturbing influence of gravity from the Giant Jupiter. Most of these bodies today have undergone repeated collisions that knock off chunks from the target body, some of which escape to become new asteroids but much (most?) re-assemble into the collided remnant or into a neighboring asteroid to form a new shape.

Some asteroids are solid in the conventional sense but may now consist of joined fragments (note shape of Toutatis, which appears to have three connected pieces); these have greater densities but may contain internal voids. Others, perhaps the majority, being lower in density, are thought to be comprised of smaller fragments. These indeed may even be a composite of sand, gravel, and small boulder sized material (probably the residue of the carbonaceous silicates that were the first solar dust aggregates). They are held together by gravity, electrostatic attraction, and possibly a form of ice. When collisions on them occur, pieces may be knocked off but these tend to rejoin the parent in a matter of days. Over billions of years many collsions have occurred, reordering the asteroids into new assemblages. Most asteroidal surfaces, whether solid or held together in a manner similar to "sand castles" made at a beach, are covered with loose debris or "regolith".

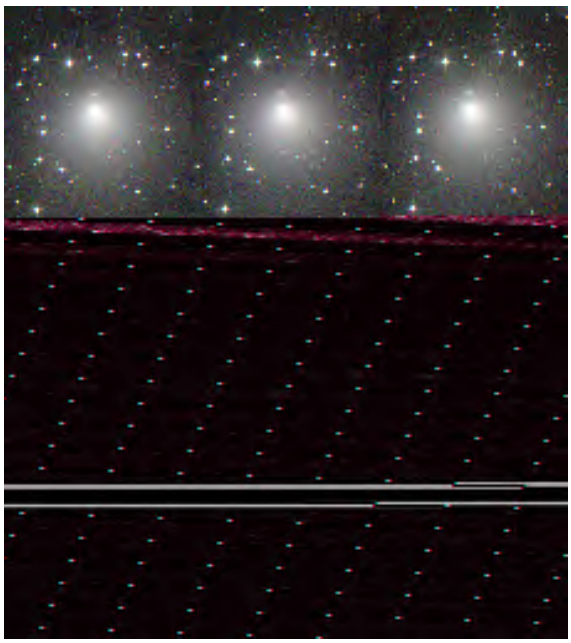
In the treatment of meteorites (page 19-2) it was pointed out that asteroids ranged from primitive carbonaceous chondrites in composition to those that seem to have a distinct set of concentric shells, with iron cores. The stony meteorites display varying degrees of thermal and shock metamorphism, with the irons being evidence that the asteroidal body may have grown large enough to melt and form an iron core before later disruption. Sources of heat energy are needed to produce these differences. Some of that energy was provided by the same radioisotopes of uranium, thorium, and potassium as currently continue to heat the Earth. But in the early Solar System there were also isotopes with short half-lifes that may have produced even more thermal energy. Two known to have been effective are Al^{26} and Fe^{60}, both now naturally extinct. Collisions between asteroidal bodies also contributed to the heat content, causing at least part of the target body to overheat locally. Another source of heat could have been from the Sun itself, both as thermal radiation and solar wind.

Asteroids, and fragments therefrom, occasionally hit the Earth, as we showed on page 19-2 dealing with meteorites and page 18-1 dealing with impact craters. So do comets strike our planet: a possible example (although it could have been a stony meteorite) was the 1908 Tunguska event in Siberia, where trees were knocked down (in a radial pattern pointing to a center of blast) over many hundreds of square kilometers by an explosion just above the surface. Great amounts of dust were thrown into the atmosphere, producing abnormal red sunsets worldwide for the next several years.

Comets are among the most spectacular of the heavenly bodies with their long, icy tail, receiving mystical significance from early observers, until later observers determined their true nature. Interest has always been high regarding comets, balls of ice and rock, because of their spectacular appearances (glowing "stars" with tails); they are well represented in mythologies and astrologers' portents.

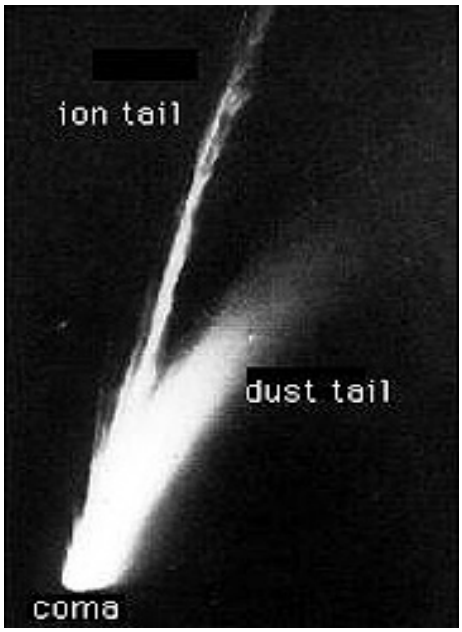
19-74: Name the comets that YOU have actually seen (through binoculars or telescope, or even naked eye) in your lifetime. ANSWER

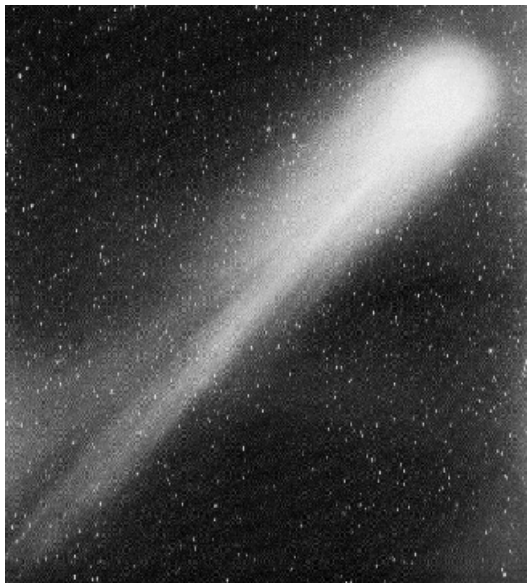
Scientists now know that comets are mainly ice balls of varying sizes (up to 10-30 km [6-19 miles]), mixed with rock debris to some extent. They travel in eccentric orbits (e.g., Kuiper Belt) within the Solar System, repeating appearances or simply flying through once, if not captured gravitationally. As seen when still far from Earth, the central coma appears as a glowing ball from 10,000 to 100,000 km (6,214 - 62,137 miles) in apparent diameter, around a much smaller solid nucleus. A good example is the comet Encke:



As a comet passes the Sun, the solar wind and other factors cause it to ablate, expanding the coma and creating a stream of particles that trail off as a long tail (up to 100 million km [62,137,000 miles] long) of dust and plasma. The tail points away from the Sun along a radial line. Most comets actually develop two tails, one consisting of debris reflecting sunglight and the

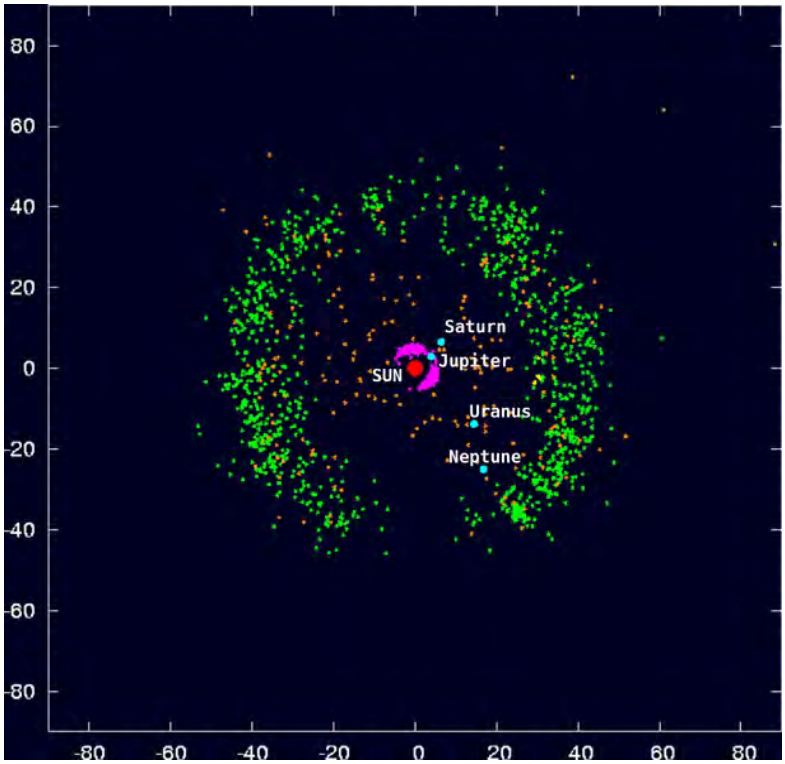
other composed of ions excited by UV irradiation. Look first at the unnamed comet and then at the tail from famed comet Halley.

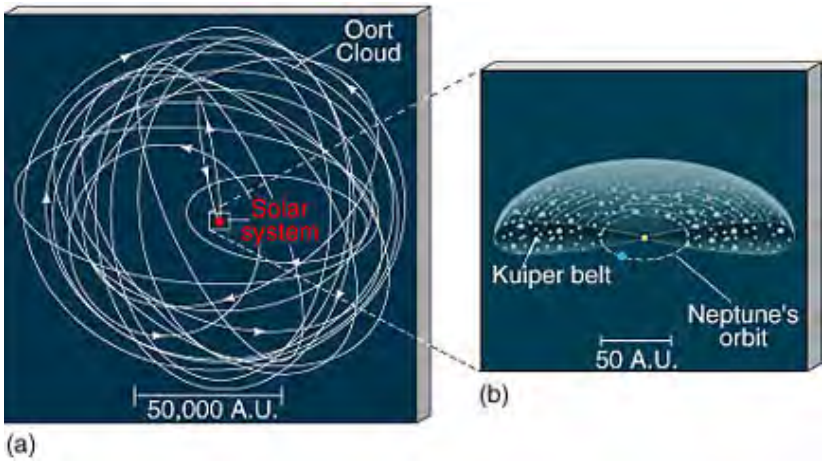




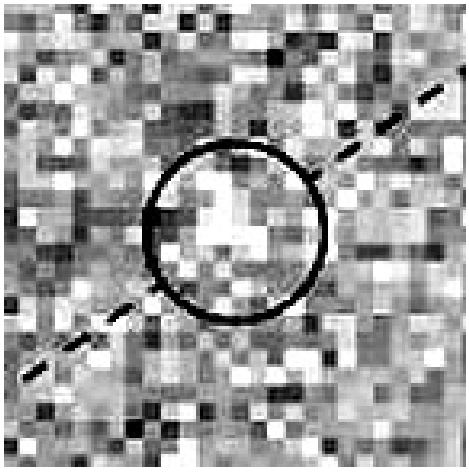
Scientists think that the nucleus (descriptively referred to as a "dirty snowball") consists of very primitive materials, organized during the Solar System's development. Spectroscopic studies indicate the presence of molecular compounds of carbon, nitrogen, and hydrogen, including CN, NH, NH_2, and HCN, which break into ions carried into the tail. A telescope view of Comet Kudo-Fujikawa using a sensor that detects hydrogen ions produced this image, in which the H ions are colored reddish:

Astronomers know the orbits of some comets well enough (through observations) to predict when they will return. Most comets come from either the Kuiper Belt (beyond Pluto) or the vast Oort "Clouds" that extend to the outer reaches of the Solar System. Some may be intergalatic but that has yet to be demonstrated. As seen in this diagram, the Kuiper Belt is a toroidal configuration that begins beyond the orbit of Neptune. Short-Period (recur within the inner Solar System in less than 200 years) comets come from this Belt. The Oort Cloud, extending between 10000 and 50000 A.U., is the source of Long-Period (> 200 years) comets and is distributed in a spherical zone around the Solar System.



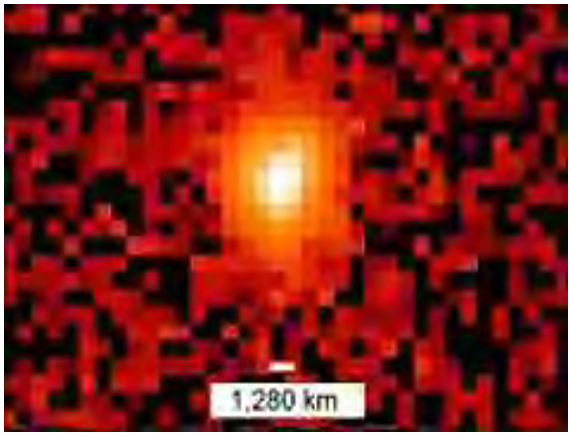


The ***Kuiper Belt*** contains Objects (KBO) that are primarily comets but seem to also include some asteroidal type bodies. The Belt was predicted to exist by Gerald Kuiper in the 1950s but the first direct imaging was not achieved until the 1990s. The example shown here indicates the difficulty in spotting such small objects far out in the Solar System; once located, the circled object moves relative to the background, allowing its orbital migration to be calculated. (The blocky squares are the individual pixels in the detector plate.)



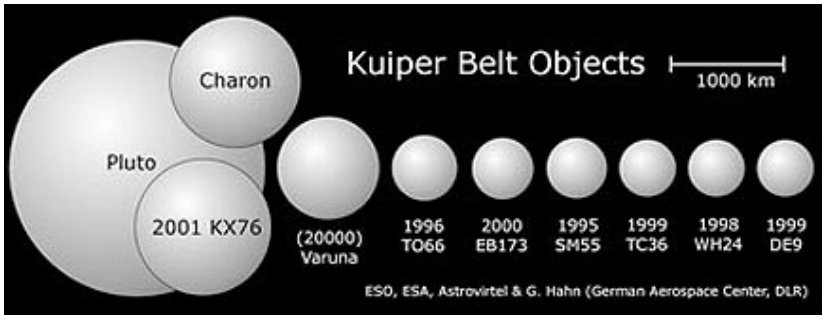
The KBOs are clustered in a disk-shaped region near the ecliptic (plane in which most of the planets orbit the Sun), lie at distances from 30 to 100 A.U., and may contain up to (an estimated) 100,000 objects larger than 20 km in diameter. Centaur-class KBOs are relatively close to Neptune's orbit and include the icy body named Chiron (discovered in 1977; at first thought to be an asteroid but now proved to be a comet with a short tail); of the estimated 400 Centaur objects greater than 100 km in diameter, at least 9 are notably larger.

One of the largest known KBO asteroids is 2002 LM60, discovered first by a ground telescope and then confirmed and studied by the Hubble Space Telescope:



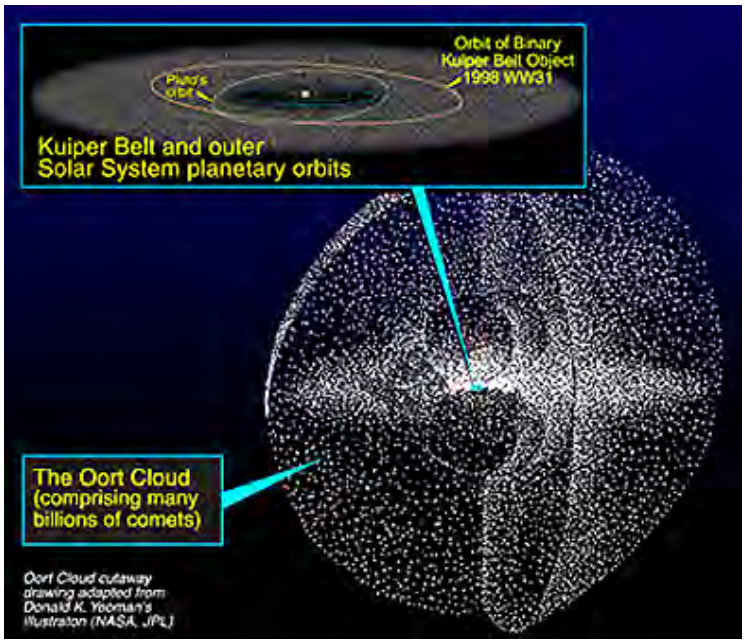
The discoverers have named this asteroid Quaoar (pronounced Kwa o whar), an American Indian name associated with the tribe that once occupied the Los Angeles Basin. The asteroid is

spherical (indicating that it was once molten) and is close to 1250 km (780 miles in diameter). Quaoar is 6.5 billion kilometers from the Sun, around which it moves in a circular orbit. It lies 1.5 billion kilometers (about one billion miles) past Pluto. Another recently discovered asteroid is 2001 KX76, named Ixion, whose size ~1200 km (745 miles). Some other large objects in this belt (using Pluto and its moon Charon for comparison) are depicted in this schematic:

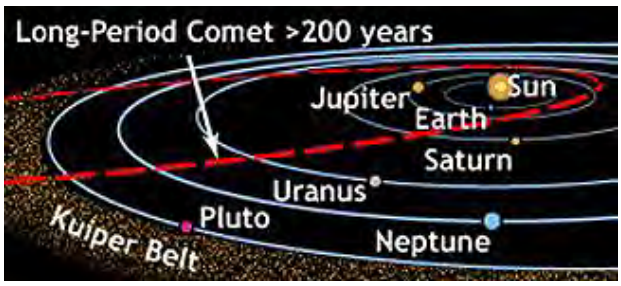


Results of observations reported in 2003 by Rodney Gomez of Brazil support the presence of two distinct Kuiper belts. One lies largely within the solar ecliptic; its asteroids are grayish. A second belt has trajectories up to 40° off the ecliptic; its asteroids have a reddish tint.

The ***Oort Cloud*** (OC) is a widespread swarm of perhaps billions of comets that orbit farther from the Sun than the KBOs. These orbits are not confined close to the ecliptic but can follow paths that occur anywhere in the "sphere" of the Solar System with the Sun at its center. As yet, no OC object *as it resides now within the Oort Cloud* has been imaged by either ground-based telescopes or the HST owing to their small sizes and great distances from Earth. However, their existence is based on firm reasoning that predicts them to be a significant part of the bodies that formed from the nebula that organized into the Sun, planets, and other objects. The artist's sketch below shows roughly the relative sizes of solar region volume containing the Oort Cloud and the Kuiper Belt:

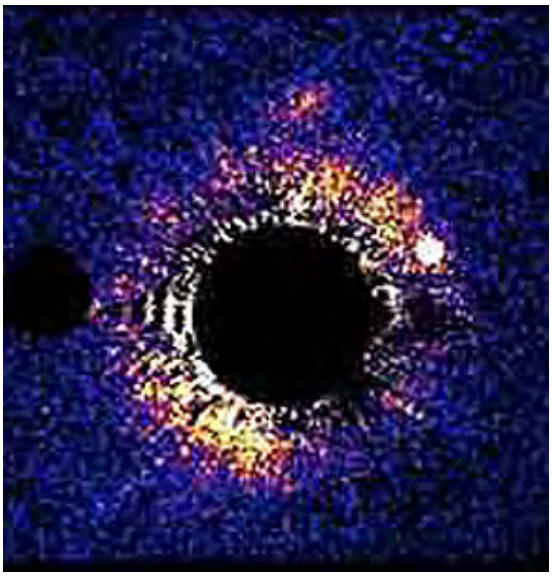


Individual Oort Cloud comets have been imaged because they have been perturbed from their distance locations and brought into the Solar System, as described by this diagram:



One of the most recent Long-Period comets to be discovered (in 1996) is Comet Hyakutake:

Kuiper Belt and Oort Cloud assemblages have been observed around other stars. These are usually referred to as Debris Disks (see page 20-11). Here is one example seen by the Hubble Telescope. A mask around the central star allows the reflecting debris, probably developed from or modified by collisions, to stand out.



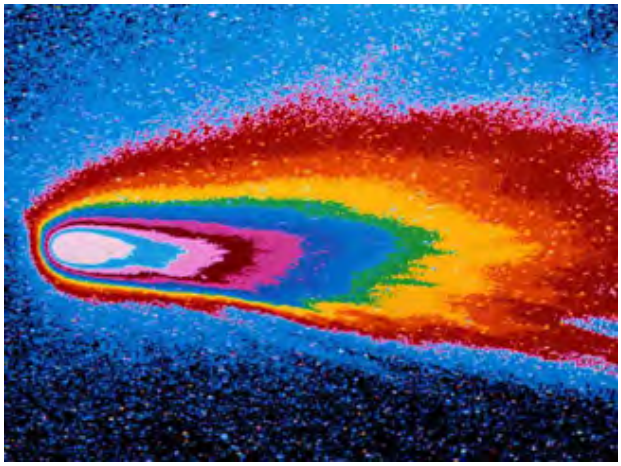
As of 2005, 898 comets have been catalogued; most now have their orbits reasonably well calculated. Of these 184 have definitely been established as *periodic* - namely, those comets known to repeat their passage through the inner Solar System at intervals of years to centuries as they orbit the Sun; these are the aforementioned Short-Period (KBO) and Long-Period (OC) types). New comets and others of the 878 may also be periodic but more observations over time are needed to confirm this. The number of comets - most still undetected - within the Solar System may be in the billions; it is not yet known whether comets also reside in interstellar space.

Most comets positioned well away from the Sun are without pronounced tails, and are best found by looking for notable displacements of small bright objects (early stage comas) relative to fixed star backgrounds in film records taken days apart. Such motions delineate the advance of comets, as well as reflecting asteroids, at high speeds through solar space.

The best known of all comets is Halley's Comet, observed and recorded in ancient times. Its periodicity, first predicted by Edmund Halley to verify Newton's Laws of Motion, causes it to reappear about every 76 (range 75 to 79) years. It is thus within the Short-period class. Having passed in 1909, as shown in this wide angle view that displays its magnificent tail, it reappeared in 1986, and many predicted it would provide a great celestial display (which generally fell short of expectations).

Debate raged in advance about sending one or more space probes to examine it close-up, since the next opportunity would not be until 2061. Although NASA decided against this adventure, Japan, the former Soviet Union, and the European Space Agency sent probes to gather data. In particular, the Italian government designed and launched a spacecraft named Giotto, which came within 540 km (336 mi) of the nucleus on March 13, 1986. Here is a close approach view:

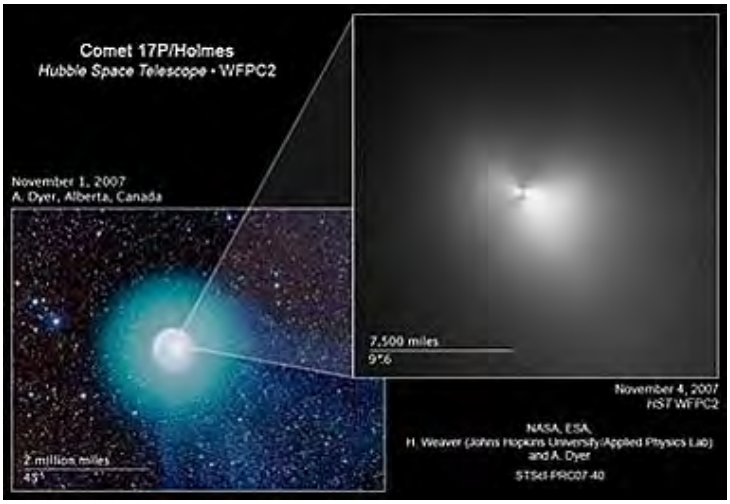
Giotto found that Halley's nucleus, measured at 16 km x 8 km (10 x 5 miles), is very dark, lumpy and of low density (0.1-0.2 g/cc). This suggests that it was then very porous, with most of the ice having ablated or evaporated away, leaving carbon-rich dust as a residue. Although not as bright as anticipated and a disappointment to ground viewers, the comet, as seen through telescopes, provided exceptional displays. We can emphasize reflectivity variations in its coma and tail, representing particle density differences, by displaying them in false color:



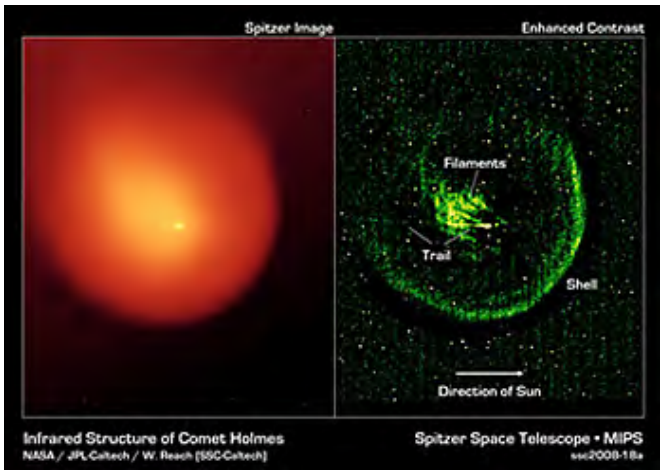
A recent comet sensation is Comet Hale-Boggs, discovered on July 23, 1995, that passed Earth as close as 85 million miles on April 2, 1997. Hailed by many as the Comet of the Century because of its size (four times larger than Halley's Comet) and brightness, it was visible in northern and southern hemispheres during much of the first half of 1997. Here is a typical view, taken on March 11, 1997, by Jerry Platt, one of many amateur astronomers who tracked this

spectacular celestial visitor.

Another comet, Holmes, first noticed in 1997, began to flare up in November of 2007. Here is a ground telescope image of this comet; the inset shows the nucleus at the time of maximum flare up:

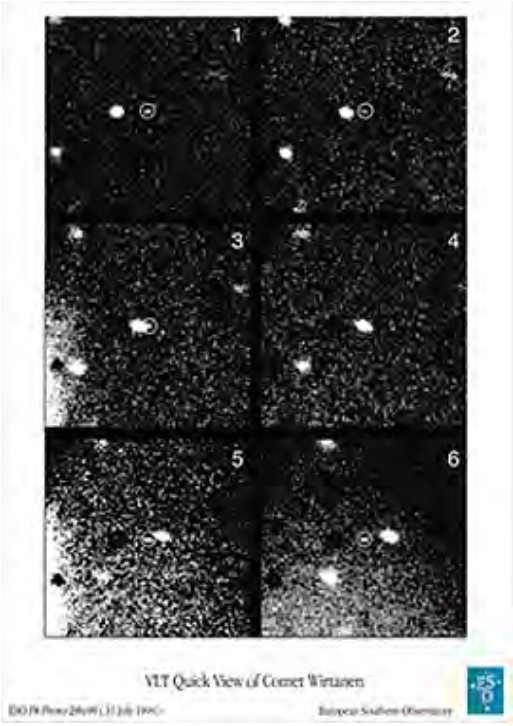


The Spitzer Space Telescope produced this pair of images; the left image shows the huge coma that appears to be made up of dust (imaged at the 22 micrometer wavelength so that this is a thermal view) that "exploded" outward during the flare up; the right image shows details at a different wavelength:



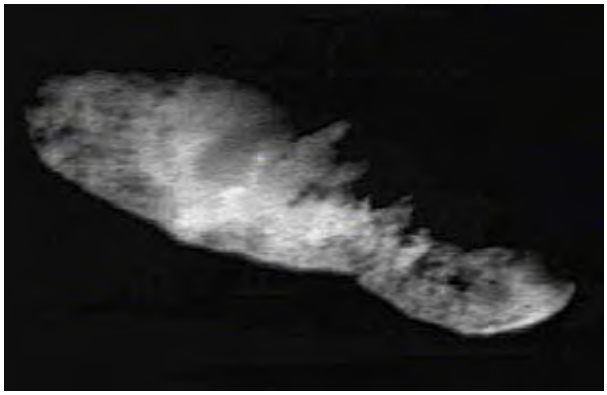
Here is a Hubble Space Telescope image of comet Holmes, again without a tail:

Most comets that are far from the Sun don't have pronounced tails, so to find them we look for notable displacements of small bright objects (early stage comas) relative to fixed star backgrounds in film records, taken days apart. Such motions delineate the advance of comets, as well as reflecting asteroids, at high speeds through the solar system. The 6 panels in the next illustration show the progressive movement over hours of the comet Wirtanen; when imaged it was about 605 million km - or more than 4 A.U's - from Earth. The comet, a tiny dot is circled in each panel. Follow its displacement from right to left (in panel 4 it is directly in front of the reference background star).

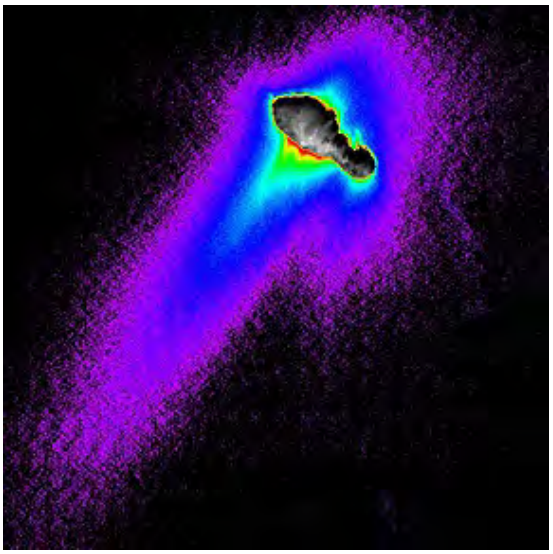


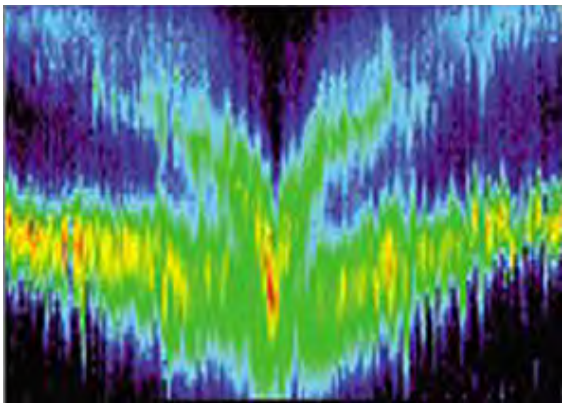
In 1999, NASA sent a probe called Deep Space 1 to test new observation techniques and if possible to gather new information about comets it would approach. It has lasted beyond its

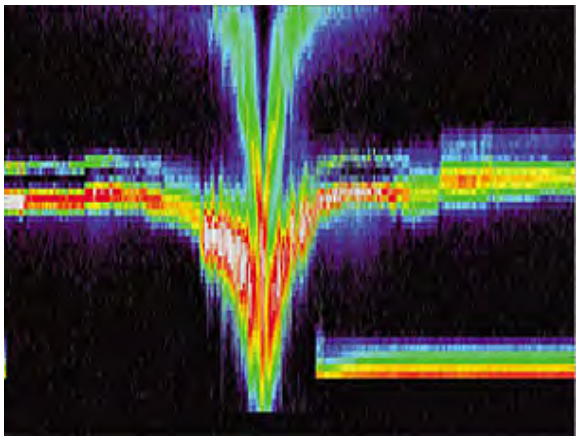
planned lifetime. One of its most spectacular achievements was on Sept. 22, 2001 when it passed within 2200 km (1400 miles) of the Comet Borrelly to image its nucleus. Concern was high that the probe, suffering now from some malfunctions, would be damaged by cometary debris. However, it succeeded grandly in getting excellent images of the nucleus. Below is the last image received before planned shutdown in which the elongated nucleus, an 8 km (5 mile) long object imaged at 45 m resolution, displays ridgelike irregularities, faults and other geologically describable features; in some ways, it resembles an asteroid but its dark surface is composed of ice and dust (note the jets of material streaming off):



Sensors on Deep Space 1 produced these images, with information described in their captions.

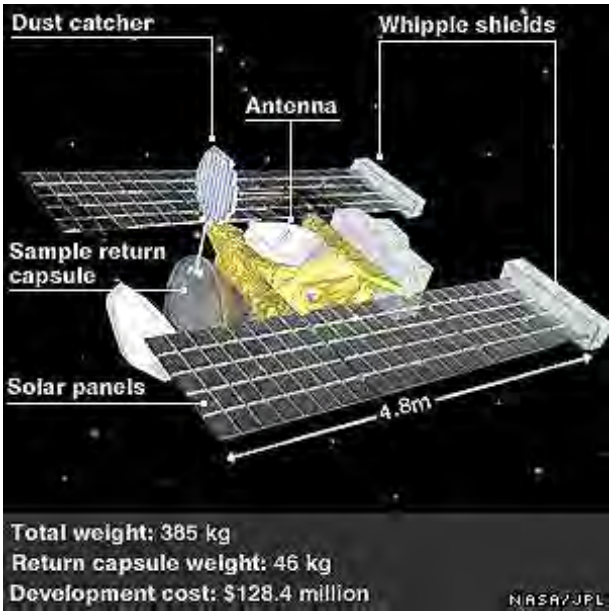




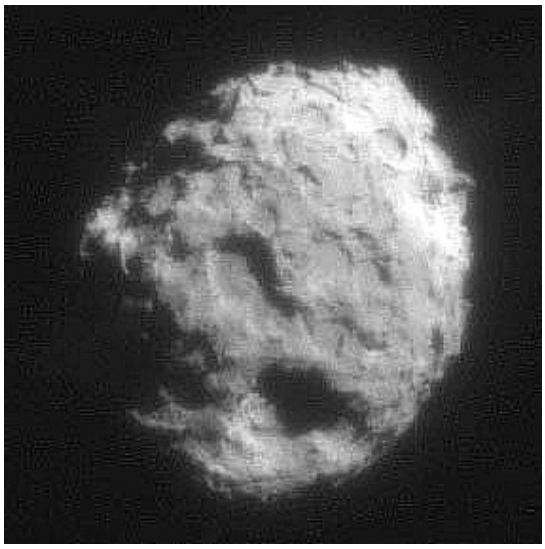


A relatively pristine comet was discovered in January 1988 by the Swiss astronomer Paul Wild (its German pronunciation is "Vilt"). Comet Wild 2 was once beyond Jupiter but a close encounter with Jupiter has carried it into an orbit beyond Mars such that it passes Earth about every 6 years. Here is its appearance in December 1990:

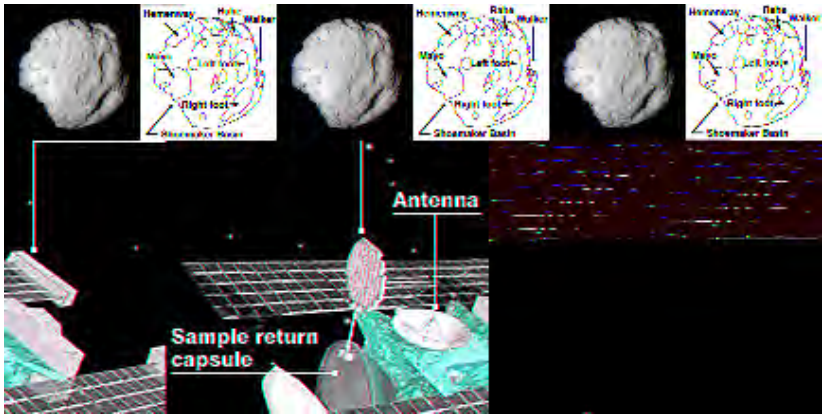
In 2001, NASA launched a probe (the **Stardust** mission) to image Comet Wild 2 and to collect samples of dust and gas as it passed through the comet's tail. The experiment was successful. This drawing shows how the particles are collected:



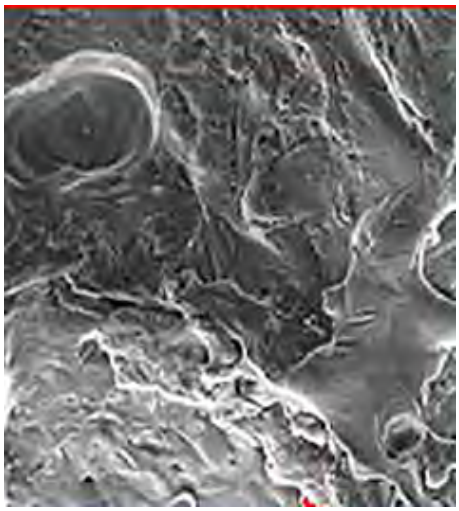
Here is a view of Wild 2's nucleus (~5.4 km [3.3 miles] in maximum dimension) from a distance of 500 km (341 miles):

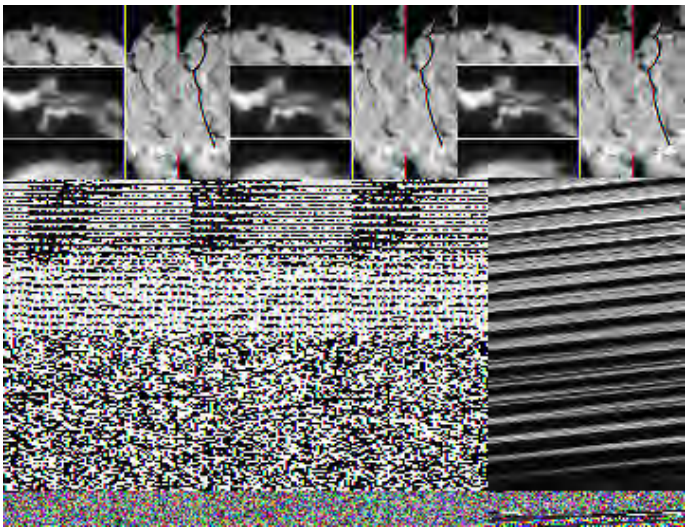


Features on Wild 2 have been named (with descriptive humor), as shown here, in which the spacecraft was somewhat closer.

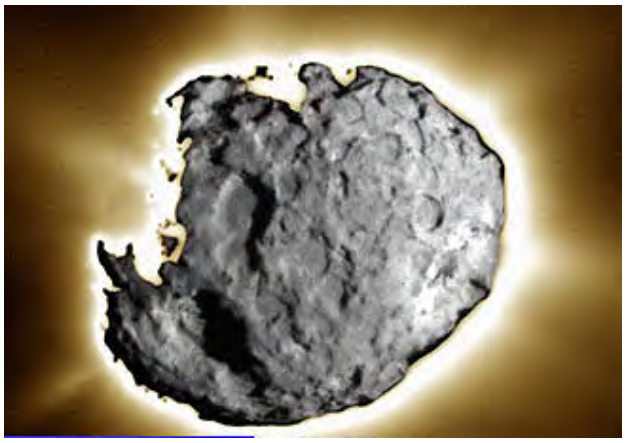


Close-up images show some distinctive features, including depressions, fractures, and prominences

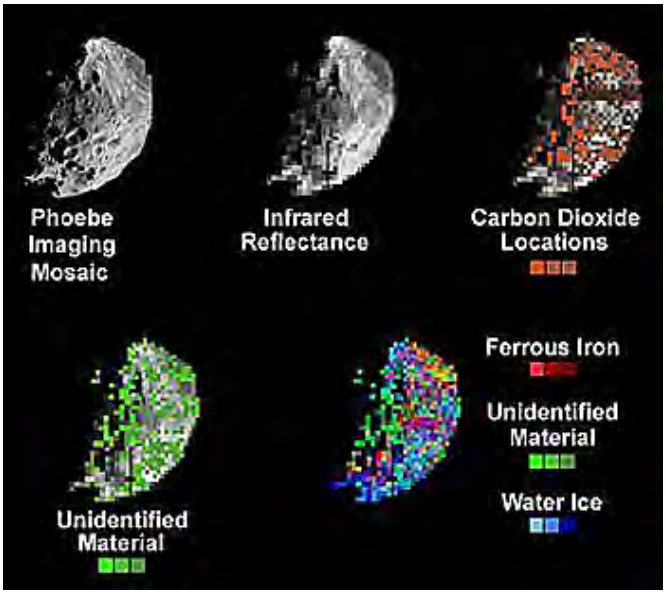




Still another image subjected to different processing showed glowing material coming off this nucleus. The cometary body apppears to consist of a mix of ice and rock together hard enought to serve as a strong solid that is not disrupted by repeated cratering action (some investigators think the pits are ablation - rather than impact - produced).



Using photometric filters, Stardust has obtained data in the visible and infrared allowing good estimates of the composition of materials at Wild 2's surface, as follows:

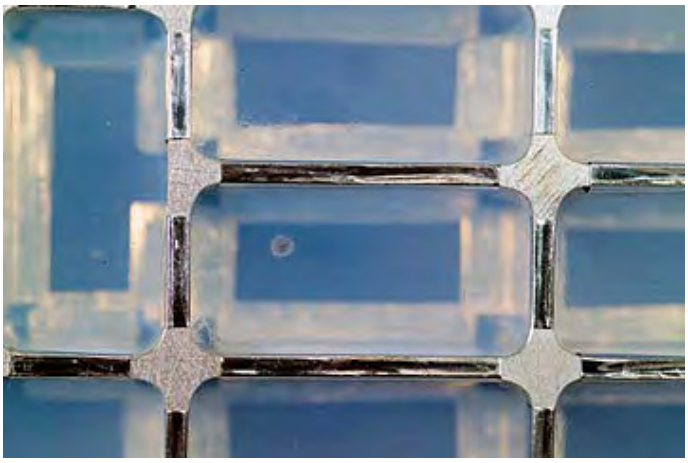


Based on the observations just described, the interpretation of Wild 2's makeup indicates that it has some similiarities to the Kuiper belt asteroids, suggesting a kinship.

Stardust successfully flew through the comet's tail and collected particles. It then fired rockets to begin the long journey home. In the middle of the night on January 15, it hit the atmosphere at about 40000 km/hr, slowed, deployed a parachute and landed saftely on the desert floor of the

Dugway Proving Grounds in Utah. Here is a photo showing it at this resting place.

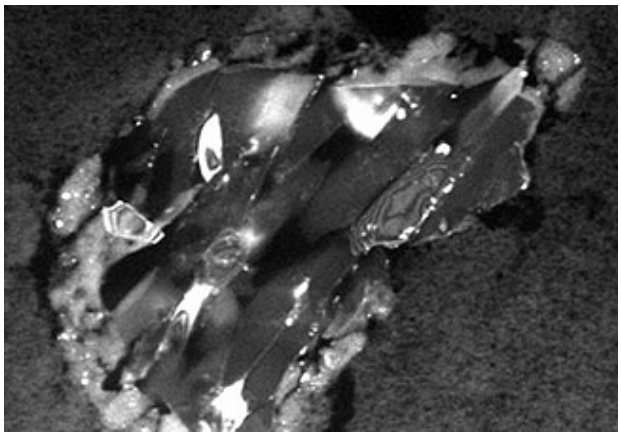
After removing the sample payload package, it was opened at JPL in Houston, with great hopes. The sample collection material is a network of rectangular "cells" placed within an aluminum framework. These cells are made up of a material called "Aerogel" which consists of a silica compound that is solid and rigid but supports a structure that gives the gel about a 95% porosity (open space) so that cometary particles can move through until finally becoming trapped in the solid component.



Note the black spot in one of the cells above. This is a cometary particle. Preliminary estimates suggest perhaps up to a million very tiny (millimeter to submillimeter) particles were collected. One noteworthy example, with associated tracks, is shown here.

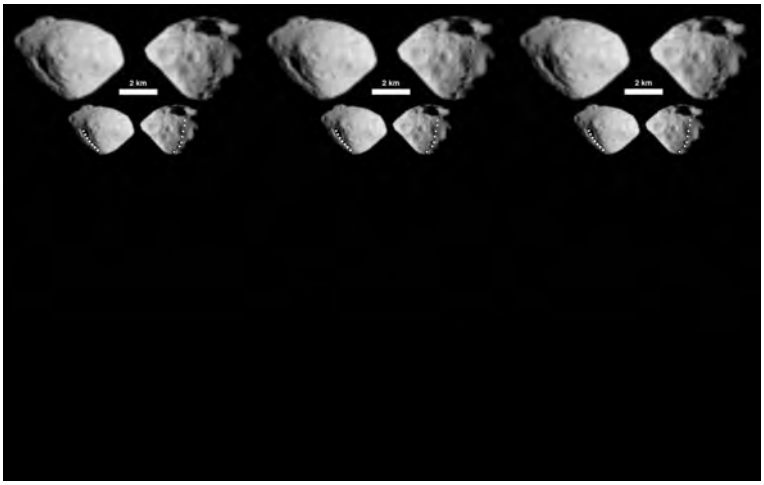
Another eye-catching dust particle impact produced a pattern in the aerogel that is remarkably similar to pictures of the ejecta tossed out from much larger impacts during the growth phase of the forming crater; this Wild 2 example may have been caused by some different process:

The first mineral identified in the captured comet debris is the olivine end member Forsterite (Mg_2SiO_4) which is on Earth a gem variety known as Peridot. Here it is:



This mission is judged a huge success. Everything worked. Several hundred chosen investigators will subject recovered grains to a wide range of tests. As information is acquired, results will appear on this page.

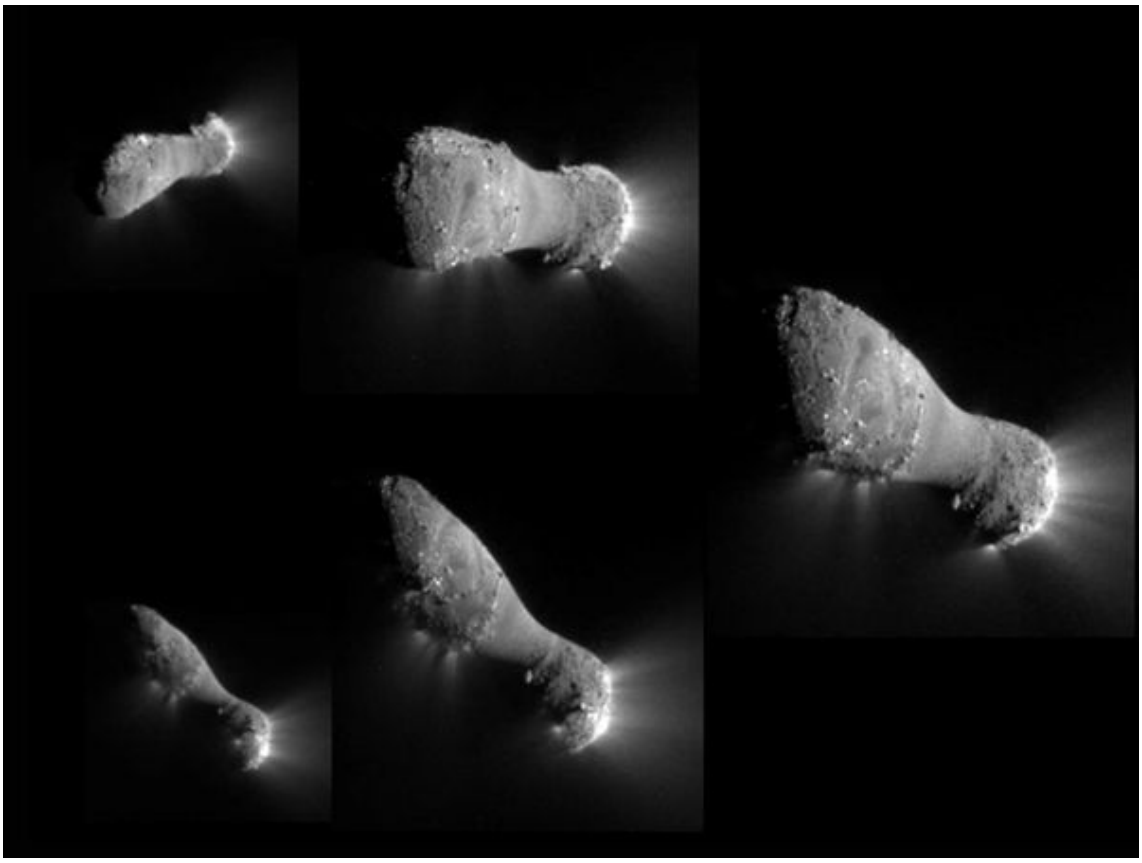
The European Space Agency (ESA) is conducting a follow-up to Giotto with a mission (Rosetta) to the Comet Churyumov-Gerusmanko (6 km diameter) after a successful launch on March 2, 2004 and planned arrival in 2014. Heavily instrumented, Rosetta includes a lander on the comet's nucleus. As it travels through the asteroid belt, it has been observing several of the asteroids. Here is its view of the small asteroid Stein:



NASA's Deep Impact Mission was launched in January of 2005 to comet Tempel 1. In July of 2005 as it near Tempel'surface, it launched a probe designed to make a large impact of the surface. This took place with results summarized in this image:

Missions to comets and asteroids remain dynamic programs because they are oriented towards learning what these bodies can tell about the early Solar System. We will mention two now underway by citing their Web sites: 1) the EPOXI mission which is the continuing flight of the Deep Impact satellite that passed by comet Hartley-2; and 2) the Stardust-NExT mission, launched in 2007 as a follow up to Stardust, to examine the modification made by impact of the spacecraft on the surface of comet Tempel-1

Hartley-2 was imaged successfully on November 4, 2010. This is one of the best images obtained during the flyby of this 2 km long comet nucleus.:

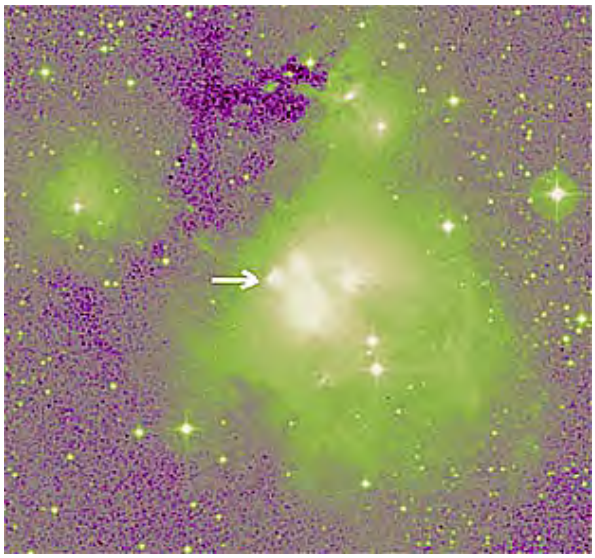


The bright streaks or jets coming from Hartley-2 are dust and gas (almost entirely CO_2). The comet is now close enough to the Sun for solar radiation to be effective in causing ablation.

So, as of the close of 2010, five comets have been visited and imaged in detail. A review of these encounters is given by this montage:

19-75: How do you think comets have influenced the Earth in history? ANSWER

What is the ultimate fate of most comets? Two possibilties are most likely. First, they may just gradually "wear away" as their icy comas are ablated or otherwise destroyed as they repeatedly pass their parent star. Second, they may be pulled into that star or one of its planets (see next page) by gravitational attraction, thus becoming totally annihilated. Evidence for this latter fate has been indirectly retrieved from a ground telescope study (McDonald Observatory in Texas) of a Herbig Be type star, LKHa234 in nebular cloud NGC 7129, 3200 l.y. away from Earth in the Milky Way (see page 20-5 for a discussion of this star type). This 100000 year-old star (about 5 solar masses) was observed over 5 days. Spectra taken of it show dramatic changes, e.g., in sodium light, during that period. One reasonable interpretation is that a comet with sodium content crashed into the star, thus dispersing its source so that sodium spectral lines disappeared. Here is an image of LKHa234 and its surrounding nebula taken during this period.



In the last twenty years or so, concern has heightened that asteroids or comets might strike the Earth, as was proved in Section 18. Those whose orbits could cross Earth are collectively known as Near-Earth Objects (or NEOs). This possibility was heightened by the comet that struck Jupiter, as discussed on the next page. Some federal and private funding has been provided to conduct a systematic search for NEOs. To date hundreds of new asteroids and a few comets

have been detected and their orbits calculated. If one is shown to have a good chance to hit the Earth, several ways to stop this have been considered. Sending a space probe with a nuclear device onto the NEO might break it up, but the pieces could continue (like buckshot) on a collision course. The best strategy would be find a way to deflect the orbital path so that the asteroid or comet would have a much reduced chance of meeting the Earth in its orbit. A nuclear explosion away from the NEO could exert enough pressure to modify its orbit. Since calculations have now indicated that an Earth-striking NEO's impact strong enough to cause serious damage may be as frequent as once every thousand years, the NEO detection program has taken on a new urgency.

All this new knowledge about asteroids and comets roaming about out there has raised troubling concerns within the Planetology community about the possibilities of impacts from these projectiles onto the Earth. This has been mention before in Section 18. Proposals to mount a systematic inventory of "heavenly bodies" and their trajectories are now given serious evaluations. The U.S. Congress has called for an effort that would give the world enough warning for our technology to devise schemes to intercept approaching asteroids and comets, steering them to a non-collision course. One group - Pan-STARRS (Panoramic Survey Telescope and Rapid Response Systems) - will utilize four telescopes (3 meter mirrors) at different localities to constantly search for these potentially dangerous projectiles.

A helpful overview of spacecraft visits to asteroids and comets is found during a JPL lecture series webcast. Access this at von Karman lectures, choosing the topic "Comets and Asteroids", June 20, 2002.

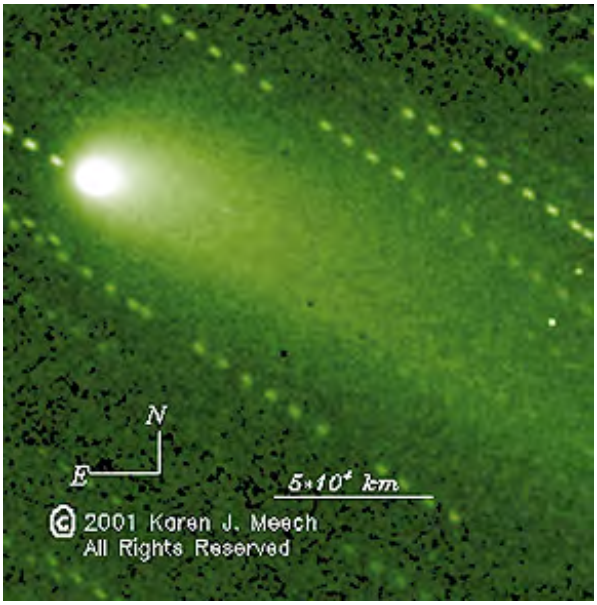
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

A special mission, called Deep Impact, was designed to have a small probe, with camera, strike comet Tempel-1, at high velocity. The amount of material ejected, and the size of the crater, will give good clues as to the nature of the comet's material, and possibly internal structure. This impact occurred on the Fourth of July 2005 and was observed by cameras in the mothership and by many observatories on Earth.

The Deep Impact Mission

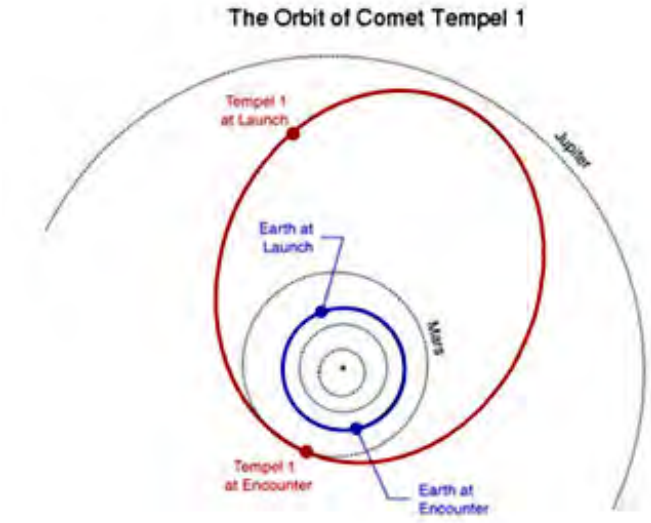
After the great excitement and beneficial knowledge from the Shoemaker-Levy event, NASA decided to stage an impact of its own, this time on a comet rather than a gasball planet. The comet chosen is Tempel-1, discovered in 1867 by George Tempel. Here is a telescope view of Tempel-1 when it was close enough to the Sun in August of 2001 to develop a tail; the bright circular area is the comet's coma.



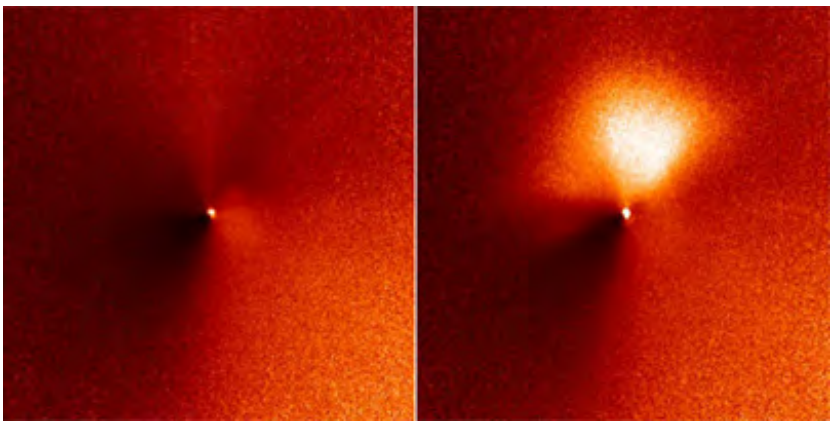
The rationale behind the mission, named Deep Impact, is simple: A probe sent from the mother ship onto the comet's surface should produce an impact crater, whose size, for the velocity and mass parameters chosen, can tell much about the nature of the material making up the surface and into the interior. A nose-mounted camera in the impactor will show progressive closeups of the target area, much like the Rangers did for the early days of lunar exploration by crashing onto the Moon. Instruments on the surviving mother ship can observe the impact and make measurements of the debris thrown off the comet for days thereafter. This mission, the 12th in the Discovery series, has a price tag of $333 million dollars. You can read more about the mission on this NASA website.

Here is the Deep Impact spacecraft as it rests completed at its fabrication site:

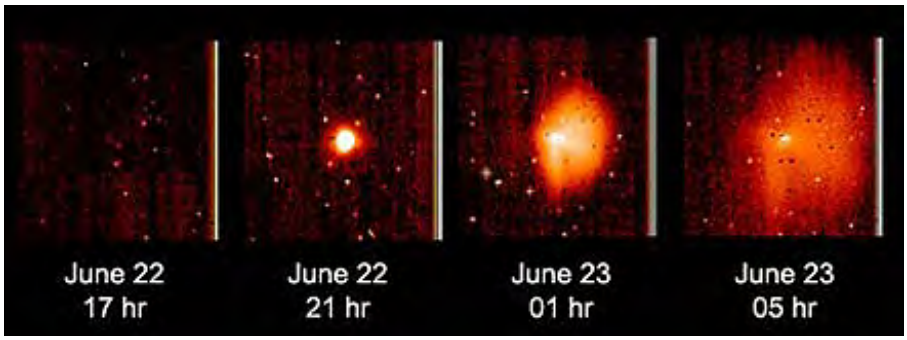
Launch date was January 12, 2005 from Cape Canaveral. The transit time was deliberately set for arrival, impactor separation, and impact itself on July 4, 2005 as part of the U.S.'s "birthday" celebration. Here is the orbital history.



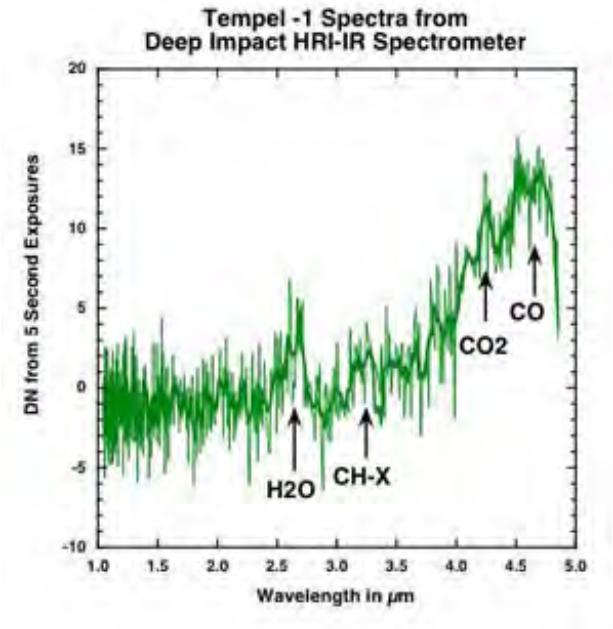
In the weeks prior to the impact telescopes trained on Tempel-1 observed events that proved to look very much like the July 4th impact. These have been noted on comets many times in the past. Two explanations seem most likely: 1) there is an internal gas release which ejects material outward; or 2) some object in space collides with the comet causing a fireball and ejecta. The first illustration below was made through the Hubble Space Telescope on June 14, 2005:



The second explosive event was on June 22 and was observed directly by the video camera on Deep Impact. The escaping plume lasted well beyond 5 hours.



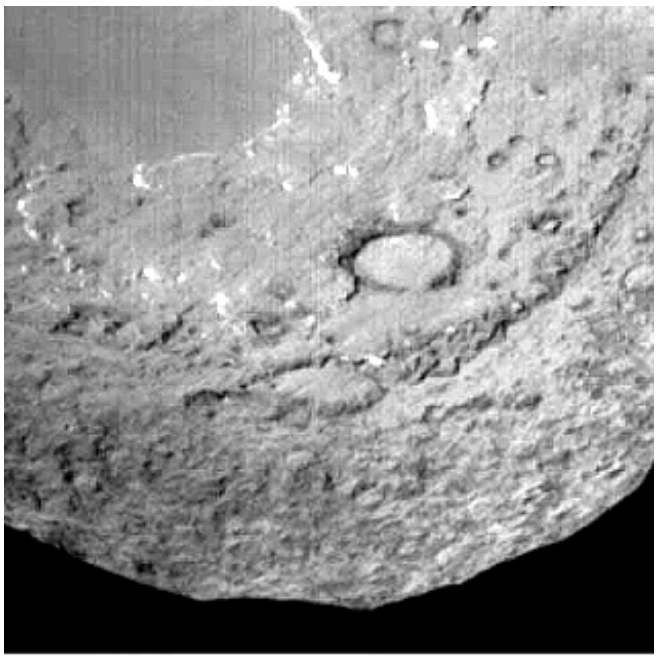
A spectrometer on board were able to determine some of the constituents in the plume:

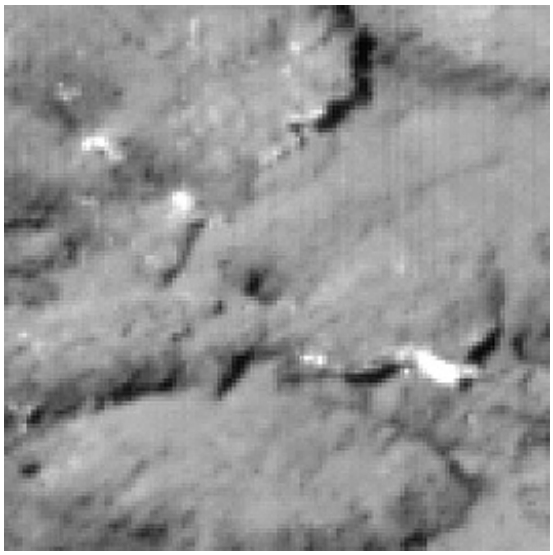


On July 3, 2005 Deep Impact approached Tempel-1. Technicians and scientists at JPL awaited anxiously the final minutes. Some 12 hours before impact the impactor was successfully released from the mother ship and hurtled its way towards the advancing comet. The collision was designed to combine the orbital velocity of the comet with the opposing velocity of the impactor to attain a net impact velocity of 10.2 km per second (6.2 miles per second) or 37360 km per hour (23000 mph). For a mass of the impactor at 330 kg (820 lbs), the K.E. generated at this speed is equivalent to 4.5 tons of TNT. At the time of impact, 1.52 AM EDT, the comet was 134,700 km (83000 miles) from Earth.

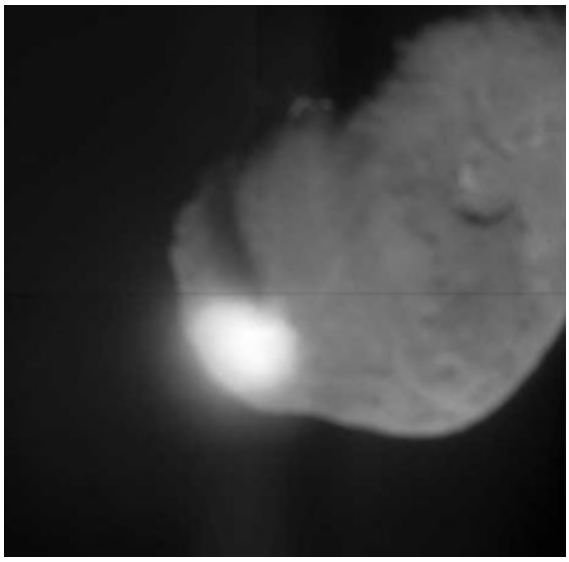
As it closed on target, this view of Tempel-1 was obtained.

The next two scenes show part of the comet's surface and then a close-up moments before the impactor struck about 60 meters from its pre-selected target.

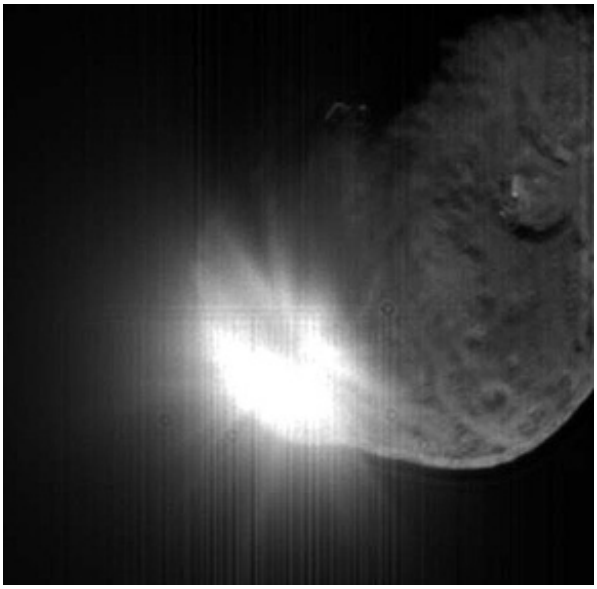




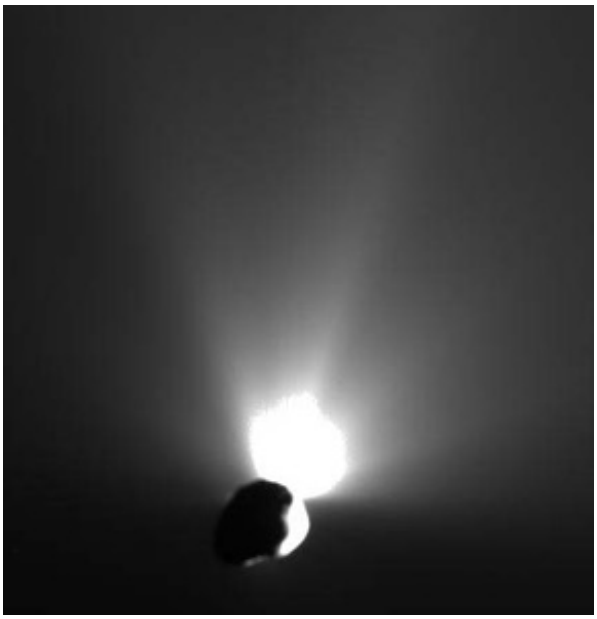
At almost the exact instant of collision, this image was acquired by the mothership.



Just seconds thereafter this image was obtained:

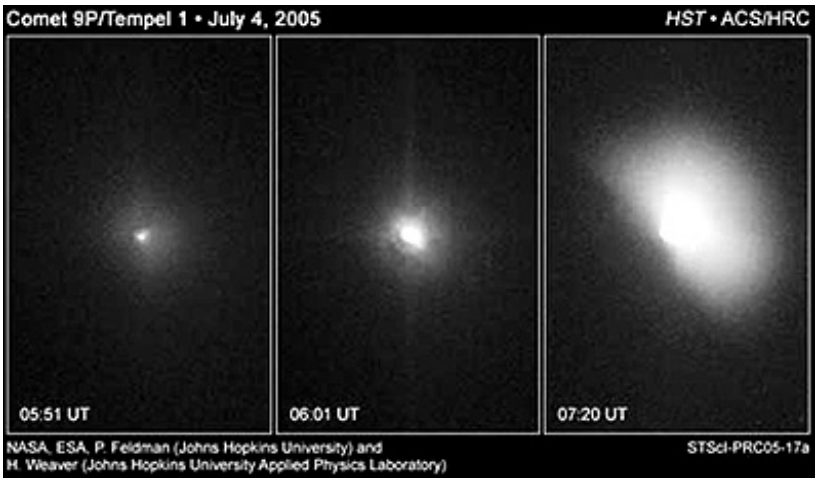


The moment immediately at impact, as imaged from the Deep Impact mother spacecraft, shows the luminous fireball at its maximum extent; the dark area is the comet, the bright is the fireball.

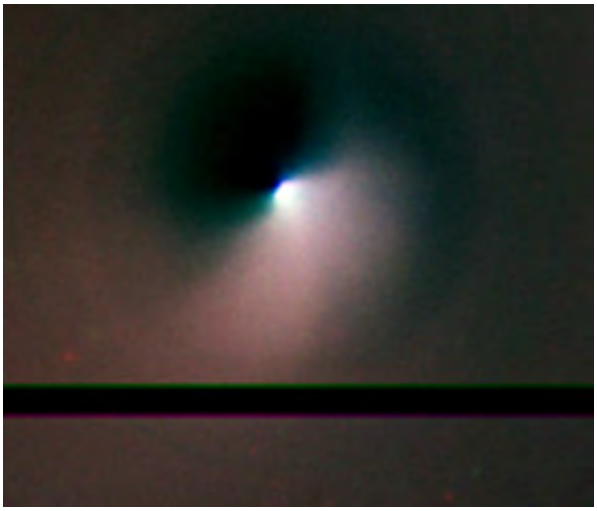


This fireball contained two sources of light: 1) superheated materials from the comet; and 2) light from the Sun reflecting from particles. As the ejecta plume continues to be maintained, it will lose the luminosity of the heated cometary substances but will continue to reflect sunlight. The plume was still glowing 5 hours after impact and will continue to be monitored from the mothership and telescopes over succeeding days. The crater itself was not visible during the first few days, attesting to the masking ability of the still dense debris in the plume.

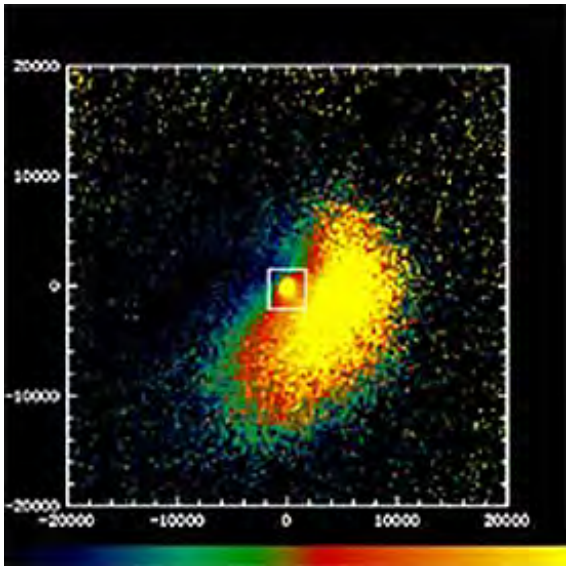
Telescopes worldwide were trained on Tempel-1. The best view from a distance was obtained through the Hubble Space Telescope:



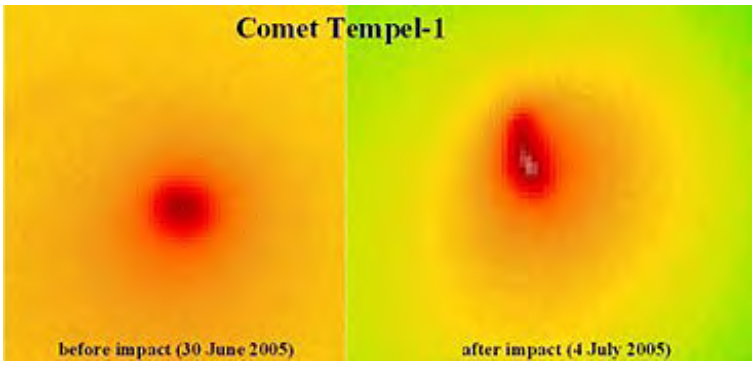
European Space Agency telescopes produced some of the most informative observations. The 8.3 m telescope at La Silla captured this visible light image that shows the dispersal of gas and particulate debris soon after the impact.

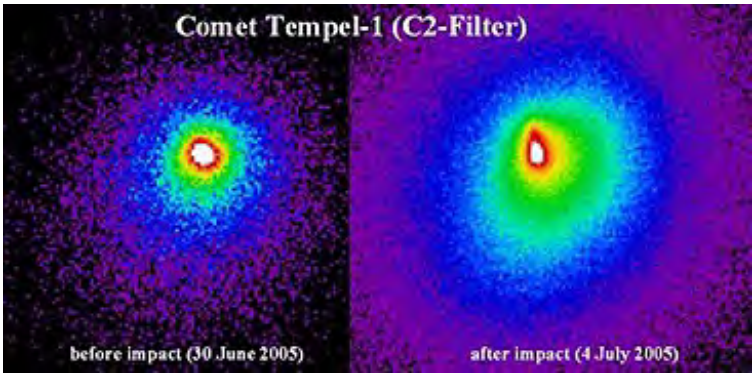


An NTT telescope yielded an image in which the ejecta plume could be roughly density-sliced to indicate variations in the amount of light-scattering debris. The coma of the comet is within the square.

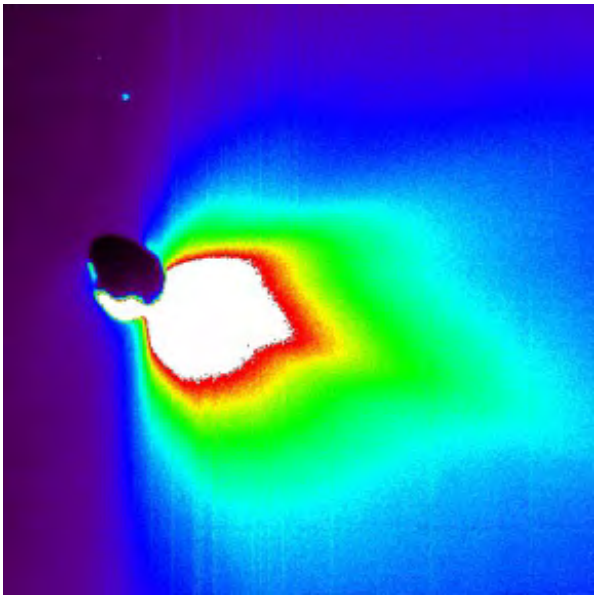


The Optical Ground Telescope in the Canary Islands examined the ejecta using a red filter and a C-2 filter, with these results.





An image from the mothership taken 55 minutes after the impact showed the extent of the still expanding ejecta curtain beyond the comet:



Interpretation of this material has led to the conclusion that most of this curtain is composed of fine dust (face powder consistency), with some water vapor/water. The picture being derived from Tempel-1's surface is that it contains more solids, and less ice, than expected.

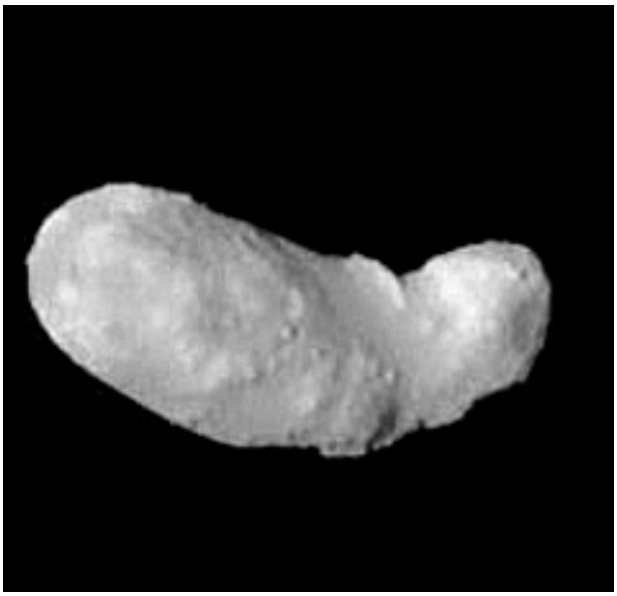
Calculation of the amount of dust ejected helps to estimate the size of the new crater as between 100 to 200 meters (up to 650+ feet in diameter).

So far, only a little specific science results have been released to the scientific community and the public. What is shown above is mainly observational. Unfortunately, the mothership is a flyby configuration and is receding from the comet. The dust may not clear enough to get a good look at the formed crater, as had been hoped. Whether any system will detect the crater at a later time is still uncertain.

But, in September 2005 scientists made this statement: Analysis of the dust made from the Spitzer spacecraft shows the presence of silicates (including Olivine), as expected, but also

small amounts of materials that appear to be clays and carbonates (unexpected). Aromatic hydrocarbons were also detected.

What is really wanted are actual samples of asteroid or comet nucleus materials. Plans for such missions are being worked out. However, as was discussed on page 19-22, on May 9, 2003, the Japanese Space Agency launched the asteroid probe Hayabusa towards a very small asteroid, Itokawa, which is only about 3 km (9000 ft) in long dimension and almost 0.7 km (2000 ft) in cross-sectional dimension. Here is a repeat view of Itokawa:



Itokawa was reached in late September 2005. A tiny rover, MINERVA, was sent to its surface to observe with its cameras and temperature probes. It failed to transmit evidence of a successful landing (either it crashed, instruments malfunctioned, or the craft capsized so that the antenna does not point towards Earth). But more ambitious (and risky) were the attempt(s) to set the mother ship down, sample the surface materials, and have Hayabusa take off for a return (with samples) to Earth for a parachute landing in Australia in June, 2010. The first try was aborted. But in November 2005 a landing and possible sampling occurred. The mother ship appears to have left the comet's surface successfully and is back in a parking orbit but details of sample procurement are sketchy and confusing. Hayabusa will attempt to return to Earth in February 2010 and release any samples in a re-entry capsule but its recovery remains a challenge.

A review of the Haybusa Mission is found at thisWikepedia website.

Thus endeth this great voyage of exploration through the Solar System! The previous 18 Sections should have convinced you that our Earth is an extraordinary, yet very livable place. But, this Section hopefully has opened your eyes to the fact that the rest of the Solar System, while inhospitable, is clearly fascinating in its own right. Recently, astronomers have discovered several new planetary systems elsewhere in our galaxy. Who knows what incredible worlds lie beyond our own? Perhaps someday we'll remotely sense a sister planet and then, when the technology develops, visit it by probes or even by our species.

In the meantime, what should be the priorities in future exploration of the Solar System. NASA, ESA, the Russians and other space programs will make these decisions. But input from the public - especially space connoiseurs - who "foot the bill" - is very much desired. The writer (NMS) has his own favorite 'top of the list': Mars! We now know enough to realize that if that planet had only been larger (thus holding more of its atmosphere) and its potential biogenic development could adjust to the decreased solar energy it receives then it may well have been a favored host for life to develop. As it is, the possibility has yet to be ruled out. Rovers and orbiters will help in the search. Man's visit would be decisive.

Some of the "flavor" of future exploration plans is covered in a JPL von Karman lecture. Access

this at von Karman lectures, choosing the topic "Space and Earth Exploration:2020, November 21, 2002.

Now it is time to expand our remote sensing horizons towards Infinity to show how this technology lies at the heart of the exploration of the whole Universe. The next Section, 20, may be the most provocative of all.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Imagine you are a famed planetary geologist who played the key role in bringing the study of impact craters into the mainstream of astrogeological science. You are among the few who truly believed in the catastrophic dangers of inevitable major collisions between the Earth and large enough asteroids and comets. Your experiences in studying nuclear explosion craters and the much bigger impact craters had you firmly convinced that disasters of that magnitude were genuine threats to mankind's safety and even survival. So, in your later professional years you dedicated much of your time to a task shunned by all but a very few in the know: making an inventory of all detectable asteroids following potential Earth-crossing orbits. Then one night at the Observatory your wife, Carolyn, whom you had trained to help in this search, noticed a shift in a small bright object as she studied photographic plates exposed through the telescope on different nights. This surely was a comet. You contacted your colleague, David Levy, for confirmation - soon obtained. You let out the word to other observatories. Soon thereafter, one of these reported that the comet had split into pieces. Calculations then led to the astounding prediction that the pieces were on a collision course with Jupiter's gravity field and would likely be drawn to it in a series of collisions. No one on Earth had ever witnessed such an event, on this planet or any other. You, the expert on impacts, would realize a life's dream of actually observing and participating in a real time collision. Then, in July of 1994 it happened and the multiple collisions were of such magnitude as to be easily detected by the Hubble Space Telescope in space and many others on Earth. The world at last had come to know of the likelihood that big impacts are part of the Solar System's past and future and that humans had better start thinking of ways to protect our planet.

This is the saga of Gene Shoemaker, probably the most influential planetary geologist since the Space Age opened. The comet, of course, has immortalized his name, that of his wife Carolyn, and their friend David Levy, by being called Shoemaker-Levy. Here is a photo of the three together:

On this page, we will recount this once-in-a-lifetime happening, which captivated millions as it happened and has pushed scientists and even the military into serious thinking about the genuine threat impacts pose to life on Earth.

Comet Shoemaker-Levy

In March of 1993, Carolyn Shoemaker, wife of Dr. Eugene M. Shoemaker* (the Dean of

Astrogeologists and chief guru of the Impact School), and an accomplished teacher turned scientist, known particularly for her contributions in the search for asteroids and comets, inspected photos taken days apart using surplus film, from a damaged batch having high sensitivity. A small bright elongate object had moved relative to star reference points. As it was imaged later at higher magnification, it was revealed that the short streak displayed a strange but distinct pattern of separated dots

About the same time, their colleague and member of their asteroid search team, David Levy, made similar observations. After reporting their find to other observatories, word came back that the odd arrangement of dots was actually due to a string of individual cometary bodies, each with a small coma and tail, stretched out for more than 700,000 km (435,000 miles) along an orbit tied into Jupiter. Then, in May of 1994, word came that this comet, by then named Shoemaker-Levy 9 (S-L 9), was on a collision course that would end in a succession of hits onto Jupiter in July of that year. This would be a unique event–the first time humans would witness a major planetary impact!

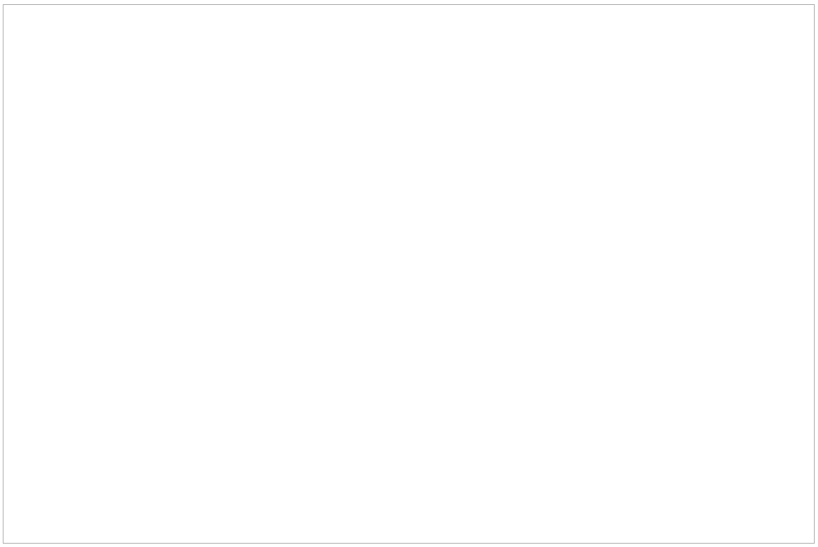
Scientists believe that Comet S-L 9 originated beyond Neptune. Over centuries of travel, various bodies perturbed its orbit towards Jupiter, so it became a Short-Period comet, whose very eccentric orbit carried it remarkably close to that planet. Calculations showed that S-L 9 had earlier passed as close as 20,000 km (12,428 mi) to Jupiter's surface, well within the Roche Limit (a distance from a planet's center, within which its gravitational forces are likely to disrupt a passing body), and had broken apart then into 21 fragments. Each nucleus was less than 2 km (1.25 mi) in diameter. The HST made this dramatic image:

Just before reaching Jupiter, the Hubble Space Telescope took this view of some of the comet's

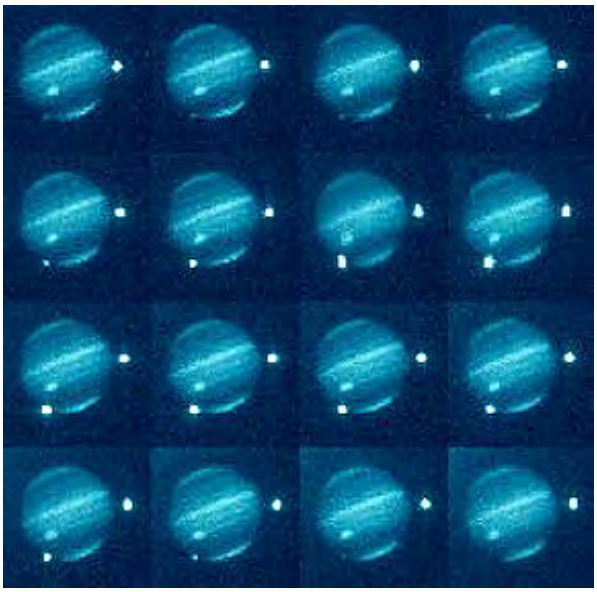
fragments, enlarged enough to show cometary material ablating off in a direction away from the solar wind:

The world of planetary science quickly and efficiently mobilized for this collision event. Calculations pointed to a high probability that it would actually hit Jupiter in 1994. As more precise estimates of just when and where the collisions would occur, scientists realized that HST and Galileo would be in position to observe at least some of the impacts, and Earth-based telescopes should also see something. Guesses as to what would happen ranged from penetration into the Jovian atmosphere without much occurring to massive fireballs. Anticipation ran high among the network of watchers.

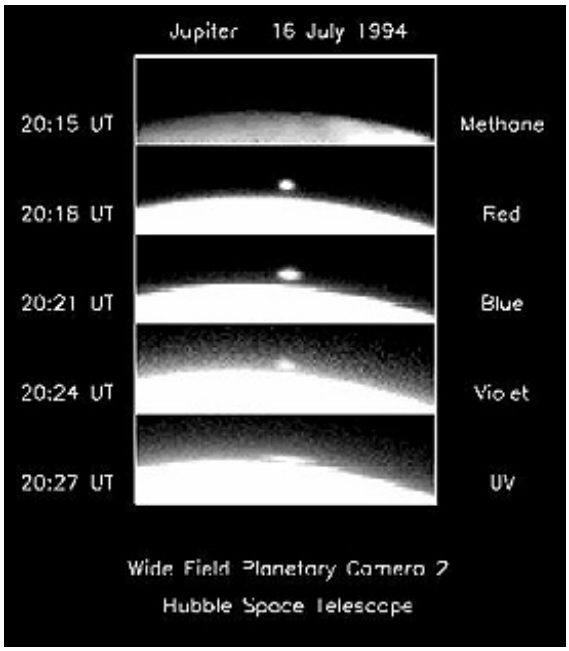
First to hit, on July 16, was the small nucleus at the left end of the string in the image above. Labeled A (with succeeding ones designated B, C, D ... through W [with I and O omitted]), this body struck Jupiter in its southern hemisphere, producing a detectable flash and subsequent plume picked up by observatories in Spain and Chile. Below, we display the sequence of color IR images made through a telescope fitted with a methane filter (which absorbs radiation at most wavelengths but passes radiation in a narrow band around 2.3 µm):



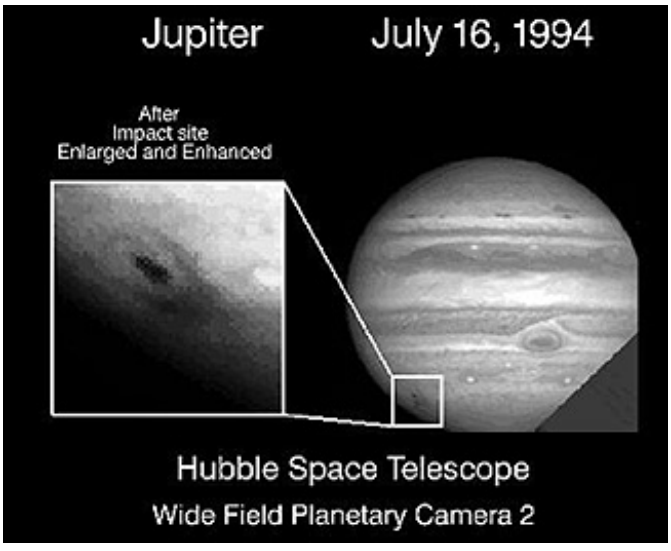
The changing bright spot on the lower left surface is the thermal flash of the impact (the other bright oval above it is the Great Red Spot and Io is visible to the right). A more detailed sequence of observations of event A was obtained through the telescope at the South Australian Astronomical Observatory



The Hubble Space Telescope obtained impact images at several wavelengths:

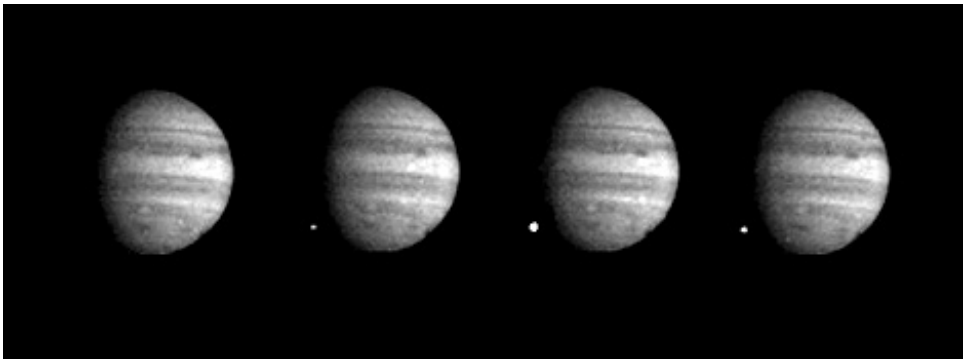


And HST obtained the first view of the impact site - a black scar, but no crater (any depression would be transient in the jovian atmosphere)

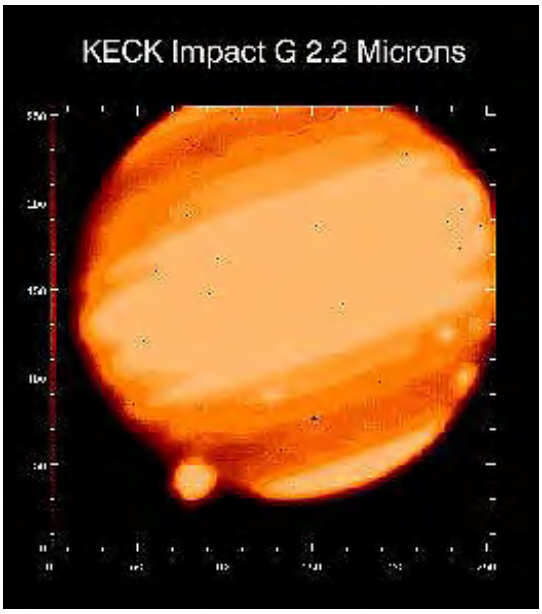


Astronomers monitored most of the subsequent impacts over the next six days either directly or

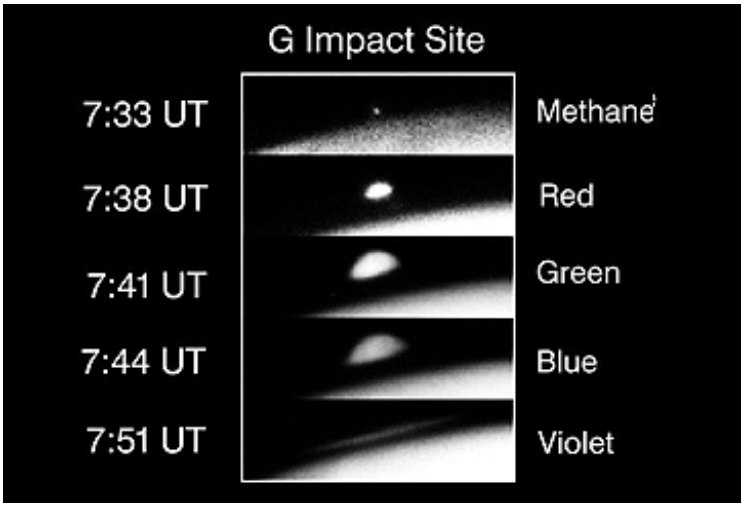
as the impact scars moved into sight during Jupiter's rapid rotation. Galileo's camera caught several on the night side limb. Next, we show W's hit on July 22, sequenced seven seconds apart, showing the flash during and shortly after impact (it looks like a satellite off the planet but is really a momentary response set against the dark surface within the shadowed limb):



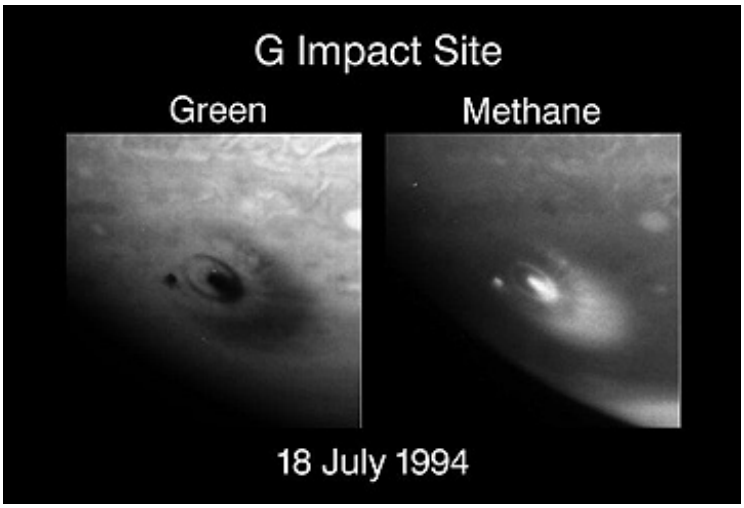
One of the biggest events was the strike of nucleus G on July 18. As imaged at the Keck Observatory:



The rapidly rising plume from this, imaged in a time sequence by HST, shows it reached a height of 3,000 km (about 1,864 mi) in about six minutes:



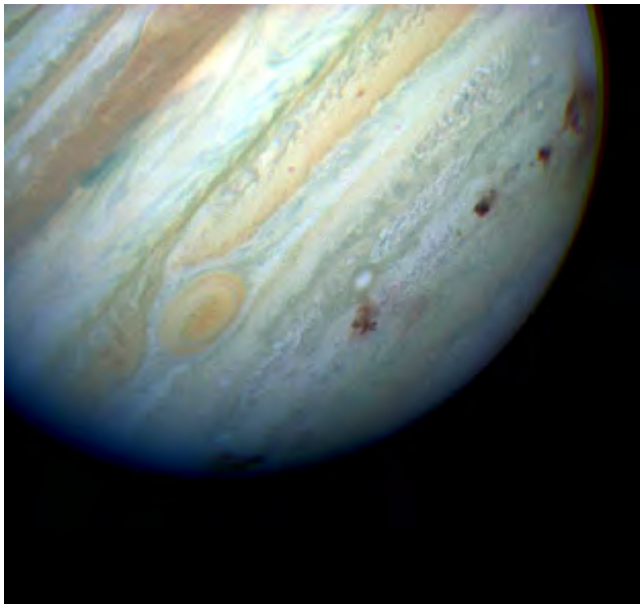
The HST imaged the scar (larger than Earth's diameter) left behind in the surficial layers, through a green and a methane-based filter:



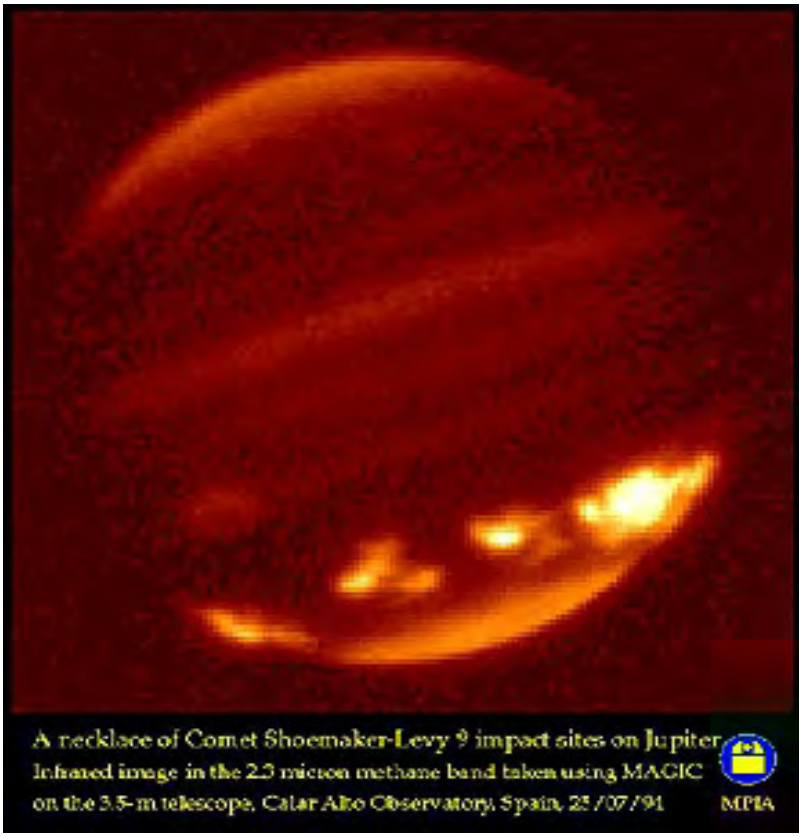
19-76: How do you explain that the impact of the G fragment, probably only a few kilometers in size, which on Earth would make a crater several hundred kilometers wide, actually made a visible scar on Jupiter that was at least as big as the Earth's diameter itself? ANSWER

Interpreting this, and other scars, shows us that each nucleus punches a tunnel into the Jovian outer atmosphere, shocking, compressing, and flashing its gases. Materials from the incoming comet and the atmosphere carry into the hemispherical plume and also sidewards (influenced by the angle of impact) as crescent-shaped clouds, in much the same manner as those associated with terrestrial impacts on Earth. The dark matter making up the scar is presumably color-altered constituents from the Jovian atmosphere,such as, HCN or sulphur derivatives. Unlike solid ground craters, those on Jupiter gradually dissipated, although their visible traces endured for several years.

The following three views of Jupiter, after all impacts had occurred, demonstrate this persistence. The first is a natural color view made by HST of one side displaying impact sites



The second and third show a color IR images, the first acquired by NASA's IR telescope on Mauna Kea, Hawaii and the second taken from the Calar Alto observatory in Spain:

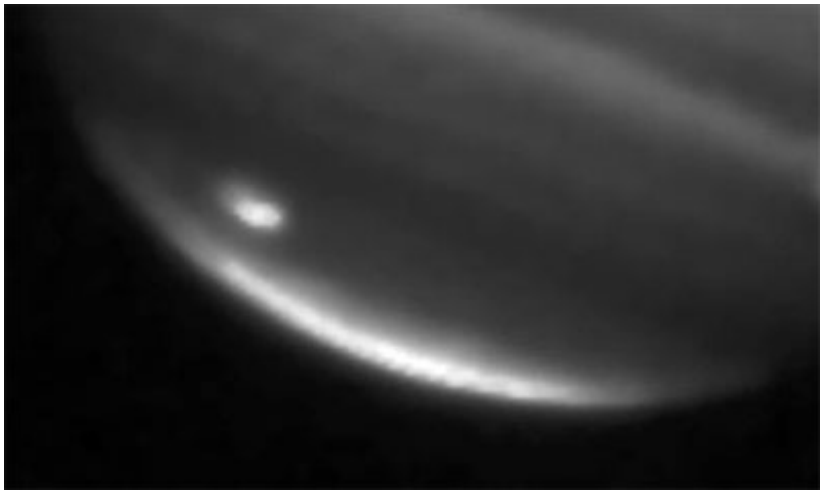


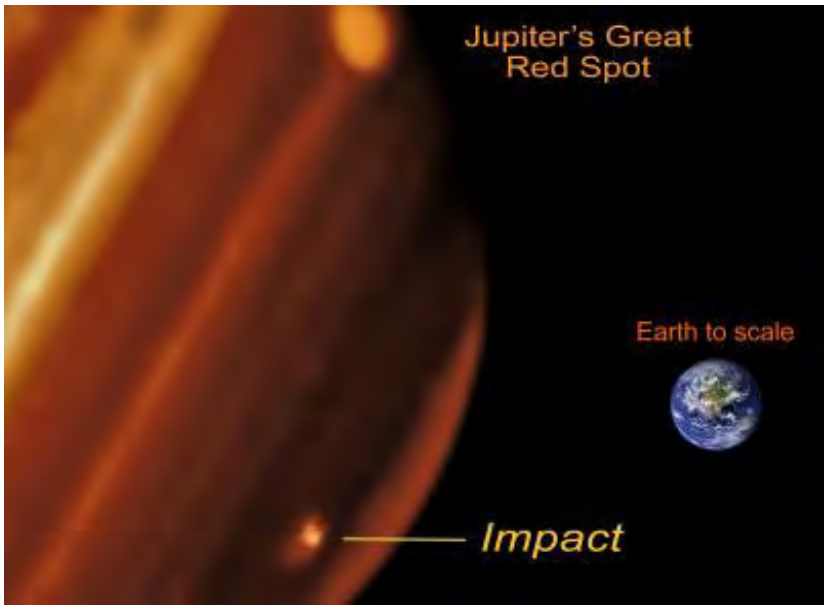
Another image, taken by HST's UV camera, brings scars H, Q, R, D, G and L (the round object in the upper hemisphere is the shadow of one of the moons) into sharp contrast:

Had any of the fragments from S-L 9 struck Earth instead, a catastrophe of a magnitude never before experienced by humanity (but certainly by the dinosaurs) would have ensued. We've been warned!

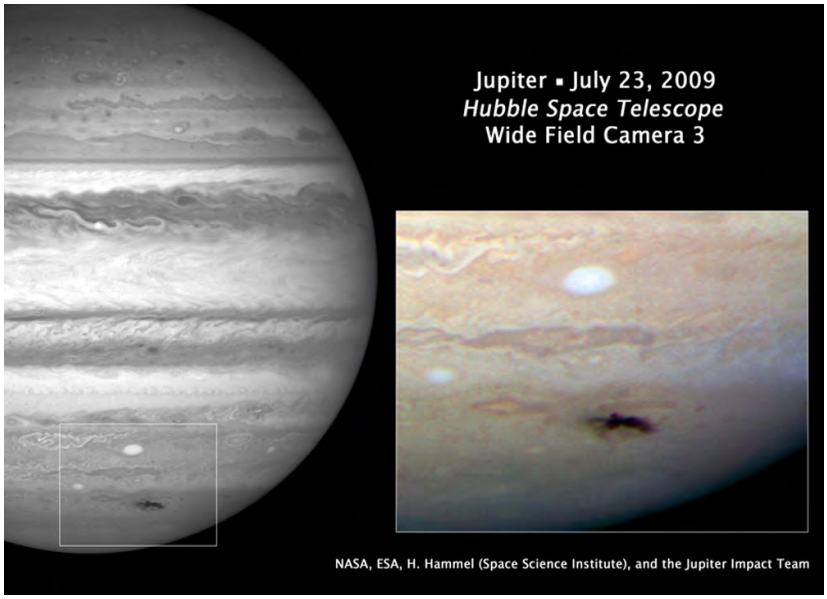
Two good Web sites that cover the Shoemaker-Levy event are: (1) and (2)

Shoemaker-Levy was thought to be a 'once-in-a-lifetime' event. It was somewhat surprising, at first, when another large impact occurred on Jupiter on July 20, 2009. A new transient spot appeared near the south pole. Here are two images:





The Hubble Space Telescope has taken a follow-up picture of the scar that shows it, like those of Shoemaker-Levy, has by then been smeared out:



The initial impact scar was the size of the Pacific Ocean. No one foresaw this impact. No object in space was observed to be closing on Jupiter. The object most likely was an asteroid or comet. The implications of this event are troubling: 1) large impacts may be more frequent than some have thought; 2) such impacts can occur in short succession; 3) impactors can be missed, even with watchful search programs; and 4) Earth is certainly not immune to being a target.

Shoemaker-Levy taught us what to expect from impacts into liquids and gases. Our concern with that mode is real: asteroids/comets could hit the Earth's oceans. But of greater danger could be impacts on land. NASA conducted a mission to comet Tempel-1 named Deep Impact, covered on the next page - that not only added information about impact on a solid surface but also gained knowledge about the nature of that surface

---

(*) **With deep sorrow, the writer (NMS) must report here the death of a good friend and teacher of impact principles, Eugene M. Shoemaker, who may rightly be called "the father of Astrogeology." He was killed on July 18, 1997 (three years to the day after the main SL-9 event; and also on my birhday), in an auto collision (impact, ironically) in the interior of Australia, while pursuing meteor craters that are so well displayed on that continent. Read about his life on this Wikipedia biography.**

This photo captures very well the Gene Shoemaker I so respect. He is laughing, something that came easy for him. He is in formal garb (receiving an honor) but with his ubiquitous western lanyard instead of a tie.

It is fitting that part of Gene's ashes were onboard Lunar Prospector so that when that probe finally descended and crashed onto the lunar surface in search of water, in a sense he too "landed" on the Moon–a long time goal of his (to be a lunar astronaut) that health problems thwarted just prior to the Apollo program. A second honor was posthumously awarded him by renaming the NEAR probe to the asteroid 433Eros as the NEAR-Shoemaker spacecraft. And, to memorialize him on Earth, the meteoritical community, including the Australian contingent, has renamed the 30 km (18 mile) wide Teague crater to the Shoemaker crater, as seen below:

I have since visited his widow, Carolyn. She is carrying on the work that so preoccupied him in his last years: looking for asteroids and comets, particular those that might be a threat to Earth.

Primary Author: Nicholas M. Short, Sr.

The Earth formed simultaneously with the other Solar System planets and the central Sun. Accretion of planetesimals produced a large body which assumed a spherical shape. Probably cool at the outset, this proto-Earth rapidly heated up, formed its metallic core within 100 million years, and was subjected to continuous impact bombardment by asteroids, comets, and meteorites. It may have had a molten exterior which quickly cooled to a crust. Very early in earth history, its Moon was produced from a glancing collision with another planetlike body. A second period of bombardment helped destroy the early crust. By about 3.8 billion years ago, rocks formed crusts of more silicic rocks embedded in a basaltic crustal layer that extended worldwide. Oceans were produced early, weathering attacked the crustal rocks to produce the first sedimentary rocks, and protocontinents began to form (from metamorphosis and melting of accumulated debris). These continental nuclei probably were moved about by processes akin to convection-driven plate tectonics. An early atmosphere consisted largely of nitrogen, with some carbon dioxide, ammonia, methane, and water. Those ingredients may have been converted to organic molecules which in turn organized into primitive one-celled bacteria about 3.85 b.y. ago. In time, living plant organisms developed the capability to photosynthesize solar energy, releasing oxygen as an end product, which gas gradually built up to present day levels, evolving more advanced life forms.

The Earth as a Planet

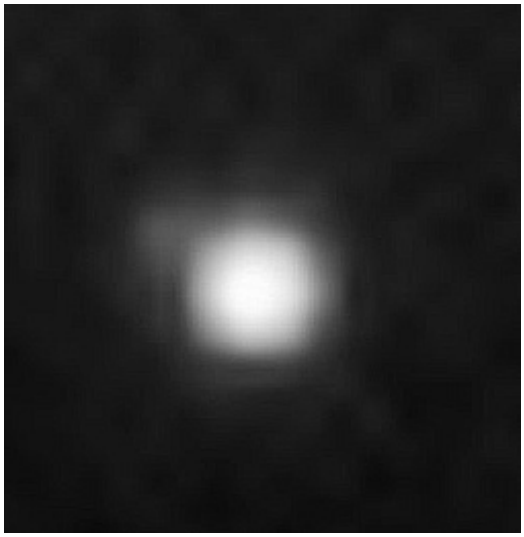
Since it is our home, we all know a great deal about Earth. Photos of the "blue" or "water" planet have appeared many times on TV and elsewhere in this Tutorial. So we are familiar with its appearance close-up as seen by a meteorological geostationary satellite.

But here is a full face view of Earth as taken from just above the Moon by the Clementine satellite:

As seen through a telescope from a distance of about 140 million kilometers, the Earth and its Moon would look something like a double star. Such a view was actually obtained by the Messenger satellite when it was on the other side of the Sun relative to Earth, so that it spots the terrestrial body because of reflected solar light, as seen here:

When seen in true color from the Voyager spacecraft near Mars, the Earth appears as a blue dot (tone largely from the oceans) of moderate brightness:

Seen from a distance by a more powerful telescope on the Cassini spacecraft as it orbits around Saturn, the best image of Earth from the outer Solar System once more shows it to resemble a star (the Moon was behind Earth at the time), with almost no details discernible - this gives an insight into how difficult it is to study the solar planets from earth-based telescopes and thus justifies why we have sent unmanned probes to all the planets to obtain close-up views:

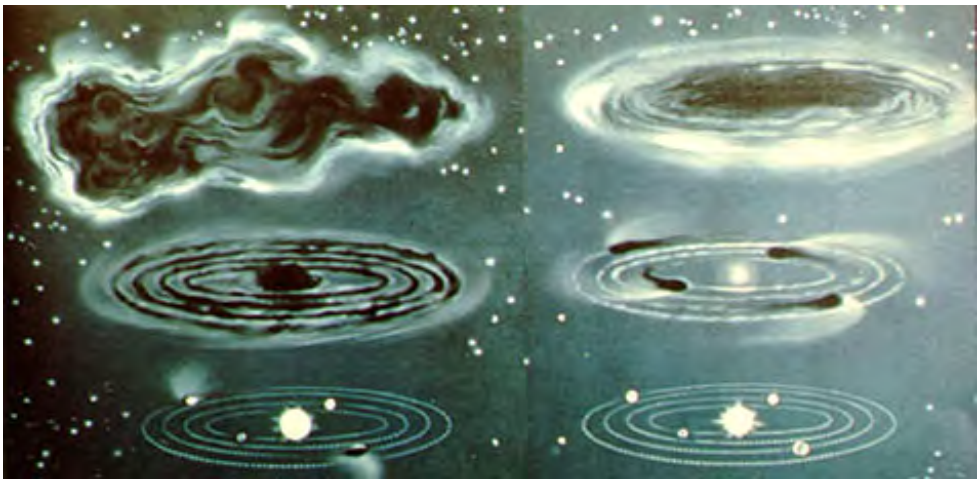


But the Cassini image was taken through its telescope and is thus magnified. How would Earth appear to the naked eye from, say, Mars. Here is the first such unmagnified image made by the Mars rover Spirit. Our planet is nearly invisible in the twilight night glow that pervades space from the Sun outward , but the bright dot - Earth - has been enlarged in the inset. Under these conditions Earth is as small as Mars appears to us as we look up from our planet.

Earth is the largest of the four inner rocky planets. It almost certainly began to organize in the earliest days of the Solar System, along with its sister planets, even as the Sun itself came into being as a ball of hydrogen-helium gas mixed with heavier elements. Some of the gas and much solids - mostly dust size - remained outside the central region of the gas-dust "cloud" that comprised the protostar system that evolved into the present day Solar System. The best estimate of when this all began, based on meteorite age data (in which the primitive meteorites are assumed to record the accumulation of the dust within the local cloud that gathered into small objects), is between 4.55 and 4.6 billion years ago. The Earth started to organize soon thereafter.

A momentary digression to mention what will be covered in detail near the end of Section 20 - the conditions that led to the origin of the Solar System and the Sun. The overview in this paragraph is germane to the next few paragraphs: Present thinking considers the System to have formed from an organizing event that affected a cloud of hydrogen and dust. That event may have been a supernova explosion of a nearby star. This would account for the presence of considerable amounts of the heavier elements. The explosion could have intiated some compression of the cloud such that gravity began to bring matter together, with the Sun being the "sink" into which the dominant hydrogen moved, along with some of the heavier elements. Rotation commenced as the cloud organized into the Sun and solid and gaseous matter beyond. Clumps of matter in the extrasolar segments of the rotating cloud began to collide and build up into a few larger protoplanets. In time, much of the initial material in the space beyond the Sun that was influenced by the Sun's gravity as it built up was swept up to produce the present collection of planets, asteroids, comets, and interplanetary gas and dust.

This illustration follows one of several similar models that describe these formative phases, as applied both to Earth and to the planetary system as a whole:



Possible source: Indiana University

The upper left panel shows the local gas-dust cloud as initially cold, but with some heating as it contracts (upper right). The middle left panel suggests that the cloud has now contracted into a protoplanetary disk, within which light H and He gases are drawn mainly to its center, where this material contracts rapidly into the early Sun (middle right). The bottom two panels indicate that material in the disk has separated into rings of denser materials which begin to heat up as their constituents are swept up into ever larger bodies that collect enough solids and gases to build into individual planets that gravitationally rework the hot solids (probably with partial to nearly complete melting) into spheres.

An actual example of a star-forming gas/dust cloud is shown in the Hubble Space Telescope image of a nearby nebula:

Beyond the forming Sun, solids start out mainly as dust-sized particles that collide and accrete into larger (meters range) bodies. Many of those further interact to grow into bodies in the kilometers range, which are called "planetesimals", some of which have been sampled as meteorites. The asteroids are an example.

Painting courtesy W.K. Hartmann; copyright

Collisions persist, with some bodies growing larger than most others, until they reach sizes

similar to present day planets or planetary cores. Thus, the bigger bodies use their growing gravitational force to attract smaller bodies that collide and accrete onto the increasingly larger bodies. The Sun, meanwhile, is organizing into a gas ball in which contraction raises internal temperatures to greater than 1,000,000 ° C , to the extent that nuclear fusion commences. The early Sun had less energy output than it does today. Its surface temperatures then were less that at present. The fusion produces a steady solar wind that drives particles outward. Most of the original H and He retained at first around the inner planets is pulled off into space, in part because the lower gravity of these bodies cannot hold on to these low atomic weight gases. The outer planets, being further away, experience less solar wind and retain large amounts of H and He, giving them more mass and stronger gravitational pull that maintains their thick gas envelopes (including a mix with gases having other compositions).

All of this probably took less than 100 million years to accomplish, meaning that the planets and Sun are approximately contemporaneous. In the Solar System spatial realm, beyond the planets, a large quantity of gas and solids remained. Meteoroids developed from these solids, some remaining at a range of planetesimal sizes (these thus are primitive, carbon-rich and water-bearing, making up the carbonaceous chondrite type of meteorite falling on Earth) with others enlarging through collisions to 100s of kilometers. Examples of these planetesimals are still mostly in the Kuiper Belt beyond Pluto. While this was going on, much of the dust experienced some degree of melting (from solar energy bursts, electrical discharge, shock waves, radioactivity[?]) that produced small spheres of Fe-Mg silicates (chondrules; see previous page 19-2) that after being dispersed are recombined into newer planetesimals.

A few planetesimals appear to have reached sizes that led to complete melting and differentiation (by gravitational settling) into silicate bodies up to planet size, most likely with iron-nickel cores. Today interplanetary space is occupied by survivors of this early aggregation history as asteroids, comets (solid silicates mixed with frozen water and/or carbon dioxide ice [the so-called dirty "snowballs"]), and smaller meteoroids (many are broken pieces of small protoplanets).

Most of the bigger bodies underwent general (usually total) melting. The Earth is believed to have experienced almost complete melting in its early days. There are several sources of the heat needed to raise these bodies to temperatures in excess of 1000° C. Heat from collisions (impact) during accretion can be large. Heat of compaction may be involved. But the main source is the heat released during radioactive element decay. Uranium and radioactive Potassium were important. In the earlier history of Earth, radioactive Al^{26}, with a short half-life but in abundance then, probably was a significant contributor.

The larger bodies once molten adopt near-spherical shapes (the result of equal pull in all directions from internal gravity). The general arrangement of planet distribution in the Solar System - four "small" inner planets (the *Rocky Group* Mercury, Venus, Earth and Mars, the latter three with thin atmospheres) and four "giant" outer planets (the *Gas Ball Group* Jupiter, Saturn,

Uranus, and Neptune with thick atmospheres)- may be unusual (in the light of recent observations of planetary systems around other stars). These planets have survived now for nearly 4.6 billion years. The giant planets influence the rocky planets in two vital ways. First, their gravitational pull counterbalances that of the Sun, thus helping to maintain the inner planet orbits. Second, the giants' large masses serve to preferentially attract small bodies such as asteroids and comets, pulling them into the gaseous envelopes around Jupiter, Saturn, Uranus, and Neptune, thus reducing the number that would otherwise hit the inner planets; early on they also attract gases in large quantities that formed the bulk of their volumes.

The position of Earth in the Solar System lineup seems fortuitously favored for its unique plethora of living creatures. It is far enough from the Sun to avoid having much of its atmosphere blown away by the solar wind. It is large enough to have retained that atmosphere (which continued to grow during terrestrial degassing and has evolved over time to support life). Its atmospheric-affected temperatures are conducive to life's origins and survival. And its surficial temperature conditions have permitted liquid water to be retained in large quantities, even in the early days of Earth's history.

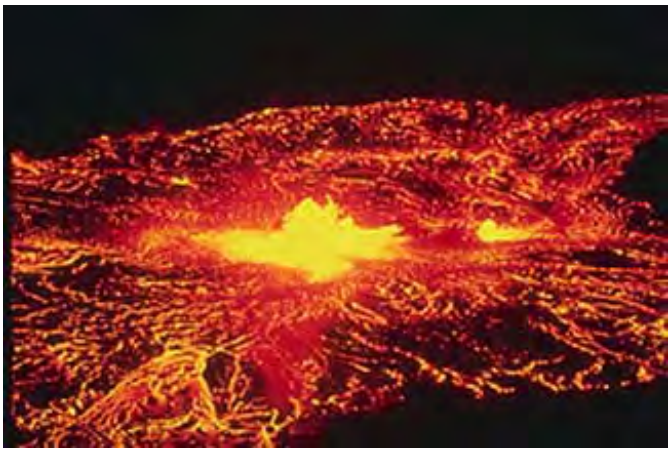
Many of the meteorites that fall on Earth come from the Asteroid Belt between Mars and Jupiter. Some planetologists consider this belt to be the remnants of a fully formed planet (which on melting provided an iron core and iron-stony inner mantle that we observe in about 10% of the meteorite falls and finds). Others hold that this belt never organized into a larger planet because of the disrupting influence of Jupiter's gravity field.

More information on the above topics can be found on pages 20-11 and 20-12.

Let's zero in hereafter just on the Earth. It built up rapidly to a fraction of its present size. As it grew, it became hotter, owing to three sources of heat mentioned above. There is reason to believe that this proto-Earth underwent complete melting in its first 50 million years, with heavier iron and nickel, and smaller amounts of other heavy elements, sinking centripetally to form the Fe-Ni core, and the bulk of materials that are dominantly iron-magnesium silicates making up the outer Earth in what remains today as the Mantle. Various planetologists argue about the "completeness" of this melting - some restricting it to just the outer layers, forming the so-called "magma ocean", others accepting general melting. The indication that the core is almost all metallic iron + nickel, a deduction based in part on the nickel-iron meteorites that presumably were involved in another, now disrupted planet, would be most easily explained by complete melting. The next illustration is a conception of the early Earth surface as a partial magma ocean with "islands" of cooler solid bodies

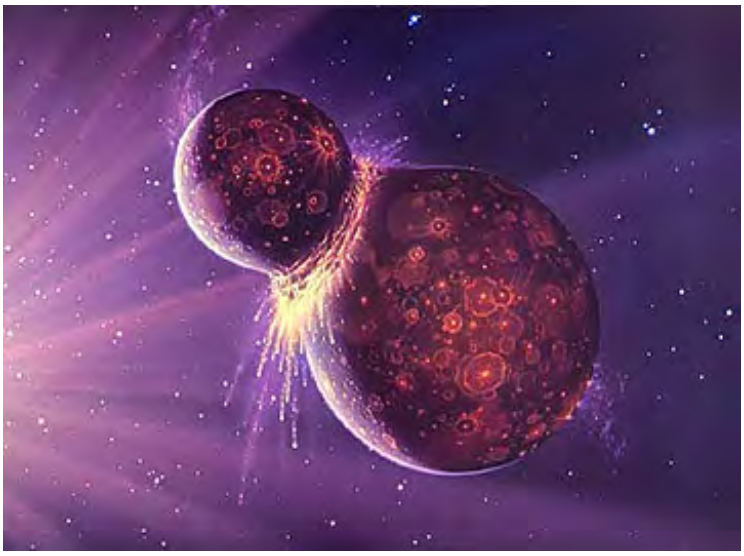
It is reasonable to assume, in this model, that the magma oceans cooled sufficiently to form a thin solid crust of compositions similar to today's basic igneous rocks (principally basalts at and near the surface and gabbros and ultrabasics [e.g., pyroxenites] at greater depths). Parts of such a crust floundered, much like the crusts on the lake in Halemaumau Crater at the summit of

today's Hawaii's Kilauea Iki volcano.



In the first billion years of Earth history (often called the Hadean Eon), the Earth's surface as it cooled continued to be bombarded by comets, asteroids, and meteoritic material at rates much greater than in more recent times thereafter. Here, again, is an artist's view of this process.

A major, near catastrophic event, seems to have occurred in the first 50 million years after the intial spherical Earth had formed and started (completed?) melting. A body (Planet X) roughly the size of Mars, i.e., another terrestrial-type planet or a huge asteroid, collided violently with Earth but more by grazing (at an angle near 30°) than direct, head-on impact, which would almost surely have totally disrupted the Earth. This is the event that tossed a consider amount of outer Earth material into space to collect as fragments in orbit. These gradually coalesced and melted into a spherical body that formed the Moon. The first stage of this is shown in the next figure, with the Moon's formation discussed in detail on page 19-6b.

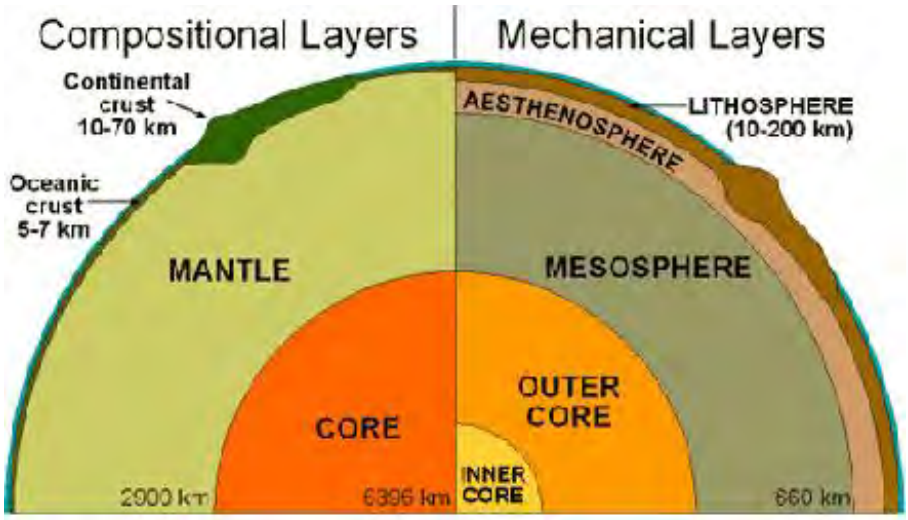


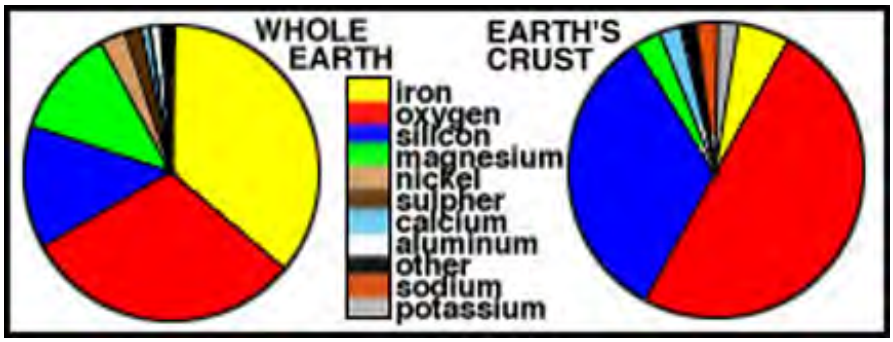
From Universe Today

Another, less plausible variant of this model has both the Earth and Planet X mutually breaking each other apart; most planetologists now think in terms of the Earth remaining intact. Some of the incoming body's material was added to the Earth (which may have been smaller then but thus grew to its present size by collisional additions). The Earth may have largely remelted and resumed core formation. Much of the ejected material circled the Earth close in (initially the Moon was about a tenth its present distance from Earth, but it has gradually receded outward owing to tidal friction losses that lessened gravitational interactions). This orbiting material rapidly organized into terrestrially-derived planetesimals - these coalesced by accretion until the just-born Moon itself melted and organized into a spherical body. Chemical similarities (including diagnostic isotopic compositions) establish this Earth-Moon kinship and are consistent with an impact mode of origin. The Moon, as it melted completely to form a very small, now solid iron core, differentiated, and formed a feldspar-rich crust (the anorthositic Highlands discussed later in this Section).

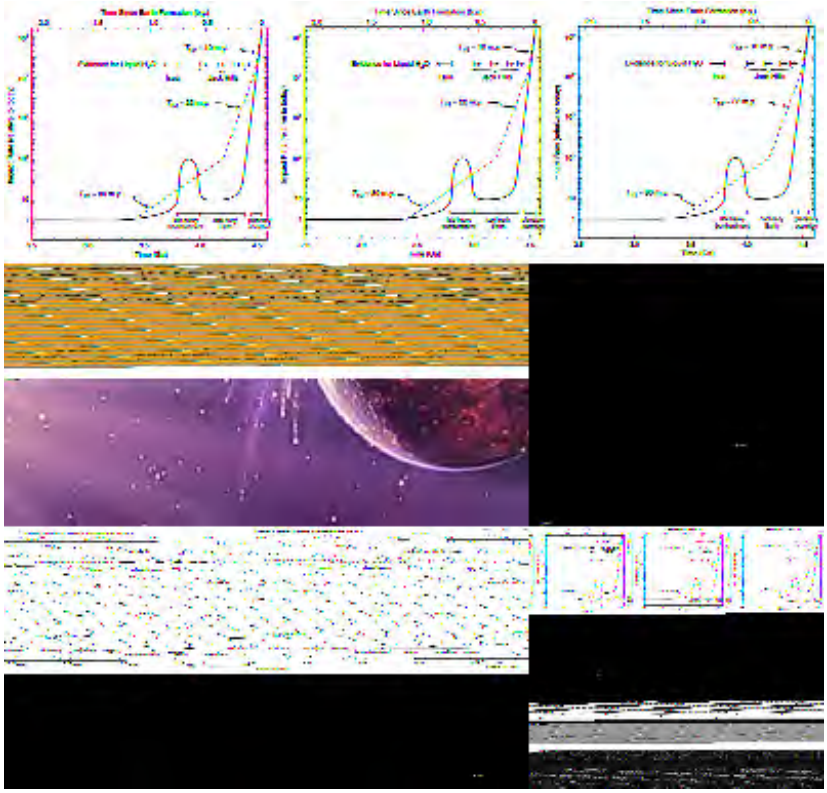
One vital consequence attributed to this collision was that the Earth's rotation axis was tilted 23°: from the orbital ecliptic by the process. This gave rise to the variations of illumination of different parts of the planet during an Earth year. During the early Hadean the Earth day was just 6 hours long because of more rapid rotation; tidal interactions with the Moon have been slowing the rotation progressively, so that earth days are getting longer.

The early Earth melted during its first 100 million years, setting into motion the process called differentiation. This process spatially separates elements according to their atomic weights and chemical affinity. During this the iron (and nickel) collected in a large interior core, above which is the mantle consisting of silicates rich in iron and magnesium (the mafics), and in the Earth's outer exterior a crust of aluminum-rich silicates (the salics or felsics). These next two illustrations describe the structure of the Earth (as it is now but was largely developed in the first billion years) and the composition in terms of total Earth and outer crust.





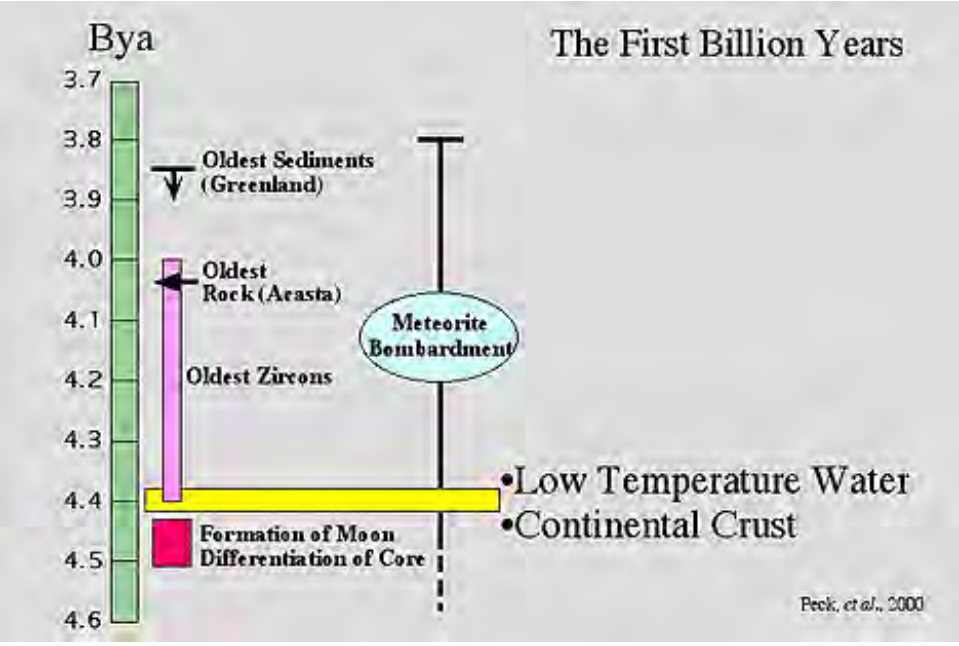
Another major event in the first billion Hadean years was what is known as the "Second Bombardment", in which a notable fraction of larger objects near the Earth were involved in frequent impacts onto the planet and its satellite. The evidence for this is inferential: Study of lunar chronology has demonstrated a heavy bombardment about 3.9 billion years ago. The Earth, a much larger target, must also have experienced this bombardment but nearly all rocks that would have recorded these events have been destroyed.



A very important part of the story of early Earth involves the transition from solid crust over the entire globe to a stage in which the Earth became the "Water Planet". Evidence is mounting that

this may have begun as soon as 150 to 200 million years after Earth's first formation. The extent of the water over the surface may have been regional or even more widespread. The source of the water was to some extent from the interior through venting from volcanoes and lava outpourings. Many planetologists believe that at least some (the proportion still debatable) water was added from comets, asteroids, and meteoroids, many of which contain significant but variable amounts of bound or free H_2O. The critical condition behind water's accumulation is that the surface and atmospheric temperatures were between 0 and 100 degrees Centigrade. The primary consequences of this water buildup and its activity through rainfall and flow are that sedimentary rocks could form and, if other factors were operative, primitive organic molecules (possibly including some "life" at the bacterial level) might have originated and survived. The presence of this early water has been hypothesized based on the discovery of ancient rocks contain the mineral Zircon (see below).

Let us delve deeper into the history of Earth' first billion years, using as a guide this diagram.



Since we conventionally decipher terrestrial history using rocks and fossils, this would seem the "route to go" in describing this first Hadean Eon (which is followed by the beginning of Archean time about 3.8 billion years ago, an Eon that lasted until 2.5 b.y. ago, when the Proterozoic Eon [about 1.9 billion years in duration] was initiated). Unfortunately, Hadean rocks of any kind are extremely rare (Australia; Greenland) and discrete fossils from that Eon have yet to be found, although indirect evidence for simple microbes exists. Much of what is statable about this Eon is speculative.

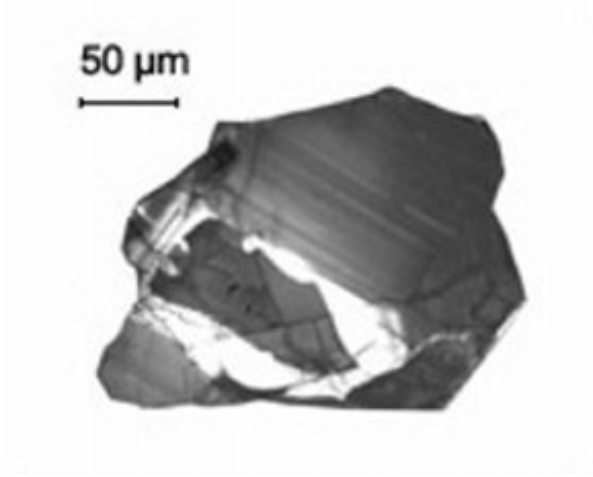
If there was a magma ocean, the crust would have started to form from outlier "rafts" that eventually were enclosed by the first thin solid crust that survived remelting. From knowledge of other terrestrial planets, that crust was almost certainly basalt-rich (a more general term is simatic, which refers to igneous rocks low in silica and high in iron, magnesium, and calcium) in composition (the idea of a thin floated feldspar anorthositic analog to the Moon has been discounted by some planetologists [but remains an alternative]). As the crust thickened, parts of it also were remelted repeatedly by large impacts and probably also by internal thermal convection currents from the mantle. There is reason to believe that a simatic subcrust exists worldwide today within the lithosphere, as a modified survivor of the protocrust.

From knowledge of differentiation mechanisms occurring in the younger Earth, it seems plausible that here and there sialic (high in silica, aluminum, and sodium) crust formed regionally. This crust could have compositions described as rhyolitic, felsitic, andesitic. Conceivably cooling conditions might even have formed granites, which are a common host of the accessory mineral Zircon (see below). Another way to form sialic rocks is by metamorphism of sediments that have been

enriched in Si, Al, Na. Regardless of mechanism, the end result was to develop clots of silica-enriched crust that rose above the general crustal elevations - these would form nuclei (analogous to the term craton in the Earth's present geology) that grew mainly by accretion to their boundaries (perhaps by obduction or terrane addition (see latter half of Section 17), especially if plate tectonics mechanisms developed early in Earth history. Remnants of these nuclei are probably buried within rocks making up the continental shields.

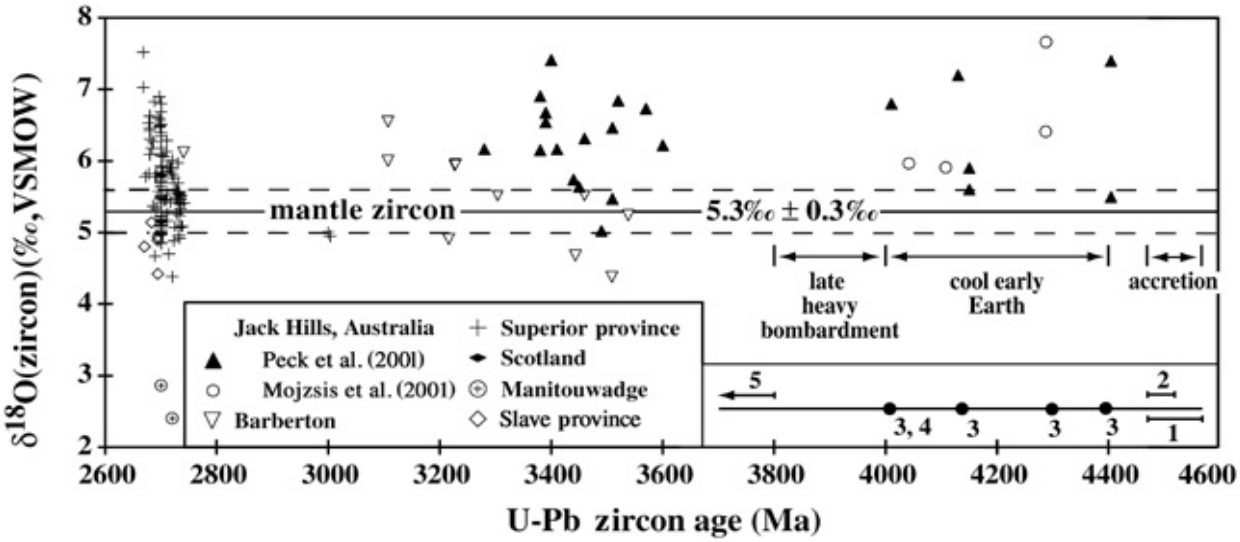
Evidence favoring some early non-mafic crustal rocks is found in the discovery of the mineral Zircon (a gemstone), $ZrSiO_4$, at various localities, mainly Australia, where non-igneous rocks have been found to contain this mineral as an accessory component brought in from weathered older crystalline rocks. Zircon only rarely forms in mafic rocks, and certainly not directly in any sedimentary type, but is a common indicator of silicic igneous rocks, principally Granite. The oldest individual zircon crystal dated so far (Uranium/Lead decay data obtained using an ion microprobe) is from younger rocks, including conglomerates deposited between 3.3 and 3.7 billion years ago, found in the Jack Hills in Western Australia. This narrow belt of folded and old Precambrian metasedimentary rocks looks like this from space:

This minute zircon's age is ~4.4 billion years, making it the oldest known solid material derived from terrestrial rocks of very ancient ages. If so, this and other zircons of somewhat younger ages presumably formed in sialic igneous rocks, probably granite, that were then weathered, released their zircon crystals, which survived and was incorporated in sedimentary rocks of varying ages much later (the youngest only 2 billion years old). The Australian zircons vary in age, with many being 4.0 to 4.3 billion years old; the one shown below is the very crystal whose age is 4.4 billion years.



Here is the outcrop from which this zircon was recovered:

In zircon-bearing Granites worldwide, water in the magma is almost always involved in the formation of this rock type. This association is one of two lines of evidence for water in and perhaps on early Earth crust. The other is even more diagnostic: the O^{18}/O^{16} ratio determined for these early Zircons. That ratio is as high as 7.4 for the oldest zircon yet found. The range of ages and their corresponding O^{18}/O^{16} ratios for dated Hadean zircons is expressed in this graph. As discussed next, this implies that its parent igneous rock may have formed in the presence of water.



Notice that the Jack Hills zircons and those from several other locations have high O^{18}/O^{16} ratios, commonly expressed as δO^{18}. (The Oxygen isotopes are analyzed using the ion microprobe along with an instrument called SHRIMP.)

Some have interpreted this as of great significance. High O^{18}/O^{16} ratios (above 5) are characteristic of water in several environments (processes selectively drive off the O^{16}, being lighter and thus relatively enrich the heavier O^{18}). The implication of this higher ratio is that liquid water was present in the environment in which the original (4.0+ b.y.) zircons crystallized. If this is ultimately proven, the conclusion is that the Earth's initial crust cooled rapidly (in less than 100 million years) and steam in the primitive atmosphere condensed to form bodies of water. That water may have been present mainly in hydrothermal solutions that affected the silicic igneous rocks (which would have formed as the end products of magmatic differentiation) containing the zircons. But a growing number of geoscientists take a larger leap and suggest that the water condensation at that time was large enough to form into bodies that would have been lakelike or even mini-oceanlike. Such waters became involved in igneous proceeses that led to granites. If that proves correct, then conditions might have existed for very primitive life to have started

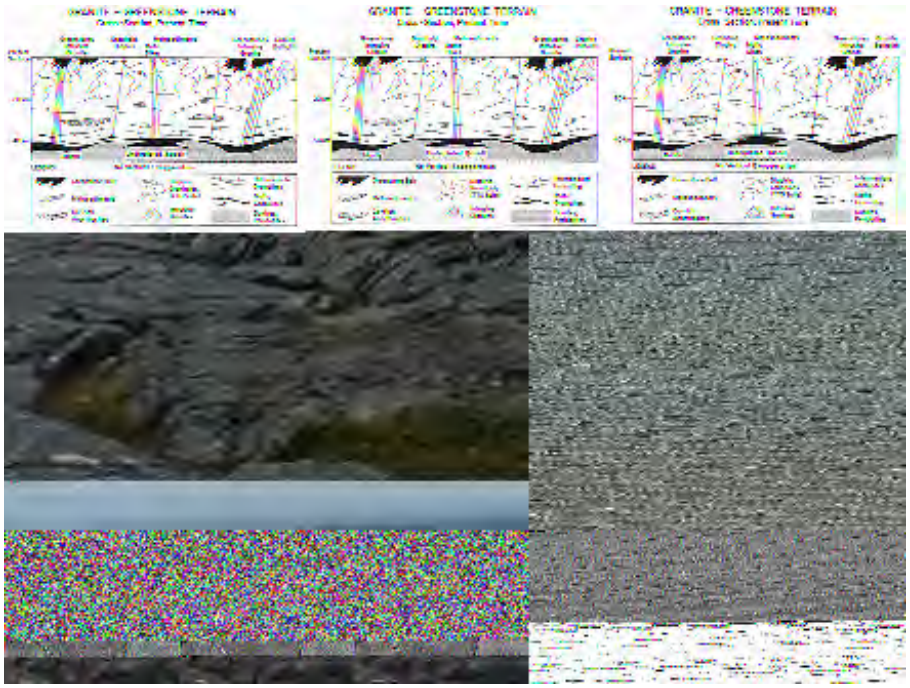
(see below) in surface/near surface settings (ponds to oceans) much earlier than now believed; but since no rocks from the early Hadean have been found, this remains a supposition.

Thus, the supposition from these zircon ages and their elevated δO^{18} values (as high as 7.4) is that sediments formed from erosion of the very early crust, collected in the protoseas, were buried, melted, and produced zircon-bearing sialic (silica-rich) igneous rocks including granite. If so, these rocks formed the nucleus of one or more ancient continents. They persisted above water long enough to erode over time while releasing the zircons of various ages that are now found in the sedimentary rocks of 3 to 3.5 b.y. age.

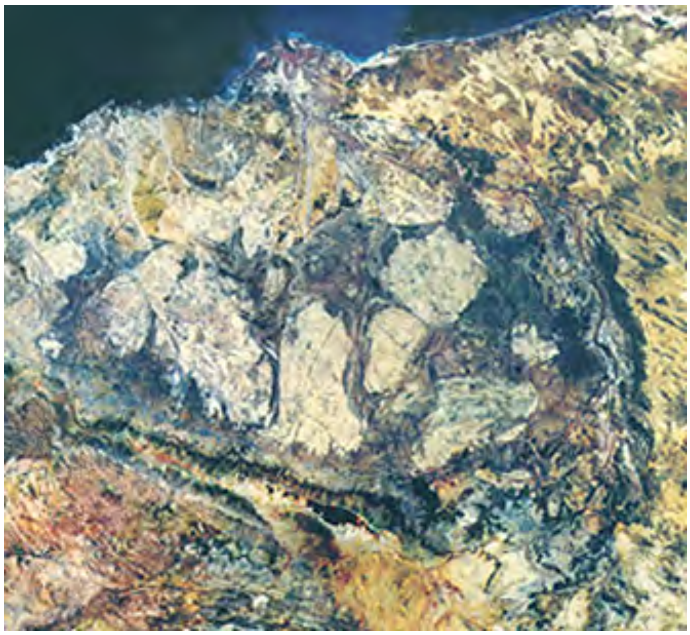
There are some surviving rocks that are about 3.8 to 3.95 billion years in age. These are metamorphic or metasedimentary. Nearly all rocks older than that - basalts, granites, probably also sedimentary, and metamorphic rocks - have seemngly all been destroyed (but geologists still seek representatives of such older rocks). Mechanisms for this destruction include bombardment, remelting, burial to now inaccessible depths, metamorphism (which resets the age-dating isotope clock), weathering/erosion, and possible subduction if the early crustal plates were set in motion (see below). A thorough review article, "A Cool Early Earth?", by John W. Valley, in the October 2005 issue of *Scientific American* summarizes the above ideas stemming from zircon analysis.

Canadian geologists from McGill University in Montreal claim to have found the oldest surviving rock formation on Earth. Located next to the eastern shore of Hudson Bay, it is greenstone, which is metamorphosed basalt. Its age is 4.28 billion years (some critics have challenged this result, citing uncertainties in the Nd/Sm age dating method). Archean greenstones are fairly common, as are Archean granites and metamorphic gneisses

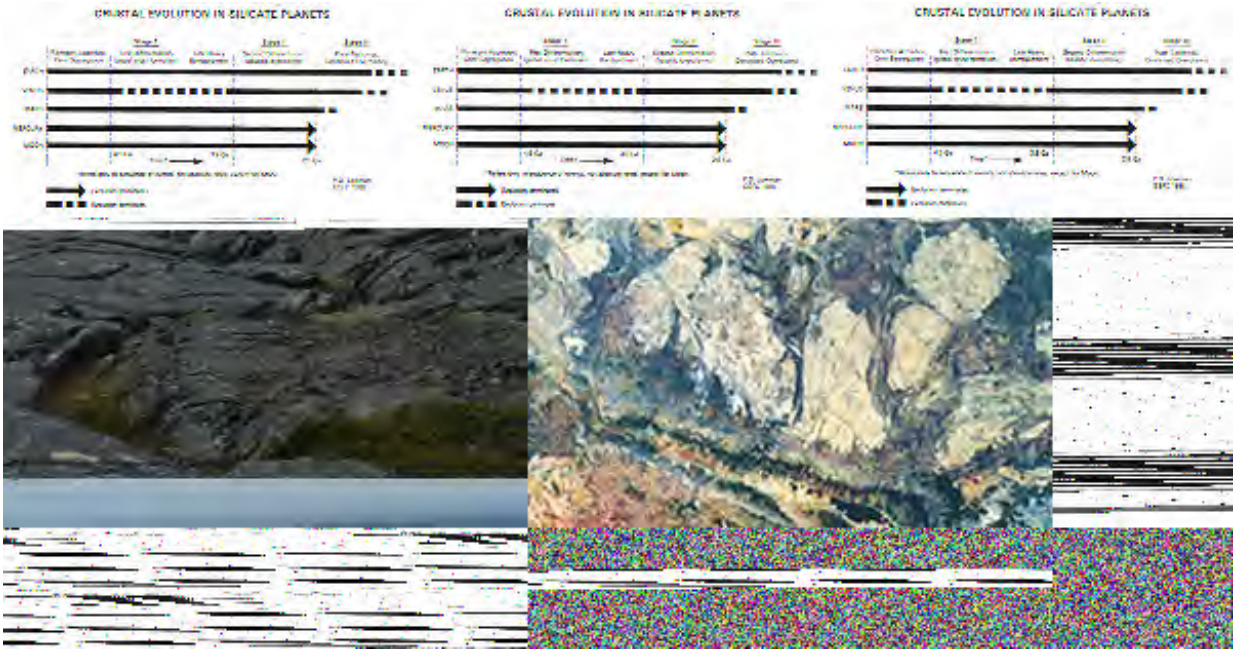
A general model for continental crust, which had widespread development throughout the Archean over much of the Earth outer layers, is indicated in this diagram produced by Dr. Paul D. Lowman, Jr (author of Section 12 of this Tutorial).



The closest actual example of this type of primitive crust is in the Pilbara district of northern Australia (shown on page 6-15). We put up here a somewhat different version of that scene which shows granitic plutons intruded into greenstones in a block surrounded by younger rocks:



Dr. Lowman has also developed a three stage model highlighting the commonalities among the four inner terrestrial planets; read the annotation in the illustration for contextual applicability to the Earth.



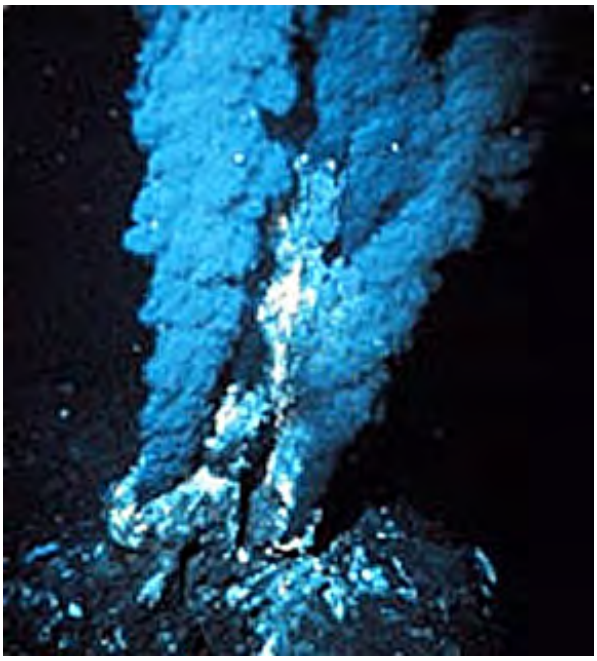
The early continents were built around basalts (ultimately, metamorphosed into greenstone) into which magmas, some differentiated into sialic granites, intruded. Sedimentary rocks formed on and around continental nuclei; these could be metamorphosed by burial and heat into a variety of metamorphic rocks including gneiss and even granites formed by melting of those rocks. Thus, some number of individual continental masses gradually increased in size; these, being lighter in density could rise above the primitive oceans where their rocks were weathered, transported, and converted into sediments that hardened and were themselves weathered or subjected to forces that led to metamorphic rocks (the Rock Cycle; see page 2-1a). A crust composed of heavier basalt within which were "continental islands" was probably in place sometime in the first billion years. That crust was subjected to circulating convection currents in the mantle of the hot Earth. This brought about fracturing of the crust, upwelling of melted mantle, and lateral movements of the crustal blocks dragged along by convection currents, i.e., the plate tectonics described on page 2-1a.

It seems probable then that plate tectonics started in the first billion year period. This figure below depicts a hypothetical early Earth distribution of protocontinents which drifted by seafloor

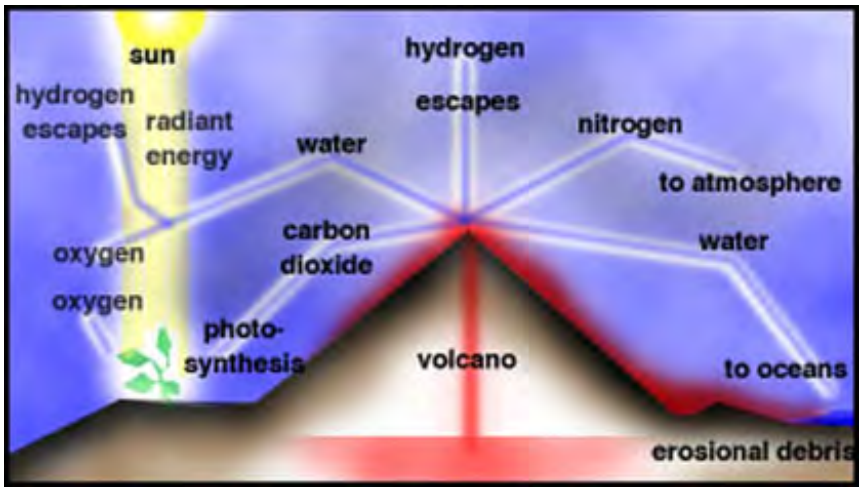
spreading until they collided into a supercontinent (Rondinia); that single supercontinent then broke up (like the later Pangaea) and its fragments made new individual continents - this process of coming together, splitting, moving in various directions over the global surface, recombining, resplitting, etc. probably has happened at least several times in Earth history.

Little has been said so far about the development of life on Earth. As mentioned above, oceans may have existed almost from the beginning but were definitely widespread by early Archean times. To appreciate the story of life's appearance and subsequent development, we need to talk briefly about the history of Earth's atmosphere.

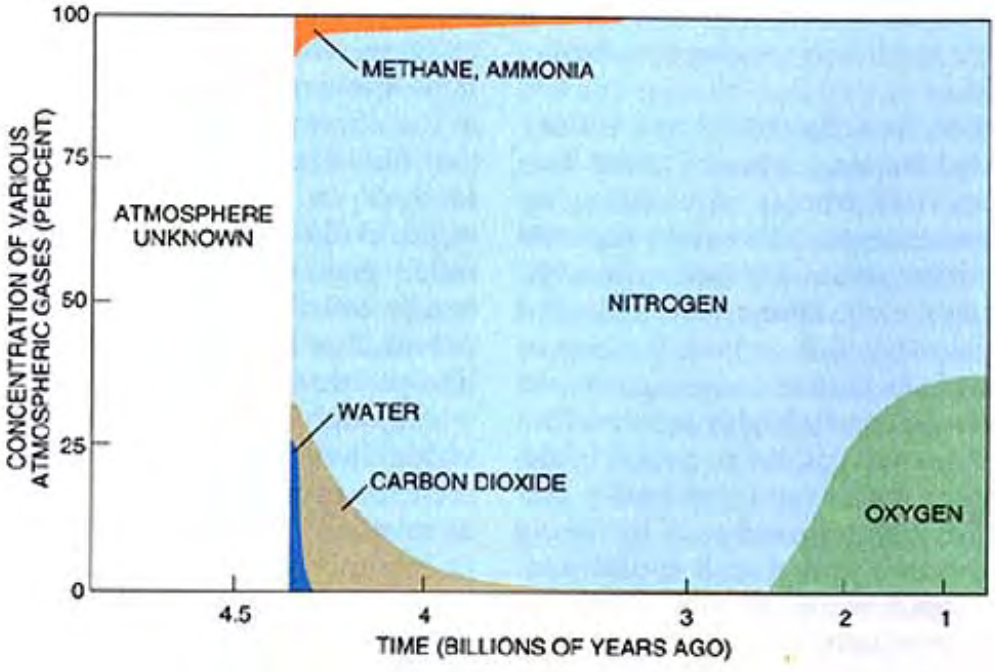
At the very outset, hydrogen and helium were the dominant gases in the forming Earth. As said above, these were largely blown away by solar winds. As the Earth melted, it degassed volatiles from its interior through the hot or molten surface. CO, CO_2, NH_3, NH_4, N_2, methane (CH_4) and water vapor were the likely principal gases that accumulated above the hot surface. Water vapor could have condensed to liquid but probably did not build up into an extensive early ocean, being more likely to revaporize as impacts continued. A dominant outlet was by volcanic vents, some being released into any water bodies, similar to the modern "black smokers" in today's oceans:



Most of the early Earth gases were released through volcanism, following a general pattern similar to modern conditions, except for the survival of any released oxygen:



The Earth's atmosphere for at least the first two billion years was very oxygen-poor and hence reducing. In time, N_2 became the dominant constituent of the atmospheric envelope that extended as a thick shell around the solid Earth. Methane and carbon dioxide persisted for some time. The carbon dioxide was utilized in part by organisms that developed photosynthesis capability.



Any oxygen released by volcanism was quickly combined with other elements, so as to have only a fleeting residence in an atmosphere. Starting about 4 billion years ago, much of this oxygen combined with iron that precipitated along with silicates to form the banded iron formation (BIF) rock that is a mainstay of sedimentary deposits until about 2 billion years ago.

This rock resulted from accumulations of ferrous Iron (Fe^{+2}) in oceans and lakes (which were more green in color than today; ferrous iron can produce that color as, for example, in a Coca-Cola glass bottle). The Iron readily combined with any available oxygen, so that the latter was always destined to be caught up in the iron precipitates (Fe_2O_3) and thus didn't remain in the atmosphere. While BIF is a hallmark of sedimentary rock formations during this extended period, other rocks also formed (shales; sandstones) but carbonates (limestones) were much less commmon. Starting about 2.3 billion years ago, oxygen levels and other factors led to common production of ferric oxides (Hematite) that made prominent red beds periodically to the present. One variety includes alternating chert layers, some rich in iron.

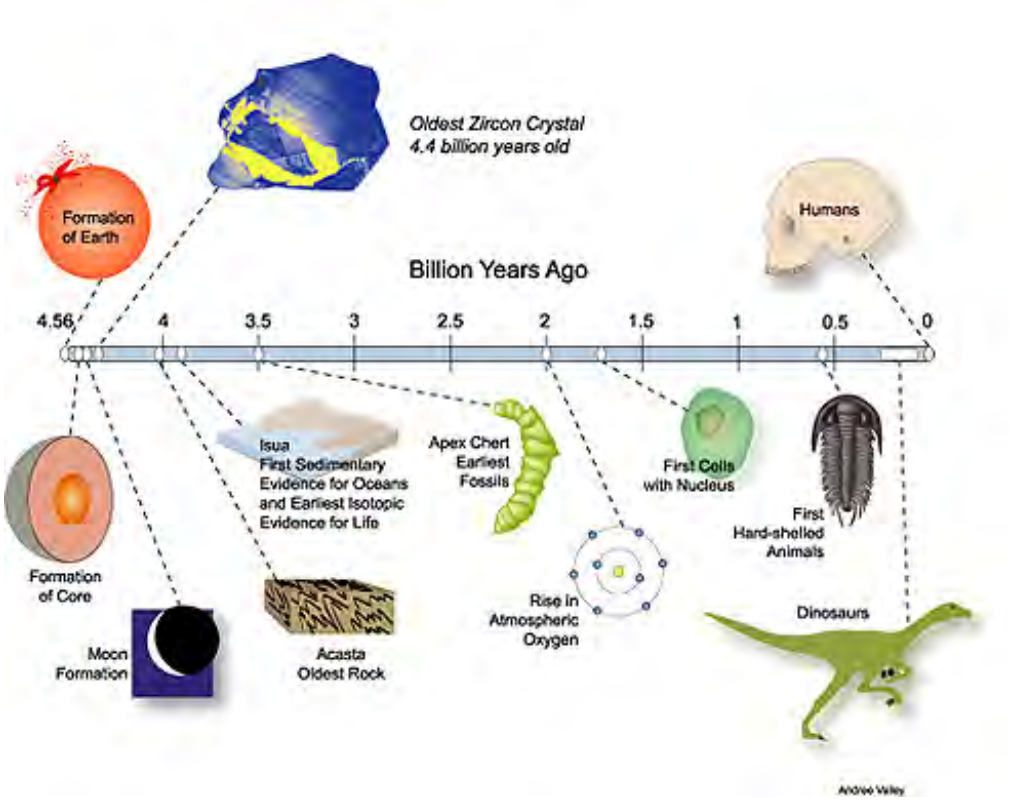
Not all BIF beds have the obvious red color from Hematite. This BIF outcrop shows alternating black (Magnetite?) and white (chert) layers (similar to the Taconite that the writer encountered in Minnesota while I was working (during a graduate summer school break) for Jones and Laughlin Steel as part of a field survey team:

As will be shown in Section 20, the origin of life is still uncertain but several models based on informed speculation have been proposed. In one (the Miller/Urey experiment), simple organic molecules were produced from atmospheric and dissolved constituents in the oceans that began to survive less than 4 billion years ago. Another cites organic molecules in carbonaceous meteorites. A third believes such molecules, and possibly even biologic organisms, were carried to Earth in comets (the "panspermia" model). Regardless, the first living organisms that have actually been observed are found in rocks formed about 3.55 billion years prior to the present. It is reasonable to surmise that rocks several hundred million years older hosted microorganisms but thermal processes have destroyed all traces; carbon isotope anomalies tend to support this postulate of life in these older rocks. They were microscopic **bacteria** in nature - these are numerically still the dominant form of life today. An abundant single-celled prokaryotic (without a nucleus) type is *cyanobacteria* (once misnamed "blue-green algae"). Colonies of trillions of these bacteria built up cabbage-like structures called stromatolites. The bulk of a stromatolite colony consists of layers of calcium carbonate interspersed with mattes deposited by the cyanobacteria. Stromatolites still exist on Earth but are rare (mainly at two localities in Australia).

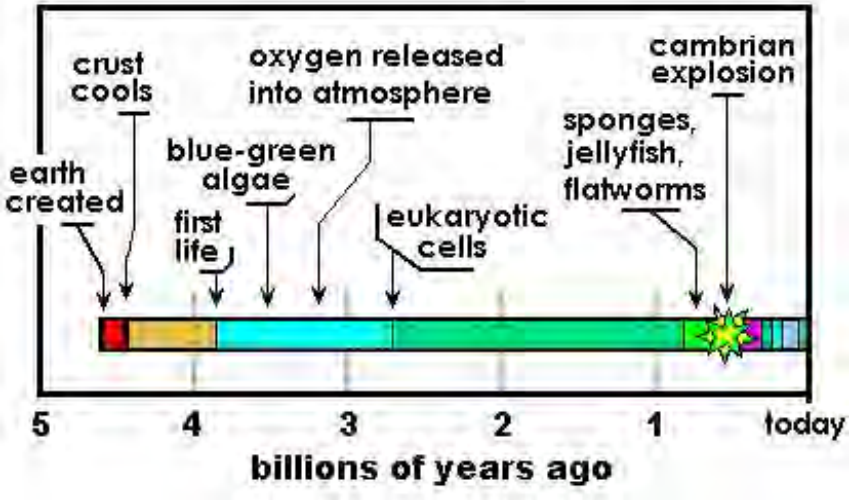
The oldest fossil utilizing hard parts consisting of a tiny spherical shell dates at about 3.1 b.y.:

Ancient rocks have disclosed other life forms including those that represent different types of bacteria, along with primitive true algae. For about two billion years after Earth's start, these bacteria and algae gradually developed the capability of receiving their metabolic energy through sunlight, which evolved into the process of photosynthesis. One of the by-products of photosynthesis is O_2. At first, this oxygen continued to combine with other elements but over time the accumulation of released oxygen exceeded the capacity to precipitate in the oceans and gradually built up a concentration in the atmosphere to its current 21% level. This had important consequences: O_3, or ozone, developed by solar UV reactions with normal oxygen in the upper atmosphere, providing a shield that greatly reduced the amount of UV reaching the Earth's surface, so that organisms sensitive to damaging ultraviolet rays could now begin to survive, flourish, and diversify. Multicelled (eucaryotic and metazoic) organisms appeared in the last two billion years as the oxygen component of the atmosphere increased. Respiration became possible as a metabolic source of energy in more advanced life forms. These became prominent in the fossil record about 750 million years before the present and underwent an explosive evolution about 560 million years ago (Cambrian time).

Many of the topics covered above can be summarized in this diagram that shows the main events and changes in the Earth's history:



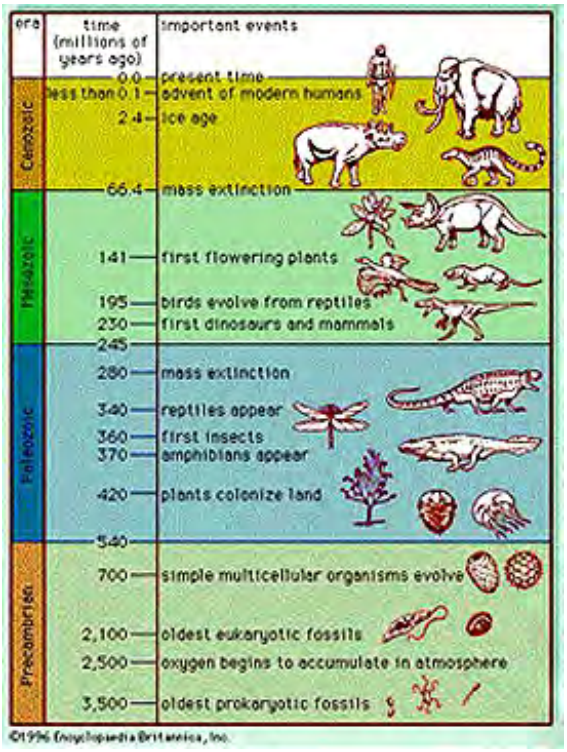
Two more illustrations expand on the ideas in the above figure. The first is a broadbrush overview of the Earth's history; the second uses the geologic time scale as a matrix in which key events in life's evolution are stated:



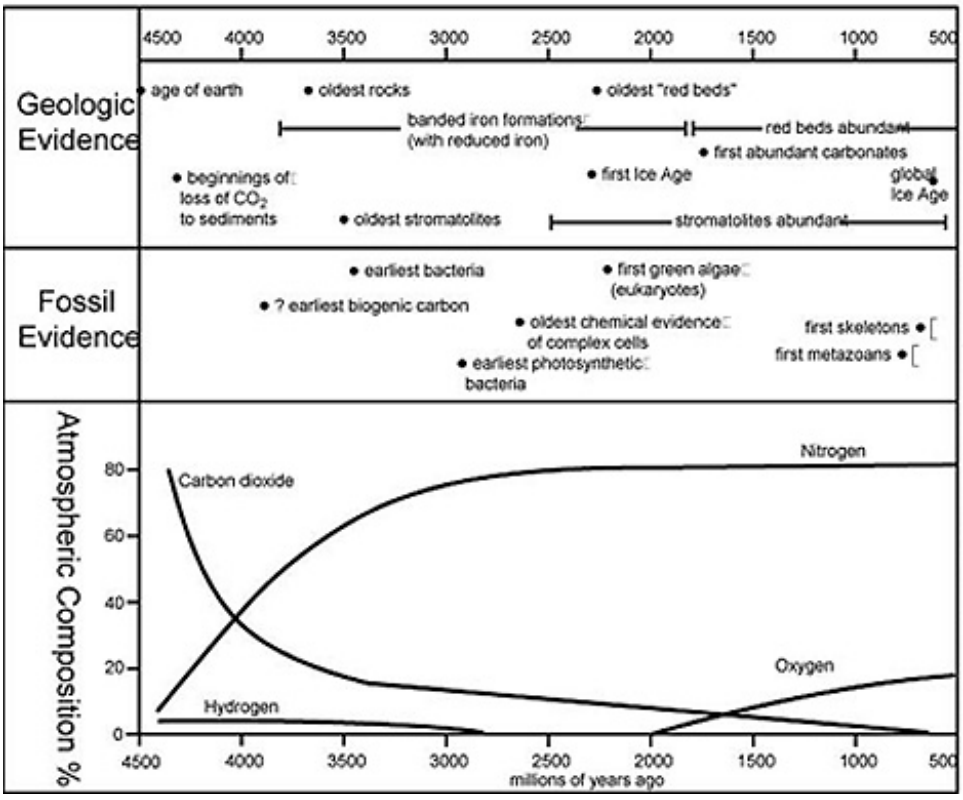
= mass extinction

Cenozoic era	Quaternary period Tertiary period	Diversification of mammals, evolution of humans
Mesozoic era	Cretaceous period	Continents mostly separated; mass extinction 70 m.y.a. kills off the dinosaurs
	Jurassic period	Dinosaurs and other reptiles dominate
	Triassic period	Continents begin to separate; diversification of reptiles
Paleozoic era	Permian period	Formation of the supercontinent, 'Pangaea'; mass extinction 250 m.y.a.
	Carboniferous period	Diversification of amphibians, early reptiles
	Devonian period	Diversification of bony fish and terrestrial plants; mass extinction 360 m.y.a.
	Silurian period	Early plants and insects colonize the land
	Ordovician Period	Mass extinction 500 m.y.a.
	Cambrian period	Diversification of marine life
Proterzoic era		Origin of Eukaryotes (1900 m.y.a.); multicellularity (640 m.y.a.)
Archaean era		Origin of life (3500 m.y.a.)

This next chart is similar both brings out more information on the first appearance of representative life forms.

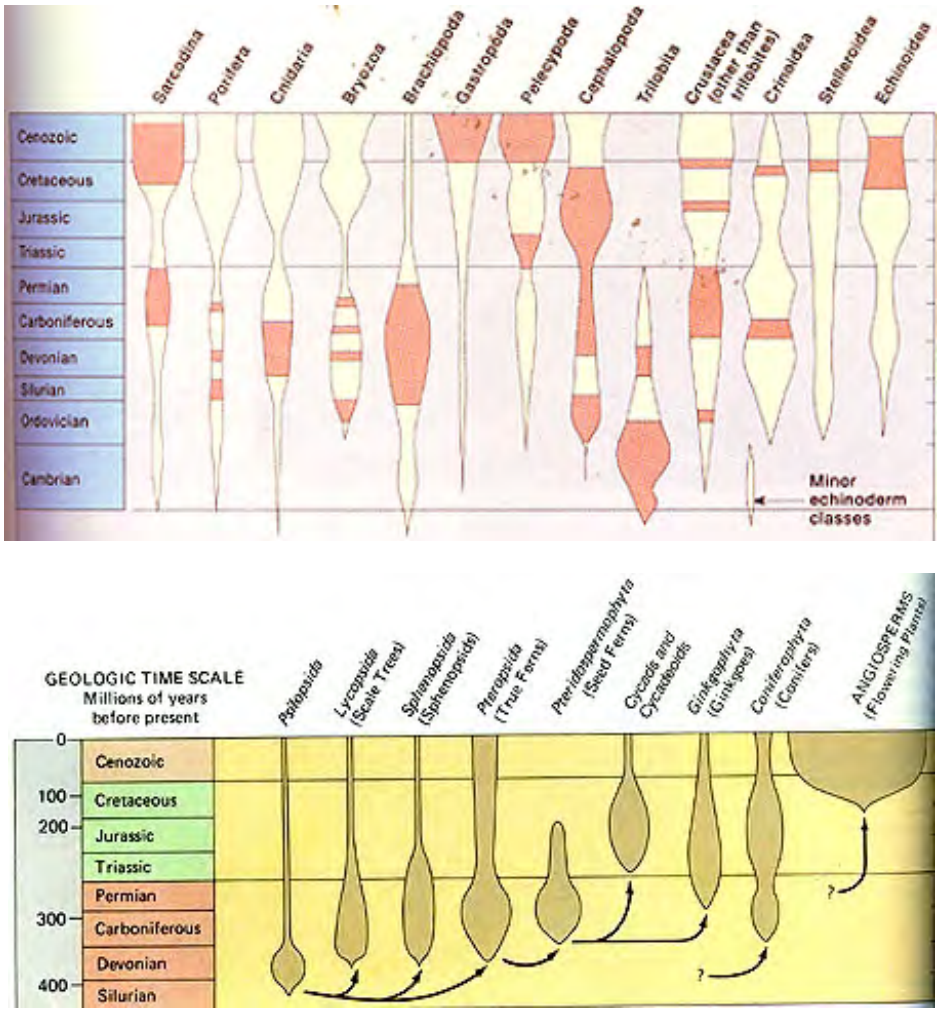


Although some information is redundant in this next chart which concentrates on Precambrian history, there are other new entries which may add to your understanding of this time span which covers nearly 90% of Earth time.



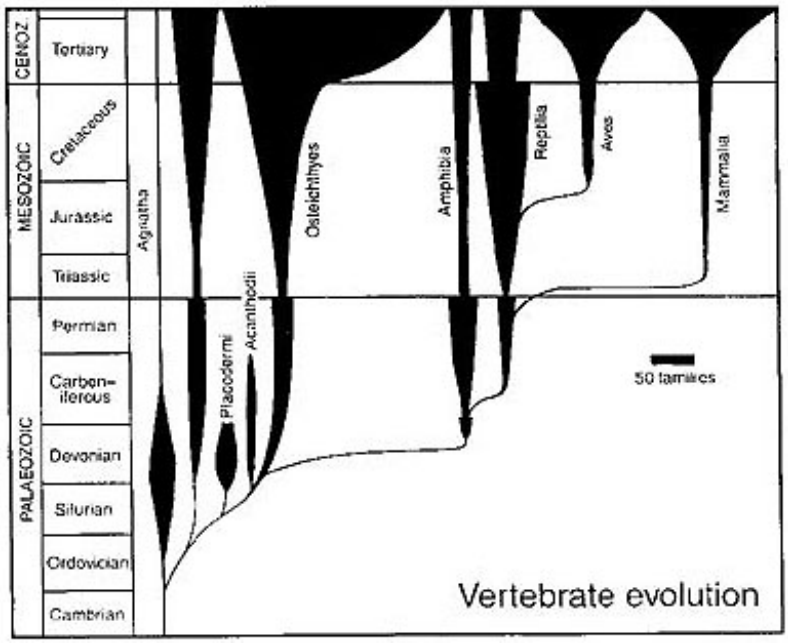
One of the benchmarks of Phanerozoic life occurred near its beginning with the great explosion of animal life forms in the Cambrian (Burgess shale fauna, mentioned on page 20-12). A preview of this expansion was the Ediacaran fauna (most creatures with hard parts) around 700 million years ago, as preserved in Australian rocks.

Invertebrate and Plant Life in the Phanerozoic (all time after the Precambrian) is summarized in these two diagrams:



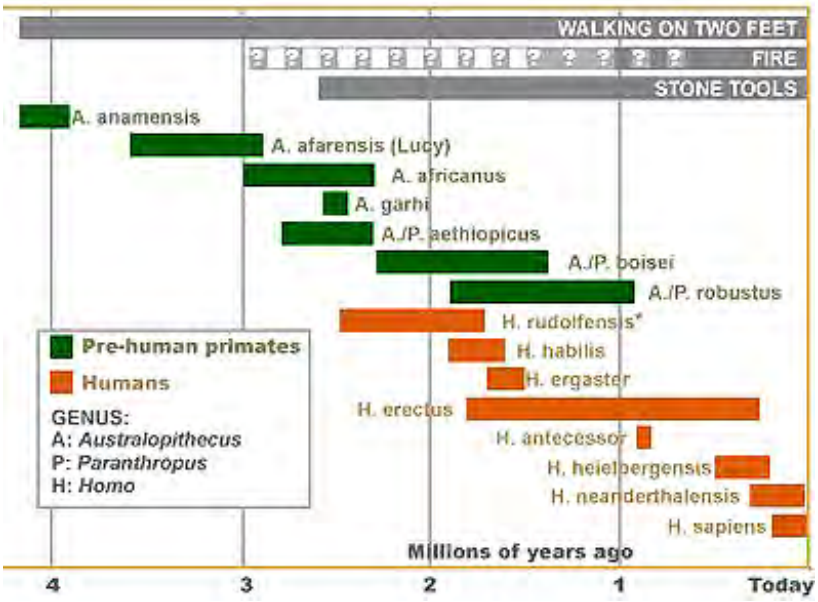
From H. Levin, The Earth through Time

Vertebrates can be traced back to the Ordovician. The first vertebrates were fish, followed by amphibians, reptiles, mammals, and birds. The evolutionary history of vertebrates is encapsuled in this diagram:



The culmination of Earth history has been the appearance of creatures with reasoning intellects - the humans. Although only illustrated for now, the following diagram summarizes the main stages in the evolution of the hominids (humans) over the past 5 million years (the Australopithecines are ancestral to the hominids). Much more about the development of life on

Earth and in the Solar System in general is covered on page 20-12.

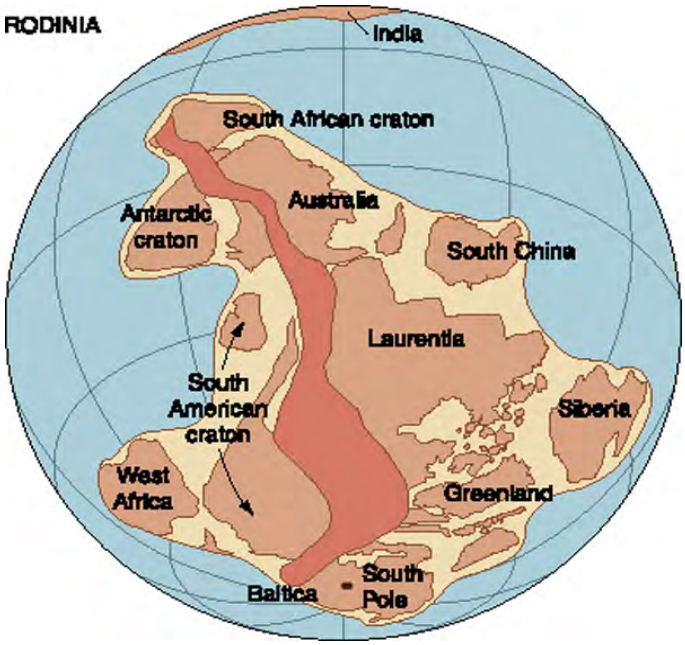


There are many Web sites that supplement and reenforce some of the ideas presented on this page. Consult these select ones if you wish: 1) Malaspina University Geology group, Canada; 2) UCLA Astrobiology Center; and 3)The Evolution of Life site sponsored by the University of Waikato in New Zealand..

Many of the features and processes that commonly operate on today's Earth, and in the past, also have now been observed on other planets and their satellites - this is embodied in the *comparative planetology* approach mentioned on page 20-1. That, again, is why Planetology has Terrestrial Geology as its foundation and principal source of methodology. We usually identify a feature on another planet by comparing its to similar looking and acting features on Earth. So, the majority of Planetologists also are, or started as, Geologists; the others have learned a lot of Geology to prep them as planetary specialists.

Let us illustrate the value of comparative planetology, using the Earth as the reference base, by examining the "Snowball Earth" hypothesis - which asserts that most or even all of Earth, including the oceans, was covered with a kilometer or more of ice around 750 million years ago. This surmise has moved to the forefront in the last two decades owing to ideas put forth by Paul Hoffman of Harvard University and others. The essence of the hypothesis is summarized in this diagram (Stages 3 and 4 - the aftermath of maximum glaciation - have been omitted).

The evidence for this worldwide glaciation and ice cover is mainly the distribution of glacial tills and other deposits in most of today's continents, shown here as they were regrouped in the supercontinent Rodinia, at the time of the Grenville orogeny 1.1 billion years ago:



Associated Banded Iron Formation deposits and unusual carbonate deposits further support the Snowball model. But, the hypothesis is still not broadly accepted, as valid objections persist. You can read more about this subject at the Snowball Earth website produced by Wikipedia.

If indeed Earth was largely covered by ice then (and perhaps earlier, about 2.5 billion years ago), then it resembles (on a larger scale) many of the satellites of Jupiter, Saturn, and Uranus that will be examined later in this Section.

With this history of Earth as a planet, we can now turn our attention to the nearest large planetary body, our own Moon, which provides an excellent example of how scientists go about obtaining information about other solar objects and interpreting the nature and characteristics of planets.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

On this page, links to other relevant planetary Websites are emplaced. A pitch is made to the reader to consider taking a look at the extensive review of Astronomy/Cosmology in Section 20 as a background to this Section - optionally switching to it here before moving on to the planets. The definition and nature of a "planet" is then considered. A table lists the major facts and parameters pertaining to the solar planets. Some of the characteristics of the motions and distribution of the planets are described in terms of the historical contributions by Copernicus, Brahe, Kepler, and Newton. Finally, a subsection is presented on how meteorites are used to determine compositions of solid planetary, asteroid, and cometary bodies in the Solar System.

INTRODUCTION TO THE PLANETS

Internet Links to Planetary Sites; Book References

Planetary exploration has become one of incredible and vast accomplishments, in which huge amounts of data have now accumulated. Much of it has relied exclusively or largely on remote sensing. This Section is one of the longest in the 27 units of the Tutorial. The intention is to provide a thumbnail view of the major missions to the planets. Despite the importance of learning about planetary atmospheres, we will not say much about the results of remote sensing of these gaseous envelopes nor do we discuss in any detail facts and conjectures about planetary interiors.

There are many sources of additional images and descriptive information. Among the best of these currently online is a repeat of Chapter 5: *Planetary Geology*, by James Bell III, Bruce Campbell, and Mark Robinson, in the 3rd Edition of the *Manual of Remote Sensing: Earth Sciences Volume*, 1996, at Marswatch. This lengthy and detailed review focuses on remote sensing approaches to planetary exploration. Its one drawback is a sparsity of images (compared with this Section 19 Overview). An excellent chronological survey of the history of space exploration is found at the Planetscapes website. A complete listing of all planetary missions can be accessed on this NASA Goddard web site. Another site worth visiting is the Home Page of the Jet Propulsion Laboratory (JPL) where you can get addresses to visit other sites dealing with terrestrial and planetary space programs. Another NASA source is the National Space Science Data Center NSSDC. Two other exceptional Home Pages are The Nine Planets, by Bill Arnett of the Lunar and Planetary Laboratory, University of Arizona (LPL) and Views of the Solar System, by C.J..Hamilton of the Los Alamos National Laboratory (Spaceart). Dr. J. Schombert of the University of Oregon offers three courses on Planets, Astronomy, and Cosmology that he has put on the Web; the first of these - The Solar System - is accessed at his AST121 site. The NASA Headquarters Space Sciences Directorate maintains an excellent Site that summarizes the major findings in both planetary and cosmological realms during the latest 9 to 12 months that can be accessed at Space Science (see its lists, especially News). Many Solar System missions were managed and conducted by the Jet Propulsion Laboratory; descriptions of Past, Current, and Future missions are obtaining by clicking on any of interest at this Missions site.

Books that treat planetary remote sensing as part of a larger review of Planetology include a now out-of-print text by this Tutorial's author (Nicholas M. Short), *Planetary Geology*, 1975, Prentice-Hall Publ., still in libraries, Murray, Malin, and Greeley's *Earthlike Planets*, W.H. Freeman & Co., 1981, and Billy P. Glass's *Introduction to Planetary Geology*, 1982, Cambridge University, Press. More recent are *Planetary Landscapes* by R. Greeley, 1985, Allen & Unwin, *The Planetary*

System, by D. Morrison & T. Owen, Addison-Wesley Publ., 1988, and *Exploring the Planets* by W.K. Hamblin and E.H. Christiansen, MacMillan, 1990.

The writer *strongly* recommends ***Lonely Planets: The Natural Philosophy of Alien Life***, by astrobiologist David Grinspoon, Harper Collins Publ. published in March, 2003 that covers much of what we knew by that time about planets in the Solar System. Its main purpose, however, is to review the conditions for and likelihood of life (from very primitive to advanced thinking creatures) in our galaxy and beyond - throughout the Universe. Although a bit wordy and redundant, this book explores almost all facets of whether Earth is unique (he thinks not!!) and under what circumstances planets containing living organisms (or once living, now extinct) can develop.

Scientists who have spent at least part of their careers studying the planetary bodies of the Solar System are called *Planetologists*. The majority of these are also *Geologists*, although some are instead *Astronomers* and *Physicists*.

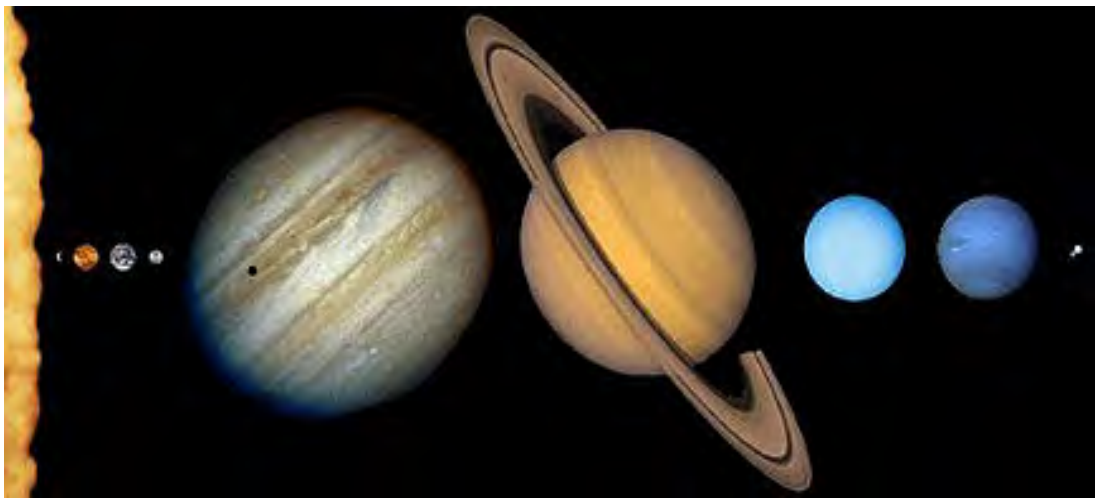
Before we start our tour of the planets, you may wish to review some of the main principles and concepts of Astronomy. If so, please skip to Section 20, which is a comprehensive review of this subject as it is subsumed into the closely related field of Cosmology.

The Nature of a Planet

We concentrate in Section 19 almost entirely on the planetary bodies of the Solar System. Other planetary systems, around different stars, have been discovered discovered in recent years, as described on page 20-11.

The central body controlling these planetary bodies is the Sun. (For background information on the Sun, check Sol and/or Sun)). Despite the large size of some of the solar planets relative, say, to Earth, these planetary bodies orbiting the Sun taken together, plus also asteroids and comets, contain only 0.14% of the mass of the Sun (99.86%). They do, however, contain most of the angular momentum of all bodies (including the Sun) in the Solar System.

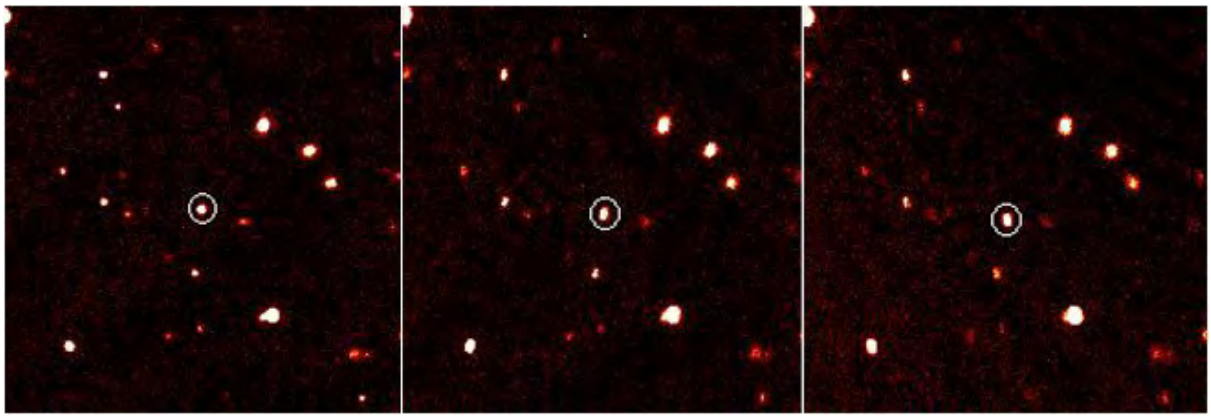
To set a framework for our survey of the Solar System's inhabitants (exclusive of the Sun - that star is treated at the top of page 20-5a), which we will consider as the "Planetary System" despite the recent discovery of more than 200 extrasolar planets (again, Section 20), look first at the illustration below, which shows the relative sizes of the nine planets, with a small segment of the Sun shown to the far left, so that its size relative to the planets is scaled (the distances between them are not actual - each planet is just placed at the same distance next to its neighbors)(this diagram dates before the 2006 demotion of Pluto within our Solar System).



19-1: Using their appearance, how many of the above planets can you name? Check the answer to see a diagram that shows the size of each named planet relative to the Sun. ANSWER

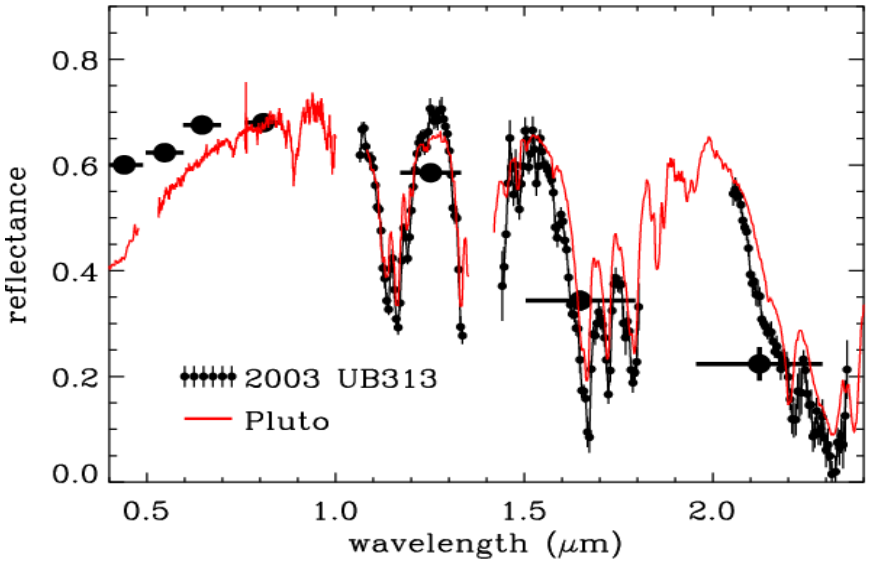
In late 2003, announcement was made by a group of astronomers of discovery of what they claim to be a possible Solar System planet, which they named Sedna, that lies astride Neptune's orbital path. This is an isolated small body (about 1250 to 1800 km; 750 - 1100 miles in diameter) that is spherical. Such a shape is suggestive of melting and reorganization into a round mass. According to the American Astronomical Society rules, this size is below the lower limit agreed to be the smallest a body can be to be named a planet. It appears to be a new class - between irregular asteroids, which can be larger, and the solar planet sizes; some rounded "moons" are in this size range which could explain it as an escaped Neptunian satellite but a mechanism to remove it from Neptune's orbital family into its own solar orbit is yet to be proposed.

Actually, another planet had been imaged in 2003 but astronomers at Cal Tech did not recognize it as such until 2006 when three images taken 90 minutes apart were inspected and analyzed. This is what they saw:



The circled bright object in the left image seems to be one body. But in the second image, two bodies are resolved. The upper body moves again in the third image, taken 90 minutes later. The lower body proved to be a distant star. The upper body was shown to be within our Solar System. It was named 2003 UB313. After its acceptance as a planetary object, it was given the name Eris. At the 2003 sighting time, Eris lay a distance of 97 Astronomical Units. (*An Astronomical Unit [A.U.] is defined as the distance between Earth's center and the Sun's center, approximately 151 million kilometers or 93,000,000 miles.*) Continuing observations show it to have a highly elliptical orbit so that its closest approach to the Sun will place it at a distance of 38 A.U. This orbit is inclined 44° from the ecliptic. Eris will take 560 years to complete one revolution around the Sun. Eris is one of several larger spherical objects that are within the Kuiper belt of asteroids. Most of those are irregular in shape. So, is the sphericity of Eris enough to qualify it as a planet, along with Sedna and others.

As a momentary digression, this plot is an example of how remote sensing - in the form of gathering reflectances as a function of wavelength - is used in planetary studies. The plot shows the spectral reflectance curve for Pluto and for Eris. Both are nearly identical, suggesting that Eris has a surface composition similar to Pluto.

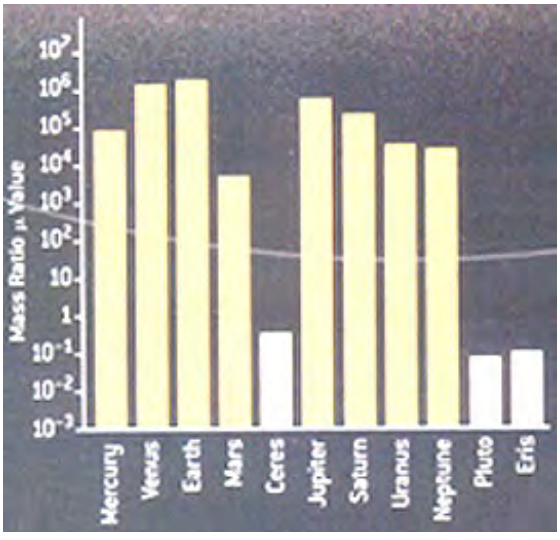


The Sedna and Eris discoveries along with detection of other similar bodies have reopened one of the great debates in Solar Planetology: Is Pluto really a planet? Its size is the smallest (2288 km diameter) of the 9 traditional ones stated in textbooks. Purests consider it not worth of planet status; some believe it, like Sedna, is an escaped moon. The controversy was the centerpiece of a debate at the 2006 meeting of the International Astronomical Union (IAU) in Prague. A committee there produced this simplified definition of a planet (applicable to any planetary system, not just that of the Solar System):

A planet is a celestial body that (1) is in orbit around a star but is neither a star nor a satellite of a planet; (2) has sufficient mass for its self-gravity to overcome rigid body forces so that it assumes (usually after melting) a hydrostatic equilibrium (nearly round) shape, and (3) has swept up most of the smaller solid bodies (e.g., asteroidal "chunks") in the spatial region around its orbit.

Item 3 is the new criterion that is to be used in defining a true planet. A key word is "most". Both Earth and Mars have some asteroidal-sized rock chunks within their orbital zones but these bodies are in toto only a very small fraction of the mass of the parent planets. Pluto has much more such bodies still in its orbital zone, so it has failed this third criterion.

Item 3 leads to another parameter that clearly separates the Main Planets from smaller bodies. This is done by calculating the ratio of the mass of the parent planet to the mass of all small body materials (excepting moons) in the orbital zone of the planet; this ratio is represented by the Greek letter μ. This plots as follows:



The mass ratios show that the 8 Solar System planets all have μ values greater than about 7000; the three spherical bodies Pluto, Ceres, and Eris all have values less 1. On this basis, the reason the IAU now contends that there are only eight major planets is obvious.

This reclassification might seem straightforward, but exceptions may cause confusion, e.g., Pluto's moon Charon. But Charon lies beyond the barycenter for the Pluto-Charon pairing. The barycenter is the center of gravity for a two body system that has one body orbiting the other. The Moon remains a satellite despite its size because of the Earth-Moon barycenter still residing within the larger Earth. In time (billions of years), as the Moon recedes the barycenter will migrate outward beyond the Earth, so that, technically, the Moon would become a planet in the sense of Charon. The large satellites of Jupiter, Saturn, Uranus, and Neptune all associate with barycenters located within their parent planet.

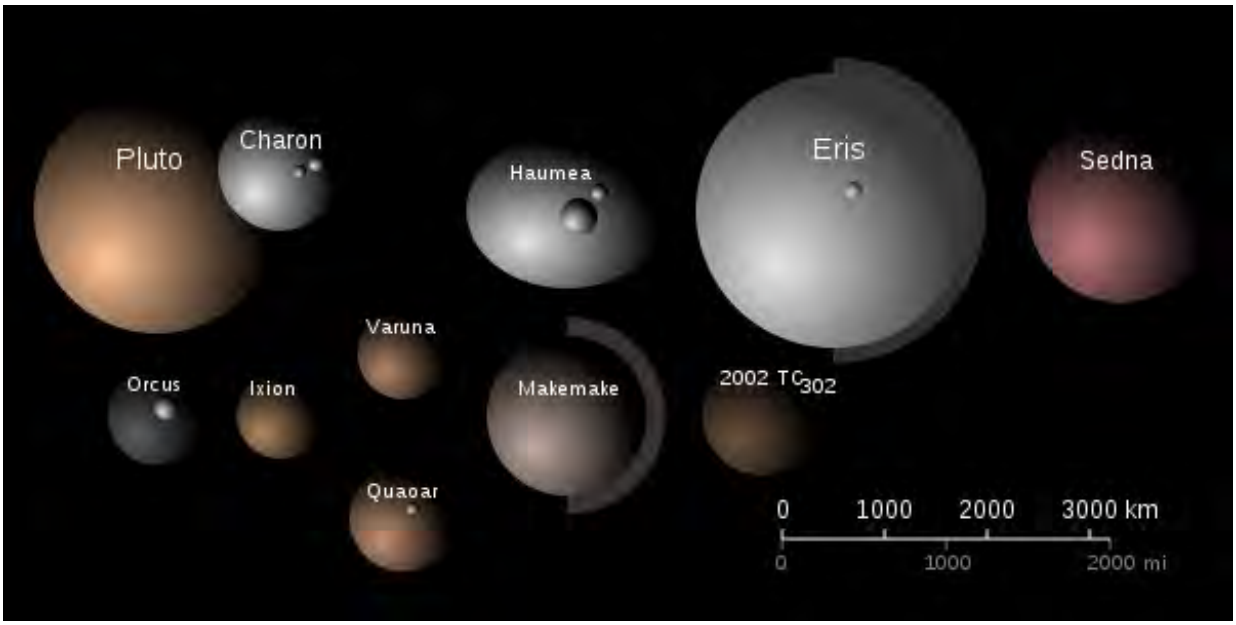
So, what did the IAU decide: They first mulled over the "strawman" concept that as of 2006 there might be 12 planets in the Solar System, with the three new ones (2003 UB313 is now named Eris) shown in this diagram. They firmly decided at this meeting against these three as 'full blown' planets, keeping them as large members of the Kuiper Asteroid Belt.

The latest thinking is that there should be three categories: 1) planets; 2) dwarf planets; 3) solar system objects. Dwarf planets must not be satellites of larger planets. Dwarf planets must be round, whereas solar system objects must be irregular in shape. Dwarf planets that are icy can be as small as 200 km in diameter; if they are rocky, the smallest size that is round is about 400 km. As of 2010 about 50 round dwarfs have been found in the Solar System.

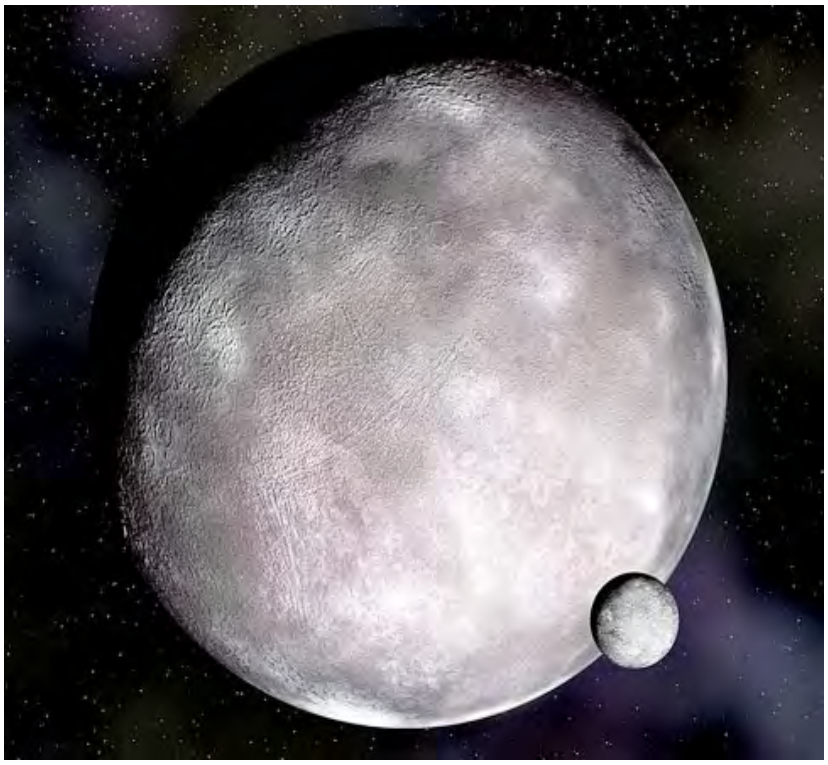
A number of small round bodies have now been found in the Kuiper (asteroidal) Belt and possibly the Oort (cometary) Cloud (page 19-22). They are part of a vast collection of large to small objects - most non-spherical - orbiting the Sun. According to the IAU definition, some smaller round ones may be eventually raised to the dwarf planet level. Here are some current candidates:

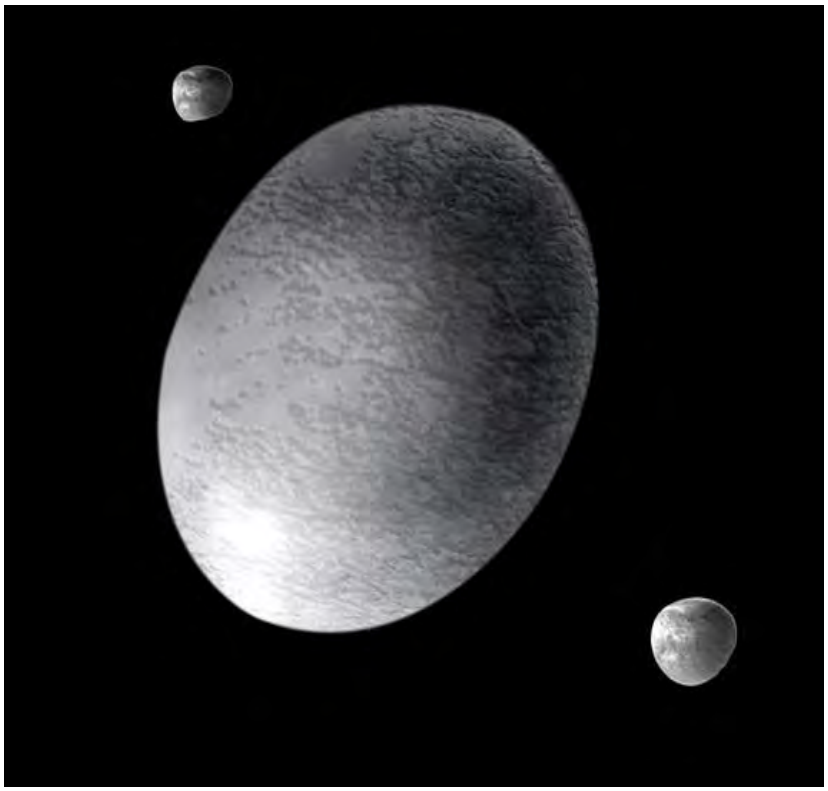
The IAU voted on August 24, 2006 to further redefine planets by size - these falling into Main (large) and Dwarf (small) categories. This enabled Pluto to retain some planet status by becoming a Dwarf Planet along with Ceres and Xena. Pluto's moon Charon remained just that - a satellite. But as in any democratic procedure, those on the losing side will continue to disagree.

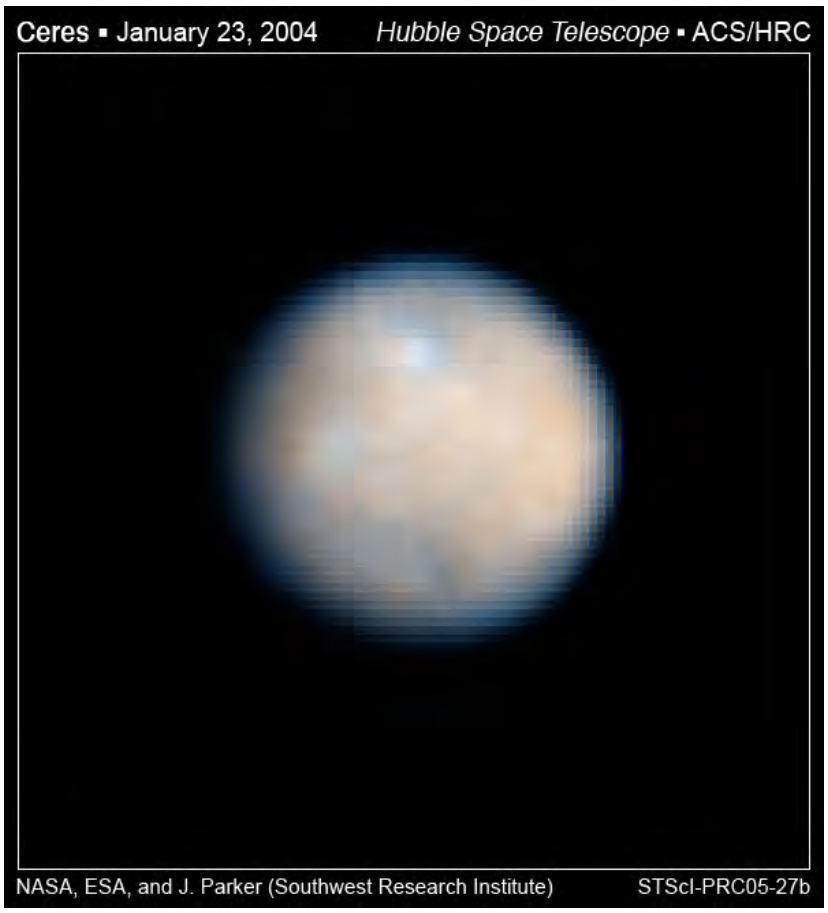
This diagram depicts some of the Dwarf planets in terms of their relative sizes:

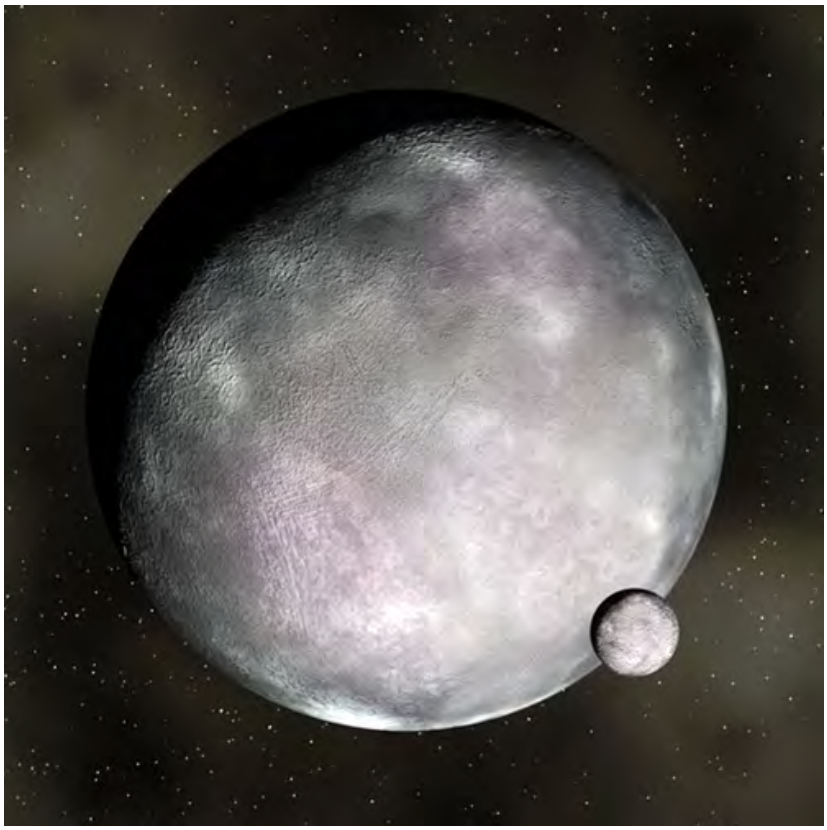


Since the Dwarf planets are small, good photo images of them are not easily acquired. Most are depicted from observations that are translated into "artist's concept", visual representations of the 'best guess' of their appearance. Below are some of these planets - read their captions to determine if each one is an actual image or an artist's concept:






Ceres ▪ January 23, 2004
Hubble Space Telescope ▪ ACS/HRC
NASA, ESA, and J. Parker (Southwest Research Institute)
STScI-PRC05-27b



The question "What is a Planet" has been addressed in the January, 2007 Scientific American article by that name as written by Steven Soter.

There seems little doubt that Earth is still the premier planet among those now discovered (within and beyond the Solar System). It is (so far) unique in having life (but not organic molecules) as a main characteristic. In the Solar System, Earth is alone in having 1) widespread water oceans, and 2) abundant vegetation. It is the presence of trees, grasses, and other vegetative cover that distinguishes Earth from the other solar planets; but from the standpoint of remote sensing this absence on the other planets makes it possible to examine them in terms of the geologic processes that led to their development (on Earth, extensive vegetation usually masks the underlying geology, as is typical in the eastern United States).

This subsection ends with the illustration below, which shows the full disc appearance of each of the four Inner or Terrestrial Planets scaled to their actual relative sizes, of which Earth is the most important (or you wouldn't be reading this):

Some Planetary Parameters

Consult the table below, which summarizes the principal characteristics and properties of the nine planets. We list them from top to bottom in the same sequence as those shown from left to right in the above illustration. To simplify, we do not include the names of the principal satellites orbiting some of these planets, but we cover them in a listing below the table.

PLANETARY BODY	DISTANCE FROM SUN (AU)	ORBITAL PERIOD (yrs)	ROTATIONAL PERIOD (days)	DIAMETER (km)	DENSITY $(gm/cm)^3$	NUMBER OF SATELLITES
Mercury	0.387	0.24	58.6	4,880	5.44	0
Venus	0.723	0.62	243R	12,105	5.25	0
Earth	1.000	1.00	1.00	12,757	5.52	1
Mars	1.524	1.88	1.03	6,786	3.93	2
Jupiter	5.203	11.86	0.41	143,797	1.34	8 R; 55 IR
Saturn	9.539	29.46	0.43	120,659	0.70	21 R; 26 IR
Uranus	19.18	84.01	0.72	51,121	1.28	18 R; 9 IR
Neptune	30.07	164.80	0.73	49,560	1.64	6 R; 7 IR
Pluto*	39.44	247.68	6.4	2,288	2.06	1 R; 2 IR

* Included in the Table despite its reclassification as a Dwarf Planet.

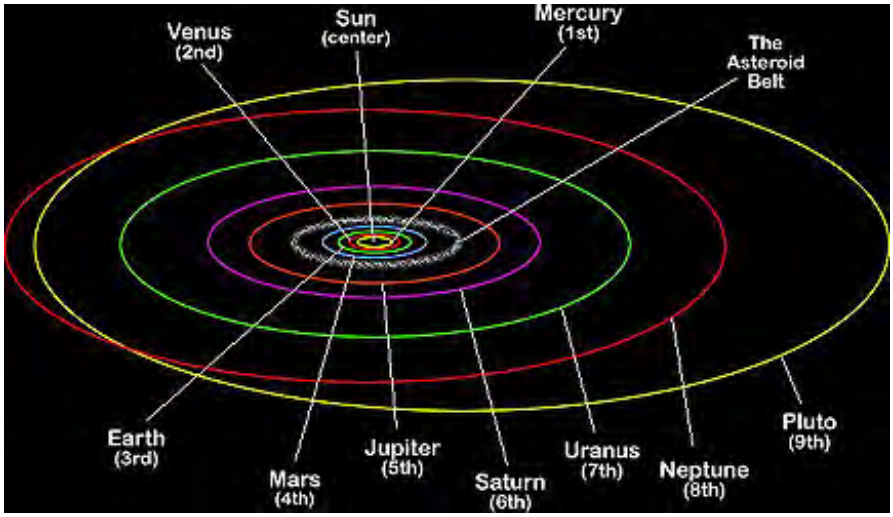
Note: For the number of satellites; that numeral left of R refers to those satellite that are nearly spherical - that left of IR refers to irregular shaped satellites (see page 19-14).

AU = Astronomical Unit, which is the mean distance (approx. 150 million kilometers, or 93 million miles) from the Sun to Earth

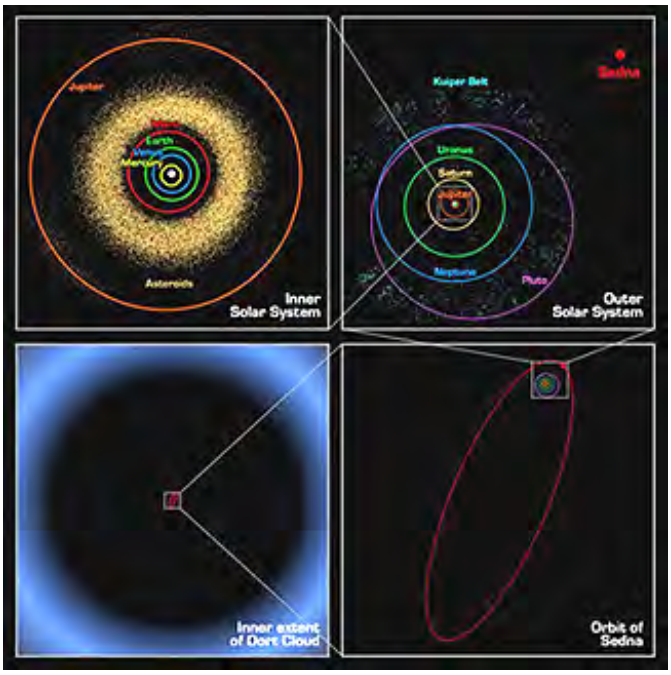
Names of principal satellites (smaller ones omitted):

- Earth: Moon
- Mars: Deimos; Phobos
- Jupiter: Io; Europa; Ganymede; Callisto
- Saturn: Mimas; Enceladus; Tethys; Dione; Rhea; Titan; Hyperion; Iapetus; Phoebe
- Uranus: Miranda; Ariel; Umbriel; Titania; Oberon
- Neptune: Triton; Nereid; 1889N1
- Pluto: Charon

The above table lists the distances of the planets from the Sun. This diagram shows these distances in terms of the orbits of the planets:

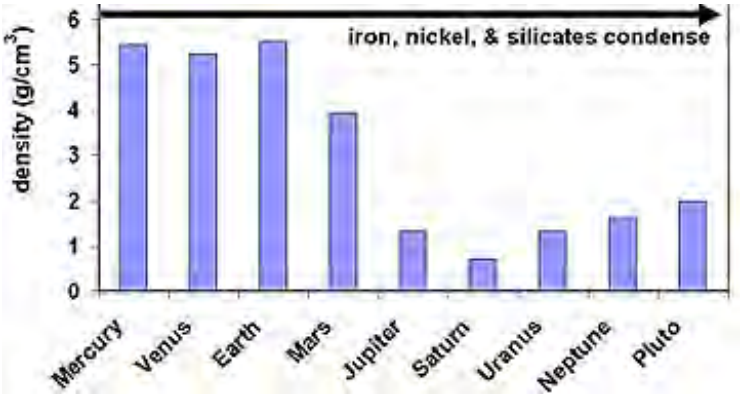


As will be described later, beyond Neptune and Pluto are Sedna and the Kuiper Belt (asteroidal bodies) and Oort Cloud (mostly comets). This diagram indicates the extreme distances of these latter features in relation to the nine planets:

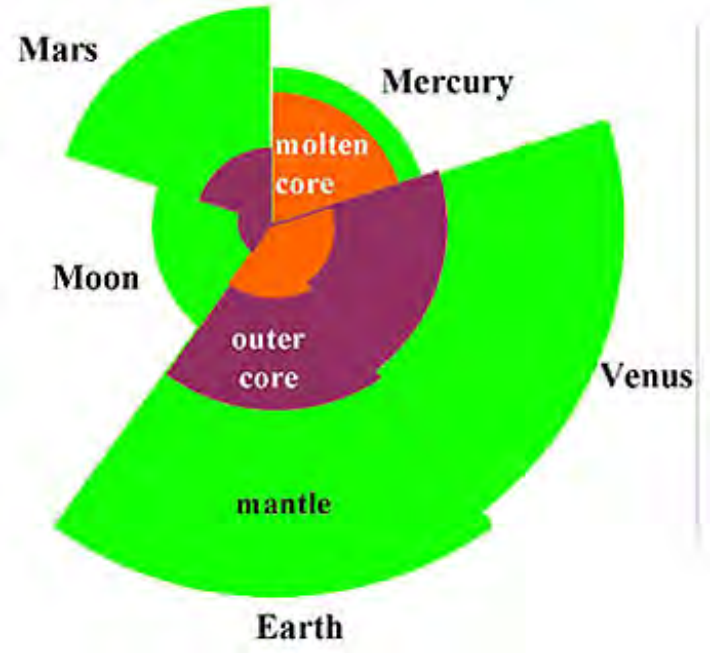


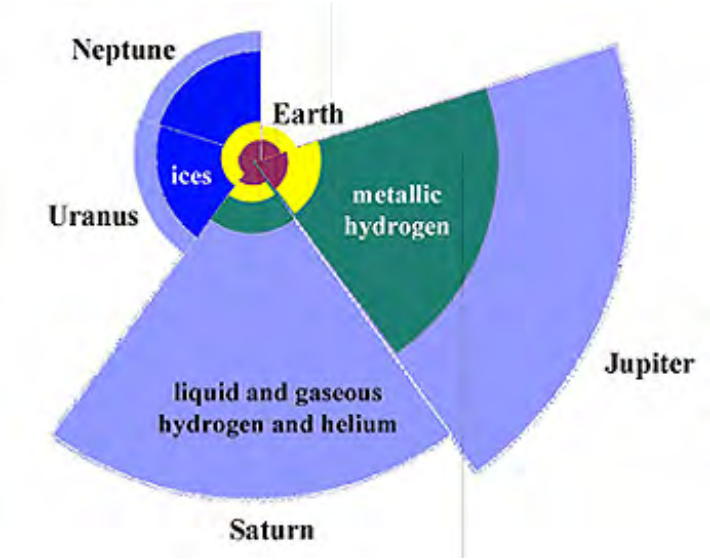
(A brief mention is made here of an interesting hypothesis maded by Dr. Harold Levinson of the Southwestern Research Institute: He postulates that in the early Solar System the planets had orbits different from the present. The Giant Planets Jupiter and Saturn had orbits such that on very infrequent occasions they would approach each other so that their massive gravitational interactions caused them to start a process called resonance. This eventually forced them to readjust their positions and also hurled Uranus and Neptune further out. About 3.9 billion years ago, the resonance also perturbed the asteroid belt objects, forcing many of these towards the Inner planets. The Moon, formed by then, underwent a cataclysmic bombardment [see subsection on the Moon] during which most of its craters were formed [extensive cratering also took place on Earth but almost all of these have been obliterated by subsequent erosion]).

The densities of the planets is one characteristic parameter. Here is a histogram that compares this property:

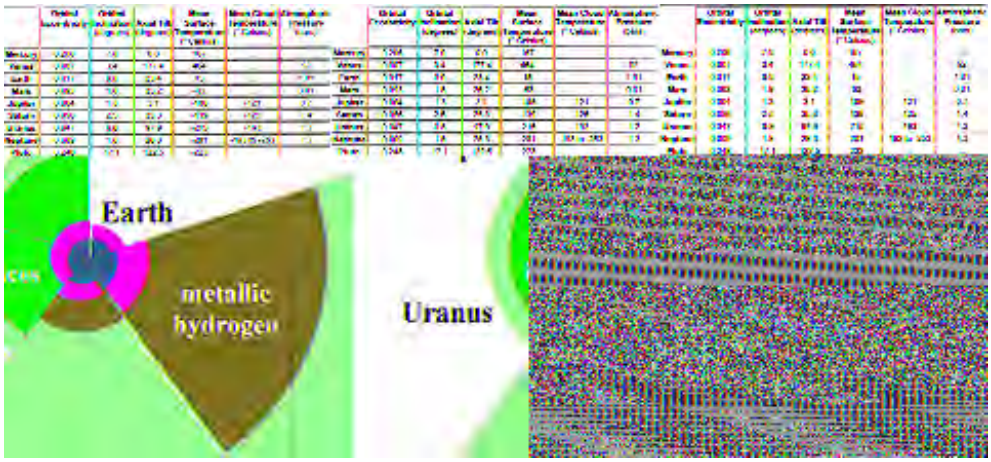


The Inner planets all show rocky materials at their surfaces (all but Mercury have atmospheres). Much of the density components depends on the relative sizes of the core, mantle (if present), and crust. The next two illustrations show the measured (or estimated) size of each component:





The next table summarizes orbital parameters and atmospheric characteristics of the solar planets.



The orbital inclination is a measure of the departure of a planet's orbital plane, in degrees, from the orbital plane of Earth around the Sun (Sol) which defines the **ecliptic**, a plane containing both the Sun and the Earth's orbit (see page 20-5, near the top).

History of Planet Studies and Aspects of their Motions

The table above shows that the four planets closest to the Sun are small compared with those beyond Mars. These are the Inner or Terrestrial (like Earth, with rocky material at their surfaces) planets. From Jupiter through Neptune, the planets are much larger (the Outer or Giant group) and have surfaces that are all gas (Pluto, the exception, may be a "maverick", possibly being an escaped satellite). Nearly all planetary satellites are either rocky or a mix of rock and ice (one, Saturn's Titan, has a thin atmosphere). The four inner planets and Jupiter and Saturn were known since ancient times; Uranus was discovered in 1781, Neptune in 1846, and Pluto in 1930. The Sun-orbiting planets are recognized by astronomical observations because they move relative to the background stars (the ancients called them "wanderers").

Despite the efforts of pre-Renaissance astronomers (e.g., the Greek, Ptolemy, living in 2nd Century Alexandria, and later Arab observers) to develop a legitimate model of the Solar System, the frame of reference put the Earth at the center of the System (geocentric model). The ancients perceived the "Universe" (for them, mostly the known planets, and other points of light called stars) as a set of concentrically nesting spheres that had the Sun on one sphere; all spheres rotated around the Earth at different rates. The "map" below is one version of the geocentric Solar System as envisioned in late Medieval times; note the descriptors are in Latin:

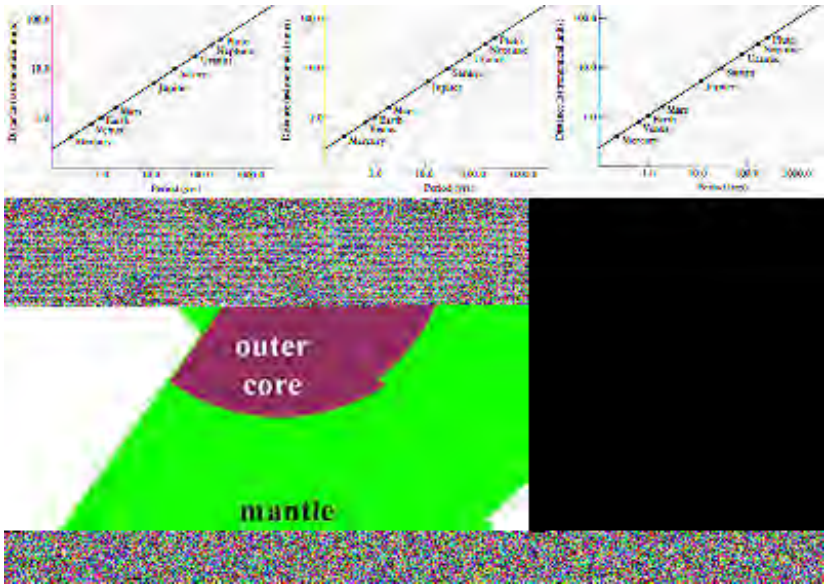
This was replaced in 1543 (date of publication) by the heliocentric model, based on work by the Polish scholar and priest Nicolaus Copernicus, who postulated that the Earth rotates and the planets revolve around the Sun. This Copernican model was largely ignored for decades, mainly from philosophical/theological objections, until observations by Tycho Brahe in the 17th Century supported the Sun-centered scheme (which, unfortunately, he rejected after conducting a flawed experiment). Galileo also made vital observations through one of the first telescopes; his discovery of satellites around Jupiter confirmed the notions of bodies revolving around a central body. General acceptance by the scientists of the times was still slow but the laws of planetary motion enunciated by Johannes Kepler (Tycho's protege) and motion in general by Isaac Newton finally led to such overwhelming evidence that scientists and other thinkers and eventually the Church acceded to this reality.

Kepler deduced from the patterns of motion that the planets revolving around the Sun did not follow precise circles but instead followed elliptical paths with the Sun at one of the two foci that characterize an ellipse. The ellipses defined by him and later astronomers were only slight departures from circularity, except for Mercury (strongly elliptical) and Pluto (which periodically crosses the ellipse traced by Neptune). Kepler's second law is derived as follows (see figure below):

Start with a line from the Sun to a planet at any locus. e.g., a, along its orbital path. After it had moved some distance a-a' along the path, it will define some given area A for the time in transit, For another segment elsewhere along the orbit, a different pattern - area B - ensues as it traverses the distance b-b'. Now, if the elapsed time between orbital transits from positions a to a' and b to b' are specified to be the same, the areas in the patterns will be equal (A = B). The law can thus be stated: Imaginary lines from the Sun to any planet sweep out equal areas in equal elapsed time intervals during different stages in the planet's revolution. Since the distance a-a' is shorter than b-b', it follows that the velocity (distance/time) of the planet moving through b-b' is greater than the speed through a-a'; in other words, planets move faster when closer to the Sun. Separate arguments based on Newtonian mechanics show that the velocities of the planets decrease progressively outward from the Sun.

(As an aside which applies both to the planets and to orbiting satellites [like Landsat], the velocity needed to achieve and maintain orbit is a balance between the forward motion vector of the moving body and the gravity vector pulling it towards its parent body [whose mass is assumed to be at its center]; thus the tendency to move away tangentially is offset by gravitational force such that as the parent, e.g., Earth, rotates such that seemingly its surface falls away from the tangential line, in fact the satellite (or planet) is pulled downward just enough to maintain the same distance to the center of mass, describing a path that produces a circular orbit [or is modified to some degree of ellipticity], even as its momentum [mv; v varies for the elliptical case] keeps it in that orbit. Like the planets, the velocity needed to get and keep an Earth-orbiting satellite in place decreases outward. Landsat moves much faster [~26,600 km/hr], and with a period [time to complete one orbit] of 103 minutes, than does a geostationary satellite. The latter, when placed at 22,300 miles [36,235 km] above ground, moves slower [24 hours to complete an orbit] over a much longer orbital path at an orbital velocity of ~11000 km/hr; when inserted so as to move parallel and over the equator, the geostationary satellite moves forward at the same speed as its nadir point on the equator and thus is stationary [no relative movement] with respect to that point on the Earth's surface.)

Kepler discovered a third relationship affecting the paths of the planets. If the orbital period P of a planet (third column in the table above) is plotted on log-log graph paper against its distance R (second column) from the Sun (taken as equal to the semi-major axis of the path ellipse), then the result is as appears below. The mathematical expression for the equation representing the resulting line is $P^2 = R^3$, the mathematical statement of Kepler's third law.



Another, rather curious relationship was put forth by Johann Titius in 1766, with later modification and promotion by Johann Bode. To formulate it, look at this sequence:

N = 0.0 0.3 0.6 1.2 2.4 4.8 9.6 19.2 38.4 76.8

Now, add 0.4 to each N term, giving this new sequence:

0.4 0.7 1.0 1.6 2.8 5.2 10.0 19.6 38.8 77.2

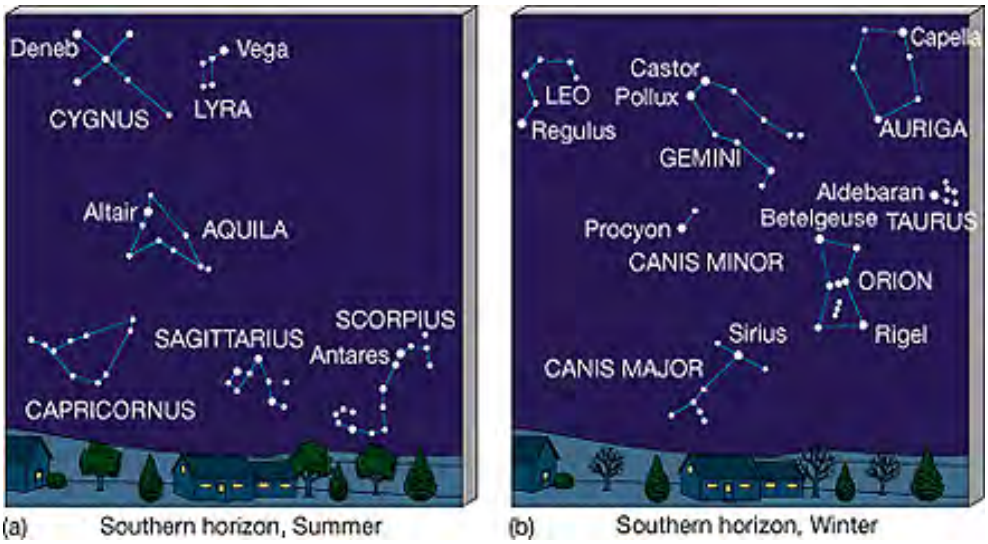
The actual positions (in A.U.s) of the planets are: Mercury = 0.39; Venus = 0.72; Earth = 1.0; Mars = 1.5; Asteroid Belt = ~2.8; Jupiter = 5.2; Saturn = 9.5; Uranus = 19.2; Neptune = 30.0; Pluto = 39.5.

For each successive planet we doubled the previous N and added 0.4; for Mars this yields 1.6 and for Neptune this results in 38.4 + 0.4 = 38.8. Remarkably, this set of numbers is closely matched by the actual distances as Astronomical Units for all the planets except Neptune which lies at 30.07 A.U. Pluto (no longer a planet), however, lies at 39.4 close to the 38.8 value. The value 77.2 suggests a "missing" planet at that distance. No obvious physical reason has yet been found for the Titius-Bode "rule", nor is the Neptune anomaly readily explanable (it may relate to an interaction with Pluto). But one consequence was a prediction that some planetary body should exist at A.U = 2.8. None was known at that time but the later discovery of the Asteroid Belt at 2.8 fulfilled the prediction. That gap is evident in the figure above, as is the anomalous position of Neptune.

The more general subject of celestial motions, or also referred to as celestial mechanics, is rather complex - beyond the scope of the Tutorial here (some aspects are treated near the top of page 20-2). But a few ideas relating to the Sun's and the planets' motions relative to Earth as the observing platform are introduced here.

These motions with respect to the Earth and to the Sun's path across the sky can be plotted on the Celestial Sphere - usually displayed as either the hemisphere north or the hemisphere south of the Equator or one's local horizon, for each case showing the expanse of celestial objects above the horizon in all directions. Thus from any location on one or the other hemispheres, there is a a local observational hemisphere on which the motions of the Sun, the Moon, the planets, and the stars appear to move as the Earth rotates each day, causing each body or point of light to trace an arcuate path across the hemisphere (best seen at night). An axis around which the heavenly bodies seem to move in circular arcs is imagined to pass from Earth to the hemisphere, with the North Star (Polaris) being very close to the point where the celestial axis would penetrate the (hemi-)Sphere; by convention, the Earth's rotational axis is placed to coincide with the celestial axis, and the Earth' North Pole is near the celestial North Pole. The position of the North Star, and hence all other celestial bodies, will vary with latitude and with

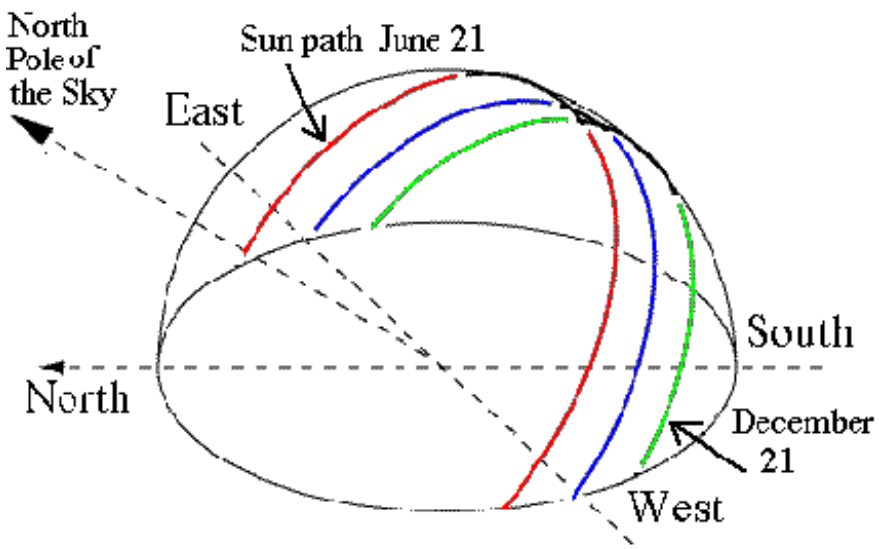
season of the year. At different times in a year different distinctive groupings of stars - known as constellations (discussed again on page 20-2) appear in the same part of the Sphere at the same time of night, are illustrated here:



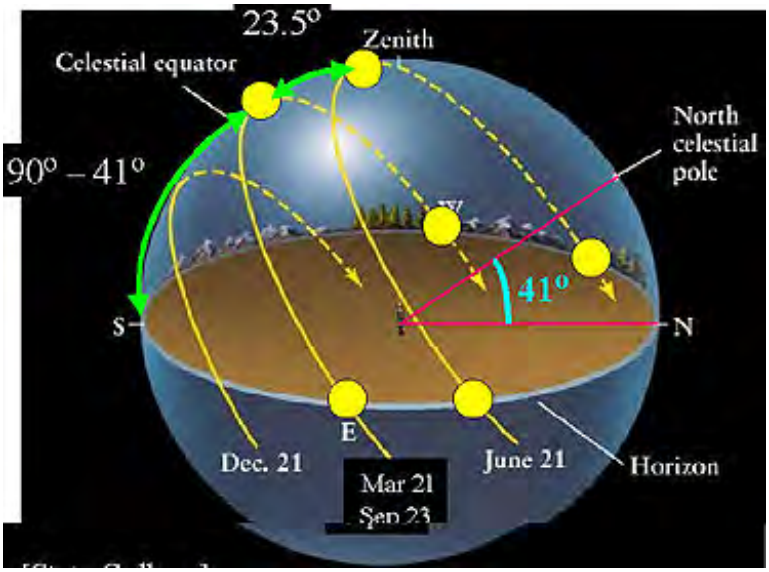
One approach to understanding celestial motions is to trace the pathway of the Sun across the Sphere as it varies seasonally and with location on the Earth's spherical surface. First, we introduce the concept of the solar ecliptic. This is the plane defined by the elliptical trace of the Earth's annual orbital path around the Sun (the orbital traces of the other planets approximately follow the ecliptic as well; Mercury at 7° and Pluto at 17° have the largest departures from the orbital plane as defined by Earth's path). A corollary: the ecliptic is also the Sun's annual path among the stars as viewed from Earth. In this frame of reference, the Earth, and also six other planets rotate in the same direction - counterclockwise (prograde) as viewed from above (as from the celestial North Pole); Venus and Uranus have retrograde (clockwise) rotations. None move in circular orbits; each follows an elliptical path, usually with small ellipticity, such that it will be varying distances to the Sun with points of closest and farthest approaches.

While all the planets orbit on or near the ecliptic, their rotational axis is tilted to varying degrees relative to the ecliptic plane (or to a line normal to [90°] the plane). Only Mercury's axis is at 0° with respect to this normal. Venus and Pluto have their north rotational pole tilted more than 90°, hence the top of the axis lies beyond (south of) the ecliptic. The Earth's axis is tilted 23.5° from the a line at right angles to (normal to) the ecliptic.

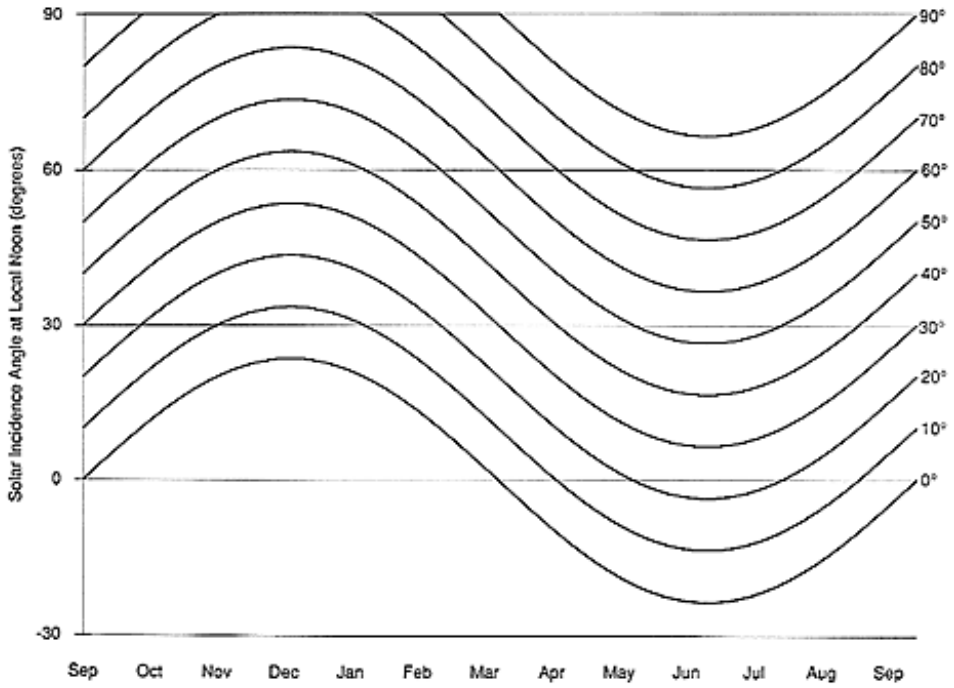
Two diagrams will indicate the seasonal and the latitudinal effects of the Sun's movement across the celestial hemisphere:



In the above diagram, the (unspecified) latitude of observation, along with Earth's axial tilt, is responsible for the celestial North Pole being about 60° down towards Earth's Equator (in this projection, an 'ellipse' marked by E, W, N, and S). The highest the Sun will reach is at Summer Solstice (~June 21); the plot of a daily path then across the sky at that latitude is a partial circle (arc) shown in red that has the greatest length, thus resulting in the maximum amount of sunshine on that day. That arc begins at sunrise (after dawn) and ends at sunset. The smallest arc is at Winter Solstice (December 21), shown in green; at that time the daylight is the least - several hours less than 12. At the Spring (~March 21) and Fall (~September 21) Equinoxes - blue line - the arc is at a median length, and the sunrise-to-sunset duration is about 12 hours.



The second diagram has a different orientation of the equator, and a set of seasonal Sun paths, for a location at 41° North. Here the Sun becomes almost vertical (would cross the zenith on the Sphere) on June 21. Thus, the Sun path will vary systematically with latitude (the effect of the Earth's axial tilt is built in). This can be plotted on a diagram:



Here's how to read this diagram. First consider your location at 0° latitude, i.e., you are on the Equator. This is represented by the bottommost curve. The starting point is September 21: the Sun is directly overhead at Noon. By December 21 (Winter Solstice) the Sun has migrated northward to 23° North latitude, reached when local Noon occurs. By March 21, the Sun has 'retreated' southward to again cross the Equator and be overhead. The Sun then continues

southward into the southern hemisphere reaching 23° South at the noontime of the Summer Solstice. It then progresses back to the Equator to repeat this cycle. Now, look the 90° curve (right side) which places you at the North Pole: From March 21 to September 21, the Sun will never set - 24 hour daylight; from September 21 to March 21 the Sun does not rise - "perpetual" night. At high latitudes long days or long nights, but some Sun setting, will occur. For a third case, consider the curve for 40° North: The Sun at that latitude is at 40° North on September 21; it migrates to about 62° North by December 21, then moves southward until it reaches its southmost position at ~19° latitude at the Summer Solstice, then reverses trend, going back to 40° North by Fall Equinox. Try your ability to interpret a curve for some other latitude.

From day to day during the year, as Earth orbits the Sun in the plane of the ecliptic and maintains a tilted axis of rotation which points in a fixed direction in space, we see four things change:

(1) the sunrise direction changes

(2) the sunset direction changes

(3) the length of daylight changes

(4) the height of the Sun at noon above the southern horizon changes

Thus, the two Celestial Sphere diagrams show that the Sun follows arcuate patterns (of varying lengths depending on time of year) across the sky as seen at any location during the day. Stars, on the other hand, will follow circular patterns, centered on the Celestial North or South Poles, over much of the Sphere. Planets, being part of the Solar group circuiting the Sun, will also move during a given night, but their paths over months and the seasons will appear as though they "wander" through the background of stars and constellations. Their shifts are not equal per terrestrial time period and are seen from Earth as influenced by their relative distances from the Sun and their locations in their elliptical orbits relative to earth observers.

Meteorites as Samples of Planetary Materials

Prior to the space program which has led to visits of unmanned probes past or onto planetary bodies and over one fabulous decade the landing of humans to explore the Moon's surface, our knowledge of the planets were largely confined to two avenues of investigation: 1) telescope observations and selected properties measurements using accessible parts of the EM spectrum, and 2) samples of one or more planets and smaller solid bodies that fall to Earth as meteorites (or, as discussed in Section 18, as large bolides - megameteorites, asteroids, and comets). The bulk of the rest of this page will be devoted to a review of meteorites, which continue to be a prime source of information about some of the Sun's planetary and fragmental-bodied associates.

"Stars" falling from the skies have been known since ancient times; rarely, stones are found that were tied to these "shooting stars". One such rock has been venerated by Islam (in its encased shrine in Mecca) for more than 1300 years. By the 19th Century, meteorites were identified correctly as samples from other parts of the Solar System. They are part of the nearly 500 tons of extraterrestrial rock material that reaches and enters the atmosphere each day. Most of that material is burned up by friction from the high speed of entry but meteoric dust can remain in the air and a very few individual blocks of material survive this passage to fall in the sea or on the ground as meteorites.

A general nomenclature has been developed to describe rocks in space that may reach the Earth's surface. If these rocks are relatively small (say about house-size or less), as they exist in space they are referred to as *meteoroids*. If they reach Earth and pass through the atmosphere, creating intense light as their outer skin is melted by friction, they are called *meteors*. If they do not burn up completely in transit, and land on the Earth's surface, they now are designated

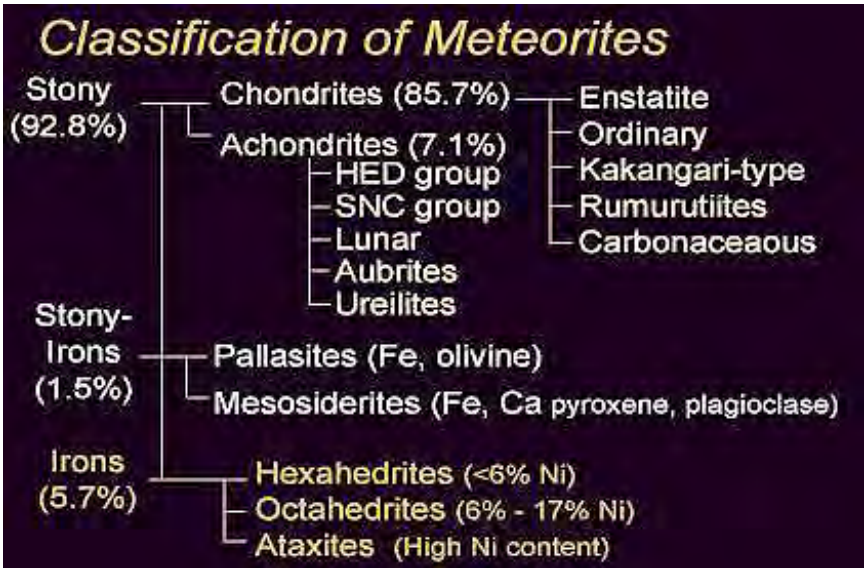
meteorites. The largest meteorite found so far on land is about the size of an automobile; most meteorites are much smaller. Much larger bodies moving in space, such as asteroids and comets (page 19-22), can strike the Earth, either as still intact bodies or broken into fragments; these will nearly always produce impact craters (Section 18) or shock-induced destruction on the ground (such as knocking down trees) if they explode in the atmosphere. Another terminology distinction: Meteorites whose passage through the atmosphere was observed and then someone soon thereafter locates the object are referred to as *Falls*; those whose passage was not observed but were eventually discovered (often by serendipity) are called *Finds*.

By the start of space exploration, nearly 1900 meteorites had been collected. That number has jumped notably (over ten thousand) when scientists exploring the Antarctic deduced that a few of the rocks scattered about the ice surface might be meteoritic debris. Patient collection has since verified this, thus providing a very effective way to find new meteorites. However, of any thousand rocks on the Antarctic surface, only about 1 or 2 prove to be meteorites. But each year, a new expedition (on snowmobiles) continues to add to the total.

In searching for meteorites, two clues call attention to certain stones as candidates to be collect and broken into to reveal indications of their nature: 1) rocks that appear to be composed solely or largely of iron metal; and 2) rocks that have a thin dark fusion crust, where friction has melted the exterior. Although the classification of meteorite varieties consists of various categories, most meteorites fall into two types - Iron and Stony - as shown here:

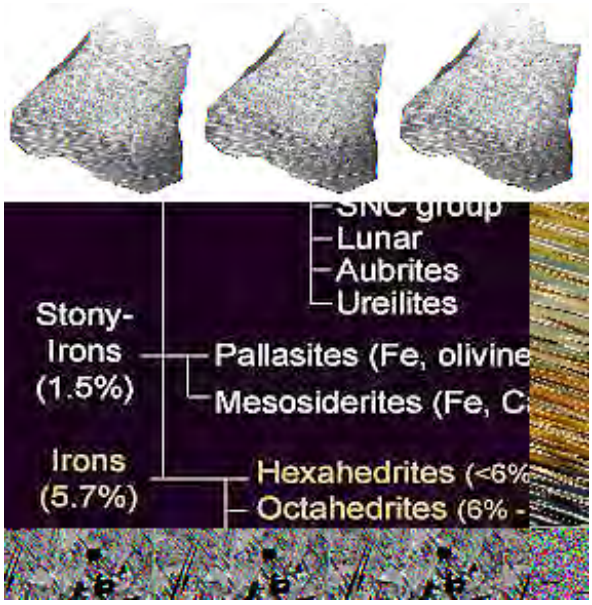
The Renazzo stony meteorite shown below as broken open reveals the typical texture of this type:

The mineral composition of meteorites is distinctive. The iron meteorites contain native iron metal alloyed with 5 to 17% nickel. The stony meteorites are composed of minerals that are common in basic igneous rocks: olivine, pyroxene, and plagioclase feldspar. together with a variety of minerals (some found only in meteorites) present usually in small quantities. Various combinations of these and some other constituents, together with distinctive textures, provide the basis for classifying the different meteorites. One general classification appears below. You can examine a more detailed classification by going online to this helpful website.



Iron meteorites (known as Siderites) are uncommon but quite distinctive. (Most believe they are the core material in differentiated (melted) asteroids. They contain one or both of the structural phases of metallic iron: Kamacite and Taenite. The Iron types are classed by the amount of nickel present and the nature of the iron phase(s). When an iron meteorite's interior is exposed, usually by sawing to create slab faces and then etched by nitric acid, some distinctive textures are often present, such as what is termed "Widmanstatten strucure" caused by unmixing of the two structural phases (the broader bands are Kamacite), displayed here at two magnifications.

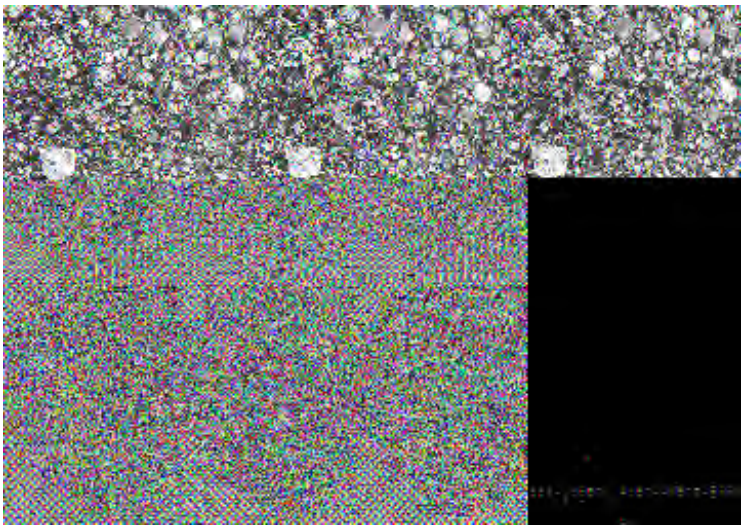
Another planar structure in the Iron meteorites is called Neumann Bands, which is a twinning mode induced by shock. Most likely, this shock effect occurred during the breakup of the parent body of which the Iron meteorite was in the interior (a core analogous to the Earth's?):



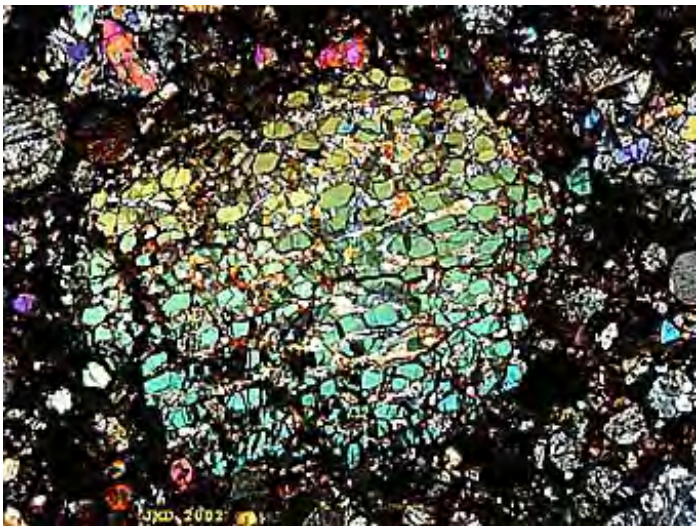
Transitional to the stony types are the stony irons, that include the Pallasites and the Mesosiderites. An example of the first is the Esquel meteorite (generally, a meteorite is named from a geographic location where it fell and was collected:

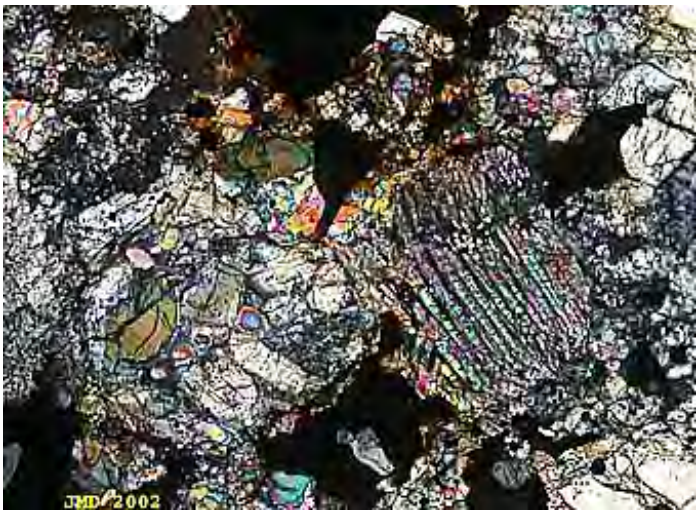
As the percentage of native iron decreases and silicates increase the resulting stony-iron meteorites are called Mesosiderites.

Most meteorites based on the percentage of Falls (those observed as "shooting stars"), which may not be the same as the percentage of Finds, since some meteorites are more likely to be destroyed by weathering, etc.), are of the type called Chondrites which in turn are grouped into classes depending on mineralogy and texture. Chondrites contain generally small (millimeters up to a centimeter) spherical bodies called *chondrules*, which most meteoriticists believe were once molten silicate droplets produced by melting of interspatial dust by one or more mechanisms such shock waves or heat from the forming Sun. They then cooled and crytallized into Olivine, often accompanied by pyroxene mineral species (Enstatite, Bronzite, Hypersthene) and Plagioclase (calcium-rich). Most chondrites contain small crystal specks of iron-nickel. The chondrules seem to be embedded in other dust and isolated crystals which incorporate the chondrules as the meteoroid or asteroid built up from the remaining materials in the dust clouds surrounding the growing Sun, over the first few million years of the organizing Solar System. This photomicrograph shows a texture characteristic of chondrites, with subspherical chondrules, crystal fragments, a few iron-nickel grains, and a fine-grained matrix:

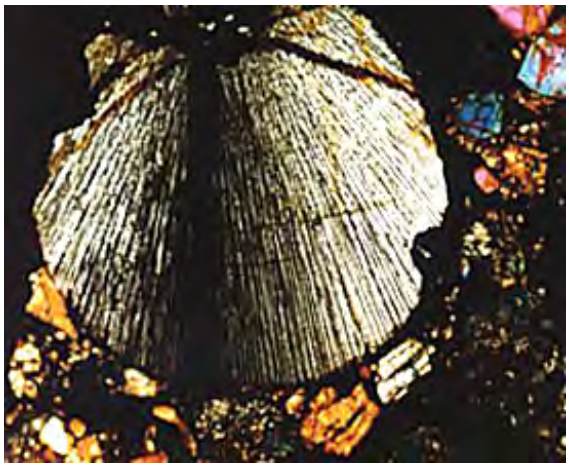


The next two figures depict photomicrographs (with the petrographic microscope's Nicols in the cross-polarization mode) of individual chondrules:

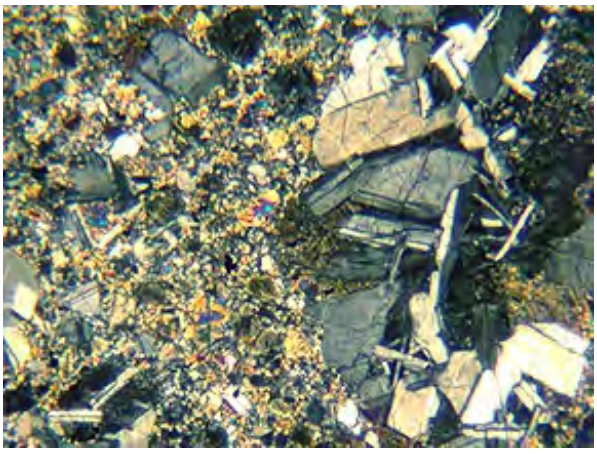




Some chondrules show a characteristic radiating structure assumed by the Pyroxene



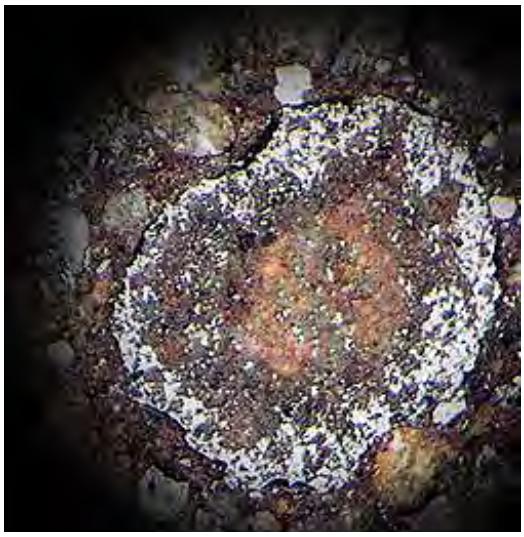
Plagioclase can be conspicuous in some chondrites, as shown in this photomicrograph.

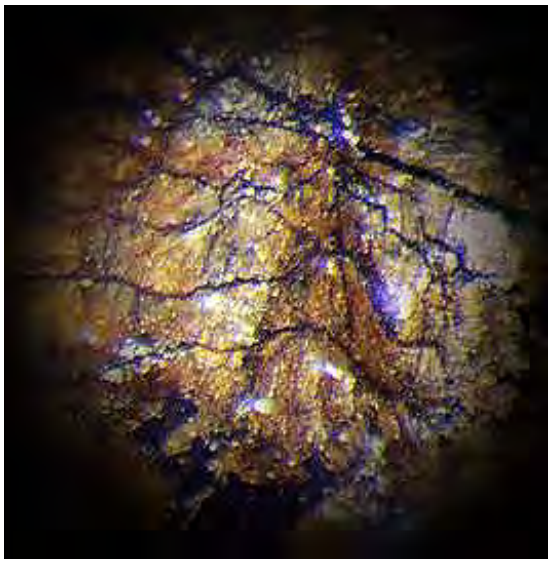


The bulk texture typical of an Ordinary Chondrite is exemplified in this slab cut into the Homestead meteorite:

Somewhat larger chondrules are present in this sample from the Brenham meteorite:

Classification of the Chondrites is determined to some extent by the particular mineral species present. However, the usual hierarchy (Type 1 through Type 6) is determined by the degree of water content and extent to which the chondrule appears to have been reheated and thus recrystallized by thermal metamorphism. Type 1 is most primitive and contains some water; it probably was never reheated after primary crystallization beyond about 300 °C. Type 6 is anhydrous, shows thermal and/or shock textures, but was reheated up to about 800°C. In the next two photomicrographs are shown 1) a ring of iron metal that accumulated when the chondrule was thermally heated to the extent that iron was melted; 2) a chondrule with veins of glass caused by shock heating.

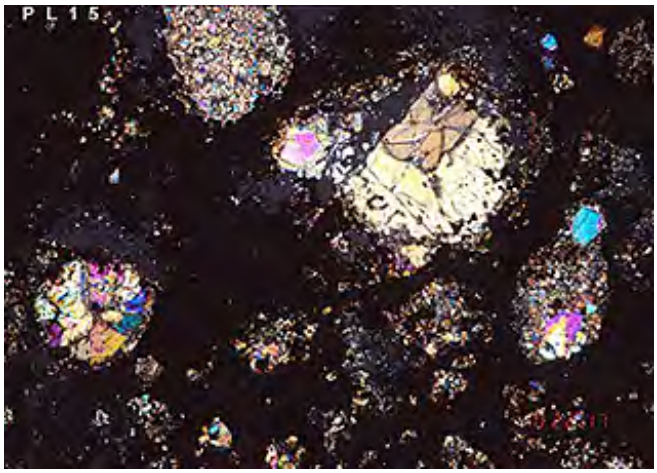




We turn now to a special class of Chondrites called Carbonaceous Chondrites. These contain up

to 6% carbon, either in elemental form or in the composition of organic (hydrocarbon compounds, including some amino acids, but not biogenic) molecules that occur within them. Low temperature minerals, such as clay minerals and serpentines, attest to the conclusion that the matrix was never subjected to the high temperatures that melted the associated chondrules. This is supported by the variable water content; some of these meteorites contain up to 11% H_2O. Many meteoriticists consider carbonaceous chondrites to be the most fundamental and primordial representatives of the solid materials available for making up the planetary system. They are thus held to be condensates of melted silicates that mixed with low temperature organic and inorganic phases which grouped into asteroids and comets, or were aggregated into the planetesimals that evolved into the planets. Here is one of the best-studied of this class - the Murchison meteorite that fell on Australia:

One of the most famous meteorite falls was the Allende carbonaceous chondrite, in which nearly two tons landed in a farmer's field in northern Mexico in 1969. It's quantity has proved to be a bonanza for researchers. Below is one of the pieces and a thin section which shows a carbon-rich matrix around the chondrules

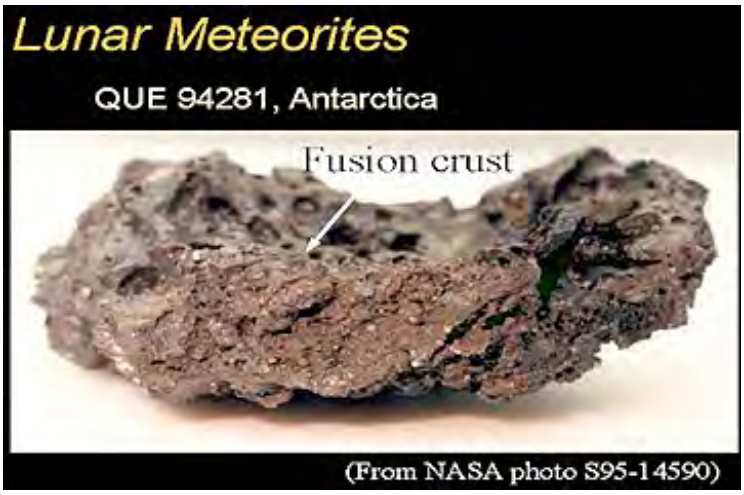


About 8% of the silicate (stony) meteorites do not contain chondrules; the group is known as the Achondrites. There are many varieties, as evident in the classification we pointed you to. Most members are thought to have come from the surfaces of asteroids. Some of these are breccias and other unusual textures may be distinctive. Eucrites are a common class and are either similar to terrestrial basalts in texture or are brecciated. The basaltlike Millbillillie exemplifies the first type here;

Two brecciated achondrites appear here:

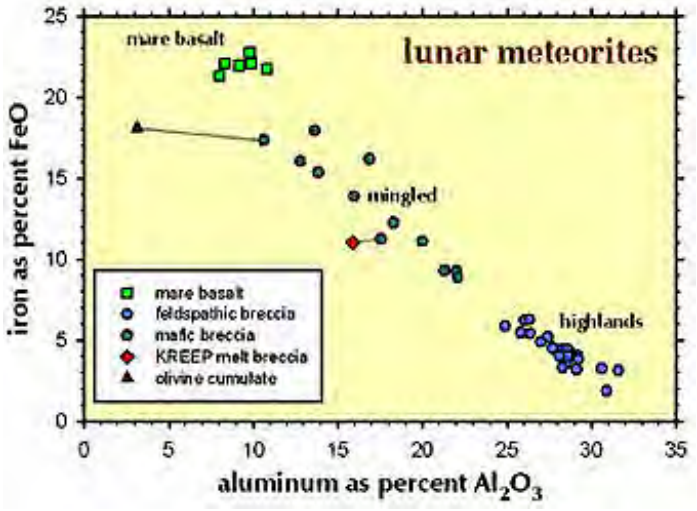
There is a growing realization that many of the Achondrites may be pieces of the Moon or Mars expelled from these bodies by impact.

Lunar meteorites may directly strike the Earth after thrown off by a lunar impact or fall after being captured in orbit. Martian meteorites need to be thrown out beyond martian gravity into orbits that may be perturbed or decay to allow eventual Earth-crossing encounters. Below are three meteorite samples of probably lunar origin, as determined by age and composition.



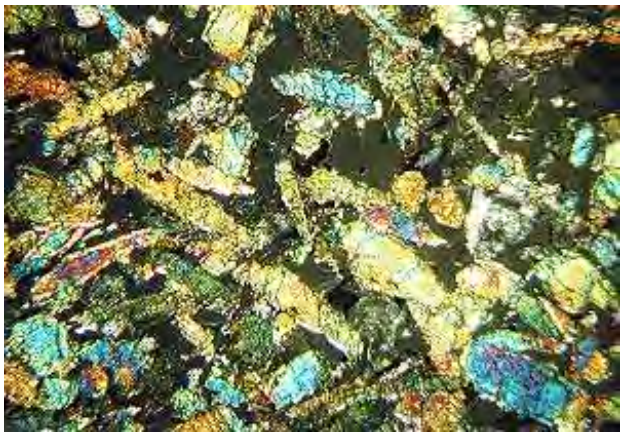
This last meteorite is remarkably like lunar regolith (the loose debris on the surface). If so, it was shock-lithified by the impact that hurled it to Earth; if not, it was probably breccia rock that was part of an ejecta blanket later lithified.

After the Apollo moon rock returns, it became much easier to prove certain meteorites to have come from the Moon; both chemical composition and isotope ratios were particularly diagnostic. Here is a plot of the composition of meteorite specimens of certain-to-probable lunar origin:



Less than a 100 meteorites are thought to Lunaites (Moon meteorites). Most are fragmental. A neat review of Lunar Meteorites is found at this Washington University web site.

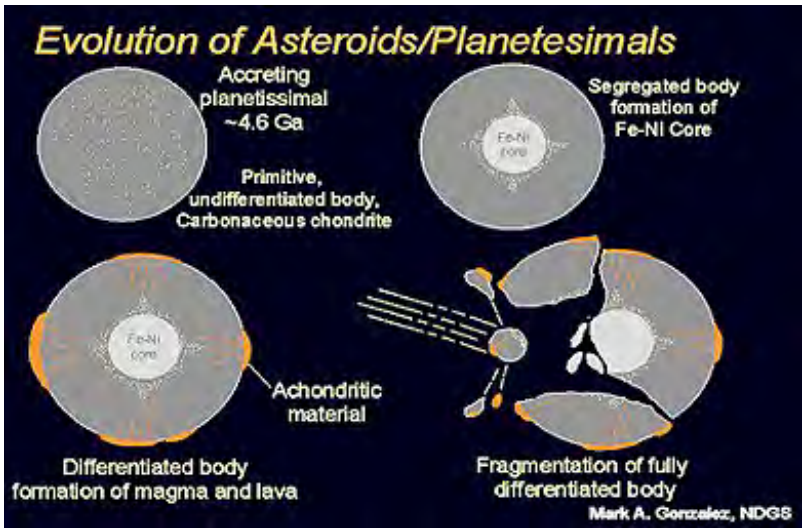
At least 35 meteorites have evidence that they came from Mars. The next figure is of a Nakhrite type meteorite of probable martian origin whereas the second illustration shows the texture of the Zagamil meteorite which is considered of martian origin. We will show other examples of these planetary meteorites on pages in this Section that treat the Moon and Mars.



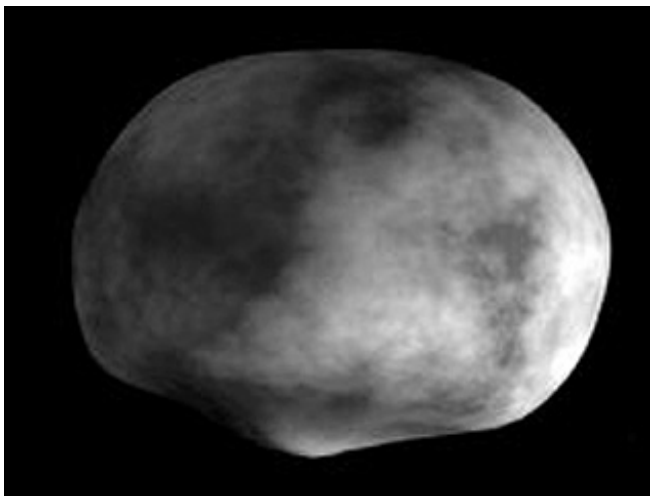
The age(s) of meteorites can be instructive. Elemental isotopes are used to date them. The chondrites give very old ages (clustering around 4.5 - 4.6 billion years), suggesting that these formed near the beginning of the Solar System. These ages are determined by Uranium-Lead and Rubidium-Strontium isotopic analysis; the presence of I^{129}, derived from Xe^{129} decay, which has a short halflife, confirms that at least some of the constituents were incorporated early in the inception of the Solar System. But there are one ot two younger ages, called exposure ages, which indicate times when the meteorite body separated from a larger host body and began its

travel through space. Abnormalities in the amount of Ar^{40} and He gas point to a time when larger parent asteroids may have broken up from collisions. Still younger ages deduced from He^{3}, Ne^{21}, Al^{26}, and A^{38} contents are associated with times when the meteoroids were traveling in their final sizes and were subject to cosmic ray bombardment.

Genetic implications of the different meteorite types are these: For those not of lunar or martian origin, there may have been four stages of organization of meteorite parent bodies (most believed to be from the asteroidal belt between Mars and Jupiter): 1) condensation of high-temperature refractory silicates, oxides, and metals; 2) separation of silicates from metals as granular particulates in the solar nebula; 3) condensationn of lower temperature or volatile phases; and 4) varying degrees of remelting of the earlier condensates. From a different persepective: 1) the Carbonaceous Chondrites are the most primitive; 2) Chondrites formed from aggregation of chondrules (melted by shock, thermal radiation, or other process[es] and dust into bodies that never became large enough to melt; these bodies may, however, have experienced collisional breakup of asteroids (thus, some of the Achondrites might be so derived), 3) the Iron meteorites may (?) be cores of completely melted large asteroids, or less likely, bigger planets that were destroyed by collisional disruption, and 4) Some of the Achondrites were made by fragmentation/reassembly of differentiated planetary or asteroidal surfaces subjected to impact bombardment, or may be shock-lithified surface rubble (such as the regolith deposits on the Moon, as discussed later). Some of the general conditions that lead to different meteorites derived from asteroids are depicted in these diagrams:



At least one meteorite has been traced to a specific asteroid, Vesta, based on strong similarities in spectral properties:



As space exploration goes on, more answers to organizational details are forthcoming, e.g., the similarity of asteroidal material to carbonaceous chondrites has been established by probes that approach or land on the asteroids.

The importance of asteroids in the makeup of our Solar System is paramount. But we will defer further discussion of these bodies until after we have examined the major planets. However, for the curious who would like some insight now, go to page 19-22.

We have said nothing on this page, nor elsewhere in this Section, about the origin of the planets and the development of a Solar System. These topics are treated in some detail on page 20-11, after astronomical principles are considered. We will start our extraterrestrial planetary tour with the Earth itself and then Earth's sole satellite, the Moon. The geological aspects of Earth were covered on page 2-1a and 2-1b, to which the curious user can refer now by clicking on this page number for a refresher review. However, the Earth does deserve a brief overview of its general nature and history as one of the planets within the Solar System.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Because of its proximity (mean distance: 385,000 km, ~238,000 miles from Earth), the Moon has always been the "heavenly body" most easily and ambitiously studied. The space program made the lunar surface an early target, with a number of exploratory, mainly photographic missions and culminating with six manned landings during the Apollo Program. On this page some of the characteristics of the Moon, its albedo, the phases of illumination, and its orbital "locking" on Earth are described. Then, some of the remote sensing studies, optical and radar, of the lunar surface are considered.

Pre-Apollo Exploration of the Moon

Lets begin with an image that places the Moon in context with its parent planet Earth. This view, taken from the Galileo spacecraft to be described later, is what an extraterrestrial space traveler might see as he/she/it approaches the Earth-Moon system.

Without question, the greatest part of early remote sensing investigations of the planets, as expected from its proximity, centered on the Moon. Indeed, its mean distance from Earth of 384,400 km (238,710 miles) meant that even before the 20th Century, dreamers could fancy someday going there for a visit (Best known fantasy: Jules Verne's *Journey to the Moon*). Earth's sole satellite is much smaller than the parent body: 3475 km (2160 miles) in diameter. (For a quick synopsis of the Moon's general characteristics, read through this review by Bill Arnett: The Moon . Then, also check with the Web Site Rosanna Hamilton's Moon Page, looking particularly at the Table of Lunar Statistics and the section on the origin of the Moon stemming from an early huge impact of an asteroid with the primitive Earth).

Lunar Statistics are important and informative. Reproduced here are most of the "vital" ones:

*** Mean Distance of Moon from Earth - 238,712 mi (384,400 km)**

*** Greatest Distance of Moon from Earth (Apogee) - 252,586 mi (406,740 km)**

*** Shortest Distance of Moon from Earth (Perigee) - 221,331 mi (356,410 km)**

*** Circumference - 6,790 mi. (10,930 km.) 0.27 of Earth's circumference**

*** Diameter - 2,160 mi. (3,476 km.) 0.27 of Earth's diameter**

*** Mean Radius - 1,079 mi. (1,737.5 km.)**

*** Equatorial Radius - 1,079 mi. (1,738 km.)**

*** Polar Radius - 1,077 mi. (1,735 km.)**

*** Mean Angular Diameter - 31' 07"**

*** Mass - 8 x 10 (19) tons (7.35 x 10 (22) kg)**

*** Mass Ratio (Earth Moon) - 81.301**

*** Volume - 2.4 x 10(9) mi^3 (2.197 x $10^{10}$$km^3$**

*** Mean Density - 208 lb/ft^3 (3.34 g/cm^3) 0.6 Earth's density**

*** Gravity at Surface - 5.31 ft/s^2 (1.62 m/s^2) 1/6 Earth's gravity**

*** Escape Velocity - 1.48 miles/sec (2.38 km/sec)**

*** Mean Inclination to Lunar Equator - 6° 41'**

*** Mean Orbital Inclination to Ecliptic - 5° 08' 43"**

*** Oscillation of Orbital Inclination to Equator - +/- 0° 9' every 173 days**

*** Inclination of Lunar Equator to Ecliptic - 1° 32' 33"**

*** Period of Revolution of Perigee - 3,232 days**

*** Period of Rotation about the lunar axis - 27.3 days**

*** Orbital Direction - east (counterclockwise from above)**

*** Mean Orbital Speed - 2,287 mi/h (3,683 km/h) 33 minutes arc/hour**

*** Daily Sidereal Motion - 13.176358°**

*** Mean Centripetal Acceleration - 0.0003 g**

*** Mean Eccentricity of Orbit - 0.0549**

*** Synodic Month - 29.53059 days (29 days, 12 hr. 44 min, 2.8 sec.)**

*** Sidereal Month (star to star) - 27 days, 7hr. 43min, 11.5sec.**

*** Anomalistic Month (Apogee to Apogee) - 27 days, 13hr. 18Min, 33.2sec.**

*** Rotational Period - 27 days, 7hr, 43min. 11.5sec**

*** Surface Temperature - day = 273 F (120° C) night = -244F (-153° C)**

*** Surface Area - 14,657,449 mi sq. (37,958,621 km sq.) 9.4 billion acres**

*** Visible Surface - 41% during lunar cycle, additional 18% from librations**

*** Parallax - 0.9507°**

*** Moon's Angular Diameter - 0.5181°**

*** Magnitude of Full Moon - -12.5**

*** Average Albedo - 0.07**

*** Estimated Age - 4.6 billion years**

*** Flight Time from Earth - 60 to 70 hours**

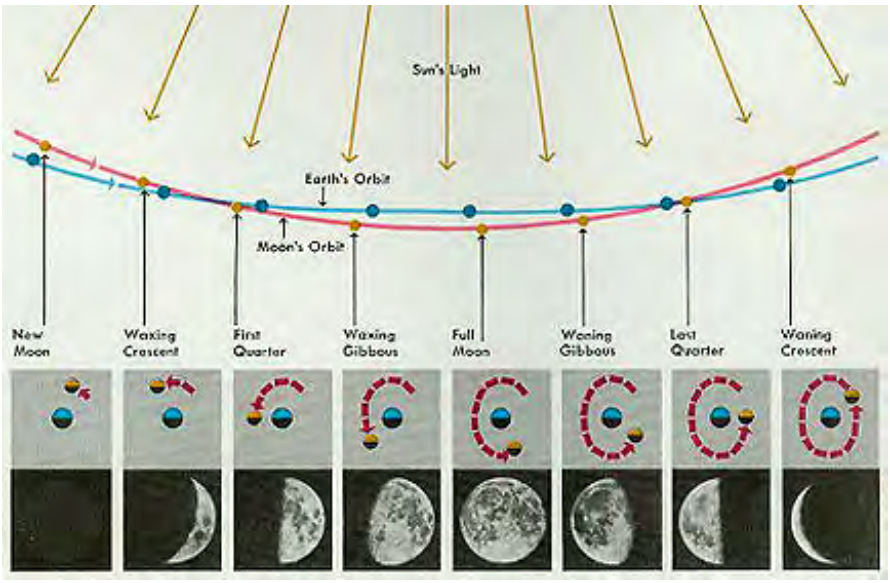
*** Increase in Mean Distance from Earth - 1.5 inches per year (3.8 cm per year)**

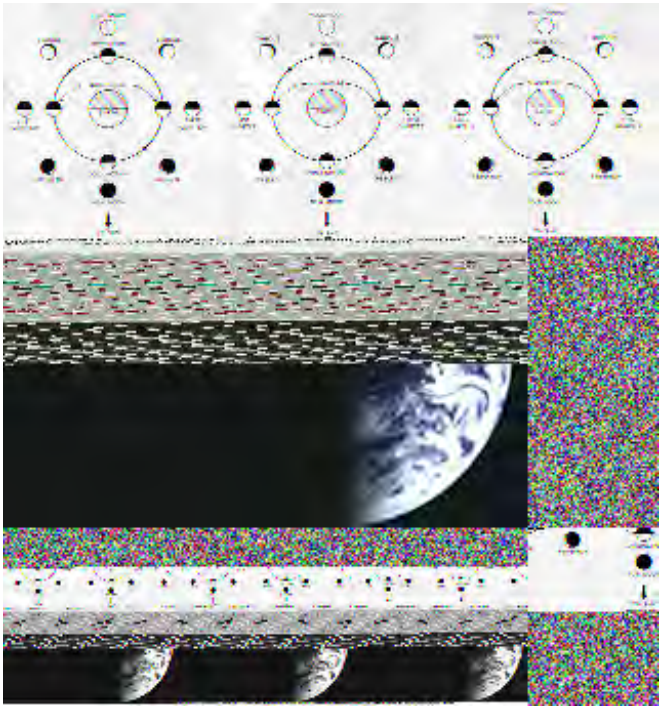
From StarrySkies.com

The Moon's orbit is not circular, but is an ellipse or elongate oval. The furthest out point (410,452 km; 252,586 miles) in the orbit is called the *lunar apogee*. The closest point (359,663 km; 221,331 miles) is the *lunar perigee*. The Earth's orbit also has an aphelion relative to the Sun, which occurs in Summer, and a perihelion, which is in Winter. The closer approach of the Earth at perihelion to the Sun increases the amount of light received by the lunar face, making the Moon seem brighter in the winter (when the atmosphere tends to be clearer) and less so in summer at aphelion.

One of the above parameters reveals why we see the Moon in the first place. The mean orbital inclination to the ecliptic of just greater than 5° means that the Moon's orbit is offset enough so that the Sun is not routinely blocked by the Moon and the Moon is not so blocked by Earth that sunlight fails to hit it. At any time of the year, say the time when this paragraph was written on August 10 of 2006, the Sun arises for the day from a direction South of East, transits in an arc towards the sky's zenith, sets towards the West. (This path is maximum at the Summer Solstice and moves progressively lower towards a minimum height in the south part of the Celestial Sphere until Winter Solstice, then starts rising in the Sphere until the Summer Solstice). The position of the Sun at High Noon varies over time owing to the 365 day year that Earth takes to orbit (revolve) around the Sun (see page 20-2). The Moon follows a similar path during any same period but is displaced 5° to the west. The first Crescent Moon (see below) reaches above the southwestern horizon early at night. Each successive night is marked by the Moon having undergone a shift along its orbit of 13.18° towards the southeastern direction of the orbital arc; the Moon arises progressively later each new night (about 50 minutes) as it moves through its sequence of phases (described below). These solar-lunar patterns of arc transits will vary depending upon the latitude of the observer.

The total intensity or brightness of reflecting surfaces at specific points or areas varies with the *phases* of the Moon (progressive waxing and waning due to illumination changes, depending on relative positions of the Sun, Earth, and Moon). The lunar photometric function, a plot of intensity versus phase angle, represents the intensity mathematically. The degree of polarization of light from the Moon also shifts with the phase angle and can show considerable variations in different regions of the Moon, again related to composition and textural changes. These two similar diagrams show how the observed phases are controlled by the relative positions of the Moon and Earth with respect to solar illumination.





19-2: What are the vernacular names applied to the phases of the Moon? ANSWER

To reiterate: The Moon rises and sets in roughly the same places on the horizon as the Sun, but about 50 minutes later each night. Every 29-30 days there is a New Moon - we cannot see it because it is near the bright late afternoon Sun in the sky. A few days later we can see a thin Crescent Moon, with its limited light surface on its eastern flank, which appears high in the western sky soon after sunset. This first visible sliver of illuminated Moon (the boundary between light and dark is the *terminator*) was actually in the Celestial Sphere and was rising during the latter part of the day, but this could not been seen because of the blinding effect of daylight from the west. Each successive night, as its phase gets greater (called waxing), the Moon becomes seen further south along its transit arc and sets a bit later until about 14 days after the New Moon we see a Full Moon. The Sun and the Moon are on opposite sides of the Earth, and after the Sun sets, the Moon rises. A Full Moon usually arises during the evening hours and, depending on when dawn occurs, may be visible much of the night. For the next 14 days the Moon's phases wane back to a crescent on its West which rises shortly ahead of the Sun, before the whole process repeats. In the summer's long daylight, a largely illuminated Moon can often be seen well into the sky in daylight. The entire sequence of phases is called a *lunation*. The amount of an illuminated Moon, at any phase, and the duration it is visible to a ground observer will change with the seasons, since in Winter darkness can last up to 14 hours and in high Summer it can be light beyond dawn for as much as 16 hours. And, the path actually

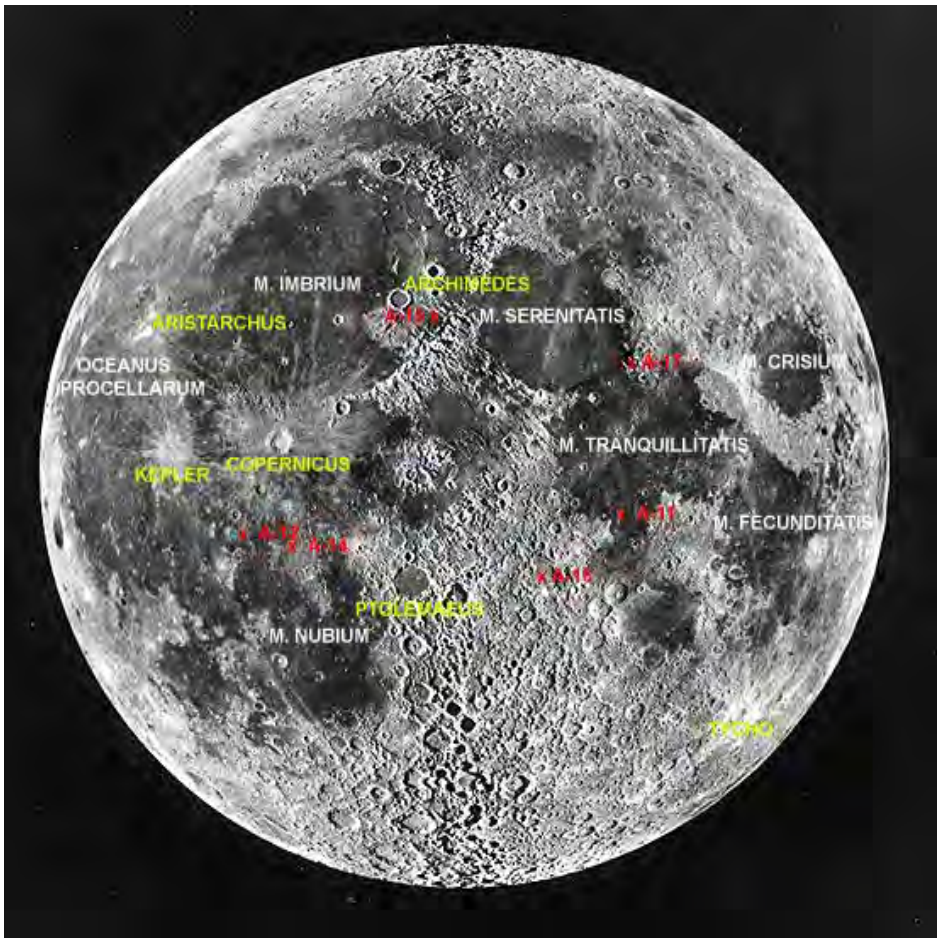
followed by the Moon across the sky depends on where the observer lives - it is latitude-dependent.

The first telephoto image of the Moon's surface was made in the 1860s. Prior to the space program, almost all lunar studies used the telescope. At first, emphasis was on optical observations (and photographs) within the visible spectrum. Among the first measurements was the Moon's *albedo*, the ratio (stated as a percentage) of the light reflected to the light incident on a surface. (The Earth's albedo is 39% on average [with slight fluctuations due to varying cloud cover] and Venus falls between 70% and 90% [from strong reflections of sunlight by a thick planetwide cloud cover].) By comparing albedos measured on possible terrestrial counterparts, the lunar albedo offers insights into the composition, texture, and structure of its surface materials. Once established, the trick is to interpret these albedo values.

We can measure the albedos of likely surface materials in the laboratory, where we can vary the textures and illumination conditions. In this way, lunar scientists postulated that the dark areas (*maria*, singular - mare) were some type of volcanic material, most likely basalt. The lighter areas (*terrae*, singular terra, or highlands) were more enigmatic. Several other volcanic types could produce albedos in the proper range, but uncertainties about particle size and other variables inhibited definitive identification. The first unmanned landers, the five Surveyors, resolved the problem. The answer was that the highlands included a different group of rock types, of which anorthosite was prevalent. Anorthosite is an igneous rock that contains large amounts of light-colored feldspars.

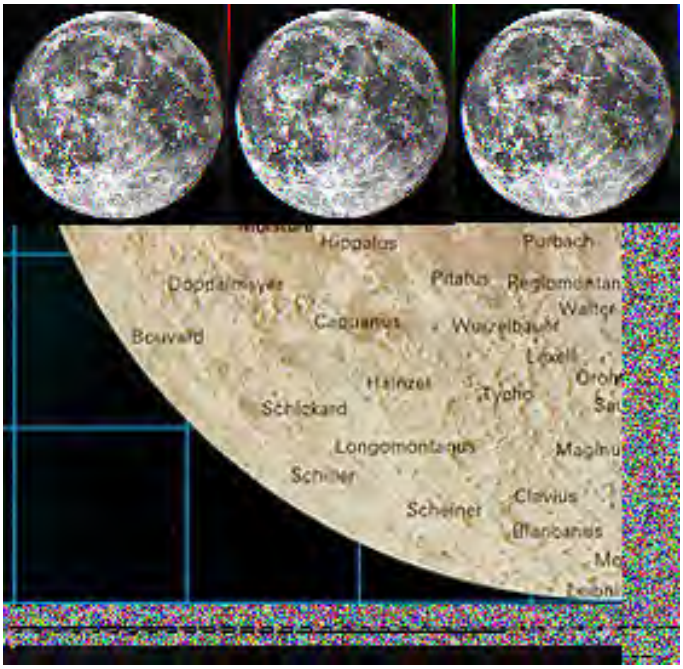
Investigators judged that the albedo values for both the maria and highlands were consistent with scattering by a granular or fragmental loose surface covering having textures associated with unconsolidated surface materials (some thought a landed spacecraft would sink into loose, almost powdery rock debris, which, fortunately, didn't happen). This covering is now called the *regolith* or also the lunar "soil". Regolith is produced by continual, repeated comminution of rock and rock fragments by multiple impact bombardments over millions of years.

At full Moon, the dark areas (maria, now known to be mainly basaltic lavas that filled huge impact basins) have albedos of 5% to 8%, while brighter areas, generally coincident with elevated terrain in the highlands (now proved to be feldspar-rich primitive lunar crust) range between 9% and 12%; these average out to about 7% for the full side facing Earth. These albedo differences are evident in this full Moon photograph, taken from the Lick Observatory in California (M. stands for Mare in the labels; A refers to Apollo Landing Sites [in red]; and major craters show in yellow):



In the above photo, the two most obvious major features are the dark patches (*maria*, or basin lavas), with a smaller number of major craters, and the lighter-toned areas (*terrae*, or highlands), with numerous larger craters. Surprisingly, good maps of these major features are seldom found in texts or on the Web. Here is one that shows many of the best-known craters; note that the maria are labeled with their English translations of the original Latin used in the above full Moon image (those labels starting with M.):

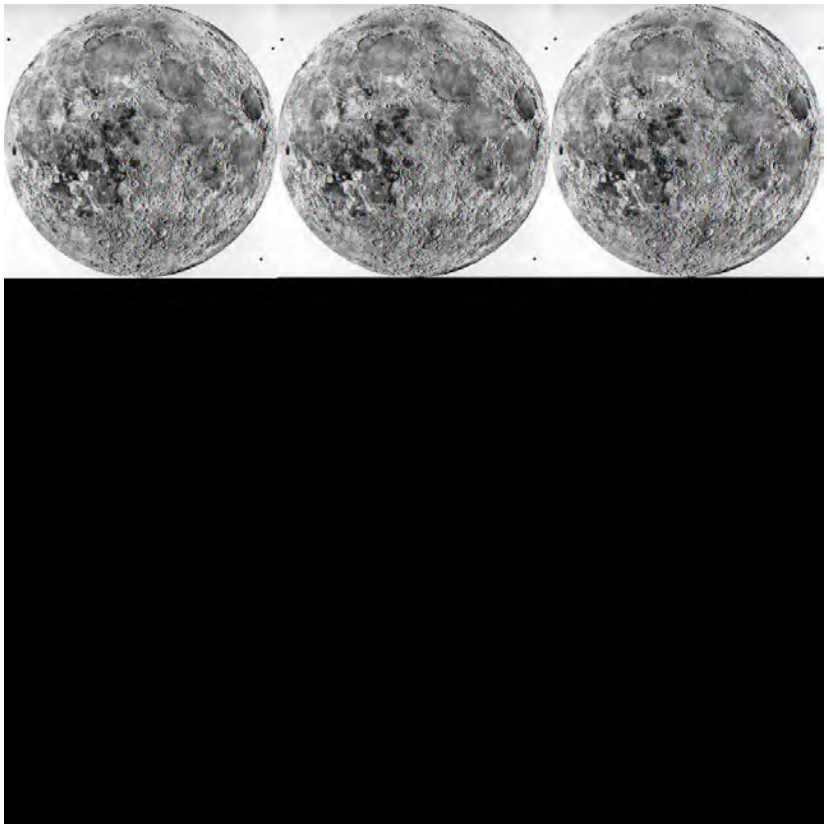
The above Lick Observatory rendition is striking in its sharpness, but in achieving the tonal balance that highlights the craters, the ray deposits from Tycho, Copernicus, and others are strongly suppressed. Tycho (85 km rim-to-rim diameter) whose very long ejecta rays are visible to the naked eye from Earth, is very conspicuous in this next full Moon image in which the lighting and processing are optimized to display these ejecta deposits. Again, Tycho is near the bottom, with a dark rim, and at the locus of the converging light rays; this implies that it formed recently enough (in the last billion years) so that the ray material has not been obliterated.



19-3: How many of the Apollo landings were in the maria; in the highlands? [ANSWER]

19-4: What would have happened to the Apollo program if the lunar surface indeed had a thick cover of powdery dust? ANSWER

This next image is a novelty that has practical value. It is an artist's painting in black and white of the lunar frontside made from a combination of telescope observations and imagery obtained from some of the lunar missions to be described. It is remarkably like the actual photos through the telescope but the lunar relief has been enhanced:



Always a spectacular event, which happens rather frequently (a few years to a decade or so), is a *lunar eclipse*. In some parts of the world a total eclipse may occur; elsewhere the eclipse is partial. A lunar eclipse takes place when the Earth is on a line between the Sun and the Moon, blocking out much of the sunlight that had been illuminating the lunar surface (a solar eclipse results when the Moon moves across the line between the Sun and the Earth). A record of a near total eclipse visible over much of the United States is this photo taken on November 8, 2003.

The Moon can change color during different seasons. The best example is the "Harvest Moon",

particularly associated with the Fall season. Then, the Moon has a distinctive orange tint. This modification from the normal yellow-white is largely the result of dust in the atmosphere scattering shorter wavelengths, owing to increased plowing as crops are harvested and fields cleared. (The appellation "Blue Moon" refers not to color but to the occasional situation when there are two full Moons in one calendar month.)

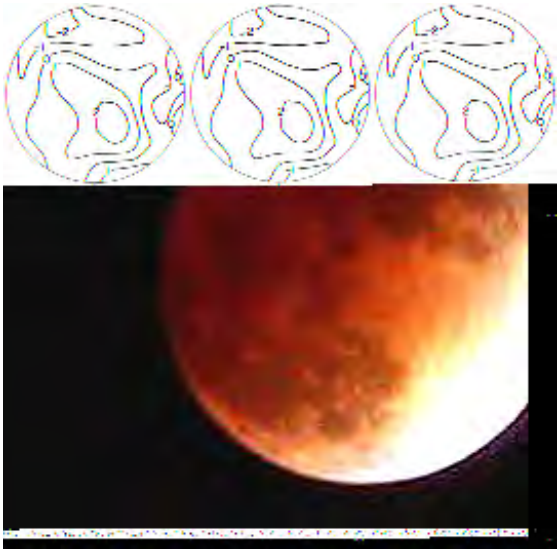
By selecting certain major features on the Moon's front face as reference points, one notes that the same face of the lunar sphere is seen during the entire period of its revolution around the Earth. This curiosity is the result of the Moon's *sidereal revolution period* of 27.3 days. The sidereal period is measured relative to a chosen star or star group. The Moon's rotation is almost exactly the same - 27.3... - as its sidereal revolution. In effect, a center point on the lunar frontside is seemingly *locked* onto the Earth, i.e., an imaginary line perpendicular to the lunar surface at that point would extend towards the Earth's center at all times during the Moon's orbital transit. This notable condition describes a state of *captured* or *synchronous* rotation. Because of this the Moon always presents approximately the same face to Earth observers. The *synodic lunar month* is 29.5 days. This increase in time by approximately 2.2 days results from the combined relative motions of both Earth and Moon forward in orbit around the Sun during which the illumination angle with respect to the Sun will be shifted; the Moon thus requires extra time for the first sliver of reflected light at New Moon to be seen from a reference point on Earth to compensate for the angular advance of the Earth-Moon system since the previous lunar month began.

19-5: What is the difference between rotation and revolution of a planetary body?
ANSWER

19-6: Most people live in the Northern Hemisphere of Earth. It is a little known fact to many such individuals that the Moon is "upside-down" in the Southern Hemisphere, with Tycho near the top and Mare Imbrium towards the bottom. Can you explain why?
ANSWER

Although the Moon is nearly a sphere, with one side always facing Earth, inspection of photographs taken on different dates discloses an apparent shift of major landmarks relative to reference point at the lunar limbs. Mapping demonstrates that almost 60% of the total surface (front and back) of the lunar sphere has been seen from Earth through telescopes. Because this seems to be a "wobbling" or "rocking back and forth", the term *libration* (from a Latin word for "balance") is applied to the phenomenon. Librations result in part from apparent (optical) displacements that bring more of both equatorial and polar regions into view. The motions involve variations in angular velocity of the Moon's revolution in elliptical orbit, tilt of its rotational axis relative to its orbital plane (inclined at 5°09' to the ecliptic), and parallax effects for different observation points on Earth. There is also a real libration owing to varying gravitational pull by Earth as the Earth-Moon distance varies and by the Moon's departure from perfect sphericity.

Like the Earth, that departure is due in large part to equatorial bulging owing to a once faster lunar rotation. In the early Earth-Moon history, the Moon was much closer to Earth [perhaps almost to the Roche Limit - the closest distance two large bodies can have stable orbits before the smaller one breaks apart from gravitational force - which for the Moon is ~2.9 x the Earth's radius or about 18000 km] but had then both a different rotation and revolution than today. Over time, tidal interaction between Earth and its satellite, which results in transfering angular momentum to the Moon, has caused 1) both the latter and Earth to slow down their rotations, and 2) the Moon to recede farther into space, expanding its orbit. However, on the Earth side (nearside or frontside; by agreement the term "backside" was declared "anathema", farside being preferred), there is an added component to the rotational bulging which occurs in the highlands (the farside surface is mostly highlands and is the manifestation of an early thicker crust). This so-called Lunar Bulge was described by G.A. Mills in 1968 as shown in this smoothed-contour illustration (the lines are in kilometer above and below a mean frontside elevation).

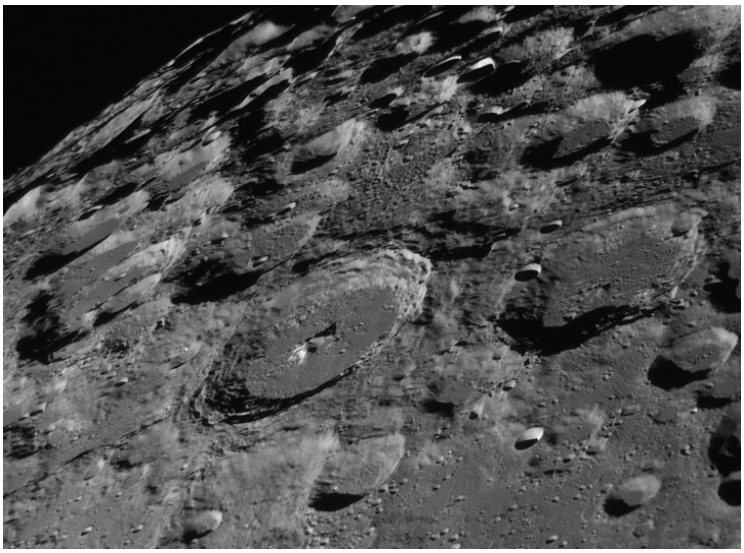


This bulge is not due to just higher crustal topography alone (note that the topographic relief is greater than 4 kilometers - almost 3 miles) that is brought out by the elevations in the flat lower maria. It appears to come from a thickening of the subcrust, with an increased density. Current interpretation holds it to have formed early in lunar history, when the Moon was much closer to Earth, as imposed by Earth's gravitational attraction that caused plastic rock to move towards the parent body. In time, the bulge was "frozen in".

Other remote sensing methods were applied to the Moon prior to the Space Age. Even before the pre-Apollo and Apollo programs, scientists examined the Moon in the multispectral mode (see a recent example on page 19-6b). The simplest approach was to use photofilters that pass limited spectral ranges. Thus, they examined the Moon in the UV, blue, green, red, IR, and other spectral intervals. This technique brought out variations in color shades that appeared to distinguish different surface units. Some maria tended to have stronger blue components, while others were more reddish. Differences between maria and highlands were accentuated. These subtle variations related to the influence of elements, such as calcium, iron, and titanium, on the behavior of reflected light at different wavelengths. Some investigators made more exacting measurements by passing the light through a spectrometer, which provided semiquantitative estimates of variations in percentages of several of the common elements. Higher values for iron, magnesium, and calcium supported the surmise that low albedo lunar rocks were most probably of basic igneous composition (silica-low, but Fe/Mg-rich).

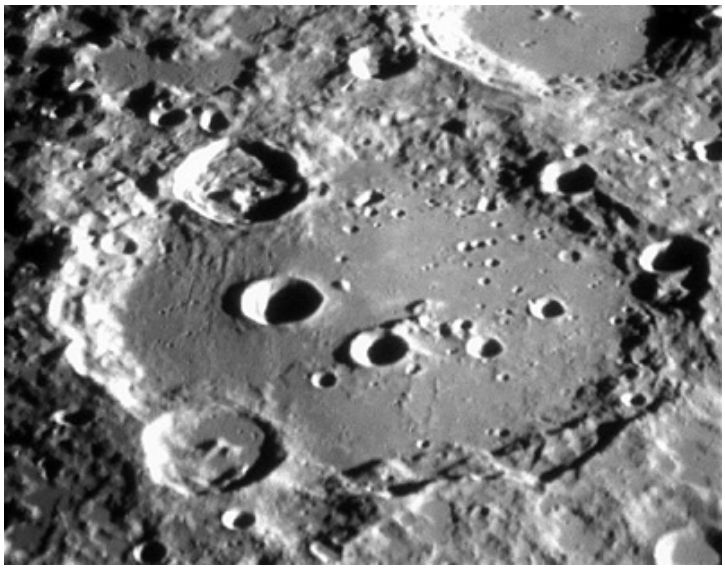
Astronauts orbiting the Moon described its color as dominantly medium to dark grays with tan overtones. We have obtained many pictures of the lunar surface, starting with unmanned pre-Apollo missions, then through numerous Apollo shots, and, afterwards, in a series of images sent back by the Galileo spacecraft as it passed Earth enroute for its rendezvous with Jupiter, and

then by the Clementine orbiter (page 19-6b). This next photo shows a typical surface of the Moon as would be seen by astronauts orbiting above. The most obvious feature is indeed the most obvious component of the lunar landscape - impact craters of various sizes:



Large craters show three distinctive features attributable to impact (discussed in more detail in Section 18): 1) A raised rim; 2) Slumped walls; 3) A central peak.

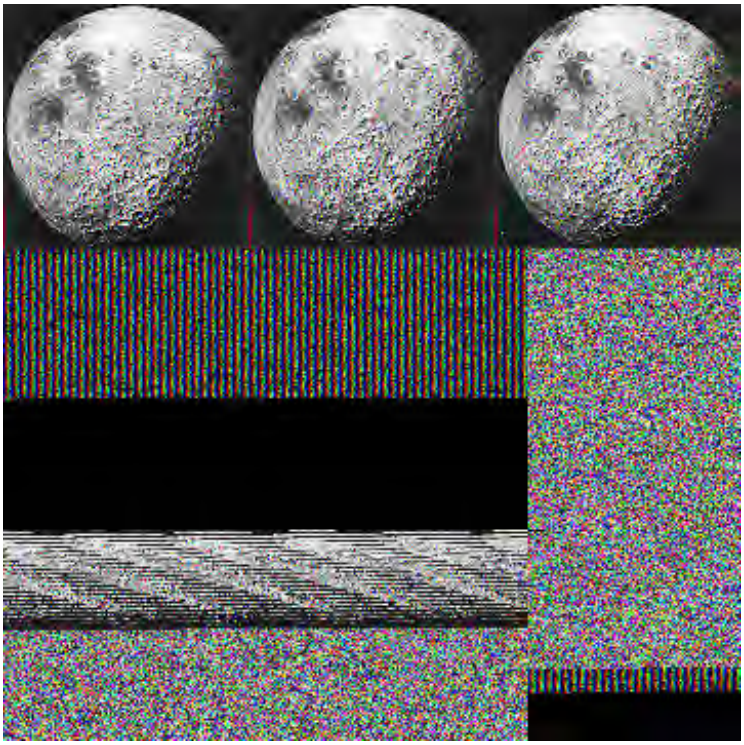
Craters come in all sizes ranging from tiny depressions less than a meter wide to basin more than 1000 km in diameter. Typical of larger craters is Clavius (diameter = 225 km) near the lunar south pole and nearby Tycho; in turn, it hosts smaller craters that were imposed on it by later impacts; its central peak has been submerged by post-impact lava infill:



Far and away, craters are the dominant geomorphic feature on the Moon. Galileo was the first to observe craters as he studied the lunar surface through his crude telescope. As later

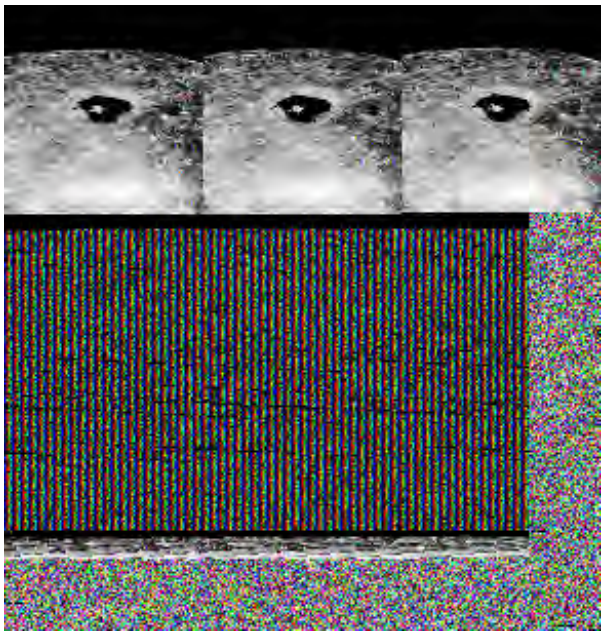
astronomers continued to look at the Moon, the explanation for the nature and origin of these craters almost universally favored volcanic activity. After all, on Earth no one suspected impact and many craters were associated with volcanoes. In 1870, the astronomer Richard Proctor first put forth arguments for impact. In 1992, the American geologist G. K. Gilbert championed impact origin for lunar craters based in part on his conclusion that Meteor Crater in Arizona was formed by a large iron meteorite. This opinion was largely ignored. Ralph Baldwin, in his book "The Face of the Moon", developed new ideas favoring impact, which gained some support. But, the belief that the dark areas on the Moon (the maria) were volcanic basalts kept the volcanoes hypothesis in the forefront. There were some scientists who passionately believed that lunar volcanism was responsible for most (nearly all) the circular depressions on the Moon. But by the start of the Apollo program in 1961, the majority of astrogeologists had come around to the impact origin for most craters on the Moon. Dr. Eugene M. Shoemaker led the way with his studies of terrestrial meteorite craters and nuclear craters. The role of impacts on the Moon was confirmed by the return of lunar samples in 1969, as will be discussed later in this Section.

First from orbiting spacecraft and then from the astronaut, humans got their first look at the lunar farside (never visible from Earth because the Moon is locked into a synchronous orbit). A typical image of the farside Full Moon, taken from the Apollo 12 Command Module as it orbited behind the earthside, shows much of the backside and part of the eastern limb seeable from Earth. Visible maria include Smythii, Marginis, and Crisium (near circular). This photo was taken when the Moon was largely dark as seen from Earth, but was between Earth and Sun

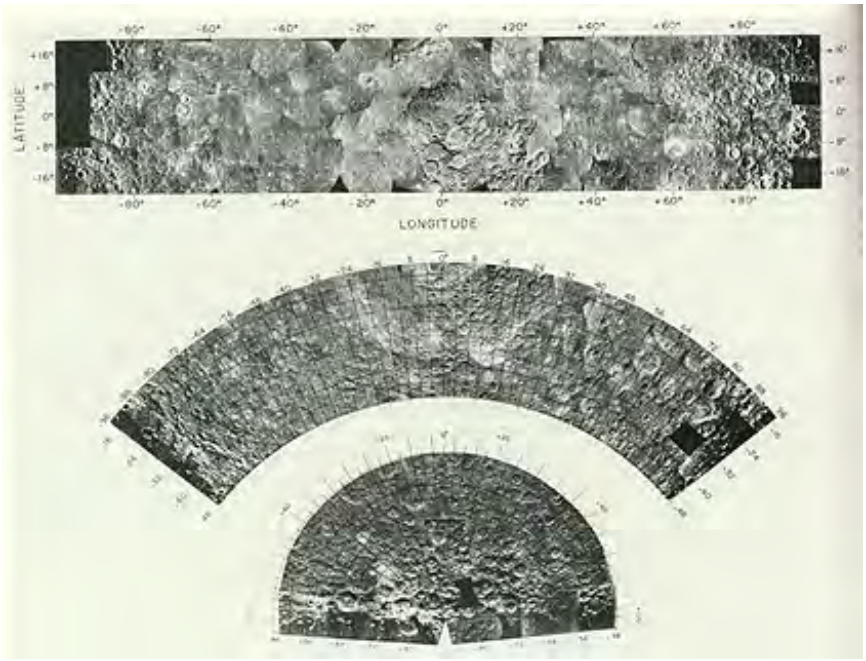


19-7: Considering the upper diagonal half of the above full Moon image to be typical of the farside, how does it differ from the full Moon view seen from Earth? ANSWER

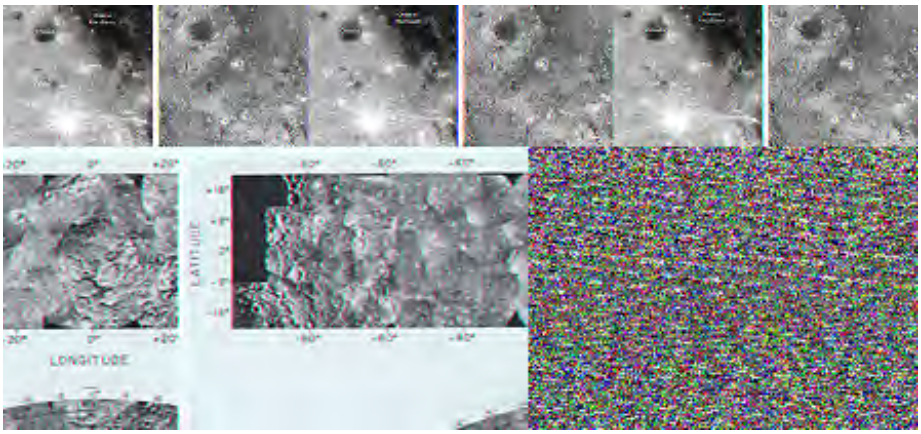
The best known crater on the far side, imaged by the Russian Lunik and Zond spacecraft, and shown here in a higher resolution photo taken by an Apollo 9 astronaut, is Tsiolkovsky (named for a Russian lunar astronomer) with its filling by dark lavas.



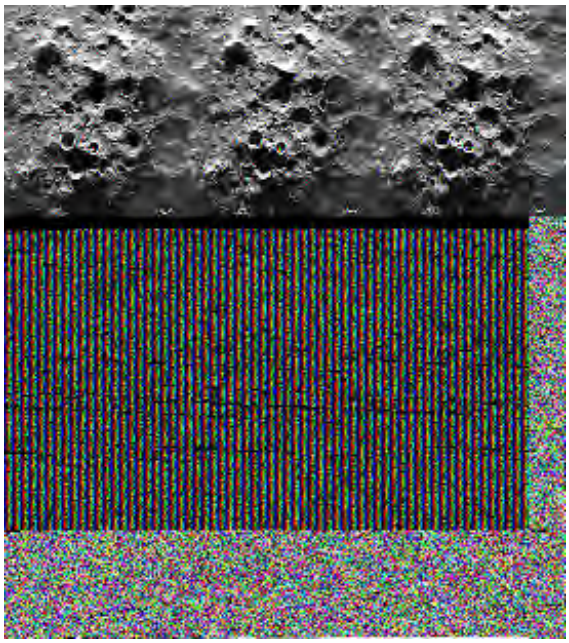
As early as 1946, radar had provided the first ever images of the full Moon and selected regions. Sent from Earth transmitters, signals returned as polarized backscatter, in which we measure time delays and Doppler frequency shifts. Here are three mosaics in Mercator (top), Lambert Conformal, and Polar Stereographic projections made from reflections of 3.8 cm pulses sent by the Haystack Observatory at the Massachusetts Institute of Technology:



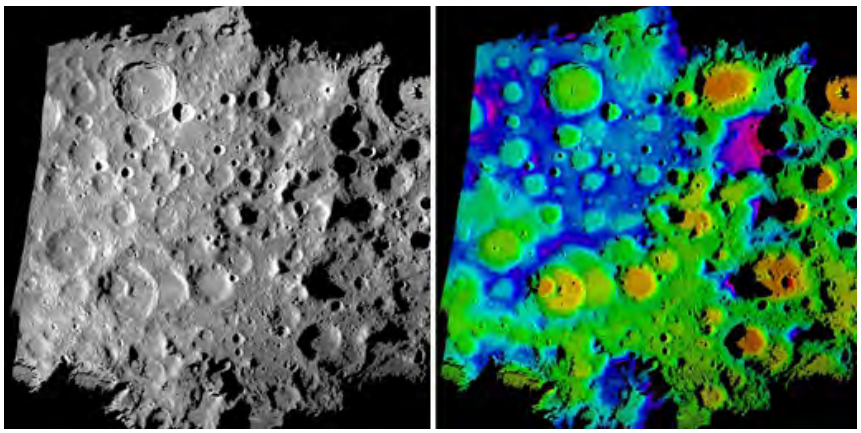
Just as was seen during our review of terrestrial remote sensing, the appearance of surface features can be quite different when optical and radar imaging are carried out. That is evident in this pair of images (visible left; radar right):



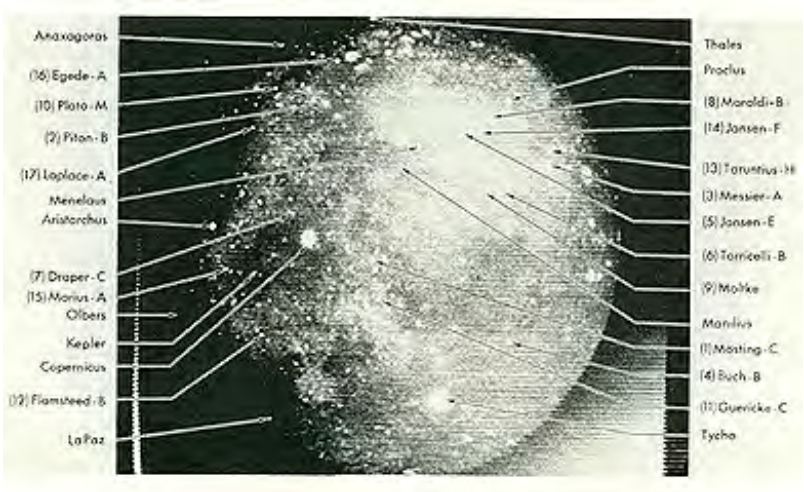
As decades past, ever better radar images were obtained from Earth. This view of the lunar south polar region, made with radar at Arecibo, confirms the statement:



Earth-based radar studies of the lunar surface continue to this day. One objective has been to use radar to map the topography of the lunar polar regions where (as explained later in this subsection on the Moon) there is a possibility of frozen water (ice) that could be extracted if a manned lunar base is eventually established. Here is a Goldstone radar mosaic of a 500 x 400 km surface around the Moon's South Pole, along with a general map of its topography (blue is low; red is high):



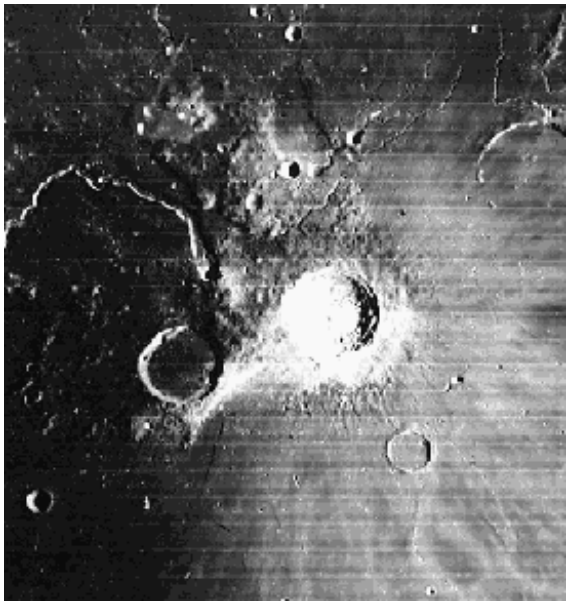
Infrared scanners, operating through ground telescopes, can acquire thermal emission images of the Moon's surface. The ideal time to sense the full Moon is during a total lunar eclipse:



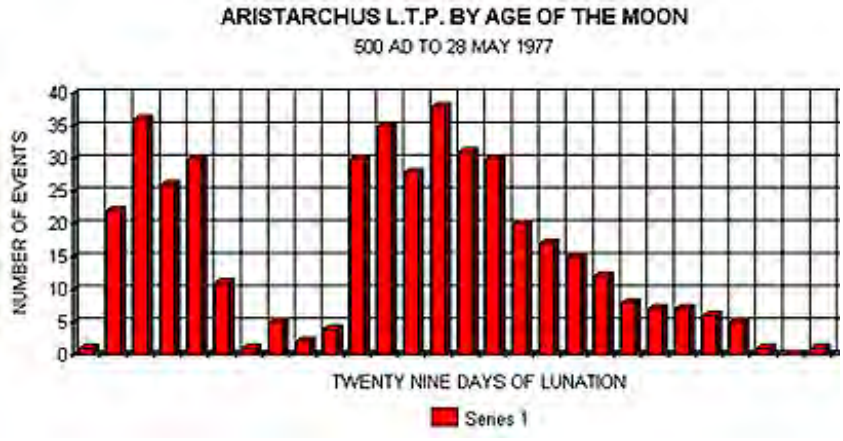
Under this condition, bright areas called "hotspots" appear, many of which, correlate with large lunar craters. These craters often contain dark lavas that absorb solar radiation (for more on this black body effect, see page 9-1) and re-emit it at higher radiant temperatures. During such times, observers also look for "lunar transients" (localized short-lived bright, often reddish glows visible to the eye) that some believe are evidence of volcanic activity and other thermal phenomena. Some people claim to see these volcanic events, even when the Moon is normally illuminated, particularly in the shadowed areas (dark phases).

19-8: Think of another possible explanation for these glowing transient phenomena. ANSWER

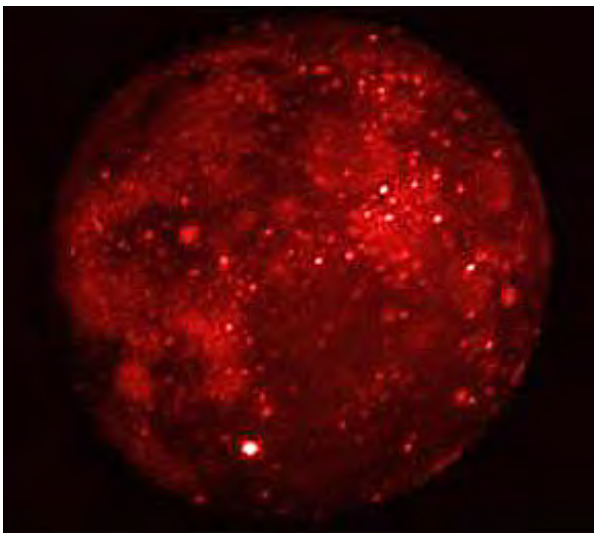
Some lunar scientists have been fascinated (almost obsessed) by lunar transients. The most frequent transient "hot spot" is the crater, Aristarchus, the brightest such feature on the Moon. Here is a Lunar Orbiter view of Aristrarchus:



Reports of events involving short-lived light emission at Aristarchus go back to the 6th Century. This plot of the days in the 29-day lunar month in which observations were recorded shows a provocative non-uniform distribution:



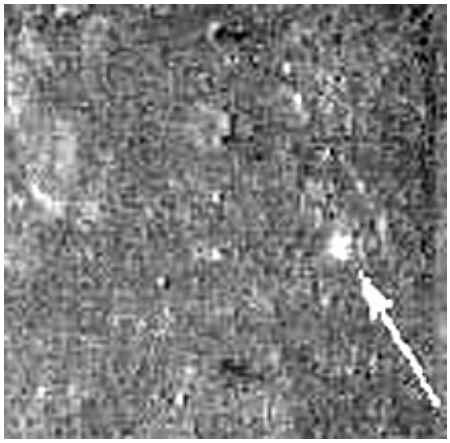
The above eclipse photo was taken more than 50 years ago from a ground telescope. In September 1999, a NASA spacecraft called MSX was in position around the Earth to image a total lunar eclipse using an IR channel. Here is the result; compare the hotspots here to those in the telescope image above.



The large bright spot is Tycho, which appears to be the hottest spot on the Moon's frontside.

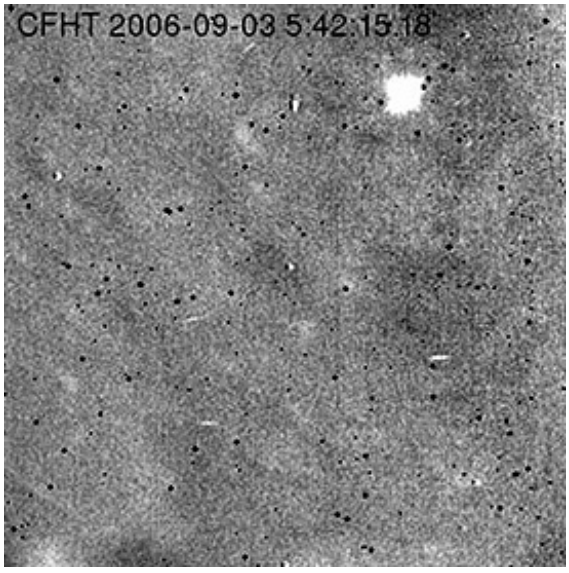
Moon watchers, especially those in the amateur ranks, have been conditioned to look for transients whenever they are observing. (Much credit for this is due to Wini Cameron, who spent most of her professional life at NASA Goddard in promoting the validity of transients). Among this category of observations is the a controversial sighting, made by an amateur, Dr. Leon Stuart, that he also photographed through his home telescope. A strong visual flash appeared to him on November 15, 1953 for just a minute or so near the lunar terminator; his photo shows this as a bright dot.

At the time he interpreted this spot to be a glow from energy released when a meteorite struck the Moon's surface. Although believable, there were doubters, and the idea remained in dispute until 2002. Interpreters of the imagery telemetered back by the spacecraft Clementine (see page 20-6b) found an image that contained the estimated point of this possible impact. Here it is:

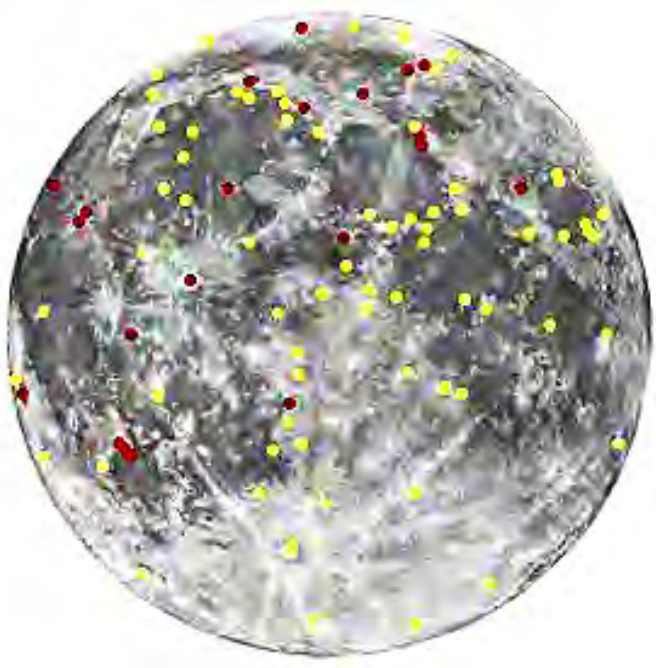


This has all the "earmarks" of a fresh crater. It is made visible by the more brightly reflecting material tossed out by the impact. From the crater size (1.5 km diameter), an incoming bolide 20 m in diameter upon collision released about 500 kilotons of energy as it dug out this now confirmed impact structure that is probably the most common type of landform on the Moon.

The detectability of a lunar impact as it happened was confirmed by the deliberate crashing of the European lunar explorer SMART-1 onto the surface. Telescopes on Earth observed a bright flash that was photographed. This bright spot greatly diminished afterwards. This is one such photo of the event, which took place on September 3, 2006:



This map shows the areas of the Moon at which a number of observations of transients make them "leaders" in the "hot spot" category:



Because most lunar observers have never seen a transient, this topic is still treated with skepticism. If transients or hot spots exist, the three most likely explanations are: 1) the light released comes from an impact, 2) there are gases emitted that are excited by particle bombardment, making them glow (usually red), and 3) some type of electrostatic discharge is involved.

This page closes with an Internet reference, to a NASA website that provides a synopsis of the Moon: Moon overview

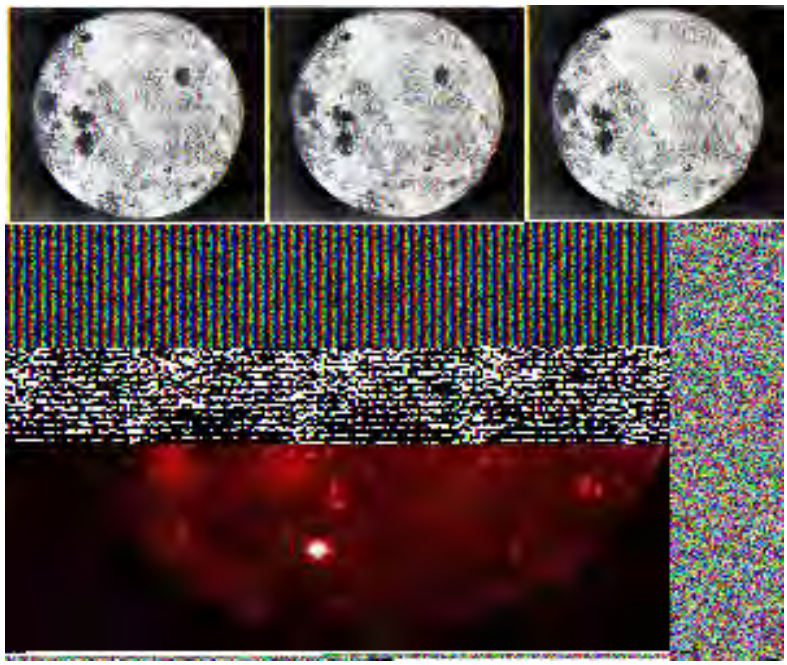
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

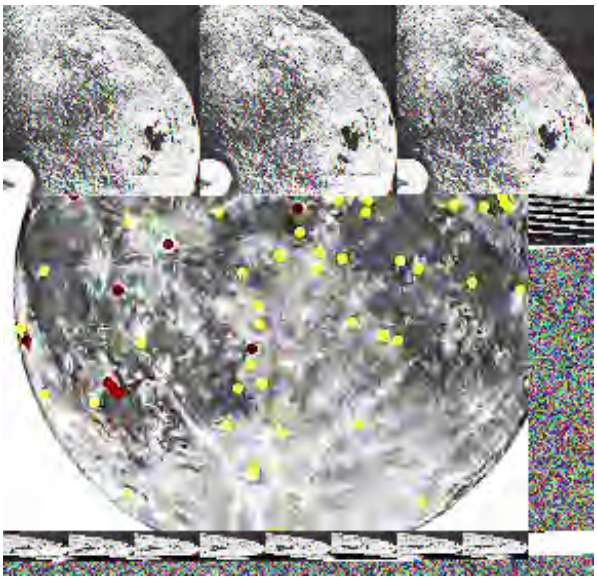
Both the U.S. and the Soviets sent spacecraft to the Moon to photograph its surface at high resolution to gather details beyond Earth-based telescopes. The Russians were the first to take photos of the farside of the Moon. The NASA Ranger series took pictures as each spacecraft approached the Moon to deliberately crash. The Lunar Orbiter satellites, using onboard photoprocessing, produced extraordinary pictures. Some examples of these are shown, highlighting both impact features and volcanism.

Early spacecraft visits to the Moon

Exploring the planets with spacecraft began with the first trip to the Moon, two years after the former Soviet Union set the space program into full motion with the launch of Sputnik I, in October 1957. In September 1959, Lunik (Luna) 2 impacted onto the lunar equator. The next month Luna 3 completed the first flyby of a planetary body, returning a series of pictures of the lunar limb and a few views of the Moon's farside, neverbefore seen by humans.



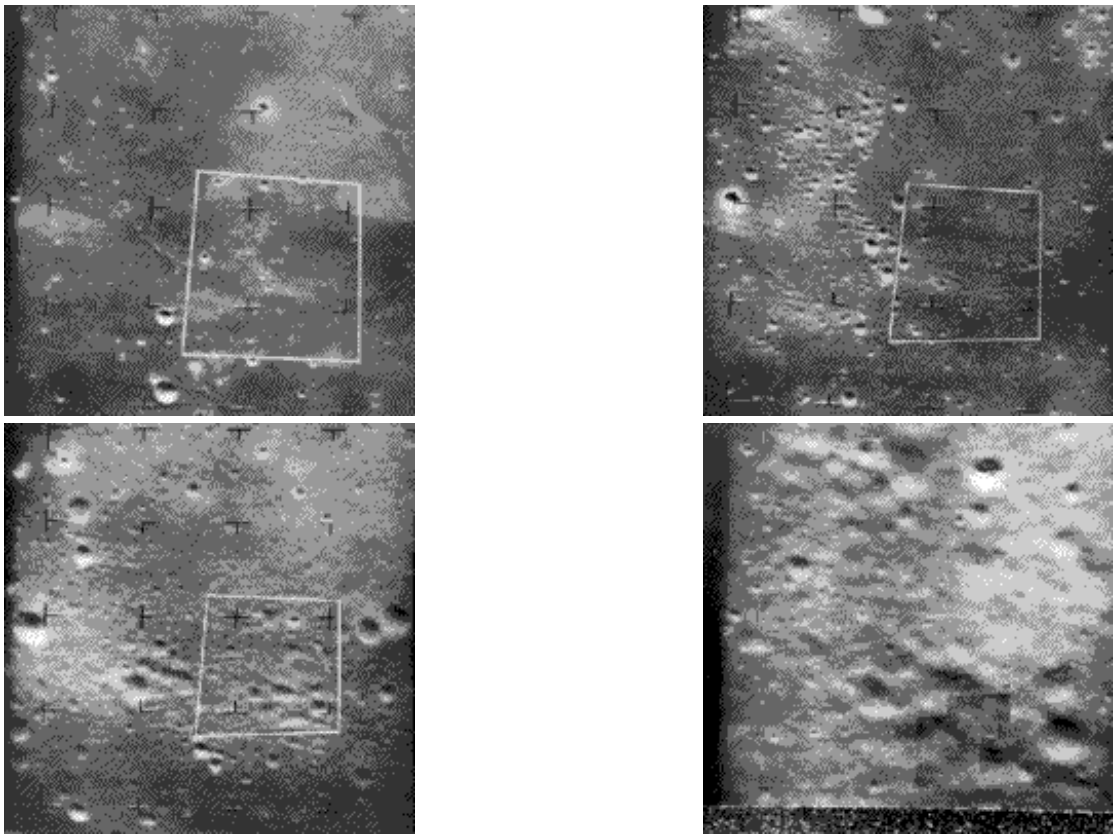
And Zond III returned images like this:



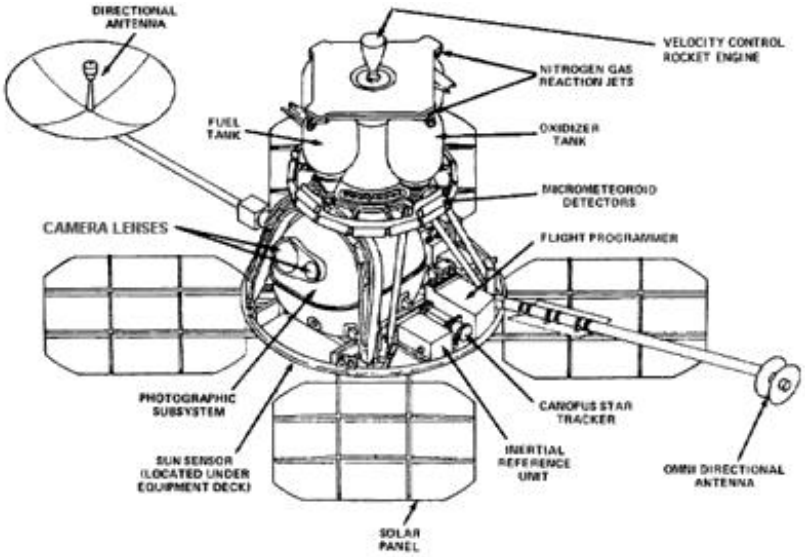
All told, the Soviet lunar program sent 20 Luna spacecraft and 5 Zond craft to the Moon; most but not all were successful. Several reached the lunar surface. Lunokhod 1 was the first unmanned roving vehicle to operate on another planetary body, landing on November 17, 1970.

The first U.S. missions to the Moon were literally a **crash program**! Between 1961 and 1965, NASA launched three Ranger spacecraft (7, 8, 9) like bullets to land (impact) onto the surface, destroying themselves in the process. Each Ranger had a television camera mounted in its nose that looked forward at the surface as it approached. They instantly telemetered progressively higher resolution images, until the moment of collision. The first image taken from Ranger 7 on July 31, 1964 was sent back when the spacecraft was 17 minutes from the lunar surface impact point. The large craters Alphonsus, Ptolemaeus, and Arzechel was evident.

The last images had resolutions of just a few meters. Here is a typical example, captured by the Ranger 7 camera. Each trapezoid frames the next scene in the succession:

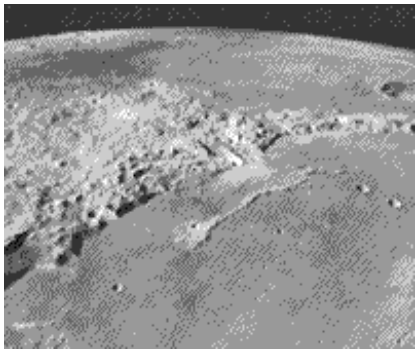


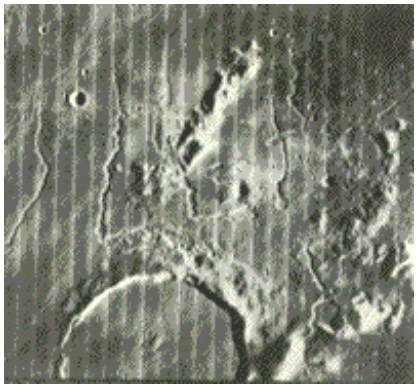
One of the hallmarks of pre-Apollo lunar exploration was the Lunar Orbiter program, in which five "photolab" spacecraft orbited the Moon between August 1966 and August 1967. The former Soviet Union was actually the first to orbit the Moon, in April of 1966, but its image returns were minimal. The first three Lunar Orbiters used orbits inclined to the equator, and the last two were in near-polar orbits. Here is a cutaway sketch of one of the Orbiters:

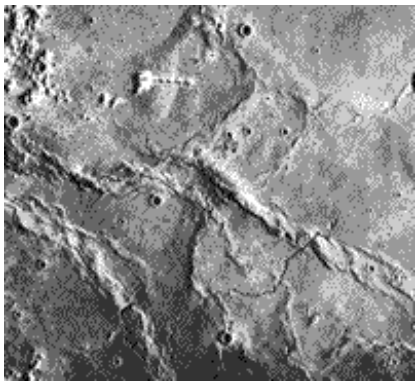


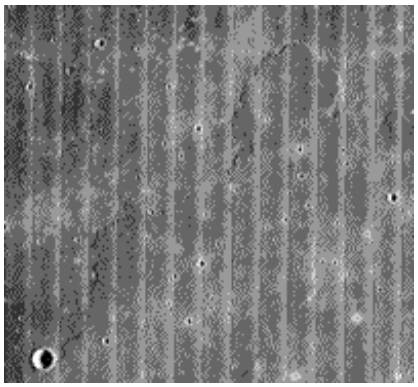
Each Orbiter had a payload consisting of two cameras holding 70 mm black and white film. One camera, with a wide angle lens, produced photos whose resolutions ranged from 8 m to 150 m (26 ft to 492 ft) depending on varying orbital altitudes. The second High Resolution camera improved these values by a factor of eight but covered much smaller areas. The photos, developed onboard as negatives, consisted of continuous strips, electronically scanned by a "flying spot" passed into a photomultiplier tube that generated an analog signal transmitted to Earth. On Earth, the signal drove an imaging device which produced 35 mm positive strips that photointerpreters joined into mosaics.

The 2,000, or so, photos from these Lunar Orbiter missions covered almost the entire lunar surface, leading to detailed geological maps that have been interpreted mainly through superposition and cross-cutting relationships (see next page). Here are four typical scenes:









In the top left is an oblique view of the Apennine Mountains, with an uplifted crust that is similar to highlands in terrain characteristics, lapped by mare lavas from younger basalts that filled Oceanus Procellarum. The right photo above shows lava plains with sinuous rilles (flow trenches)

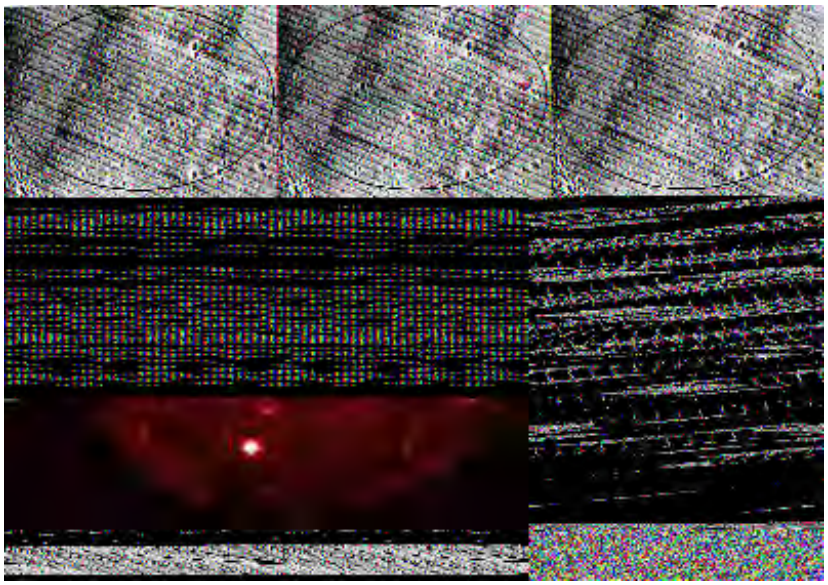
in Oceanus Procellarum and part of crater Prinz. Each strip in the mosaic is 4 km (2.5 mi) wide.

The bottom left photo shows mare ridges (squeeze-ups) and another rille. The photo at the bottom right seemingly is a bland view of a flat mare surface, except that a close look discloses an upper flow unit that is marked by a distinct front that creates a lobe marking the advance. These structures clearly signify basin-filling by multiple extrusions.

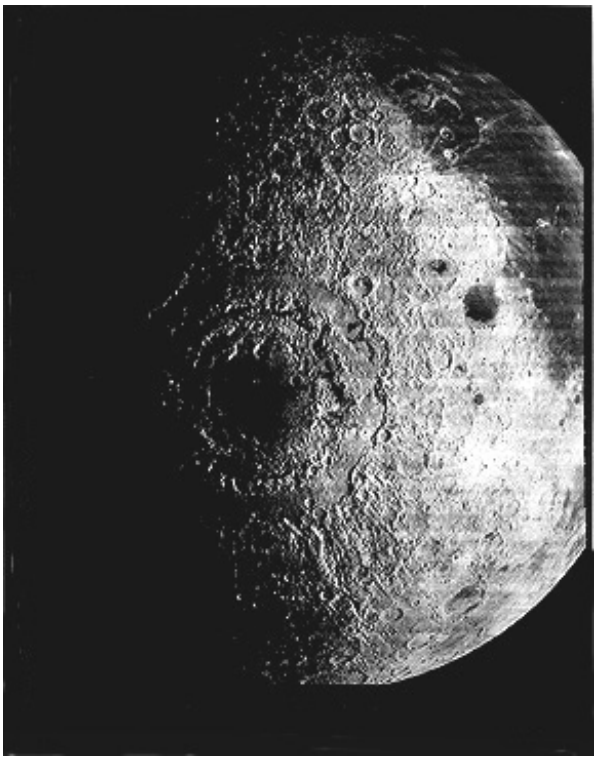
19-9: Speculate on plausible origin(s) of mare ridges and sinuous rilles. ANSWER

A primary goal for the Orbiters was high resolution photos of candidate landing sites. The scene below (5 km by 8 km, 3 miles by 5 miles) documents an elliptical site in Oceanus Procellarum, near the Apollo 12 touchdown point, considered acceptable despite a seemingly high crater density.

19-10: Would you feel comfortable trying to land the Lunar Module (LM ['Lem']) in this particular site? ANSWER

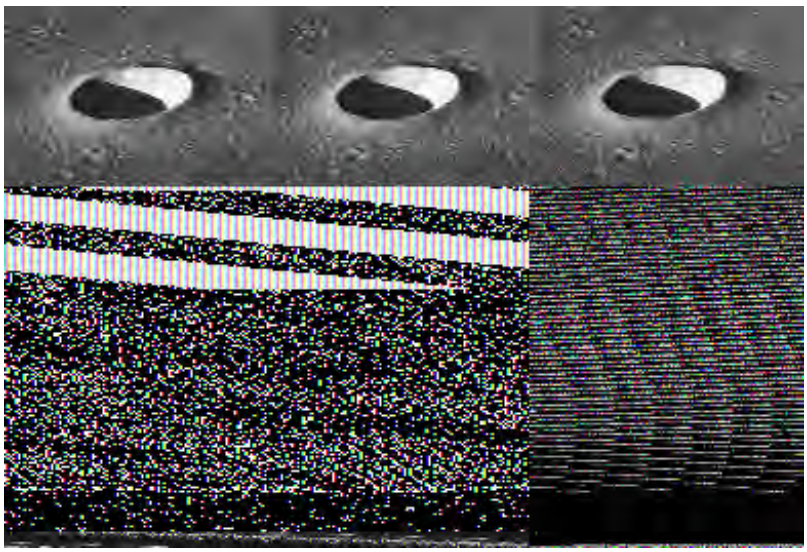


Constructed from photo strips acquired by Lunar Orbiter IV, at an average altitude of 2,700 km (1,678 miles), the next classic image shows the great lunar "bullseye," known as the Orientale Basin, a multiringed (perhaps four; the outermost with a diameter of 900 km [559 miles]) impact structure invaded by lavas (Mare Orientale). This feature lies just beyond the visible western limb of the Moon; the western Oceanus Procellarum is at the upper right.

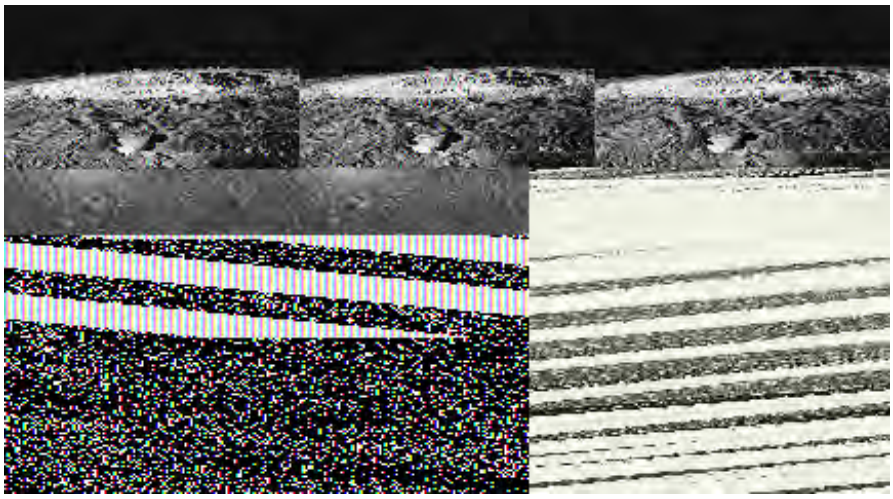


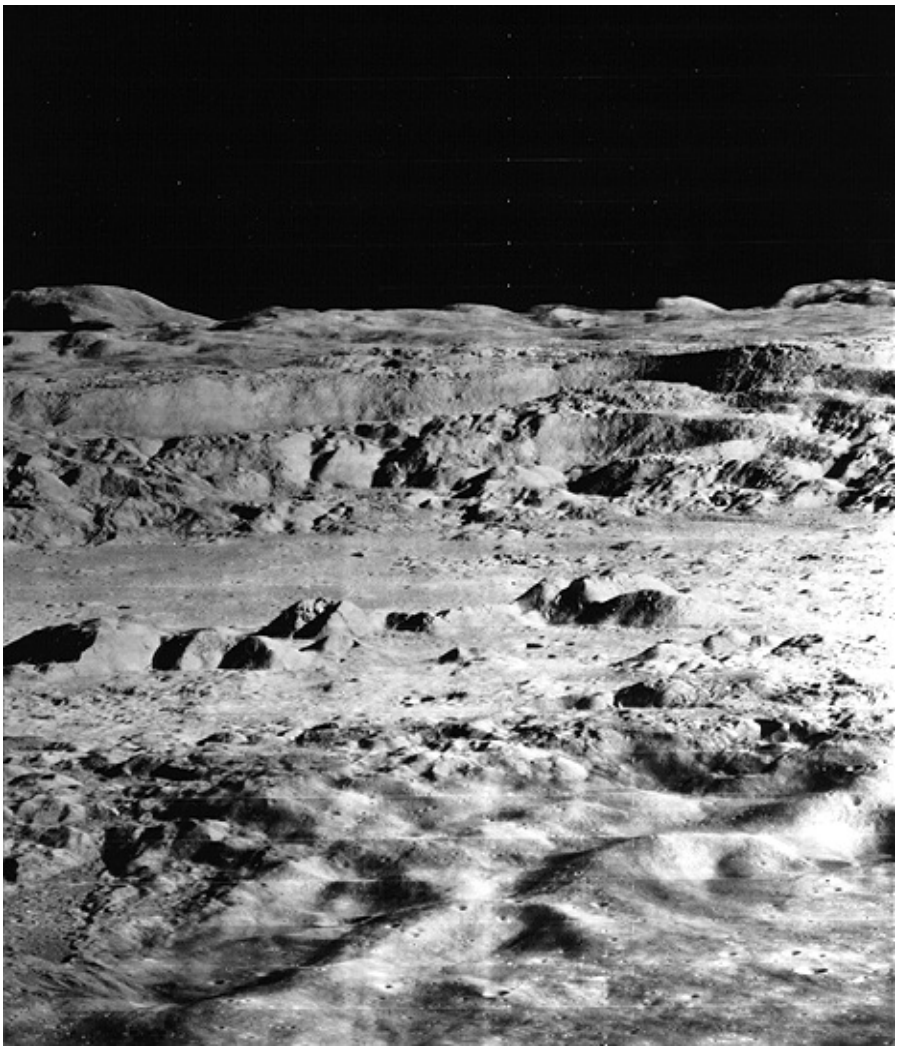
19-11: In what one way is the Orientale impact basin unlike any impact structures known on Earth? ANSWER

The Orbiters confirmed that impact cratering is a (perhaps "the") dominant process operating at the lunar surface. This next image is actually a photograph made by an astronaut from a spacecraft orbiting the Moon. This is a classic young smaller crater, with an intact rim and well-preserved ejecta beyond it.

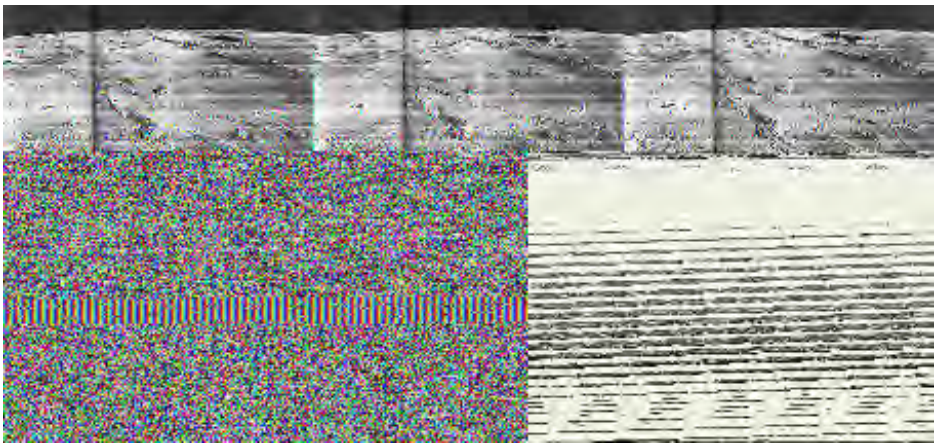


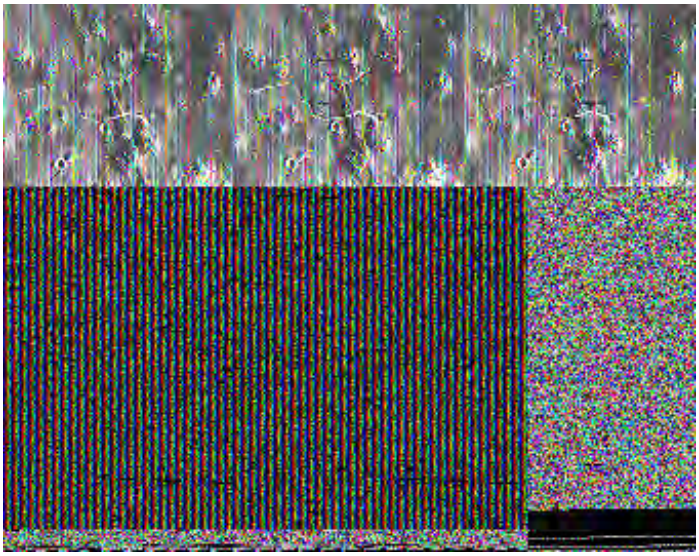
Copernicus is one of the most studied craters on the Moon. Here it is in a Lunar Orbiter II panorama. Below that is the famous close-up view of the Copernicus interior - this image was named "Picture of the Year":



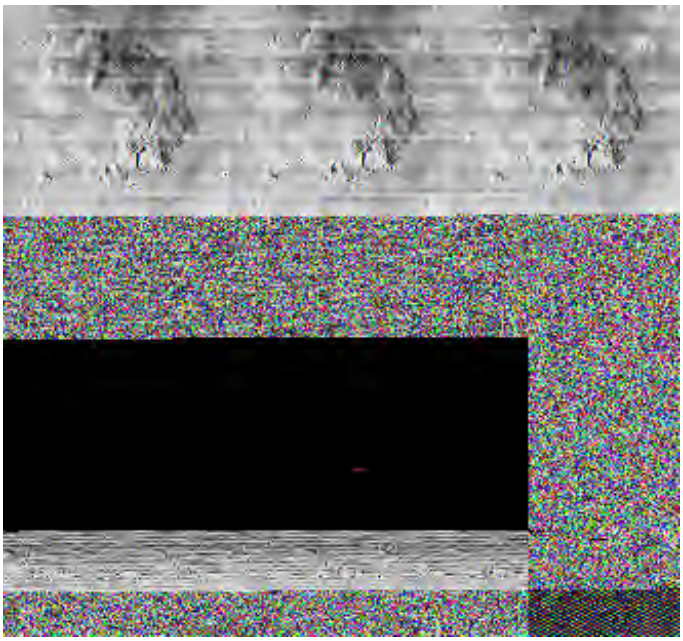


But, signs of volcanism were ubiquitous: photointerpreters indentified some volcanic calderas, lunar rilles were either flow channels or collapsed lava tubes, domes were evident, and the floors of many large impact craters contained clear indications of volcanism. Best known among confirmed volcanic features on the Moon is the Marius Hills seen in these two Orbiter views:

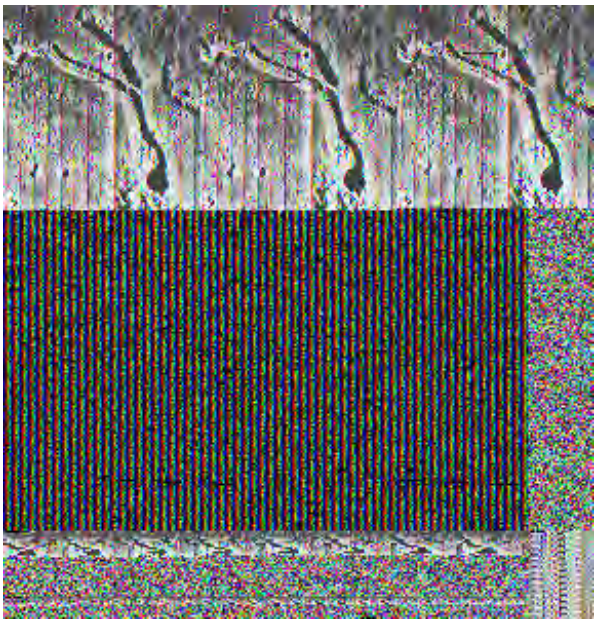




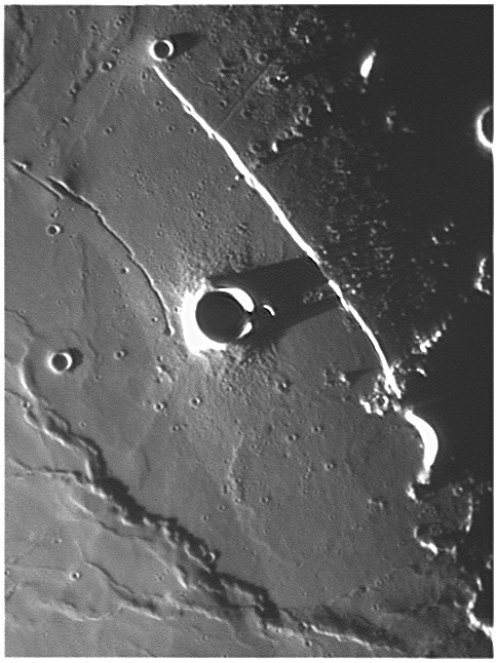
However, such constructive volcanic features are rare on the Moon. Stratocones have not been found. A few domes, such as the individuals found in the Marius Hills or the domes atop a single larger dome at the Rumker Hills (shown below), are characteristic of upwellings of basaltic lavas.



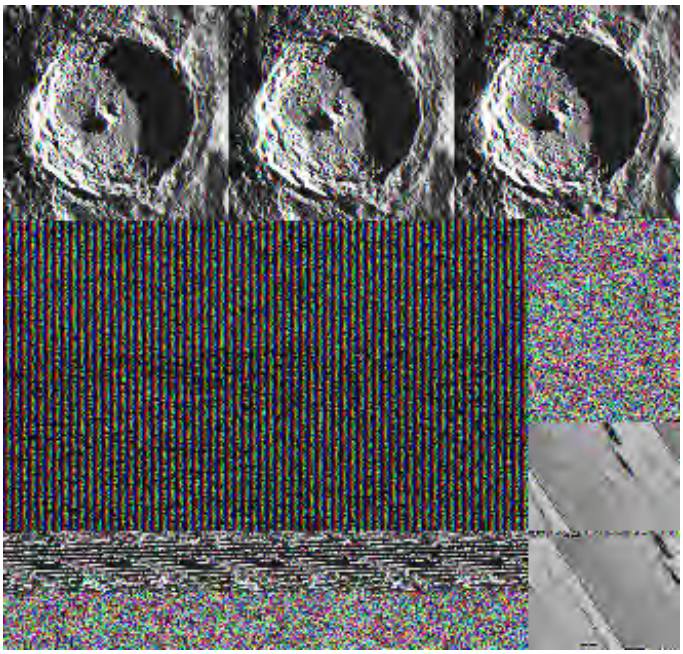
Some of the rilles on the Moon are very much like volcanic channels on Earth. The Schroeter Valley is a good example:



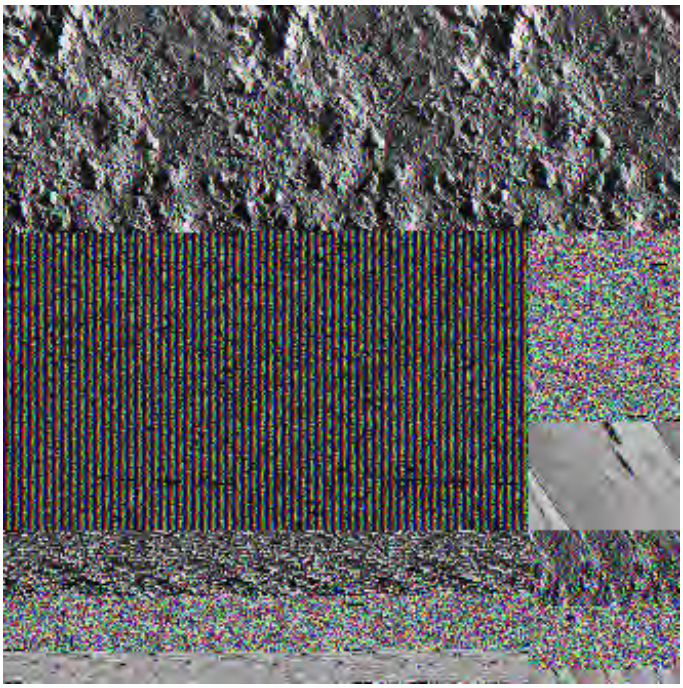
Both volcanic and tectonic features are shown in this next image which shows part of Mare Nubium. A large wrinkle ridge (lava pushup) is subparallel to a long escarpment, Rupes Recta, that is the surface expression of a fault scarp 115 km long, and up to 300 meters high:



Many large lunar craters contain structures that are almost certainly volcanic features. This Orbiter V view of the interior of Tycho (see previous page; one Full moon image shows a dark interior, the similar image below it has a bright interior, resulting from the method of photo processing) shows, on a larger scale, many features similar to those on terrestrial caldera floors:

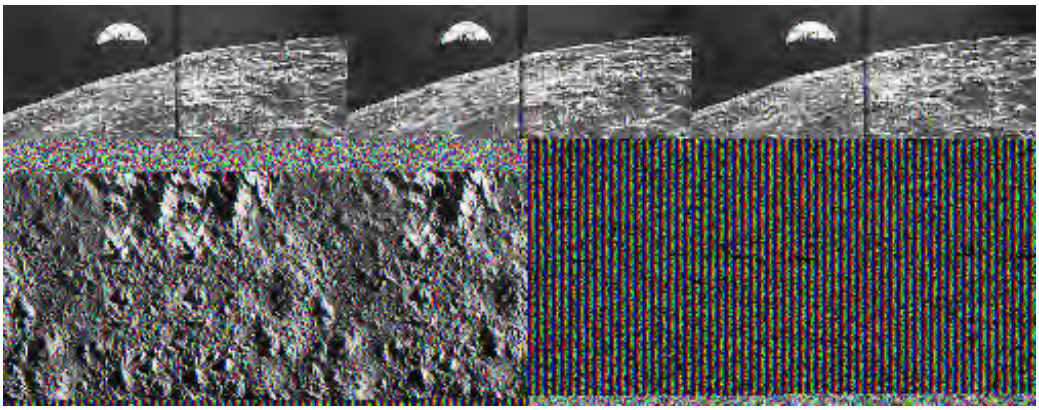


This Orbiter V close-up view of the interior of Tycho shows details of volcanic structures such as tumuli and squeeze-ups much like those on terrestrial caldera floors. such as tumuli. The lavas that enter the crater form in the lunar subcrust after the impact unloads rock from above causing hot rock below to melt and migrate to the new crater floor (much like blood flows to a surface wound):



19-12: There is one conspicuous volcanic feature on the floor of Tycho. Describe it in general terms. What is its origin? ANSWER

Lunar Orbiter I sent back a view of Earth from above the Moon's surface. This is reputed to be the very first such image of our planet from "way out":



In 2008, the data used to produce the above image of Earth were digitized and reprocessed to generate this image, showing more detail:

The Orbiter program paved the way for the eventual Apollo landings. But it was necessary - and wise - to land unmanned probes on the surface to learn first hand its properties. Hence, the Surveyors.

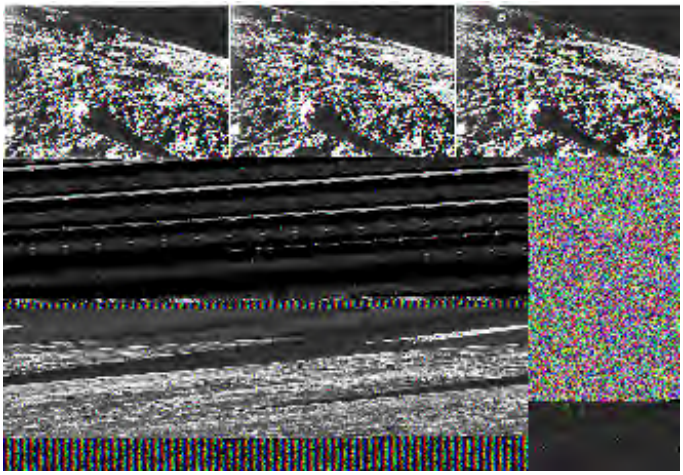
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

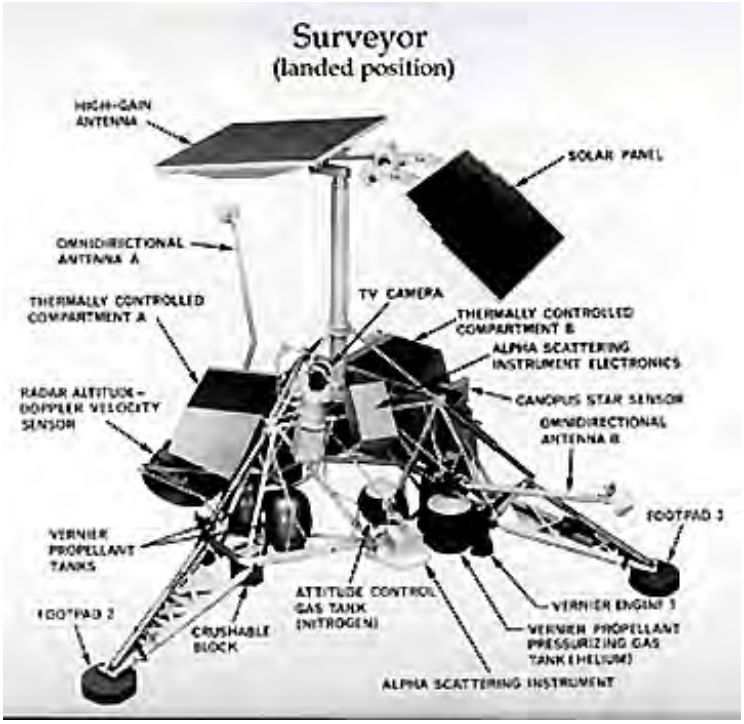
We continue with the results from a new class of moonprobes - the landers. The Soviets were the first to set down intact and transmit TV pictures of the immediate site area. The U.S. followed with its Surveyor program: Of 7 attempts to land, five succeeded. Instruments on board determined the properties of the surface materials including a partial chemical analysis. The second part of this page describes how the Moon is mapped using basic geological techniques. The objective was to establish distinctly different deposits that could be treated as stratigraphic units. The law of superposition ("the younger material lies atop older") worked well. But a new principle was also applied: Relative ages can be estimated with counts of crater frequencies: older units had more craters per unit area.

Unmanned Lunar Landers

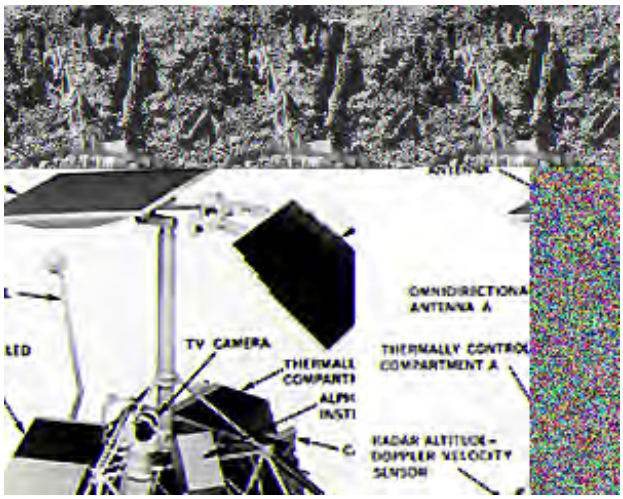
Extraordinary as these orbital photos were at the time, NASA scientists had much more to learn about the surface from a close-up vantage. Landing a viewing and sampling system on the surface was a high priority in pre-Apollo days. The former Soviet Union achieved this first with their Luna 9, in early 1966, as documented in this view of the rock-strewn surface of the mare lavas in Oceanus Procellarum.



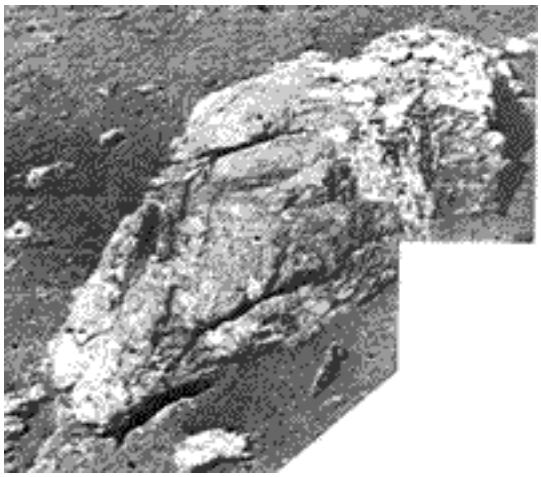
The U.S. quickly followed in May of 1966, when Surveyor 1 touched down in the western Oceanus Procellarum. Shown below, the instrument module rested on four pod-footed legs. Surveyor's TV camera provided pans of the immediate site and nearby features.



The retractable scoop could dig into the surface soil (regolith) to retrieve samples for analysis. It also answered an important question: Is the soil firm enough to support heavy landers of the future? This was visually confirmed by pictures such as this:

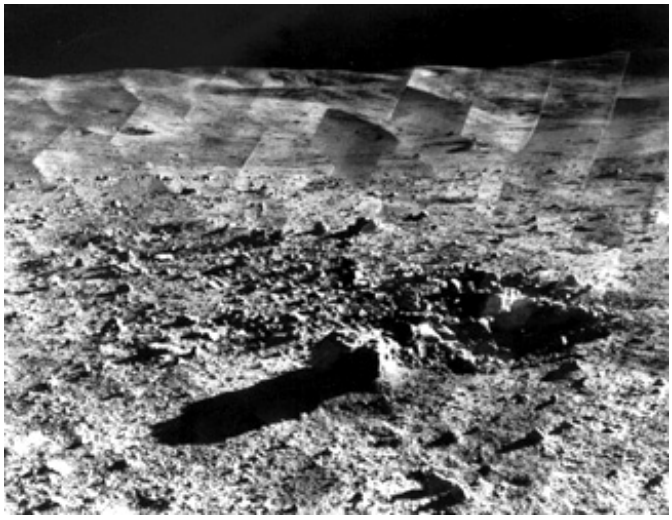


Surveyor 1 obtained close-up views of some individual larger rocks, such as shown below. This rock shows mottled dark and light patches that, based on Apollo sample returns, suggest it to be a lunar breccia containing both highlands and mare rock fragments.

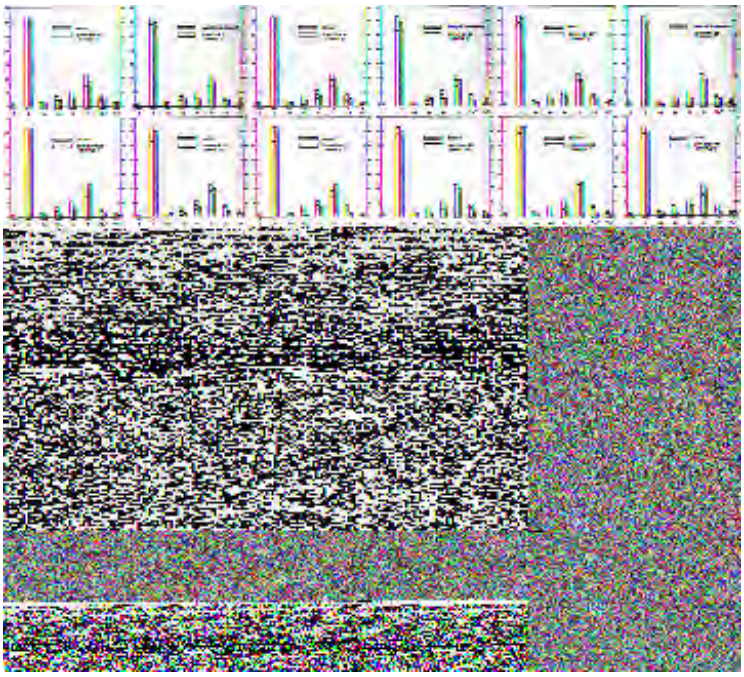


Four more Surveyors (out of seven) successfully landed. All but one set down in the lunar equatorial belt:

None of these landers sank, as feared, into weak surface materials. Surveyors 3 and 7 had an extendable pantograph arm that served as a soil mechanics surface sampler. Surveyors 5, 6, and 7 had magnets to collect metallic particles. The image mosaic below shows a panoramic view at the Surveyor 7 site:



They also carried alpha-particle backscatter instruments to sense outputs from certain elements (Fe, Ti, Na, Al, and Mg). The first solid evidence of feldspar as a primary constituent of the highlands came from Surveyor 7's analysis of its immediate vicinity. Here is a chart of chemical data for Surveyor 7's sampling of the area near Tycho:



19-13: Comparison of the VII results with these common igneous rock types show that none of the latter really fit (come close to) that of the Surveyor's analysis of highlands materials. How might this be explained? ANSWER

19-14: Summarize the principal results (most important conclusions) of the Ranger, Orbiter, and Surveyor programs. ANSWER

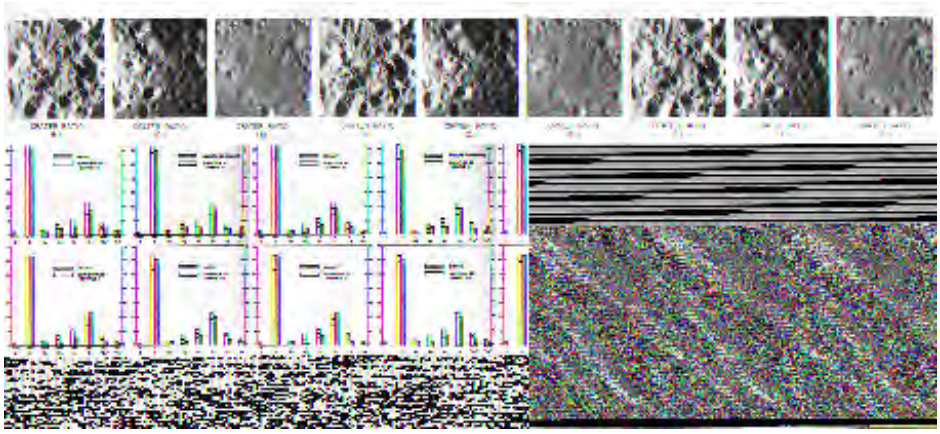
Mapping the Moon's Surface Units; Lunar Stratigraphy

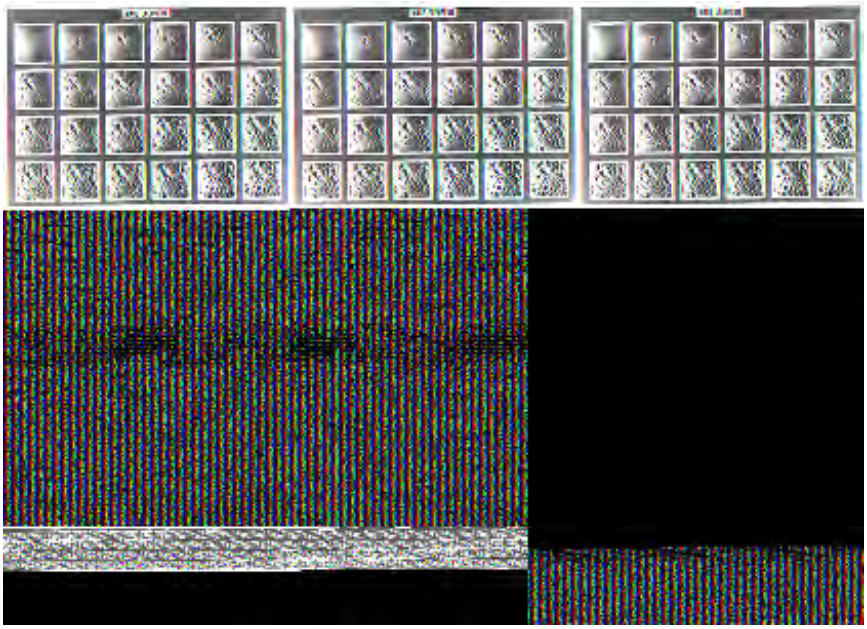
Early in the space program, when a lunar landing was the prime goal, scientists urgently needed information on the detailed morphology and other conditions on the surface. They planned and executed a series of missions that led to high quality geological maps of the entire front side of the Moon. Geologists at the U.S. Geological Survey in Flagstaff, AZ and elsewhere had already begun this mapping in the days before Lunar Orbiter, by direct telescopic observations; a noble effort that involved peering for hours through an eyepiece in a cold observatory, accompanied by sketching each item seen in the advancing fields of view.

These maps were the first with any detail for any of the other members of the Solar System. The principles used to construct them were derivatives of the procedures applied to aerial photos (except that, at the time, field checks [actual ground truthing] were not an option to validate map accuracy). They first defined units on the basis of albedos or relative reflectances, essentially by their tonal variations, as observed visually or in photos taken through the telescope. Two fundamental concepts in stratigraphy were the chief guides: 1) the law of crosscutting (features or structures cut into the terrain were younger than the materials making up that terrain); and 2) the law of superposition (deposition or extrusion emplace materials on top of the terrain later than those beneath, with each overlying layer being progressively younger than the layer(s) immediately below). The most common exception: conditions where structural deformation places older units on top of younger ones at thrust faults or where a fold lies on its side do not seem to occur on the Moon; there is no evidence for the compressive forces needed to produce these effects.

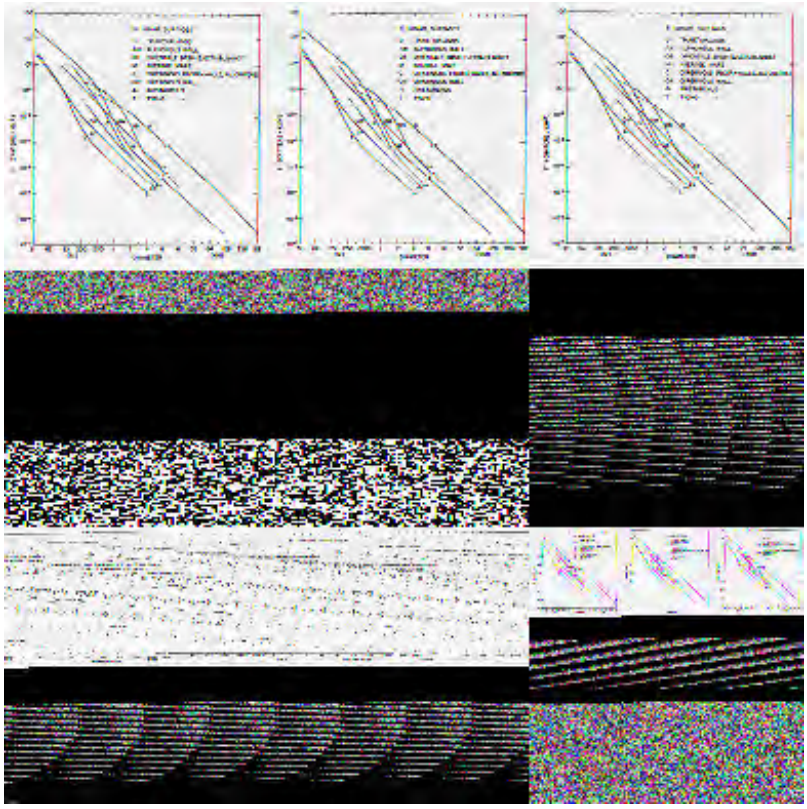
However, scientists developed a new concept, using crater frequency and morphology, specifically for the Moon as a major tool in determining relative ages. Sparsity of craters on Earth preclude its use on our planet but crater "dating" works well on the planetary bodies imaged by missions beyond the Moon. The notion underlying crater frequency (or density, i.e., numbers of craters of all discernible sizes within some arbitrary counting area [typically 10^6 km^2]) is straightforward.

The simple idea is this: as time goes on, impacts will create more craters within the area, until they occupy all the space, i.e., the surface is saturated. In this model, older surfaces have more craters, and thus a higher frequency of craters of all sizes. Or, restated, assuming a continuing flux of impactors over time, young surfaces have few craters; old surfaces have many, The number of larger craters diminishes with size, and there are ever greater numbers of progressively smaller craters. However, over time the surface can be saturated with craters, i.e., there is no room to fit new ones and continued bombardment only stirs up earlier ones. These two illustrations should help to visualize this; read their captions for brief explanations:





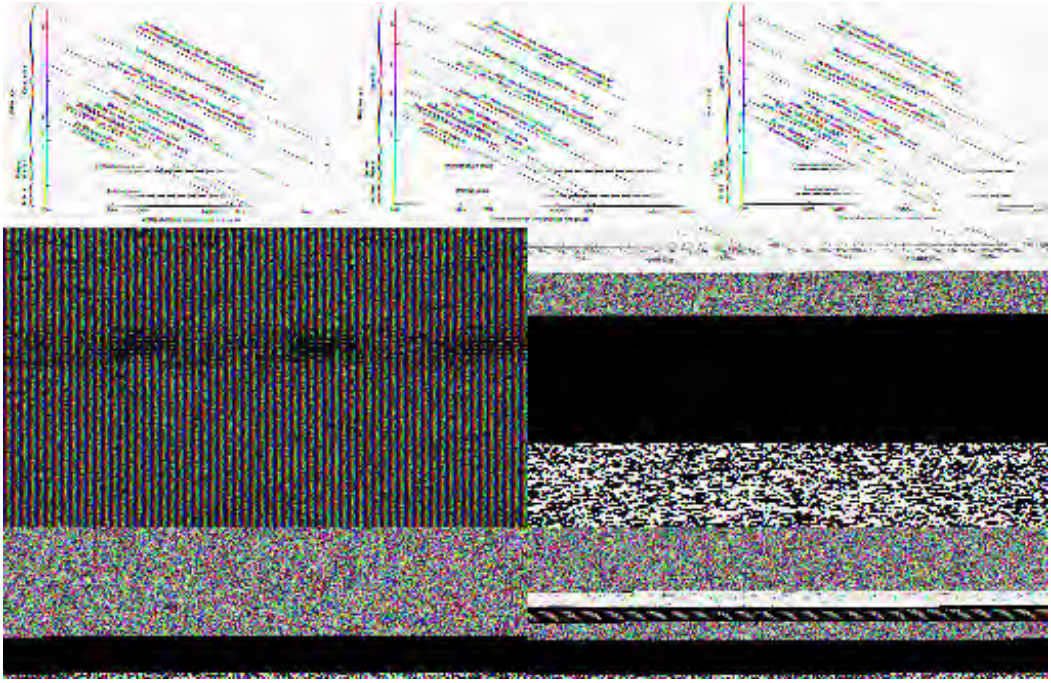
This relationship is evident in this plot of crater frequencies for eight areas of the Moon, whose relative ages are expressed by the position of each plot.



The pure uplands surface (curve U) is proposed as the oldest, whereas those associated with the interiors of Aristarchus (A) and Tycho (T) are the youngest. This means that for any given crater size (say, 250 m [812 ft] diameter), there are more craters per unit area (extrapolating to the ordinate or left vertical axis) in the uplands (about 10^{-1}) than in the maria (M) and, in turn, more mare craters than inside Tycho T (about 10^{-3}). Most of the plots are nearly parallel to one another, indicating the same relative proportions of craters at all size ranges. Scientists confirmed these relative surface ages, generalized here as decreasing to the left of the plot, by radiometric age dating of Apollo lunar samples collected from terrains whose ages had been estimated earlier by this crater-frequency counting method.

19-15: Why does the highlands or uplands (U) plot stop abruptly at a crater diameter of about 500 meters? ANSWER

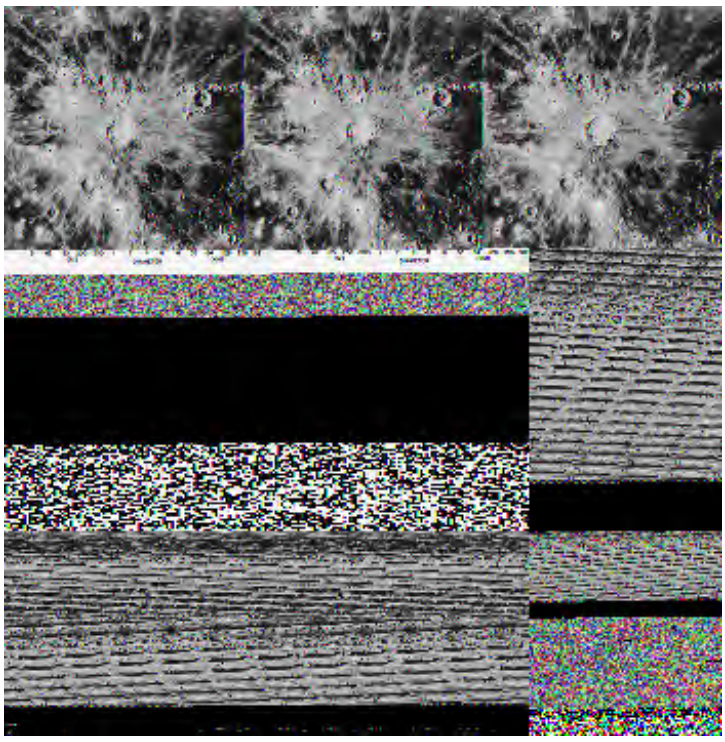
The next plot is a refinement of this technique that relies on modifications of crater shapes and degrees of degradation.



The relative age decreases upward on the ordinate; i.e., the youngest craters are at the top. A fresh crater should be characterized by a sharply defined rim, well-preserved interior terraces, a distinct external ejecta blanket, and, if large, persistence of secondary rays. The degree of degradation depends on the length of time since its formation and on its size. Consider the category s*trongly subdued*. If such a crater were 50 m (164 ft) in diameter, it would belong to the Copernican class (see the stratigraphic classification below), falling at about 3.8 (in relative age units; **not** in billions of years) on the age axis. If it were 500 m (1640 ft) wide, it would belong in the Erastothenian age group.

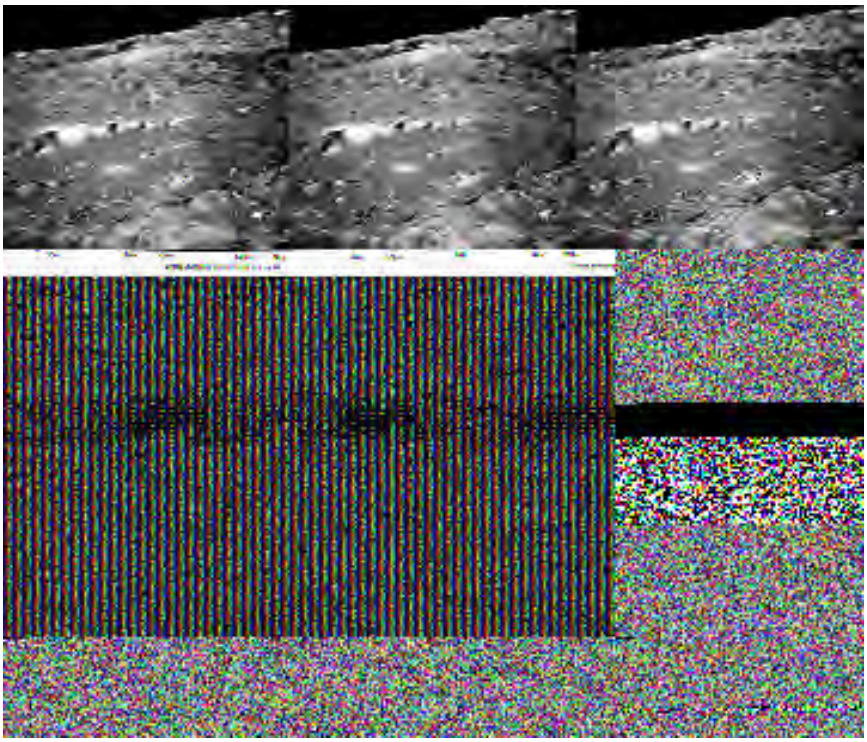
19-16: Looking at the inclined dashed lines, what is happening as you go from lower left to upper right? ANSWER

The classic example of lunar geologic mapping is the region around the great crater Copernicus. Look at this Earth-based telephoto image of Copernicus and its surroundings:

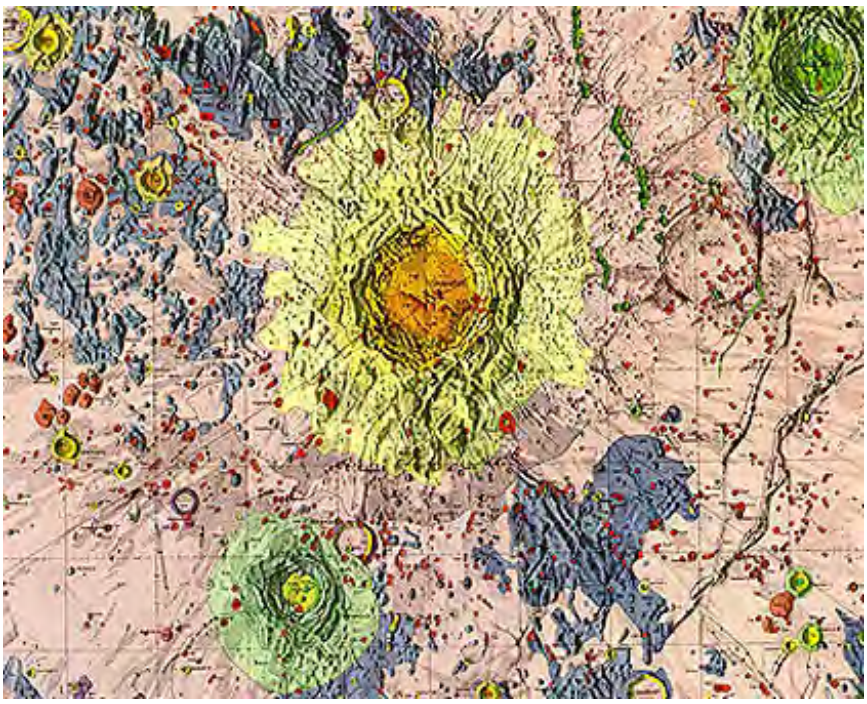


19-17: How many distinct (different) mappable units can you pick out in this photograph? ANSWER

The most obvious feature is the crater with its slumped walls, central peak, and the ejecta blanket and rays that extend beyond. This crater is considered the "type locality" for impact structures on the Moon. Apollo 11 astronauts photographed the interior and rim of Copernicus as a major target of interest.



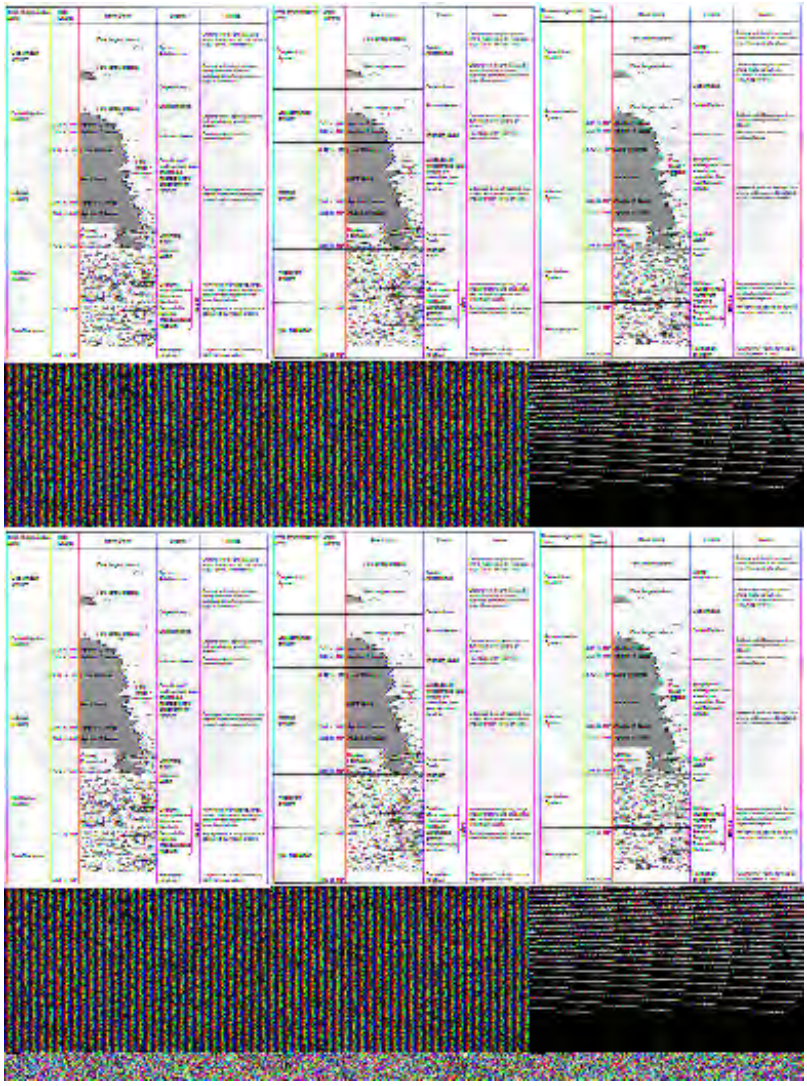
Now, examine the color-coded map made by astrogeologists from the U.S. Geological Survey (USGS) Branch of Astrogeology (Flagstaff, AZ), largely by long hours of observations through telescopes:



The oldest units are in *blues* and *purples*, being highlands-like terrain, consisting of ejecta blocks tossed over the lunar landscape from the impact of an asteroid that formed the Imbrium Basin to the north. The mare lavas (*lavender-pink*) filling the edge of Oceanus Procellarum largely "submerge" these units, known as the Fra Mauro formation. Small *red-orange* patches are lava domes. Craters cut these lava plains, including Erastothenes (the large one northeast of Copernicus, with its well-defined hummocky ejecta blanket), rendered in *green*, but one crater,

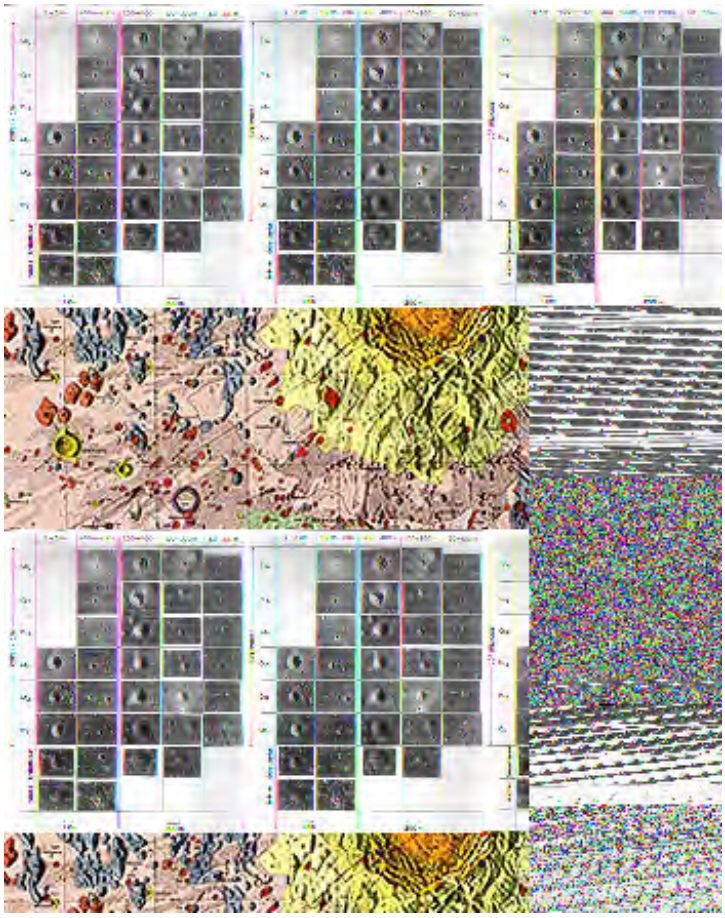
Stadius, to the right of Copernicus, is older, being largely covered by lavas, so only a faint outline remains. Ejecta (*yellow*) surround Copernicus and a younger central fill (*orange*) is basaltic lava. Small *red* units are dark-halo craters of probable volcanic origin. Many of these relationships are difficult to see in the photograph but are visible to the geologists peering through their telescopes.

General mapping of the lunar surface in this way, re-enforced by radiometric age dating of returned Moon rocks, has led to a system of lunar units classification that follows the conventional stratigraphic approach used on Earth. Thus, scientists assigned period names to time intervals and relatively ranked all the events within these spans. For distinct units, they gave formation names. The stratigraphic column depicted in Earthlike Planets (by Murray, Malin, and Greeley, 1981) summarizes these classes. In our version, we replace period names by their equivalents in System (time and stratigraphic) terms. The Copernican System/Period began about 1.1 billion years ago, based on radiometric dating of ejecta from Copernicus.

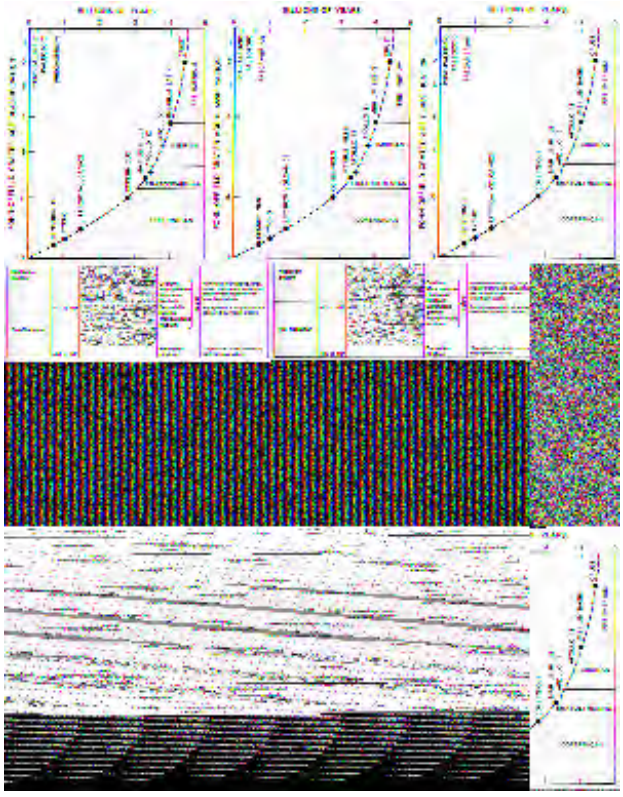


19-18: Why are there specific dates associated with some intervals in this stratigraphic column and not with others for which particular events are defined? ANSWER

N. Trask (USGS) has put together this next illustration to show the relative degradation of craters over time as a function of size (follow the 200-400 m group as a guide); on the left are symbols related to age nomenclature defined above:

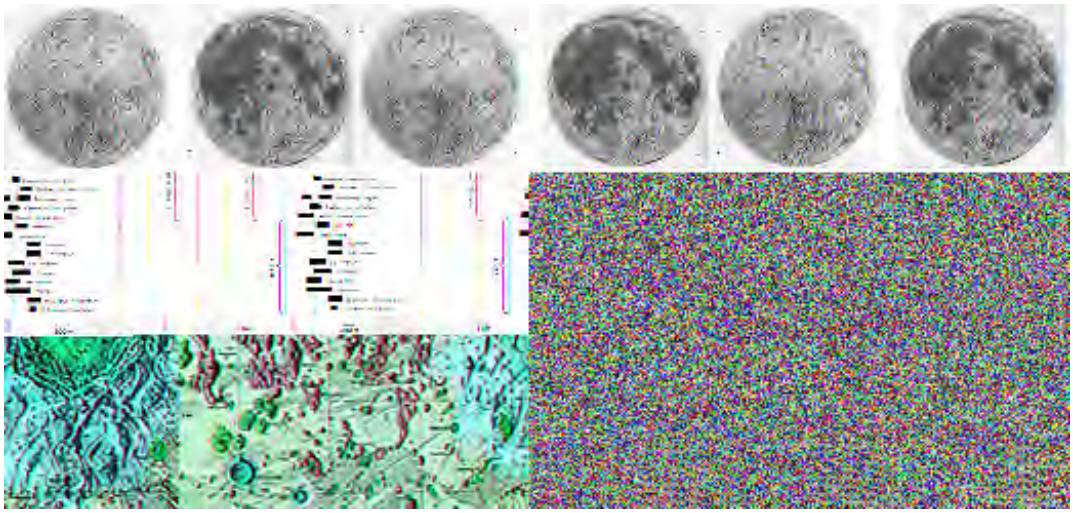


H. Pohn and T. Offield (USGS) have constructed a general curve relating their interpretations of the degree of crater modification (freshness versus degradation) to the stratigraphic time units established in the above diagram. They have used Apollo sample radiometric dates to calibrate the position of the curve to the time intervals involved. They used crater frequency surfaces adjusted to size ranges in deducing the ages of different regions.



Pohn and Offield also produced a more detailed breakdown of relative ages for additional specific features (the black bars are measures of uncertainty). again using crater morphology (age of a region is controlled the oldest craters present) :

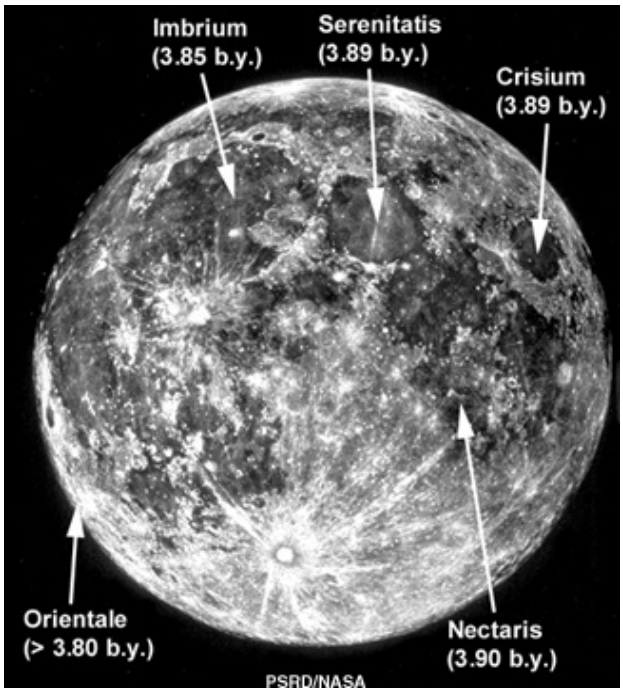
In this last diagram, three major events are indicated: 1) imposition of large basins and deposition of regional ejecta blankets; 2) evolution of lunar plains and extensive highlands cratering; 3) emplacement of the maria. More generally, the history of the Moon falls into four major phases or stages: 1) formation, general melting, and development by differentiation of a lighter crust; 2) intense bombardment of that crust, producing most of the larger basins and many craters still surviving in the Highlands; 3) widespread extrusion of basaltic lavas (about 3.8 to 4.0 billion years ago), filling basins, some larger craters, and low areas embayed against or within the Highlands; and 4) reduced but still common additional impacts (such as Copernicus) on both solidified maria and highlands, continuing to the present-day (Tycho is likely less than a billion years old). Stages 2 and 3 have been visualized by an artist in the following diagram:



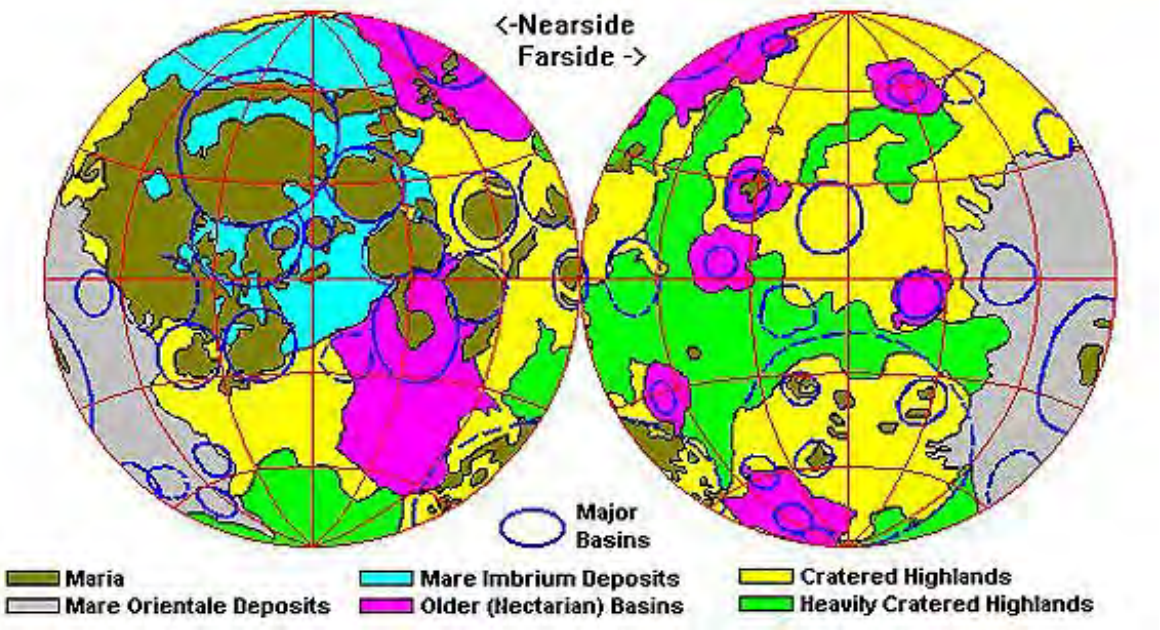
Of special interest is the multi-ringed (successive rims) of the Imbrium basin (upper left), similar to the Orientale basin (not shown); these rings may be partially submerged by later mare lavas.

The ages of the major impact basins have now been established, using radiometric dating methods on ejecta (as rocks collected during Apollo missions) traced to each basin. The near identical age values indicates that the basins were formed almost simultaneously during a brief

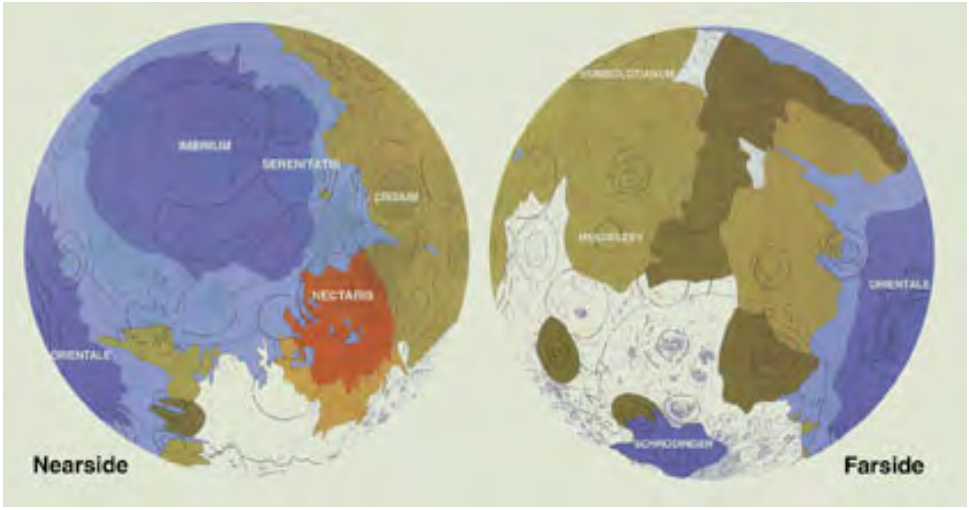
period of intense bombardment by asteroids.



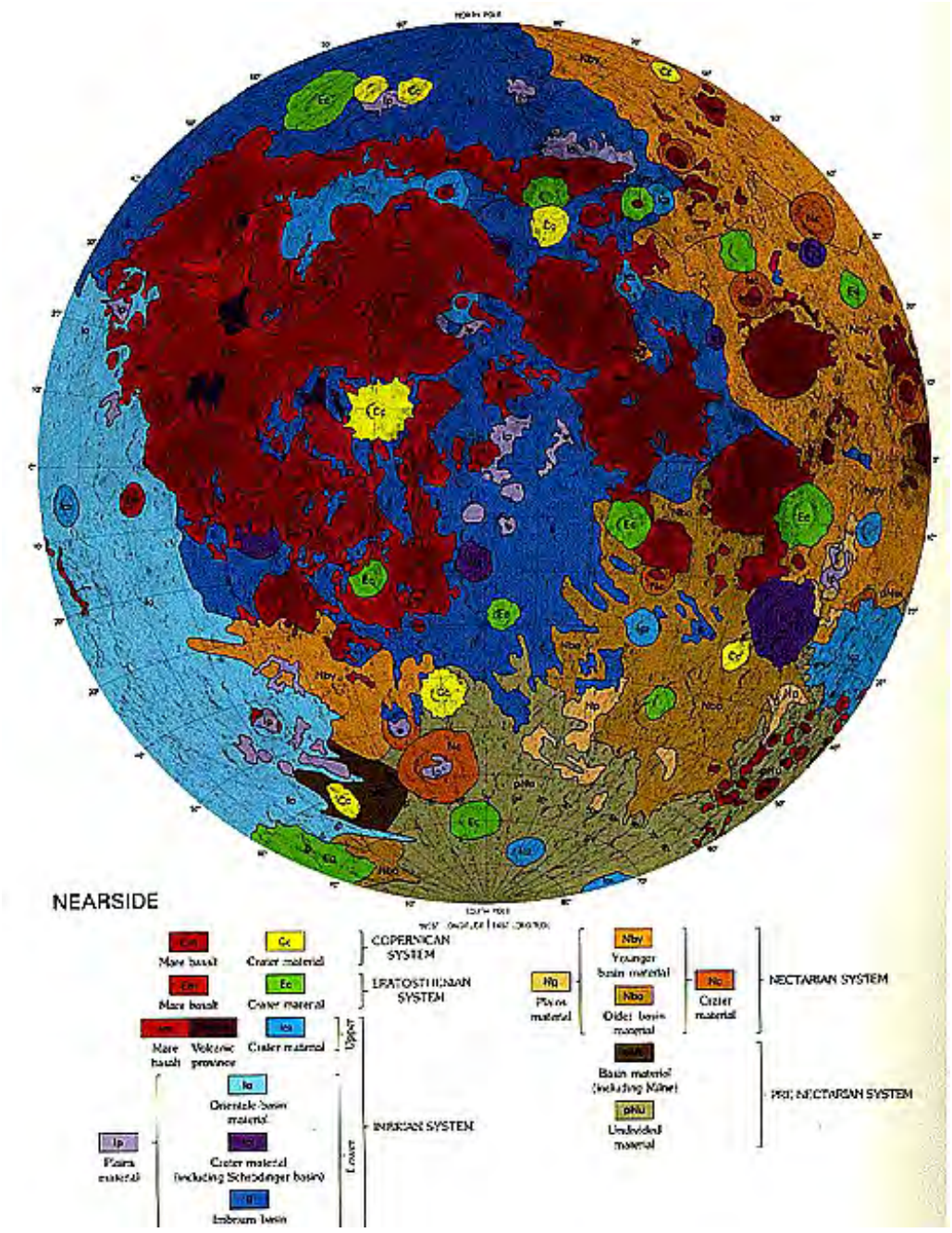
The geology of both near and far sides of the Moon can be generalized in terms of the types of units (cratered; ejecta-covered; mare lava-filled) at the surface. This map shows the results (youngest units in upper left of legend, oldest, lower right):



This next illustration shows two maps (near and far sides of the Moon) in which several major stratigraphic units are distributed. These are: Pre-Nectarian - dark brown; Nectarian - brown and red-brown; Imbrian - medium and dark blue:



We close this page with the published U.S. Geologic Survey map of the front or near side of the Moon. The legend with named units and time periods is below it. To get your bearings, the yellow patch near upper left center is Copernicus; the yellow patch near the bottom is Tycho.

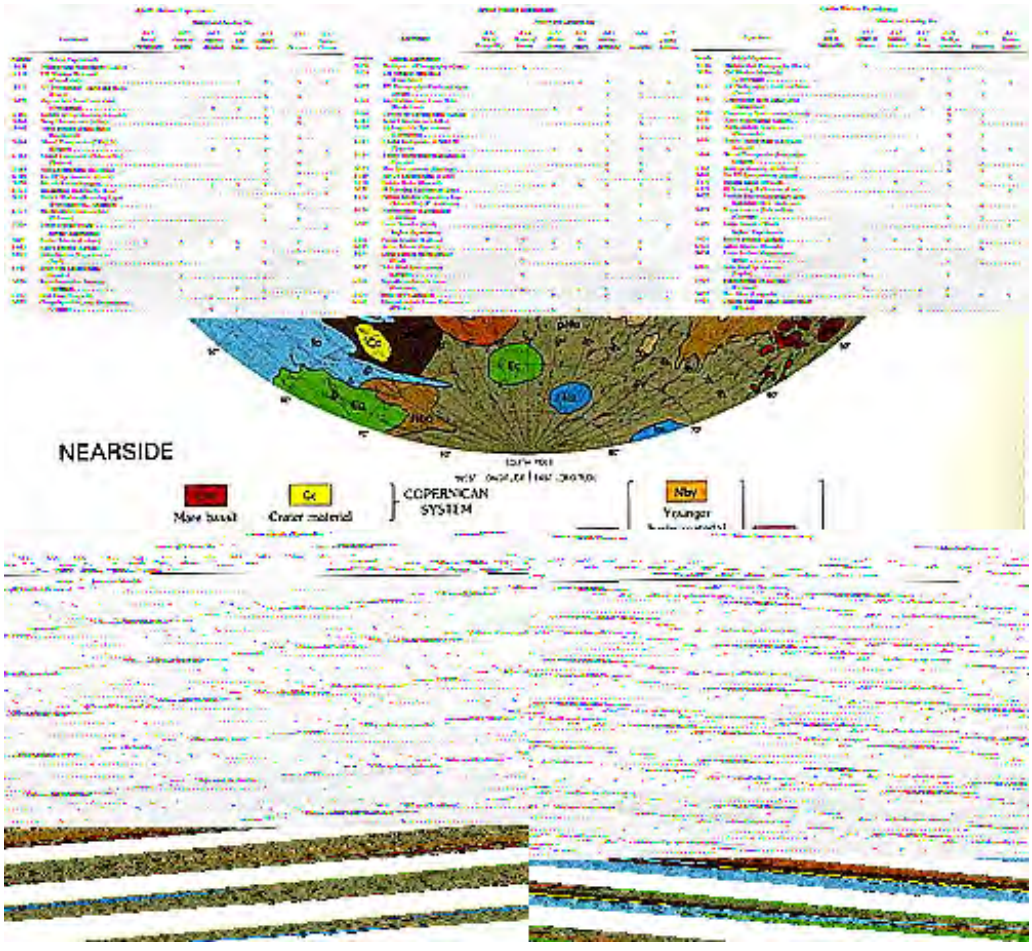


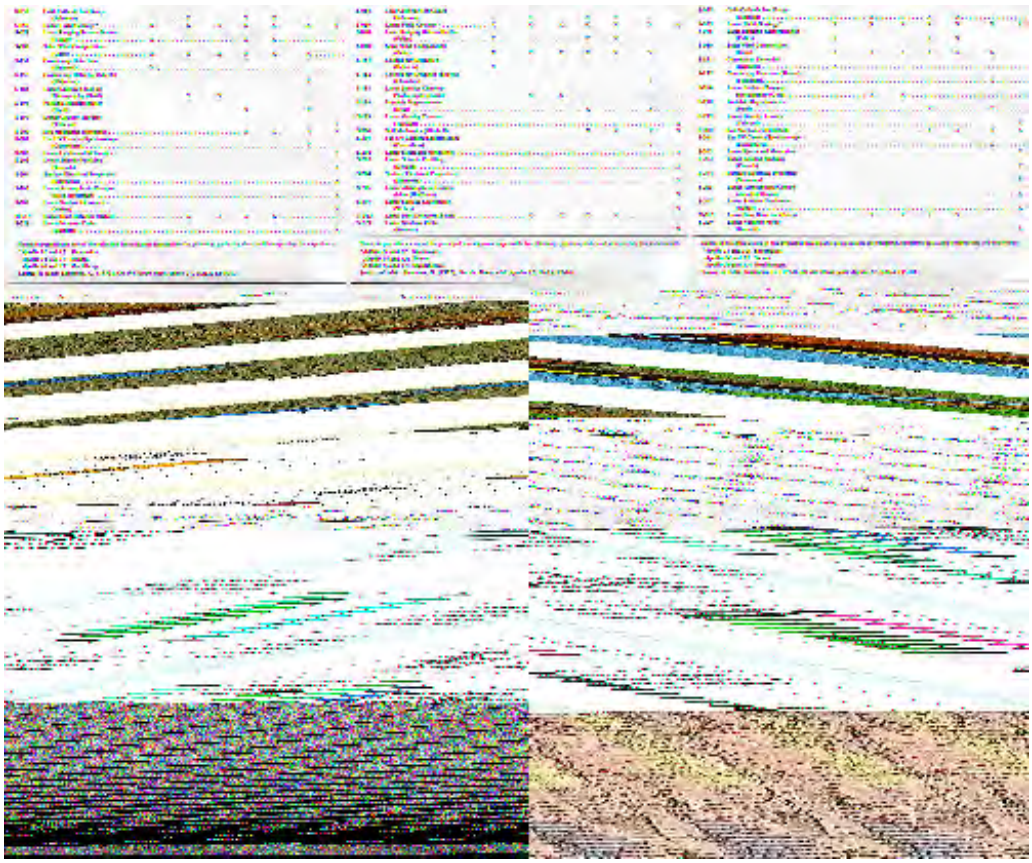
Primary Author: Nicholas M. Short, Sr.

The lunar astronauts not only photodocumented their landing sites and collected rock/soil samples but also deployed numerous sophisticated instruments at the ground station. Some acquired data during the lunar stay but most telemetered results back to Earth and certain equipment, such as seismometers, continued to operate for years. The ALSEP group of ground instruments is described, with several examples of results given. A second group of instruments were mounted in a bay on the Command and Service Module that orbited the Moon while the ground crew performed their program. These instruments made measurements of lunar gravity, magnetism, and chemical composition, together with monitoring some specialized physical phenomena.

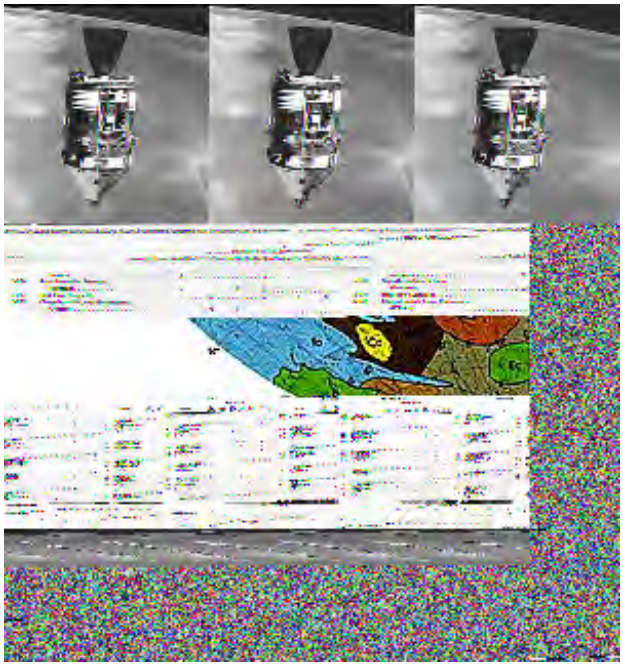
Apollo Instrument Experiments

The top priority for Apollo 11 was just to get there (and back). Once successful, the principal effort was directed towards collecting rock/soil samples. But Apollo 11, and all of the subsequent 5 missions that touched down on the Moon, carried a wide variety of scientific instruments selected and designed to measure a diverse set of physical and chemical properties of the Moon - locally, internally, and in its external surroundings. These fell into two categories: equipment mounted in the SIM (Scientific Instruments Module) of the orbiting CSM (Command and Service Module) and surface-emplaced equipment . The instruments used in both activities and the missions on which they were employed are summarized in this table.





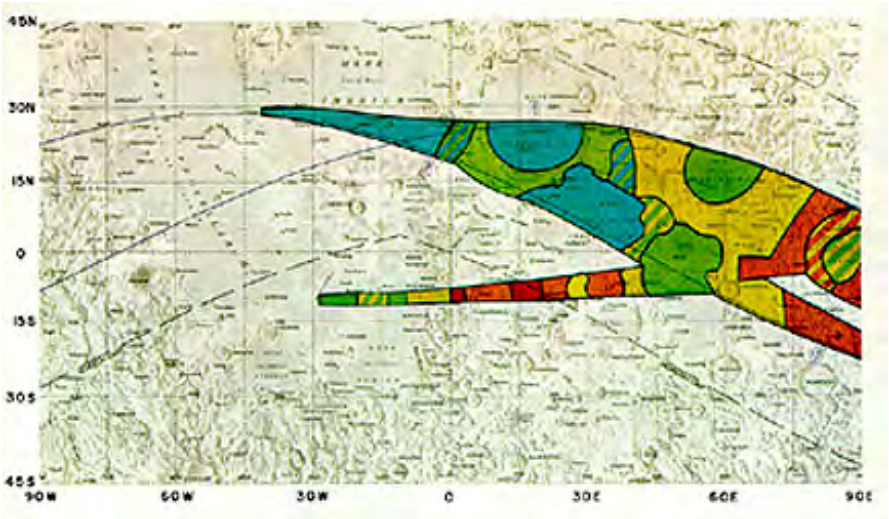
The CSM opens its SIM outer sheath to expose the instruments to the lunar surface, as seen in this photo:



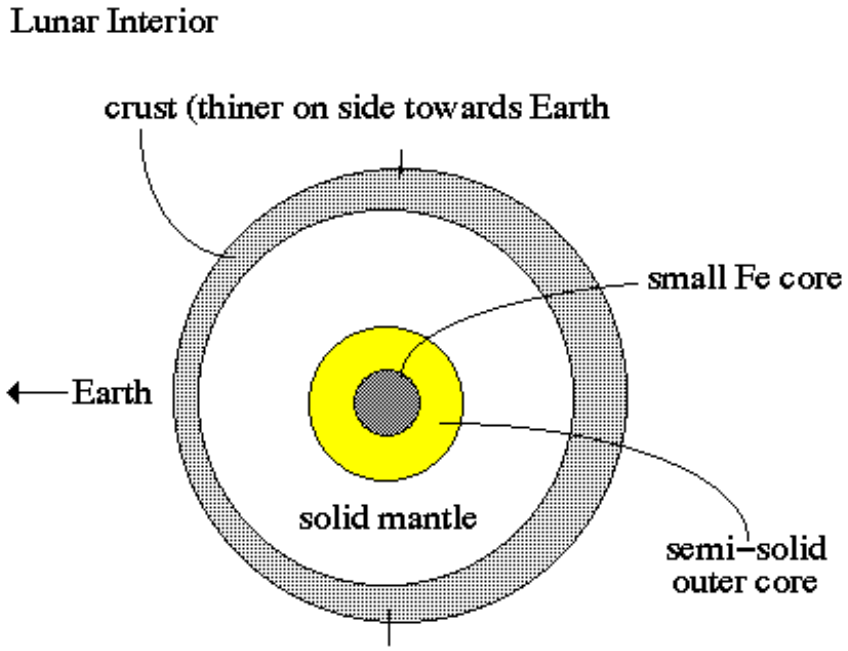
The instruments for the Apollo 15 SIM are labeled in this diagram:

Many significant results have stemmed from operating both the orbital and the ground instruments. But in this Tutorial we are limited to just a few examples. First, several orbital findings, shown as images with minimal comment.

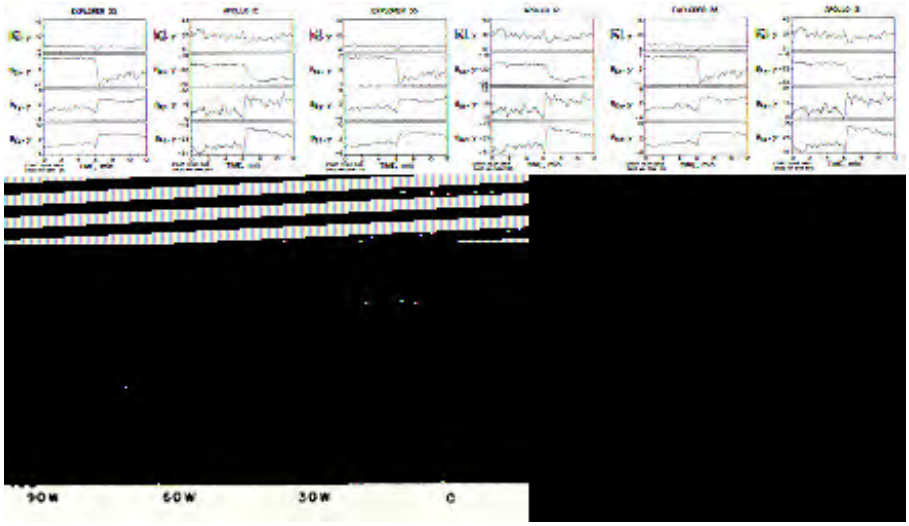
The x-ray fluorescence spectrometer is capable of determining fairly accurately the amounts of a number of elements, including most of the major ones in rocks and some minor or trace elements. This plot shows the variation in Al/Si ratio along paths traversed by the Apollo 14 CSM.



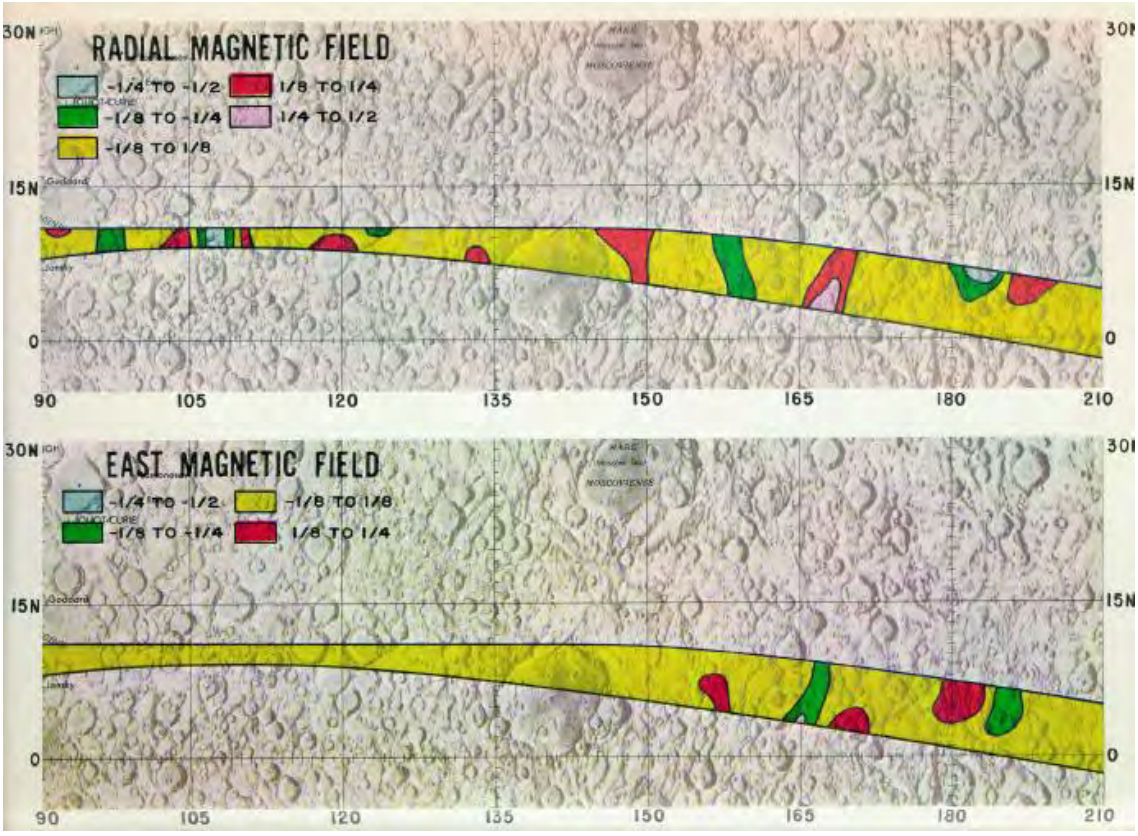
The Moon has a weak magnetic field but one not caused by rotation of a melted zone like that on Earth. It has a small metallic core (about 20% of the radius of the entire Moon) but that is not now suffiently fluid (the outer part may be "mushy") to generate an internal field is probably ruled out. This core was developed when the Moon had accreted and then probably melted more or less completely. The differentiation, like that of the early Earth, yield a crust (anorthositic), a thick mantle, and a small iron core. This illustration is one model of the Moon's interior:



The present weak field is a "fossil" or paleomagnetic one. The source of the field may be remanent magnetism in the lunar rocks themselves, possibly induced by interactions with solar and/or terrestrial magnetic fields or induced in the rocks when/if the lunar core was once fluid in the early history of the Moon. That the Sun's magnetic field plays a large role in the current state of the Moon's field is demonstrated by the similarities (general correspondence) in the B magnetic field vector onboard an Earth-orbiting Explorer 35 and simultaneous measurements on the ground at the Apollo 12 site.

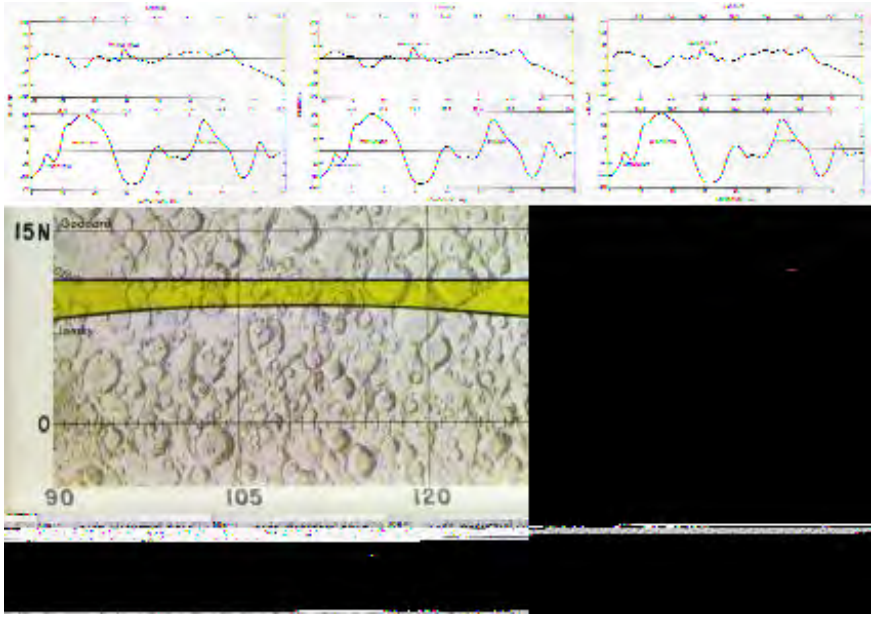


Magnetic field strengths as low as 38 and high as 341 γ were measured at the Apollo ground stations. Thus, there is notably variation from one area of the Moon to another. This is what Apollo 16 measured from orbit using the SIM magnetometer to show surface variations of two magnetic components:



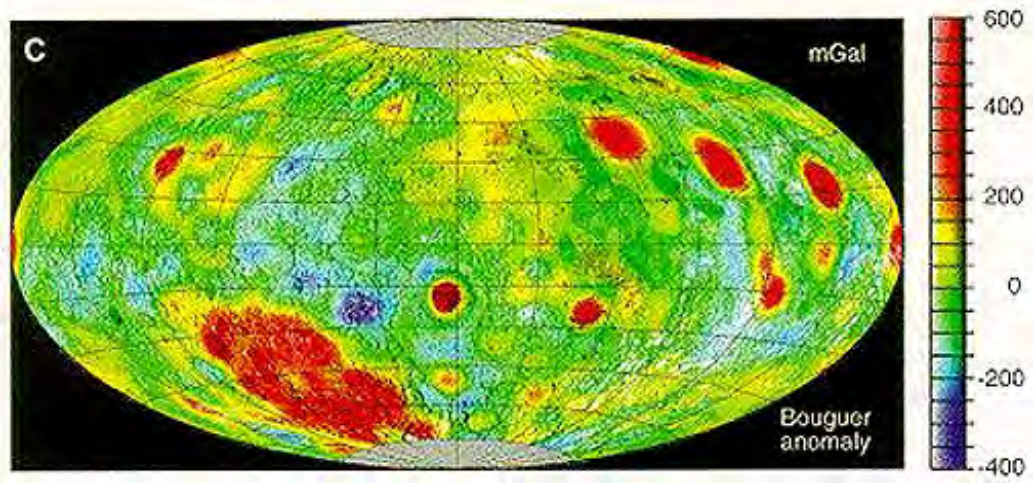
Laser altimeters operated from several of the CSM's. Here is an Apollo 15 profile along one complete circumlunar orbit, expressing differences in elevation as ΔR's, or variations in radius from the mean lunar spheroid.

Variations in lunar gravity were another important parameter to measure in detail. One way to do this is to use radio tracking to determine changes in spacecraft speeds which result almost entirely from fluctuations in gravitational pull. The S-band (radio) Transponder on Apollo 15 was used to obtain these measurements, from which these two profiles (and others) resulted:

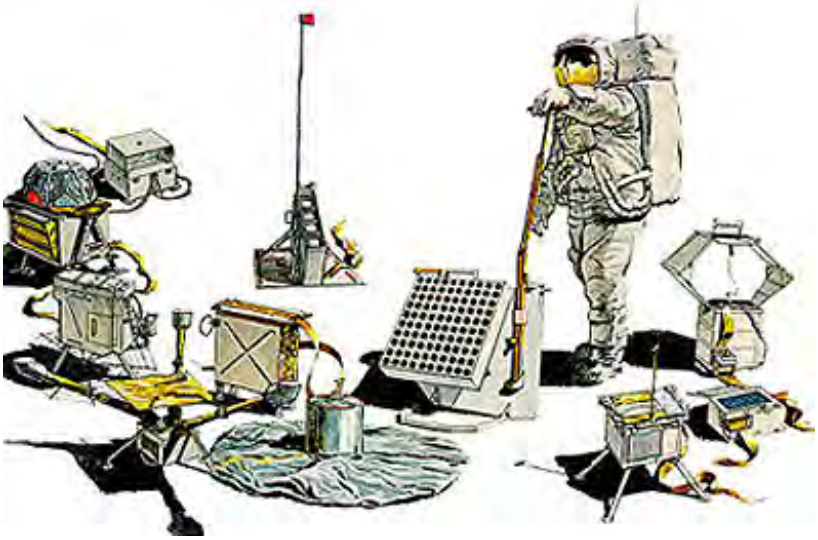


The gravitational highs over Mare Serenitatus and Mare Crisium correspond to features earlier identified and named *mascons* (for mass concentrations). These gravity highs are believed due to increases in rock densities at depths where pressure converts the gabbro feeding the surface basalts into a denser rock called eclogite (see illustration at bottom of next page).

This map covering the entire Moon shows in red areas of mass concentration or mascons; these tend to be found underneath major mare-covered basins:

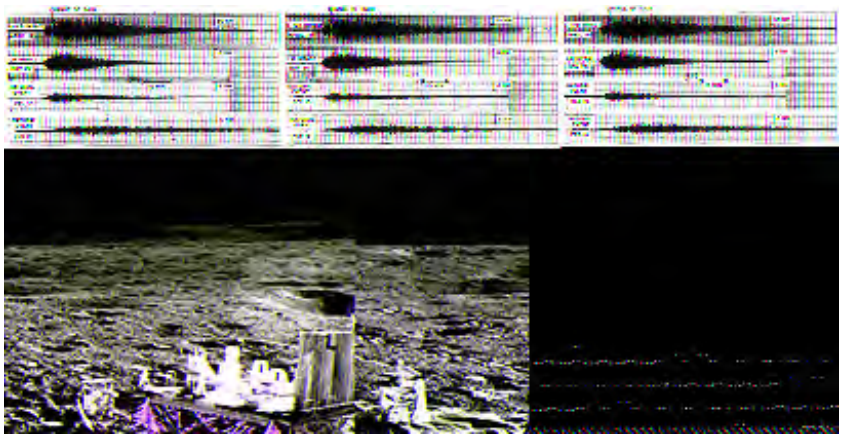


Now, turrning to experiments conducted on the lunar surface by the astronauts. Each Apollo LM carried a compact package of ground-deployable instruments known by the acronym ALSEP (Apollo Lunar Surface Experiments Package). Two good reviews of the ALSEP program are found at these Internet sites (1) and (2). Several instruments were on most flights but individual Apollo missions usually carried new ones, as evident in the table above. This next illustration, credited to a 1970 article in the National Geographic Society magazine, shows some of these instruments:



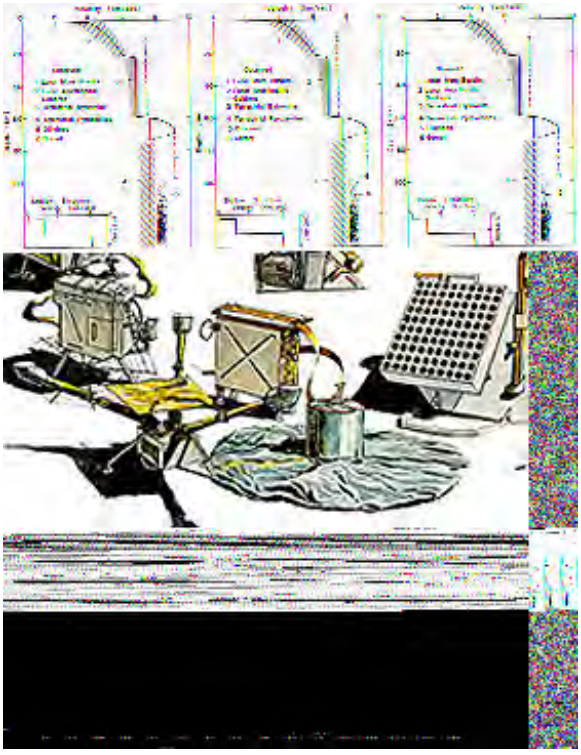
In front of the astronaut is the Lunar Ranging RetroReflector (LRRR) (incredibly, there to reflect a laser beam sent from Earth). Then, clockwise from his left are the Lunar Surface Gravimeter (LSG), the Lunar Atmospheric Composition Experiment (LACE), The Lunar Ejecta & Meteorite Experiment (LEAM) the Passive Seismic Experiment (PSE), the Charged Particle Lunar Environment Experiment; the Lunar Surface Magnetometer (LSM), the Suprathermal Ion Detector (SID), the Cold Cathode Ion Gauge, the Solar Wind Experiment (SWE), the Heat Flow Experiment (HFE), and the Active Seismic Experiment (ASE). Some ALSEP instruments are stand-alone, with their own power source. Others are tethered by cable to a central power source, the Radioisotope Thermoelectric Generator (RTG), shown here at the Central Station point at the Apollo 14 site.

These many sophisticated instruments (some costing in the $millions to insure their spaceworthiness) have returned vast amounts of data that have enlarged our understanding of the Moon's interior and surface environment. One experiment must suffice to illustrate this versatility. Below are seismic records (seismograms) obtained by the Passive Seismic Experiment. The top two are produced when a spent Saturn IV-B casing was forced to crash on the Moon and when a LM (left behind as the CSM returns to Earth) was allowed to decay in orbit and hit the Moon. The bottom two records are of natural (and small; magnitudes of 2 or less) moonquakes from the interior.



All these records differ from those associated with terrestrial earthquakes in lacking S (secondary or shear) waves. They also last for much longer than earthquakes - up to an hour, as though the Moon were "ringing" like a bell. The bulk of the moonquakes originate from depths of 600 km or more (attributed to the seismic signals bouncing back and forth in the low velocity surface layers [ejecta blanket]). They tend to occur in swarms associated both with preferred spatial locations and specific time intervals (probably accounted for by stresses induced by terrestrial tidal forces).

Seismometers at Apollo 12, 14, 15, and 16 continued to operate for years, providing a wealth of data about the Moon's interior. One model for that interior is shown in this diagram:



Note the two velocity discontinuities at about 20 and 60 km. Velocities within this interval are consistent with anorthosites. The higher velocities deeper than 60 km are associated with a proposed pyroxene-olivine upper mantle. Velocities between 4 and 20 km are indicative of a fractured basaltic flow sequence responding to load pressure. The shallow velocities (lower left inset) are attributed to the lunar ejecta blanket units discussed on the preceding page, overlain by comminuted regolith.

A final word: Heat probe measurements showed a higher heat flow range, averaging 2.8×10^{-6} W/cm^2, than was predicted before the landings (Earth's average is ~6.2×10^{-6} W/cm^2). This implies a hotter interior than expected, perhaps related to higher uranium (heat-producer through radioactive decay) concentrations.

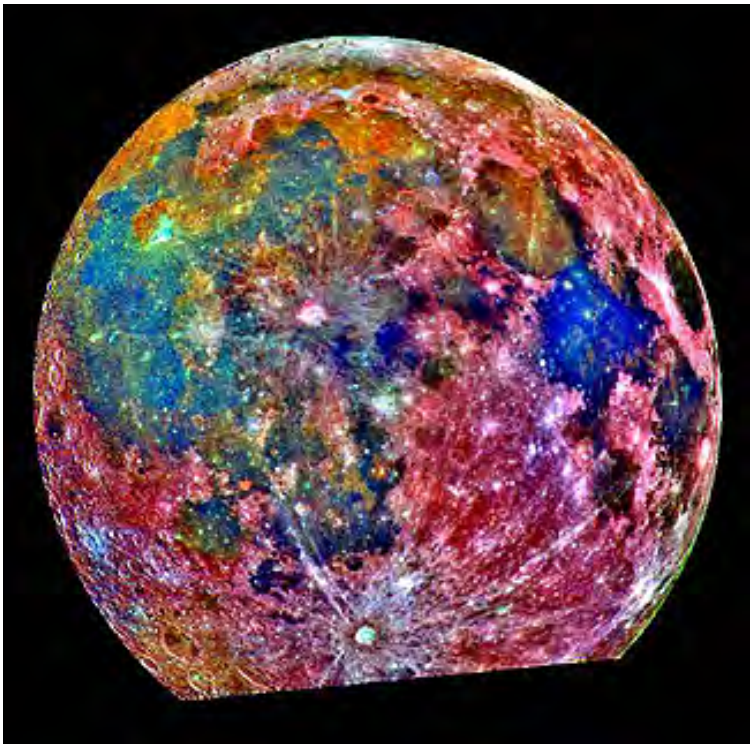
Primary Author: Nicholas M. Short, Sr.

Following the Apollo program, public interest, and to a lesser extent that of many planetary scientists, seemed to have notably waned. One reason for this was the increasing exploration of the other planets (Mariner; Pioneer; Voyager, and other programs). However, scientific investigations of moon rocks continued unabated. In the 1990s, a renewal of scientific interest in the Moon led to NASA programs to return there with orbiting spacecraft that extended instrumental analysis to the entire surface. Results of the Clementine and Lunar Prospector programs are presented. Of tantallizing significance was the discovery of considerable water ice at the Moon's South Pole, which reopens the possibilities for setting up a station or base for long term stays that would permit carrying out extended experiments. Speculative models for the origin of the Moon (now most scientists consider an impact on Earth to have been involved) are strongly supported by its exploration over the last 30 plus years.

Post-Apollo Lunar Exploration

After Apollo interest in further scientific examination of the Moon seemed to rachet down almost to "nil". But as various nations have succeeded in establishing viable space programs, the Moon once again has been targeted for renewed exploration during the first decade of the 21st century.

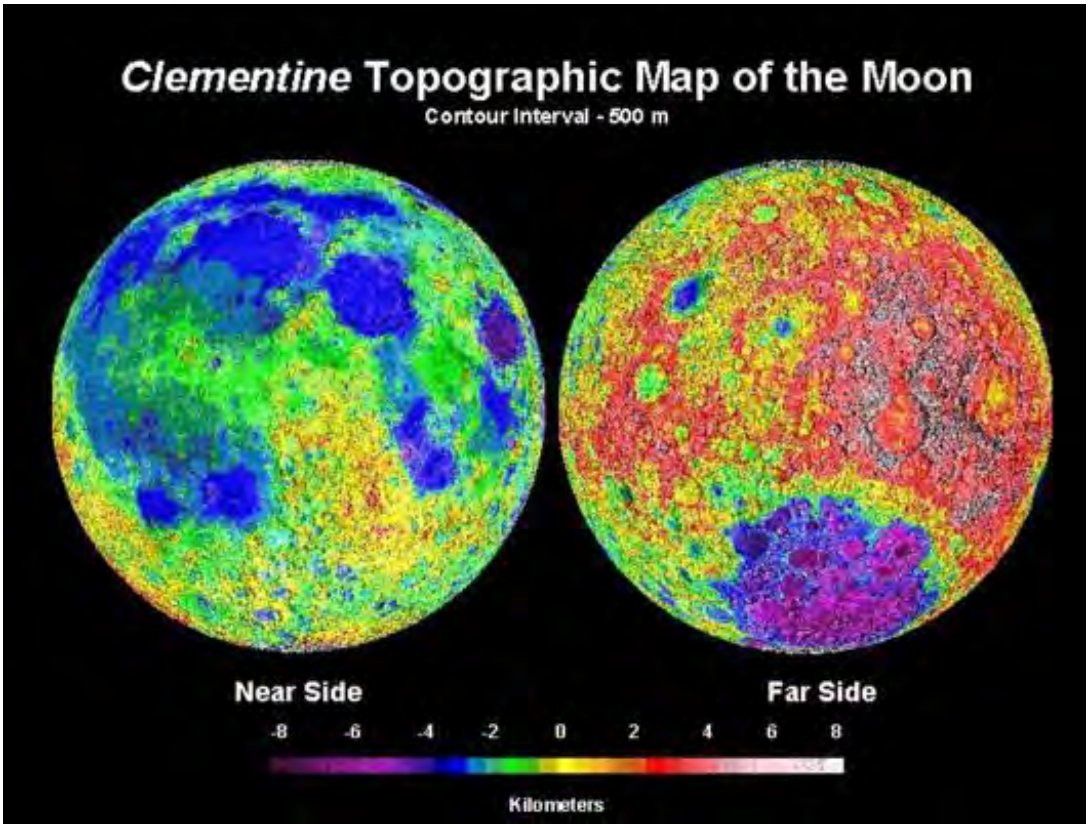
The views presented so far highlight the two dominant characteristics of the lunar surface: 1) the mare/highlands dichotomy, and 2) the abundance of circular features, nearly all being impact craters and basins but some of probable volcanic (caldera) origin. This next scene emphasizes both characteristics by showing an exaggerated false-color image of the front side of the Moon, taken by the multispectral vidicon onboard the Galileo spacecraft (described later in this Section). The highlands, with their higher reflectances, appear in shades of red and orange and the lower reflectance maria are in blues and greens.



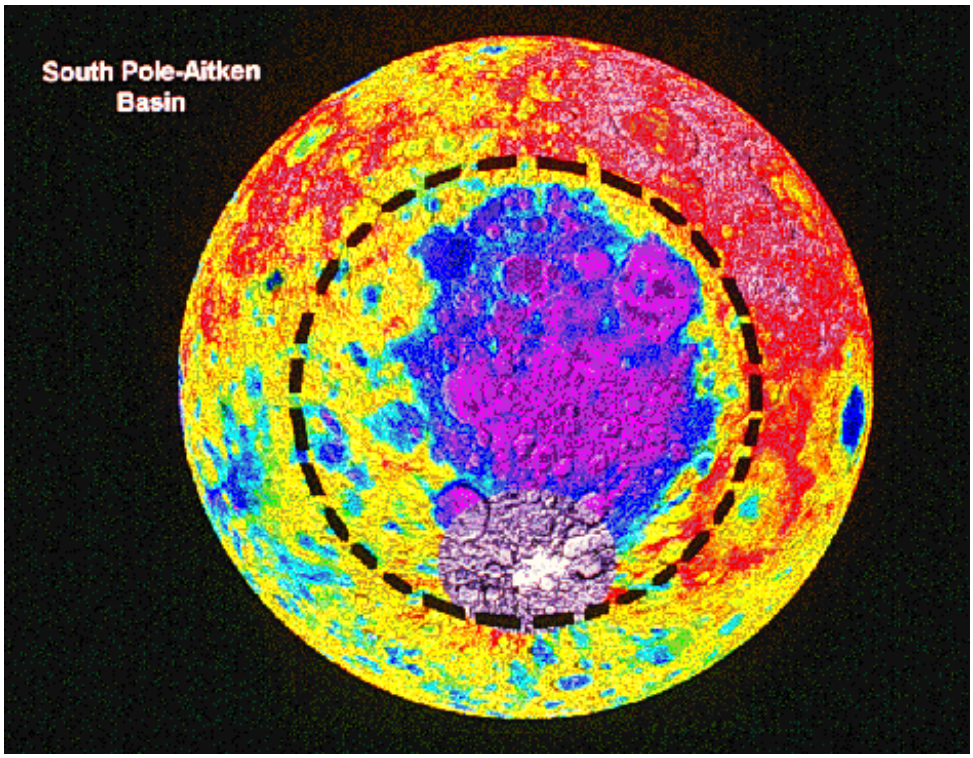
After Apollo, the Moon was not specifically revisited for 22 years, until an unmanned spacecraft,

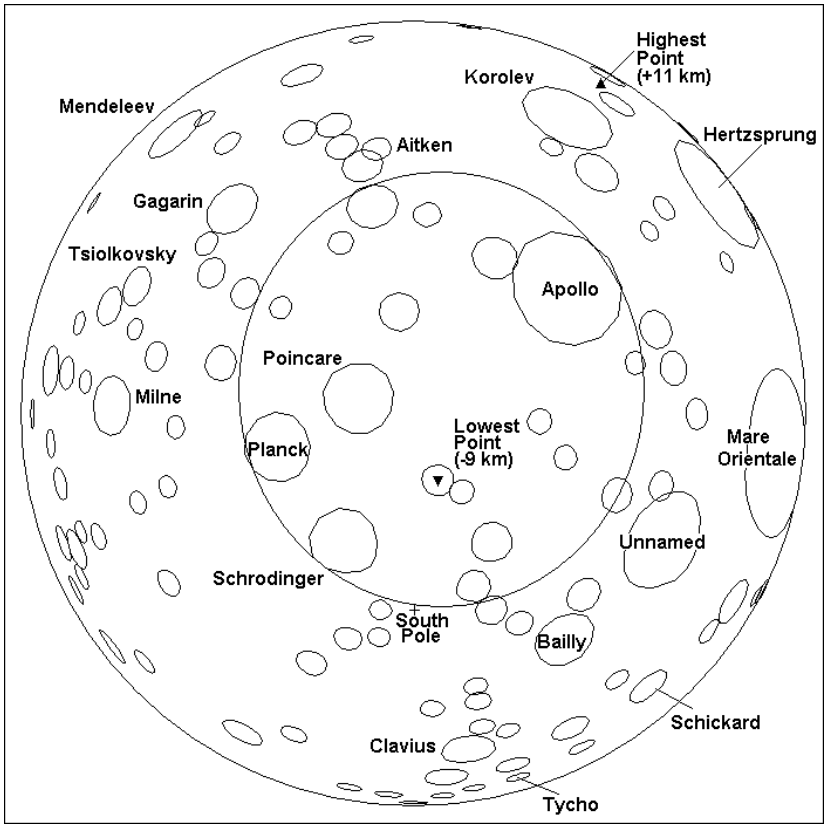
Clementine (funded by the Department of Defense), orbited it to conduct mapping studies between February 19 and April 21, 1994, using UV/Visible, Near IR, and High Resolution Cameras, Lidar (a radar altimeter), and a radar-like unit that transmits in the S-band radio frequency (2.293 GHz, or 13.19 cm wavelength).

Look first at a topographic map of the front and far sides of the Moon, in which stereo data provided elevation differences from high resolution photographs and radar altimetry data, acquired by the Clementine spacecraft as it orbited the lunar surface. In the Far Side view, note the low topography around the South Pole.

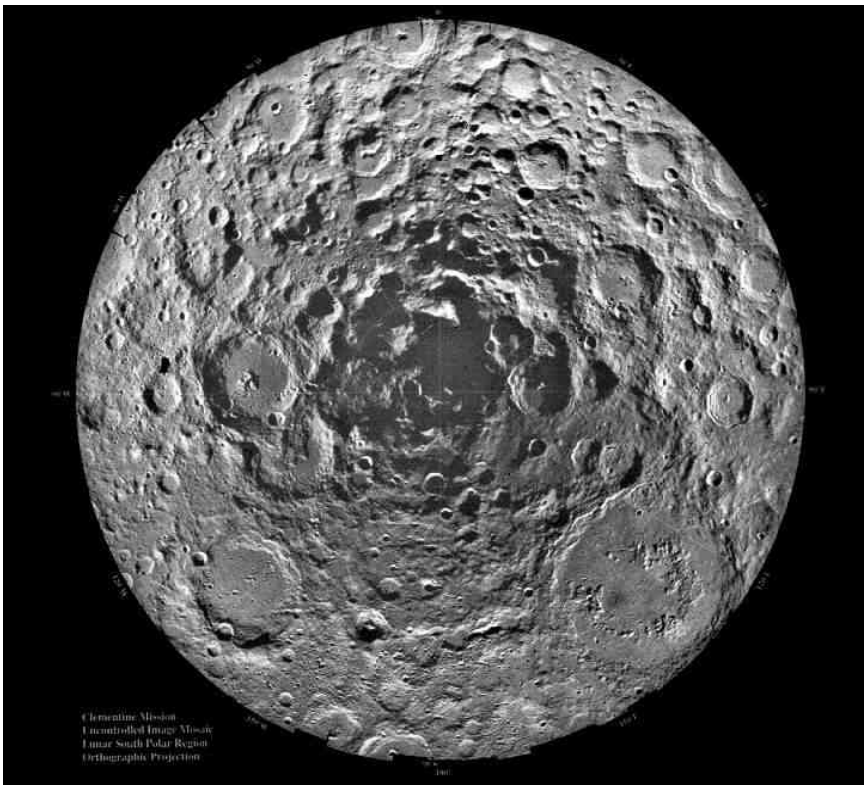


Clementine has produced topographic maps of the polar regions of the Moon. This one covers the South Pole region. It is notable for showing the largest impact basin on the Moon, and perhaps in the Solar System. This is the South Pole-Aitken Basin, with its main craters named in the map below. It is approximately 2500 km (1550 miles) in diameter and 13 km (10 miles deep). Being old, it has been much degraded by subsequent impacts.





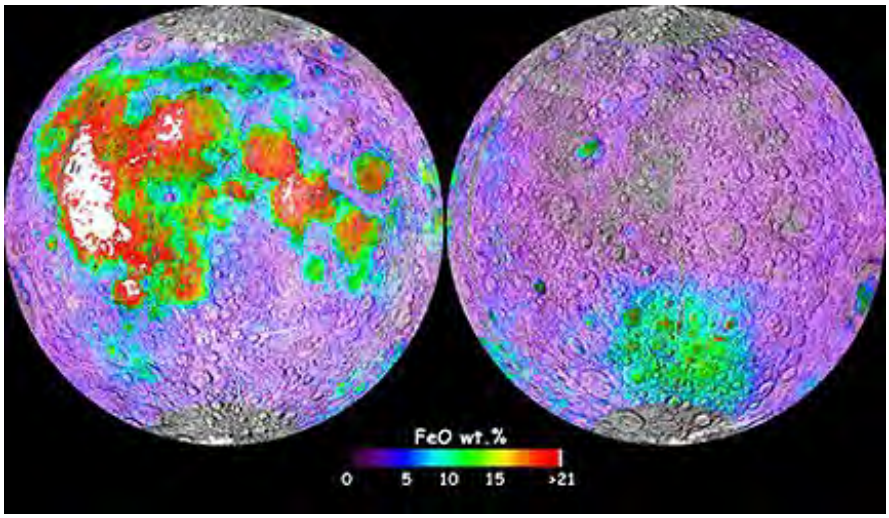
The next image, made as a mosaic from Clementine images, also shows the South Polar region. But it is hard to associate (compare) this with the above maps, since it is centered on a different point and has a different orientation. This demonstrates the problems one can have in comparing lunar scenes that have features in common - there are so many craters that fitting one map into another can prove visually difficult.



As explained in the next paragraph, image data at various wavelengths can be used to map compositional differences in much the same manner as with multiband data obtained by Landsat and other terrestrial spacecraft. Below is an image of the 40 km wide crater Aristarchus that is found in the southeastern part of Oceanus Procellarum. The composite image is constructed from three ratio images (input bands in units of micrometers [µm]): 0.750/0.415 = red; 0.750/1.00 = green; 0.415/0.750 = blue. The dark gray surface is mare basalt; the reddish unit is ejecta from Aristarchus; the light blue is probably anorthositic rock (common in the highlands) exposed in the crater interior:

Among specialized products were more detailed maps of lunar topography (elevations) and global maps of the distribution of several chemical elements, such as iron (Fe) and titanium (Ti), determined by analyzing reflectance variations at 0.75 m m and 0.95 m m, where these elements absorb irradiation. Most of the iron is actually in the form of FeO (reduced iron). The Clementine

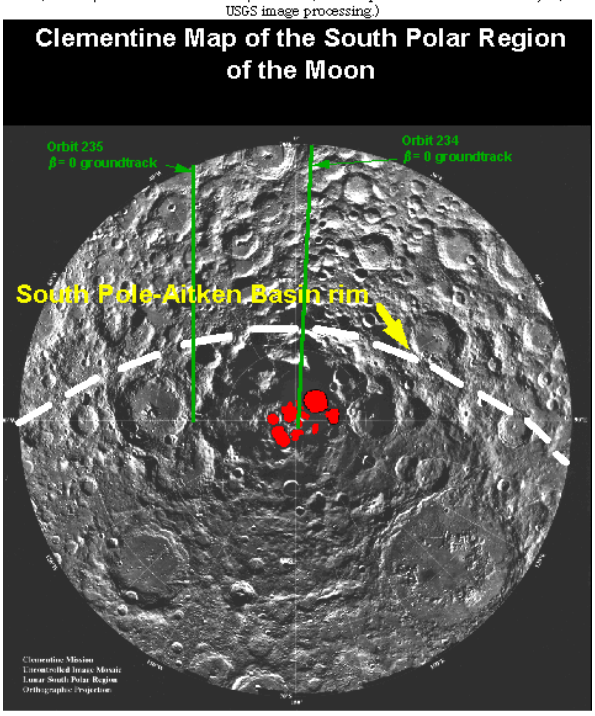
results when plotted as FeO are thus:



While iron is widespread, its maximum concentrations are in a broad region on the nearside, roughly coincident with the vast lava outpourings into Oceanus Procellarum and several other mare basins.

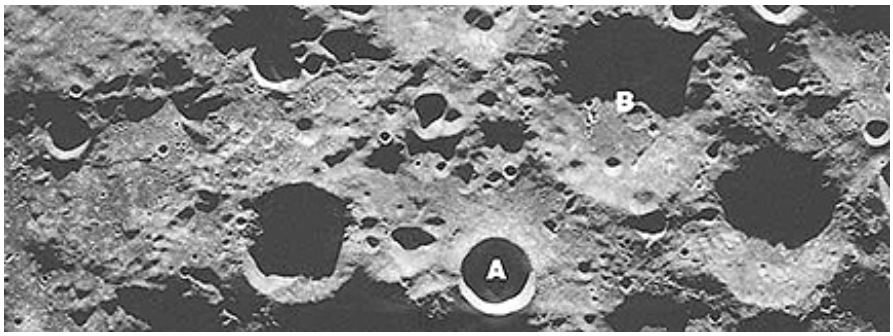
Clementine made a controversial discovery, which, if proved correct, has major implications for humans returning to the Moon. Its S-band radio unit detected abnormal reflections from the rim of a huge crater (basin) around the lunar South Pole, in areas permanently sheltered from the Sun's rays, as seen in this Clementine image:

(Courtesy of Lunar and Planetary Institute, basemap from the Clementine Project; USGS image processing.)



These reflections could be due either to water ice or to some abnormal surface roughness condition. If indeed ice is present in significant quantity, then this precious material (which supplies water needed for life and also oxygen, when broken down by electrolysis) might allow future astronauts to establish a manned base on the Moon. Transport of sufficient water and oxygen for long stays is presently beyond the space program's technical capability.

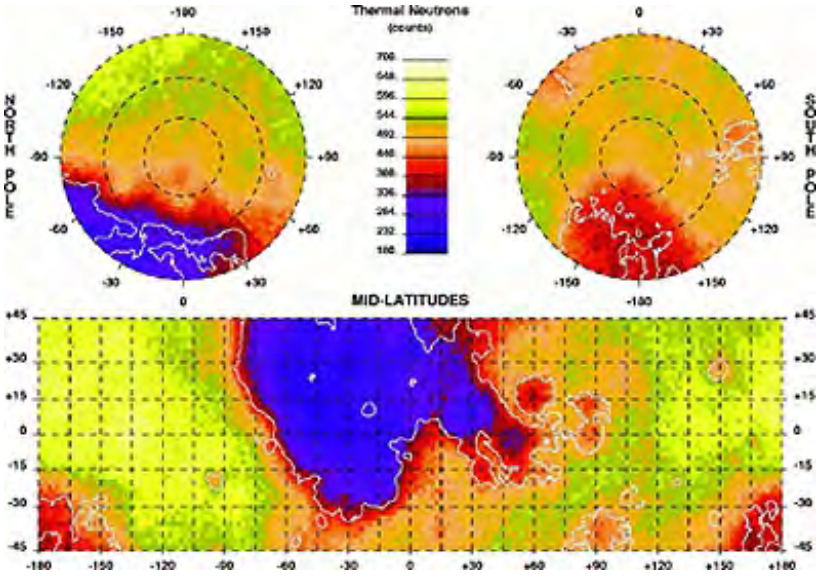
Because the South Pole region is a candidate for an eventual lunar base, radar units from Earth have returned high resolution images of the polar terrain, without so far having confirmed the presence of ice. Here is an image made at 13.2 cm wavelength, from the Arecibo radar dish in Puerto Rico (no ice was identified at the 20 m resolution of this system).



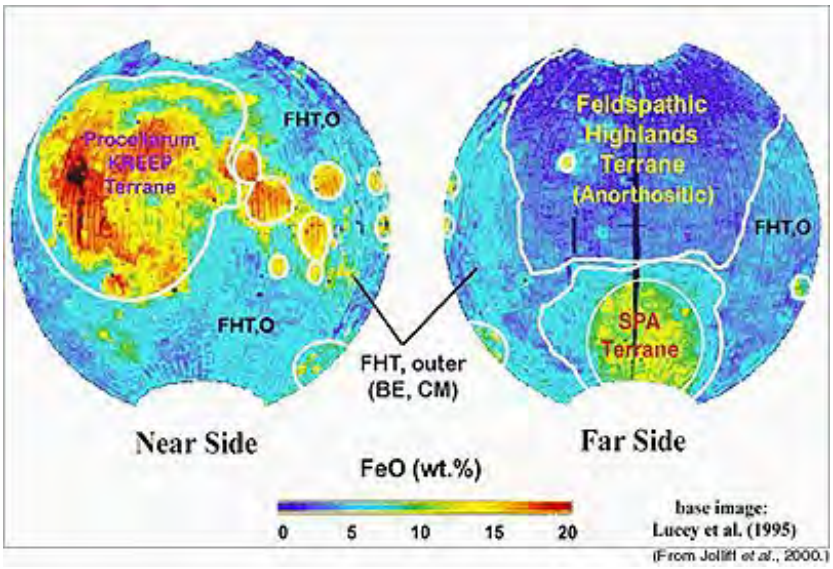
The observation of possible ice, and other intriguing results of Clementine's compositional mapping, has led to a follow-on mission. For the first time in 25 years, NASA has returned to the Moon with a small, but versatile orbiting satellite, called Lunar Prospector. The entire mission including data analysis is another effort by NASA to achieve high scientific returns at relatively low cost (for LP, $65 million). Launched on January 6, 1998, by an Agena rocket, Prospector now is operating in a 100-km high circumlunar polar orbit, from which it can map the entire Moon over a 3-year lifetime in more detail than Clementine provided. Here is an artist's sketch of the spacecraft:

The spacecraft, just 1.4 m (4.5 ft) high and 1.2 m (4 ft) in diameter, weighing 300 kg (660 lbs), receives its power from solar cells that surround its exterior. An S-band radio sensor designed to measure lunar gravity employing a Doppler effect procedure, sits on top of a conical communications antenna (top). At the end of the 8-ft boom or mast extending to the front left, a Magnetometer/Electron Reflectometer will conduct improved measurements of the Moon's magnetic and particle fields. At the end of the left rear mast is the Gamma Ray Spectrometer, which can detect these elements: U, Th, K, Fe, Ti, O, Si, Al, Mg, and Ca. On the right boom are the Alpha Particle Spectrometer that will measure radon gas to assess lunar radioactivity as a clue to volcanic and other current events, and the Neutron Spectrometer that will determine the presence of hydrogen and can detect water ice (its confirmation from Clementine results is a major goal).

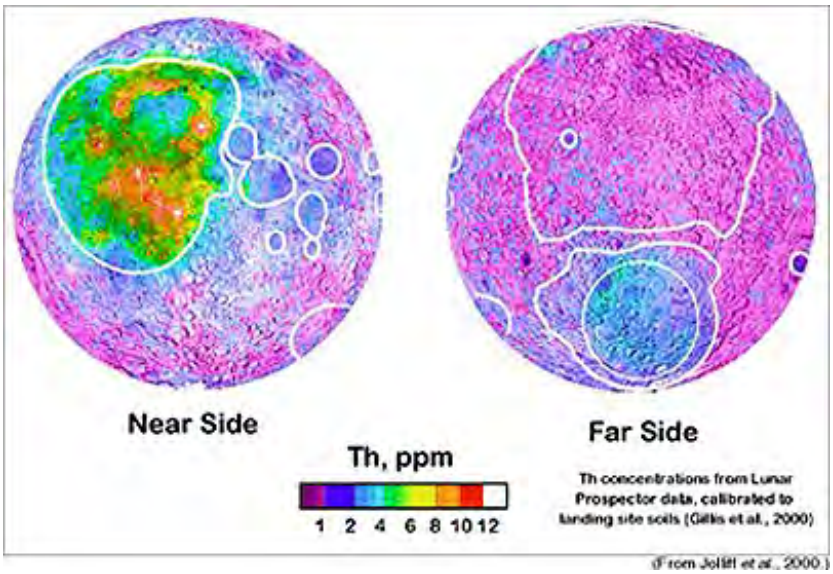
A plot of the varying thermal neutron flux, as determined by the Neutron Spectrometer, show a wide area of low neutron counts (resulting from high neutron capture) associated with the maria on the frontside and near the North Pole and higher counts in the highlands.



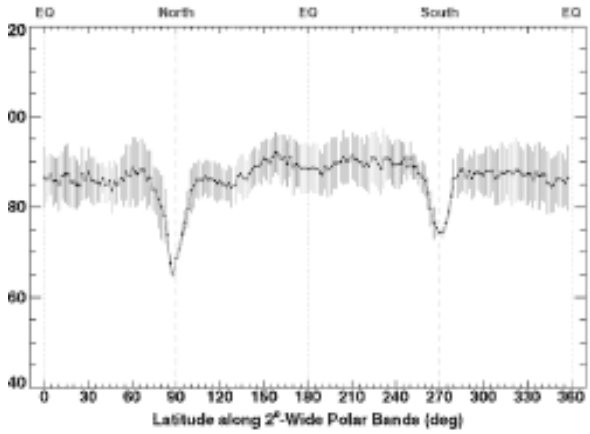
Compare the distribution of Fe as determined by Lunar Prospector (below) with the same coverage by Clementine shown above:



Information on the distribution of radioactivity on the lunar surface was one goal of Lunar Prospector. This map shows that the element thorium is highest on the front side of the Moon, mainly in the highlands south of Mare Imbrium. The correspondence with the Imbrium Basin suggests that the basaltic lavas that filled it were enriched in Th. Note that corresponding highland surfaces on the farside are lower.



The first results on Lunar Prospector's detection of ice were released during an exciting press conference, held on March 5, 1998. Around both poles, the neutron spectrometer has indeed detected neutrons, released from hydrogen by natural cosmic ray bombardment of water ice in craters that have permanently sheltered shadow zones. The drop in neutrons emanating from the Moon is clearly maximal around the poles as seen in this plot.



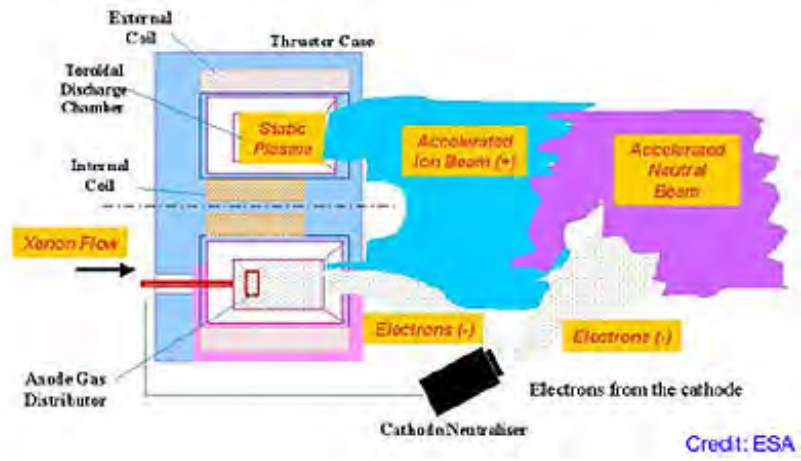
Los Alamos National Laboratory

The initial estimate of the amount, to be determined more accurately with later observations, is 30 to 300 million metric tons (recent thinking has raised the upper limit to perhaps as high as 3 billion tons). If melted, this larger number would fill a "lake" 10 square kilometers in area (3.1 x 3.1 km) to a depth of 10 meters. Surprisingly, the North Pole region contains about 50% more ice than its southern counterpart. The source of the water ice is probably residues from cometary bodies that impacted the polar regions, forming craters but allowing much of the comet mass to survive embedded in the target. The implications are encouraging for future exploration of the Moon, to the extent that we can establish and occupy a manned base facility over extended time because of the availability of vital water (for consumption and as a source of hydrogen, suitable as a fuel). Landing in polar regions is technically more difficult but doable. The dream of a permanent observation post on our satellite is now much more feasible.

More details on Lunar Prospector are given at the National Space Science Data Center Web site and the Mission Management Home Page at NASA Ames Research Center. As NASA accrues and releases data and maps, we will place them in the Web version of this Tutorial and in later CD-ROM versions.

An important mission to the Moon is ESA's SMART-1 spacecraft. Launched September 27, 2003 as Europe's first venture in exploring beyond Earth, the spacecraft, using a novelion (Xenon gas) propulsion system, proceeded slowly to the Moon and then arrived in November of 2004.

Stationary Plasma Thruster



SMART-1 is Europe's first lunar mission and has provided some significant advances to many issues currently active in lunar science, such as our understanding of lunar origin and evolution. The mission also contributes a step in developing an international program of lunar exploration. The spacecraft was launched on 27 September 2003 on an Ariane 5, as an auxiliary passenger to Geostationary Transfer Orbit (GTO), performed a 14-month long cruise using the tiny thrust of electric propulsion alone, reached lunar capture in November 2004, and lunar science orbit in March 2005. SMART-1 carries seven hardware experiments (performing 10 investigations, including three remote sensing instruments, used during the cruise, the mission's nominal six months and one year extension in lunar science orbit). The remote sensing instruments will contribute to key planetary scientific questions related to theories of lunar origin and evolution, the global and local crustal composition, the search for cold traps at the lunar poles and the mapping of potential lunar resources.

This low cost satellite orbited the Moon for nearly two years gathering information about surface composition. Its instruments are:

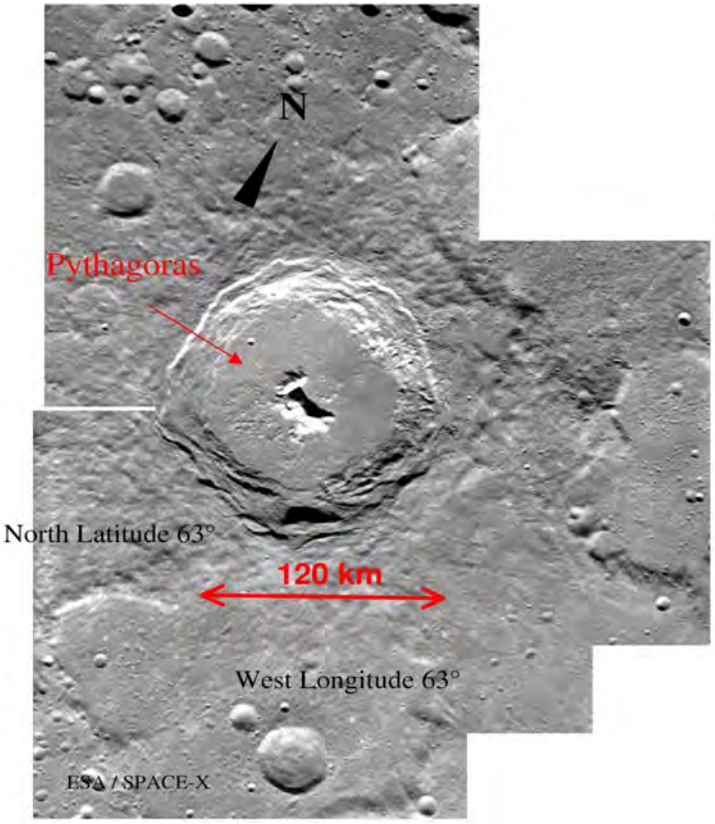
AMIE: A miniaturised color camera with a resolution of 40 meters for lunar surface imaging.

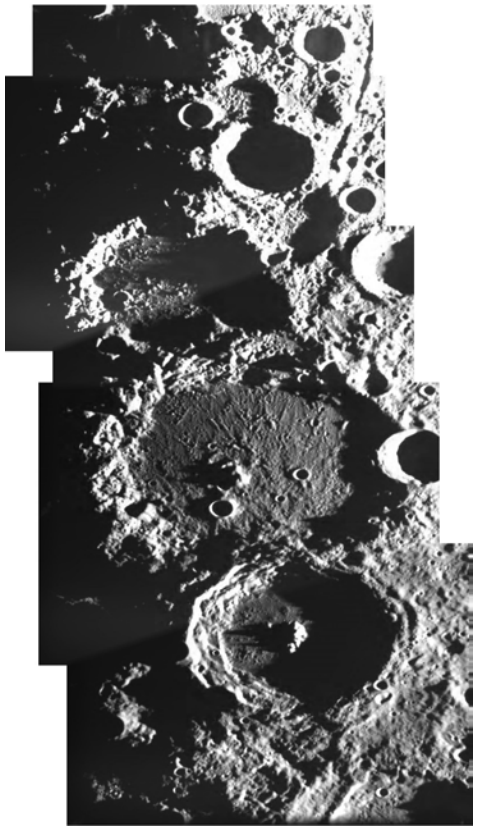
SIR: A near-infrared spectrometer for lunar mineralogy studies.

D-CIXS: A compact X-ray spectrometer to perform fluorescence spectroscopy and map the Moon's surface elemental composition. It also performed observations of celestial X-ray sources while en route to the Moon.

XSM: An X-ray monitor to support D-CIXS with measurements of solar X-ray emission for calibration. It also observed solar flares while en route to the Moon.

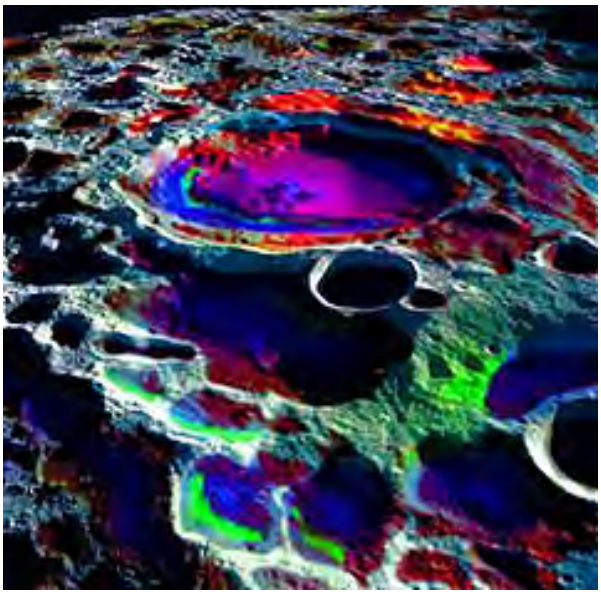
Here are two images made by AIME of the lunar surface.



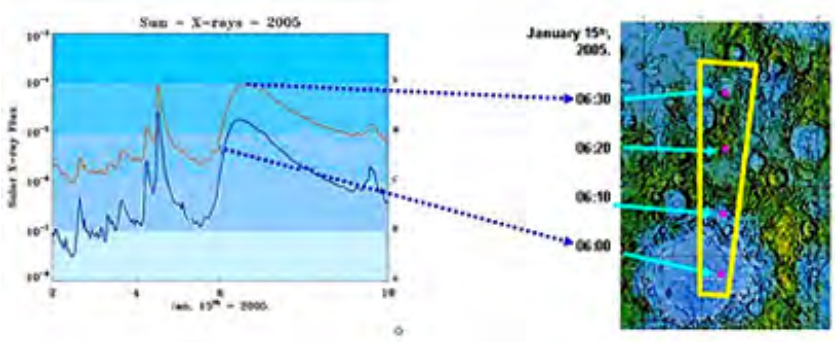


Below are an AIME view of the central peak of Crater Zucchius and ejecta around Mare Orientale:

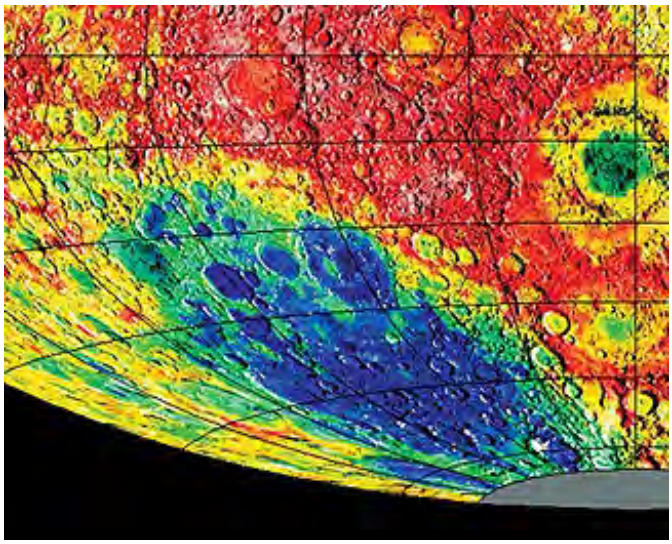
Here is a SIR mineral map of a cratered area on the Moon:



The D-CIXS detects X-ray fluorescence of minerals that are being excited by X-rays from the Sun. This is a map of Calcium distribution within a small area of the Moon around Mare Crisium:



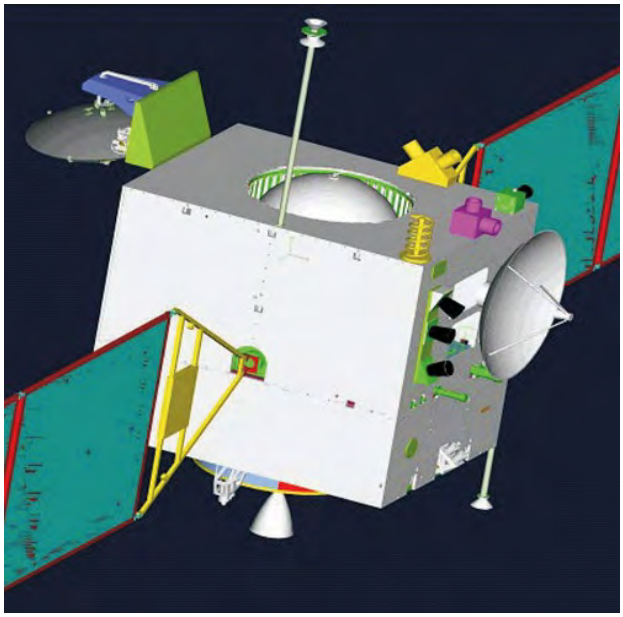
SMART-1 spent a lot of time looking at the polar regions, partly to search for evidence of possible water ice. In so doing, it provided information on the Aitken impact basin, some 2500 km in diameter, reputedly the largest such structure in the Solar System. Here is a map of the South Polar region showing this structure:



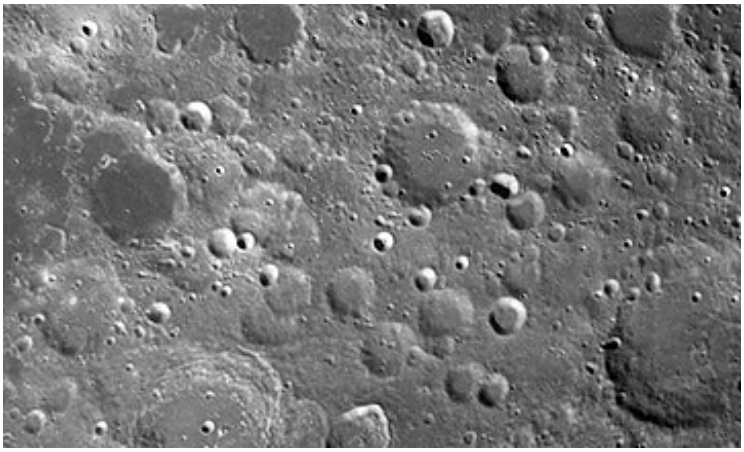
SMART-1 ended its mission by being deliberately impacted onto the lunar surface on September 3, 2006.

The Chinese and Japanese have sent probes to the Moon in 2007 and 2008 respectively.

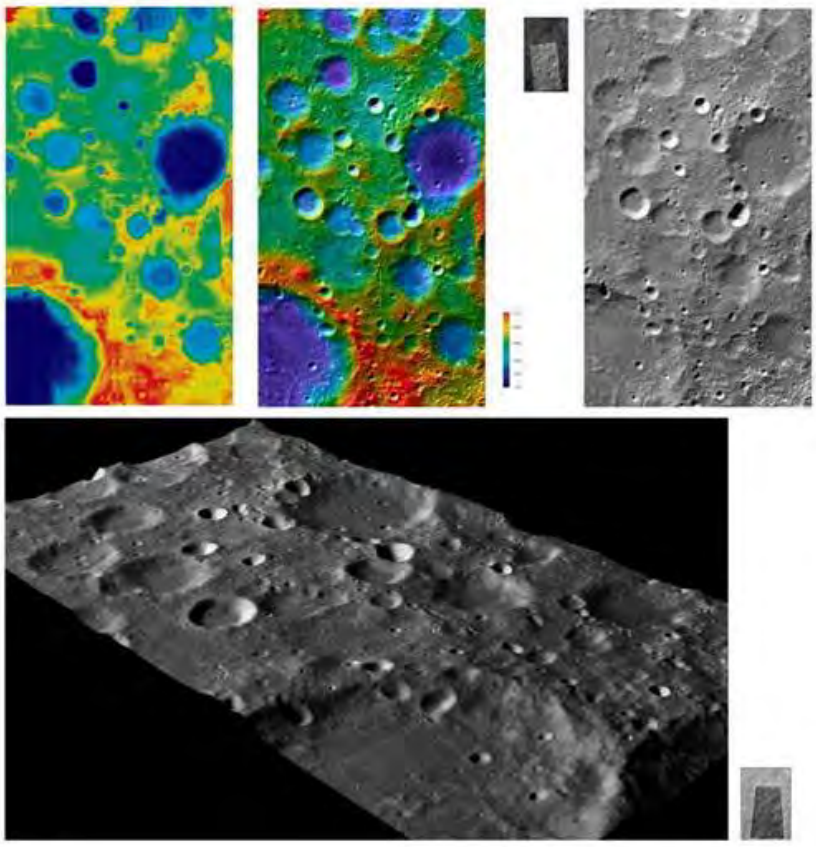
This is an artist's view of the China's Chang'e-1 spacecraft, which was launched on October 24, 2007:



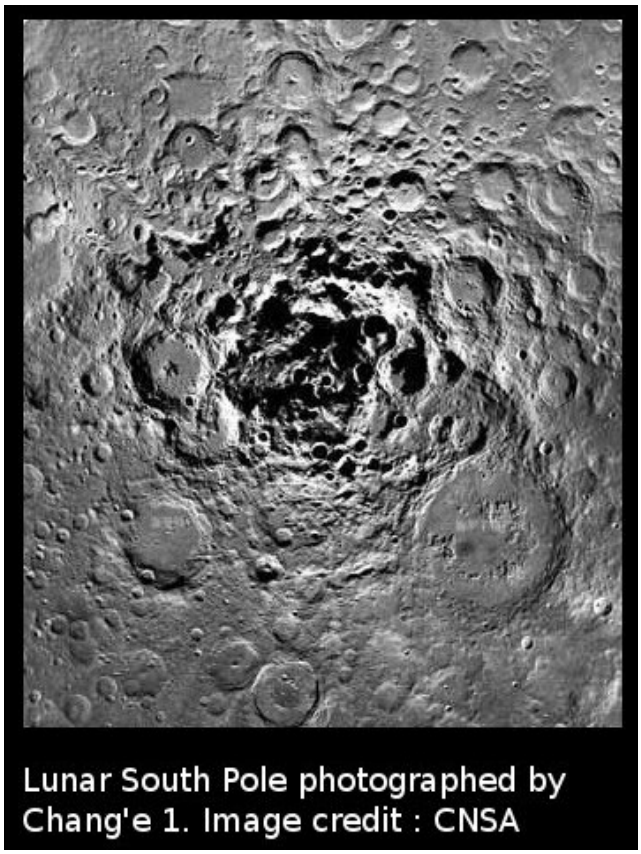
The first image from Chang'e-1 is shown here:



This panel shows the types of products coming from Chang'e-1:



The contrast in this Chang'e-1 image of the craters at the lunar South Pole is interesting:



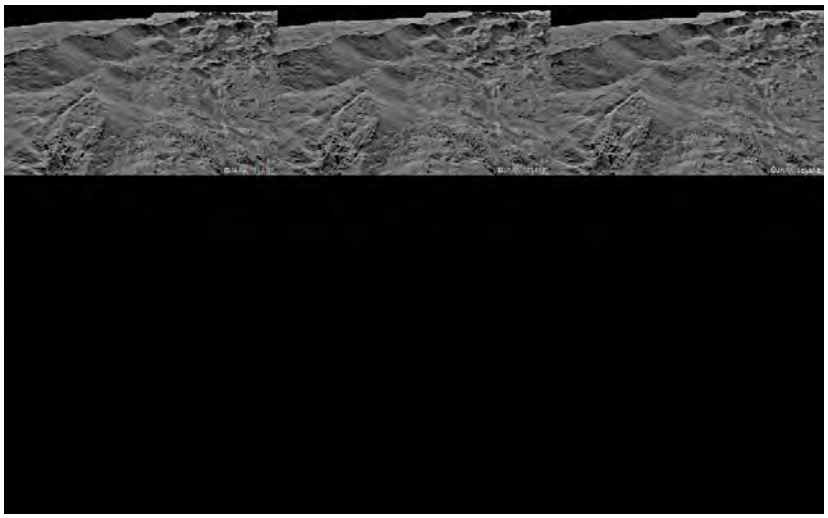
Lunar South Pole photographed by Chang'e 1. Image credit : CNSA

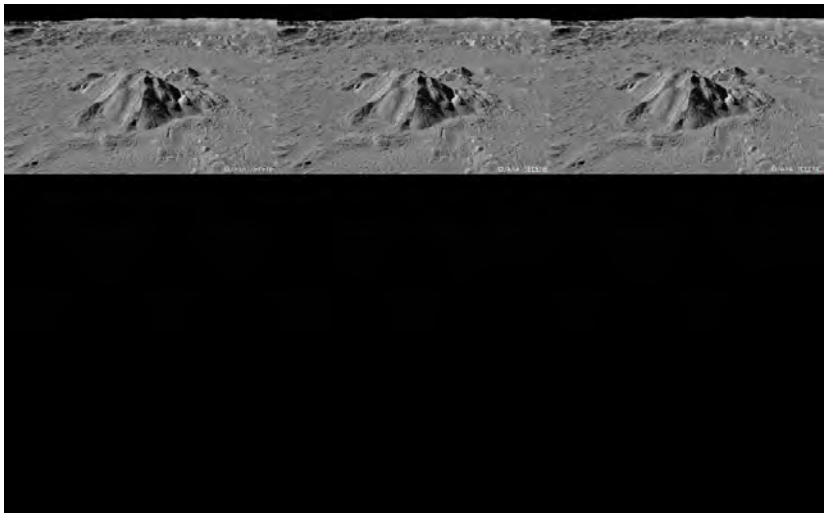
Chang'e-1 was sent to the lunar surface (crashed) on March 1, 2009.

Japan's spacecraft, Kaguya (Selene), is primarily an imager but it does have 13 separate instruments. Here is how it looks:

These are typical images of the lunar surface as captured by Kaguya:

Here are Kaguya images of the rim and the central peak of the crater Tycho:

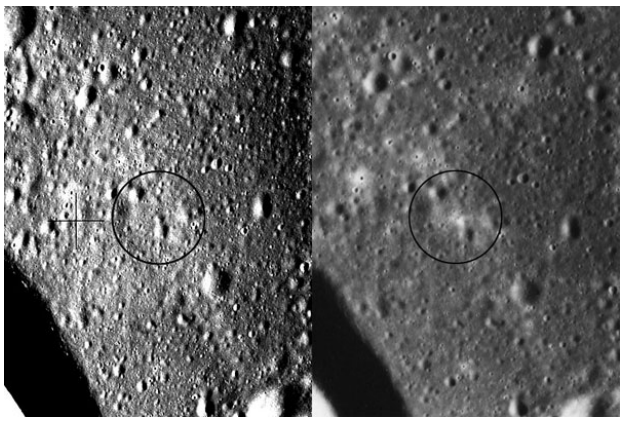




This is Kaguya's image of Mare Muscovensis on the lunar far side:

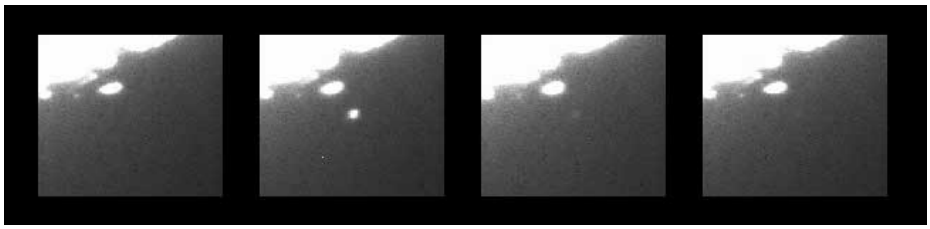
Unlike the earlier U.S. orbiting imagers, Kaguya can produce quasi-color images:

An interesting Kaguya image shows the disturbed soil around the site where the Apollo 15 LM took off more than three decades earlier:



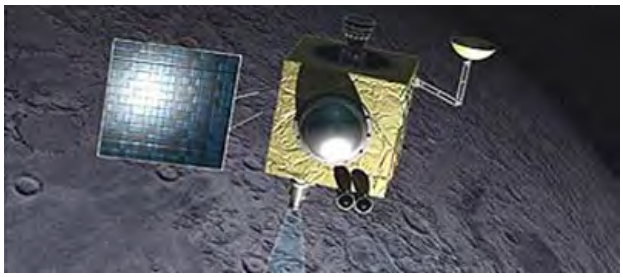
One of Kayuga's task was to search for water, especially in the polar regions. Its instruments dedicated to that search came up with no positive evidence of water. At face value, that may seem discouraging but perhaps the sensors were inadequate for that purpose. The question remains moot.

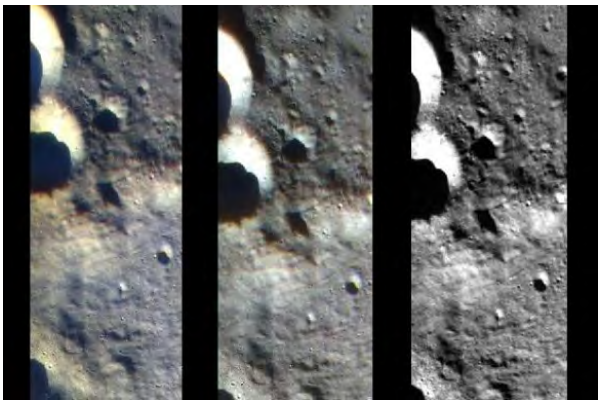
Kayuga ended its mission on June 10, 2009 by crashing onto the lunar surface. The crash was observed from Earth. This is a sequence seen through the Anglo-Australian telescope:



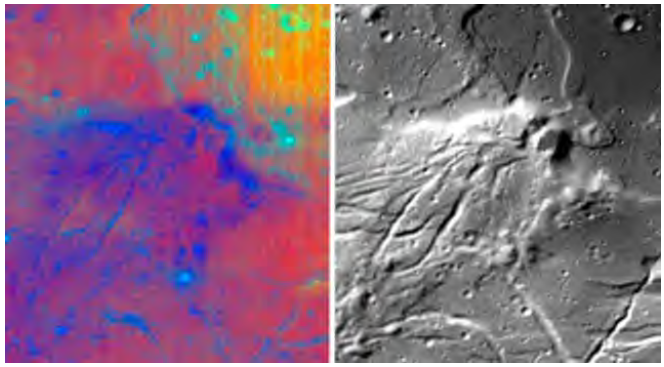
The impending crash itself was monitored by Kayuga as its approached the surface. Here are two of several images sent back just before contact:

India launched its first outer space probe towards the Moon on October 22, 2008. Named Chandrayaan-1, the spacecraft has multiple instruments from several nations. NASA's contribution are the MiniSAR and M3 (Moon Mineralogy Mapper, which will gather data in both the Visible and IR parts of the spectrum). Data from Chandrayaan (shown below) started to be received by mid-November:

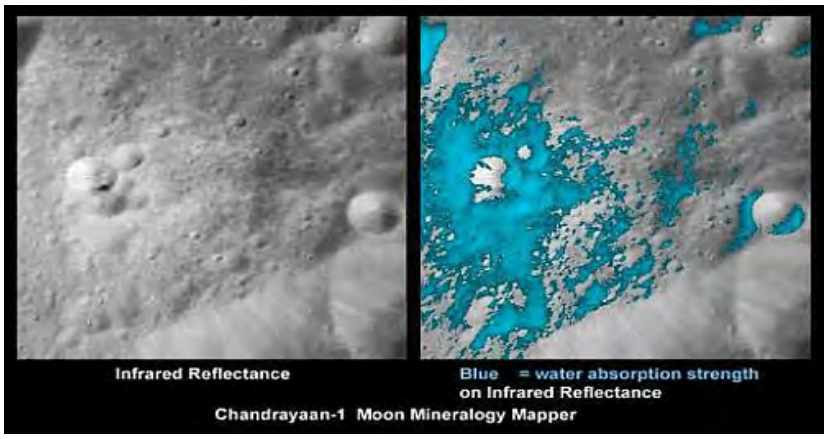


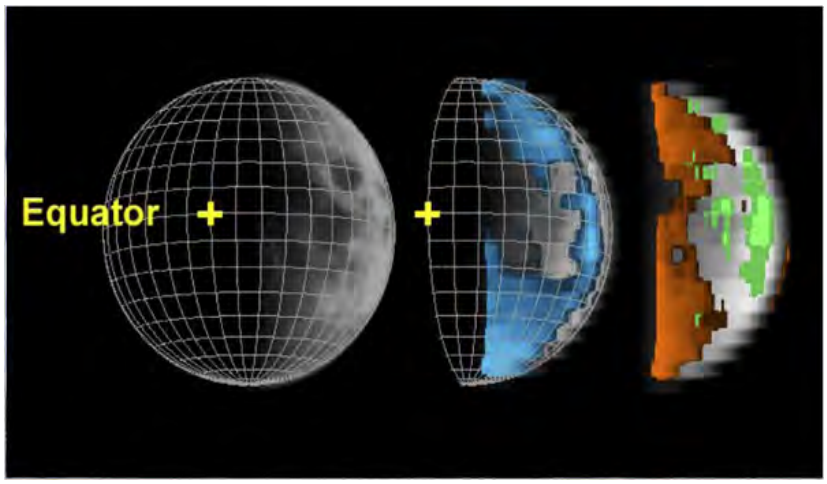


NASA's JPL has an instrument onboard Chandrayaan-1, called the Moon Mineralogy Mapper (m3). One of the first products released to the public is this image in which iron minerals are shaded green:



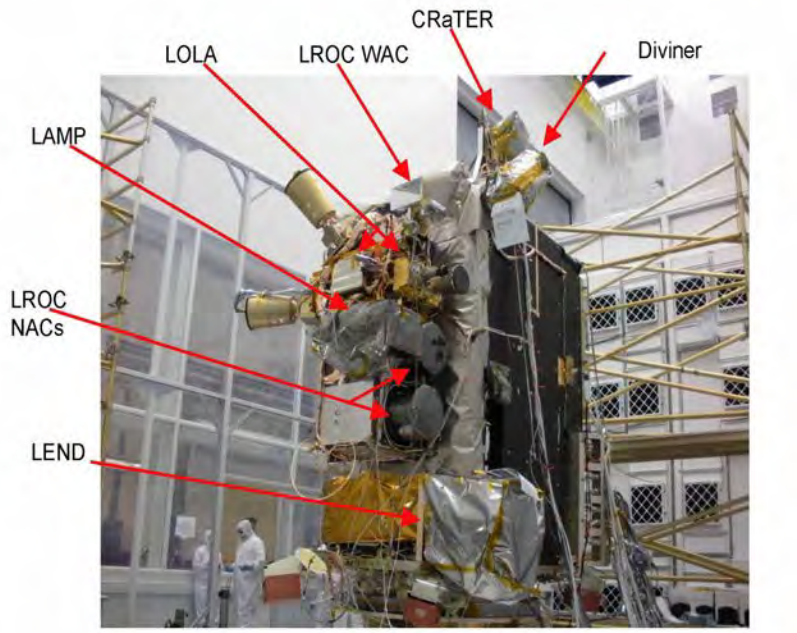
In September of 2009, NASA and several scientific groups made headlines with the announcement that water has been detected over most of the lunar surface. The amounts are miniscule by earth standards. The soil spread over any area comparable to a football field in size hold a cumulative amount of water sufficient to fill an ordinary drinking glass. But the discovery is important since it indicates that water does occur in non-polar latitudes. This illustration, based on Chandrayaan-1 data, indicates typical distribution:





NASA has mounted two missions in 2009, namely LCROSS (Lunar Crater Observing and Sensing Satellite), which will try to validate the evidence for water ice at the poles ultimately by crashing a probe onto the surface, and LRO, the Lunar Reconnaissance Orbiter.

The two satellites were successfully launched together by an Atlas 5 rocket on June 18, 2009 and entered orbit with perilune at 35 and 50 km respectively. The LRO spacecraft is pictured below. These are its onboard instruments:



The Cosmic Ray Telescope for the Effects of Radiation (CRaTER) will characterize the lunar radiation environment and determine its potential biological impacts. CRaTER will also test models of radiation effects and shielding, which may enable the development of protective technologies.

The Diviner Lunar Radiometer (DLRE) will provide orbital thermal mapping measurements, giving detailed information about surface and subsurface temperatures (identifying cold traps and potential ice deposits), as well as landing hazards such as rough terrain or rocks.

The Lyman Alpha Mapping Project (LAMP) will map the entire lunar surface in the far ultraviolet. LAMP will search for surface ice and frost in the polar regions and provide images of permanently shadowed regions illuminated only by starlight.

The Lunar Exploration Neutron Detector (LEND) will create high resolution hydrogen distribution maps and provide information about the lunar radiation environment. LEND can be used to search for evidence of water ice on the Moon's surface, and will provide space radiation

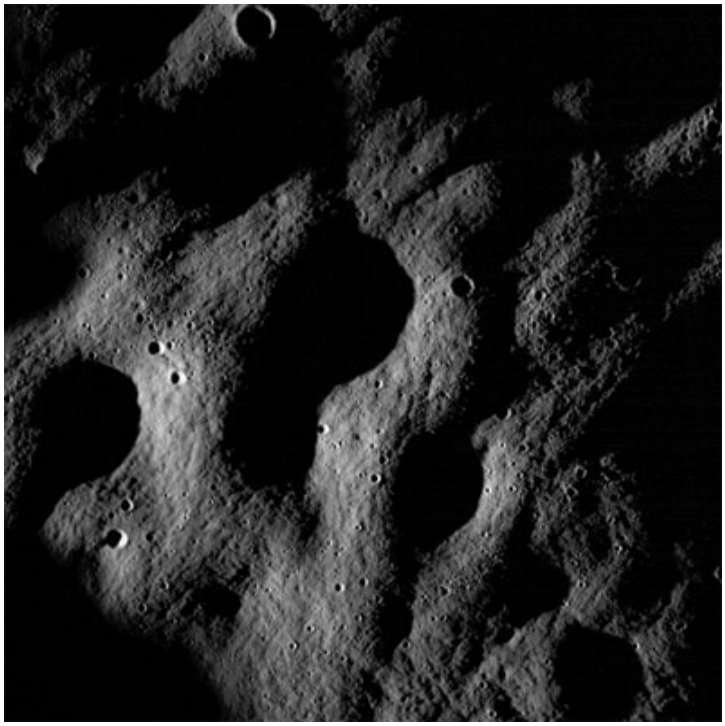
environment measurements useful for future human exploration.

The Lunar Orbiter Laser Altimeter (LOLA) will measure landing site slopes, lunar surface roughness, and generate a high resolution 3D map of the Moon. LOLA will also identify the Moon's permanently illuminated and permanently shadowed areas by analyzing Lunar surface elevations.

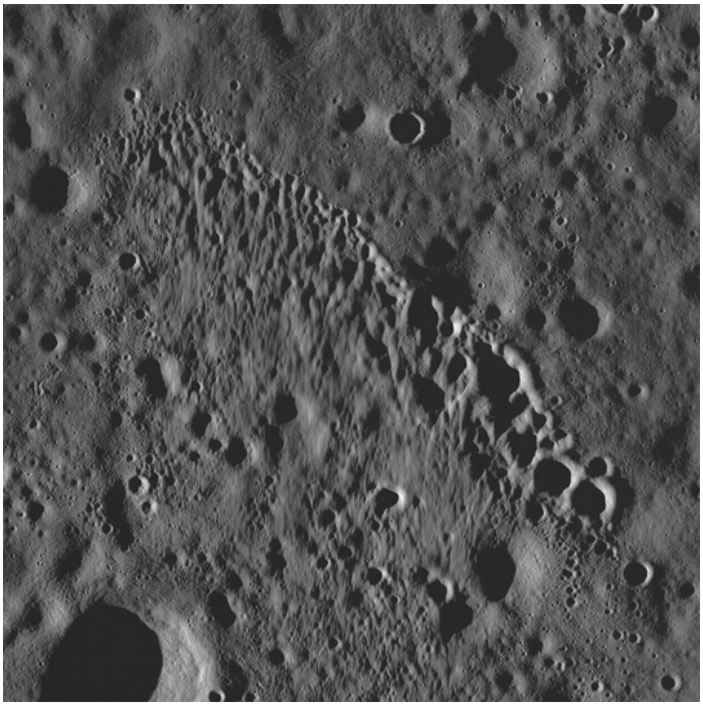
The Lunar Reconnaissance Orbiter Camera (LROC) will retrieve high resolution black and white images of the lunar surface, capturing images of the lunar poles with resolutions down to 1m, and will image the lunar surface in color and ultraviolet. These images will provide knowledge of polar illumination conditions, identify potential resources & hazards, and enable safe landing site selection.

The Mini-RF technology demonstration's primary goal will be to search for subsurface water ice deposits. In addition, Mini-RF will take high-resolution imagery of permanently-shadowed regions.

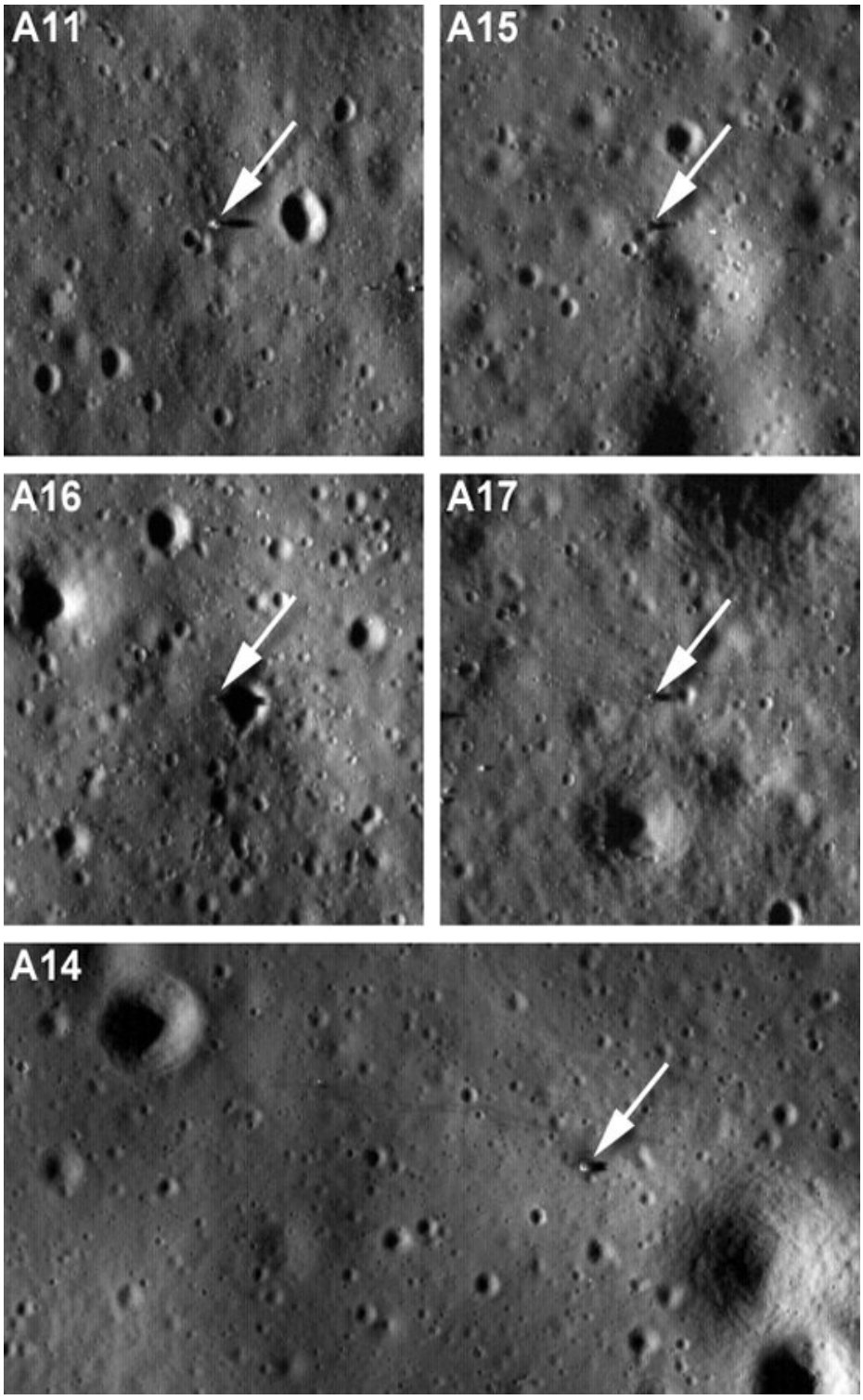
The first LRO image was released on July 2, 2009. It shows a scene 1800 meters across in cratered terrain within Mare Nubium:



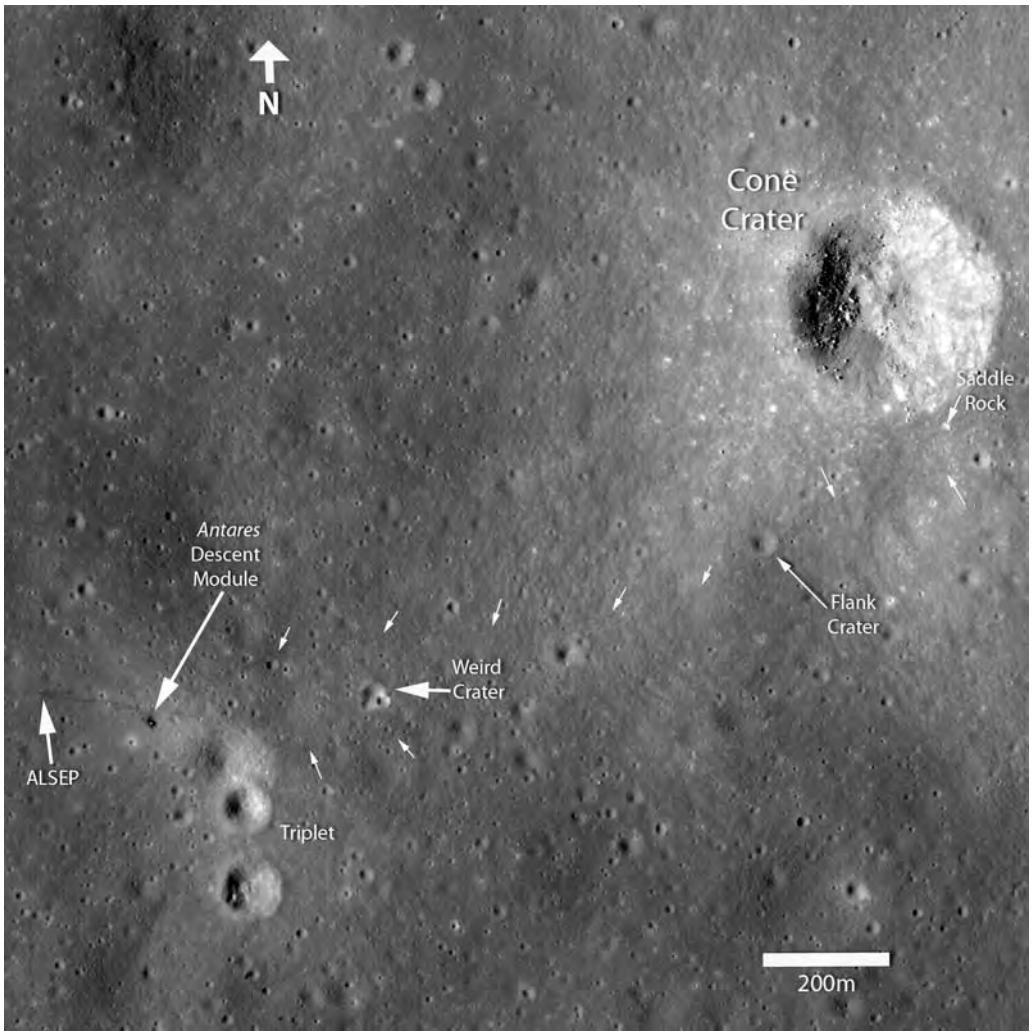
This image shows part of the crater Hahn; below it is a series of secondary craters formed from ejecta from a distant large impact crater:



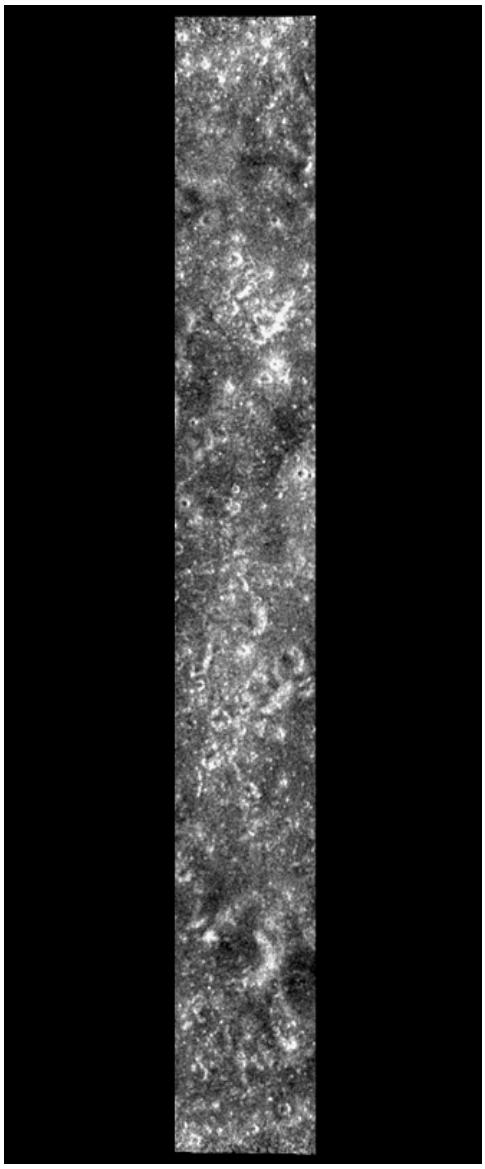
The montage below shows five of the Apollo landing sites; arrows point to the base equipment from which the LMs were launched.:



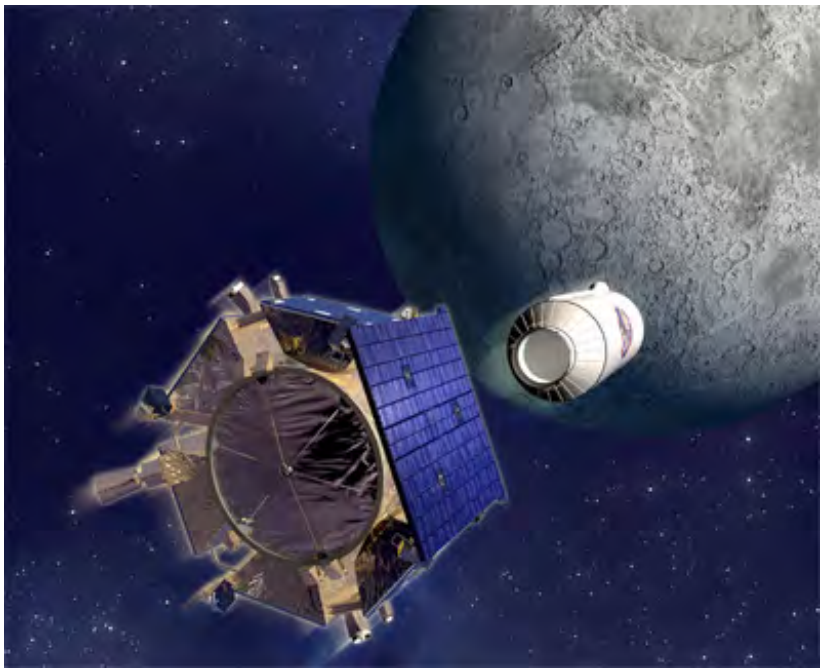
So, how good is the image resolution obtained by LRO? In this next image, again of the Apollo 14 landing site, some of the tracks made by the walking astronauts are visible as dark streaks:



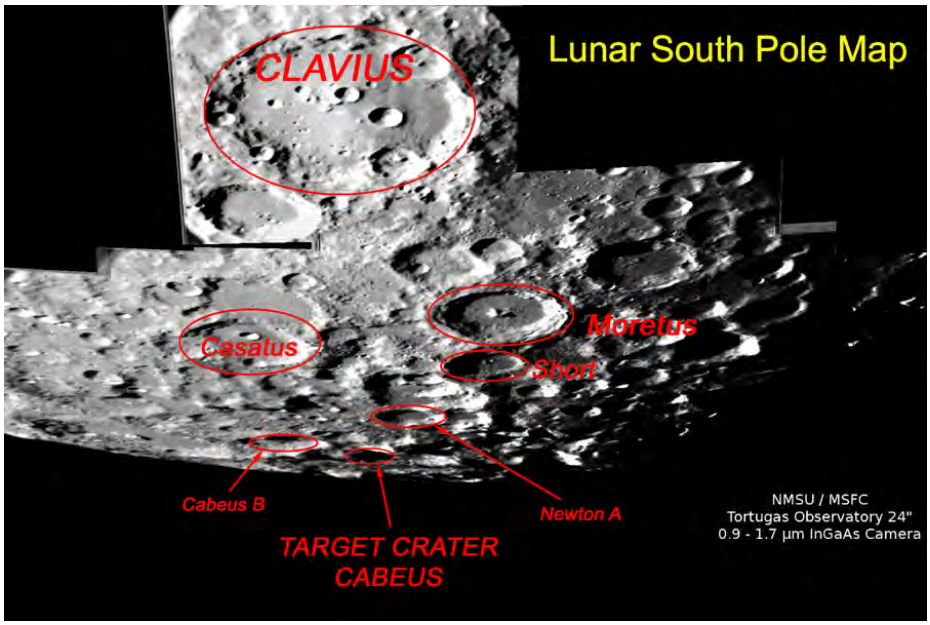
LRO carries Mini-RF, a SAR-type radar instrument whose main task is to search for ice, especially in the polar regions. Here is a radar strip of terrain near the South Pole.



LCROSS can be treated as a separate mission. After establishing orbit, sensors (visible and spectrographic) on LCROSS were programmed to examine craters in the polar regions seeking more detailed evidence that any of these contain water. The final event, at 4:35 AM PDT on October 9, 2009 was a one-time-only "shoot the moon", in which the Centaur rocket upper stage was made to crash onto the south polar surface at a small crater that looks promising as containing ice. The LCROSS remained nearby in orbit to observe the collision and the ejecta spray or plume produced by the first impact Following that, this shepherding spacecraft was deliberately crashed onto the same Cabeus A. Here is an artist's sketch of LCROSS as the impacting projectile is released:



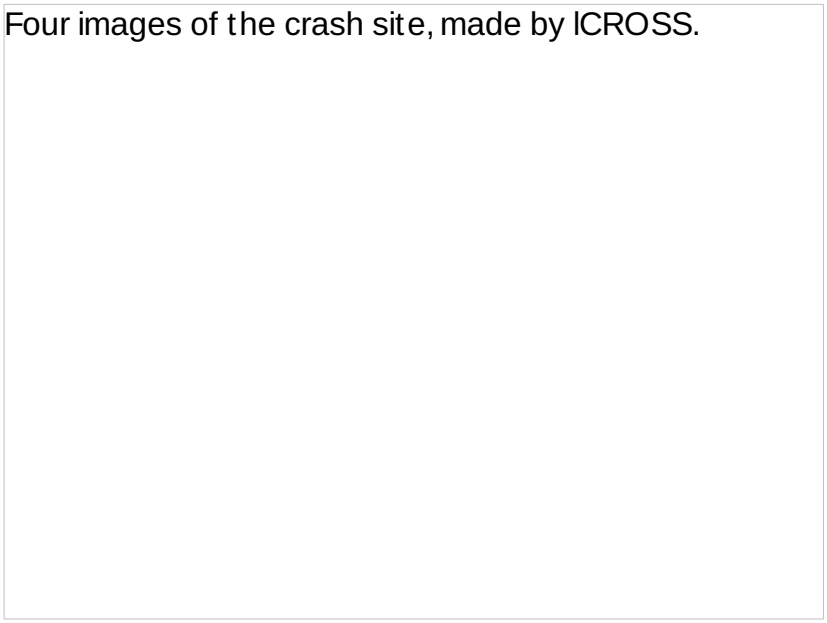
The target for LCROSS was the Cabeus crater, shown below. Note the nearby 'Short' crater. Regretably, it was not named to refer to the present writer of this Tutorial. Instead, it honors James Short, an 18th century British mathematician who built several innovative telescopes.



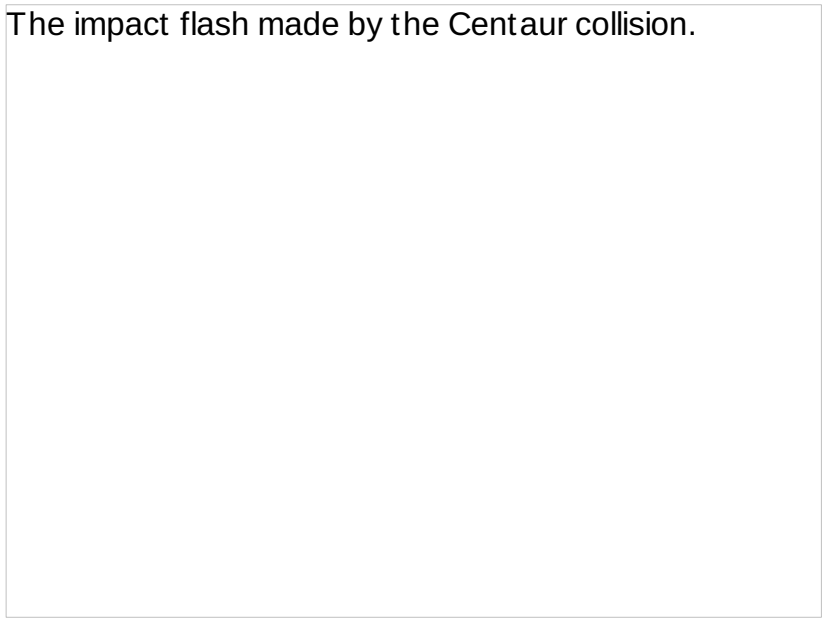
Expectations were high that the collision and resulting plume would be "spectacular". Professional and amateur observatories on Earth were trained on the polar region. No one in fact saw much of anything. LCROSS itself picked up a brief flash but no obvious plume. (The writer explains this as a consequence of the dark material in any such plume lacking notable contrast against the dark surface of the crater itself.) The spectrograph did pick up a strong signal but this will take awhile to be analyzed.

In the next three images, we show first the target crater, then a sequence of four images made by LCROSS just after Centaur impact, and finally an image which shows the Centaur impact as a bright flash:

Cabeus, the target crater.



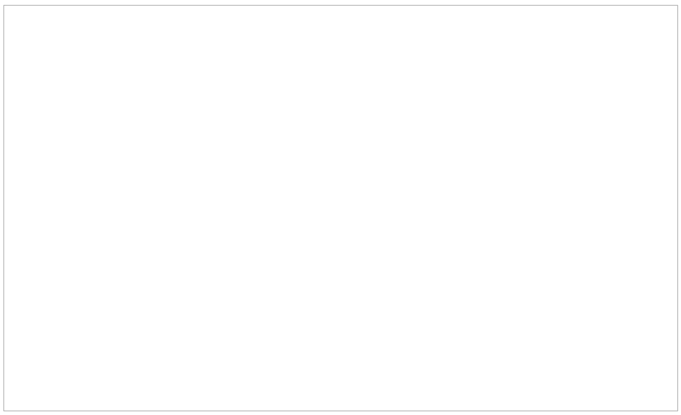
Four images of the crash site, made by lCROSS.



The impact flash made by the Centaur collision.

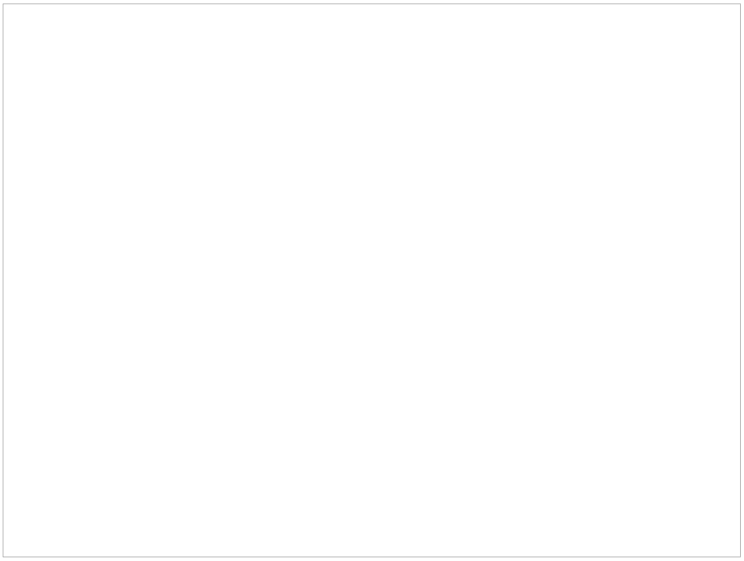
Some two weeks after launch, an image of the faint plume was processed and released to the public. To make it visible, the scene had to be strongly contrast-stretched, making the lighter tones almost white. Here it is:

Another processing, which narrowed the contrast range, shows how little difference there actually was in the tone level of the plume compared with its shadow background:

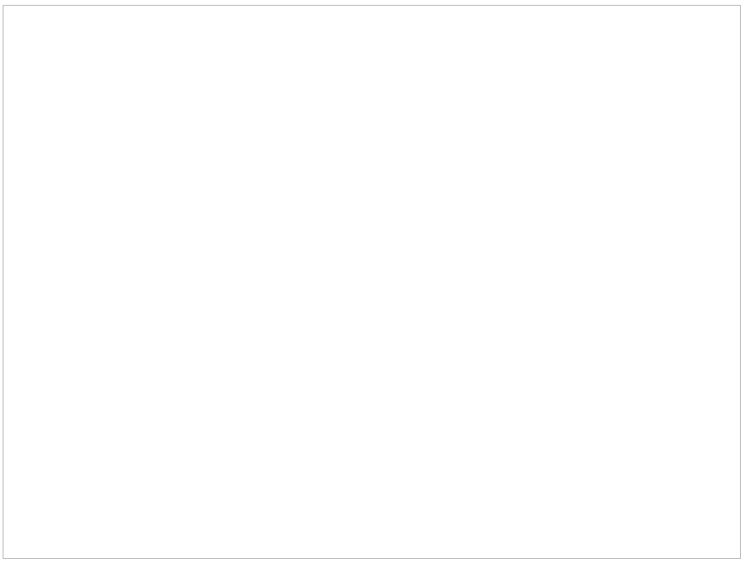


The resulting crater has been imaged. It is at least 20 meters wide, with a surrounding ejecta blanket between 100 and 200 meters in width:

On November 13, 2009 at a press conference held at Ames Research Center in California, the LCROSS Science team announced (with great enthusiasm) that instruments on board LCROSS had detected even more water (greater amounts than in some terrestrial deserts) than anticipated. The preliminary analysis pins the amount to be at least 24 gallons. The water appears to be present as discrete ice particles scattered within the surface soil of the shadowed part of the Cabeus crater. This is the spectrum for the thermal Infrared instrument, with the yellow bands representing the bands related to the water.



Although not as pronounced and definitive, the UV spectrum also shows water peaks.



The bottom line: **THERE IS DISCRETE WATER ON THE MOON**. Whether it is confined to ice in the polar regions, or is more generally distributed over lower latitudes of the lunar surface, is yet to be determined. But the implications for water being available at future lunar bases is distinctly *positive*.

The lunar probes shown above are all providing new data that will help in planning for eventual manned Moon missions. The question now being asked: Is such a resumption of exploration by humans affordable in the current world economy?

Humans Return to the Moon: The (now mothballed) Master Plan

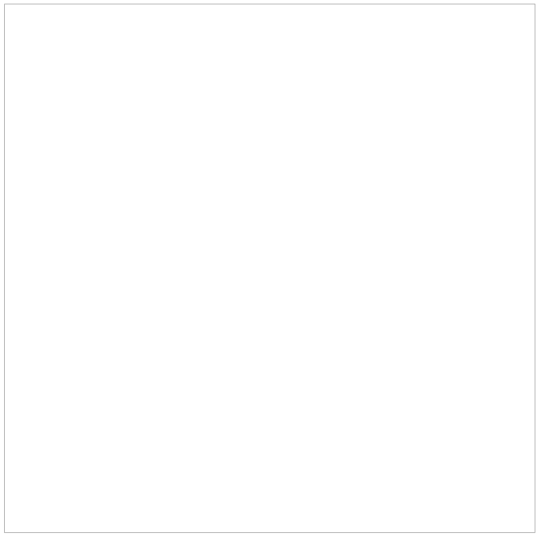
As mentioned earlier in this Section, in the mid-2000s NASA and other space agencies started planning a manned return to the Moon under the mandate given by Pres. George W. Bush (which, unfortunately, has now been canceled by Pres. Barack Obama). Despite this reorientation of the space program, it is instructive to review below the master plan for America' return to the Moon, since someday it might be reactivated:

A new, more versatile Space Transportation System will be needed. The first landings would probably be more like the Apollo ones but in time it is hoped to establish a permanent (or at least long term) lunar base where astronauts can subsist and explore for extended stays. Four things are essential in making a safe, flexible base: 1) a means of replenishing oxygen; 2) water; 3) source(s) of power; and 4) suitable shielding from extralunar radiation.

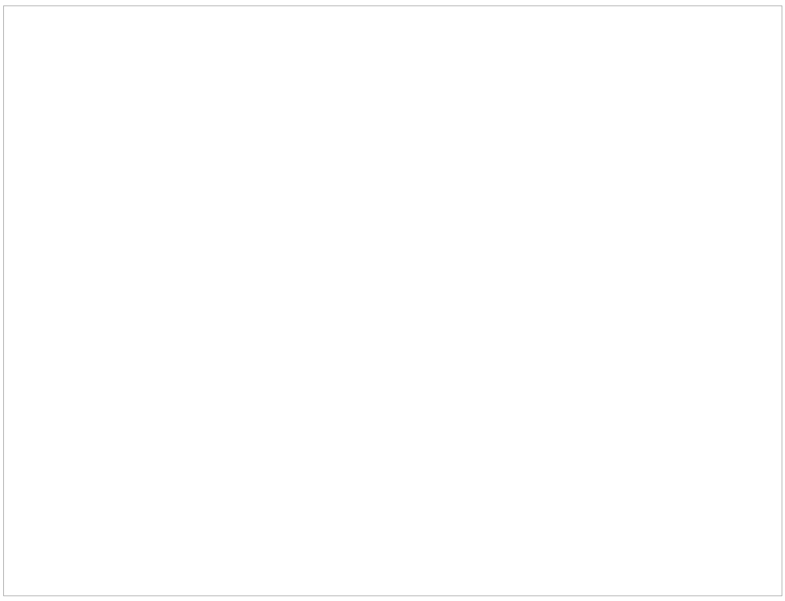
Oxygen, in principle, is extractable from the lunar silicate minerals but a reliable, practical means of obtaining this is yet to be worked out (in May 2005, NASA issued a Call for Proposals for innovative solutions). Water can, in part, be recycled from sources (such as astronaut urine) brought with the explorers. But, if substantial water is found near the polar regions, extraction should not be too difficult - thus the base would likely be located at high latitudes. Power requirements can be met with nuclear generators and/or with efficient solar arrays. Shielding may prove difficult since the base units (presumably separate from the landing craft) need to be of light materials. Still, growing experience should aid in selecting radiation-absorbing outer components of the base.

There is another strong argument for selecting polar regions for the base besides the water potential. Placing astronomical observatories at either or both poles would allow almost ideal observing conditions (for some applications better than the present Hubble Space Telescope since systems and components would be state-of-the-art). Nearly all of both celestial hemispheres would be accessible, whereas locating an observatory at lower latitudes would have some light interference from earthshine. But exploration would be curtailed somewhat by dependence on a polar station.

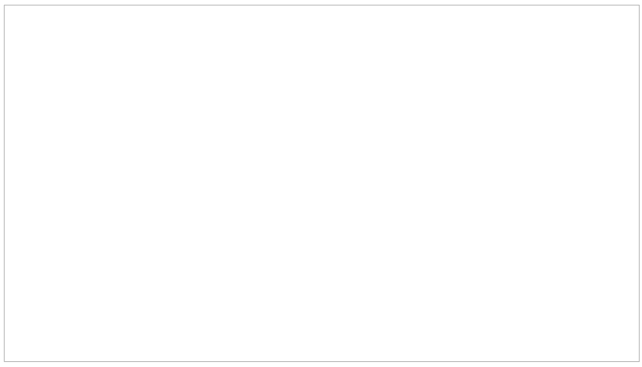
A recent National Academy of Sciences report offers another cogent argument for resumption of Moon visits: The observatory that would eventually be built would be of immense value in astronomical studies and in continual monitoring of the Sun. But, even more valuable to earth-dwellers, 21st Century technology operating through telescopes would produce copious data on the Earth itself. While geostationary satellites can do some of that, the fact that any part of the non-polar Earth is bathed in sunlight every 24 hours makes "observation on demand" feasible. The value of the Moon as a platform was demonstrated during Apollo 16 when astronaut John Young pointed a geocoronagraph towards Earth, getting this image that shows a shroud of low density excited hydrogen around much of the planet:



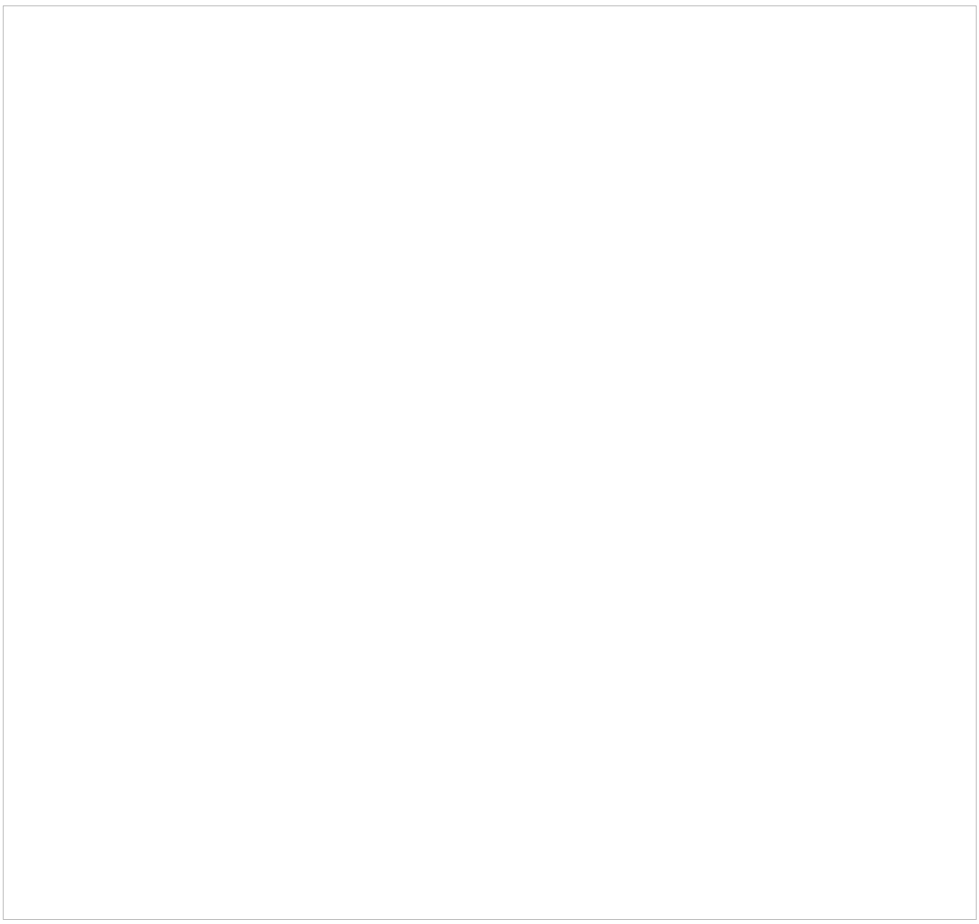
Establishment of a Moon base would be a giant step in mankind's renewal of outer space exploration. Among its benefits, it could serve as the launching site for a trip to Mars. On September 18, 2005 NASA made its first public announcement of how its approach to the Moon landings (and probably Mars later) will be made. There is a striking similarity to the Apollo approach in that landing craft will be on a large multistage rocket, with the main thrust section falling back to Earth after putting the manned vehicle on its journey. This vehicle and a companion for sending material to build a lunar base shown here, with other existing vehicles side by side for comparison:



A closer look at the rocket that would carry a crew of into lunar orbit is pictured here. At its tip is the Crew Exploration Vehicle (CEV).



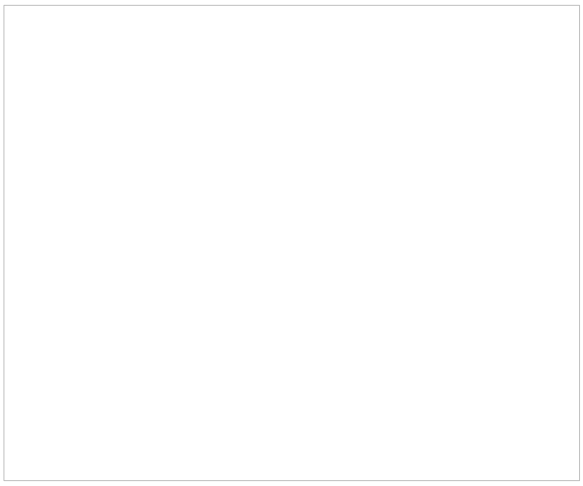
This panel diagram shows the sequence of events or stages now envisioned in the current plan for renewed lunar exploration:



As the lunar trip gets underway, the Departure Vehicle (jettisoned after burn) and Lander group have mated with the Crew Exploration Vehicle (CEV) capsule, as shown here:

The Service module remains unmanned during the days spent on the lunar surface. The lunar landing craft, housing all 4 astronauts (but, eventually, able to support 6 astronauts), as envisioned on the Moon's surface, is displayed here:

The landing units descend to the surface, much like during the Apollo program, with the larger unit consisting of a braking rocket and fuel. After the lunar stay is ended, the upper crew unit fires its rocket to put it into orbit and eventual docking with the service unit.

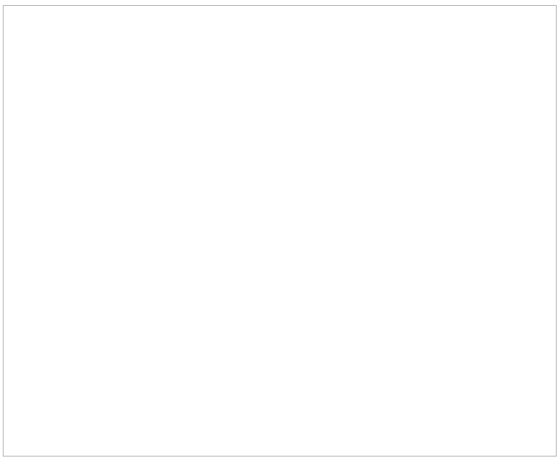


The schedule calls for the CEV system, without the Lander units, to be ready to fly sometime after 2010. It could replace the phased-out Space Shuttle program and will be NASA's means of servicing the International Space Station. This will provide extended experience in CEV use up to the first Moon landings.

The CEV docking with the ISS

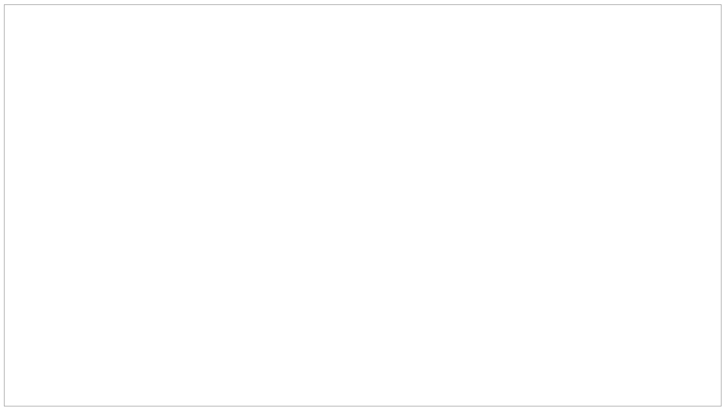
In the master plan (now in 'limbo') a crew of 4 would descend on the CEV and stay for (at least) a week. Over time, the stay will be longer as the astronauts build a lunar base capable of sustaining the mission for weeks to months. This will provide the needed experience for prolonged missions that would take place on Mars at a later time. When the crew returns to Earth in the detached capsule, it will have the capability of landing either on land or at sea. If no

serious damage occurs, the CEV can be used up to 10 flights.

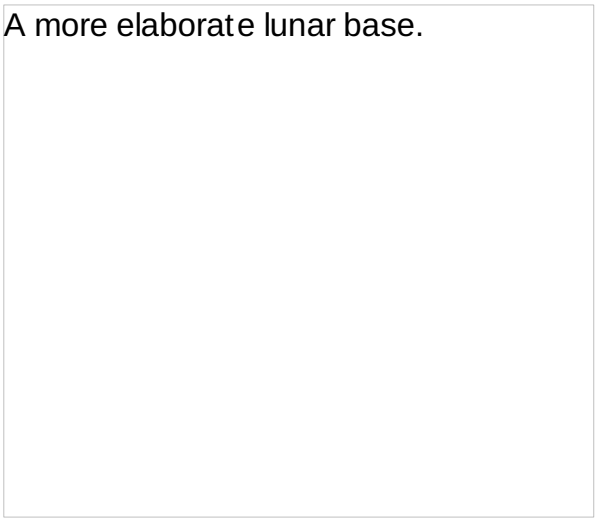


NASA, and even its critics, together recognize that a Moon exploration resumption followed by Mars exploration (which would gain from the lunar experiences) may be vital to keeping the American space program healthy enough to press forward, rather than wither and diminish by loss of dedicated personnel.

Imaginative futurists being a common breed today, one could predict that no sooner did NASA announce plans for a Moon Base, the Internet would start accruing many "artist's conceptions" of the layout of such an endeavor. Here are two:



A more elaborate lunar base.



Two active websites offer more discussion of such a base: 1) Colonization of the Moon, and 2) Solar Voyager.

Summary of Lunar Exploration Science

This subsection includes both a review of the sequential or evolutionary history of the Moon and

a survey of ideas pertaining to the origin of our lunar satellite.

19-23: List or mentally note what you think were the principal findings of the Apollo program? ANSWER

The list you created - and that in the answer - certainly does not tell the whole story. Here are some key items that complement the listing:

1. Probably the top achievement, by consensus, is this: that human intellect and commitment combined to make the technical feat of astronauts reaching and landing on the Moon - and RETURNING safely - six times stand out as perhaps the greatest accomplishment of mankind to date.

2. Close behind is the singular success in the face of dire adversity of bringing back the Apollo 13 astronauts to Earth after the near fatal catastrophe that almost doomed them to a dire ending in outer space.

3. The many orbiting and lander missions leading up to Apollo 11 proved the value of unmanned flights designed to gather critical data as part of an overall exploration program.

4. The question of the Moon's origin has been settled by analysis of lunar samples, which disclosed a strong kinship with Earth such that the Moon had to be derived from its terrestrial parent - most likely from a huge impact event (see below).

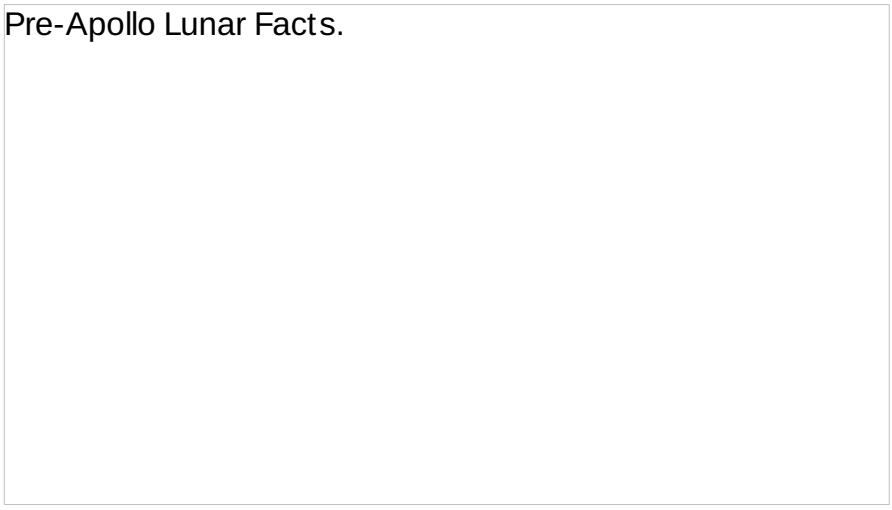
5. Impact and volcanism, which dominate the lunar surface, are adjudicated to be among the most fundamental of planetary processes.

Another of these top achievements is/are model(s) of the lunar interior. We cite this diagram that is one of the early popular versions (Dr. Anthony Ringwood, of Australia) of the Moon's structure. By now, you should have learned enough to explain the meaning of each major layer in the outer part of the Moon (note: the quartzo-feldspathic layer at the top proxies for the felsitic rocks typified by sample 12013).

A similar but more recent model assumes the outer half of the Moon melted - forming the so-called "magma ocean" - early in its history and then underwent differentiation to produce the present general layering:

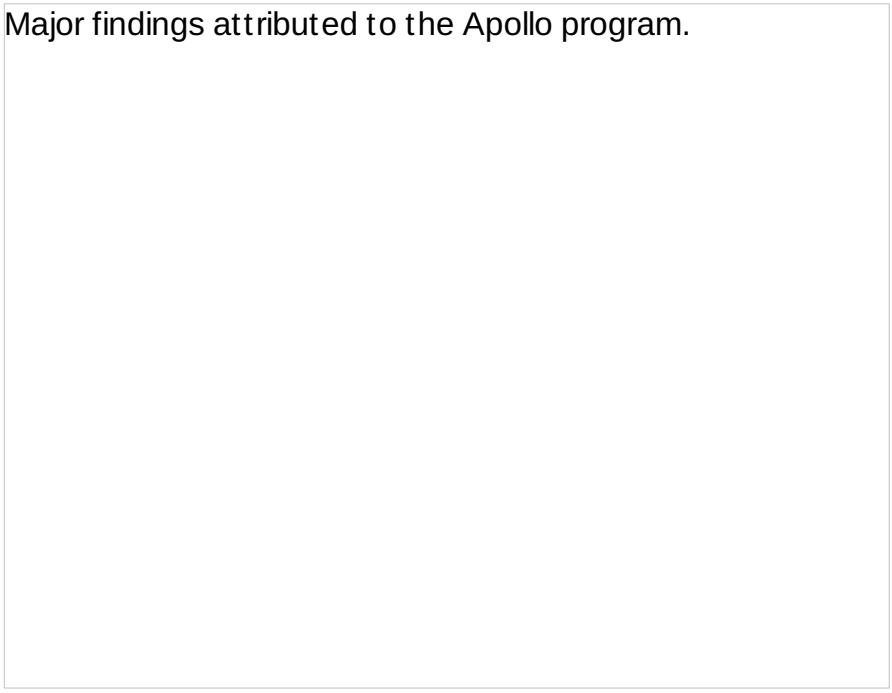
This brings us to two general topics: first, a summary of lunar evolution and then a survey of the Moon's possible origins. The first and second topics are both included in an Internet site that considers an Apollo-based Evolution of the Moon. written by Harrison H. ("Jack") Schmitt. Some of the ideas and illustrations from that site are used here.

We begin that review of the Moon's history or evolution by showing first a chart that summarizes what was known prior to Apollo:

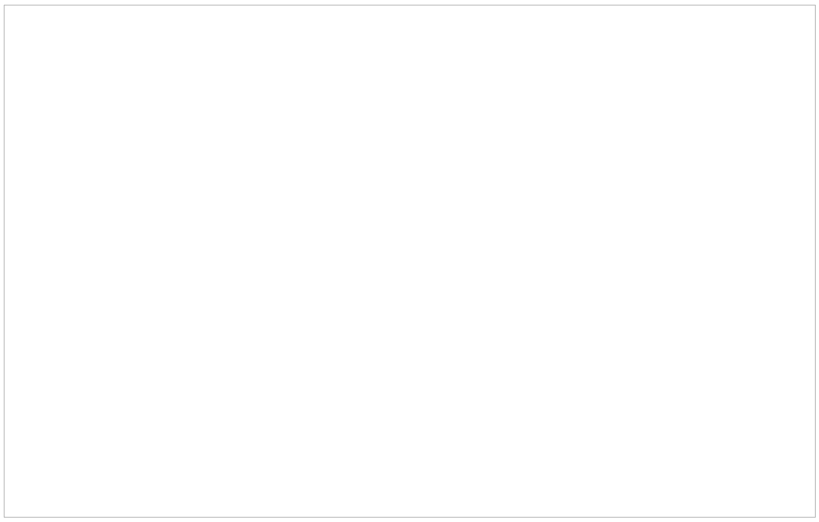


These salient points were determined both from Earth-based telescope observations and from lunar orbiters and landers. The presence of a lunar soil or regolith was confirmed by the Surveyors. The next chart encapsulates the main information on the time-marked evolution of the Moon arrived at from all sources utilizing both Apollo human observations and lunar sample analyses by Principal Investigators and other scientists:

Major findings attributed to the Apollo program.



The ages cited in the above chart are determined from relative cratering frequencies, calibrated by radiometric dating of the Apollo samples. In this model, a feldspar-rich moonwide crust forms from the magma ocean. Two periods of intense bombardments by asteroids, mini-planets, and comets produce major basins which tossed materials from the highlands crust as first and then also mare surface over most of the lunar surface. These formed eventually consolidated deposits of large to small blocks and fragments making up interleaved "ejecta blankets" from 100s of meters to several kilometers thick. Off-loading and other processes mobilized subsurface rock (largely basaltic [high Fe, Mg, Ca and low Si] in composition) that melted and invaded the surface filling the maria and the interiors of larger craters. Cratering began early in lunar history, reached a maximum around 4 billion years ago, and has tapered off since. This next chart describes the changes and conditions associated with the Moon's outer reaches at the outset of the main period of basaltic lava extrusion:



The ideas expressed in these charts can be presented in a different way, as shown in this timeline chart (again, courtesy of H.H. [Jack] Schmitt):

Timeline chart for major lunar events.

This diagram has a deceptively inconspicuous word, "cataclysm", which calls attention to a major defining event in lunar history. First proposed in the early 1970s by Dr. Gerald Wasserburg of Cal Tech, the Lunar Cataclysm (also called the "Late Heavy Bombardment [LHB]), occurred during a hundred million year span centered around a 3.9 billion year age. Wasserburg found that glasses in lunar breccias had prevailing radiometric ages (Argon-40/Argon-39) in that time frame. These glasses are best interpreted as resulting from impact processes. He, and later workers, attributed the impacts that caused large basins such as Imbrium, Serenitatis, and Crisium, together with many of the smaller craters, to swarms of asteroids striking the lunar surface. (Later work on lunar meteorites [found mainly in the ice surfaces of the Antarctic], which represent sampling over the entire lunar terrains, have confirmed this clustering of ages around 3.9 b.y.) The bulk of the asteroids are believed derived from the Asteroid Belt between Mars and Jupiter; these asteroids are postulated to have been perturbed out of their prevailing orbits by a "resonance" process related to Jovian gravitational influence.

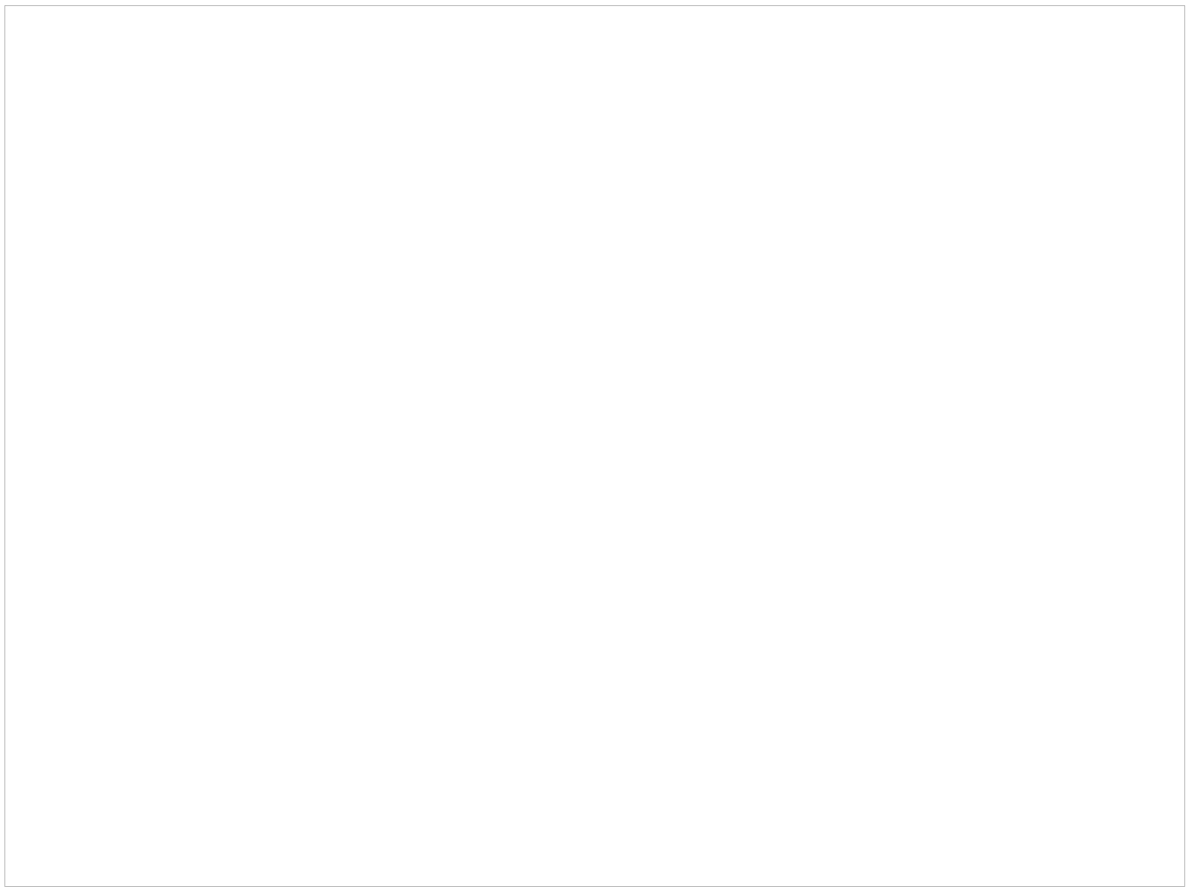
The LHB has a strong corollary implication for the Earth itself. The asteroids would also have hit Earth during the bombardment period. This would have had a profound effect on the early crust of Earth, but since almost all of that crust has since been destroyed by subduction and erosion the evidence for a corresponding terrestrial LHB has been erased.

The Moon has obviously changed its appearance over time. This set of three paintings depict (left to right) 1) the surface following the major bombardment up to the present, 2) the surface after the Imbrium lavas were emplaced, and 3) the early lunar surface before lavas entered the impact basins.

Three stages of lunar evolution.

Jack Schmitt has a most interesting Internet site in which he uses various illustrations to show

the progressive development of the Moon from its earliest history through the late stages of basaltic emplacement around 3 billion years ago. The site, accessed here, is in .pdf format, which requires Acrobat Reader. His figures on lunar evolution take into account much of the research done in the last 30 years, so it is worth a try to move through this site. To entice you to work through his sequence, we put up here the last (and most complicated) of his model diagrams which has added mare basalt emplacement from a period ending 2 billion years ago. By going through his Lecture 8, you will see the evolutionary steps taken to get to this stage (after which the major changes are associated with small to large impacts).



Another summary of lunar history has been put online by Prof. Stephen Dutch of the University of Wisconsin-Green Bay. From his web site we have extracted this figure and key:

Steps in the developmental evolution of the Moon.

A. Initial Accretion of the Moon, probably from debris launched into Earth orbit by a mega-impact.

B. In the last stages of accretion, so much heat accumulates that the outermost 100 km of the lunar crust melts to form a magma ocean. Light feldspars rise to accumulate an anorthosite crust.

C. Late impacts excavate giant basins. One of the earliest is the South-Pole-Aitken Basin.

D. Mare Nectaris and other bains form.

E. Mare Imbrium forms.

F. Mare Orientale forms

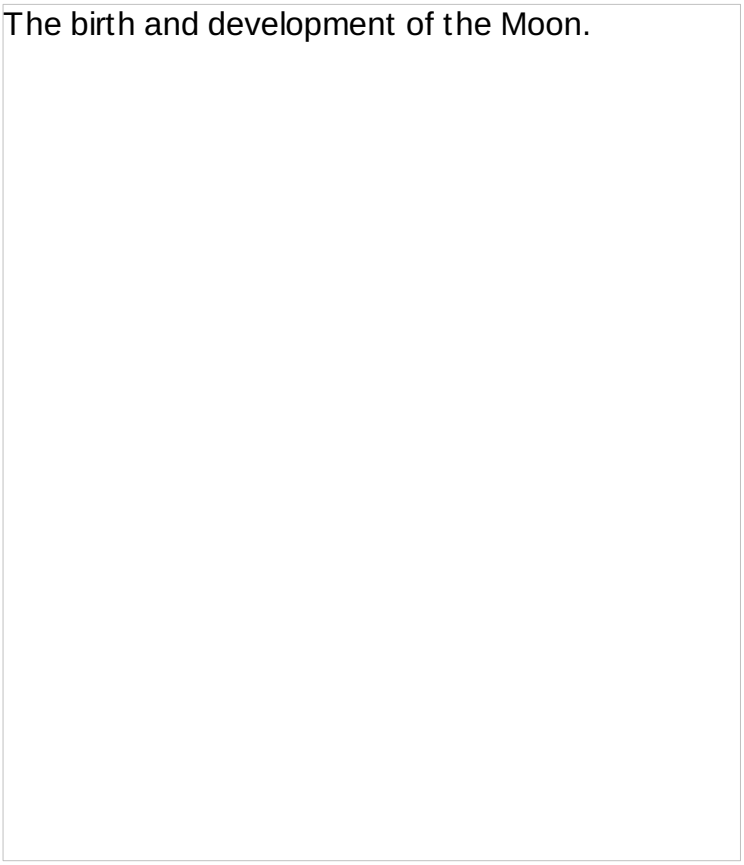
G. Mare basalts erupt and flood many of the impact basins.

H. Since 3000 Ma, only a few large rayed craters like Tycho and Copernicus have formed.

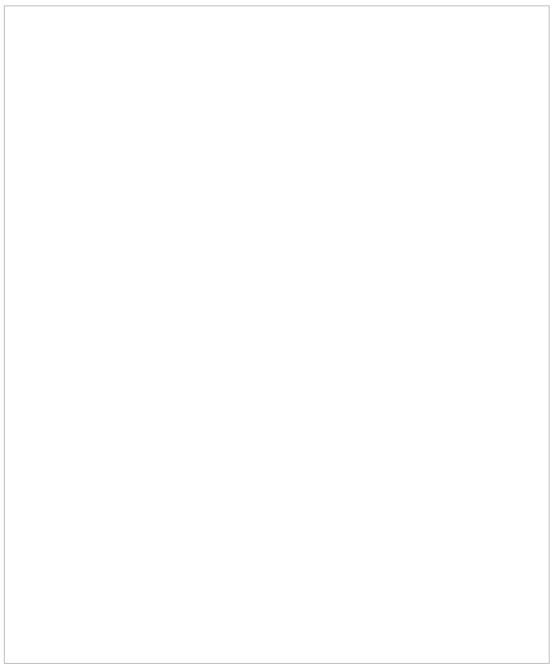
The origin (formation) of the Moon has always been a prime topic for conjecture and scientific insight among selenologists. Four main schemes for lunar origin existed before the Apollo program brought back lunar samples. One view had the Moon form from leftover debris as the Earth itself built up by aggregation. A second idea holds that debris which makes the Moon was tossed off the Earth in the latter's early days when our planet was spinning (rotating) much faster. A third proposal claims the Moon is a captured small planet once more distant from Earth. The fourth ascribed its formation to material wrenched from the Earth's outer crust by a massive impact leaving the Pacific Ocean Basin as a scar equivalent to a huge crater (a model that would need revision and probable discounting after the ideas of plate tectonics and continental migrations took hold). None of these hypotheses adequately explains the observed balance

between the combined angular momenta of the Moon and Earth which theory indicates remains constant since the two bodies became linked. Despite its greater rotation speed in the first few hundred million years of Earth's existence, this still is not enough to foster co-accretion. Nor is the speed sufficient to fission off the debris. But, that spin was too fast to allow capture of a passing body.

By the 1970s, with the Apollo data now in hand, impact had gained favor as an integral part of lunar formation. Several impact models has since been proposed. This diagram neatly summarizes this idea and the key features of subsequent lunar evolution:



All lunar genesis models are constricted by the two Apollo observations that the entire Moon is deficient in iron (Fe can be high in some mare basalts, but is very low in its interior, with no, or a small, iron core) and by the low percentages of the volatile elements sodium and potassium. That the Moon was derived from an impact of giant magnitude on the early Earth is supported by the strong similarity in oxygen isotope compositions in the two planetary bodies. The first model was developed by scientists associated with Harvard University. But, their head-on collision model has since come up with energy and compositional problems. The most recent variation on the general impact model is illustrated by the succession of steps shown in this diagram which is the result of a computer simulation of a huge impact into the protoEarth but oriented at that moment so as to glance against or sideswipe the outer layers of an Earth whose crust had not yet fully developed. Look at this computer simulation of such an event:



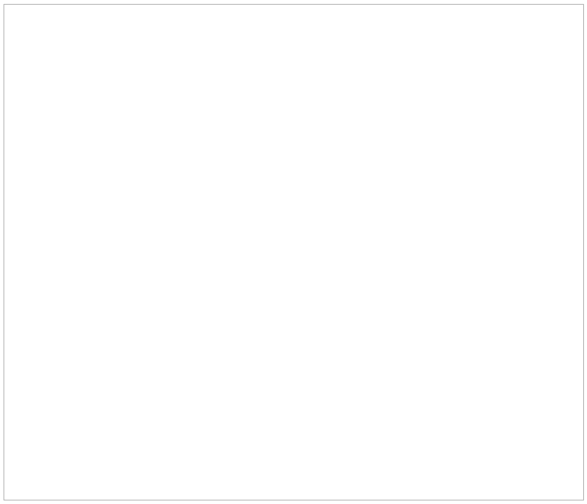
The model and some variants, collaboratively developed by scientists at the Southwest Research Institute (William Ward and Robin Canup; others) and the University of Arizona (A.G.W. Cameron, Jay Melosh, William Hartmann; others), considers the impact to have occurred late in the formational history of the Earth, but probably prior to the differentiation that formed an early terrestrial crust. At this time, a part, perhaps much, of the outer Earth may have been molten. A Mars-sized asteroid or small planet (about 10% of the present terrestrial mass) struck the Earth at a glancing angle. Although the Earth survived total disruption, much of the outer shell on one side was tossed into space, but held to the Earth by its larger gravity. The fragments in the ejecta plume are affected by rotational forces from Earth and within 24 hours have organized into a near circular orbit. In time these fragments (whose composition mirrors that of the primitive Earth's outer shell(s)) began to collide until the Moon was built up to its present size, large enough for it to have melted and reshaped into a sphere, developing an anorthositic crust. The Earth, still forming, healed its "wound", resumed its organization during subsequent remelting into a near-sphere, and went on to fully differentiate into the crust, mantle, and core that has survived to the present day.

The advantages of the swiping impact model are these: 1) a proper relation between Earth-Moon angular momentum comes out of the calculations; 2) the high heat of such an event boils off all water and some of the volatile elements sodium and potassium; 3) the similarity of refractory element composition between Earth and its satellite is explained; 4) only the outer mantle and any early crust are involved; 5) temperatures in a glancing event would have been higher (up to 18000° K); 6) a larger fraction of the Earth target would be ejected into orbit; 7) differences in composition could be due to incorporation of some of the impactor body, which likely varied somewhat from Earth.

The resulting Moon may have been much closer to Earth, perhaps as near as 29000 km (18000 miles). This first Moon would have appeared to occupy much more of the sky than today. It is now known that the Moon is receding at a rate of about 2.4 cm/year (around an inch), to its present average distance from Earth's center of 384000 km (240000 miles). Extrapolating back in time for 4.5 billion years yields this early proximity value (which, however, may exceed the Roche Limit - the closest distance two large planetary bodies can be without one at least being disrupted).

To close this subsection, there are literally thousands of informative and often exotic images of the Moon, taken by various remote sensors. Perhaps none can better convey the human emotions of having triumphantly landed astronauts on the Moon than this heart-throbbing photo taken by Michael Collins from the CSM of the about-to-dock LM containing Neil Armstrong and Edwin "Buzz" Aldrin, with Mother Earth looking so distant in the background, yet as history

shows returned to successfully by these intrepid Apollo 11 explorers and ten others who set foot on the Moon's surface (watched over by five comrades in orbit) in subsequent missions:



Two very readable popular accounts of lunar exploration are *The Moon Book* by Bevan M. French, 1977, Penquin Books, and *Lunar Science: A Post Apollo View* by S. Ross Taylor, 1975, Pergamon Press.

Reluctantly, we must take leave of our local satellite to begin an impressive journey through the Solar System. We start with the two innermost planets–Mercury and Venus.

PREVIOUS NEXT

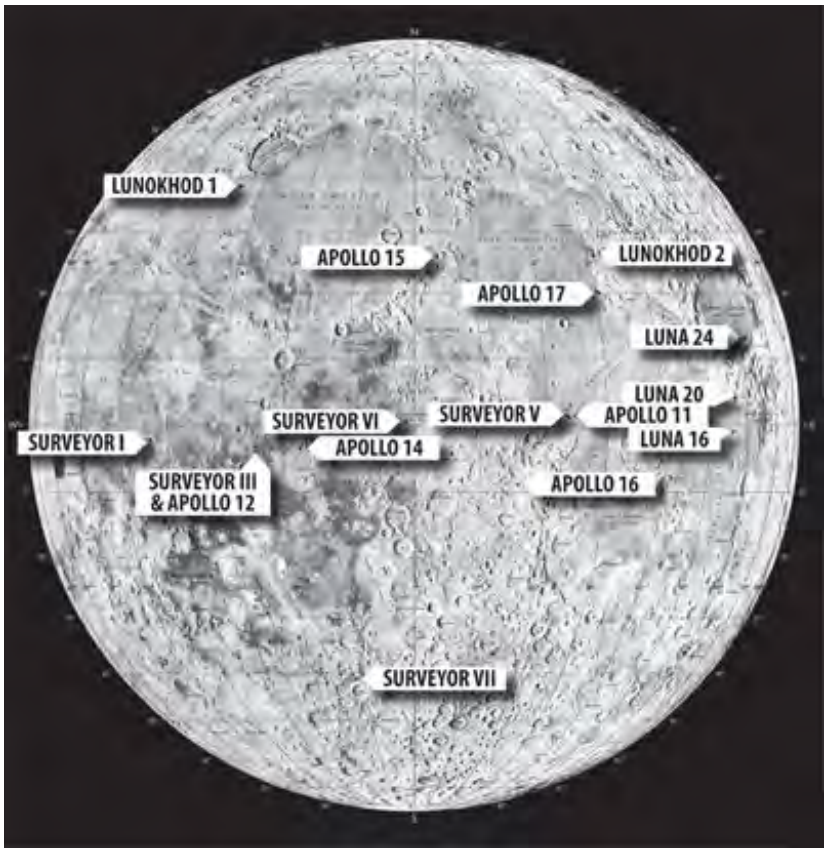
Primary Author: Nicholas M. Short, Sr.

The pinnacle of the lunar exploration programs of the United States - the response by NASA and many other organizations to the challenge of President J.F. Kennedy to land astronauts on the lunar surface and return them safely to Earth - was the highly successful Apollo program - a series of 9 flights to the Moon, six of which actually reached its surface. This page recounts the achievements of that effort - perhaps the greatest sojourn in human history. The collection and return of 810 pounds of rocks and soil was the top accomplishment. These rocks are mostly basalts, anorthosites, and breccias that are mineralogically similar to but chemically different from Earth rocks. The majority of lunar samples have formation ages between 3.7 and 4.2 billion years - older than nearly all terrestrial rocks in Earth's crust. Various instruments were deployed and extensive photographic documentation from both the ground and from the orbiting Command and Service Module added to the triumphs of the missions. But, the ultimate success was to prove, by doing, that space explorers can live and operate on other planetary surfaces.

The Apollo Program - Man on the Moon

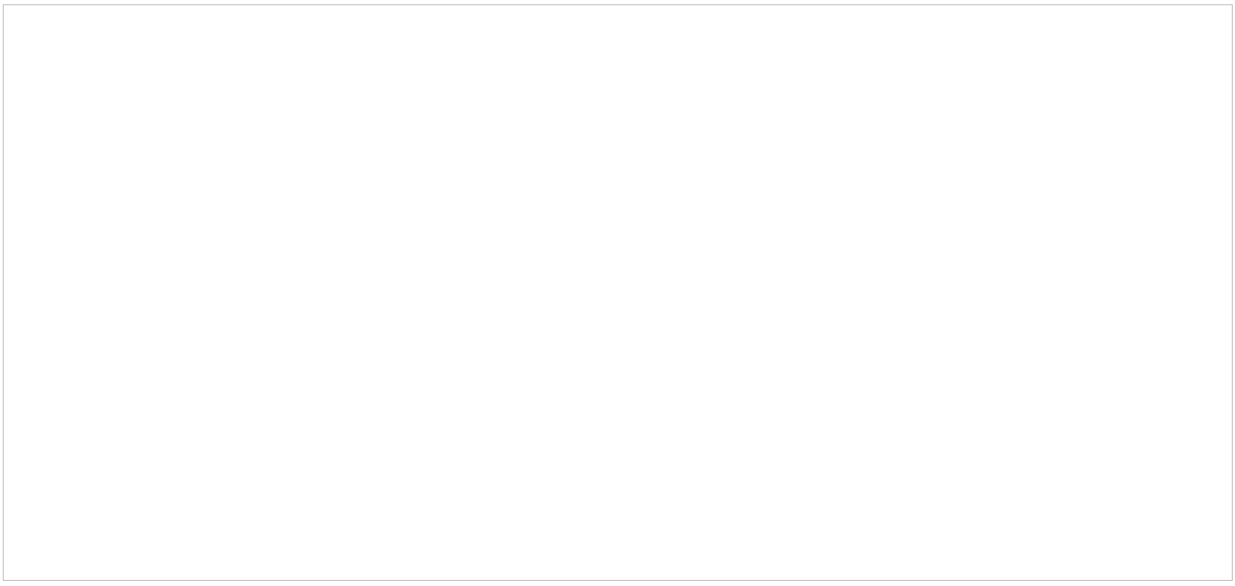
Either before or after reading through the next three pages, you may wish to consult an excellent online summary of the Apollo Expeditions to the Moon.

The culmination of the great push to the Moon, announced in 1962 by President John F. Kennedy as a major goal for all humankind, were the six manned landings between July 1969 and December 1972. The location of each site (A), along with the locations of the earlier Ranger (R) and Surveyor (S) spacecraft and the Soviet Luna (L) series of rovers, are plotted on this reference map of the Moon's front side:

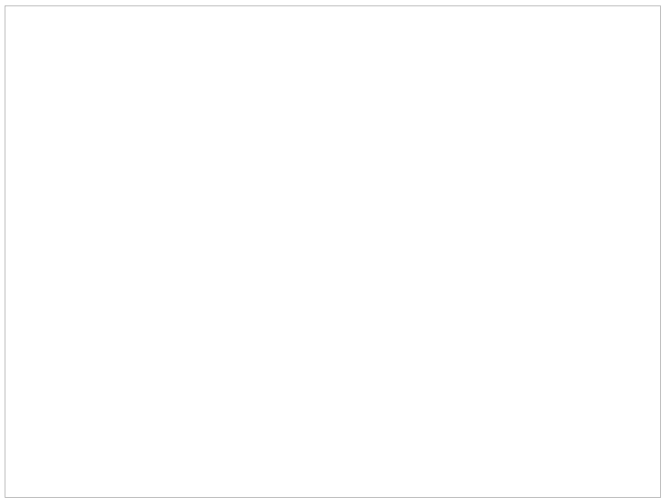


The Lunar Orbiters had earlier imaged each site at various times; examples are seen in this next set of panels:

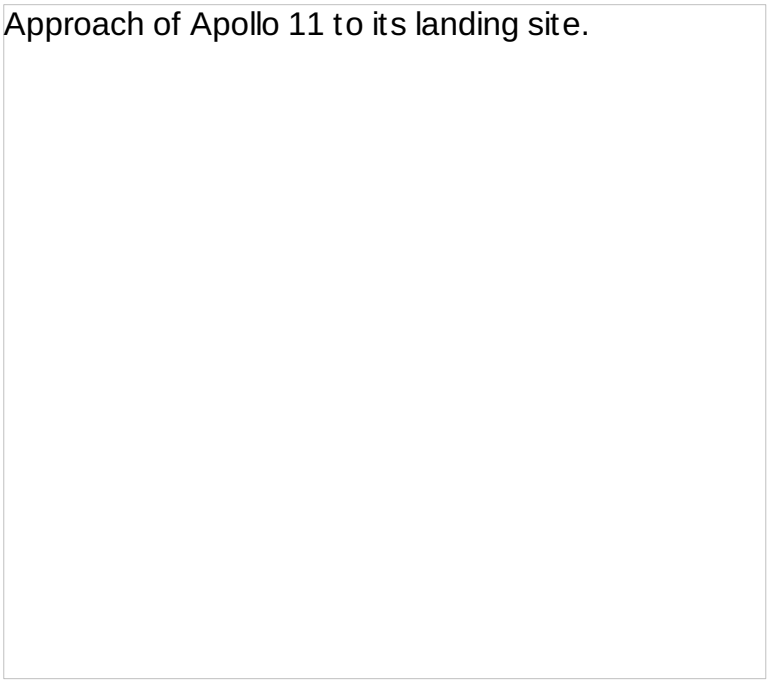
The following table summarizes the science rationale for visiting each site:



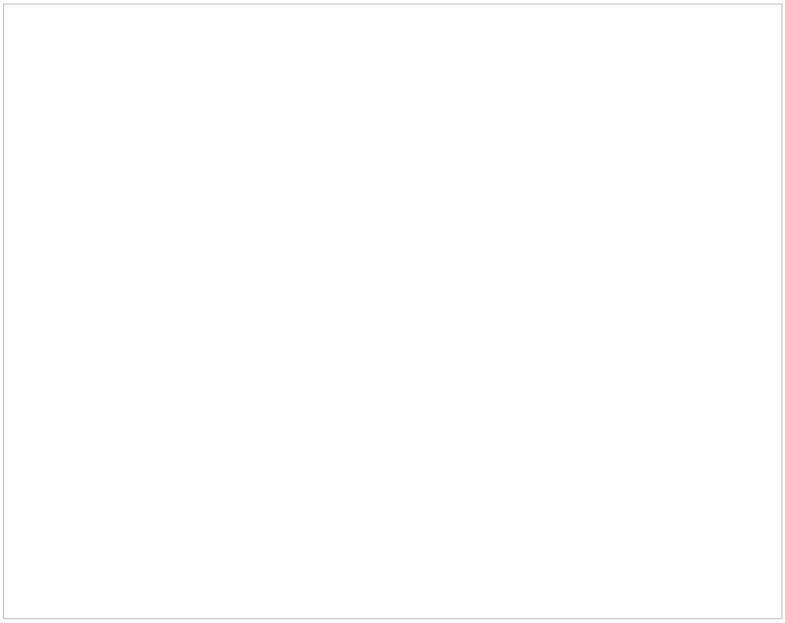
The Apollo 10 astronauts had previously reached and orbited the Moon without landing during Christmas of 1968. This emboldened NASA to send the first trio of astronauts, Neil Armstrong, Edwin "Buzz" Aldrin, and Michael Collins, in the summer of 1969 to attempt a landing near the south edge of Mare Tranquillitatis, chosen for its flatness and relative sparsity of larger craters. Here are the three "immortals" photographed before their journey that began on July 16, 1969:



Their final approach was captured by a photo out the LM window which showed both the landing site ahead and the many guiding landmarks (given quaint names as was the custom) visible ahead:



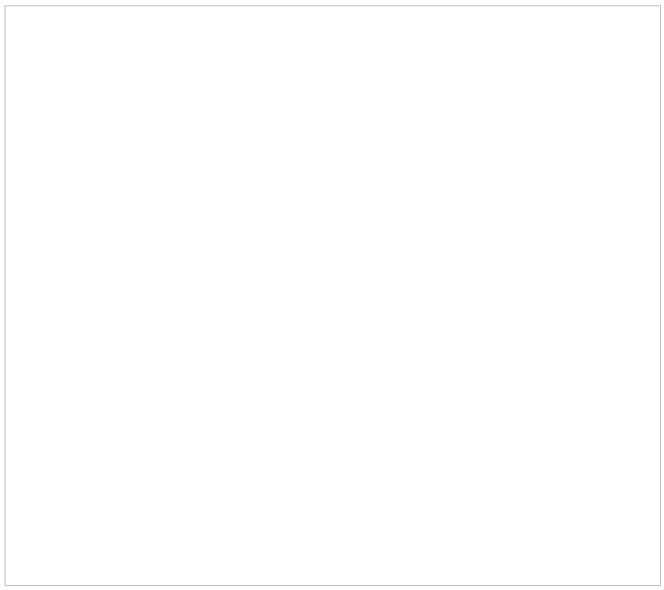
The touchdown point (LM) at Apollo 11 is plotted on this Lunar Orbiter photomap:



It was mid-afternoon (4:17 PM EDT) on July 20, 1969 that the Apollo 11 LM touched down safely on the lunar surface, much to the relief of the program people at Houston Mission Control and to an anxious but excited world. The descent phase is captured on this Space.com video site, narrated by Buzz Aldrin (it may not work on your computer, but try it; of course, you need to be connected to the Internet). There was drama and uncertainty in the final two minutes: a warning alarm sounded repeatedly (it was an overloaded computer) and descent fuel was down to its final minute). Here are six photos that show the scene sequentially just before touchdown:

The final approach.

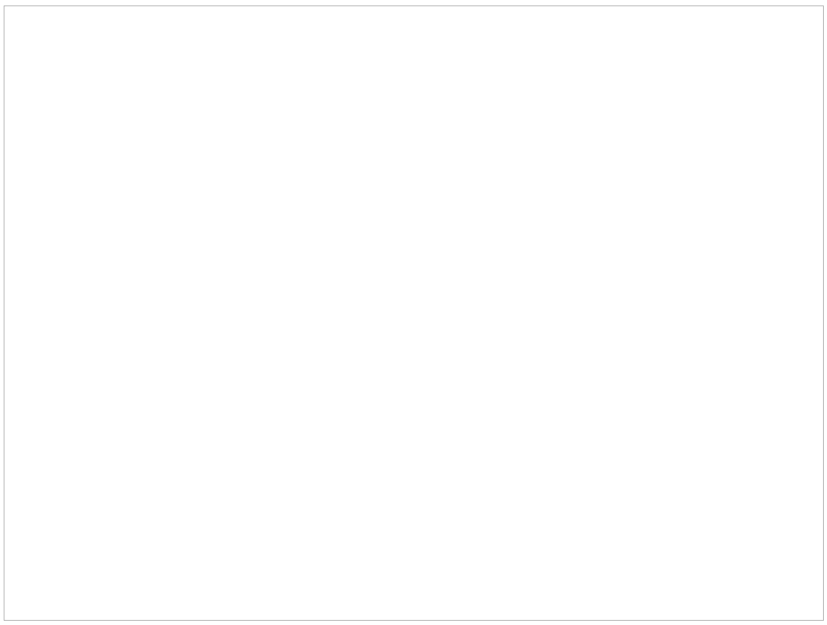
Six hours later came one of the greatest moments in human history, immortalized in this TV image of the first astronaut to descend to the surface. (If you have forgotten his name and his famed quotation, check the caption; also, face the question: If he were first out, who was holding the TV camera?) Of course, you should remember his "immortal" first worlds: "One small step for (a) man; one giant leap for mankind!" (The 'a' in parenthesis is what he intended to say but this is not heard in the broadcast audio.)



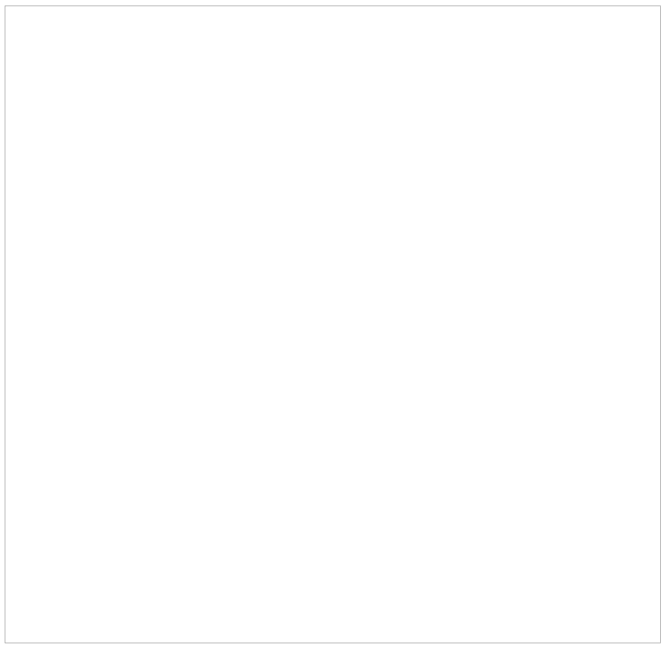
This next photo captures an indelible image of what most Anericans consider one of the greatest triumphs of the science and technology of the USA, and a tremendous source of national pride - the planting by the astronauts of the American flag (appearing unfolded by making it as a metal sheet which preserves the "flapping" in the breeze) in the lunar soil.

Buzz Aldrin and the American flag.

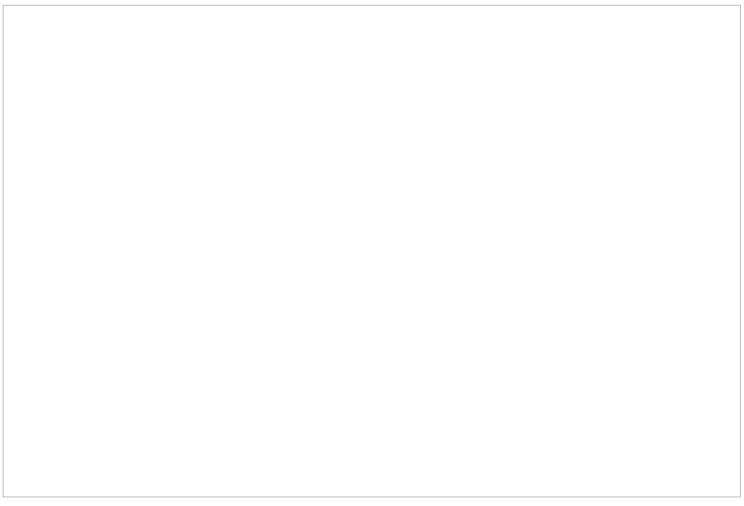
The two astronauts spent only two hours on the surface (this was the planned duration since there was uncertainty about their safety in the hostile environment of the lunar surface). During this time, they collected rocks and soil and emplaced instruments including the first of three seismometers:



A view of a small, local crater near the Apollo 11 LM is typical of mare scenery at that site:



The upper part of the Lunar Lander (the LM) blasted off safely, leaving behind the base, which is just visible in this Lunar Reconnaissance Orbiter image:

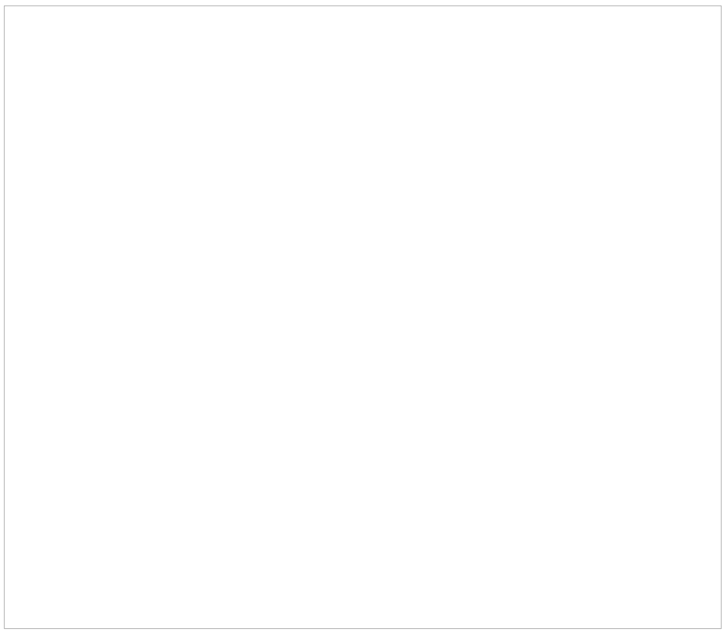


One of the most famed of lunar landing photos is shown next. It is often cited when arguments resurface about the wisdom and need to return to the Moon. The photo is self-explanatory:

Astronaut footprint at Apollo 11.

Each astronaut involved in excursions after landing was well-equipped to operate in the low gravity (1/6th that of Earth), airless surficial environment, as depicted in this figure:

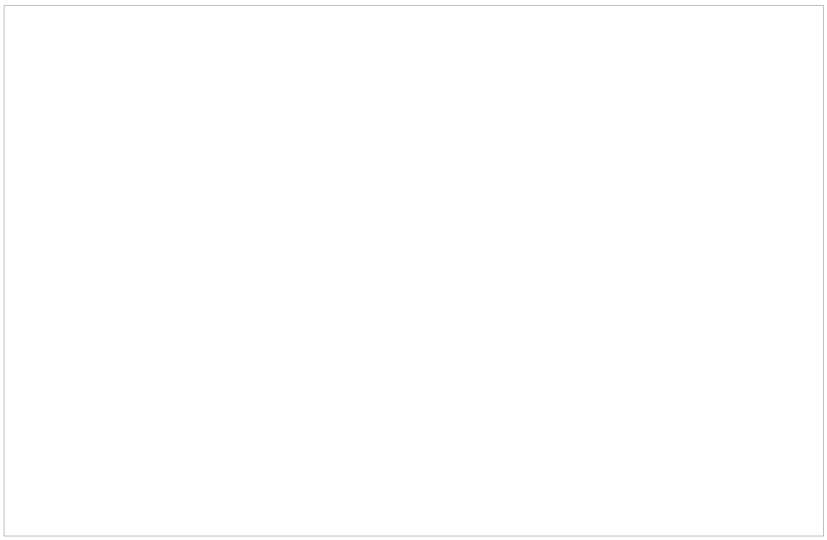
Apollo 11 and 12 sites were both in mare terrain. Apollo 11 rocks were mostly mare basalts extruding 3.7 b.y. ago. Apollo 12 also sat astride ejecta within a ray coming from the crater Copernicus. Apollo 12 represented a remarkable achievement in navigation, setting down within easy walking distance of the Surveyor III spacecraft remaining from an earlier mission,; pieces of Surveyor were returned to Earth to study the wear on its surface from micrometeorite bombardment. Here it is, with the LM in the background:



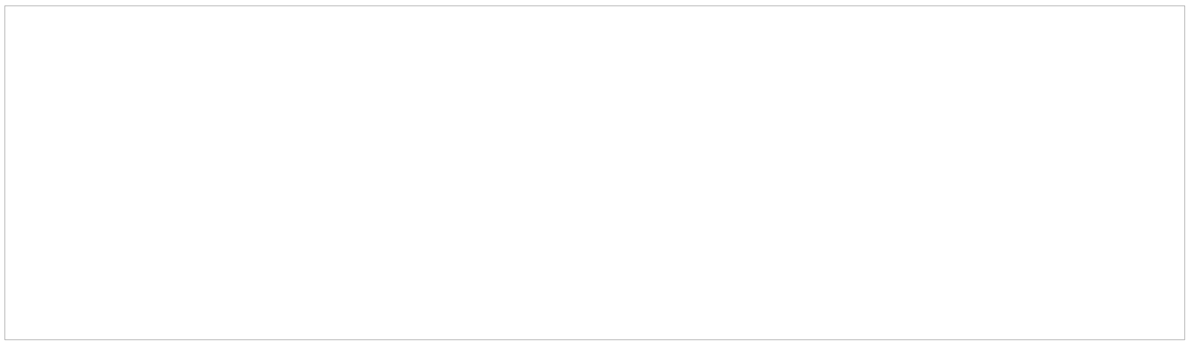
19-19: What happened to Apollo 13; did it "go" and if so why didn't it touch down on the Moon? ANSWER

In the opinion of many, the salvaging of the Apollo 13 mission ranks Number 1 on the list defining *the capacity of dedicated humans to master modern technology gone astray, requiring extraordinary self-control and innovation to avoid a disaster and save precious lives*. The story is told in the movie masterpiece "Apollo 13" with Tom Hanks as Astronaut Jim Lovell - a must-see.

Apollo 14 was in Fra Mauro ejecta emanating from the huge cratering event that produced the Mare Imbrium Basin about 3.9 b.y. ago. Basalts from Oceanus Procellarum were also present. Some of Imbrium Basin ejecta was scattered about Cone Crater, in the top view of this next illustration; a close-up of this ejecta is in the bottom view:



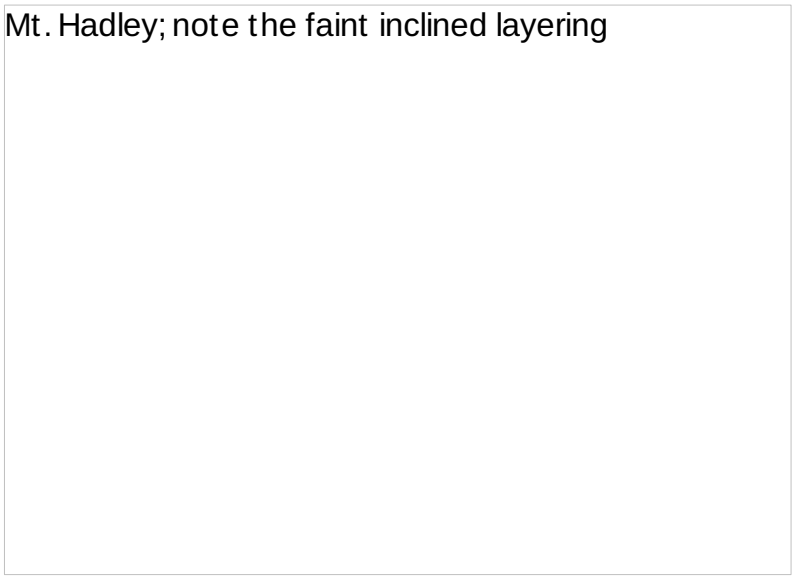
Apollo 15 landed on basalt flows near the Apennine Mountains, made of rim rocks, which are lunar crust, pushed up when the Imbrium Basin formed. These mountains, shown in the background of the next (top) view that also pictures the LM, rose as high as 4.2km (about 2.8 miles). The highest on the Moon, Mt Hadley (bottom), reaches to 4.5 km (14765 ft).

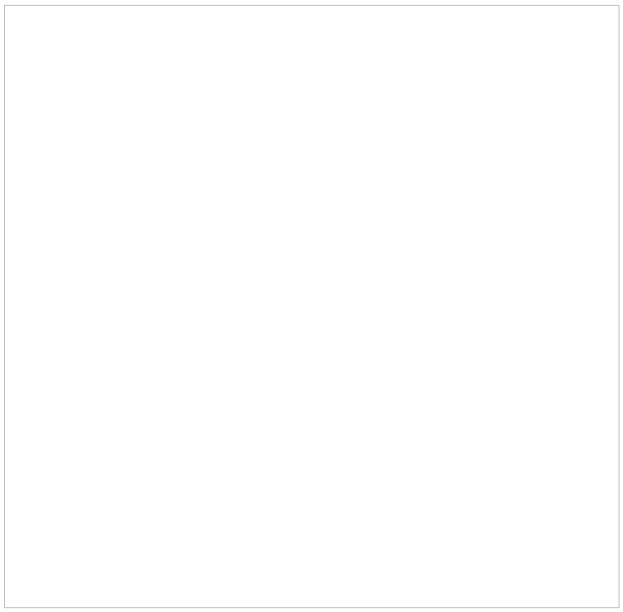


19-20: What might have caused these mountains, since the Moon does not have a plate tectonics system like on Earth. ANSWER

If the mountains are indeed uplifted lunar crust, then they expose rocks that were buried at every other Apollo site. Unfortunately, their distance from the Apollo 15 landing area precluded direct examination. But telephoto lens pictures revealed one distinctive characteristic. In the next two images are photos of Mt. Hadley; in the top one faint layering (inclined to the left) is evident and in the contrast-enhanced lower photo this uniformly-spaced set of layers seems to be crosscut by a second set of lesser dipping (lower angle) layers.

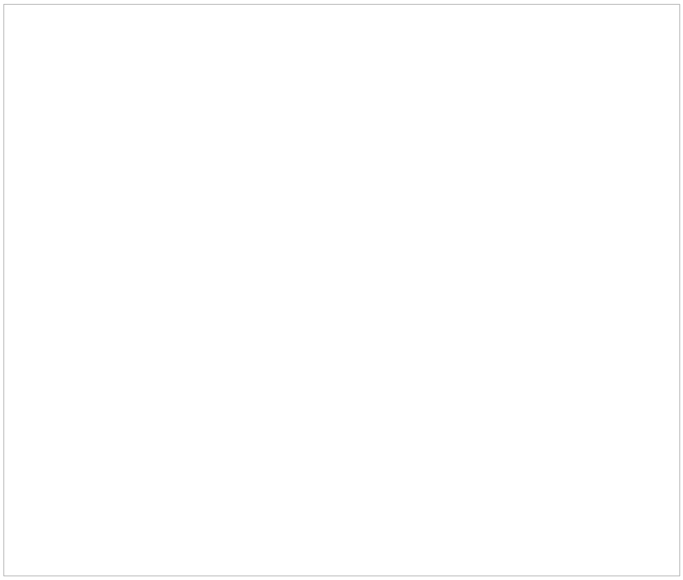
Mt. Hadley; note the faint inclined layering



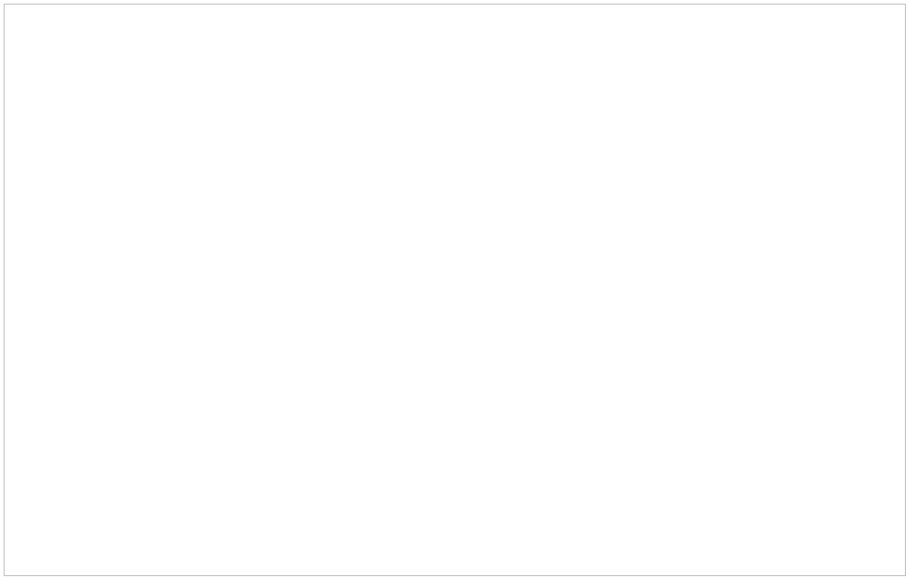


The precise nature of these layers is still the subject of speculation. But the telephoto below of a terrain extension named Silver Spur in the Hadley Mountains shows a sequence of thick (20-30 m) layers that may be a strong clue. These are dark and resemble the layering in flood basalts

such as those in the Columbia plateau. Dr. Paul Lowman (author of Section 12) interprets them as the succession of flow layers involved in filling Mare Imbrium.



A mobile, self-powered vehicle, the Lunar Rover, was first used at the Apollo 15 site to transport the astronauts on long (kilometers) excursions to vist places too far for foot excursions. The Rover, also called the "Lunar Buggy" is equipped as indicated in this schematic:



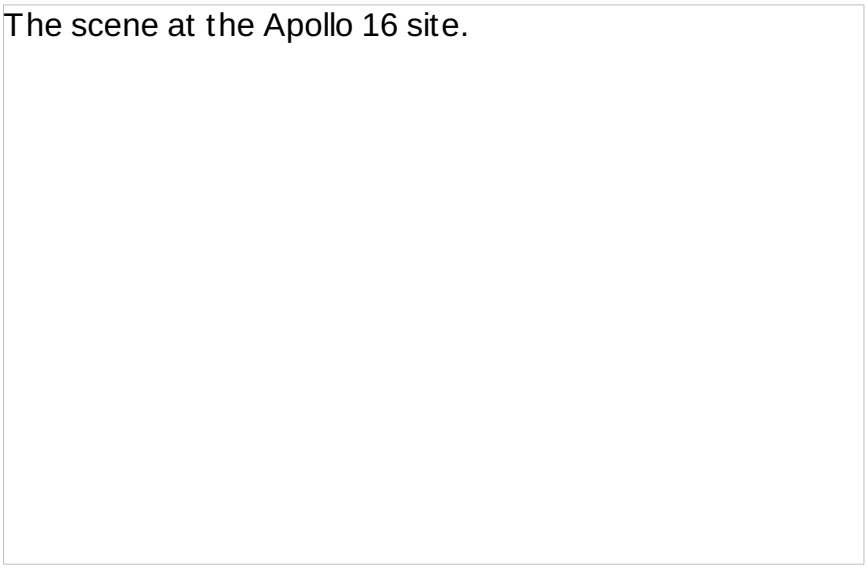
The Rover is shown in this next photo that also shows the Lander and astronaut James Irwin: when

James Irwin and the Lunar Rover.

Apollo 15 also carried the third seismometer emplaced on the Moon. After it was deployed and sent back good data, the simultaneous operation of the three working instruments permitted calculations of the seismic properties of the lunar interior. Several models have been proposed: here is one put forth by M. Toksöz (other investigators argue for a small metallic core); read its caption for more details.

Apollo 16 was the only lunar manned mission to visit true highlands terrain. The spacecraft landed in the Descartes Mountains; the site is shown in this Apollo photograph:

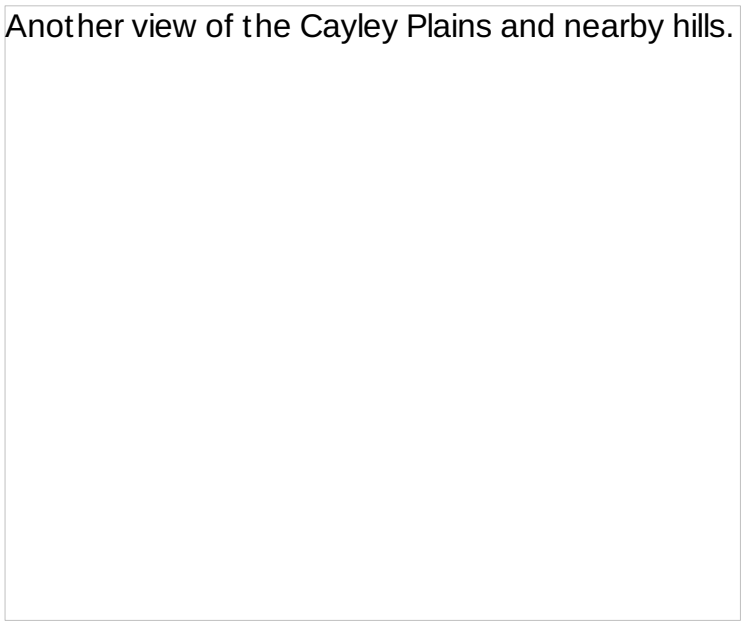
This Apollo 16 view encompasses the plains, a terrace, and low mountains in the background:



The scene at the Apollo 16 site.

The Apollo 16 astronauts main goal was to sample Highlands units, both in the hills (the Descartes Formation) and the lower units consisting of the Cayley Formation, making up part of what the USGS astrogeologists believed to make up the Cayley Plains. As interpreted by planetary geologists, these supposedly were volcanic deposits, perhaps of pyroclastic nature, but they were shown during and after the mission to be consolidated impact ejecta. A small crater and hills beyond, as shown below, are typical of this highlands landscape:

Another view of the Cayley Plains and nearby hills.



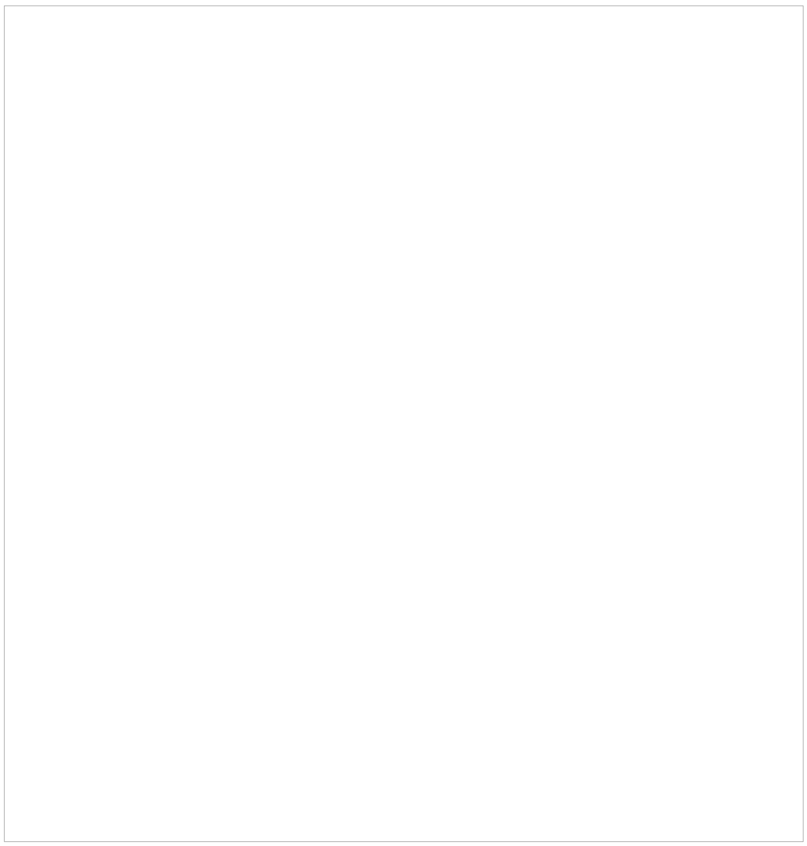
This Apollo 16 sample is typical of the breccias occurring at the site:

An Apollo 16 breccia sample.

The largest breccia sample found at the Descartes site is House Rock:

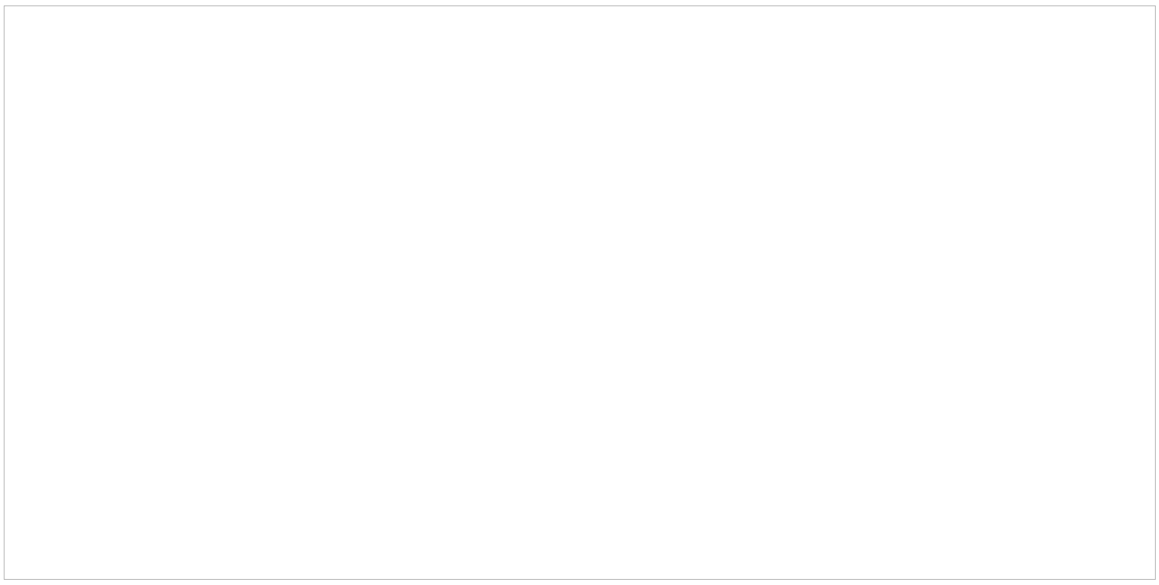
House Rock.

The Apollo 17 landing site was in the Taurus-Littrow region, where the last visit to the Moon occurred in December, 1972. Here is the approach to this site as photographed from the orbiting Command Module.



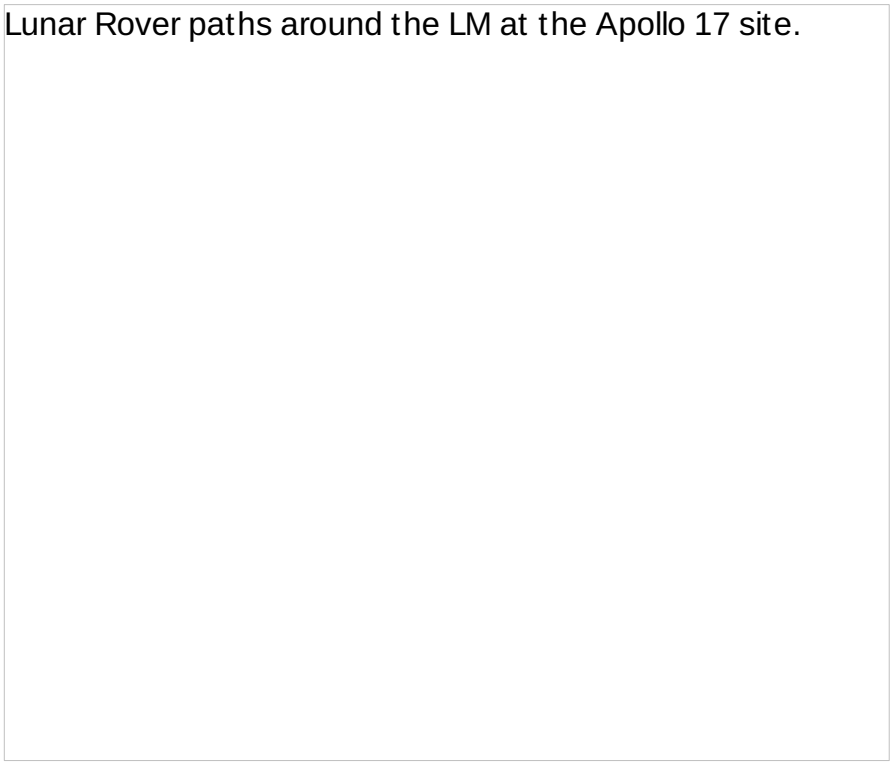
This vertical photo shows the landing site for Apollo 17 and some key features in the area including the dark, flatter volcanic units which proved to be mare basalts in nature.

Apollo 17 also sought volcanic pyroclasic units, predicted to occur around the Taurow-Littrow Mountains but, except for the notable orange layers (glass droplets splashed out of mare lavas before they hardened) within the *regolith* (debris "soil"), the rocks exposed in the *massifs* (high hills) were again made of ejecta. Some of the rocks at the 17 site were huge, as indicated by this view of "Split Rock" which consisted of a consolidated ejecta (a breccia); astronaut-geologist Harrison H. "Jack" Schmitt is standing nearby:

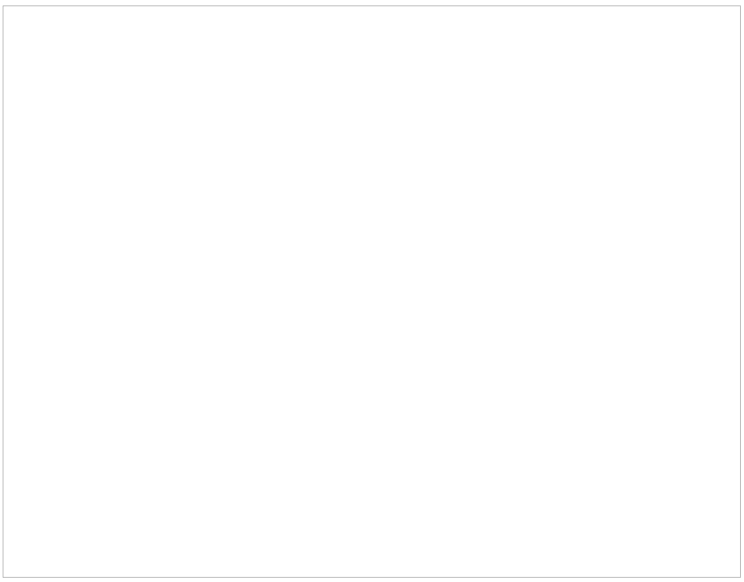


Against the backdrop of the site photo, this map shows the paths followed by astronauts Jack Schmitt and Eugene Cernan as they used the Lunar Rover to traverse to various sampling and instrument emplacement locations.

Lunar Rover paths around the LM at the Apollo 17 site.



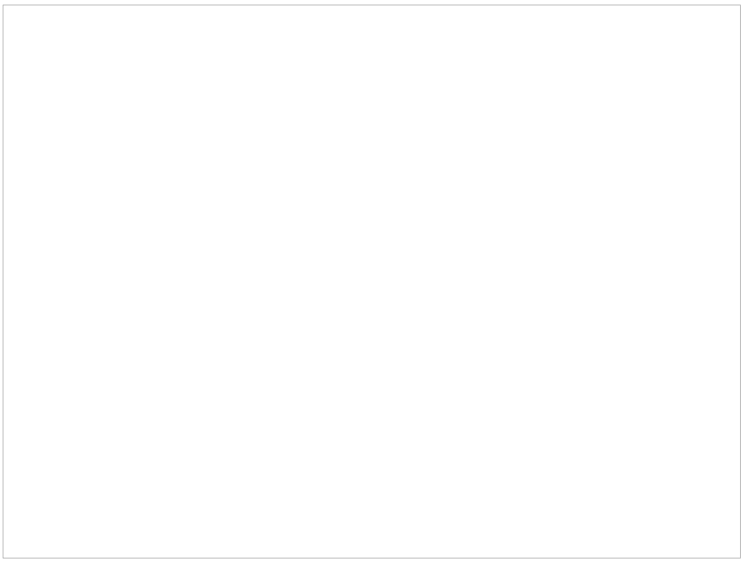
Familiarize yourself with the geography of this scene, and note especially the patterns of the massifs (hills in the highlands) that stand out as bright against the darker central lava-covered plains around the touchdown site. Now, look at this ground scene:



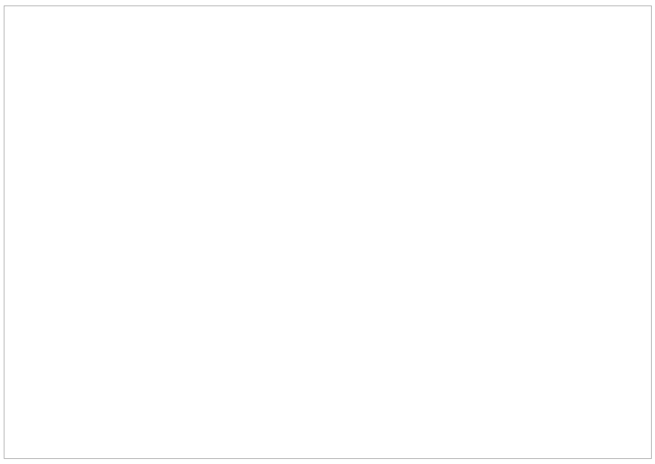
Astronaut Jack Schmitt (a geologist and the only career scientist to walk on the Moon) took the photo, partly to show the Lunar Module (LM) sitting safely on the plains against a background of a massif (mountain block) composed of breccias (see below). Astronaut Eugene Cernan is at the controls of the mobile Lunar Rover, used for excursions up to several miles from the LM.

19-21: Scroll back to the two photos showing the massifs at Apollo 17 and to the vertical photo taken from the orbiting Command Module. Can you tell which photo shows the North and which the South massifs? ANSWER

One of the prime goals for discovery at the Apollo 17 site was evidence of volcanism other than invasion by mare basalts. The U.S. Geological Survey had forecast that volcanic deposits transported above the surface should be present. There was a moment of high excitement when one of the astronauts suddenly spotted an orange layer in the regolith (surface "soil" debris), as photographed by him here:

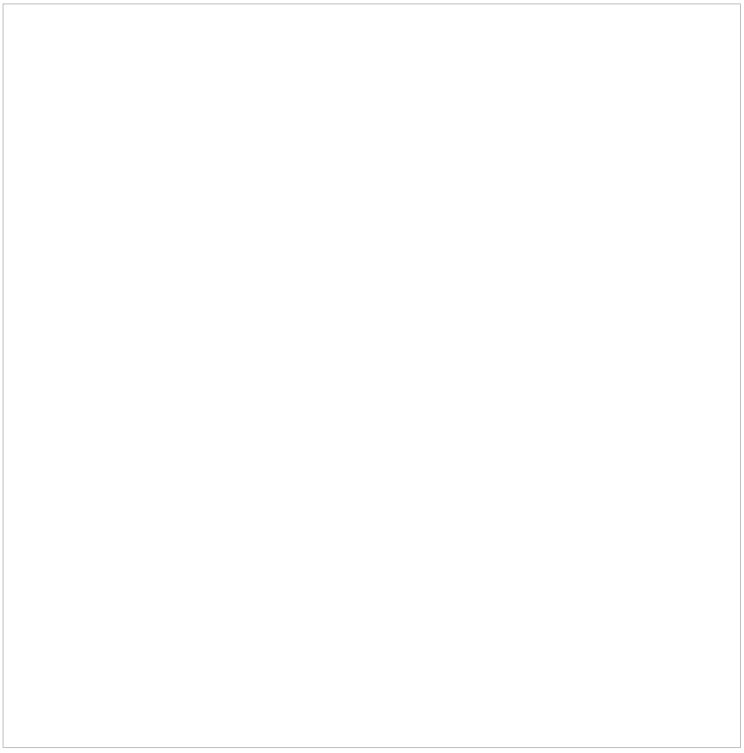


When samples of the orange layer were studied after return to Earth, the material was seen to consist of small black glass spherules and chips of orange glass. These are believed to be splash droplets tossed out of cooling lava during impacts into still fluid target lavas (probably with a solid surface crust). This is what they looked like back on Earth:



You may be interested in this Web site prepared by Jack Schmitt in which he has summarized his Apollo 17 experiences, with emphasis on the observations and results made during the mission.

Apollo 17 also shed some proof on a prediction made by the writer (NMS). Just before the Apollo 15 mission, he had attempted to calculate the thickness of the ejecta deposits that he had proposed as the principal surface units in the Highlands. The basis for this calculation was simply to determine the amount of material that would be excavated and tossed out of lunar (impact) basins and craters larger than 1 km on the frontside of the Moon. A terrestrial cratering model was assumed. The result indicated thicknesses on the Highlands to range between < 1 km to just over 3 km. Maximum thicknesses would be near the edge of large basins; the central Highlands would contain, on average, about a kilometer of ejecta (also referred to as the ejecta blanket or megabreccia unit on the surface of which normally develops, both in the Highlands and the Maria, a thin surface layer - the regolith, produced mainly through constant diminution resulting from extralunar debris that impacts as micrometeorites). After Apollo 15's seismograph began operating, indications favored a widespread low velocity layer on the Highlands which seemed to correspond to this ejecta unit. The writer published this generalized isopach (thickness) map shortly thereafter:



After the failure of Apollo 16 to find extensive volcanic deposits, the group within NASA responsible for site selection planned to meet to finalize selection of the Apollo 17 site. Rumor had it that the USGS would guarantee Taurus-Littrow as rich in volcanics. I consulted the above map and concluded that there would instead be about 3 km of lunar ejecta at and around the

site (not counting the local mare basalts). I organized an "insurgency", after I had convinced Dr. Wm. Muehlberger (U. of Texas) who was on the selection committee to champion my argument against the USGS recommendation (my choice was Copernicus, the potential Apollo 18 site). Bill argued vociferously. But it came down to 'Nick Short vs USGS' - guess who won! Guess who was wrong! Read on.

On landing, astronaut Harrison H. (Jack) Schmitt noticed that North Massif was a 1.5 kilometer high mountain made up entirely of breccia (consolidated ejecta). A small portable seismometer then determined that there was a kilometer and a half of low velocity (ejecta) below the site. These two observations totaled 3 kilometers, either a remarkable confirmation of the isopach map at that point or a very lucky coincidence.

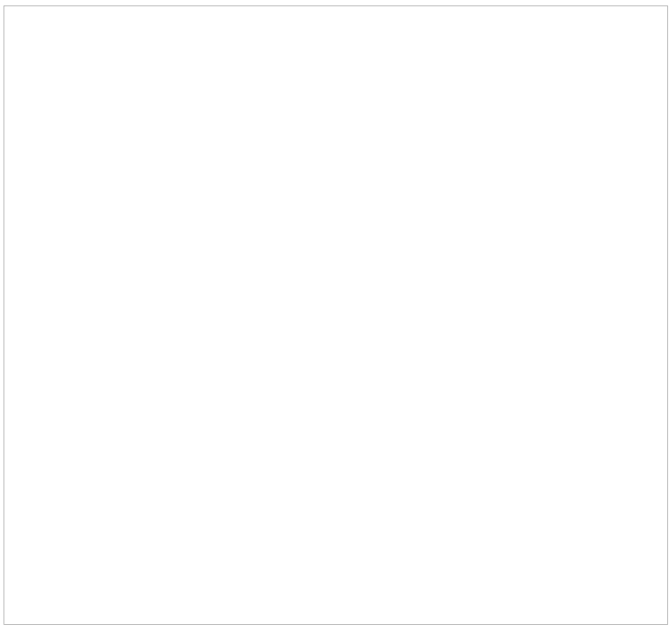
(NOTE added on March 25, 2003: The writer (NMS) spoke with Jack Schmitt when the latter visited Bloomsburg University to lecture, and I was able to gather additional details on the stratigraphy. First, he couldn't recall saying I was proved right. He said that the ~1.5 km ejecta deposits above the valley floor were from the Imbrium Basin. Below the floor the seismometer "thump" signal indicated a kilometer and a half of mare basalts. But that was covered by an earlier ejecta interval derived mainly from older Serenitatis Basin impact excavation was beneath that; its thickness exceeded a kilometer. If this is a correct update than the 3 km total is inaccurate [2.5 km being more likely] since the total would thus be 4 km of mixed breccia/flow units above, presumably, lunar highlands rock. But the 2.5 km ejecta prediction still is close.)

The dominance of breccias (most if not all being consolidated lunar ejecta) at Apollos 16 and 17, and the prevalence of this type of rock in the Fra Mauro formation at Apollo 14, supports the idea that the lunar highlands is composed of up to several kilometers of ejecta breccia, probably overlying an anorthositic crust. The unconsolidated regolith at the surface is a superficial deposit. Breccias likely were covering the crustal surface when huge impacts - the basins - removed them for redistribution, followed by invasion of the mare lavas that remain today.

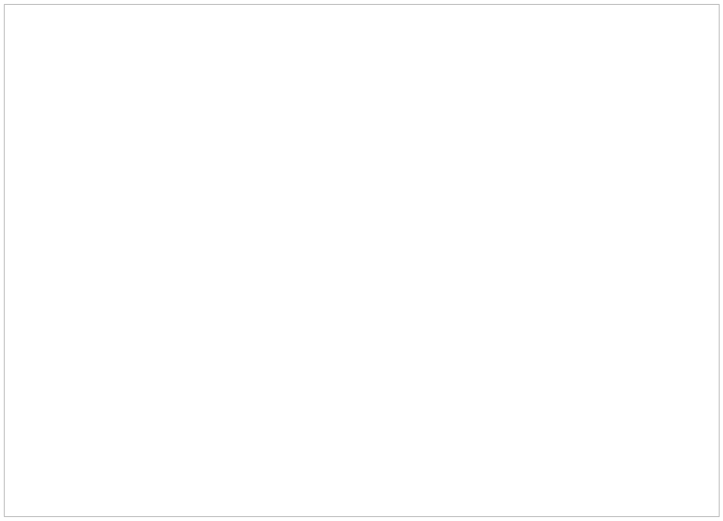
The Apollo program remains the paramount achievement of Man in Space. The greatest reward from the journeys to the Moon lies in the nearly 368 kilograms (810 pounds) of rock samples collected by the astronauts. Almost three decades later, scientists continue the most intense analysis and scrutiny of any natural substances taken from a planetary body. Starting in 1970, every year in March, hundreds of geoscientists meet at the Johnson Space Center near Houston, Texas, for the annual Lunar and Planetary Science Conference, to report on new findings and exchange hypotheses on the interpretation and implications of recent data. From examining the lunar rocks several fundamental ideas have emerged: differentiation of a primitive planetary body, the nature of its early surface, the pre-eminence in non-mare rocks of shock effects from impact cratering, the age and history of the Moon, and its origin. The prevailing opinion is that it derived from accretion of terrestrial debris hurled into space after being ejected from a huge impact on the early Earth that resulted in crustal/mantle materials attaining orbital velocities. A variant of this model, held by some, is that much of the material came from the incoming body, probably an asteroid whose size approached that of Mars).

The appearance of typical lunar rock specimens is shown in the next four photos:

First is a specimen of lunar basalt from the Apollo 12 site; it is fine-grained and pitted by micrometeorite bombardment. This is the prevalent rock type in the maria.

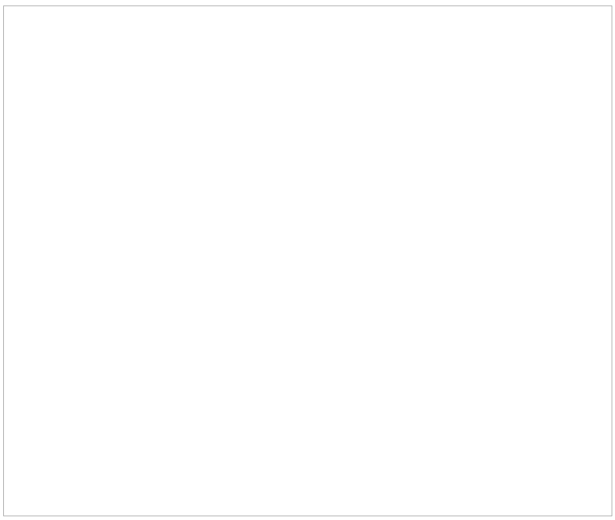


Next is an anorthositic gabbro, a whitish rock dominated by feldspar. It is probably typical of the bulk of the original lunar crust, representing crystals of plagioclase that floated upwards in the cooling magma that developed during the Moon's first general melting. Note the black glass that coats part of the sample, presumably plastered on the rock as it was transported from an impact crater that made the glass.



Third is an unusual rock found first at the Apollo 12 site (sample 12013). It contains two color phases: light (shown here) and dark:

Apollo sample 12013; the KREEP rock.



The dark phase is a breccia composed of light-toned fragments in a basaltic matric. The light phase is described as a felsite (fine-crystalline equivalent to a granite) and has a higher silica (SiO_2) content than most lunar samples. It contains potassium feldspar and needles of quartz. Because it is high in potassium (chemical symbol K), the rare earth elements (REE), and phosphorus (P), it is representative of the lunar rock class called KREEP. (Note its position in the diagram at the bottom of this page.) Other examples were found at later Apollo sites, particularly Apollo 14. Current thinking holds it to be the top differentiate of a magma that gave rise to a more general anorthositic crust beneath. A variant is that this very old rock may be a sample of the original lunar magma which suggests that crust formed from it was more granitic.

Finally, from the Apollo 17 site, the large central rock is a breccia, that is, a rock composed of fragments of other rocks that accumulate from various (usually distant) sources which are then welded together by heat and pressure.

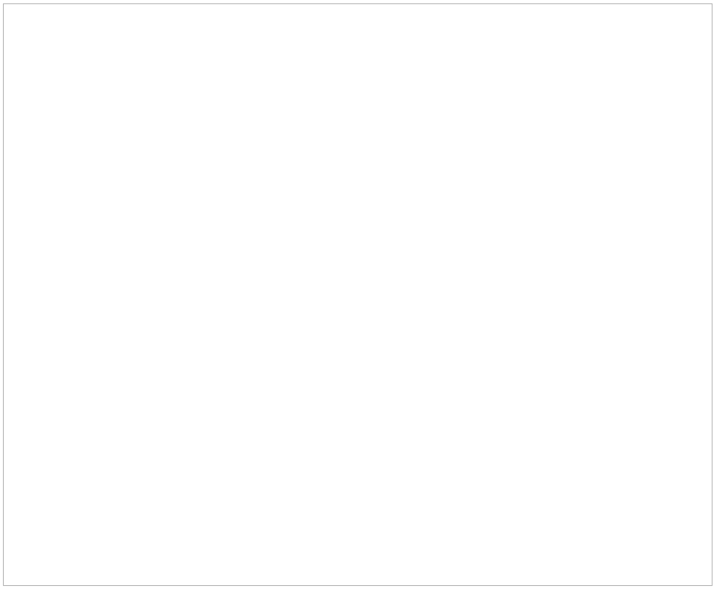
A lunar breccia from the Apollo 17 site.

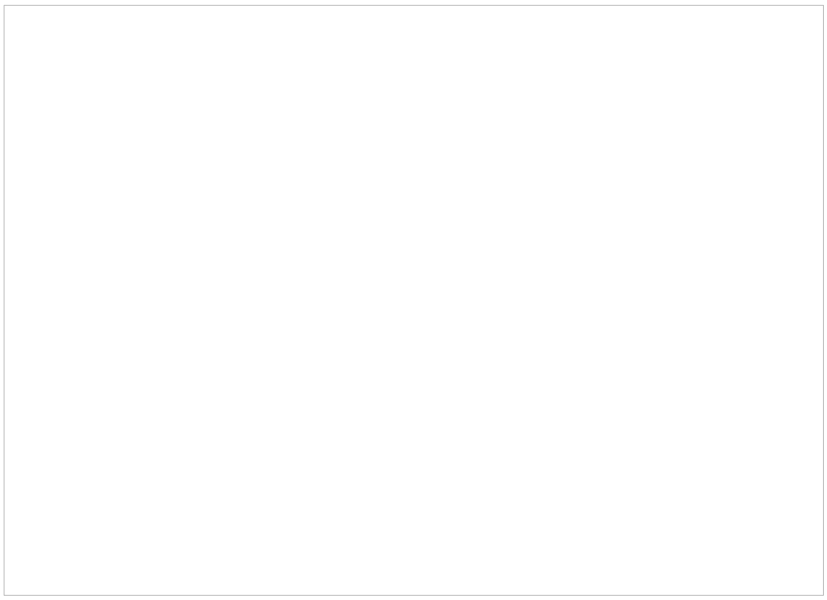
These fragments represent impact debris from more than one area of the Moon that makes up the principal deposits of the outer layers of the Moon, i.e., part of the 1-3 km (about 0.62-1.86 miles) thick, lunar-ejecta blanket. The dark material in this specimen is basalt, the solidified end product of iron-rich lava that fills lunar marias. The lighter rock fragments probably originated from the lunar highlands; Surveyor chemistry and Apollo samples show these rocks to have a high percentage of grayish feldspars (Ca-Na aluminosilicates) that cause the highlands to appear lighter in tone (higher albedo or reflectivity). The smaller rock samples surrounding the large specimen are vesicular basalt pieces (dark) and individual highlands rocks (light), collected nearby on the same mission.

At every Apollo site, the surface was covered with fine fragmental material, the lunar soil (technically called *regolith*) which lay atop the lunar ejecta blanket in deposits from meters to tens of meters thick. This material is the accumulation of debris brought in from near to distant

sources after cratering tossed fragments of a wide size range to varying distances beyond the rims. The deposits were then further comminuted by constant micrometeorite bombardment.

An important task for the astronauts was to drive drills into the regolith that allowed core to be removed intact. Here is Astronaut David Scott in process of coring at the Apollo 15 site and below that are recovered core samples from the Apollo 12 site.



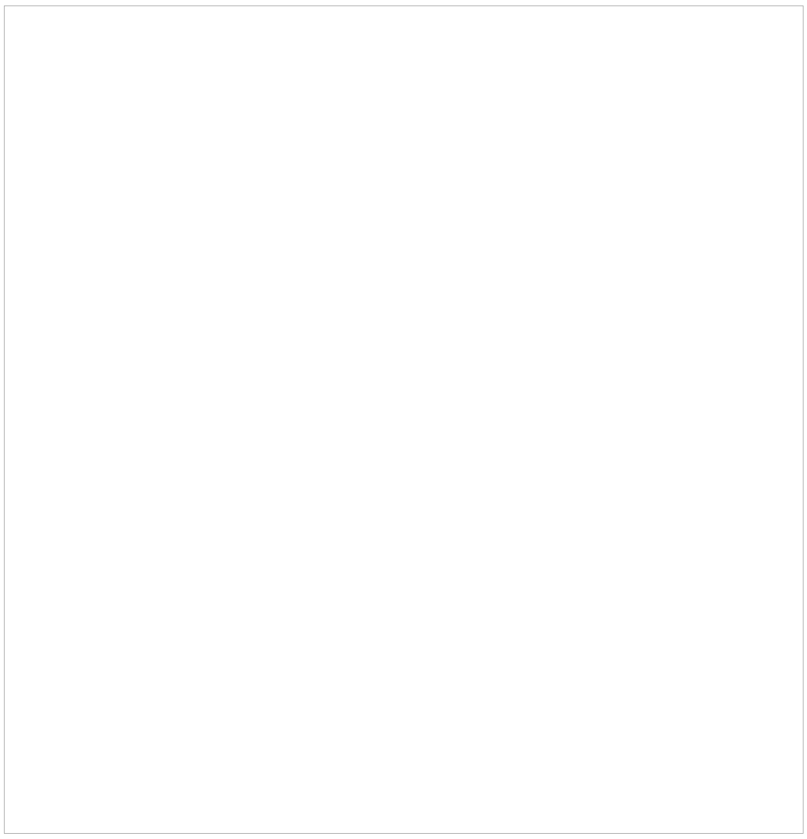


The core samples were analyzed as intensely as the rocks returned from the Moon, but they told a different story. Their components roughly indicated the relative contributions from nearby versus faraway sources. Here is a plot of the relative proportions of different lithologies of coarser fragments in one core from Apollo 11 and several core samples from Apollo 12.

Comparison of lithologies in several Apollo 11 and 12 cores.

Basaltic rocks and breccias were predominant, indicating local sources provided most of the material. Anorthosites were carried to the sites from distant highlands. When mapped in detail, the cores clearly show multiple layers that differ in composition and size distribution, as expected whenever larger impacts tossed out significant ejecta to their surroundings.

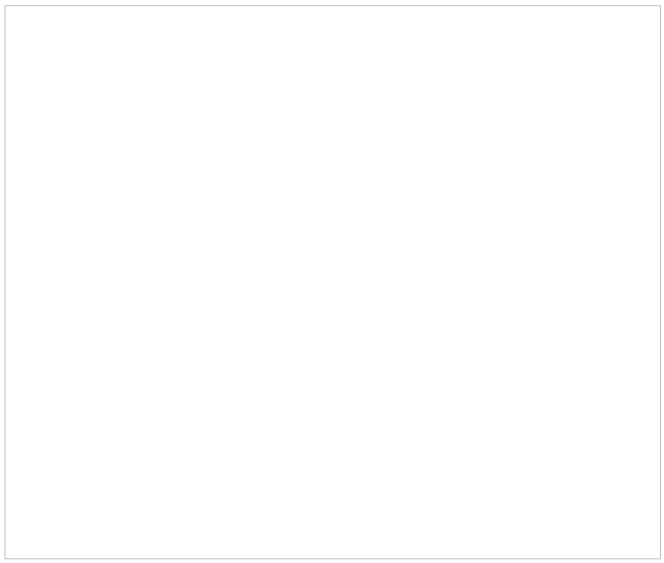
Analysis of lunar samples brought to light several distinct features. Some breccias when individual fragments were examined disclosed definitive evidence of shock effects, indicating that these rocks were debris involved in major impacts. Shock effects were found also in lunar soil samples. The writer (NMS) in 1969 was one of the original 142 Principal Investigators selected to study the first lunar samples; my task was to search for shock effects, to be expected because of the prevalence of impacts into the outer surface. However, my earlier experience with shock features in basalt subjected to nuclear explosions led me to predict that these effects would be harder to find than in more silicic rocks such as granites. This proved true. Still, effects attributable to shock were observed rather frequently. This group of six photomicrographs (reread the page on shock metamorphism [p. 18-4]) depicts some examples (check the caption for descriptions):



I also prepared a second montage showing shock effects in color. Again, the description of each picture is in the caption.

The upper right photo warrants further discussion. This is a small rock fragment found in a breccia sample that almost certainly was a part of the now-lithified outer ejecta blanket described above on this page. The fragment consists almost entirely of pieces of plagioclase feldspar. The writer first set eyes on this fragment, in a "Eureka!!" moment in late October of 1969, while working on my P.I. tasks. As I peered through the microscope at this this fragment, I was suddenly aware of how much its texture was like that of the shock-lithified sandstones ("instant rock") described on page 18-3. Shown below is a variation: instant rock made up entirely of feldspar fragments welded together by the implosion tube method described on page 18-3.

Compare this texture to that of the 'Eureka' lunar fragment that appears here:



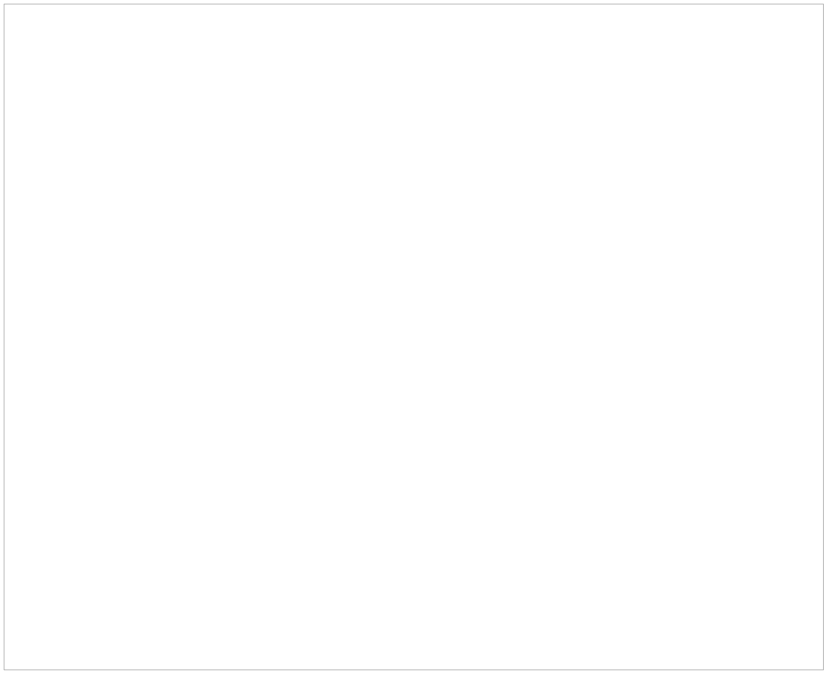
I immediately formed a prescient postulate. What I was looking at is a piece of the highlands regolith that had been converted into instant rock by an impact that hurled this fragment a long distance from the highlands site only to land in the forming regolith at the Apollo 11 site. If this were correct, then I was in fact confirming the argument made by A. Turkevitch and his colleagues (see answer to question 19-13) that the highlands was made up of mostly anorthosites (feldspar dominant rock). This conclusion I reported during the first Lunar Science Conference (1970), as did John Wood of the Smithsonian Astrophysical Observatory.

Mineralogically, moon rocks are similar to several types found on Earth, except that the individual minerals are almost entirely unaltered by hydrothermal solutions, i.e., they are very "fresh" looking under the petrographic microscope. Basalts, anorthosites, and breccias (fragmental mixtures of several different rock types) are prevalent. The most common minerals are calcium plagioglase, several pyroxenes, olivine, and ilmenite (an iron-titanium oxide). Three new minerals, unknown on Earth, have been found: tranquillityite, armacolite, and pyroxferroite.

Chemically, the lunar rocks were mostly in a class by themselves, being different from their terrestrial rock type counterparts. They are highly variable in iron (as FeO), with mare basalts being richer in this metal than are terrestrial basalts; they are low in volatiles (including potassium and sodium), and are totally anhydrous (meaning that water was not present when they formed; water found on the Moon is discussed below). Compared with Earth rocks, they were exceedingly fresh, showing almost no signs of alteration. As two examples of their chemical specificity, examine these diagrams which plot 1) Fe vs Mg, and 2) the ratio of potassium (K) to uranium (U) versus changing potassium content; note that both meteorites (chondrites and carbonaceous chondrites) and terrestrial igneous rocks plot in different areas of the diagram

than the lunar rocks. However, the meteorite class Eucrites plots partly within the lunar samples field, suggesting that these are actually ejecta from the Moon that reached Earth.

Mg vs Fe for lunar samples.



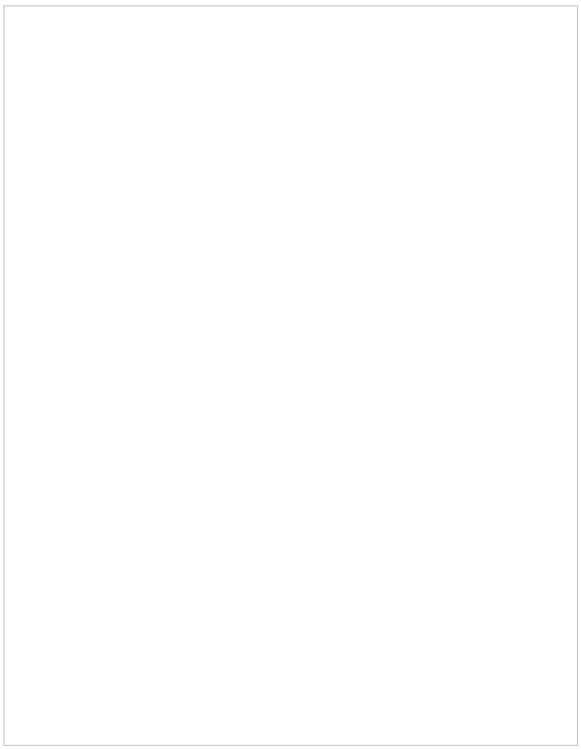
19-22: What conclusions can you draw from this K/U vs K (ppm) plot? ANSWER

Geologists would, of course, favor getting additional samples from other locations on the Moon, so as to better define the variability of its exterior composition. Possible future expeditions to the lunar surface (unmanned; manned) may fulfill that desire. Meanwhile, a few meteorites among the thousands being found on the ice during collecting trips to the Antarctic have been confirmed as coming from the Moon. Two examples are shown below:

Another lunar meteorite.

Here is a more recent discovery of a lunar meteorite in the Antarctic:

None of the lunar rocks collected during Apollo show ages as old as some meteorites (4.6 billion years [b.y.]). Radiometric dating (U/Pb; Rb/Sr; K/A decay methods) gave both model ages (times when initial materials appear to have formed) and formation ages (time when these materials melted and solidified to their present state). These two diagrams show Rubidium (Rb)-Strontium (Sr) ages in (A) and Potassium (K)-Argon (A) ages in B:

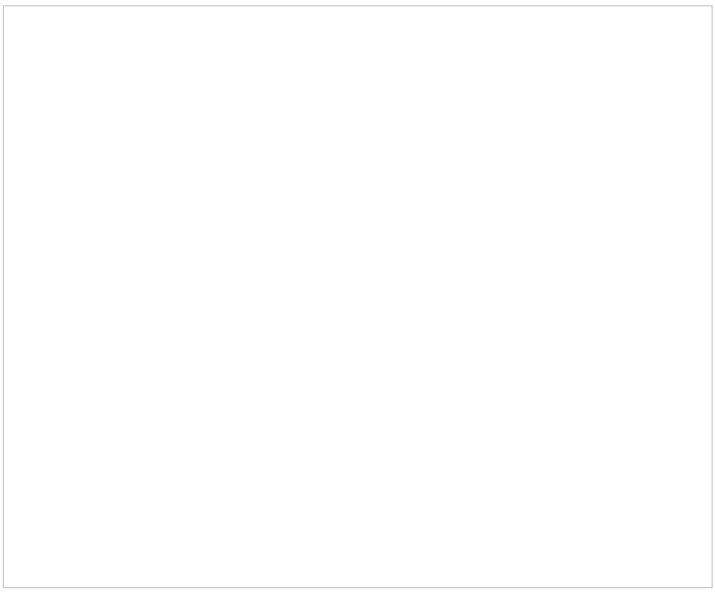


The youngest basaltic rocks came from the Apollo 12 site (exception: Sample 12013 [Apollo 12 site] is more than 4 b.y. old - it is an exotic emplaced as ejecta from the highlands); the oldest from Apollo 16 and 17 . The maria formation ages spread from around 3.2 to 3.85 b.y (none older). The major ringed basins such as Orientale and Imbrium, most now filled with basalt, seem to have formed about 3.9 b.y. ago whereas Serenitatis may be as old as 4.2 b.y. Thus, these basins took several hundred million years or more to fill.

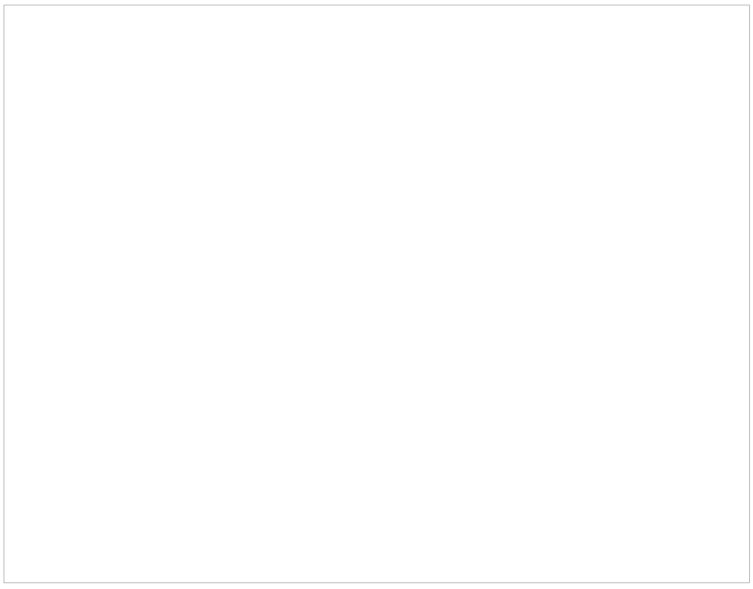
Many highlands rocks, mostly anorthosites, show ages in the 4.0 to 4.2 b.y. range but the primordial crust formed sometime between 4.4 and 4.6 b.y. ago. No model age (older) was as old as the 4.6 b.y. estimated for the Earth, but the Moon is still considered nearly that old. It could be a 100 million years or so younger if the Earth impact origin remains as the mode of origin for the Moon (see page 19-6b).

A few large lunar craters have been formed during an era beginning less than 1 billion years ago. Copernicus may be about 800 m.y. old and Tycho perhaps less than 300 m.y. in age. This is inferred by extrapolation from crater counts on surfaces elsewhere whose estimated ages have been calibrated by radiometrically dated rocks.

The Moon and the Earth differ distinctly in the ages of rocks found on their rocky surfaces. On Earth very few rocks are older than 3.2 billion years; on the Moon very few are younger than that. This plot summarizes the age distribution:



As the Apollo program progressed, many planetary geologists argued for making Copernicus a key part in the Apollo mission series. None through Apollo 17 actually went to a very large crater. A visit to one of the "biggies" would add valuable information about rocks and morphological features associated with these large structures. Before Apollo 17, there was debate about where that last mission would go. Impacters supported Copernicus but problems as to where to land safely (the interior near the central peaks was proposed) caused it to be rejected in favor of Taurus-Littrow. Although Apollo 18 had been cancelled before 17 succeeded, the astronauts on that mission took this photo of Copernicus in hopes that the abbreviation of the program could be circumvented.



Copernicus - or a comparable large crater - is still there, awaiting a visit in a hoped-for renewed manned lunar exploration program, that is a "must do" (this writer's opinion) in the foreseeable future, now that China and other countries are planning their own lunar programs. Meanwhile, with the success of the unmanned Rovers on Mars (page 19-13a), which proved the technical feasibility of traversing uneven surfaces and doing mineralogical/chemical analyses, sending a fleet of at least several to key locations of special interest on the Moon seems almost mandatory.

PREVIOUS RST NEXT

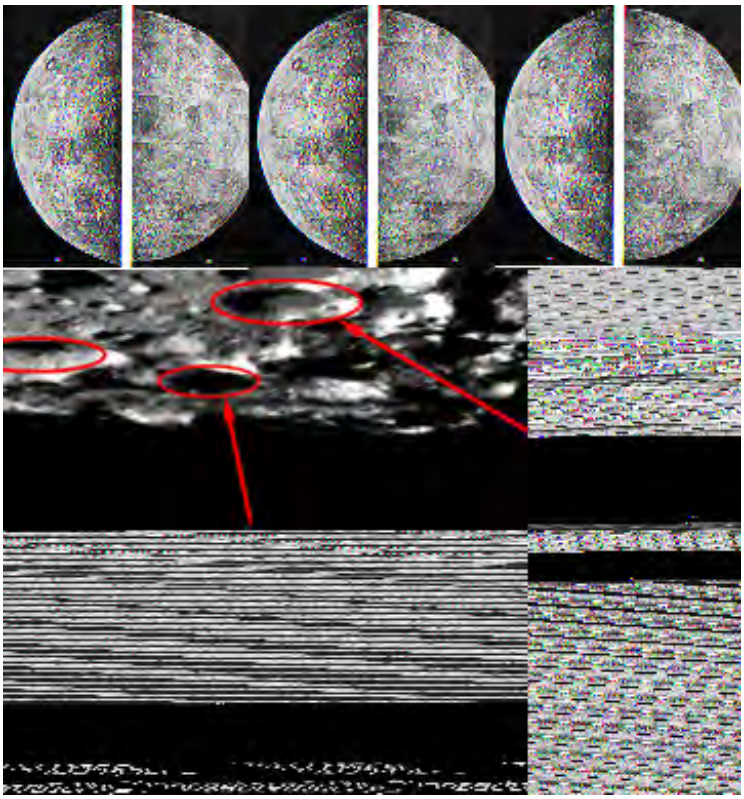
Primary Author: Nicholas M. Short, Sr.

On the next three pages we will journey to the two planets between Earth and the Sun. The closest-in is Mercury, which looks like a larger version of the Moon except for the near absence of young basaltic (mare-like) lavas. Next is Venus which normally is completely enshrouded with a very hot gas (carbon dioxide) and clouds of condensed sulphuric acid. But landers have reached its surface (surviving only for hours). More importantly, radar, first from Earth, then from orbiting satellites, has given us remarkable views of surface features - some similar to Earth counterparts but many almost unique.

The Inner Planets: Mercury and Venus

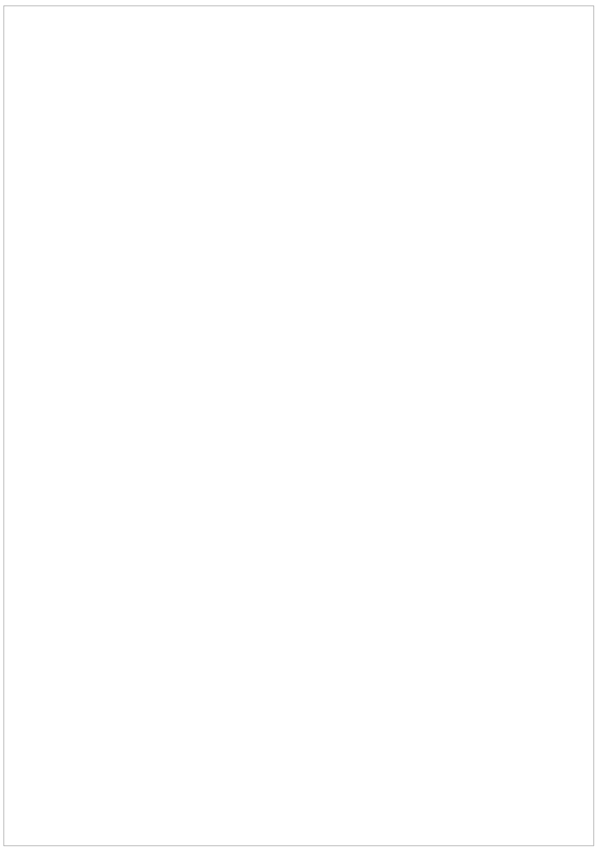
Mercury orbits at an average distance of 0.387 A.U. relative to Earth but its high eccentricity (0.205) results in a perihelion of 0.307 A.U. and aphelion of 0.467 A.U. Its diameter is 4880 km (3030 miles) compared with 3478 km (2160 miles) for the Moon and 12756 km (7921 miles) for Earth; volumewise Mercury is 1/50th that of Earth and three times larger than the Moon. Mercury's density (5.44 gm/cc) is close to Earth's (5.53 gm/cc), suggesting it too has differentiated into a crust, mantle and large iron core but presence of a weak magnetic field implies that, while the core has maybe largely solidified, it has a molten component. Without an atmosphere, Mercury is heated directly by the Sun, to which it is the closest planet, to a mean surface temperature of 180° C (on the perisolar side, this temperature rises to as high as 425° C; on the nightside, it can drop to - 170° C). This innermost planet has periods of revolution and rotation of 88 and 59 days respectively, in a ratio of 3 to 2; this relation is said to be spin-orbit coupled.

The surface of Mercury bears no resemblance to Earth but to the casual eye is almost a twin of the Moon. This is immediately evident in this view which shows the eastern and western hemispheres of the mercurian surface imaged by TV vidicon cameras onboard the Mariner 10 spacecraft launched on November 3, 1973 first to Venus (see below) and then past Mercury on March 29, 1974 and around the Sun twice to return for additional Mercury flybys in September 1974. and March 1975



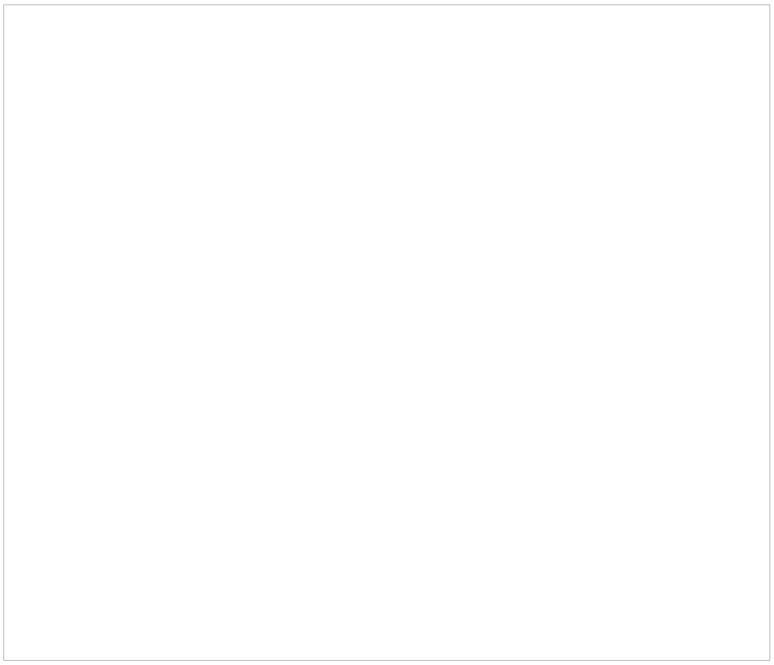
The three encounters permitted 45% of the planet to be imaged. This next view shows part of Mercury as rendered in quasi-color.

Craters, once again, are the prevalent geomorphic feature, with some impact basins exceeding 200 km (124 miles) (largest is the multi-ringed Caloris Basin at 1300 km [807 miles] diameter, part of which is seen in this next image):



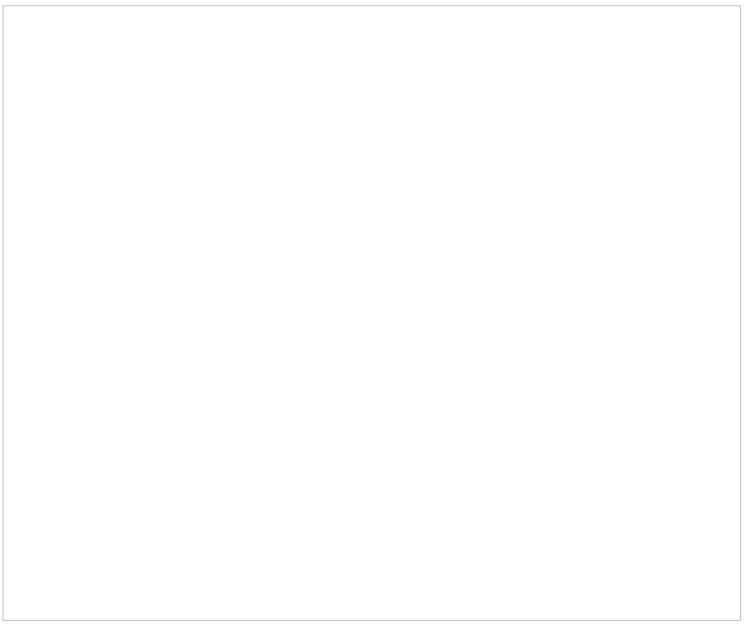
19-24: How many rings can you make out for the Caloris Basin? ANSWER

The dominance of craters is evident in this view of part of the southern hemisphere:

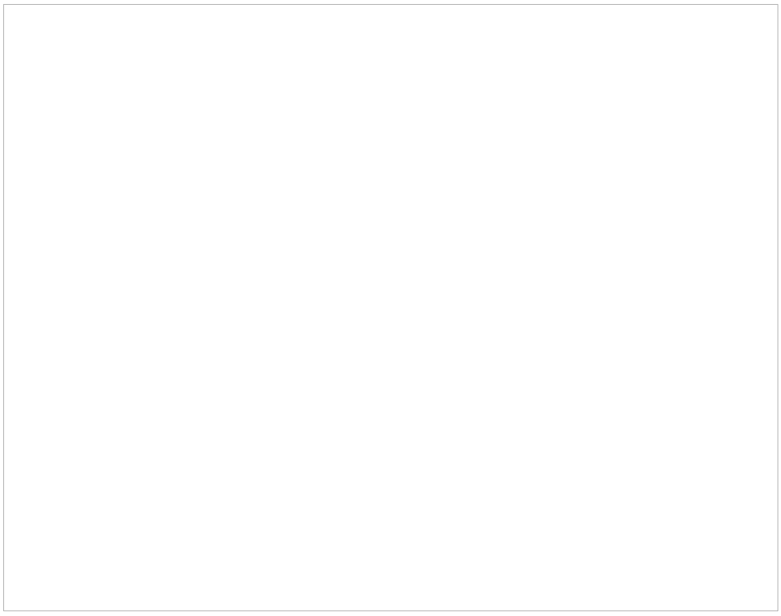


Unlike the Moon, distinctly basalt-filled maria are sparse, although some small darker patches have been seen. This implies that the general second melting that occurred on the Moon about 3.8 - 3.9 billion years ago did not happen on Mercury whose surface probably is even older and preserves the same period of impact devastation associated with the lunar highlands.

However, the bulk of the Mercurian surface is described as being a relatively flat volcanic plains made up of iron-rich lavas. Some plains units may be original crust. The most common terrain type is Intercratered Plains as seen here:

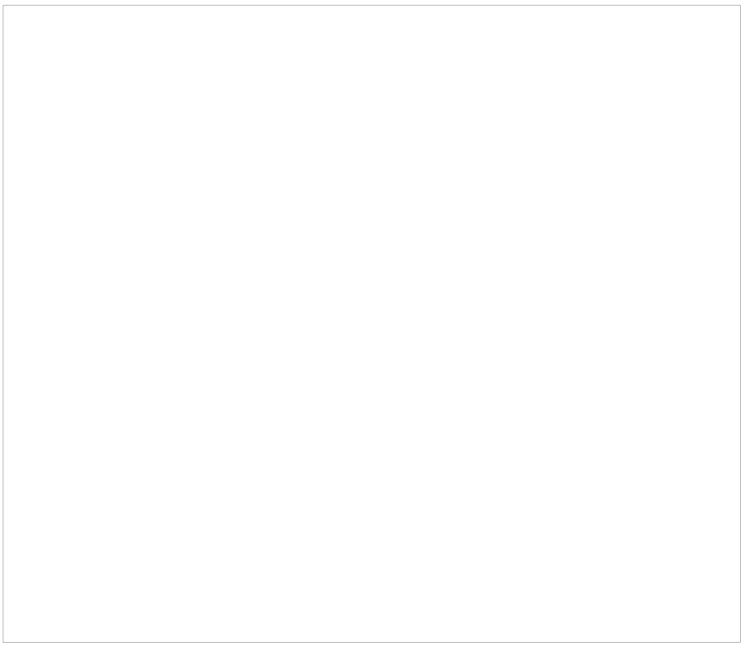


A second low relief unit has been termed Smooth Plains (although likely volcanic in origin, its mode of emplacement is debatable). It resembles visually the lunar maria and also shows a notably reduced population of craters, suggesting this Plains unit is younger than most of the mercurian surface and partly fills many older craters.

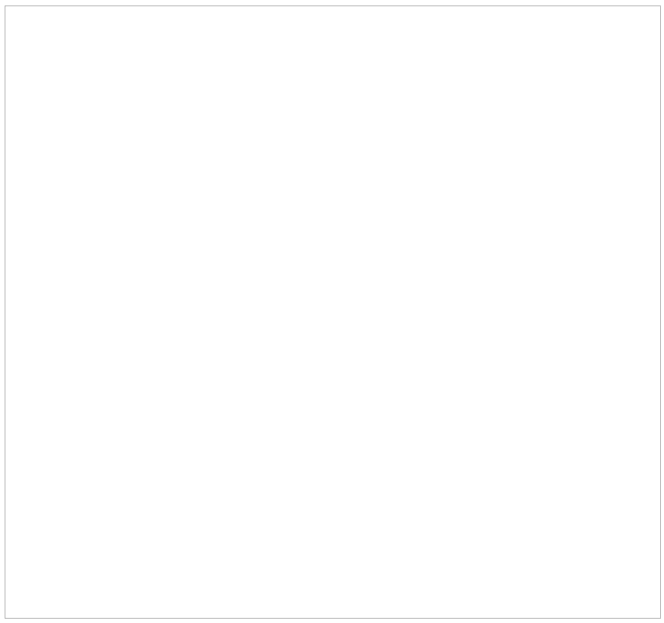


This reduction in crater numbers is further evidenced in this image:

Some regions on Mercury are rugged, with large hills and lineations, as seen here. They may be a mix of ejecta units and volcanic structures.



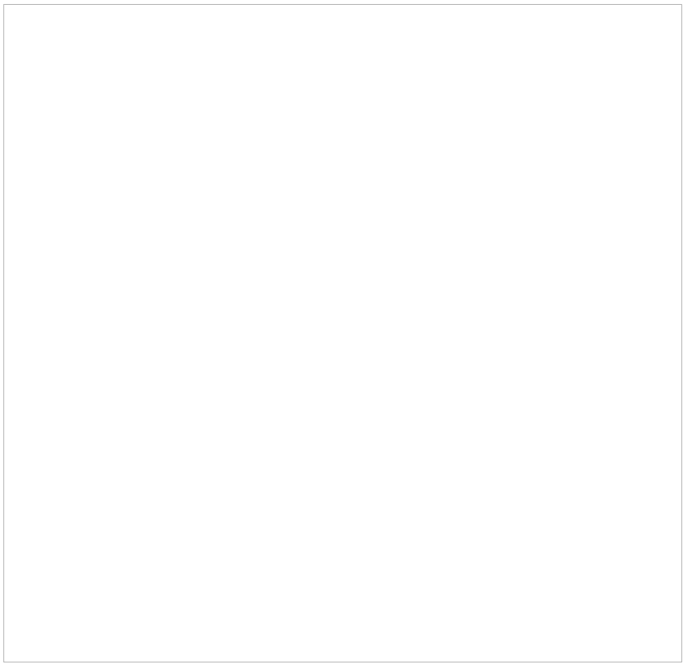
Mercury, in some differences from the Moon, displays occasional structural features indicative of compression. One example is this fault, interpreted to be a thrust in nature, that forms a scarp 3 km high:



The history of Mercury is twofold: volcanic events that produced plains units and impact cratering that have greatly modified the terrains dominated by flows.

19-25: Discuss major similarities and differences between the Moon and Mercury. ANSWER

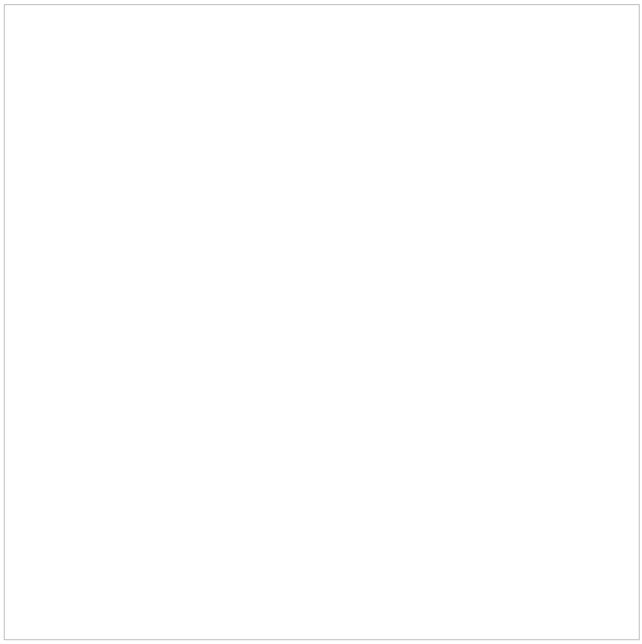
NASA has renewed the Mercury Exploration program with the August 3, 2004 launch of MESSENGER, which will make several passes near Venus and the Sun (for gravity boosts of orbital velocity) and then reach orbit around Mercury in 2011. This sophisticated satellite looks like this (artist's rendition)



Among its proposed instruments are 1) MDIS = Mercury Dual Imaging System; 2) GRNS = Gamma Ray and Neutron Spectrometer; 3)XRS = X-ray Spectrometer; 4) MAG = Magnetometer; 5) MLA = Mercury Laser Altimeter; 6)MASCS = Mercury Atmospheric and Surface Composition Spectrometer; 7) EPPS = Energetic Particle and Plasma Spectrometer. Included in its scientific goals are learning more about Mercury's crust, core, and magnetic field. You can follow this mission by tracking its home page

Messenger made its first fly-by pass on January 14, 2008. The MDIS produced images of the side of Mercury that was not seen during the Mariner 10 mission. Here are three of the first series of panoramas and close-ups:

. .



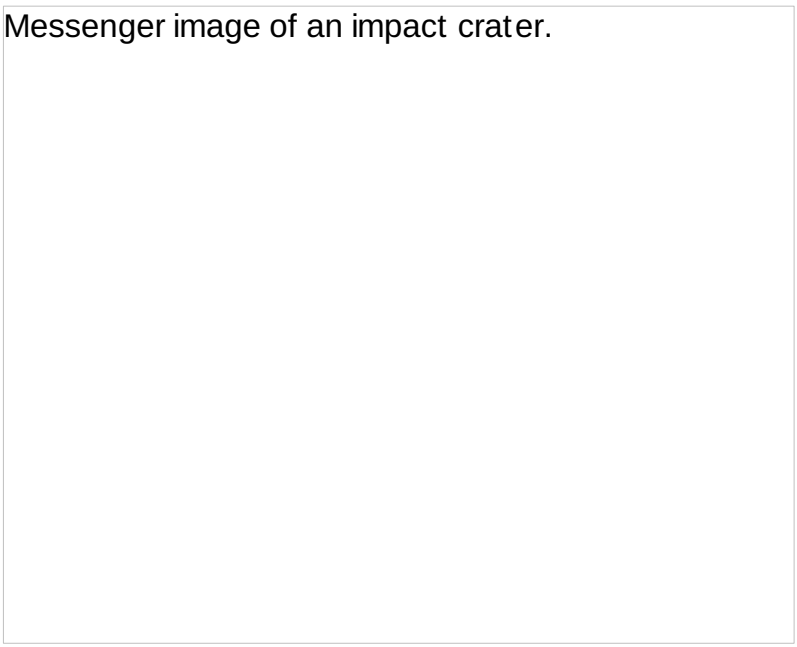
Ridges on the mercurian surface.

The south polar region of Mercury.

The impression given by these images is that the previously unseen mercurian landscape is much like that envisioned decades earlier after the Mariner flyby.

Messenger is sending back improved images of mercurian craters. This one is comparable to the Gosses Bluff crater in Australia inasmuch as it has a prominent inner ring analogous to the central peak that forms from rebound as the impact shock waves force up material from below. Note the smooth areas in and out of this ring, which could be volcanic material extruded after impact.

Messenger image of an impact crater.



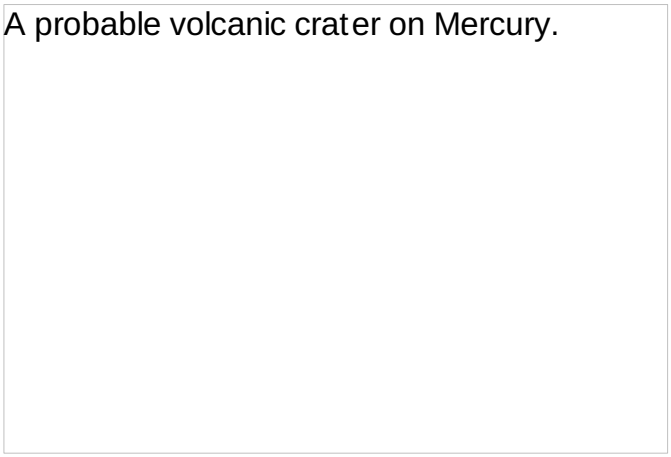
Two other similar mutliringed craters appear in these Messenger images:

Other ringed craters on Mercury.

Other ringed craters on Mercury.

One crater imaged during this first flyby is almost certainly volcanic in origin. Note the radial grooves: these are depressions, a feature associated with tensional stresses resulting in an updoming by magma. The crater itself (below) is typical of the morphology of a volcanic caldera. However, the association with the grooves, which appear to be grabens, does not have a comparable counterpart on Earth, so this feature may be unique. A plausible explanation: shallow magma could have domed the crust, causing tension that led to the grabens, followed by volcanic eruption and caldera collapse.

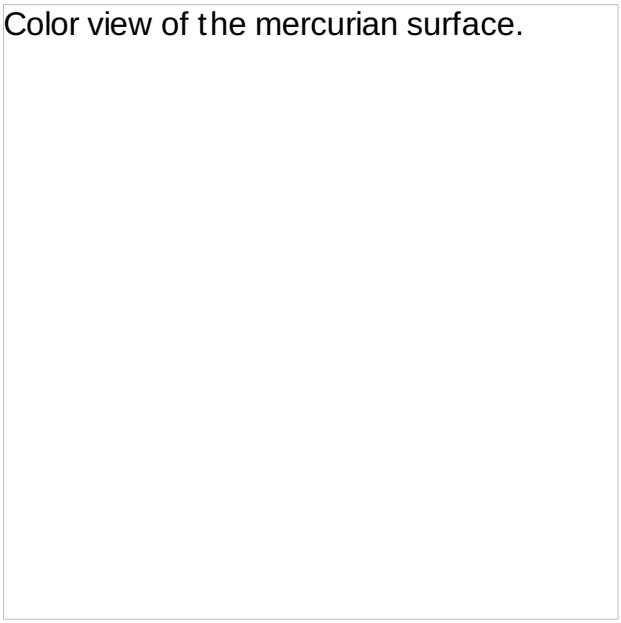
A probable volcanic crater on Mercury.



The image below shows surface features that may have a volcanic origin. Two craters near the bottom each have narrow dark rims. Melt glass from an impact is an alternate hypothesis. Possibly, the craters have exposed a pre-formation dark layer in the mercurian near-surface.

Messenger afforded a better look at the Caloris Basin. It actually is about 100 km wider than earlier estimates, confirming it to still be the largest megacrater in the Solar System. As such, it dug deep into the crust, probably into the mantle, and has brought this material to the surface. Chemical analysis of the ejecta will establish the composition of the upper interior of Mercury.

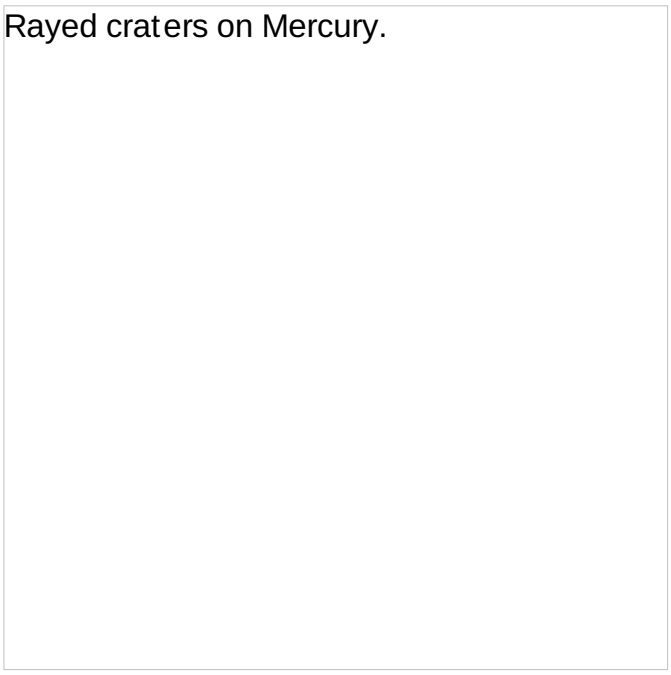
The Messenger images can be in quasi-natural color. Here is one example:



Messenger confirmed that the planet's core is large, extending to about 60% of the radius. This makes the mercurian interior different from Earth and Venus. The outsized core needs to be explained. One hypothesis postulates that Mercury was once larger and has lost some of its outer layers, by process(es) still speculative.

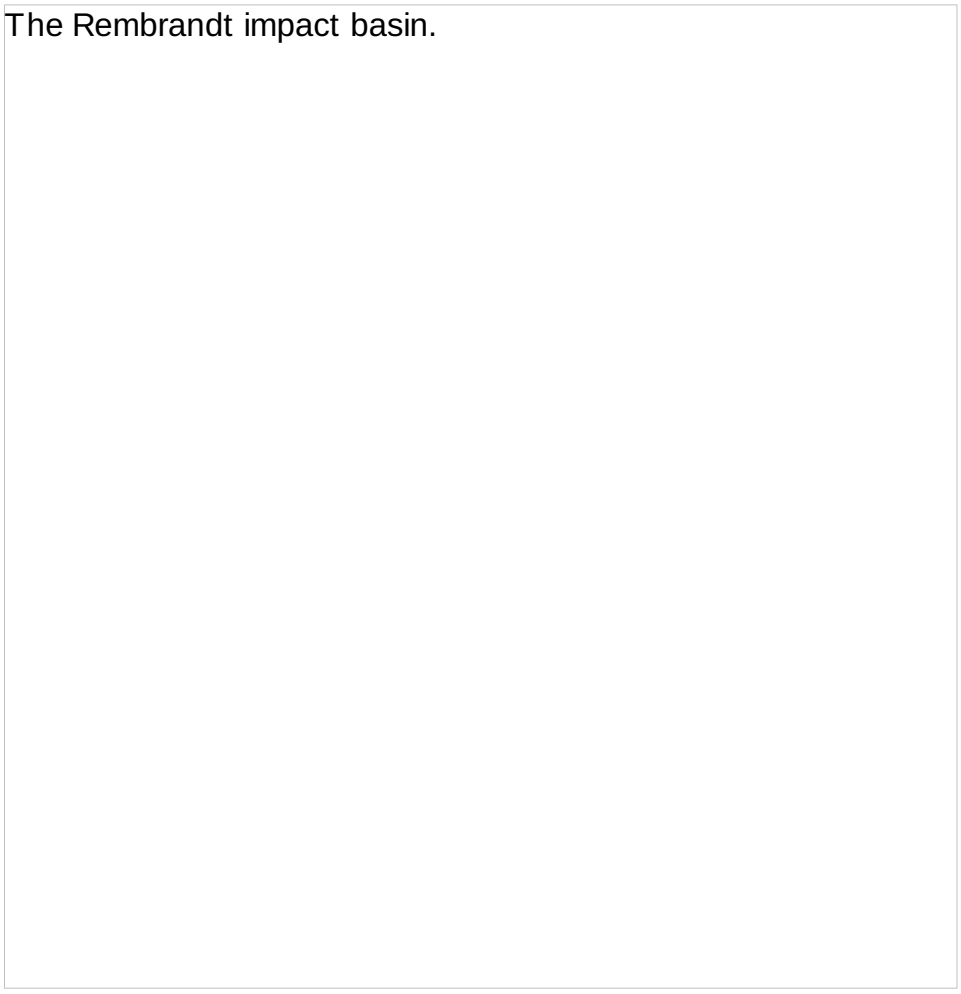
The second Mercury flyby took place on October 6, 2008. Messenger imaged surfaces never seen in previous passes (but examined by earth-based radar). This view shows several prominent craters which appear fresher (younger), from which ejecta rays emanate:

Rayed craters on Mercury.



This flyby discovered another, previously unknown large (715 km [425 miles] diameter) impact basin, given the name "Rembrandt":

The Rembrandt impact basin.



Another flyby occurred in September of 2009. In this color image, a bright (colored yellow) apron around a crater is most likely impact ejecta but possibly could be ash from a volcanic eruption. In the central image is a dark blue area which shows faint traces of a rim of a crater 180 km in diameter.

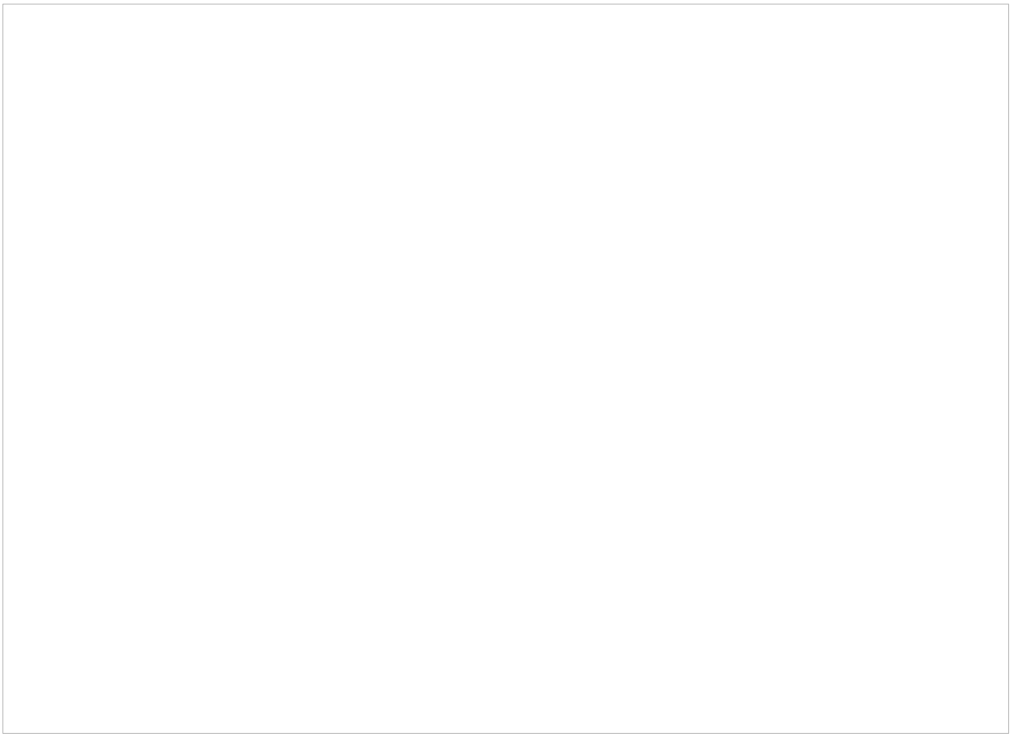
Both Messenger and Mariner provided evidence of Mercury's interior. Its density of ~5.4 gm/cc indicates a large core made up of a heavy metal - most likely iron. Compared with Earth, Mercury's core is larger, its mantle is less thick, and its crust is at least 100 km thick on average. This is a reasonable model of the interior:

The interior of Mercury.

The planet Venus is nearly the size of Earth. Unlike the Earth, it does not disclose its surface features, as these are perpetually shrouded with clouds. Its high reflectivity (albedo of 0.71; it is evidenced in the night sky as the bright "Evening Star" easily seen from Earth) implies a dense atmosphere, so that knowledge of its surface has depended either on landers or on cloud-penetrating radar.

Often the brightest "star" in the sky, Venus is Earth's closest neighbor. It was considered special to the ancients, some of whom thought it to be a planet. Along with the four moons of Jupiter, the phases of Venus led Galileo (who observed it with his remarkable telescope) to use these discoveries to argue for the heliocentric concept of the Solar System, i.e., the Earth and the other planets all revolve around the Sun. The phases could be best explained with the fact that,

like our Moon, to an Earth-based observer the progressive lighting of the venusian disc results from the changing position of Venus as it too revolves around the Sun. This image, made as a montage of successive observations of the venusian phase stages, demonstrates the progressively greater illumination of the venusian surface as seen from Earth:

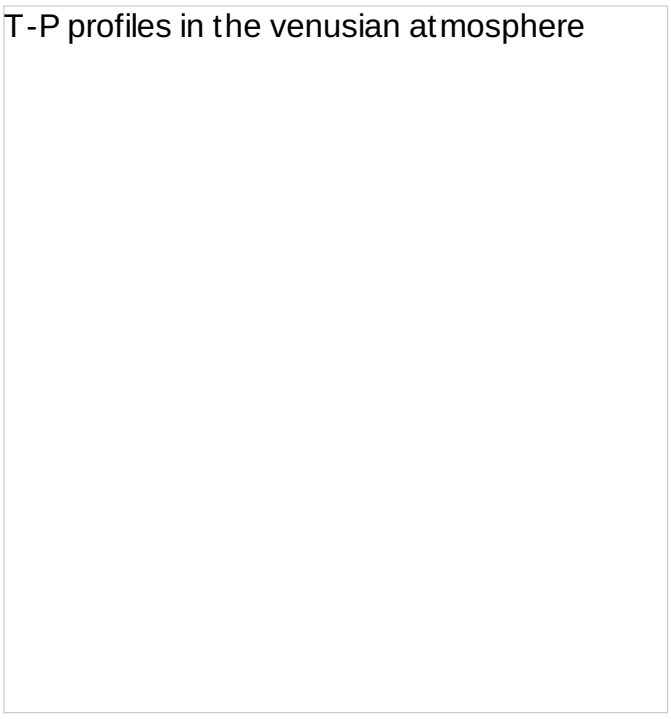


Venus, which lies at 0.72 A.U. from the Sun (and comes as close as 44 million km to Earth), has a period of revolution around the Sun of 225 Earth days and a very slow 243 day rotation which is retrograde (spins clockwise as seen from the north pole instead of the counterclockwise motion of Earth and most other solar planets), so that a sunrise would appear to begin on the western sky to an observer on the planet. Being slightly smaller (diameter: 12,100 km) than Earth, and 88% of Earth's volume, Venus has sometimes been called Earth's twin, but on close examination of its atmosphere and its surface, both quite different from Earth, the similarity in size is coincidental. Venus' interior is constructed of a solid iron core (thus no magnetism), a thick mantle, and an outer crust whose major features are relatively young.

The inpenetrable cloud cover masking Venus's surface was first breached by Earth-based imaging radar beams sent from antennae at the Arecibo Observatory in Puerto Rico, the Goldstone Tracking Station in California, Haystack in Massachusetts, and others. Wavelengths vary from 3.8 to 70 centimeters. Interference techniques using Doppler shifts process the reflected signals which offer some information on dielectric constants, surface roughness, slopes and rather crude estimates of elevation differences. Surface resolutions (areal) can be as low as 100 km (62 miles) or can be better than 3 km (2 miles). Here is an image showing variations in intensity of a backscattered radar beam transmitted to Venus at 12.5 cm from the Jet Propulsion Laboratory's Goldstone Tracking Station in the Mojave desert.

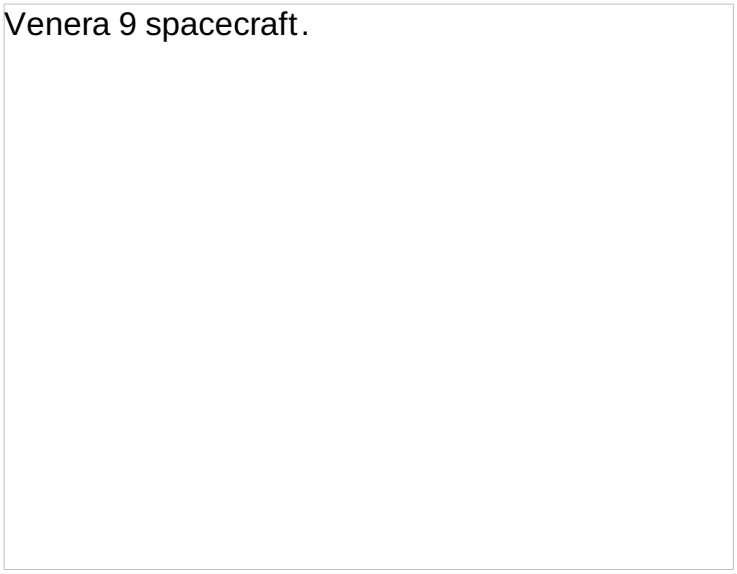
A series of missions by both the U.S. and the Soviet Union have unlocked some of its mysteries. Exploration of Venus by flyby probes was part of NASA's Mariner program which also included trips to Mars and the above-mentioned Mercury passes. Mariner 2, with its infrared and microwave radiometers, was the first American interplanetary probe, launched on August 27, 1962. Passing Venus as close as 41,000 km (25460 miles), it determined an approximate temperature for the outer cloud deck of ~500° C. Mariner 5, in 1967, came within 10,150 km (6300 miles), using these and a UV sensor to add more to the database.

About this time, the first Soviet probe, Venera 4, descended by parachute through the atmosphere in an attempt to touch down on the surface. It apparently was crushed by the dense atmosphere (~90 atm) and high temperatures but did return information confirming that CO_2 makes up about 97% of all gases present (very little water) and detecting droplets of sulphuric acid in the outer cloud deck. Venera data (refined by later Pioneer data) also lead to a general profile of temperature and pressure distributions in the atmosphere:

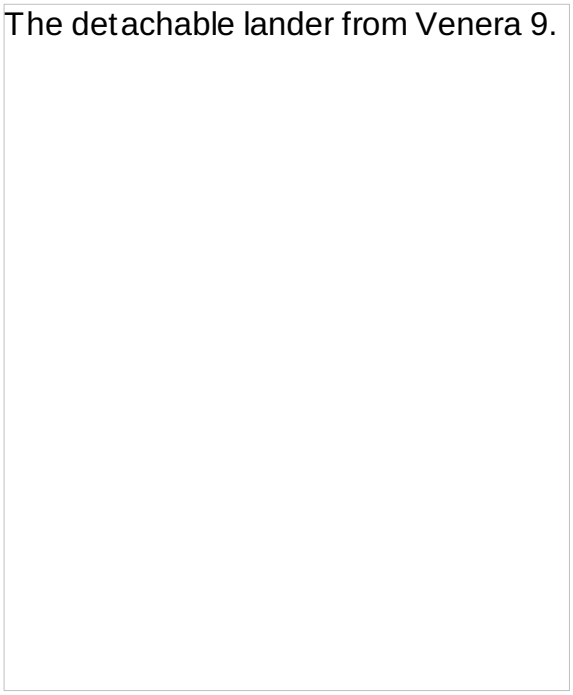


After two more failures, Venera 7 reached the surface and survived for 23 minutes in 1970. It gave the first specific surface temperature (750°K) and pressure (90 to 100 atm.). Venera 8 also succeeded in 1972, adding chemical composition data on radioactive U, Th, and K from analysis by a gamma ray spectrometer that suggests local rocks are potassium-rich (4%) basalts

containing 200 ppm Uranium and 650 ppm Thorium (both major heat sources). Measured surface temperatures were ~470° C. Four more Veneras reached the surface between 1975 and 1982; each carried a photographic system that returned pictures of the immediate surroundings. Venera 9 is shown in these photos, first of the entire spacecraft and second of the lander.

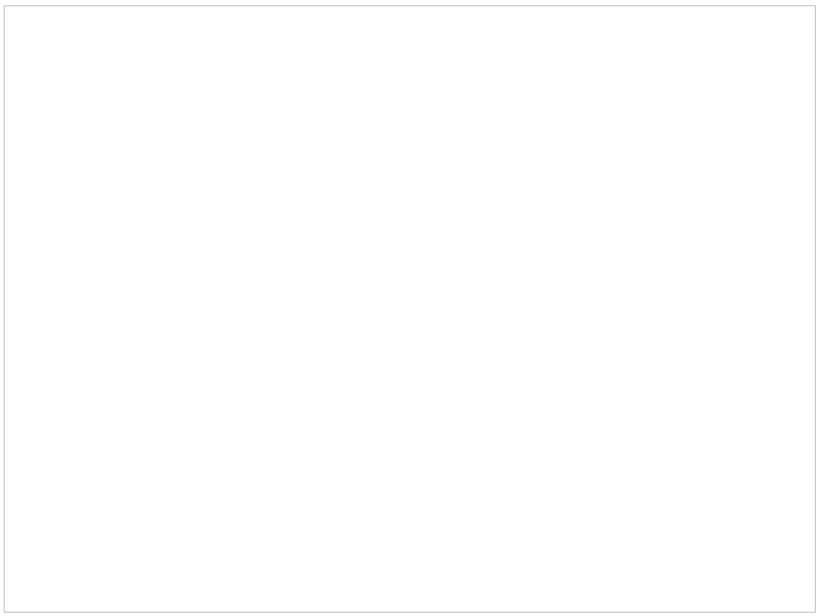


Venera 9 spacecraft.

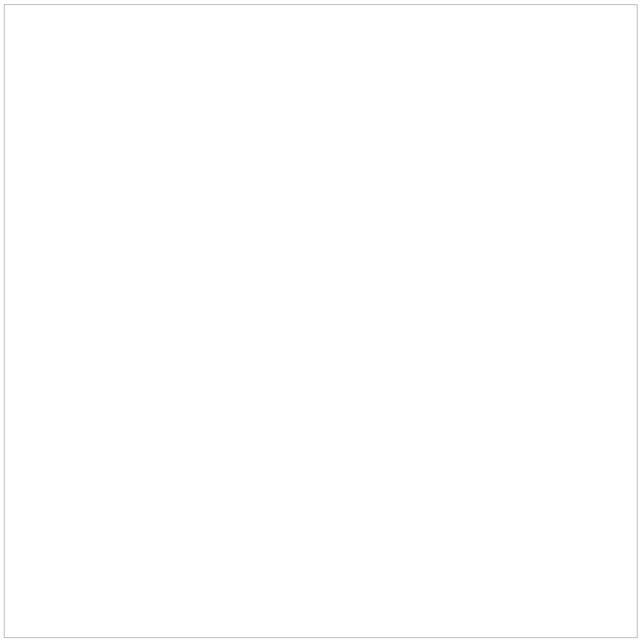


The detachable lander from Venera 9.

Two views, taken from Venera 9 and 10, disclose a rocky surface; note in the upper image a distinctive rock that reminds some viewers of a MacDonald's hamburger.

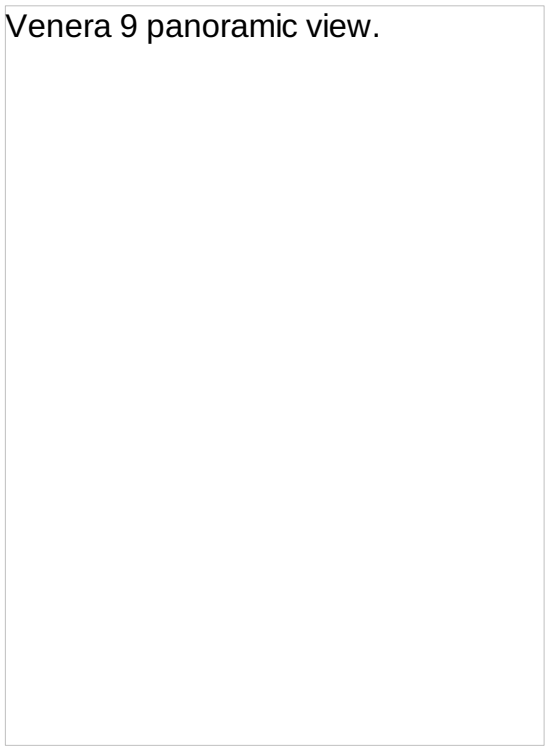


This is a close-up view of some of the Venera 9 rocks.



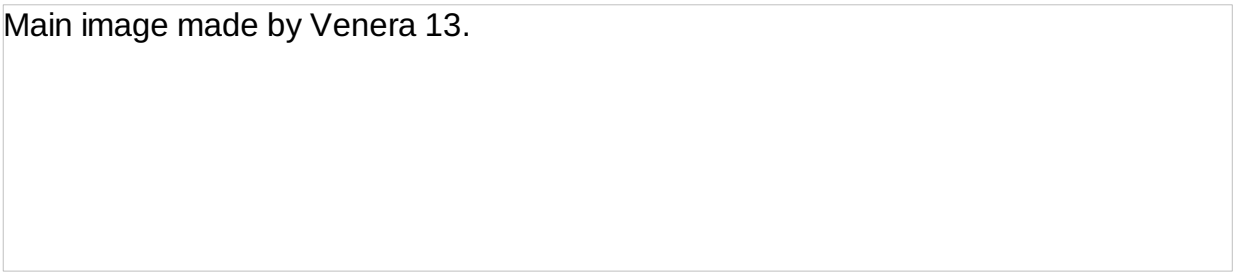
Here is another Venera 9 image, a panoramic view. Note its darkness, the result of much less sunlight reaching the venusian surface.

Venera 9 panoramic view.

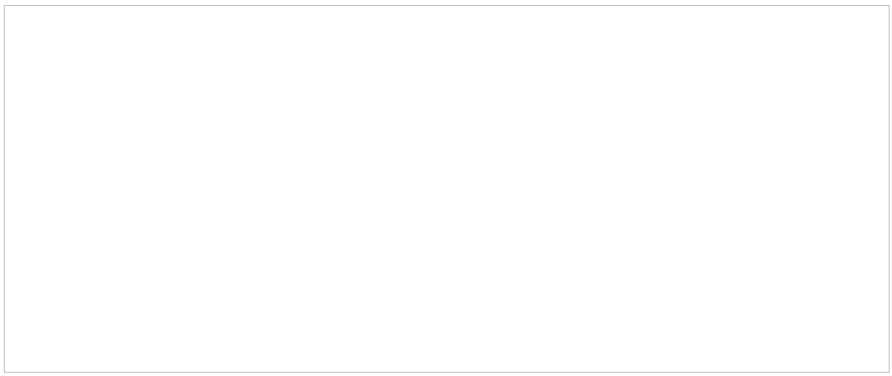


The first color images of the venusian surface were obtained by Venera 13 and 14.

Main image made by Venera 13.



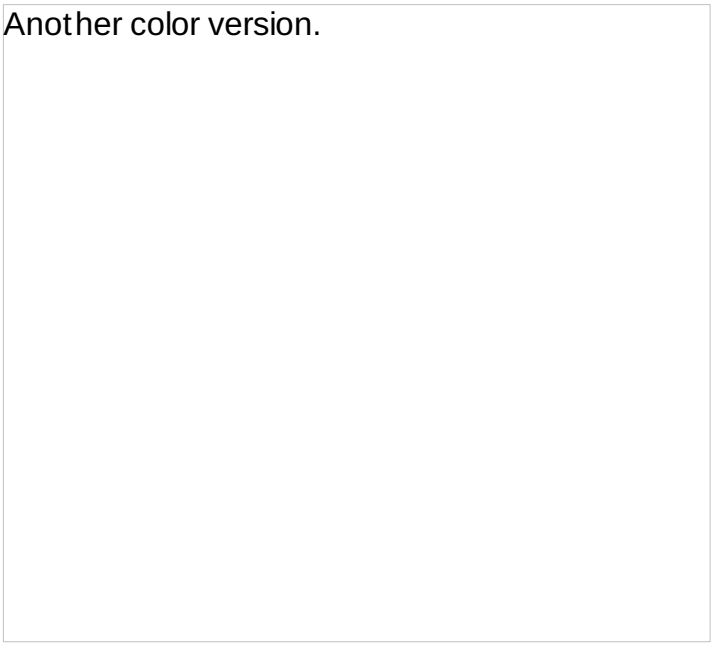
A view in enhanced color from Venera 13 suggests an iron-rich oxidized surface:



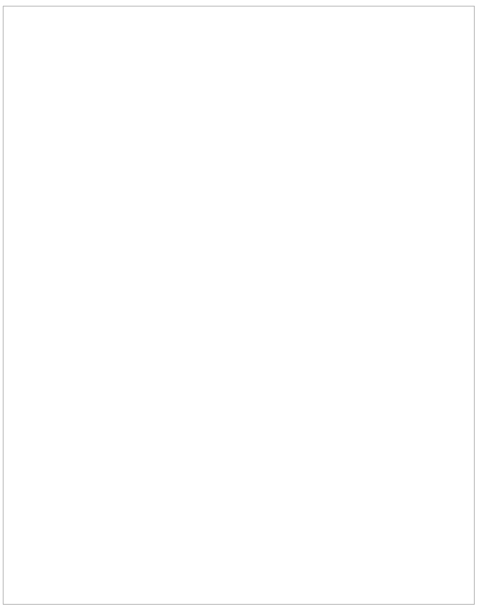
19-26: Ignoring the reddish iron surface stain, what does the other dark rock remind you of (in terms of rock type)? ANSWER

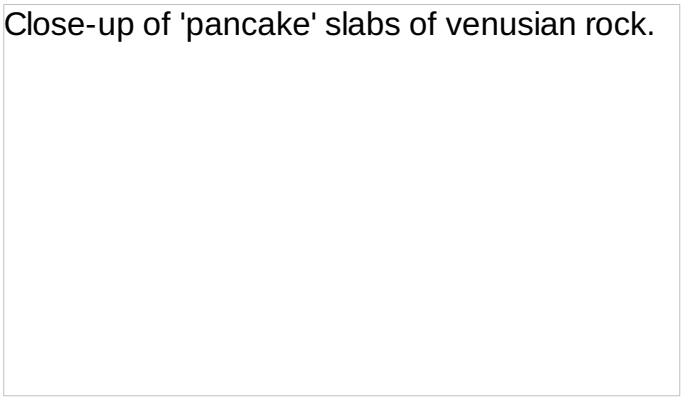
There is some debate as to the actual color of the landscape if one could stand at the venusian surface. This view made by Venus Express emphasizes the color based on the Venera results:

Not only the surface appears reddish, but also the sky (blue and green light is filtered out).

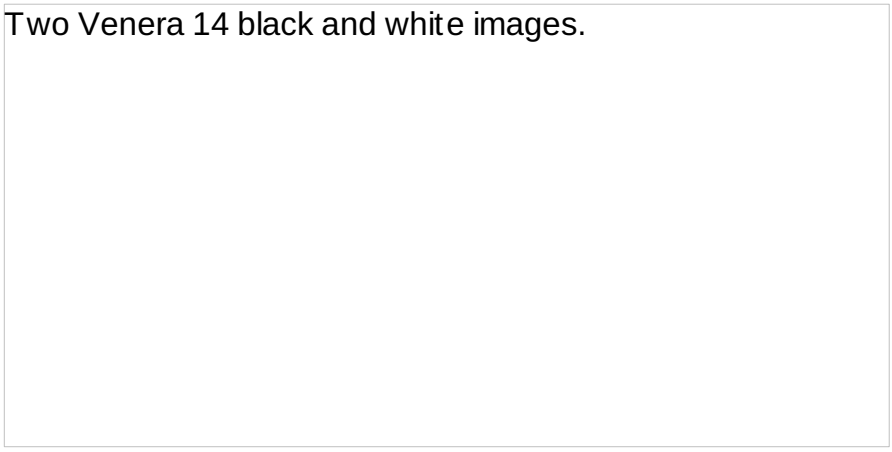


The next pair of images were taken of the rocks around the Venera 13 landing site.





Venera 14 also landed successfully and operated for well over an hour. Here are its main views:



The next U.S. probe to Venus was Mariner 10, arriving in February 1974. Using a special UV filter, its imaging camera was able to penetrate the CO_2-dominated atmosphere to detect cloud swirls that emphasized concentrations of excited carbon monoxide, suitable as markers of the general circulation patterns (winds up to 370 km/hr [230 mph] within the gas envelope).

This rendition, using blue instead of the near true color seen above, helps to define the cloud-rich from the cloud-poor parts of the atmosphere

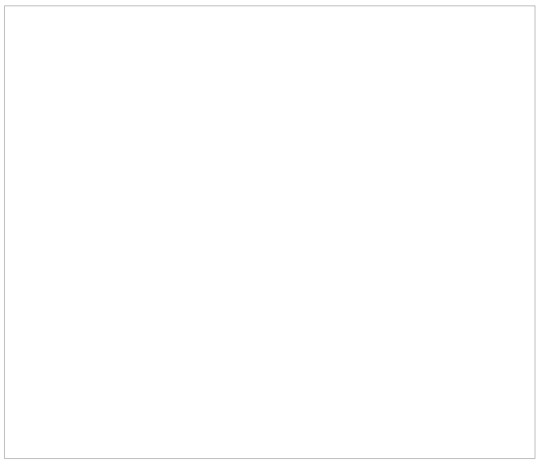
Mariner 10 image of Venus in the UV.

Mariner 10 found a circulating pattern of clouds - a vortex - developed over the southern polar region:

The Mariner 10 view of the north polar vortex was even more dramatic:

The north polar vortex.

Two Pioneer Venus spacecraft - the first an orbiter; the second a multiprobe - arrived in late 1978. Pioneer Venus 1 (Pioneer 12 in the multi-planet series), shown below, entered orbit around Venus with a diverse compliment of instruments, listed beneath the illustration:



Cloud photopolarimeter - measured the vertical distribution of the clouds

Surface radar mapper - mapped planetary topography and surface characteristics

Infrared radiometer - monitored IR emissions from the Venusian atmosphere

Airglow ultraviolet spectrometer - measured scattered and emitted UV radiation

Neutral mass spectrometer - evaluated the composition of the upper atmosphere

Solar wind plasma analyzer - measured properties of the solar wind

Magnetometer - examined Venus' magnetic field

Electric field detector - studied the solar wind and its interactions with the Venusian atmosphere

Electron temperature probe - examined the thermal properties of Venus' ionosphere

Ion mass spectrometer - measured the ionospheric ion population

Charged particle retarding potential analyzer - Studied ionospheric particles

2 radio science experiments - mapped Venus' gravity field

Radio occultation experiment - helped characterize the atmosphere

Atmospheric drag experiment - upper atmosphere density measurements

Radio science atmospheric and solar wind turbulence experiment

Gamma ray burst detector - monitored gamma ray burst events

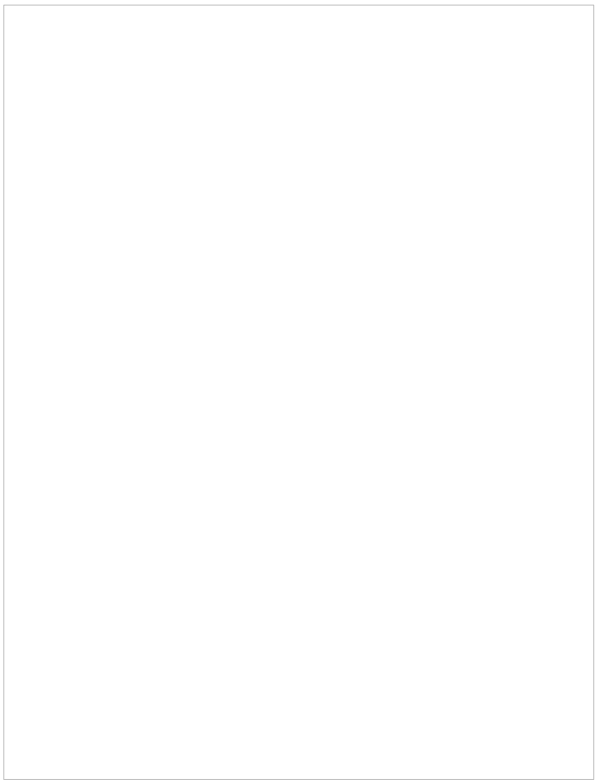
After a six-month journey, Pioneer Venus 1 entered an elliptical orbit around Venus in December 1978 and began a lengthy reconnaissance of the planet. The spacecraft returned global maps of the Venusian clouds, atmosphere, and ionosphere, measurements of the interaction between the atmosphere and the solar wind, and radar maps of 93% of the planet's surface. In 1991 the Radar Mapper was reactivated to investigate previously inaccessible southern portions of the planet. In May 1992 Pioneer Venus began the final phase of its mission, in which the periapsis (orbital low point) was held at 150-250 km until the fuel ran out and atmospheric entry destroyed the spacecraft. Despite a planned primary mission duration of only eight months, the probe remained in operation until Oct. 8, 1992.

Pioneer Venus 2 consisted of a bus which carried one large and three small atmospheric probes. The large probe was released on Nov. 16, 1978, and the three small probes on Nov. 20. All four entered the Venusian atmosphere on Dec. 9, followed by the bus. The small probes were each targeted at different parts of the planet and were named accordingly. The North probe entered the atmosphere at about 60° latitude on the day side. The Night probe entered on the night side. The Day probe entered well into the day side, and was the only one of the four probes that continued to send radio signals back after impact, for over an hour. With no heat-shield or parachute, the bus survived and made measurements only to about 110 km altitude before

burning up. It afforded the only direct view of the upper atmosphere of Venus, as the probes did not begin making direct measurements until they had decelerated lower in the atmosphere.

Among the Pioneer Venus results of atmospheric measurements were: 1. The atmosphere was found to circulate in large planetwide systems, much simpler than the circulation patterns on the Earth; 2. The atmosphere, which has winds moving at speeds in excess of 300 km/hr, appears to be decoupled from the rotation of the planet itself, which is much slower on a daily basis (it takes 244 Earth days to make one complete rotation); at these speeds it would take about 4 Earth days for an object inserted into the atmosphere to be driven completely around the planet by atmospheric circulation; 3. A collar of polar clouds was discovered, which may be part of a large atmospheric circulation vortex; 4. At least four distinct cloud and haze layers were found at different altitudes; 5. The haze layers contain small aerosol particles, including sulfuric acid droplets.

Both Venera and Pioneer Venus data confirmed the small (1.5%) but important role of sulphur in the venusian clouds. These clouds are composed of droplets of sulphuric acid, sulfur particles, and SO_2 (this last constituent accounts for absorption in the UV that explains the darker bands. The sulphuric acid is derived from reactions of water vapor in the atmosphere with sulphur compounds released from volcanoes (no evidence yet of active ones today). One model, by Keven McGouldrick, shows a possible chemical scenario:

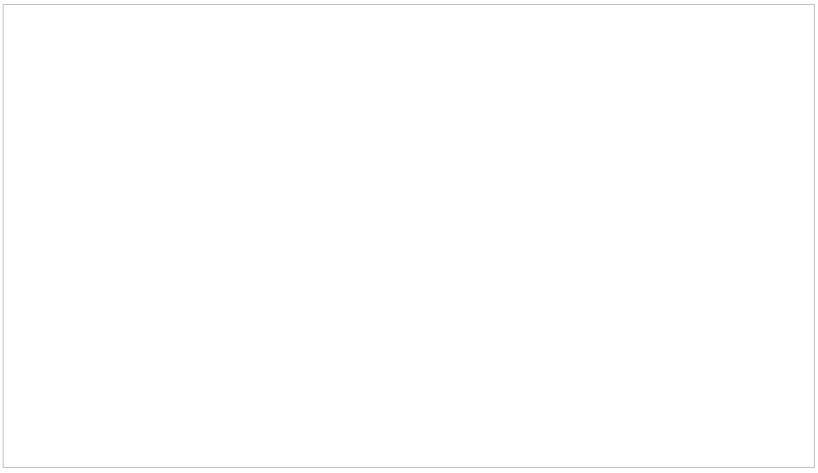


Although this suphuric acid has a strong corrosive effect on the surface and on manmade probes that pass through the atmosphere, possibly to land, it is the CO_2 that most influences the planet. The gas came largely from volcanoes. Over time it built up concentrations in the atmosphere much higher than on Earth. This has led to a classic example of the "Runaway Greenhouse" effect (see Section 16). The result is that heat from both the surface and the Sun has been trapped in the gaseous envelope causing the high observed temperatures. Variations in CO_2 levels within the atmosphere were measured during a flyby by the Galileo satellite enroute to Jupiter. The Galileo instrument used was NIMS (Near Infrared Mapping System);

Various probes through the atmosphere have revealed why the observed cloud patterns seem to be "bent" in a V profile. The upper atmosphere conditions cause winds near the equator to move faster than those in polar regions, causing a drag effect that is almost independent of the planetary surface retrograde motion. Technically, the phenomenon of differential cloud movement is the result of combined 4 (earth) day equatorial and 5 day Rosslyn wave interactions.

Pioneer Venus 1 provided good quality radar imagery of most of the venusian surface. A radar altimeter was used. The image below is of a part of that surface, rendered in strong contrast to emphasize slopes:

The next two Pioneer Venus 1 illustrations are of parts of the venusian surface that have been colored to represent changes in elevation.



Venus Pioneer 1 data were eventually organized into a map the covers almost the entire surface of Venus (some polar data are missing). This provided the first map of this planet. Science teams had "fun" in giving names to various features that were then officially adopted. The later Magellan maps (next page) have added more names.

Map of Venus with named features.

Even these earlier radar images pointed to a relatively flat Venus but with some highlands exceeding 6 km (3.7 miles). Two continental-sized areas of higher elevation are Ishtar Terra and Aphrodite Terra. These contain most of the mountains that may not be primarily volcanic in origin (are more beltlike). Mountain terrain comprises 10% of the venusian surface; 70% are upland plains; 20% are lowland plains.

The Soviet Venera 15 and 16 orbiting spacecraft (1983) carried imaging radar (8 cm wavelength) capable of 1-2 km (0.6-1.2 miles) ground resolution that gathered coverage over about 25% of the planet. The scene below is in the general Ishtar Terra region of northern Venus and shows the eastern part of Laksmi Planum, the wrinkled Maxwell Montes, and the large crater Cleopatra, a scene well over 2000 km (1240 miles) wide at the base. Beneath it is an enlargement of part of the image, showing Maxwell Montes in more detail.

19-27: What is the most conspicuous geologic feature in this scene? ANSWER

Starting with these first views of the venusian surface, and amplified by the Magellan observations, a nomenclature for landforms and features on Venus evolved. Here is a table which summarizes these:

Venus landforms.

The Category column indicates how individual names, based mainly on Greek mythology, are chosen.

Now on to Magellan - one of the most successful NASA missions ever!

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

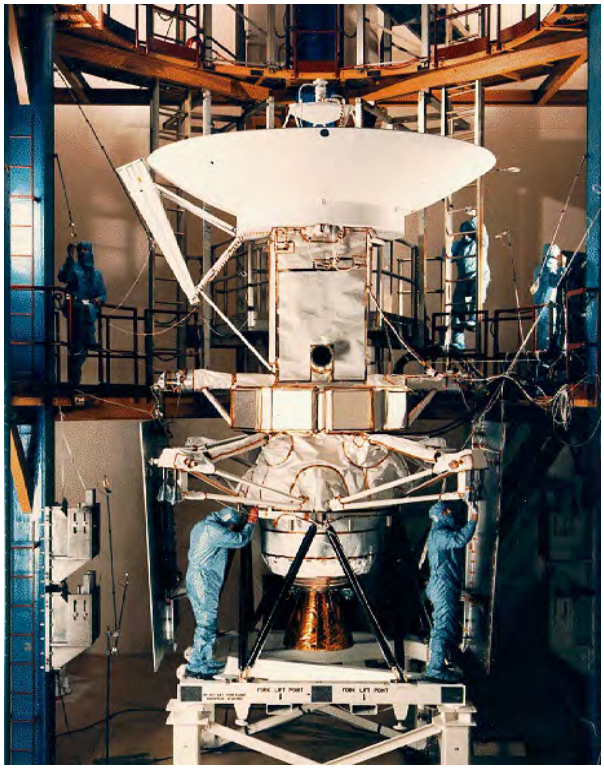
One of the most productive interplanetary probes ever conceived, JPL's Magellan was launched in 1989 and arrived to orbit around Venus in 1990. Its primary sensor was an imaging radar that over time provided views of nearly the entire surface at resolutions ranging between 120 and 360 m. It also could map elevations to +/- 50 meters and had a radiometer that determined brightness temperatures to +/- 20° K. This page elaborates on the first-order observations from the mission and shows some full views of Venus and a Venus surface map. It also describes the abundant examples of volcanism and volcanic features that dominate the venusian surface.

The Magellan Mission

One of the supreme triumphs of planetary exploration was the Magellan program, developed and run by the Jet Propulsion Laboratory, to study Venus close-up by penetrating its cloud cover. This was the first spacecraft ever launched from a Space Shuttle (Atlantis), on May 4, 1989, to another planet.

Magellan almost didn't happen! The original concept for a spacecraft to Venus was VOIR - Venus Orbiting Imaging Radar. After a design was completed and a go-ahead from NASA Headquarters was sought (meaning "funding approved"), a budget squeeze during the Reagan administration forced its seeming cancellation. However, the JPL Venus team was undaunted. They began to search for already existing spacecraft components, some of which were found, that cut costs. Their next plan submission was also denied. They persisted and made further cost savings, so that finally the plan fell within current budgetary constraints and Magellan was approved. This superb spacecraft was built.

Here is Magellan at the Kennedy Space Center awaiting loading into the Shuttle:



Magellan's primary instrument was a multimode Radar Mapper (2.385 Ghz, or 12.6 cm wavelength). In the SAR imaging mode, looking between 18° and 50° off-nadir, it could capture scenes with resolutions between 360 m and 120 m (1181-394 ft), depending on its position within its elliptical, near-polar orbit (at altitudes between 275 and 8,443 km (171 to 5,246 mi above a mean venusian radius of 6,051 km [3,760 mi]). It first established orbit on August 10, 1990 after 1 1/2 loops around the Sun. Its altimeter mode achieved a vertical accuracy of better than 50 m (164 ft) within a ground cell of 10 km (6.2 mi) diameter. Operating in a radiometer mode, the radar could sense surface radio-emission, whose signals can be converted to brightness temperatures with an absolute accuracy of ±20° K. Investigators gleaned information on mass distribution (causing gravitational anomalies) from Doppler frequency variations due to gravity effects that varied orbital speeds. Even as Venus rotated slowly beneath (one complete day every 243 Earth days), during stages of its orbit closer to the planet, the radar imaged surface swaths between 17 and 28 km (10.6-17.4 mi) wide. Through the first cycle lasting 8 months, it mapped 84% of the surface. In the next 16 months, that percentage rose to 98%. Additional coverage provided repeat looks in search for possible transient or short-term changes. After several adjustments to lower orbits, the spacecraft finally burned up in the venusian atmosphere, in mid-October, 1994.

Image strips covering thousands of kilometers, especially after being joined as mosaics, provide stunning views of a fascinating venusian surface that are still undergoing thoughtful interpretations. Although Venus no doubt formed concurrently with Earth, its surface today is largely younger than one half billion years (Earth has some surficial regions older than 2 billion years). Planetologists base this estimate on venusian crater frequencies. Even though not uncommon, the numbers of resolvable impact structures are consistent with 1) destruction of the much larger numbers from the first 4 billion years, most probably by active processes that removed them by lava overplating (resurfacing) and by still arcane erosional actions, and 2) asteroidal flux rates for the last 5 hundred million years, in line with estimates from other planetary surfaces. Effects of volcanism are conspicuous, with thousands of small volcanoes detected, along with many lava channels. Although fracture zones and sets of close-spaced ridges are evident, no direct indications of terrestrial-like plate tectonics are discernible. Surface water, if ever present, left no signs of stream or ocean activities, and would have escaped from the planet (traces are present in its atmosphere) as Venus heated up, until a massive "runaway greenhouse" warming effect overwhelmed the planet. The slowly rotating atmosphere seems to have caused some wind streaks and dune-like deposits on the surface.

The gallery of Magellan images is vast. We show only a select few here but you can access more at JPL's (Magellan Home Page).

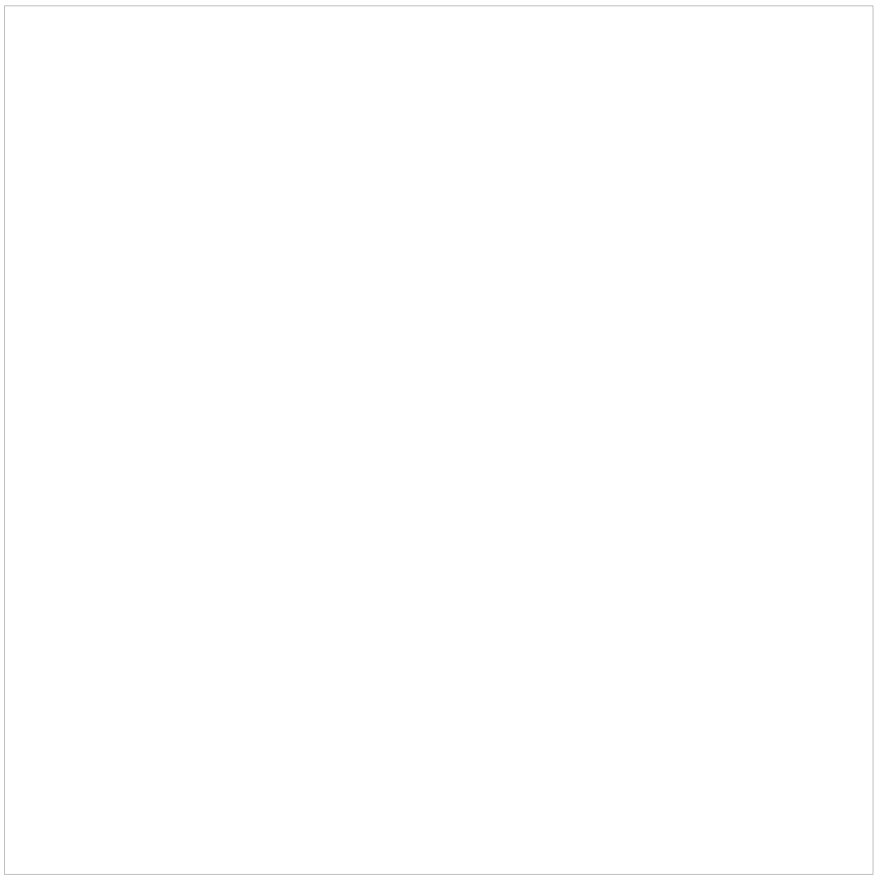
To familiarize you with some of the major features and their locations on Venus, look at this shadowed relief map of the non-polar regions of the planet with the key geomorphic features labelled, as prepared by the U.S. Geological Survey; this shows more details than the map on the previous page:

Shaded relief depiction of the venusian surface with principal localities named.

Next, consider this color-coded relief map of nearly all of Venus, on a Mercator-like projection, derived by integrating imaging and altimeter data.

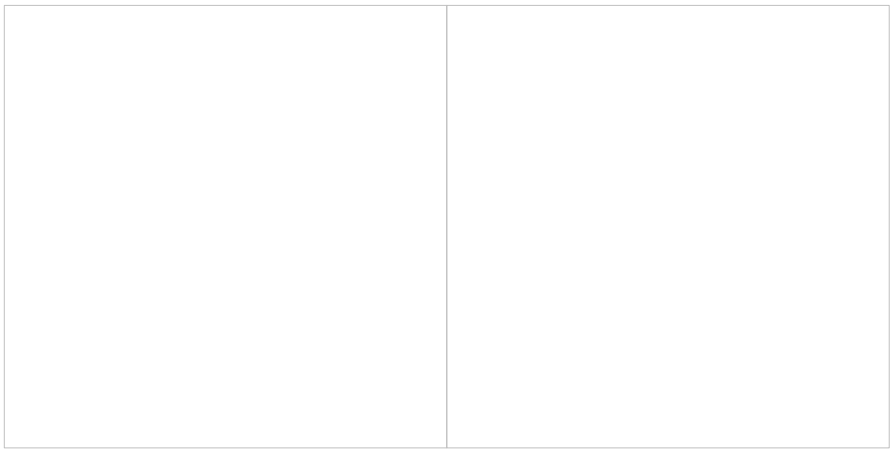
Blues represent the lowest surfaces followed by greens, then yellows and oranges with red being highest. The greatest elevations are within Maxwell Montes (top left), that includes the high point of the uplands known as Ishtar Terra. Another high region, near the equatorial center, is called Aphrodite Terra. Beta Regio, near the central left, is also elevated. Two blue regions in the north are low plains, called Sedna Planitia, below Ishtar, and Atalanta Planitia, well to its east. A large curved channel south of Aphrodite is known as Artemis Chasma.

Now look at a hemispherical projection that lies within this full map. It was made by the U.S. Geological Survey using Magellan topographic data. A vertical line through the center lies at 180 degrees east. Try to identify the high central region (in pink); refer to the first relief map above (hint: think of a lovely goddess).



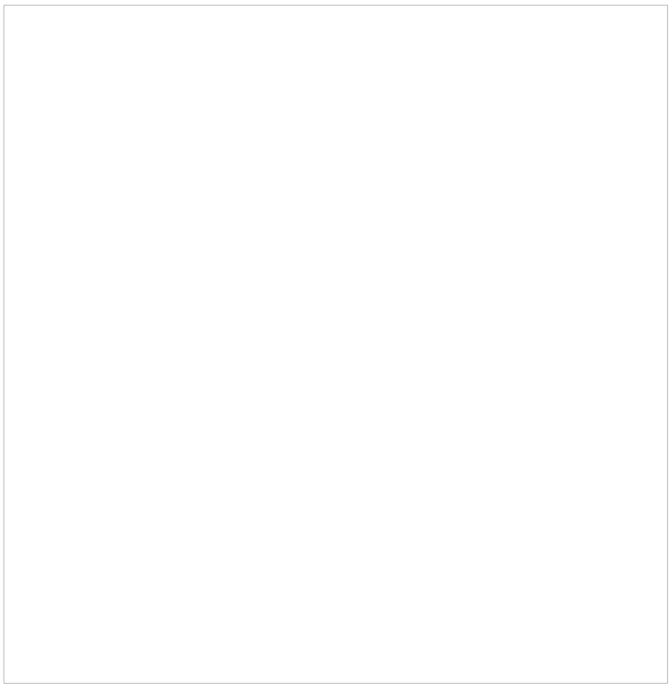
19-28: Try to identify the high central region (in pink), enlarged from the first map (hint: think of a lovely goddess). ANSWER

Here are four other topographic maps located as indicated in the captions:



We can also display (below) this same hemispherical segment as a quasi-natural color image of

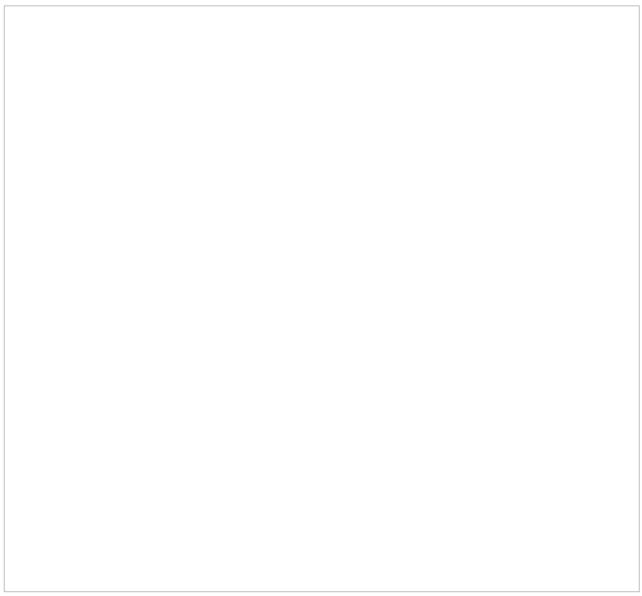
a mosaic of rectified Magellan scenes. There is no direct proof that Venus has this much red (the choice of assigned colors was a rather arbitrary, best guess) but, if so, the presence of oxidized iron could account for such tones. The dark, blackish low areas are presumably basalts.



19-29: Once again, try to orient yourself in this image relative to the shaded relief map and localities described on that map. ANSWER

Magellan carried a microwave experiment (managed at MIT) from which a map of thermal emissivity (see Section 8) could be derived, as shown here. Note that the lowest emissivities (in blue) are found in the highest parts of the venusian surface, implying that the rock types there were other than basalt.

Thermal emissivities help to pinpoint "hot spots" and "cold sinks". Here is an emissivity map of the Atla Regio region on Venus which indicates the details procured by the radiometer:



Magellan's greatest revelations were a wide variety of volcanic features. It is not an exaggeration to refer to Venus as the "Volcanic Planet". Most of the planetary surface has been judged by geoscientists to be less than a half billion years old. Some volcanism appears to be recent and the possibility that there is even now some activity cannot be dismissed (although no changes were observed during the mission which last until 1992). Impact craters (next page), while present, are uncommon, with perhaps 1000 large enough to be resolved by Magellan's radar; this is consistent with the presumption that the present venusian surface is relatively youthful, as the crater flux by then would have greatly diminished (as extrapolated from terrestrial crater frequency since the end of the Precambrian). There have been older surfaces, some parts of which may still persist at the surface, but these are largely "paved over" (resurfacing) by the continuing activity.

This (cluttered) map shows the major volcanic features and their locations on the venusian surface

Volcanic features on Venus.

The first volcanic feature we will look at is characteristic of the plains regions. Here at Lakshmi Planum are several light and dark surfaces that are interpreted to be equivalent to the basalt flow types known as *pahoehoe* (smooth lavas; somewhat specular surfaces) and *aa* (chunky lavas, better backscatterers [lighter tones]) - counterparts to common Hawaiian lavas.

Multiple lava flows on Lakshmi Planum.

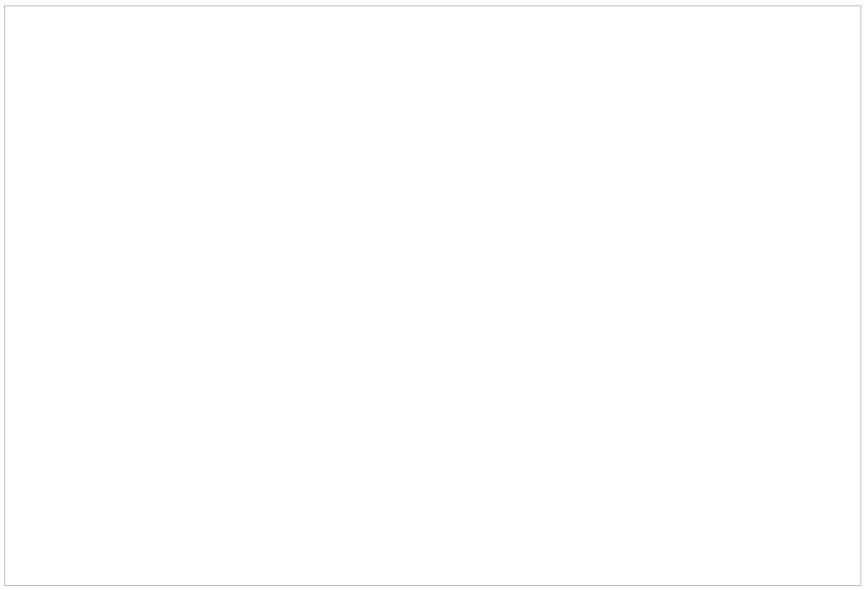
A series of Lava flows emanate from the Sils Mons volcanic source:

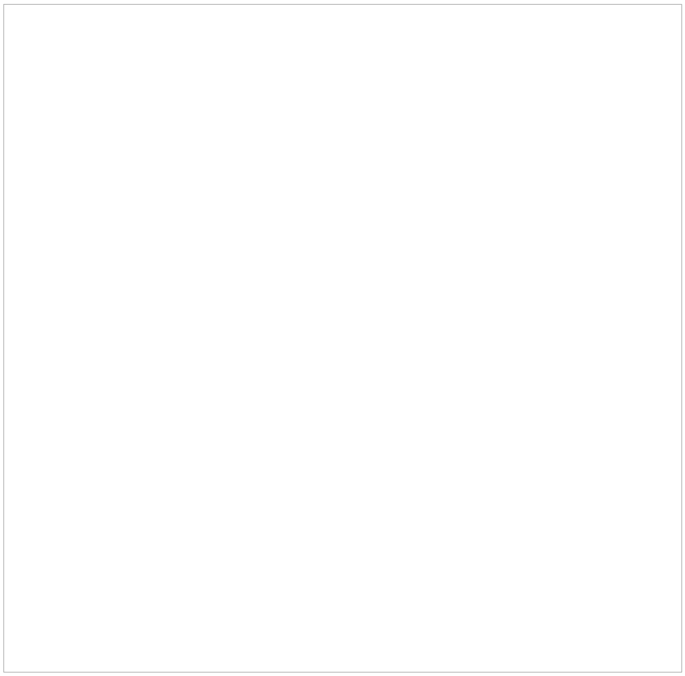
This radar image shows a long channel filled with volcanic flow material, over which a younger flow has straddled; note volcanic material on the right. This is the Ammavaru flow sequence in the Lada region. The scene's dimensions are 450 by 630 km:

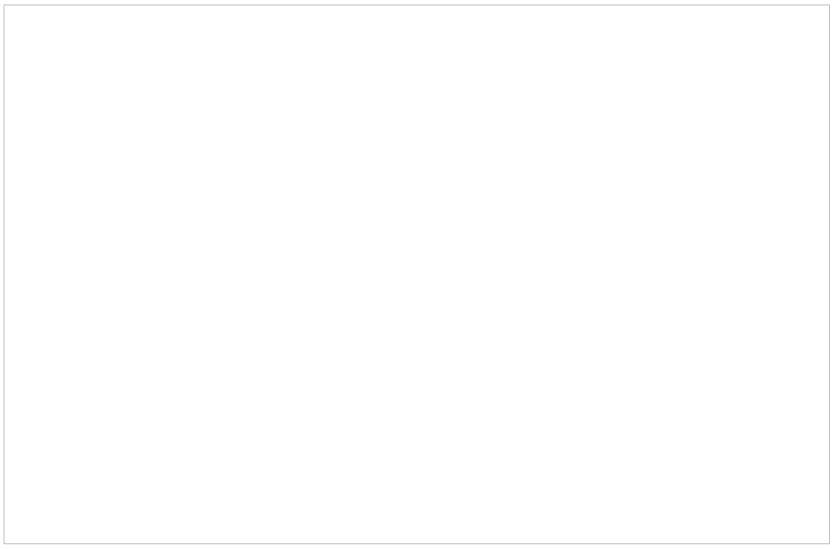
Ammavaru flows in the Lada region.

One of the longest flows (1000 km long) occurs as Myletta Fluctus in Lavinia Planitia.

Like the Moon, thin channels and sinuous rilles have been found on Venus. Here are three examples (check captions for description):





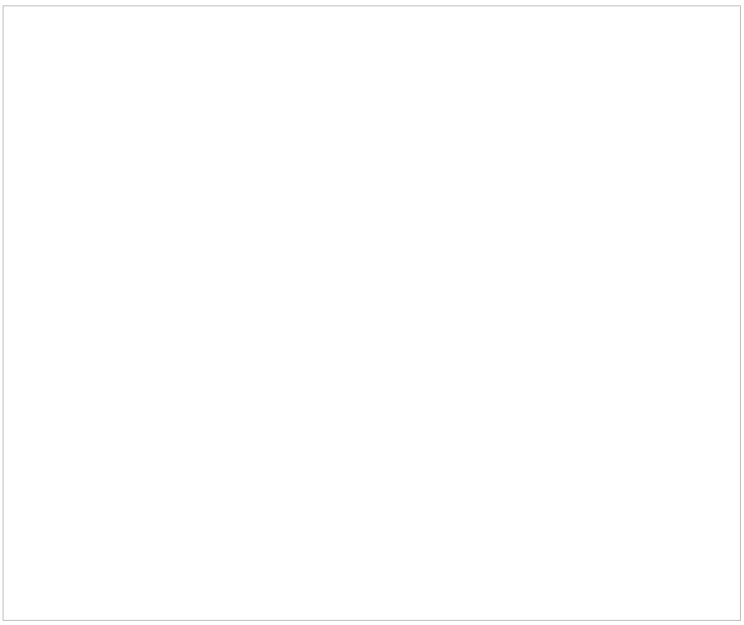


Flows can sometimes be traced to shield volcanoes (with central calderas) as exemplified by Theia Mons, 4 km high, with a central caldera measuring 75 by 50 km and surrounded by a lava field reaching 800 km in maximum dimension.

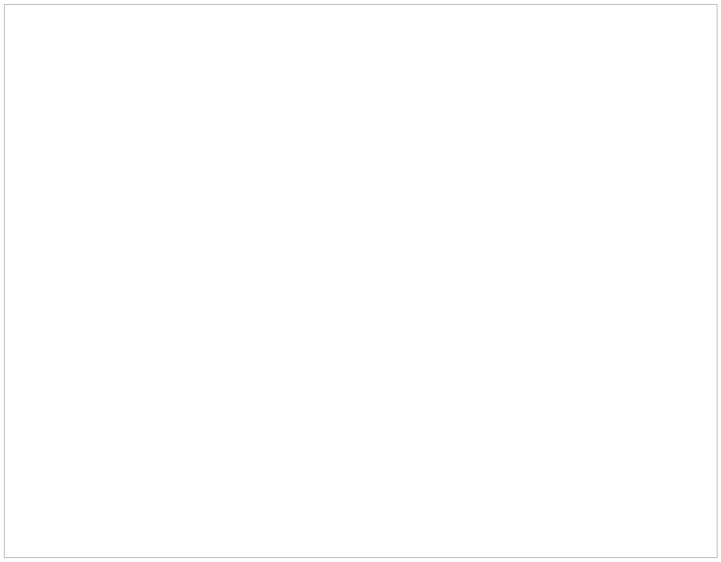
Theia Mons.

Another major volcano, seen in this colorized rendition, is Sapus Mons (1.5 km high; 120 km at its base, on an upwelled domical surface 1000 km in diameter) in the Alta Regio region.

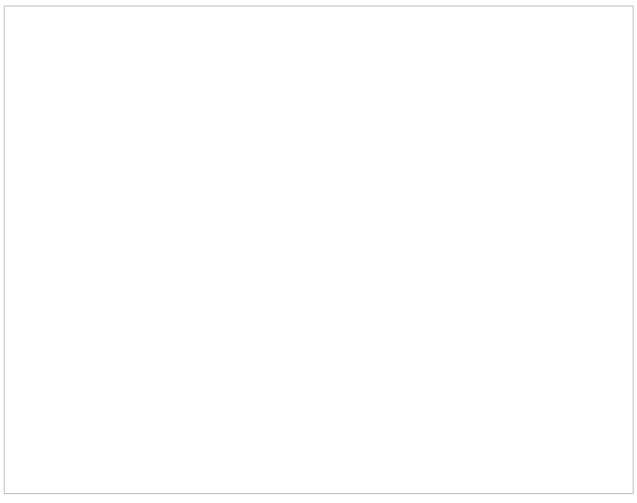
Perspective color views of this type of volcanic structure, made by applying altimetry data to the radar image, show it to be much like a shield volcano, with a broad base and often a central depression. This rendition of Sif Mons, about 2 km (1.2 miles) high and covering an area of nearly 300 km (200 miles) in diameter, illustrates this:



One of the highest mountains on Venus is Maats Mons which reaches to 8 km (5 miles) above the mean venusian elevation; its shape is transitional to a stratocone, suggesting its lava may have differentiated into that of an intermediate silica content. A younger lava flow from the main volcano appears as bright flow:



Topographic maps made from Mariner data can bring out structures that are large volcanoes:



Some of the larger volcanoes have summit calderas. The largest on Venus is Sacajawea Patera (140 km [89 miles] in long dimension):

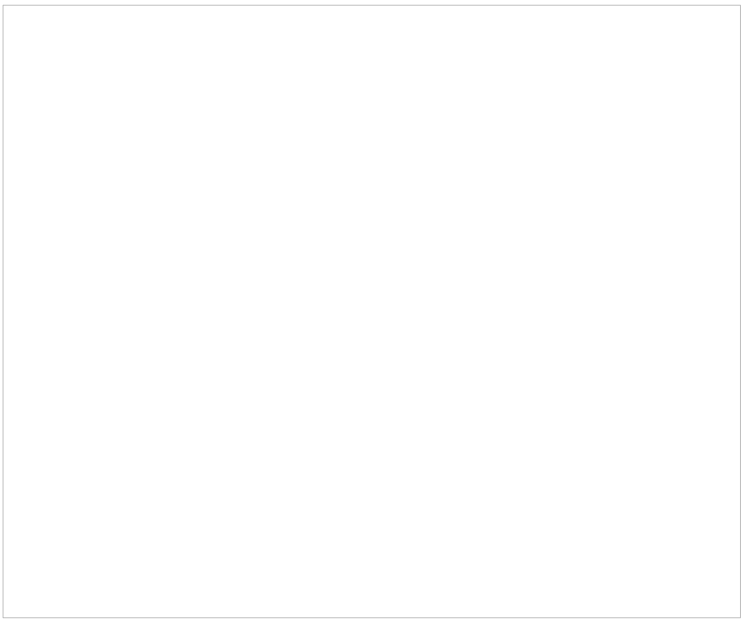
Sacajawea caldera.

Similar to that is this 30 km wide caldera

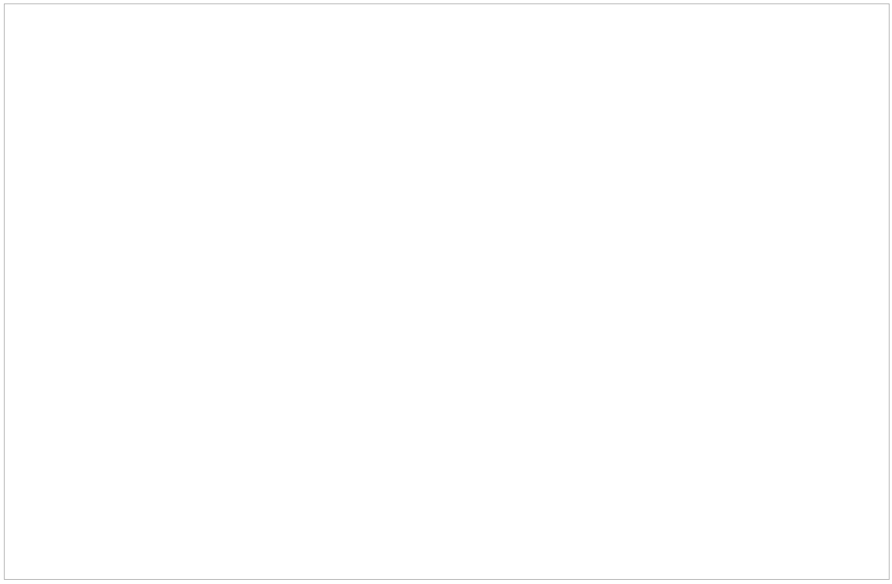
A volcanic caldera on Venus.

Such calderas often look almost identical to large impact structures (discussed below). A case in point is the circular depression below which could have been identified as such except for the prominent lava flow emanating from its side. (Note: another interpretation considers this to be a genuine impact crater filled post-impact with shock-generated melt that leaked out.)

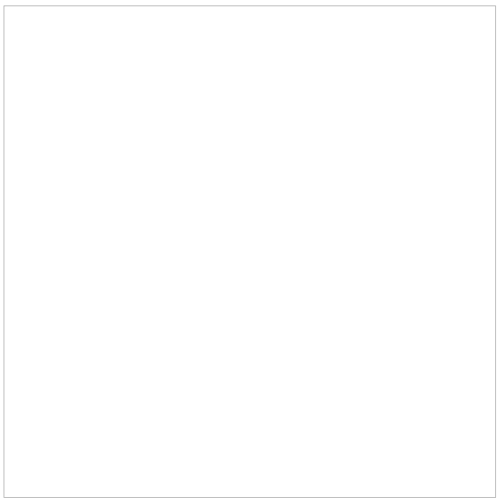
A variant of the caldera type has been given the descriptive name of "tick volcano" because of its resemblance to the insect of that name. Emanating radially from the crater walls are ridges that form the "tick legs".



While these large shield volcanoes are uncommon on Venus, there is a much larger number of smaller volcanic structures (those that rise up from the surface). This map locates most of these by category, and also includes the principal volcanic fields which contain features like those shone above:

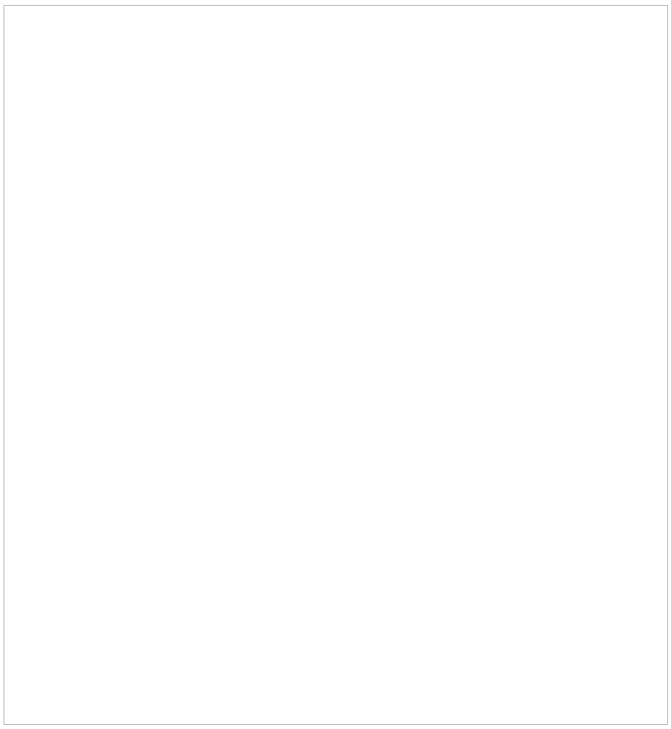


Cinder cones and stratocones (Vesuvius-like) are rare on Venus. Here is one example of a swarm of cones (each about 2 km wide) on the plains that are larger than terrestrial cinder cones but not typical of stratocones.



19-30: If large volcanic edifices like, say, Mount Rainier or Fujiyama were to occur on Venus, what would that imply? ANSWER

Typical small shield volcanoes occur in swarms such as is evident in these views:



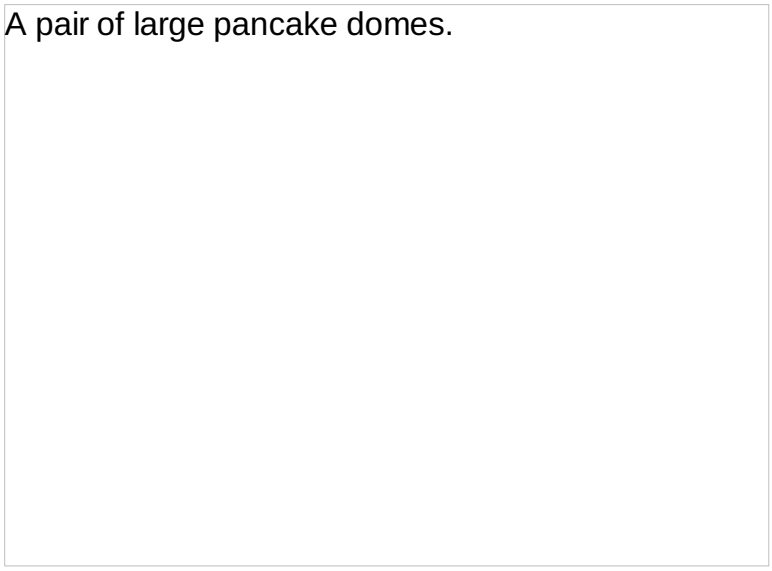
A swarm of small shield volcanoes.

Irregular flattened shield types, called fan volcanoes, are built up by several overlapping outpourings are illustrated by this example:

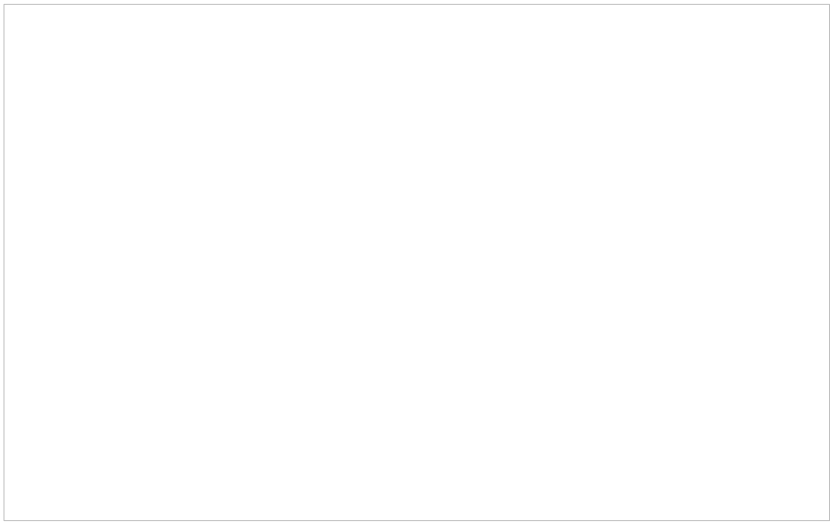
Fan Volcano.

Domical hills (Tholi), shieldlike in structure, as much as 25 km (15.5 mi) wide and up to 750 m (2,460 ft) high, dot the plains of Alpha Regio. Astrogeologists believe these pancake-shaped features result from upwelling at tubular vents of lavas that spread uniformly in all directions.

Here is a closer look at two such domes in Tinatin Planitia; the larger is 65 km wide:



Using laser data, a pancake dome in Alpha Regio is shown in a perspective view:



Some of the small domes develop distinctive flows around them that have reminded some venutian planetologists of "sea anemones", to which that name is colloquially applied.

An anemone-like volcanic structure.

A recent report from Venus Express scientists offers some evidence that at least a few venusian volcanoes are active. Hot spots have been detected both on the surface and in the atmosphere. This Express image shows one of the surface structures whose temperature is elevated compared with its surroundings.

A possibly active venusian volcano.

For a good summary of venusian volcanic activity, check out the Wikipedia website on this subject.

On the next page, we finish our tour of Venus now with examples of some other landforms, some having a volcanic connections, others of a different nature and origin.

PREVIOUS NEXT

Primary Author: Nicholas M. Short, Sr.

Recent Innovations about the Concept of "*Universe*"

Dark Energy and an Accelerating Universe?

A major clue to the open vs closed vs flat Universe seems to have been found in recent results from HST and ground "super-telescopes" observations. In fact, some cosmologists believe this assemblage of new facts to be in composite the most important and enlightening discoveries about the Universe since Hubble's observations more than 80 years ago. Some of what is presented in the next group of paragraphs has been extracted from an excellent review in the NOVA series on PBS, this November 2000 program being entitled "The Runaway Universe", from the references cited below, and from a recent book - *The Accelerating Universe: Infinite Expansion, the Cosmological Constant, and the Beauty of the Cosmos*, by Mario Livio, 2000, J. Wiley and Sons. The latest volume on this subject is *The Extravagant Universe*, by Robert Kirschner, Princeton University Press. A quick overview of the accelerating Universe idea is given at this Web site which deals with one of the currently competing concepts, Quintessence, a special form of the Dark Energy that seems to be powering this acceleration.

The basic premise behind this so-called runaway Universe is summarized in this diagram prepared by the Space Telescope Institute in Baltimore, MD:

Following the initial inflation during the earliest moments of the Big Bang, the strong gravitational forces, during the extended time when the Universe was smaller and matter/density was higher, led to attraction forces that caused a net deceleration in expansion (or, said another way, with time expansion should slow its rate). But a momentous discovery in the late 1990s has shown that, as the Universe continued to enlarge and its density reduced the gravitational effect, at some stage in cosmic time the Universe then began to re-accelerate (or, in effect, the deceleration rate itself slowed so that the galaxies started to move apart faster). Current speculation points to a repulsive force, probably in the form of "dark energy" analogous to that underlying Einstein's Cosmological Constant, that is gradually dominating the driving forces that

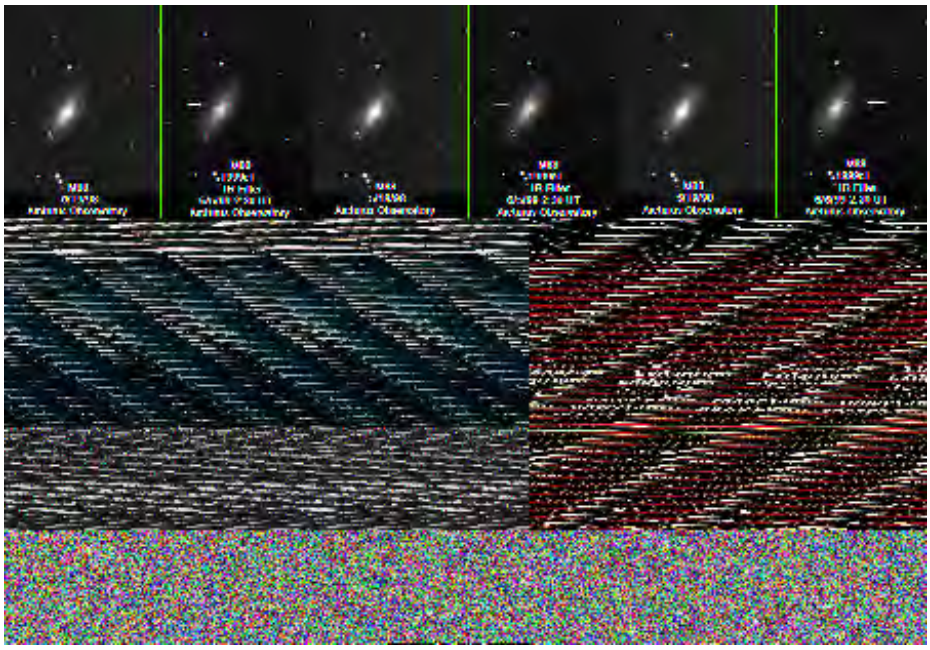
determine relative expansion rates. Succinctly put, Dark Energy is a form of vacuum energy which has three defining properties: 1) it appears to be uniformly distributed (smooth) throughout the Universe, has been steady over cosmic time, and has a very low density (estimated at 10^{-26} kilograms per cubic meter in intergalactic space), 2) in contrast to the attractive force of gravity, it acts as though it impresses a negative pressure (producing the aforementioned repulsive force) on all matter in the Universe; it has been called "repulsive gravity" since it acts to counter the effect of conventional gravity; and 3) despite its relative weakness, it's influence increases in proportion to the increasing size of the Universe. Its net effect, at present, has been to cause the rate of Universe expansion to accelerate during the second half of Universe history. In the next few paragraphs, these ideas are expounded at some length.

Dark Energy has so far proved difficult to detect directly and thus to pin down as to its nature. (It is sometimes referred to as "Phantom Energy".) The prime supportive evidence for this renewed expansion acceleration is being found through determination of redshifts of certain Supernovae. The results from analysis of more than 60 Type 1a Supernova events lying between 4 and 9 billion light years from Earth (time-distance regions [intervals] in which the Supernovae included in the study are still close enough to be easily observed) imply faster rates of expansion at some stage in cosmic time than heretofore calculated from deep quasar studies. The Supernovae appear fainter than theory predicts for their distance from us, implying that they have moved farther than expected in the time involved, which would be explainable by an increase in expansion speeds at some stage. (As we shall see this could have occurred after expansion started or perhaps it is a "hangover" from Inflation.)

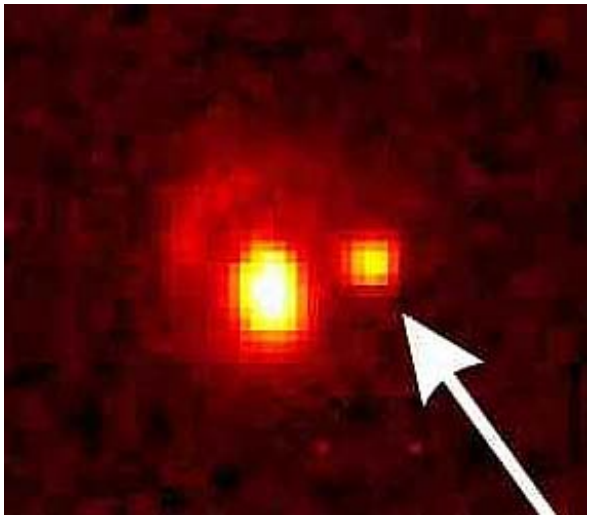
In fact, such rates indicate the Universe appears to be expanding *too fast* for it to be slowing down enough to finally contract. This speaks of an *open* Universe that will expand forever and leads to the corollary that there must be insufficient total mass (perhaps only ~ 30% of the amount needed and most of it being Dark Matter [the visible radiation components of the Universe such as galaxies comprise only 4%]) to ever close it). In order to account for this apparent acceleration, a form of energy that induces repulsion is assumed to exist.

Several ideas in the preceding paragraph are among the hottest topics of discussion in the astronomy and cosmology communities. In 1998, a conference on "Missing Energy in the Universe" focused on many relevant aspects. The announcement of evidence for an expanding Universe was judged the top scientific story of that year. Three review papers in the January 1999 issue of *Scientific American* present details relevant to this growing concept of an ever expanding and accelerating Universe. The interested reader should consult these papers (either in the library or at the three Internet sites cited as links below) but a brief synopsis of each is given in the next sequence of paragraphs.

The first article, **Surveying Space-Time with Supernovae** (Hogan, Kirshner and Suntzeff), enlarges on the nature of Supernovae. Supernovae are detected routinely by taking telescope photos of small segments of the sky at different times, superimposing them as computer images, compensating for differences in observation conditions, and subtracing later from earlier images (nulling out the same features that persist), and identifying residuals that have the appearance of Supernovae. The astronomers goal is to find a Supernova that is still in its early (front-end) stage of explosion and then follow its period of high luminosity over the next several months. These images show a "before" and "after" sequence in which a Supernova appears in the M88 galaxy in 1998 but this had diminished a year later:



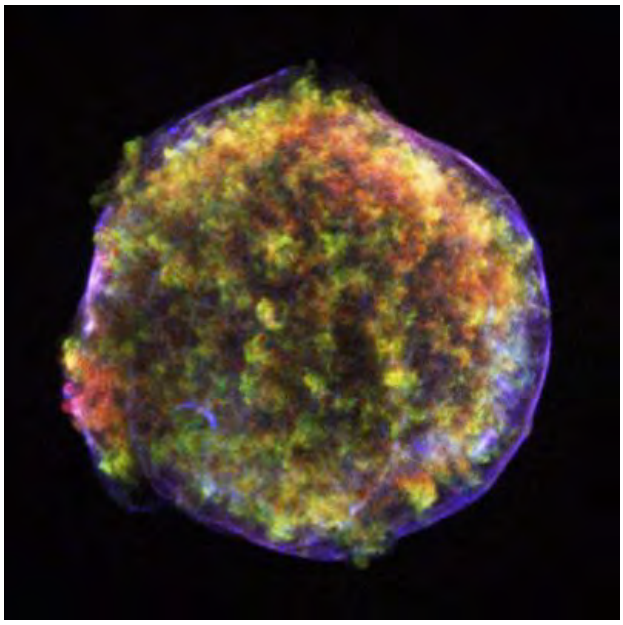
Here is another example of a more distant galaxy (bright center) with a second bright area off center that is the Supernova.



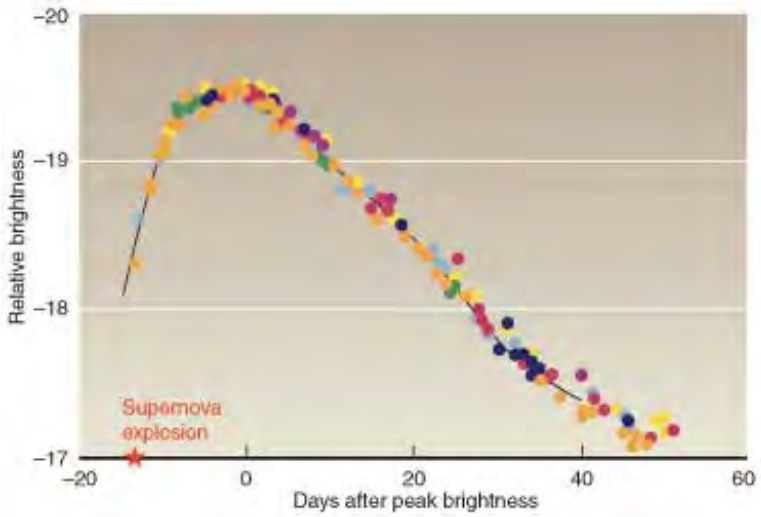
The main thrust of the article is a review of a particular type of Supernova. These, called Type Ia, are as small as ordinary stars that, after reaching the White Dwarf stage, continue to receive matter drawn from a companion star. This influx of new material onto the Dwarf increases its total mass until temperature rises to a level forcing a sudden thermonuclear explosion that results in a form of Supernova which lives for a short period whose duration depends on the mass. In turn, the mass determines the brightness. Generally, the amount of mass in the Type 1a that is involved in the Supernova is very similar in all such Supernovae, namely it is about 1.4 solar masses. When its temperature reaches about 10,000,000 ° K, its mass undergoes nuclear fusion causing the dwarf to explode rapidly. Thus, the energy released, in the form of detectable luminosity, is almost the same for every Type 1a event - meaning that these serve as nearly constant "standard candles". The luminosity decreases at a steady pace, in stages, so that the decline in luminosity is systematic. A Type 1a Supernova therefore is an excellent means of judging distance between it and Earth observers. If such a Supernova is detected almost at its outset (or its stage is known), the departure from expected luminosity output and the observed output is a function of its distance from Earth. The decrease in expected output follows the inverse square law that governs light source objects; thus a lamp of known brightness, observed first at position 1 becomes 1/4th as bright when moved to position 2 that is twice as far away.

A Type 1a Supernova occurs, on average, about once every 300 years in a galaxy but because there are many galaxies in which individual exploding stars can be resolved, a new one is found approximately once a month. Using this approach, astronomers have found a significant number

of Type 1a Supernovae lying at distances of 7 to 4 billion light years from Earth (as determined by redshifts) that actually occurred at those times in the past, when the Universe was 1/2 to 2/3 its present size. Here is a Type Ia Supernova:



The typical visual absolute magnitude of Type Ia supernovae is Mv = -19.3 (about 5 billion times brighter than the Sun), with little variation. Type Ia supernovae have a characteristic light curve, their graph of luminosity as a function of time after the explosion.



The curve shows a rapid buildup of luminosity during which the explosion occurs, then further buildup to a peak, followed by gradual diminution of luminosity. Near the time of maximum luminosity, the spectrum contains lines of intermediate-mass elements from oxygen to calcium; these are the main constituents of the outer layers of the star. Months after the explosion, when the outer layers have expanded to the point of transparency, the spectrum is dominated by light emitted by material near the core of the star, heavy elements synthesized during the explosion; most prominently isotopes close to the mass of iron (or iron peak elements).

Using a Type 1a Supernova as a "standard candle" (defined as a stellar light source whose luminosity or intrinsic brightness can be known with a high degree of accuracy and hence can be a means for estimating distance from Earth) has been a challenge, but work by several groups (the Perlmutter team; the Kirschner/Schmidt team, and others) has over the last decade achieved the ability to estimate brightness to within +/- 90% of actual value. These groups have found that Type 1a's brighten and then dim at variable rates. By analyzing brightness curves and applying theoretical corrections for the fade frequency, they have come up with a model that leads to a value for the brightness that allows it to act as an absolute magnitude event, for

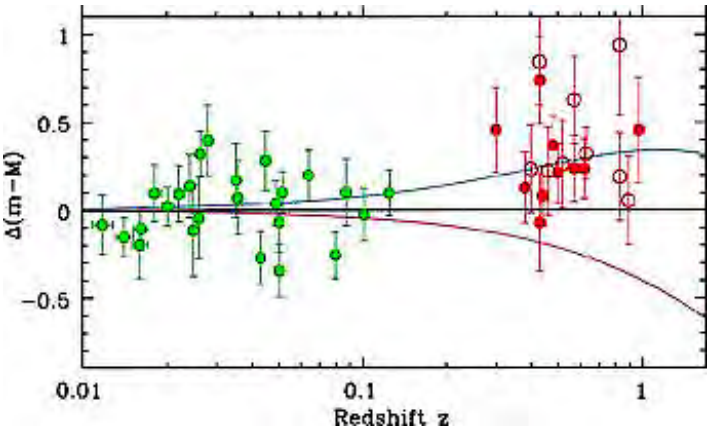
which distance (from Earth) can then be calculated with high accuracy. This distance should at face value indicate how far the star's radiation has traveled.

The surprise in these observations is that the Type 1a explosions were up to 25% less bright than they should be at the distances first postulated, i.e., these are fainter than expected assuming a simple slowing down of expansion owing to gravitational attraction. This relies on the postulated assumption that the Universe has continually been decelerating. Using redshift values for stars in the Supernova's galaxy, the expected rate of expansion for the galaxy relative to Earth as the observing point can be determined for the continuous deceleration case. But, the diminution from the expected luminosities implies that the Type 1a star and its galaxy are really farther away (and hence appear to us as less luminous than predicted from a steadily decelerating Universe) and must have reached their more extended positions because of increase in expansion rate at some time during the Universe's enlargement. Thus, those closer in are moving relatively faster which, since these appear to us as they are much later in cosmological time, indicates that the rate of deceleration of the Universe is in fact slowing down. Let's repeat for clarity: The observations are tantamount to this possibility: the Universe has not decelerated uniformly - when smaller, gravity slowed it more - but some accelerating process has reduced the deceleration rate beyond that expected with expansion, so that the divergence of space between galaxies is increasing faster than absence of that process would predict - something is pushing them farther apart more rapidly than the outward "coasting" of space expansion would suggest. To explain this, if subsequent studies continue to support the conclusion, requires either the presence of some type of energy that counteracts expansion or a different Inflation model.

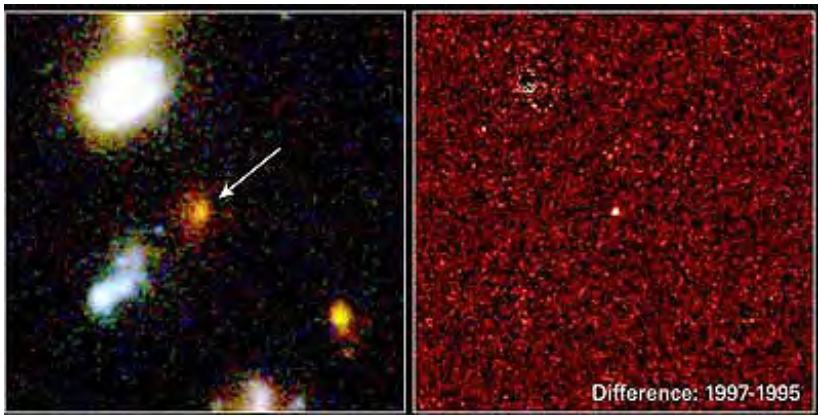
This familiar analogy should help you to visualize this variable expansion: You are standing still in a race car with the motor running. With brake released, you hit the pedal hard. The car accelerates rapidly to full speed. Without any braking, the car is allowed to gradually slow down. After coasting some distance, you touch the pedal again, this time softly, and start to counter the deceleration by renewing the acceleration. This pretty much describes the Universe's situation as of now.

Let's examine this idea of renewed acceleration in terms of redshifts. This expansion of space causes the light waves to become farther apart, and thus longer, and red shifted. The more time light has been traveling, the more time there has been for space to expand, and the greater the red shift. Thus more distant objects have a larger red shift. The redshifts for the Supernovae were greater than predicted. The accelerating expansion of the universe was inferred because distant Supernovae are fainter than expected based on their red shift. This means that they are farther away than one would expect based on the linear increase of redshift with distance. This is interpreted as evidence that the Universe is expanding more slowly in the past, so that the redshift of distant objects is less than one would expect. When the light from distant objects began its journey, the universe, and therefore space, was expanding more slowly. Thus there was a smaller contribution to the redshift than for light that traveled more recently. Nearer objects are less subject to this slowdown, because the universe was expanding faster when light left them. Thus theirred shift is comparatively larger, in proportion to their distance from us.

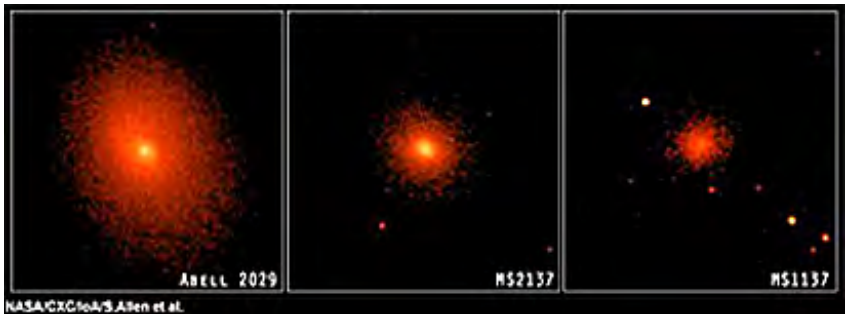
The diagram below shows the relation between shift in star luminosity and redshift. What it shows is that Supernovae 1a luminosities (m - M) measured within galaxies whose redshifts were near 1.0 (about 5 b.l.y. away) (plotted in red(were less than expected from extrapolation of the line passing within other galaxies at very low z values (hence nearby and visible as they were recently). Although other explanations for the departure are possible, renewed acceleration is held to be the most likely.



Astronomers would like to confirm that the pattern of expansion rates pictured in the top diagram on this page (slowing - then accelerating) extends even further back in time beyond the 5-7 billion l.y. observations just described. In 1995, the oldest Type 1a Supernovae known, lying near an elliptical galaxy, at a distance of ~11.5 billion l.y., has had its redshift measured, with results consistent with the acceleration model. It is shown as a bright incandescent cloudlike object in the left image below; the right image is a time difference overlay of the scene for a two year period in which most objects cancel out leaving a bright spot that is the Supernova remnant centered at the arrow point within the envelope of ejected materials. Its intrinsic brightness is greater than expected for its age and distance, consistent with the idea that it was slowing down during that time interval rather than speeding up as indicated by expansion rates in the younger Type 1a Supernovae; this supports the argument that during the first half of the Universe's history space itself underwent progressively slower contraction but since then has been experiencing expansion..

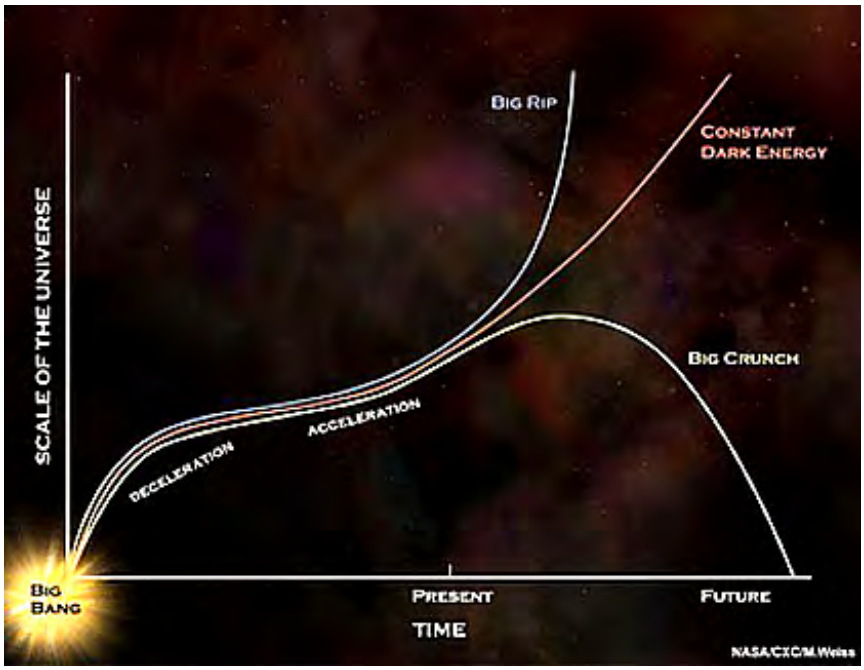


A group of astronomers under study leader Steve Allen at the Institute of Astronomy at Cambridge University was successful in amplifying our knowledge of acceleration by analyzing data for 26 galaxy clusters at varying distances from about 1 billion to 8 billion l.y. away. These clusters are each surrounded by huge balls of hot gas. Three of the 26 are shown here, in images made by combining Hubble and Chandra data:



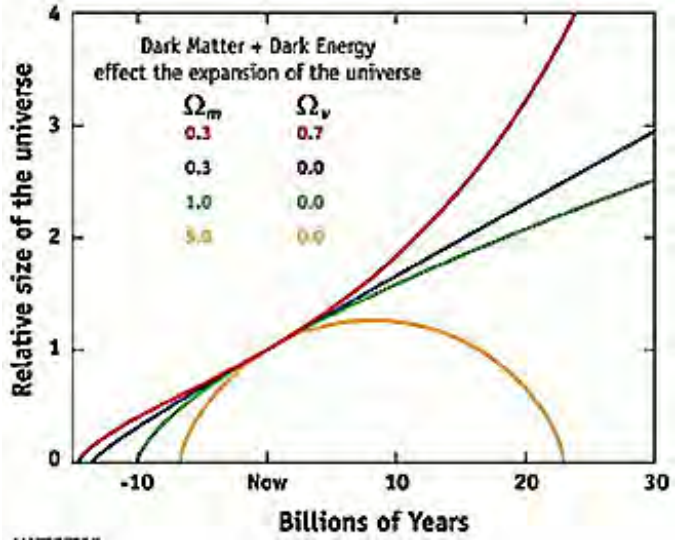
From theoretical considerations, they deduce that these clusters should have fallen apart and dispersed by now, yet they are intact. They ascribe the surviving coherence of the clusters to the influence of Dark Matter. They have calculated ratios of the mass of the hot gas (which

comprises about 85% of the luminous mass in a cluster) to the inferred mass of dark matter, using Chandra X-ray data. By assuming that the relative amount of gas is roughly the same proportionately at every cluster, these gas balls serve as another indicator of a luminous source of specifiable output, analogous to the Type 1A Supernova as a "standard candle". They calculate the distance to each using the redshift z values. Their findings: the oldest clusters show deceleration but about 6 billion years ago this reversed and the younger clusters are now accelerating (are out too far; their present distance being more than expected from a non-accelerating situation). A constant dark energy (**behaving very much like Einstein's Cosmological Constant**) is the best fit to the apparent renewal of acceleration. Their take on the possible models for the Universe's future is summarized in this diagram:



They conclude that neither the "Big Rip" (wide dispersal of the Universe's materials as everything cools down) nor the "Big Crunch" (return of deceleration and collapse of the Universe) are likely, but the Universe will continue to expand coherently into the infinite future but observers on Earth (if we survive <u>well</u> into the future) will loose "sight" of the farthest objects near the cosmic horizon.

So, what does this accelerating Universe hypothesis tell us about the type of Universe that this one appears to be. Look at this diagram - in which the four curves are further identified in terms of Omega (Ω) values (see below for definitions):



You saw a similar plot (credited to J. Silk) on the preceding page. The one labeled "Accelerating"

most closely resembles the Lemaitre Universe in the plot on page 20-9. For the time being, then, the evidence points to this type of spherical, open Universe as representing the one we've got. Comment: Was the Abbe Lemaitre (previous page) guided by Divine Providence, by scientific prescience, or was he just plain lucky?

In the second paper, **Cosmological Antigravity** (L. Krauss), the presence of a repulsive force (actually, some type of still mysterious energy) pervading all of space (including the so-called empty part that comprises the dominant portion of the Universe's volume) which offsets gravitational attraction can account for the observations and many of their ramifications. This energy is similar in important respects to the Cosmological Constant concept proposed by Albert Einstein more than 70 years ago to explain how a static Universe (the prevailing model at that time) can retain a constant size when gravity is acting to pull mutually on all matter, including the galaxies, such that matter subjected to attractive forces would ultimately collapse inwardly. In his view, there must be a repulsive gravitational force needed to stabilize this static Universe. Einstein later abandoned his idea (saying "It was my biggest mistake.") when evidence for expansion and the Big Bang became almost completely accepted. In retrospect his notion of a Cosmological Constant (which we will refer to by the arabic letter L; in equations containing it, the Greek letter Lambda Λ is used) once more has merit but at the time it was first conceived did not satisfactorily explain a Universe whose expansion characteristics were still poorly known. Like Einstein's Cosmological Constant, this arcane repulsive energy is postulated to be real largely on the basis of necessity - the need to account for the acceleration if the above Supernovae observations are sustained. The time at which this acceleration began to act is currently unknown with precision; it may be a vestige of the original Inflation or may be a later (second) inflation whose cause for inception is still a mystery.

So, it now appears that Einstein's Cosmological Constant has been rehabilitated, but used in a different way. As Einstein conceived of its use (in his static Universe), it was a term added to his field equation for relating his geometry of spacetime to the distribution of energy and matter:

$$G_{\mu\nu} = 8\pi G T_{\mu\nu}$$

where **G** is Newton's Gravitational Constant, **T** is a Stress-Energy tensor, and **μ** and **ν** are mass and energy terms. When his Cosmological Constant **Λ** is added to this fundamental equation, and multiplied by $g_{\mu\nu}$, the spacetime metric tensor, the above equation now incorporates this repulsive force which he decided was needed to counter the collapsing effect of gravity:

$$G_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}$$

The Einstein Cosmological Constant has been revised to become a new term: $\rho_{vac} g_{\mu\nu}$, the energy density of the Universe (the vacuum energy described on this page), that enters the equation on the right side:

$$G_{\mu\nu} = 8\pi G T_{\mu\nu} + \rho_{vac} g_{\mu\nu}$$

The reason for placing the modified ρ to the right is that it now is tied into the energy side of the equation rather than the spacetime term on the left. So, Einstein's Cosmological Constant has changed in its makeup and placement but still represents a force working to overcome gravitational effects.

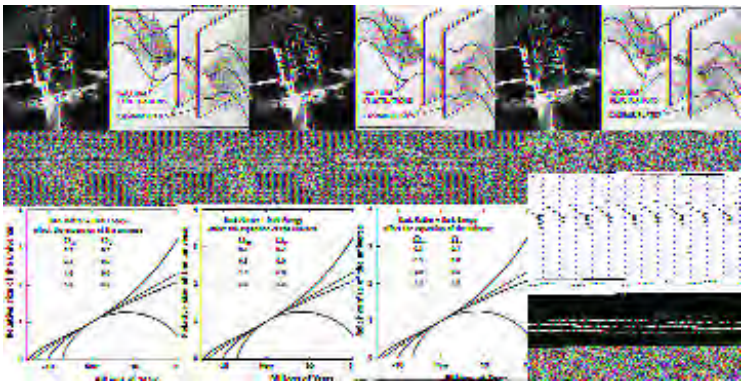
A positive Λ (L) value generates long term repulsive forces that act as though they produce a *negative pressure* (-P). Another way to look at this is to refer to it as "antigravity". This negative acceleration means that the (Open) Universe will continue to expand forever without ultimately slowing down. L is associated with another parameter, Ω (Greek letter Omega). Omega is the ratio of the density of matter/energy in the Universe to the amount actually needed to produce a Flat Universe (Ω = 1); it is also the ratio of gravitational energy to the kinetic energy of all

matter/energy extant. Although a Flat Universe seemingly fits many observations (but may be illusory in that we are only observing a perhaps small part of an infinite Universe which appears to be flat - like a tiny area on a large sphere or on a hyperbolic saddle), when an inventory of all matter and conventional energy throughout the Universe is estimated, the amounts fall far short of that needed to achieve the true flat state. In fact, the value of Omega, as suggested from Supernova observations, seems to be less than 1, which corresponds to expansion following hyperbolic geometery. (L [Λ] itself is finite in this model in contradistinction to the value in the Table on page 20-9). If this holds up, it becomes necessary to account for this lower Omega, since matter/energy now identified falls way short of the required amounts. Thus enters the present day equivalent of Einstein's L, that is, some form of energy whose characteristics are only crudely known and existence is yet to be proved.

This is the energy of empty space (the so-called *vacuum*), a seeming oxymoron in that space is then not really empty if it contains energy; see this footnote (*) for comments on the types of vacuums and voids referred to in Cosmology. The type believed to account for all possibilities is the "false vacuum", which is characterized by some form(s) of energy. Quantum theory holds that zero energy is not possible in any part of space, even after electromagnetic, thermal, nuclear, and other forms of energy are completely removed. The energy content is distributed within a constant vacuum density (new energy is just balanced by increased spatial volume) and this energy acts to offset gravity. Quantum theory suggests several of its characteristics, which are the subject of the research field of *quantum cosmology* and its subset, *cosmongony*, (the study of cosmic origins and history). This virtual **vacuum fluctuation energy** is a (maybe "the") form of Dark Energy, i.e., neither emitting or absorbing light - which means that it does not give off any radiation that can be detected. The energy is inert in that it doesn't react per se to change matter even though it influences it by causing repulsion. **Vacuum energy (also called Zero-Point Energy) is brought into play by exceedingly brief quantum fluctuations that create virtual particles and corresponding anti-particles that develop momentarily - and seemingly spontaneously - in the apparent vacuum of space and then mutually annihilate, releasing energy in a form that drives the acceleration by a repulsive force.** This phenomenon, in which particles seem to be created from "nothing" (the Aristotelian concept of 'potency' - something has the potential of coming into existence - seems a good analog drawn from metaphysical thinking.), can be conceptually explained from principles of quantum physics.

The basis for the conclusion that the vacuum of the Universe is not truly empty of everything is the Heisenberg Uncertainty Principle. Thus, consider any finite portion of this vacuum; it can be totally devoid of any form of matter. But it contains at any instant of time the above Vacuum Energy, in what theory says is a very large quantity (the actual amounts are presently unknown). The Uncertainty Principle requires that statistically the probability is not zero that from this energy a particle-(antiparticle) pair can emerge at any instant and will likely last momentarily before annihilation. Thus, this vacuum unit contains virtual energy that has a finite potential to transform into a virtual particle. But, only rarely does a particle or antiparticle survive. It is still speculative, but with growing favor, that the end product of such a virtual fluctuation appearance is the form of energy now included in the concept of Dark Energy (it may comprise the entire amount of D.E. or other forms may also be present). Based on its influence on matter, particularly in the galaxies, best estimates of the quantity of Dark Energy in the vacuum are around 10^{-26} kilograms per cubic meter, equivalent to that available in only a few hundred Hydrogen atoms.

That virtual particles exist has been verified by the Casimir effect. Consider this experimental setup:



A pair of metal Casimir plates are place in close-spaced proximity. The entire apparatus is enclosed in a near-perfect vacuum. Virtual particles are assumed to pop in and out of the enclosed space. These have characteristic wavelengths. Those between the plates cannot function or propagate because the plate separation is less than the wavelengths of the particles. Those particles outside are not thusly restricted so that they can strike the outer plates producing a weak but measurable pressure. This results in an imbalance of forces that causes the plates to push together. Since no other sources of this attractive force can be accounted for, the conclusion is a verification that virtual particles were present and therefore exist.

As was stated on page 20-1, these virtual particles (quantum fluctuations) are good candidates for the condition that led to the Big Bang. In this concept, a virtual particle of near-infinite density became unstable and "exploded" into the Big Bang event. A thought experiment leads to this notion of two alternatives: 1) there was a pre-Big Bang "void" (see page 20-1 for definition) that contained countless numbers of virtual particles, many being stable indefinitely, some undergoing changes that did not involve Big Bangs, and one (or more, the Multiverse hypothesis, covered later in this Section) that underwent a Big Bang that produced the Universe humans inhabit, or 2) there was a truly empty void (no space or time characteristics) in which a unique virtual particle, the singularity, was "created" and then "detonated" into the present Universe. That Universe contains the countless numbers of virtual particles that come and go without any being involved in another Big Bang (at least so far). Quantum physics allows for the existence of virtual particles in either of these two cases, but does not clearly explain how and why they came to be. (Even metaphysics fails to provide an adequate explanation.)

This much is clear: Near-empty space is continually invaded by myriads of these individual (virtual) particles that have only fleeting existence but the process continues constantly as long as space exists and releases vast quantities of the above-mentioned, as yet undetermined, form of energy that powers the acceleration. The value of L that influences this process may be constant (or could vary - as yet undetermined). In the early days of the Universe matter/energy density was very high but has continued to diminish with expansion. A few billion years ago, its value dropped below the (constant?) energy density associated with L, so that now its repulsive force is taking command and is causing expansion to speed up, an effect implied by the fainter Supernovae observations.

The following four paragraphs attributed to D.A. Stenger (1996) have been taken off the Internet and added here as supplementary information:

In General Relativity, spacetime can be empty of matter or radiation and still contain energy stored in its curvature. Uncaused, random quantum fluctuations in a flat, empty, featureless spacetime can produce local regions with positive or negative curvature. This is called the "spacetime foam" and the regions are called "bubbles of false vacuum" (these describe the infinitesimal volume of space at the outset of a Big Bang). Wherever the curvature is positive a bubble of false vacuum will, according to Einstein's equations, exponentially inflate. In 10^{-42} seconds the bubble will expand to the size of a proton and the energy within will be sufficient to

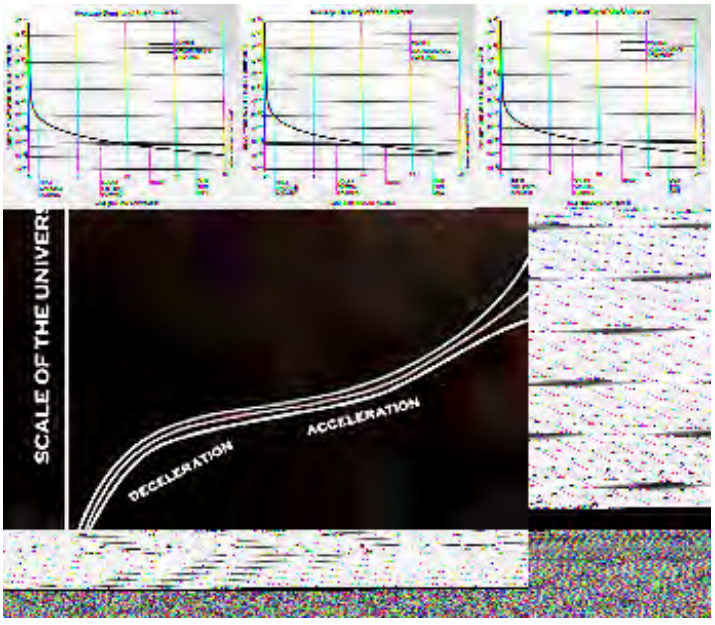
produce all the mass of the universe.

The bubbles start out with no matter, radiation, or force fields and maximum entropy. They contain energy in their curvature, and so are a "false vacuum." As they expand, the energy within increases exponentially. This does not violate energy conservation since the false vacuum has a negative pressure so the expanding bubble does work on itself.

The forces and particles that appear are more-or-less random, governed only by symmetry principles (like the conservation principles of energy and momentum) that are also not the product of design but exactly what one has in the absence of design.

As the bubble universe expands, a kind of friction occurs in which energy is converted into particles. The temperature then drops and a series of spontaneous symmetry breaking processes occurs, as in a magnet cooled below the Curie point and a essentially random structure of the particles and forces appears. Inflation stops and we move into the more familiar Big Bang.

This concept of the Universe experiencing ever decreasing density (the result of enlarging volume) of Ordinary Matter/Dark Matter with time while the repulsive energy (that part of Dark Energy equivalent to the Cosmological Constant or Quintessence [see below], or some similar form yet to be defined and quantified) is increasing, is presented in this simple generalized plot (Note: it is based on a Universe age of around 12 billion years [too young by about 1.7 b.y,]). The important feature is the notion of the **two curves crossing sometime in the past with the Cosmological Constant energy now in greater amount than the matter curve**.

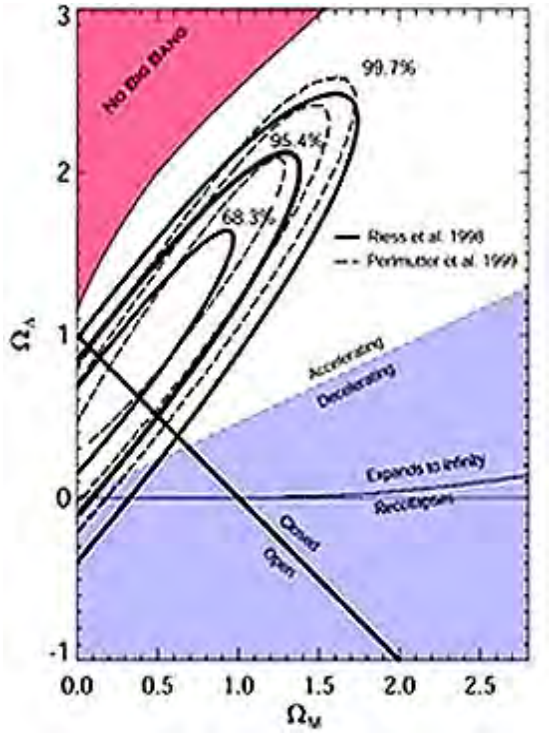


Verifying the existence of **Dark Energy** in space itself and determining its nature (properties) and origin is now near the top of the list of priorities for further observations and theoretical explanations by cosmologists and astronomers. Dark Energy may be the most fundamental and extensive of all physical components of the Universe. As shown on page 20-9, it may be, in reality, the very ingredient that makes up the estimated 73% of the total mass/energy of the Universe. It, as has been shown, determines the expansion rate of the Universe and thus its ultimate fate. Despite this importance, its detection so far has proved elusive but future experiments offer promise for its belated "verification", followed by plausible interpretation(s) of its significance.

We can restate a few logical inferences about Dark Energy that may eventually be substantiated, modified, or discarded. First, Dark Energy is presumably the most fundamental physical "thing" in our Universe, and perhaps in the "void" that exists beyond this Universe's ever

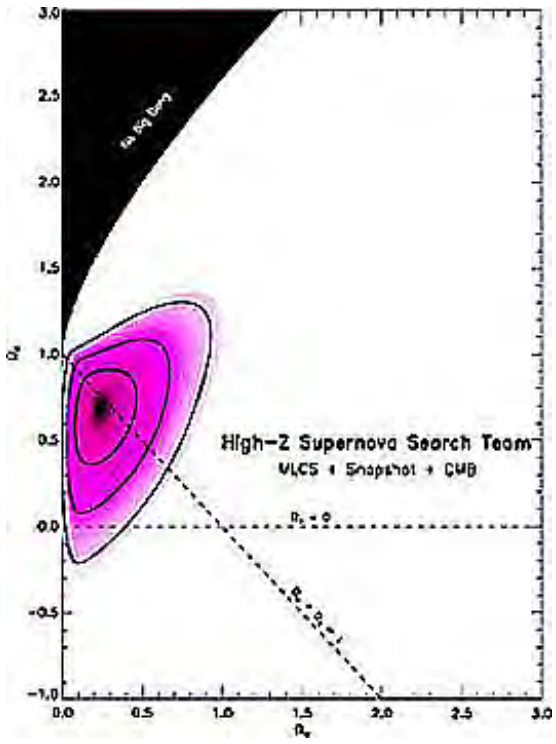
expanding outer boundary. Second, following the Einstein equivalency equation $E = mc^2$, Dark Energy can itself be converted into Dark Matter under the proper circumstances, and this may be reversible. Third, Ordinary Matter bears a still unclarified relation with Dark Matter/Energy but after the Big Bang Baryons and other forms of Ordinary Matter came into existence coincidentally with the Dark Matter/Energy that pervades the Universe. Fourth, as covered in various paragraphs on this page, the nature of Dark Energy is still undecided; it may be best connected with ideas of vaccum energy, with Inflation, with Quintessence, with negative pressure, or a combination of these or something yet to be conceived. Fifth, Dark Energy seems to have played a key role in forming galaxies and affecting their evolution and dispersal.

The key point made in this paper - and the heart of the new ideas emanating from the evidence for an accelerating Universe - is that there seems to be some preferred value for the Ωs of Dark Energy (L or Λ) and the other forms of matter and energy. As discussed earlier and above, there are various estimates of these Ω values. Here is a plot of Ω_Λ versus Ω_M



Ω_Λ is the ratio of the energy density contained in the cosmological constant to the critical energy density of the universe. Ω_M is the ratio of the energy density contained in matter (Baryonic and Dark) to the critical energy density of the universe. (The critical energy density is that which would yield a spatially flat universe)

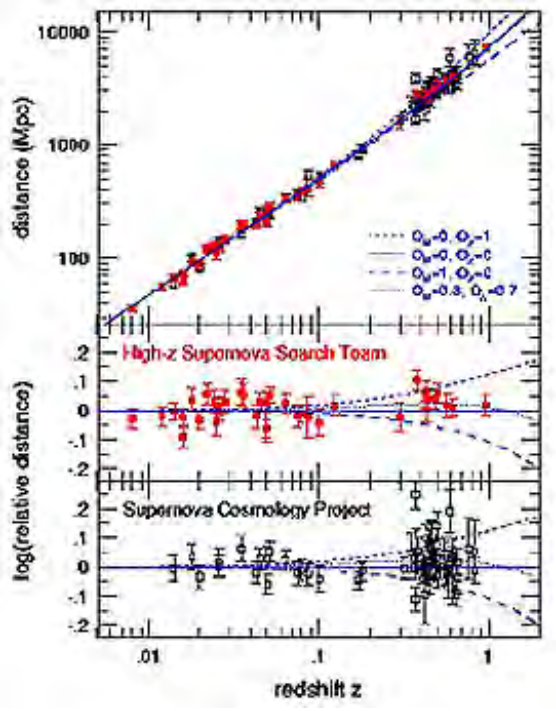
The plot above shows boundary conditions for Open and Closed Universes, as well as for Accelerating and Decelerating Universes. In the white field are contours that represent the confidence limits for the statistical locations of the range of Omega values that can be related to Supernova 1a measurements that fall within the plot. In the plot below, the data collected by the High-z Supernova search team have been used to define a field of most likely Omega combinations. The maximum falls close to $\Omega_\Lambda = 0/7$ and $\Omega_M = 0.3$ values. This is in agreement with the current estimates (as percentages) for Dark Energy and all other forms of energy/matter.



The presumption in this Ω_Λ vs Ω_M plot is: Ω_Λ plus (**+**) $\Omega_M = 1$, which is the case for a Flat Universe, the model now most in favor.

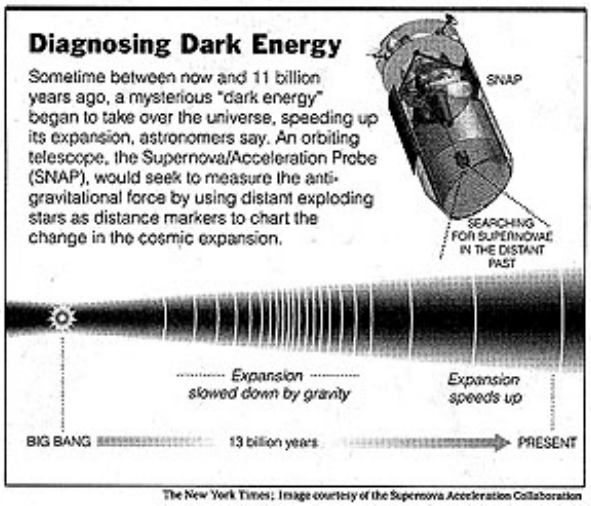
This second paper included an elegant version of the graphs shown above - the Krause diagram. To be readable, it must be entered at a large size, which will slow your downloading. So, we have placed it as an option to select by clicking on page 20-10a. Read the comments on that page which also apply to the above graphs. The notion of a Cyclic Universe is also expounded on the page

Having become familiar with these new models for an accelerating Universe, it is now propitious to show the following plot.



This figure contains data obtained by the High-z Supernova Search Team and the Supernova Cosmology Project. The upper plot shows Supernova redshift values versus distance in Megaparsecs to the SN 1a events monitored as described above. Four different sets of plots based on values for Ω_Λ and Ω_M are evident in the upper right where they diverge slightly from

one another (at lower z values they are superimposed). This narrow spread is hard to see at this scale. The curves diverge when the log of the relative distances forms the ordinate. Some indication of the acceleration trend is evident but the plot points and their statistical spreads are not sufficiently separated to be decisive. More and improved observations, especially farther out in space, will be needed to convincingly confirm the Accelerating Universe model. A future satellite, SNAP (Supernova Acceleration Probe), specifically dedicated to SN 1a data collection, has been proposed. Some information about SNAP is given in this figure:



The third paper, **Inflation in a Low Density Universe** (Bucher and Spergel) presents an alternative to postulating an L-related repulsive energy. It reaches a similar conclusion that the Universe is Open and space is hyperbolic in its pattern of expansion. The model presented, the Open Inflationary Theory, is a variant of the Standard Inflation model of Guth and others. In the standard model, the inflation is related to the **Inflaton Field** (IF), which refers to particles that exist during inflation that result from quantum field oscillations (much like the virtual particles described above). Imagine a curve shaped like a broad, open U within an X-Y plot in which the vertical describes the potential energy changes and the horizontal changes in IF. (Note: In other, similar diagrams, the parameter for the vertical axis is called *energy density* and the horizontal axis is labelled *Higgs Field* (in which the inflaton is equivalent to the [possibly now detected] Higgs Boson).

At the onset of inflation, the IF moves down the curve towards a minimum. This process can take place within the infinite entity that is conceptualized to be everything (a continuum without bounds) within which one to many individual Universes can come into existence. In the Open

Inflationary model, there is a warp in the curve near high potential energies in which the IF can be trapped, as though in a local trough, the so-called "false minimum" or "false vacuum". Many such "troughs" exist - each a potential Universe. When certain quantum processes occur, the IF state may "tunnel" out of this trough and proceed down the regular U surface towards the minimum. Each time this happens, a true inflation occurs and a Big Bang ensues (note: Big Bangs only describe the growth of Universes, not the cause of the conditions that existed before inception). The IF, by its nature, imparts an antigravity force which leads to expansion. This process can occur anywhere within the continuum and not simultaneously, so that numerous universes (multiverses) can form and growth, much like the multitudes of bubbles in water approaching the boiling point.

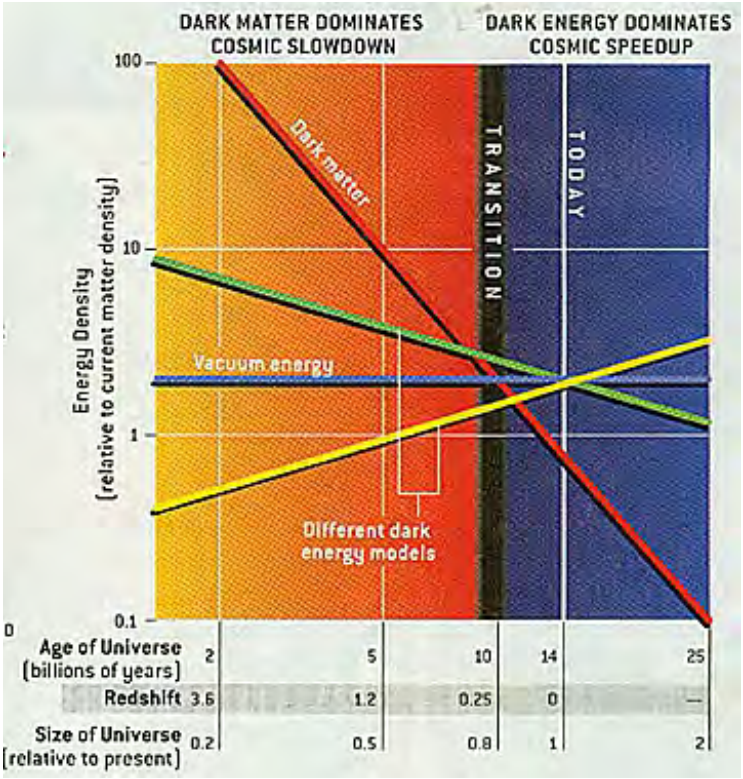
In a January 2001 paper in Scientific American, J.R. Ostriker and P.J. Steinhardt describe *the Quintessential Universe* in which the vacuum energy is modified so that it can interact with matter. It produces a dynamic quantum field, described by their term "quintessence, in which the presence of matter reduces the rate of acceleration below magnitudes predicted for cosmological constant expansion. In the long term future, say 30 to 50 billion years from the present, the quintessential expansion may either continue to increase or may be slowed down and even reversed by the influence of matter, even if that is then dispersed into a much greater volume.

One of the "curiosities" still intact after the proposal in the 1990s of expansion acceleration is that it seems to be relatively recent in the cosmological time scale. In other words, this acceleration has only in the **last 5 billion years** or less (initial estimates of 7 billion years have now been lowered to more recent age) reasserted itself (amounting to a "second coming", following the first abrupt acceleration or Inflation in the early moments of the initial Big Bang), i.e, it has increasingly acted to counter the effects of Dark Matter gravity and may become even more dominant in the future.

One might logically surmise that this repulsive force associated with Dark Energy (whose density is thought to be constant over Universe time) has existed all along but was less than the matter-related gravitational forces when the Universe was smaller during its first half of existence. Now, with further expansion, and greater separation of galaxies and reduced mass density, the repulsive force began about 5 to 7 billion years ago to exceed the attractive force (gravity) leading to a renewal of acceleration such that the pushing apart overtakes the earlier slowing of expansion when gravity exerted more effect on the Universe's growth. Although not yet proved, it seems that this repulsive force from Dark Energy will become increasingly stronger relative to gravity so that acceleration of expansion will be ever greater (this implies that the galaxies will eventually become very far apart and increasingly dispersed).

Another consequence of this incremental expansion is that it causes the age of the Universe to increase insofar as it is determined from the Hubble constant. This offsets the anomaly first reported by investigators using the HST to refine the value of H_O in which their best estimate at first was 12 billion years - a value contradicted by independent estimates of the ages of the farthest galaxies seen so far as 13-14 or more billion years and by the ages of stars, some of which give evidence of being older than 12 billion l.y. But, recalculating Hubble ages with the acceleration parameter included yields ages from 14-15 billion years, which removes the discrepancy inherant to the star age anomaly.

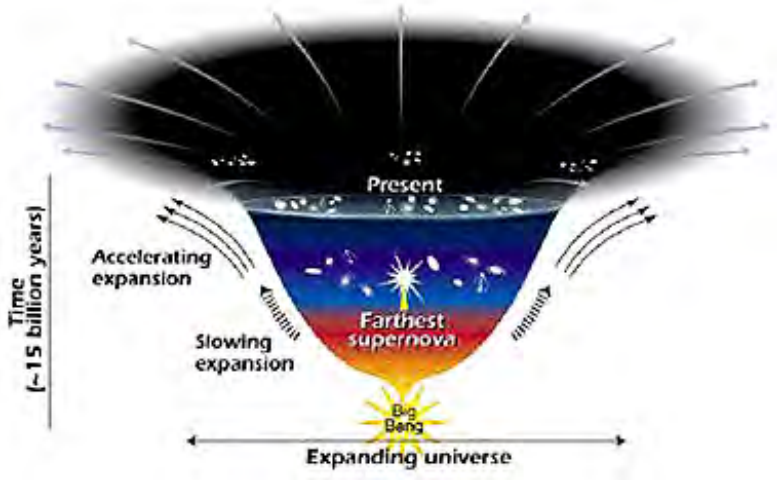
As the consequences of an inflation-deceleration-acceleration Universe gained more widespread favor, workers considering this reached the conclusion that the two principal components of the Universe - Dark Matter and Dark Energy - were the principal players in the observed expansion history. Thus the small amount of Ordinary Matter (~4%) plays an insignficant (but essential) role. This diagram summarizes the current thinking on this "new" Universe:



From Riess and Turner, Scientific American, February, 2004

Thus, the decreasing Dark Matter (red) and increasing Dark Energy (yellow) curves cross about 4 billion years ago, at a redshift value when the Universe was about 4/5ths its present size. Thereafter (to the right) the Dark Energy (in terms of Energy Density) has continued to increase relative to decreasing Dark Matter (again, relative to Energy Density since it is likely that the equivalence of matter and energy [in the Einstein sense] noted for ordinary energy and matter holds for the always dominant Dark forms). .

To sum these preceding paragraphs, recent evidence now suggests an Open Universe that tends towards the Flat type at one observational scale but may in fact be expanding infinitely in the hyperbolic space mode when envisioned at a greater scale. Both special forms of Inflation and the possible existence of a great quantity of repulsive energy may be involved. Unless huge amounts of matter/energy, having sufficient attractive power to more than counterbalance the Dark Matter expansion, are discovered in the future, the Open model - with space increasing infinitely - is most likely to be the favored scheme. If the constituents of the Universe survive forever in some form (see below), then as they move ever outward, they will always be creating new space (that which holds particles) out of the Void, in perpetuity. One consequence of the Accelerating Universe model is that it produces a density that almost exactly matches that predicted from the earlier Inflation mode. Track down the January 1999 and January 2001 issues of *Scientific American* to improve your insights and understanding of these complex ideas. Also, consider this general diagram that summarizes some of the above ideas.



Espousals of competing hypotheses such as above are a hallmark of scientific inquiry. As new evidence rolls in and specialists re-interpret existing evidence, major challenges to the prevailing theories arise; many die away but sometimes older ideas are abandoned and strikingly new concepts emerge. This has been especially true for Cosmology. With the launch of powerful, versatile new space observatories and much better ground-based telescopes, the wealth of new data over the last 20 or so years has prompted refinements of earlier models and explanations or their overthrow and advancement of better substantiated ones. This is true for Dark Energy. As of the middle of 2003, the notion of Dark Energy seemed plausible. Then, in December of 2003 a group of astronomers from ESA's Space Research and Technology Centre presented results of their interpretation of data from XMM-Newton, an X-ray observing spacecraft. The gist of their argument is based on studies of a group of galaxies about 10 billion light years from Earth. These appear to be giving out much more X-radiation than galaxies do today. If this is verified by further observations, the implication is that the early Universe was much denser than hitherto surmised. Also, their interpretation is that there were fewer galaxies then, and that all galaxies seem to continue growing by conversion of Dark Matter into Ordinary Matter. The higher densities during the first part of cosmological history run counter to the "concordance" model that requires decreasing Dark Energy and a lower average density. Instead, much of that fundamental component of the Universe could have been Dark Matter, one consequence being that Dark Energy had "condensed". However, this new hypothesis so far does not explain the apparent acceleration of the present Universe.

These new interpretations (and speculations) described above have also turned attention to the nature and role of gravity in the expansion/deceleration-acceleration history of the Universe. One of the aforementioned articles in the February 2004 issue of Scientific American, "Out of the Darkness" by Geori Dvali, reviews the latest thinking on gravity and its involvement in the story of this history. We review the main ideas in the paragraphs below:

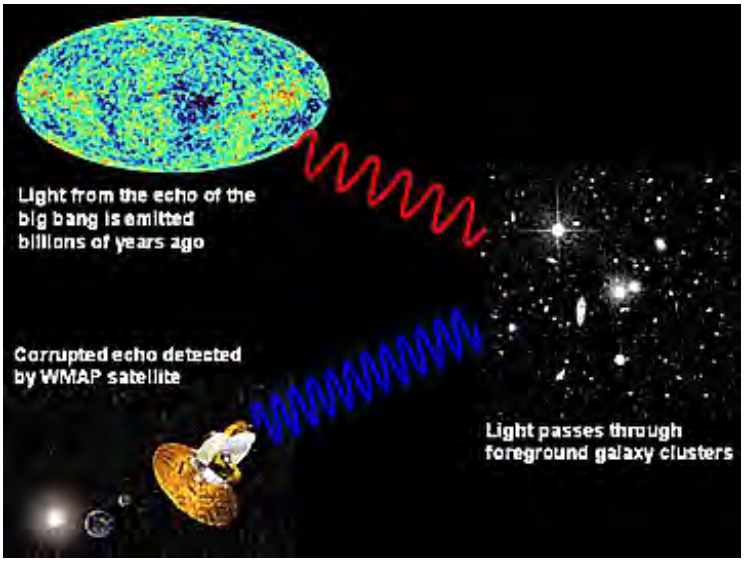
His prime thesis is that there may be two kinds of gravity: the conventional attractive mode (+) with which we are familiar and a repulsive form (-), possibly being the explanation for the Cosmological Constant or the Quintessence modes described above, but in Dvali's proposals of a different nature described in this and the next paragraph. The argument is made in the article that the gravity we observe at Earth and even galaxy scales is of the + variety whereas at much larger scales the - variety takes over. Thus gravity does not behave the same when the entire Universe is considered as a unit. This repulsive gravity can be deduced, and its properties predicted, when matter and energy are examined at quantum scales - this is the so-called quantum gravity concept that many physicists feel will tie all particles and forces together in the Theory of Everything looked at several times in this Section. This implies that the standard Laws of Physics break down, or function differently, at the largest scales.

Dvaili draws upon Superstring Theory and the concept of the (Mem)Brane we were introduced to on page 20-1 to account for the dual behavior of gravity. He examines the possibility that the Graviton, which provides the force particle that accounts for attractive gravity as it flows between particles, is capable of moving in various ways within and into/out of the Brane (see

pages 20-1a and 20-1b). Gravity within the Brane (a flat but three-dimensional Universe) operates over galactic scales as an attracting force and follows the inverse square law. In four dimensions, gravity follows an inverse cube law whose effects do not reduce its ability to hold material and energy together when galaxies or galaxy groups define the scale. But when the postulation of multiple dimensions beyond four is introduced, these extra dimensions can accommodate the participation of gravity in a way that allows its to induce a repulsive force when the large scale Universe is considered.

Gravity becomes weakened when the Gravitons are able to move outside the Brane (the flat Universe). They escape into the extra dimensions, which can be of more than one proposed type: 1) curling (into closed loops) within the scales near quantum levels (Standard model); 2) infinite and curled (Randall-Sundrum model); 3) infinite and straight (Dvali model). Gravitons, according to this hypothesis, are the only particles capable of "leaking" out of Brane space into truly empty space beyond. Those (less energetic) Gravitons that have low momentum (yielding long wavelengths) are the ones now escaping (some can return to the Brane but the net effect is a loss). This loss, on a grand scale, is reducing the overall ability of Gravitons to decelerate the Universe, and at some time in the past the influence of the overall attractive effect of gravity, now weakened as the net strength of the supply of Gravitons is diminished, has affected the Universe's behavior according to Einstein's General Relativity in the sense that the energy-density of the Universe modifies space curvature. The changing curvature is expressed as a geometry that gradually favors acceleration overcoming the restraining force of conventional gravity, which becomes unable to pull back on both matter and energy (dark and luminous). The Dvali mechanism obviates the necessity to presume the repulsive force implied by either the Cosmological Constant or Quintessence; there are in fact several other models that lead to acceleration (see Dvali's article).

There are astronomers who have questioned many of these ideas concerning Dark Matter, Dark Energy, renewed expansion, etc. As is so true of all Science, each new idea begets several variations or even alternatives. Here is an excellent example: Extremely small but significant variations in the Cosmic Background Radiation temperatures are cited as evidence for these invisible (and still undetected) forms of matter/energy. Their distribution is said to control the regions where galaxies first formed in the early, much smaller Universe. But Dr. Thomas Shanks (University of Durham) and colleagues have studied this distribution in terms of where major galactic clusters occur in the Cosmos and find a positive correlation. He cites a mechanism known as the *Inverse Compton Scattering Effect* in which the hot gases within the clusters will energize CBR. This has the effect of shortening the wavelengths of the microwave radiation, as shown in this diagram:



The implication is that the anomalous irregularities in the CBR are actually related to modifications imposed by interaction with galactic gases as faraway radiation passes through high concentrations of gases in galactic clusters.

We close this provocative subsection with several ideas presented in a 2006 paper by Biermann and Kusenko in *Physical Review Letters* which is concerned mainly with Dark Matter. They point out that the first hints of such Matter are traceable to the 1930s as an explanation of gravitational effects on Ordinary Matter. This latter possesses insufficient mass to provide the needed forces to hold galaxies and the entire Universe intact. Dark Matter seemingly possesses no electric charge and does not react to the strong and weak forces involved in binding atomic nuclei into the atoms of Ordinary Matter. What is the true nature of Dark Matter remains a mystery but Biermann and Kusenko speculate it is largely (perhaps entirely) composed of sterile neutrinos (whose existence is postulated from Supernova studies), which differ from ordinary neutrinos in not being influenced by the fundamental weak force. And, the proof that neutrinos have a tiny but real mass, plus the fact that they are so abundant, further points to their ability to influence gravity. If Biermann and Kusenko's idea proves correct, or at least is part of the story, then neutrinos might lay claim to being the most *fundamental* of the known (and hypothesized) particles. The writer (NMS) speculates that these neutrinos in some way also participate in Dark Energy, as such being in a form that is a logical extension of Einstein's argument for the interconvertibility of matter and energy. It is ironic that the most abundant category postulated as making up the physical Universe - the 74% that is Dark Energy - has to date proved so elusive and intractable to direct sensing that cosmologists have yet to confirm its nature and, even more rudimentarily, its very existence. That it ought to exist is deduced from models in which Dark Energy's presence helps to explain the observations accruing from the spurt of knowledge that is driving modern Cosmological thought.

The paragraphs above may seem a bit rambling, with "uncertainty" being the watchword. Is there anything really firm that can be said about Dark Energy? A press release on January 18, 2007 about research done by a Danish group of astronomers does seem to clarify several key points. This group has been analyzing Supernova distance/rate of expansion data, as have others. The group now takes a position that, if valid, does enlighten. First, they confirm the expansion rates being reported. Second, they find that there is no need for "fancy theories", that the rate predicted by Einstein's Cosmological Constant is consistent with recent observations. Third, Einstein's notion of Energy of a Vacuum is probably closest to the truth as of now. But, of course Vacuum energy seems an oxymoron. Unless one assumes that there is no true (total) vacuum, such that no part of space is actually "nothing". But what is the source of Vacuum energy remains a mystery for the moment (NMS: virtual particles, but what are they?). Stay tuned!

Multiple Universes

After these next paragraphs were written, the front cover of the May 2003 issue of Scientific American was enblazened with the title "Infinite Earths in Parallel Universes Really Exist" by Dr. Max Tegmark of the University of Pennsylvania, which is the lead article inside. It is a most profound summary of present-day thinking about whether many (infinite numbers may be possible!) Universes (Multiverses) in infinite space could actually ***exist****. The writer (NMS) has now read this mind-blowing ("Excedrin headache" level) article three times and I am just beginning to fathom new insights. For those reading this page, this publication is a challenge I toss to you to read at least once. Multiverses are also tied into the concept of Membrane (Branes) theory (M-theory), which will not be discussed in this subsection because of its complexity; nor will the implications of 11-dimensions be explored.*

From quantum theory and other considerations, there may be a large number (conceivably an infinite number) of individual Universes, also called *parallel Universes*, none of which we can at present be aware of since they lie beyond the outer limits of our Universe - the edge of which we have yet to make contact with (the Cosmic Horizon). In some models, most, if not all, such "bubble Universes" never touch even as nearby ones expand but if two interact, they can experience tremendous energy effects; other models find ways to either eventually interact

physically or at least contact in some way. Each Universe is considered to form like a bubble in boiling liquid, with the bubble expanding. Many bubbles form at different times and grow until dissipated; the analogy is limited in the sense that, unlike bubbles in boiling water in a teapot, there is not confining structure (the teapot) but the water is infinite, the bubbles form and disappear infinitely, and the water itself is really just empty "void" (see below).

Some (most?) Universes do not survive Inflation to expand as does the Earth's Universe. During the first stage of Inflation protogalaxies in a given Universe are "in touch", but at speeds of Inflation greater than light speed, they may lose contact. With the end of Inflation, some galaxies re-establish contact but there are today parts of our Universe that are too far away from each other (in opposing directions from Earth) to have received light from one another in the time elapsed since the Big Bang. In this bubble model, Omega begins at zero (0), then after leaving the false vacuum trough it rises to 1 during full Inflation (yields a flat Universe) and then decays to values less than 1, giving rise to hyperbolic expansion (post-Inflation), one outgrowth of which is the acceleration noted as expansion is traced back in time. This acceleration is a consequence of the IF (Inflation Force) being maximum at the bubble boundary, diminishing inward towards the center. Needless to say, this Open Inflationary model is speculative and in the future may not stand up when tested by further observations and calculations but it does offer a way out of having to depend on the bizarre Cosmological Constant energy to account for acceleration.

The precise Universe of which we are aware that has emerged is "the one we've got". Most of its controlling properties are well known, although numbers describing some, e.g., total mass, remain to be fixed in detail. Other Universes, making up a collection of independent Universes existing in what is called the **Multiverse** (Multiple Universes) with different properties may have been possible. This term encompasses all such theoretical (but possible in the light of quantum Cosmology) parallel Universes. If certain fundamental properties and constants were to be only moderately different than the ones now identified and quantified, the nature and history of these alternate Universes could prove difficult to observe. And, in fact, significant changes in certain prime parameters from the values they actually have might even have denied the Universe a successful existence.

For our Universe, more than 30 such parameters (see previous page) have been cited as critical variables whose values can tolerate little or no significant differences from what they are. Among these are: the four Universal Constants (speed of light; gravitational constant; Planck's constant; Boltzmann's constant); constants associated with the four fundamental forces (Gravitational; Strong; Weak; Electromagnetic); rate of expansion; particle/antiparticle interactions; neutron/proton abundance ratios; H/He and H/Deuterium ratios; balance between nuclear and electron forces; temperature and density variations during the first minute; total number and density of neutrinos; mechanisms of star and galaxy formation; element synthesis, and others.

Our Universe seems to follow the previously mentioned Goldilocks dictum: not too hot, not too cold, just right. If chance alone (quantum control) were the governing determinant, the odds are enormously against all of the above factors - those that allow our Universe to work smoothly in its large scale operations - being just right. Most ordinary people, and some scientists, believe that there must be a superintelligence, be it called God, the Intelligent Designer or some other analogous name or concept - had to play the pivotal role in establishing and structuring a scientific Universe. However, the Evolutionary Principle of Natural Selection has been proposed as *sufficient* to explain the origin and development of the Universe. (The essence of Darwinian evolution applicable to living things is carried over into the changes applicable to inanimate things.)

Thus, the conditions that operate in our Universe are "fine-tuned". None can vary by too much from their narrow, critical specific values or the Universe would have expanded too fast or too slow, or might not ever have organized into the present assemblage of matter. The fact that it did is remarkable, but had things been much different, the conditions favoring life would not have occurred, presumably at least not on the grand scale that is represented not only on Earth but in

countless planets in innumerable galaxies. This situation lies at the heart of the **Anthropic Principle**. The so-called weak version of this principle includes the idea that since intelligent beings (us, at the least) indeed exist, the governing conditions for our Universe must have been been such as to permit life to develop or, put another way, because we are here contemplating our existence, conditions (the scientific ones that we now have verified) had to be right for our Universe to succeed to the stage of evolving knowing creatures (note the similarity to the question "If no one is in the forest to hear when a tree fell, did it really make a sound"). The strong version states that life has depended on the right combination of physical constants, parameters, and conditions to allow life to begin and flourish and that these were pre-ordained by some independent "mind" such as is attributable to a Creator.

Now, let us expand the challenging, mind-boggling thesis, mentioned earlier on this page, that ours may not be the only Universe. The key word is "Multiverse" which connotes the astounding notion that, theoretically, quantum cosmology allows for the possibility that many Universes may actually exist. Multiverse can have two connotations: First, there could be multiple Universes existing simultaneously within a void continuum, but these are not in contact so that their existence cannot (presently) be confirmed by any known methodology. Or, second, there is only one Universe at the moment but it is just one of many (infinite) that have "come and gone"; it started with a Big Bang but eventually collapses on itself and restarts with another Bang; this is the Cyclic Universe model.

We have already alluded to the fact that scientific exploration is limited by the speed of light to detecting only those boundaries of our own Universe out to distances in the expanding space which contain galaxies, etc. that have transmitted radiation to us within the time frame of the Big Bang, i.e., are no older than the apparent age of the Universe (best estimate, ~13.7 b.y.). Paradoxically, there likely are parts of this Universe that are presently beyond this *time horizon* and won't be detected until later (for example, in a billion years astronomers will then see that part of the Universe that involves this additional time interval).

If the Universe began as many cosmologists now believe - from a sudden, extremely favorable and positive fluctuation of a quantum state in the so-called "false vacuum" in which a virtual particle instantaneously sprang into existence - containing within it all the potential energy needed to produce the matter and radiation now making up the Universe - then there is a real theoretical likelihood that the process could be repeated again and again, many times (perhaps infinitely). Thus, Universes could be springing forth from the endless (infinite in both space and time) false vacuum that nevertheless contains the "quantum capability" of energy fluctuations bringing individual Universes into existence. Each one establishes its own space which enlarges after its own time-clock is started. Because of the horizon limitations, none of the Universes are (can be) in contact with any of the others. In fact, they all may be drawing apart from one another as the Multiverse continuum itself expands at (or possibly greater than) the speed of light. They all exist independently from one another. We have no known way of communicating with other Universes (for that matter, we are very limited in communications within our own Universe) and, unless some future "breakthrough" in theory and practice that facilitates this communication happens, we will never advance beyond developing plausible models that hint that multi-universes are the norm. To add to the enormity of these concepts, there may be a number (perhaps to infinity) of individual Multiverses which in turn never contact each other. (All this is easy to state verbally but next to impossible to prove - at least for now.) More information about this aspect of multiverses, as the theory of chaotic inflation applies to it, is found at this Wikipedia web site.

Suppose that there are one or more Multiverses, each expanding and creating its own space as it enlarges. But, what then can be said about the "void" between non-touching individual multiverses? Could it qualify as a form of intermultiverse space? This would seem to be reasonable as just being a matter of scale: Since we treat intergalactic space as part of our Universe's space (the galaxies serving as reference points), so perhaps could we consider intermultiverse space to be analogous. Then, all of "superspace" would be expanding, accounting for why we will never be able to find and communicate between "verses". The fancy

word now fashionable for that "void" into which each Universe enters and expands is **Hyperspace** (the term can have alternate meanings; but it usually connotes the notion of higher dimensions of space beyond the customary 4-dimensions). There could be an infinity of such Multiverses and no real boundary to superspace. Between each Multiverse, the "void" would not be a true vacuum but a "false" vacuum containing the potential (perhaps by means of virtual particle fluctuations) for another Multiverse to flash into existence. Multiverses are a provocative idea but frustrating in that there appears to be no scientific way of ever proving their existence. A Multiverse itself - a collection of parallel Universes - may be dimensionally infinite.

For more about Multiverses, refer to this Wikipedia review on the Internet. A more theoretical treatment on the Net examine the Many worlds. Parallel Universes are a popular topic in both scientific and non-scientific fiction. Somewhat more philosophical treatments are on the Web at sites hosted by Discover Magazine and Paul Davies. For those who like to read books, these three by Dr. Michio Kaku are recommended: Hyperspace; Parallel Worlds; Beyond Einstein - these also cover string theory, Kaku's specialty.

The idea of a single Universe that repeats itself - the Cyclic Universe - has been around for at least 80 years. Albert Einstein was once of the first to espouse this, based on his reasoning at the time that gravity would eventually stop expansion, draw all matter together again in a new singularity (the Big Crunch), and reinitiate a new Big Bang; this could happen many times into endless eternity. The concept was abandoned after others showed his idea violated thermodynamics (specifically, there were entropy problems). Since then, other proposals for repetitive Universes have been put forth. They are too detailed and complicated to consider in this Tutorial but several are described at this Wikipedia website.

One of the current champions of Cyclic Universes is Sir Roger Penrose of Great Britain. In his model the beginning of each Universe is the customary Big Bang. Expansion continues for many billions of years. Matter cools, draws farther apart, and is sucked into Black Holes. Over time the Black Holes themselves evaporate into radiation so that the final products are photons. The net effect is that mass is destroyed. One aspect of massless particles is that they have to travel at light speed. From the particles viewpoint time seems to stand still and space contracts to nothingness. To these particles the Universe would appear as though infinitely small. An infinitely small Universe acts as though it were a singularity. Like the traditional Big Bang singularity this one will explode, intiating a new Universe which repeats the process. However, some remnant of the previous Universe(s) is found within highly uniform regions dispersed in the Cosmic Background Radiation; discovery of these regions (claimed by Penrose and colleague Vahe Gurzadyan) is confirmatory evidence. There is no limit to the number of times this scenario happens. Needless to say, most cosmologists are skeptical about this model.

After having read these last few paragraphs, some of you may think that the Multiverse idea, in either of its modes, is a preposterous, wild speculation of overly imaginative minds, in effect an excursion into a dreamworld of unbridled unreality. Isn't one Universe really enough!! But, to the trained scientific mind, the concepts underlying Multiverses and Hyperspace are plausible and scientifically sound. Multiverses may exist and do not violate any fundamental laws or principles of Science. Or they may not exist and Science is confined only to this (our) Universe. At the present time, the nature of Multiverses can only be conceptualized in the mind and through theoretical calculations. Their existence cannot be directly determined experimentally since, in some models at least, we can never contact them (other models invoke collisions between Universes or travel to one another through wormholes). But as the human mind becomes more sophisticated, and new knowledge accrues, attempts to better understand the Multiverse possibilities and to verify them will continue and probably increase. There is a metaphysical side to this speculation: If Multiverses indeed exist, and if they were created by an Intelligent Designer, WHY? Wouldn't one be sufficient? Doesn't multiple Universes tend to argue against such a (wasteful) designer? These ideas are explored on page 20-12 and more fully on page 20-12b.

Dark Flow

Just when the writer (NMS) had thought that enough has been said on this and the preceding page about the concept of "dark" as applied to the Universe, he discovered (on March 17, 2010) on the Internet reference to the latest findings about "Dark Flow", about which he had never heard before. Although still in the early stages of development, this new concept is causing a stir among cosmologists, some of whom remain skeptical about its reality. We will attempt to summarize its essential ideas in the next several paragraphs.

The discovery of Dark Flow was announced in 2008 by a team of astrophysicists, at Goddard Space Flight Center and several universities, led by Dr. Alexander Kashlinsky. They were investigating the behavior of galaxies in a 20 degree segment of the celestial sphere as these move relative to the Cosmic Microwave Background radiation. To their surprise, the motion calculated for about 1500 galaxy clusters located outward from Earth to about 2.5 billion light years away (in the initial study galaxies out to 6 billion l.y. were included) was all in the same direction and at speeds greater than expected. The galaxies "appear" to be moving toward the southern constellations of Hydra, Vela and Centaurus. When first announced, these galaxies were traveling relative to the CMB at speeds up to 2 million mph (more than 3 million kph). Now, that speed has been reduced to "between 1 and 2 million mph". This motion is different from the outward expansion of the universe (which is accelerated by the force called dark energy). The movement is not random as predicted in expansion models that consider large-scale motions, but unidirectional; thus it moves in a preferred direction.

The galaxies are all sources of high energy X-rays. By observing the interaction of the X-rays with the cosmic microwave background (CMB), which is leftover radiation from the Big Bang, scientists can study the movement of clusters. The X-rays scatter photons in the CMB, shifting its temperature in an effect known as the kinematic Sunyaev-Zel'dovich (SZ) effect. The temperature increases in the CMB are small, but consistently greater than the average values. Those temperature anomalies can be used to calculate velocities.

The Kashlinsky group's work shows that these temperature fluctuations are shifted in the direction of a number of galaxy clusters. This implies that the clusters are moving with reference to the rest of the universe, a motion that is therefore independent of its expansion. The result is significant because it cannot be accounted for by the known distribution of matter; something outside of the visible Universe seems to be pulling on the clusters.

The speeds associated with these preferred movements are greater than those calculated for standard expansion models, by small but consistently higher amounts. If this is valid, and confirmed by further observations, the question is why? Three explanations are under consideration. First, the motion could be as as yet undetermined consequence of the action of Dark Energy. Second, it could in some way be an expression of residual motion tied to Inflation. Third, it could be the result of attraction forces (gravity) resulting from huge amounts of matter beyond the observable Universe; this matter resides either in our own Universe or in a second (or more) Universe (if the Multiverse model is real).

These next comments are copied from "PhysicsWorld.com":

According to Kashlinsky, the answer to the source of Dark Flow might be found in the Universe's infancy. "What we believe is happening is that we are measuring the influence of the parts of space–time that were there before the inflation of the Universe."

The theory of inflation suggests that the early universe was like cosmic foam made up of many bubbles. One of these bubbles underwent a huge and rapid expansion, when the Universe was just 10^{-36} seconds old, to become the universe that we see today. It is our neighbouring bubbles that Kashlinsky holds responsible for these migrating galaxy clusters.

"Inflation says that if you go to sufficiently large scales you should see structure that is outside this bubble, outside of our Universe. If we live in such a world these structures would cause this

motion and we suspect that this is what we are seeing," Kashlinky said.

As of early 2010, the notion of Dark Flow remains intriguing but still quite controversial. More measurements, particularly in different parts of the sky, are needed before it gains acceptance. But if it is finally proved beyond doubt, the implications are highly significant, especially the surmise that there may be another Universe.

How Universes Start

The Big Bang remains the favored explanation for the inauguration of our Universe. As stated on page 20-1, the Big Bang could have originated from a moment of instability of a virtual particle. Such particles could have been ubiquitous in Hyperspace. So multiple Big Bangs are plausible.

A variant of these ideas involving Multiverses has Universes related to Black Holes as starting points for new Universes. The question then becomes one of origin of the Holes themselves. To be a Black Hole, a nearly infinite amount of matter must be somehow concentrated in a space that can range from Supergiants through a few tens of kilometers for Holes that are residual to star destruction processes (at the Neutron star stage) to essentially point (dimensionless) sizes resulting from other causes. Our own Universe contains countless Black Holes of varying sizes. Those of point size may be equivalent to singularities and could have arisen by some (as yet unknown) process in the vacuum. Under the right circumstances, a Hole may explode, initiating a new Big Bang.

Another concept is an offspring of the Inflation process. As this early phase proceeds, a segment containing a false vacuum would develop a bubble-like bulge that is connected by a neck-like extension known as a "wormhole". This can break from the parent expansion to drift free as a "Daughter" Universe that undergoes its own development, which includes its own inflation and expansion. There might even be a single "Mother" Universe (probably not ours) from which countless Universes are spawned. This "Super Universe" has always been and serves as the master control for the perpetual series of quantum fluctuations that lead to multiverses. Each remains "out of touch" with other Universes, never interconnecting or physically contacting in the Super Universe within which space only has meaning (to the finite, limited human mind) to the inhabitants of each Daughter Universe. Time in any such offshoot Universe will, in principle, be measured from its explosive conception, but time throughout the super system may be infinite.

As stated above and the previous page, the quantum cosmological idea of myriads of virtual particles emerging from the vacuum has been seized upon as an explanation of the origin of the Universe purely on the basis of a theory that does not require a "Creator" or "Intelligent Designer". In one view, virtual particles normally will encounter virtual antiparticles and will annihilate before any Universe can start. But, statistically (Quantum Physics relies heavily on probabilities) one can argue that on rare occasions, these particles (not matter in the usual sense but some still unknown energy state capable of transforming into the condition associated with a singularity) will not be destroyed but can organize into a singularity, which then reaches a critical state that leads to a Big Bang.

Drawing on some of the above ideas, the Russian cosmologist, Andre Linde, now a professor at Stanford University, has put forth still another model he calls *chaotic inflation* which leads, in principle, to many Universes. It relates, in part, to the idea of Primoridal Chaos mentioned on page page 20-1. He envisions an infinite void with within which the energy associated with the false vacuum prevails. Fluctuations involving virtual particles are continuously occuring throughout. Each is associated with some set of scaler properties which may or may not produce the fundamental parameters needed to originate and control a Big Bang. A huge number of different combinations of these properties/parameters are possible and individual events are likely to utilize conditions that vary considerably. As a result, the vast majority of these fluctuations fail to achieve the appropriate conditions that allow a Big Bang and instead are wiped out by annihilation leading to a zero end result. But, a small fraction produce the

prerequisite set of parameters that favor a successful Big Bang. Of those, some - probably most - will have parameters that cause their Universes to expand too rapidly or too slowly to shape and evolve into ones similar to ours in which life can develop. So, only a few survive.

In mid-September, 2002, Professor Linde and his wife, Professor Renata Kallosh, announced that they have experimented mathematically with their earlier models and have devised still another scenario that requires the observed Universe to eventually cease acceleration and then contract to a Big Crunch, perhaps as soon as the next 10 to 20 billion years. Although Linde has been an advocate of Inflation, he notes that the recent second acceleration models run into conflicts with some of the mathematical aspects of Supersymmetry and Superstrings (discussed on page 20-1). As these two cosmologists explored the consequences of Dark (repulsive) energy, they found reasons from their calculations to consider that this energy, whose values must necessarily be greater than zero, could in time revert to negative, or less than zero, values. If this hypothesis holds up, the acceleration will slow down and convert to deceleration leading to a final collapse. They also point out that we may be seeing only a fraction of our full Universe; parts not seen may offer evidence of this potential contraction in future time. They also consider this new model to allow many Multiverses that end up contracting (crunching) to singularities that re-explode. Thus, the notion of individual multiverses coming and going, with an infinite number of repeats, is one possible consequence of their proposed model.

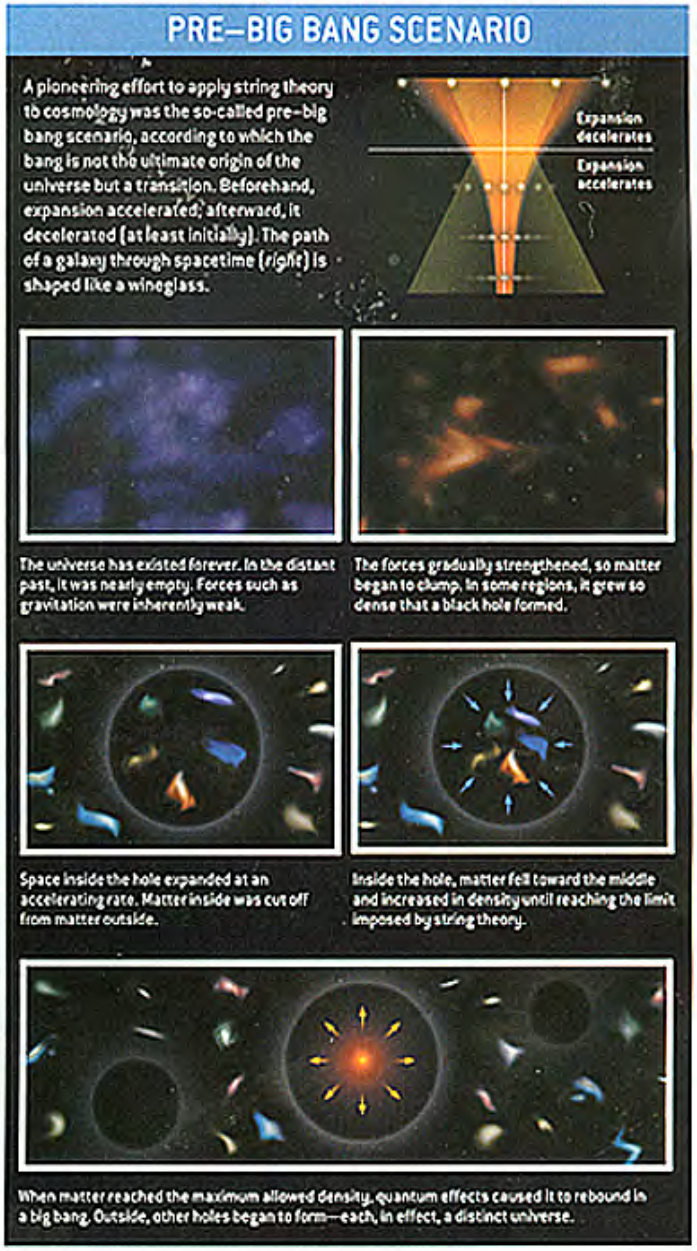
In 2006, another imaginative variant was proposed by Paul Steinhardt of Princeton University and Neil Turok of Cambridge University: its essence is that our Universe expands and then contracts over an average time of 300 billion years; its contraction leads to the crunch that explodes in a Big Bang; this process continues repeatedly so the present Universe is just one of many before and a number (perhaps to infinity) in the future; the governing factor is the evolution of Lambda, Einstein's Cosmological Constant, during the successions. The ideas are intriquing enough to warrant a summary of their hypothesis from the 'Space.com' web site for May 8, 2006. This is reproduced (with style editing) on page 20-10a under the title of "The Cyclic Universe".

Thus, there could be a range of Multiverses (in terms of size and properties) and ultimate fates of each that come and go over time, with one or some that have just the right parameters to form persistent galaxies and stars that synthesize the elements responsible for the variety of features, including organic matter, that facilitate a functional Universe comparable to ours. These surviving Universes are not necessarily identical to each other but show variations in size, composition, structural features and age, and even components without counterpart to ours, which result in different histories (including presence of life). Such Universes come and go (some may even experience Big Crunch multicycling). Suffice to say, the Chaotic Inflation concept, along with other Multiverse models, is presently impossible to test and verify but the Quantum Physics that "allows" these conceptions requires that we seriously consider these alternatives.

A model that has gained many adherents in cosmological circles is that of *Quantum Tunneling*, first proposed by Alex Vilenkin and then added to by collaboration with Andre Linde. In their view, the "nothing" that could have existed prior to appearance of a Universe is truly nothing. In Quantum Physics it is possible for subatomic particles to escape the nucleus or overcome repulsion barriers and then reappear in locations (and states) they do not customarily occupy. Or, on a grander scale, Quantum principles hold there is a finite probability that moving subatomic particles ordinarily stopped by thick physical barriers (such as a concrete wall) can reappear on the other side; the likelihood is small but real, given enough time. This has been called "tunnelling". Extrapolating this to the inception of a Universe, from "empty space", tunnelling permits an entry of a virtual particle across the quantum barrier that prior to that moment prevented its emergence. The Heisenberg Uncertainty Principle states that a particle's position is indeterminant at any instant. But, statistically, it can cross this barrier (perhaps some form of energy), through the tunnel, become a physical entity containing all the energy needed to make a Universe that separates into the negative energy of repulsion (causing expansion) and the positive energy (including all cosmic mass that develops). This can repeat, leading to Multiverses.

Of course, not every Universe would be successful beyond its early moments. Inflation may fail, the fundamental constants may be "off" so that expansion is abnormal, or other conditions could prevent a Universe from experiencing a proper development. Some Universes may be short-lived; others expand so slowly that factors promoting galaxy and star development inhibit attainment of a situation favoring life. But many astronomers and cosmologists now believe that there is a natural inclination - in a sense, a purpose - for Universes to "try" to develop in a manner that points to occurrence of Life as a "goal". A good espousal of this seeming teleological control of Universe formation is the book by Lee Smolin, *The Life of the Universe*, Oxford Press, 1997. Dr. Smolin presents the thesis that Universes are evolutionary - "natural selection on a cosmic scale". For an adventure of the mind, read this book.

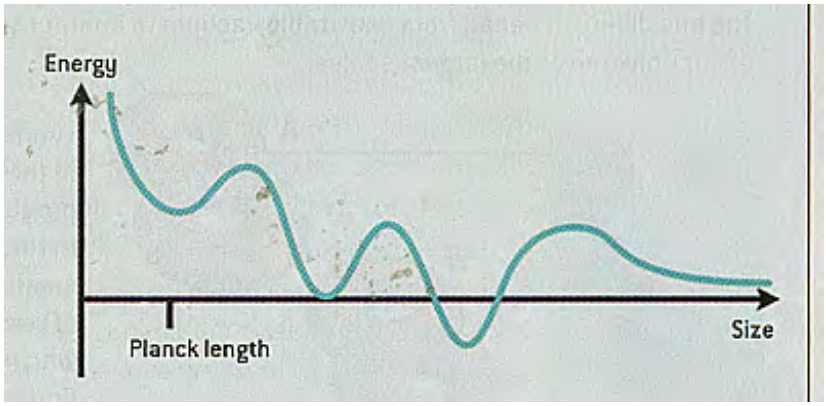
Again, after most of this page was written, another very intriguing and provocative article appeared in the May 2004 issue of Scientific American. Entitled "The Myth of the Beginning of Time", it was written by Gabriele Veneziano, the originator of String Theory. He uses that theory to help explain the moment of the Big Bang but he also belongs to the group of cosmologists who now believe that time is eternal (no begining - no end) and an infinite Cosmos has had repeated "creations" of individual Universes of which there must be multitudes. The subject matter is intense and rather mind-boggling, so no attempt is made here to summarize it. The reader who really wants to know current thinking about Multiverses and the openness of time should track this article down (several readings recommended). There is one figure in the article that summarizes much of his model, that is copied and reproduced here to provoke you into having a "go" at it.



The essence of this idea is that various forms of matter exist, each composed of strings with appropriate frequencies. This matter exists for all time throughout an infinite void defining the Cosmos. The matter is not uniformly dispersed so that periodically it clumps into discrete concentrations that form Black Holes of various sizes. A Black Hole while expanding spatially also increases in density towards its center until it reaches a threshold that results in a *rebound*, amounting to an explosion analogous to the Big Bang of our Universe (this situation is repeated many times throughout the infinite Cosmos). For a while thereafter the BB explosion goes through decreasing expansion rate (but the model does not yet treat the recently "discovered" apparent increase in our Universe's expansion). Veneziano prefers this model but does describe an alternative, related to the Brane concept, which goes by the name "Ekpyrotic Scenario" (read the paper for details).

The September 2004 issue of *Scientific American*, devoted entirely to Albert Einstein and the implications of his enormous contributions (mentioned briefly in the Preface to this Section) contains a paper entitled "The String Theory Landscape" by R. Bousso and J. Polchinski. It is far too challenging a read to summarize well here - you are encouraged to check it out yourself. But a few key ideas from it are stated here without much elaboration. The paper points out that String Theory may succeed in providing the fundamental knowledge needed to integrate gravity with quantum theory and also has the ability to provide a believable model for the Universe. But "Universe" has the meaning not just *our* known Universe but includes all possible Universes that arise from actions caused by vacuum energy changes in the *universal Universe* (This is my

phrase, meant to convey the idea that there is some kind of infinite [not spatially bounded] state [not quite a 'void'] in which many finite Universes, some like ours - others much different and many are transient, can have existences of varying duration.) String theory extends Kaluza-Klein theory (see page 20-1) to at least 6 added dimensions, none apparently large enough to presently be detected. These have a great variety of topologies (shape-determinant). The geometry chosen tends to adjust the associated vacuum energy to a minimum. For a quantity Bousso and Polchinski term "six-dimensional manifold", which is a string theory term, when a variable such as the overall size of the manifold is allowed to vary in a quantum event, the change of vacuum energy with size follows this general curve:



From Scientific American, September 2004

In this graph, the vacuum energy can decrease to states that place it above (+), at (0), or below (-) the abscissa. The energy change will strive to reach a minimum (one of the troughs) but which one depends on whether there is sufficient energy to climb out of a trough over a peak to the next trough. These minima are states that specify whether the vacuum energy fluctuation ends up in a forming Universe in a positive state (Universe will expand), negative state (it crunches or collapses), or, if near zero in a stable state.

When the topology of String Theory space is considered, a "landscape" of "mountains" and "valleys" (this holds for varying two dimensions but more dimensions may be involved and can't be shown in a 2D surface representation. But, to keep it simple, the 2D case (for which "landscape" is appropo) is considered to contain a huge number of mountain peaks and valley troughs. Entry into a trough may result in an unstable condition and failure to produce a Universe but there are some valleys in which stability (vacuum energy near zero) leads to long survival of the Universe that ensues. Empty space really has virtual particles and energy densities - it is not a true void. Early calculations came up with huge energy densities - very puzzling but proved outlandish as string theory improved and reconciled with quantum mechanics. But, a new determination of an energy density quantity called Λ_p, which is one Planck mass per cubic Planck length, showed a very large number of possible values with those near $\Lambda_p = 10^{-120}$ being closest to maximum stability. Various events continually occur in the boundless Universe but most end up in states of instability. However, those near 10^{-120} can survive and grow as expanding, long-lived bubble Universes.

Within any of these, new fluctuations may cause new Universes; a large number of Universes can keep popping up for short to long times everywhere in the (super)Universe that has infinite dimensions. Various surviving Universes (within this infinity and mostly unconnected and not interacting) will each have their set of specific parameters dictating their behavior. Most of these will be "inhospitable" for development of life; a few will have parameter values that permit life to develop and evolve. Our Universe, since we know it exists, must have had a vacuum energy density between $\Lambda_p = 10^{-180}$ and $\Lambda_p = 10^{-120}$, almost but not quite zero. In this model, at any given moment there can be countless bubble Universes undergoing various fates (growth, collapse, short-term existence, or almost instantaneous failure). But, if the void within its infinity is itself expanding, we should never be able to "see" or contact other bubble Universes.

One aspect of energy density needs further elaboration: There is a parameter called *critical energy density*. This has a value such that it just controls an efficient expansion of the Universe. Too low and the Universe expands too rapidly; too high and it collapses too early. At Planck time (10^{-43} sec), calculations suggest that the difference between Λ_{crit} and the actual Λ_p was only 1 part in 10^{60}. This condition at the outset of the BB was crucial to establishing a flat Universe (one whose geometry allows the sum of the angles in a triangle to be exactly 180°)

So, what is the future of the Universe, as can be predicted from what we now know about the timeline of the past Universe. The renewed cosmological expansion described above seems to rule out the Big Crunch - the model that was based on gradual deceleration, increased influence of gravity, eventual reversal of direction of galactic expansion, gradual collapse of matter inward, with all matter then "crunching" into some small volume that finally ends up as a singularity (with a probable new Big Bang; this expansion-contraction could be repeated many times). Instead, the accelerating expansion now observed predicts that matter will forever move "outward", probably for trillions of years. Matter draws apart and gradually cools as the energy of the Universe is dispersed. Thermally, absolute zero is approached for the evolving widespread matter - this condition has been called the "Big Freeze". Galaxies and stars eventually burn out (Red Dwarf stars are the last to go). This model implies that there may have only been one Universe.

From the above on this page, and in preceding pages, one can safely conclude two things: 1) a great deal more new knowledge of the Universe has accrued in the last 20 years, and 2) there is still a great fermenting of this knowledge into a gamut of hypotheses to explain the observations that are still "on the table" as theoretical cosmologists offer their variations of plausible explanations. If these last paragraphs touch upon the arcane or occult and confuse you (as I freely admit to having my uncertainties), I can only suggest you read the appropriate above-cited articles (at least twice) for, hopefully, more insight than provided here. A bit of metaphysics or "astrophilosophy" is still inherent in the speculations. Let's go off the deep end and plunge into the realm of the abstract.

---

*Conceptually, a "true" vacuum contains *absolutely nothing*, neither matter nor energy in any form, and never will; the term "false" vacuum applies to space that can be totally empty at times but in which matter and/or energy is potentially present and can be manifested at any moment, either as short-lived particles that self-destruct and can recur repeatedly or as an event set in motion [e.g., the Big Bang]. In all likelihood, there is no "true" vacuum, namely "a complete nothingness", that has a physical or spatial location. Beyond the limits of our Universe, the "Void" is conceptually filled with virtual particles, some of which could be neutrinos. Thus, such an absolute vacuum apparently cannot exist, despite the metaphysical inclination to think of the Universe or Multiverse [see above] as expanding into that which is characterized by a real "true nothing" capable of becoming something; however, philosophers can imagine such a total nothingness as an essential component in their notion of Being.

PREVIOUS rst NEXT

Primary Author: Nicholas M. Short, Sr.

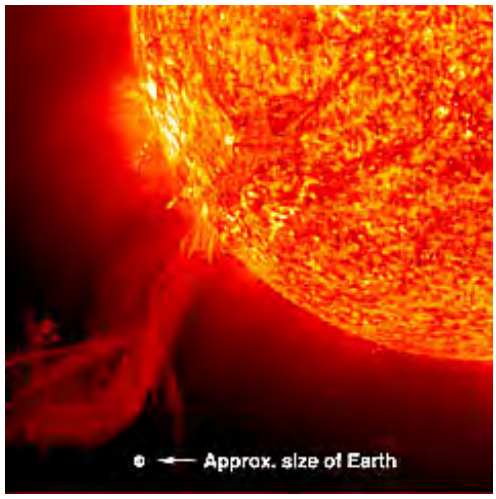
Models for the Origin of Planetary Systems

This "crash course" on Cosmology culminates with the following synopsis of the methods and results of astronomers' search for other planetary systems and consideration of the origin of our own Solar System, plus a quick look at several of the latest, provocative, and especially exciting theories (or brash speculations) on the start of life on Earth and, by extrapolation, the presence of other intellectual beings in our Universe that might be "out there".

There are several JPL movies which you may wish to view before continuing on this page. Access through the JPL Video Site, then the pathway Format-->Video -->Search to bring up the list that includes "Close up View of Plametary Birth", October 8, 2001; "Planet-forming Disks", January 11, 2002; "Beyond the Planets", February 4, 2003; and "Pointing the Way to Exoplanets", December 11, 2003 (this last is a full hour lecture). To start each one, once found, click on the blue RealVideo link.

As this page unfolds, one dominant idea regarding how planets form around stars will be developed. This idea is summarized in the diagram below so as to serve as a frame of reference for the prevailing views on planetary formation. The diagram simply shows a disc of gas and dust in a narrow plane around the forming star:

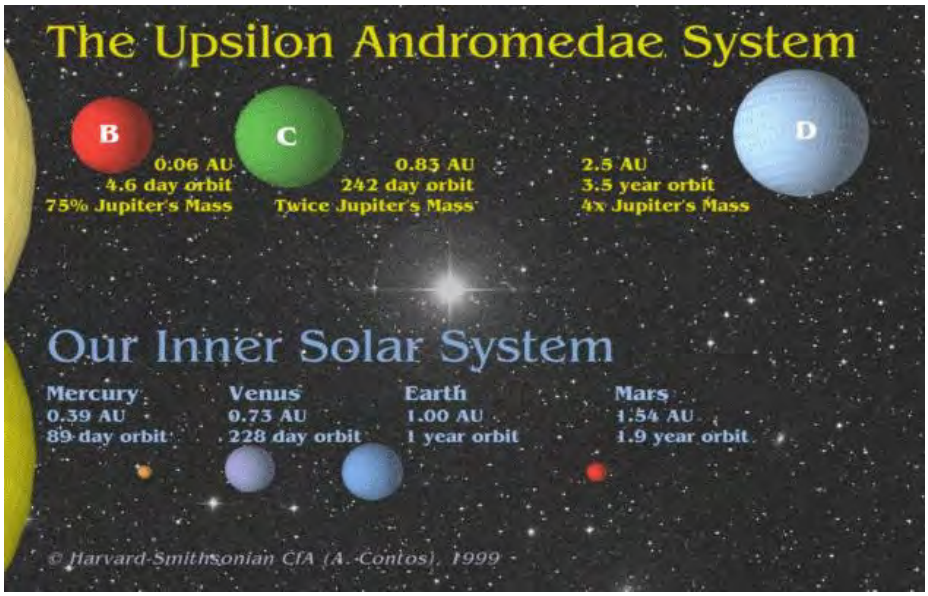
Two hallmarks in particular distinguish planets from the stars they orbit: First, they usually show marked differences in composition, being either gas balls (whose temperatures are well below fusion levels) with other elements besides Hydrogen that have rocky cores or they are dominantly rock with many elements making up usually silicate minerals. (Some have satellites with frozen water or other frozen gases in their outer shells.) And, at least one planet so far is distinguished by the presence of millions of compounds of carbon - the organic molecules found on Earth. Second, they are significantly smaller in diameter (hence volume) than their parent star. This SOHO image of a part of the Sun, with a solar prominence, illustrates this size difference, as displayed by adding a drawn sphere the size of the Earth to allow comparison:



This huge size difference between the Earth and its normal G star Sun is humbling in its stark truth. The great disparity in size also makes it clear how difficult it will be to find Earth-sized planets around nearby stars - and even more of a technology challenge as astronomers entertain hopes of finding stars elsewhere in the Milky Way galaxy and galaxies beyond.

It is natural for humans to wonder if there is life elsewhere in the Milky Way, and by implication in other galaxies. The starting point in searching for life is to prove the existence of other planets and inventory their characteristics. In the last decade the hunt for planetary systems has intensified. The first extrasolar planet (generally the term "**exoplanet**" is now in common use) was found in 1993 by Penn State University astronomers. A more definitive sighting occurred in 1995 - an object orbiting the star 51 Pegasi, which lies 50 light years from Earth in the Milky Way. The closest exoplanet found so far is just 10.4 l.y. away from us, orbiting the star Epsilon Eridani. By June 2000, planetary bodies had been detected in at least 88 other stellar (~solar) systems; as of October 2010 the latest count was about **400** individual planets associated with these and other systems -most being Giants and of low density. (Note: A few astronomers have disputed some of these observations.) There is a growing "feeling" (but not yet a consensus) that planets are the **norm** around many - perhaps most - stars, at least those of masses similar to the Sun. This is the basis of an argument with profound implications: **Statistically, there seems to be enough star/planet occurrences in the Milky Way, and by inference other galaxies, for a reasonable likelihood of some having hosted the evolution of life, and with a plausible probability of at least 10% *intelligent life*.** Where is it? - this will be discussed two pages later.

Analysis of one such system - Upsilon Andromedae -indicates it to have 3 planets (a second triple planet system was recently discovered); 8 other stars elsewhere have at least 2 planets. The Upsilon Andromedae system, diagrammed below, consists of Giant (probably gas-ball) planets (much smaller ones presently are very hard to detect), all within orbits whose distances from its star are comparable to that of the four small terrestrial planets in our Solar System:



In August 2004, a star, mu Arae d, just 50 light years away, was shown to have three planets, two larger gaseous ones and a third having a size, mass (14 Earth masses), and orbital characteristics (rapid revolution around its parent) that indicates it is likely the first rocky (inner Solar System-like) planet yet found. Then, at the end of August, announcement was made of star 55 Cancrie, 41 l.y. away and Cliese 436b, 33 l.y. away having a planet with 18 and 20 earth masses respectively.

Much attention was given in late September 2010 to the announcement that at least one of the 6 planets orbiting the Red Dwarf Gleise 581, about 20 light years from Earth, had conditions that could favor inception of some form(s) of life. Read about Gleise 581g at Space.com.

So far, only a very few possible planets have actually been seen (see below). Generally, planets are small relative to their parent stars and also have low luminosities. Nearly all of the 200+ planets claimed to have been detected are deduced to exist from their interactions with their parent star, involving perceptible movement of the star's position. Almost all discovered so far are large - Jupiter-sized or greater - and are gas balls. Several are extremely close to their star (at distances less than Mercury's orbit; one is thought to be as close as 5 million miles from its star and rotates rapidly around its parent body). Many planets have much more elliptical orbits than those moving in the Solar System. It is hypothesized that most planetary systems will consist of multiple planets but the smaller ones are presently still invisible to us and do not significantly distort the motions of their host star.

A high point in the current search for planets occurred on June, 2002 when several groups of astronomers announced jointly the discovery of up to 20 new planets - including at least two of Jupiter size - in the Milky Way galaxy, whose stellar parents reside at distances ranging from 10.5 to 202 light years from Earth. The closest in has been provisionally named *Epsilon Eridani b* (its star is the one actually cited in the *Star Trek* series as the location around which Mr. Spock's home planet, Vulcan, was orbiting). The rate of planet discovery seems to be accelerating. With it is the growing belief by cosmologists that planets could well exist in the billions, i.e., they are the inevitable result of processes that take place when most stars are born. Thus, planets may well turn out to be the norm - the expected, and perhaps the most significant products of nebular collapse in stellar evolution. (One recent estimate concludes that as many as six billion giant planets exist within the Milky Way galaxy.) The proponents of SETI (Search for Extraterrestrial Life; seeking primarily radio signals that have non-random character [perhaps some form of mathematical organization]) have been galvanized by these recent discoveries. (Some cosmologists are convinced that it is only a matter of time - probably during the 21st Century - until *contact* with other intelligent beings is achieved.) We will return to the SETI endeavors near the bottom of this page.

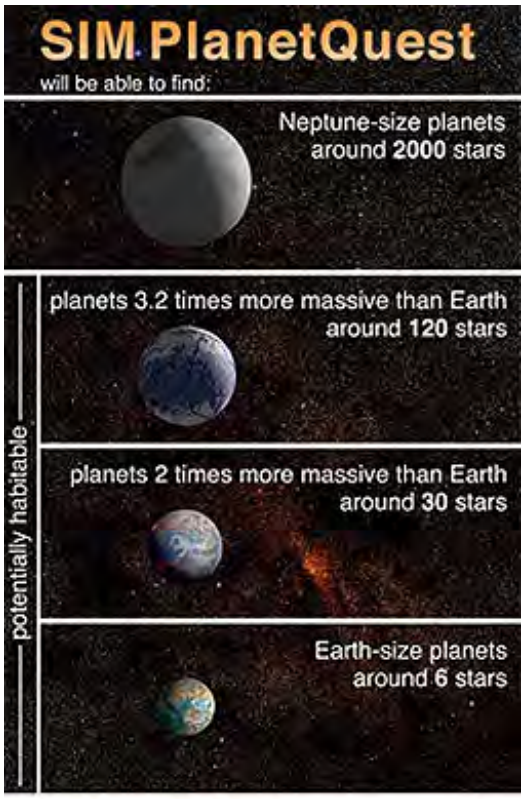
Current search for extra-solar planets is restricted to the Milky Way and (in principle) galaxies

close enough to Earth for an individual star to be resolved to the extent that changes in its motion can be measured. Gravitation attractions from orbiting planetary bodies cause the central star to wobble. This is the basis for the three prime methods currently being used to detect anomalies in a star's behavior that lead to the inference of one or more orbiting bodies.

The method that has so far been the most successful in locating (invisible) planetary bodies is called the *radial-velocity* technique. A component of a star's wobble will potentially lie in the direction on-line to the Earth as an observatory. This to and fro (forward-backward) motion causes slight variations in the apparent velocity of light. That, in turn, gives rise to small but measurable Doppler shifts in the frequencies of light radiation from specific excited elements, as expressed by lateral displacements of their spectral lines. From the wobble magnitude and period, the approximate orbit and mass of its presumed cause - the orbiting planet - can be calculated. This method is sensitive to wobble velocities as low as 2 meters per second. Jupiter, for example, causes a wobble velocity of 12.5 m/sec imposed on light radial from the Sun. Generally, this method, applied to nearby stars, will detect mainly large planets close to their star but in March of 2000, two planets about Saturn-size (1/3rd the mass of Jupiter) were found in this way.

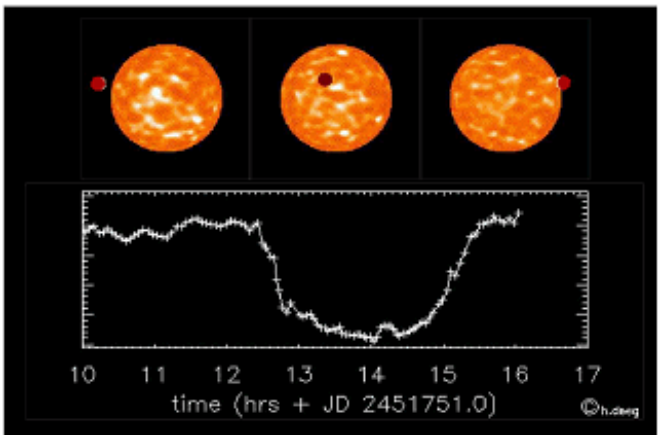
The second method, *astrometry*, also relies on star wobble but depends on measuring side to side displacements by direct observation through periodic sightings. This determination of relative shifts can be done on photographic plates taken of part of the sky at different times, commonly using the same telescope. But, considerable improvement in measuring shift results from increasd resolution achieved by positioning two telescopes physically some distance apart but keeping them joined electronically. This permits application of interferometry such that the two telescopes act as though they were one large one. Resolution as sharp as 20 millionths of an arc second (an arc second is 1/3600ths of an arc degree on the sky hemisphere [0° at the sealevel horizon; 90° at the zenith near Polaris in the celestial northern hemisphere]). The Keck interferometric telescope in Hawaii will soon be operational. This should facilitate detection of even smaller planets in nearby space or large planets in stars 10s of light years away.

As we have seen with the HST images in this Section, resolution (and clarity) capable of observing wobbles is significantly improved by operating one or more telescopes in space, above the distorting atmosphere. Resolution is also improved by using a pair of telescopes mounted on a single boom but separated by many meters, combining the image signals using the interferometric method. SIM (Space Interferometry Mission) is a NASA probe slated to fly sometime after 2015. Its two telescopes will be 10 meters apart. This will lead to a millionth of an arcsecond resolution, capable of inferring the presence of planets just larger than Earth-size or of large planets with far out orbits. Here is the spacecraft: A study of potential for success of the SIM approach comes up with this estimate of planets expected to be detected:

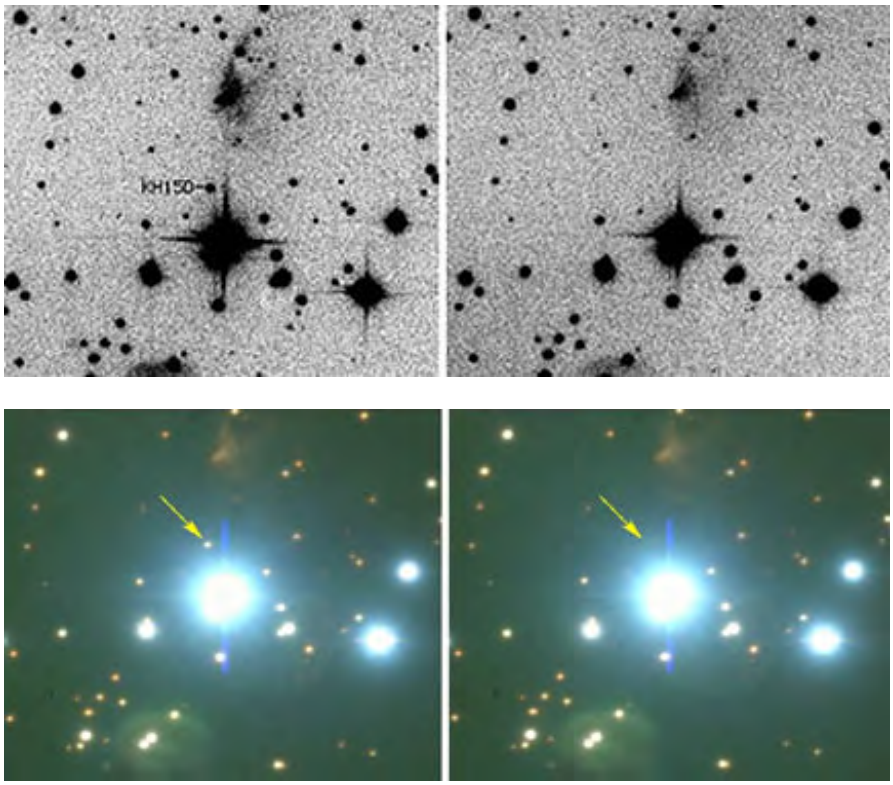


Interferometry, discussed in the Section on Radar, is an important technique for enhancing resolution. Its principles will not be discussed on this page but the reader is referred to these two websites for some insight: Wikipedia and Absolute Astronomy.

A third method is called *transit photometry*. When a planet's orbit takes it on a path where it passes across the face of the star under observation, the body will block out a small amount of radiation (usually visible light) for the period of transit. Sensitive detectors can note the slight diminution (up to about 2%). To distinguish this from a "transient event" of other origin, the astronomer needs to establish some regularity (reproducible at fixed intervals) of the drop in radiation, which will depend on the nature of the orbit (ellipticity; distance, etc.) Depending on planet size and proximity, the drop in stellar luminosity will be a few percent or less (an accurate determination helps to establish the planet's actual size). Such an effect was first noted in 1999 when a giant planet (earlier found by the radial-velocity method) passed in front of a star (HD 209468) whose light intensity underwent a drop of 1.5%. In a June, 2002 meeting, two other groups using telescopes in Chile have reported 3 and 13 possible transit detections, but these observations have yet to be confirmed independently. This is a generalized plot of the type of signal or signature that is produced by the transit method:



A fourth method has just had its first success in June of 2002. Examine this pair of images, shown first as a photo negative plate and then in color:



This is called the eclipsed or "winking" star method. In the left image (a photo negative), a Milky Way star KH 15D (2400 l.y. away; about the size of the Sun) is visible behind a much closer (or larger) star. In the right image it is totally absent, a condition lasting for about 18 days, and then it reappears. This on-off cycle occurs every 48 days. The eclipsing body could not be another star nor is it likely to be a huge (star-size) planet. The interpretation is that there is a cloud of asteroids and dust in a smeared-out clump orbiting KH 15D which block the starlight when a clump passes across the parent star; speculation considers that there may already be one or more planets formed from this debris. There may actually be two clumps (symmetrical pairing) at opposite positions in a single orbit; this has yet to be confirmed. A further anomaly: examination of photographic plates taken many years ago (although at limited intervals) does not detect this on-off phenomenon.

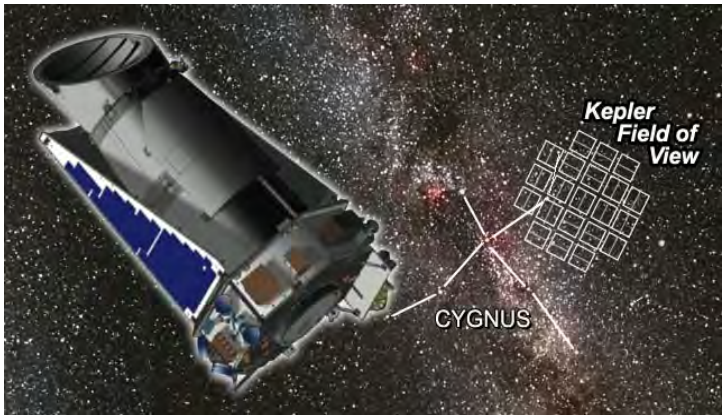
A fifth method is still in the experimental planning stage. The Terrestrial Planet Finder program at Princeton University is developing a special type of "cats-eye" mirror that will greatly reduce the effect of the luminous parent star. When this technology is deemed ready, they hope to persuade NASA or some other agency to use mirrors on several telescope-bearing satellites flying in formation and spaced to utilize the principle of interferometry to improve resolution and to detect small planetary objects.

Because (in part) of limitations at present in observational capabilities (mainly, telescopes are not yet powerful enough), there are inherent biases in finding planets of various sizes and locations around their parent stars. Thus, there is a tendency to find planets located around stars smaller than the Sun, since these planets (usually large) have enough mass to induce noticeable wobble in the less massive parent stars. Likewise, very large planets bring about increased wobble; the wobble also increases for planets close to the parent. Thus, so far Red Dwarf stars (which are quite abundant) have been the types associated with a disproportionately higher number of planet discoveries, since these stars are prone to greater stellar-wobble.

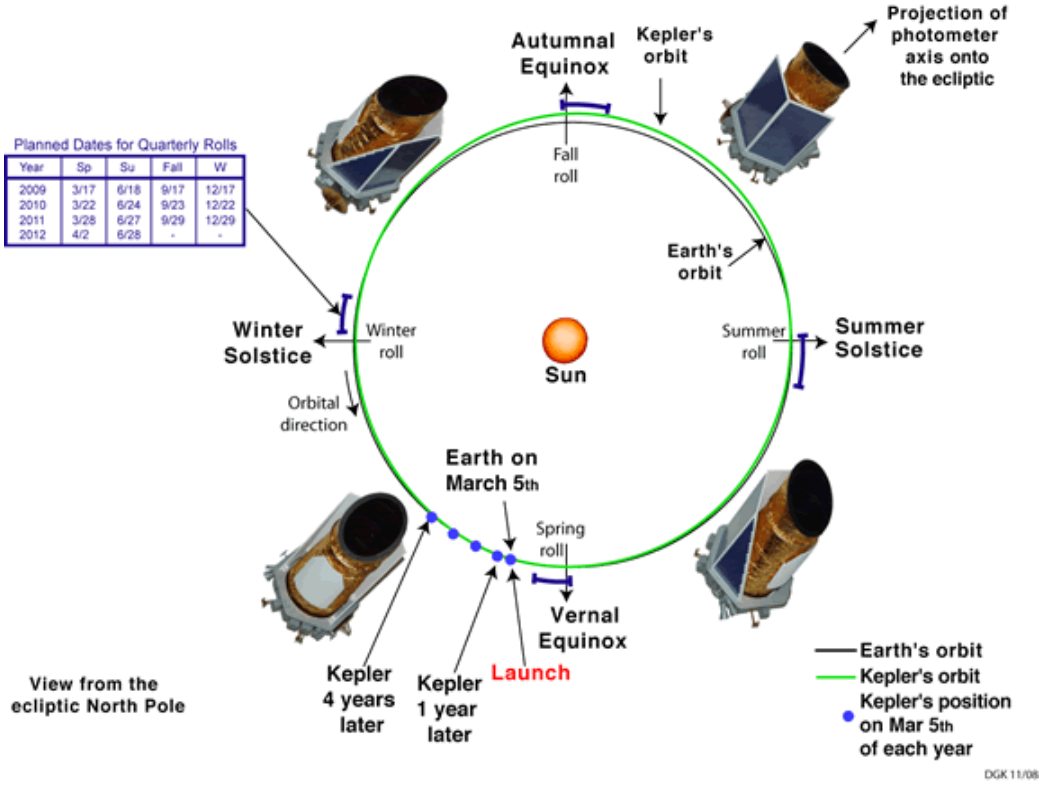
The ultimate dream is to directly visualize individual planets. This may be possible using several HST type spacecraft flying in formation ("clusters") with separations of a few hundred meters to hundreds of kilometers. In one mode, data will be combined using interferometric principles. Light from the central star can be blocked out by specialized image processing, leaving a residue of low luminosity orbiting bodies detectable by resolution- and radiation-sensitive interferometry. Both NASA and ESA each have in the planning stage such a mission (called The Terrestrial

Planet Finder and Darwin, respectively).

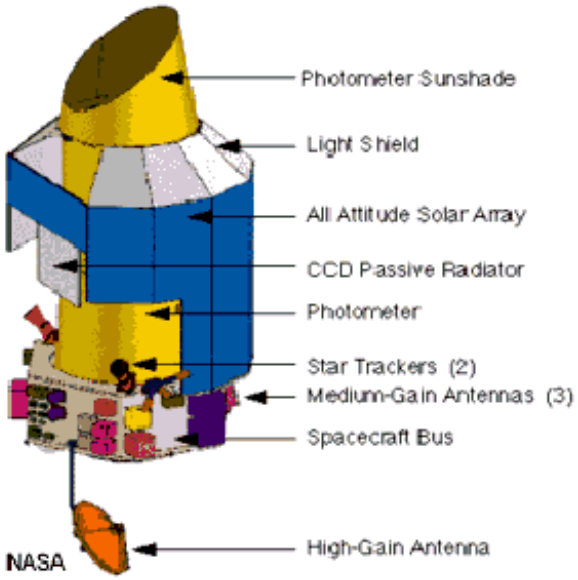
One of the most important missions since the Hubble Space Telescope is **Kepler** observatory (named after the German astronomer who formulated the laws of planetary motion). Using the transit method, Kepler will systematically search a small segment of the Milky Way in search of Earth-sized planets. Kepler was launched on March 6, 2009 at 10:58 PM EST from Cape Canaveral on a three-stage Delta II rocket and eventually drift about 45 million miles away so it won't have to contend with the reflected light of the moon and Earth.This overview illustration symbolizes the mission.



The spacecraft will occupy what is termed an "Earth-trailing orbit". Thus it will circle the Sun at a revolution cycle slightly greater than the Earth itself. This diagram amplifies this

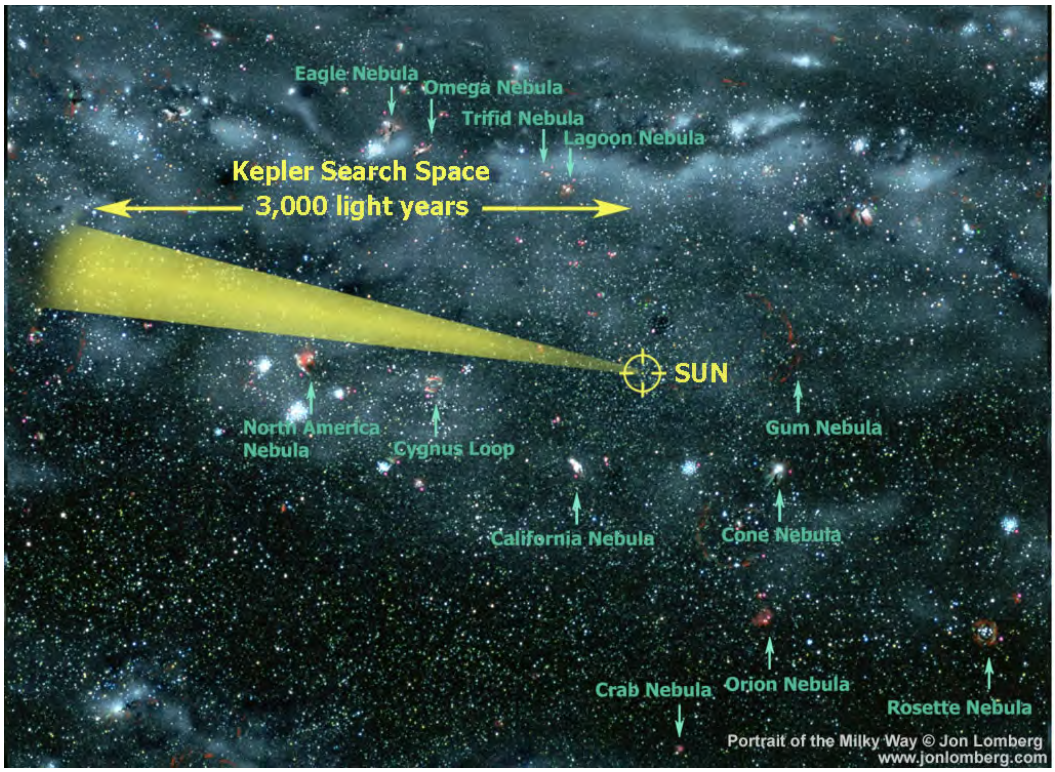


Here is a sketch of the Kepler spacecraft with its main components and a photo of the spacecraft in its fabrication facility:



The Kepler telescope has a 95 megapixel (detector) camera that can integrate brightness measurements to single out variations as small as 0.01% (capable of detecting Earth mass-sized planets). These are mounted in 42 panels, shown below, that help the instrument function as a photometer. Using its 0.9 m mirror telescope, up to 100,000 stars nearby in the Milky Way can be conveniently monitored over time and individuals with multiple planets should reveal the relative number of stars that possess planetary systems. The expectation is that at least 1000 new planets will be discovered, including some that are Earth-sized.

The region of the Milky Way to be monitored is located near the Cygnus constellation. Kepler will seek out stars that reside between 10 and 3000 light years from Earth. These diagrams are appropriate:



The next paragraphs, italicized, have been taken off the Internet from several sites that discuss details of the Kepler mission.

A transit occurs when a planet crosses in front of its host star as viewed by an observer. These transits very slightly dim the brightness of a star which allow for the detection of extrasolar planets. This change in brightness is very difficult to detect for terrestrial-sized planets, such as Earth, because they only dim their host star by 100 parts per million (0.01%), over a passage time lasting only 2 to 16 hours. That's equivalent to measuring from several miles away the change in brightness caused by a flea crawling across a car's headlight. Adding to the difficulty of the task is that dimming can be caused by events other than a transit, such as sunspots. Also, it is estimated that fewer than one star system in 100 will have planets properly aligned so that they pass between the star and Kepler's camera. In order for an extrasolar transit to be observed from our Solar System, the orbit must be viewed edge on. Any detectable change must be absolutely periodic if it is caused by a planet. In addition, all transits produced by the same planet must be of the same change in brightness and last the same amount of time, thus providing a highly repeatable signal and robust detection method. Because any planet in the habitable zone will require an orbit close to that of one Earth year, Kepler will need to observe any transits discovered amongst the sample size of 100,000 stars for at least 3.5 years to determine if the transit is periodic enough to be a planet.

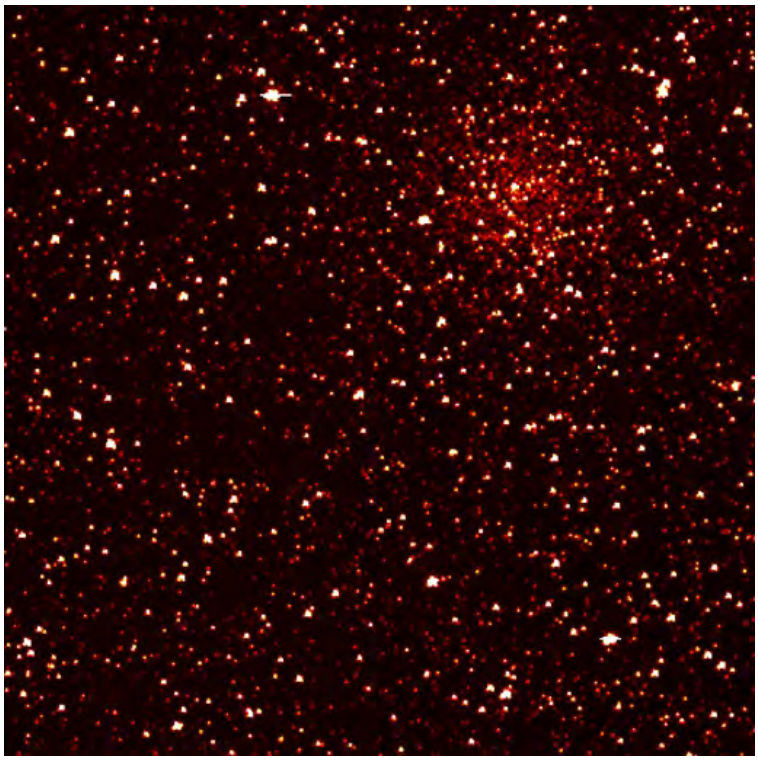
Over a four-year period, Kepler will continuously view an amount of sky about equal to the size of a human hand held at arm's length or about equal in area to two "scoops" of the sky made with the Big Dipper constellation. In comparison, the Hubble Space Telescope can view only the amount of sky equal to a grain of sand held at arms length, and then only for about a half-hour at a time.

Once detected, the planet's orbital size can be calculated from the period (how long it takes the planet to orbit once around the star) and the mass of the star using Kepler's Third Law of planetary motion. The size of the planet is found from the depth of the transit (how much the brightness of the star drops) and the size of the star. From the orbital size and the temperature of the star, the planet's characteristic temperature can be calculated. From this the question of whether or not the planet is habitable (not necessarily inhabited) can be answered.

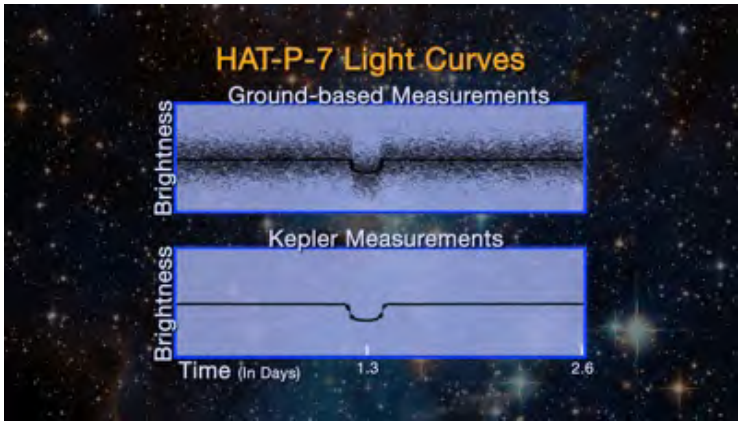
Most of the stars in Kepler's survey are relatively close, from tens of light-years to 3,000 light-years away. The first planets to be discovered in the coming months are likely to be more of the

same gas giants that have been found so far. The earliest likely announcement that another Earth-sized planet has been found may not come until December 2010 (assumes at least one repeat revolution of 300 or more Earth days after the first detection of a transit; this could be longer if planet orbits further out).

The first successful images from Kepler were obtained in April; here is a typical scene:



One of the first successful observations was of an occultation of star HAT-P-7 by a planet calculated to be a bit larger than Earth. The ground-based measurement, shown below, was notably noisier than the one by Kepler beneath it;

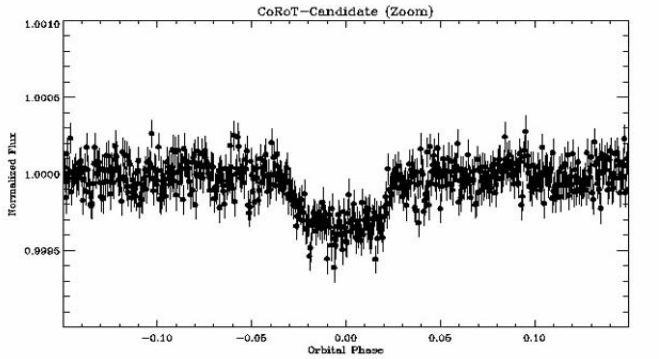


In sum, the Kepler Mission might not be able to directly determine whether or not we are alone in the universe, but it will be able to tell us if we have neighboring planetary systems, containing planets that might be capable of sustaining life. When compared to all the stars in the universe, even one discovery amongst the relatively small sample space of 100,000 stars will be significant enough for us to rethink our meaning and place in the universe. Proof that there are planets elsewhere whose size and characteristics are within the range of Earth, and therefore may be favorable for life will have profound philosophical and theological impact on mankind. (Of course, if none are discovered this too would be quite meaningful as it would favor the argument that "we" may indeed be alone - although this does not rule out possiblities elsewhere in the Universe.)

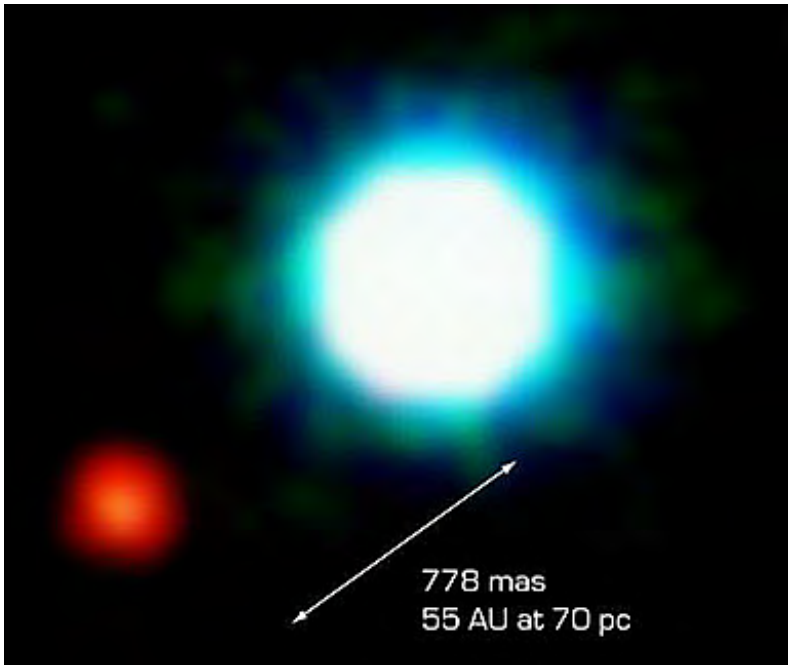
More about this mission is on the Internet at NASA's Kepler mission overview and at a Wikipedia

website.

Sponsored by the French and ESA, COROT was launched on December 27, 2006. Its primary mission is to search for planets using the transit method. It has already found several in the Jupiter size range. A recent discovery is COROT-Exo-7b, which has a mass about 5 times that of Earth and an estimated diameter of 1.7 times the Earth. Here is a plot of data received that illustrates how transit data appear:

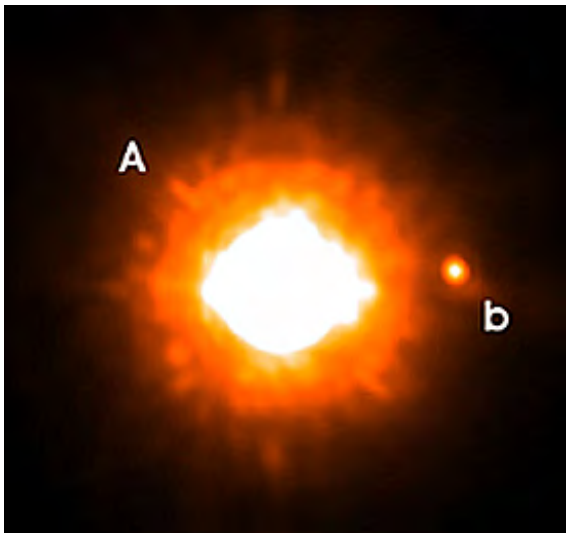


In September, 2004 a reasonable claim has been made by astronomers using the European Space Agency telescope in Chile of having seen the first actual planet. Look at this infrared image of their observation:

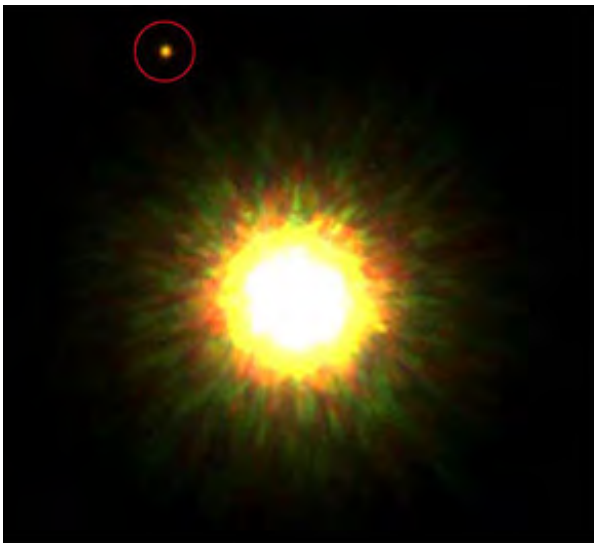


The star, 2M1207, just 50 light years away, is a brown dwarf (too small to initiate Hydrogen fusion). At a distance twice that of Neptune from the Sun is a reddish object five times the size of Jupiter but is cool (less than 2000° C) and has a spectrum that includes heavier elements. This object is likely a planet but final confirmation must await future observations of its changing positions as it orbits (there is a small possibility it is another object beyond the dwarf star).

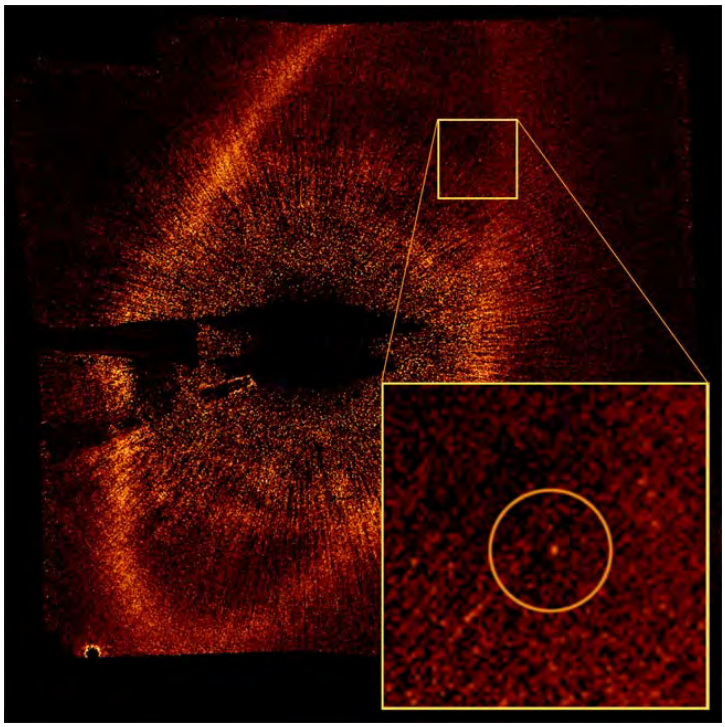
A stronger case was presented in March, 2005 by a group of astronomers using the European Southern Observatory. They have obtained a picture of GC Lupi with a definite planetary body distanced about 1 Neptune orbit from the star (a; with an extended atmosphere). This planet (b) is about 2.5 times the mass of Jupiter and has a surface temperature between 3 and 4 thousand degrees Celsius. Here is a telescopic view:



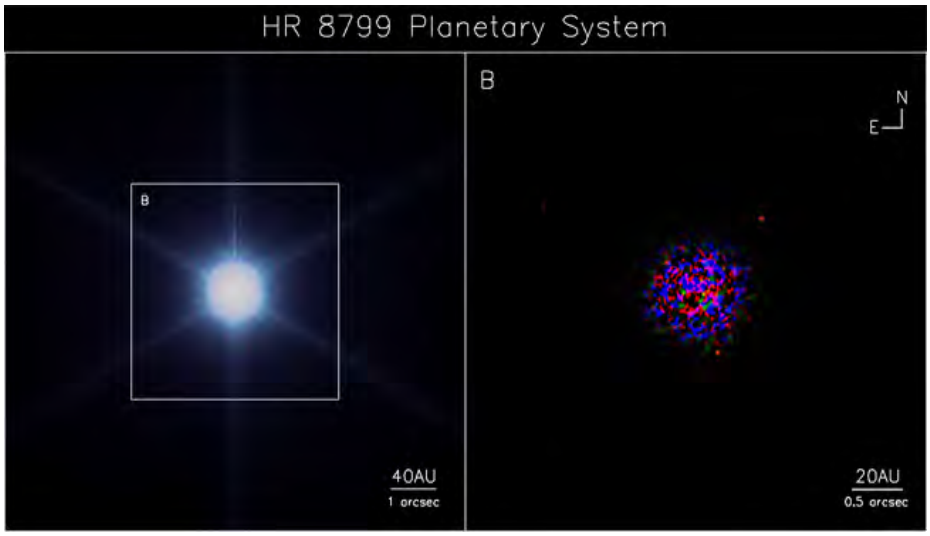
In September of 2008, astronomers released this image of a young star about the mass of the Sun, located just 500 light years away. The luminous body circled near "11 o'clock" is judged to be a large planet about 8 times the mass of Jupiter:



The case for each of these being actual planets is still being debated. At least one may be a brown dwarf companion star. In November of 2008 two papers in the journal Science offer evidence supporting planets actually observed. The bright star Formalhaut, 25 l.y. from Earth, was imaged by HST: there is a dot about 119 A.U. from the star that appears to be a Giant planet; note the broad ring of luminous dust and debris (asteroidal?) around the star (masked out).



The second paper presents imagery showing three planets around star HR8799, 130 l.y. away. The planets are comparitively young (estimated age: 600 million years), as they are still quite hot (accounts for their redness): Read the caption for details.



Statistically speaking, the number of such planetary systems in the Universe should extend into the millions within individual galaxies and the billions when the whole Universe is considered. It would logically be likely therefore that non- or weakly-self-luminous bodies, i.e. planets, are the norm orbiting around a central star for at least some of the size classes on the Main Sequence of the Hertzsprung-Russell diagram. As such stars proceed through their developmental stages (before they leave the Main Sequence), planets seem the inevitable outcome of the formational processes involved in star-making. So far, however, when the number of stars that have been studied using any of the above detection techniques are used as a base line, only about 5% have yielded evidence of associated planets; this number is probably a minimum for determining the actual percentage that do have planets since smaller ones cannot yet been recognized as present.

Two scientists, C. Lineweaver and D. Grether of the University of New South Wales in Australia, have recently published a study that relies on reasonable probabilities to estimate the number of planets just in the Milky Way. They argue that, of the approximately 200 billion stars they calculate to be the total population of the M.W, about 10% or 20 billion consist of stars similar to our Sun and most likely to have favorable conditions for planet formation. Assuming that, of

these, at least 10% will produce giant, Jupiter-like planets; thus their earlier number estimates 3 billion giant planets. Such large planets would almost certainly be accompanied by smaller ones formed out of the materials (they call "space junk") associate with the parent star. These giants help in the collection process that leads to smaller companions. But, more importantly, the giants serve as the principal attractors that gravitationally pull comets and asteroids into them (remember the Shoemaker-Levy event discussed in Section 19) and thus function as "protectors" of the small planets by minimizing the impacts these receive. Now, in a more recent presentation at the 2003 International Astronomy Union in Sydney, Australia they have raised their estimate to perhaps as much as 30 billion giant planets and a similar number of earthllike planets. This bodes well for future hunts as observational technology improves. Although this may seem "wildly optimistic", the likelihood of life on planets (see below) continues to rise dramatically with the increases in estimates of planetary occurrences - especially if one presumes that planets are the **norm** around stars in size ranges no greater than 10 solar masses.

For a while, astronomers assumed that most stars with planets would be relatively small - Sun-sized to perhaps 10 solar masses. These stars last for billions of years and thus favor the eventuality of life if planets developed during the stellar formation process. Now, several notably larger stars in the Milky Way have been found to have large planets. So, planetary formation is a function of process primarily and may have little to do with how long its star can survive. But the really big stars, even with planets, would burn their fuel and destruct long before evolution would likely foster even primitive organic matter.

The exoplanet systems probably show a wide range of individual types. Our Solar System is but one of many combinations of small-large, rocky-icy-gaseous planets. The results mentioned above are biased towards larger planets and give no real indication of the actual number in their systems. Some systems however may be almost like a binary star system in which the second or more planets approach brown dwarf mass but still too small to initiate nuclear burning.

The expectation is that planets have formed over most of the time that stars have developed in galaxies. One star pair (one a Pulsar; the other a White Dwarf) in a globular cluster within the Milky Way some 5600 l.y. away (in the constellation Scorpius), has been shown to be perhaps as old as 13 billion years. This is based on the sparcity of elements of atomic numbers higher than Helium. The large planet (4 times the diameter of Jupiter) now associated with it, is almost certainly the same age since prevailing theory holds planets to form roughly at the same time as parent stars. It verified as a true planet, it must be a Hydrogen/Helium gas ball similar to Jupiter. It is likely devoid of life owing to the turbulent history reconstructed for this star pairing and to the absence of life-forming elements. Even if some forms of primitive life did form on it or associated satellites, those would have perished because of harsh conditions that prevailed in its later years. But the chief implication of this observation (reported in July 2003) is that planetary formation can be traced to the early days of the Universe and, as carbon accumulated from the many early supernova explosions, some planets may have developed life of some type(s) since the first few billion years of cosmic time or even earlier.

A group lead by Sean Raymond of the University of Colorado has recently published results of a planet formation-distribution model that simulates (in the computer) conditions likely to produce planetary systems similar to those now being found around stars. One typical end result is shown in this diagram:

The lower row shows the case where an earthlike planet forms beyond the orbits of one or more giant planets. Also, smaller ice-crusted planets can form at even greater distances. These, or precursors, may collide with the earthlike planet(s) to contribute water that produces oceans which make the planet(s) habitable for life. The Colorado researchers conclude that at least one earthlike planet develops in about a third of the planetary systems that are produced by the model (which can generate different variations by changing key parameters).

In nearly all the planet discoveries, the orbital pathways (around ecliptic zones) of the planet(s) in a system coincide well with any residual planetary debris disks that were observed. This supports the model described below in which planetary formation takes place in the disk that develops around the parent star as it contracts and begins its life as a Hydrogen fusion object.

Nearly all the planets found to date are much larger than Earth, being similar to Jupiter as gas balls. This may be an observational bias: our present techniques can only detect larger planets. What still seems unusual is the indication of many big planets orbiting much closer to their parent (relative to the Jupiter/Saturn/Uranus/Neptune distances from the Sun). Two ideas have been postulated for this: 1) these planets were once placed much farther from their parent and have been drawn closer by the star's gravity (and thus are near their demise); and 2) the protoplanetary disks that now show these close-in stars may have been smaller relative to the Solar System dimensions, and hence the stars actually were formed nearer to the parent.

These planets are almost all in the "Giant" category, since they are large enough to produce presently detectable effects on their parent stars. But a recently reported study (July 2007) has pointed up a paradox: Giant planets that are more distant than a few Astronomical Units are rare in other planetary systems (in the study none were found around stars that should have such systems). Unlike our Solar System, giant planets so far detected are nearly all relatively close to their parent stars. The reason for this is presently uncertain.

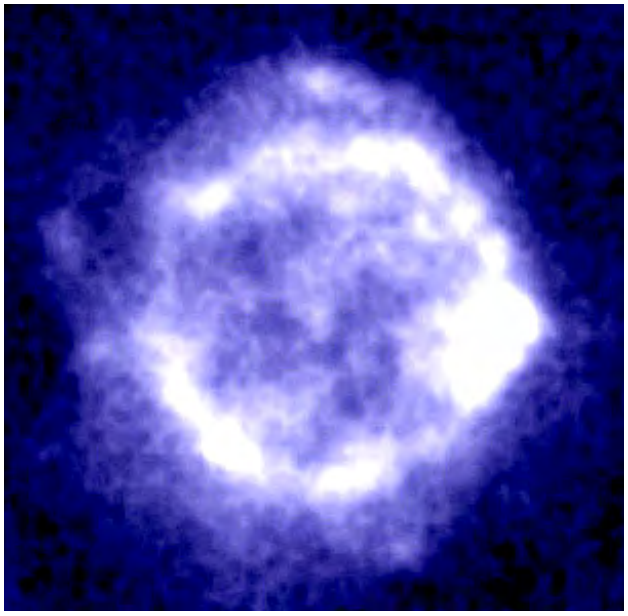
On April 24, 2006 came the announcement that the first planet similar to Earth in size and in conditions for life had been discovered. Three planets, with masses 5, 8, and 15 times that of Earth, have been found orbiting the star, a Red Dwarf, Gliese 581, located 20.5 l.y. from Earth. The 5 Earth mass planet, designated Gliese 581c, shows convincing evidence of a rocky or icy sphere whose diameter is 1.5 times that of Earth. It lies only about 11 million km (7 million miles) from its parent star (which has a much lower radiant energy output than the Sun). Because of this proximity, the planet may be locked onto the star such as to always have one side facing it (like the Moon). The planet has an atmosphere (composition TBD) and a temperature range from 0 to 40° C (32 to 104° F). This range would allow liquid water to exist at the surface. While much more needs to be learned about Gliese 581c, it seems to have passed the "Goldilocks test" - conditions favoring life may indeed exist there. More about this planet is available at this Wikepedia web site.

Another Red Dwarf star, Gliese 436, has a single detected planet about the size of Neptune. Gliese 436b's surface temperature is ~310 ° C. Its estimated density suggests water as the

main constituent (hydrogen-rich atmosphere). Because of high interior pressures, some of the water is ice, but at temperatures approaching 100° C, possible because its freezing temperature is elevated by the greater pressures.

While astronomers involved in planetary studies exercise caution about conclusions that specify the number of earthlike planets to be expected from various estimation procedures, they do propose that that number should be in the millions. Whether such planets also harbor life is much harder to pin down numerically but the statistical approach suggests that a significant fraction of the earthlike planets would possess the proper conditions. How much of that is intelligent life is still guesswork. The current sample of 1 (us!) is the only established data point. But if the reality is actually a much larger number, then, purely from statistical logic, we should expect that some of these intelligent civilizations should be more advanced than ours. Why we have yet to "hear" from them remains uncertain (but now SETI improves the chances for this) unless there is some fundamental reason that makes space travel, even from nearby stars, very difficult.

Now, let's turn to consideration of the ways in which planets form. For planets in general, both terrestrial and gas envelope types, dust maintained by the parent star's gravity must be present in sufficient quantities to collect as cores or to comprise the main body of the planet. There is plenty of dust in galaxies, mixed with gases from which stars emerge. The source of the dust has been somewhat problematic but a prime candidate is exploding stars (Supernovae) large enough to synthesize Silicon, Oxygen, Iron and other heavier elements. A recent observation strongly supports this:

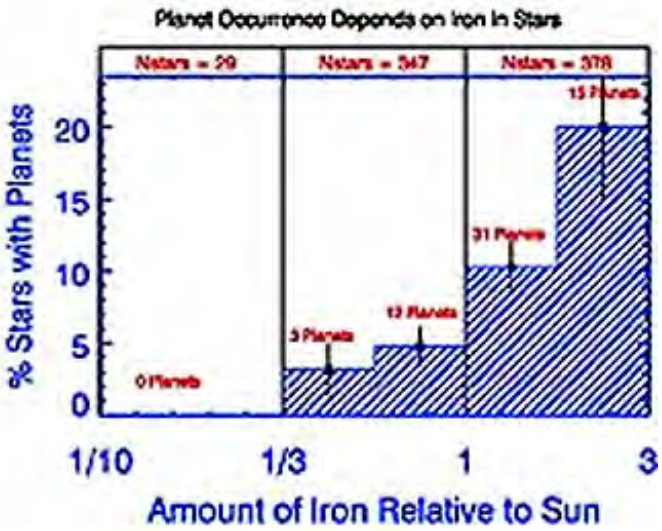


This 300 year old Supernova is hard to see in Visible light because of a superabundance of dust. When imaged in submillimeter light, the above pattern stands out. The brightest areas are advancing gases with large quantities of dust that give off light at these wavelengths.

One essential requirement for a planetary system to develop is that it forms during the organization of a central star (possible exception: a captured planet, probably rare). Also critical is the availability of dust and gas. The processes involved can be somewhat varied but are sensitive to a relatively narrow range of conditions. The sequence of formative events probably begins during the T Tauri stage of developing stars in which the conditions are favorable. These stars have notable dust clouds (particulate nebulae) that can be monitored in the Infrared. Some evidence indicates the clouds will begin to reorganize their tiny particles into large clots, which can grow to planet size, in about 3 million years. But, detection of cold nebular material at longer wavelengths suggests the dust can take 10 million years or more to build up any planets that may result.

Another important factor, recently reported, is that stars which have a relatively high content of

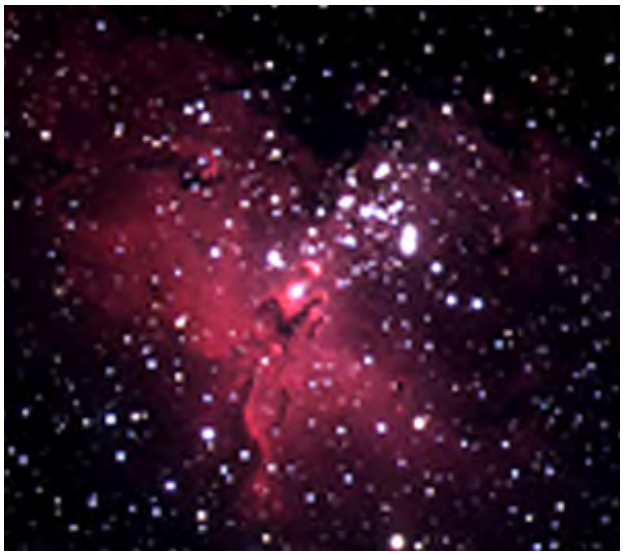
Iron (from gases enriched by repeated mixing of Supernova explosions over time) will have a much greater likelihood of producing planets. Iron is a measure of the metallicity of a star (page 20-7). The Iron may be needed to develop planetary nuclei (most ending in planet cores) that help in the gravitation attraction that drives accretion through collisions and infall. Stars with three times the Iron content of the Sun have an estimated chance of having planets set at 20% (This comes from a study of 754 nearby stars in the current (on-going) inventory of which 61 have detected planets (this amount to a probability of about 8% for all stars of mass less than 10 times the Sun having auxiliary planetary bodies). The results of this study are depicted in this graphical diagram:



The paradigm summarizing the processes involved in the formation of the Sun and its planets probably applies (with variations) to most other planetary systems in general. The first realistic notion of how planets form was proposed by Pierre Laplace in the 18th Century. In its modern version, both stars and their planets are considered to evolve from individual clots or densifications within larger nebular (cloudlike) concentrations dominantly of molecular Hydrogen mixed with some silicate dust particles that spread throughout the protogalaxies and persisted even as these galaxies matured. In younger stars, much of the Hydrogen and the heavier elements are derived from Nova/Supernova explosions that have dispersed them as interstellar matter that then may initiate clouds or mix with earlier clouds. Such nebulae are rather common throughout the Universe, as is continually being confirmed by new observations with the Hubble Space Telescope.

One of the best studied and, in itself spectacular, is the Orion Nebula, seen here:

Below are two views of nebular materials associated with the famed Eagle nebula (page 20-3): Top = full display of the M16 (Eagle) nebula (note the dark dust areas; the white dots are stars lying outside this nebula); Bottom = details of the temperature variations in the dust making up the solid particulates in the Eagle nebula as imaged by the European Space Agency's (ESA) Infrared Space Observatory (ISO) at two thermal infrared wavelengths, in which red is hot and blue is cooler (about - 100° C):

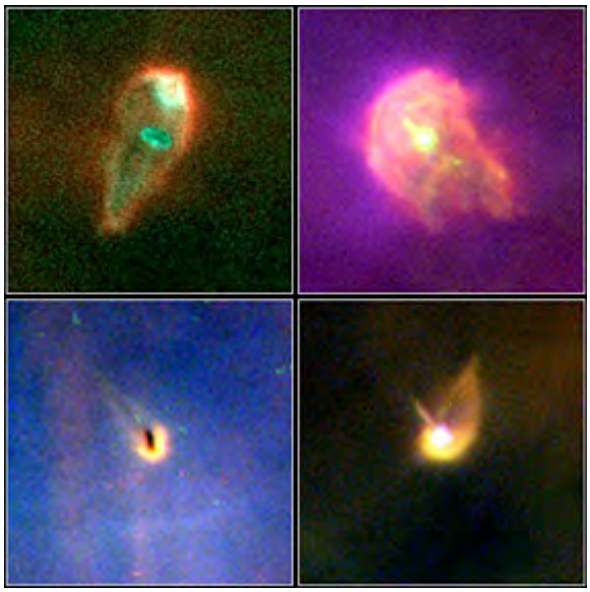


As individual stars start to develop within these gas and dust clouds, in many instances that dust will organize into a protoplanetary disk (see third figure below). The NICMOS Infrared camera on the Hubble Space Telescope has observed a prime example of this stage, in which the glowing gases moving into the central region where a protostar is building up are cut by a band of light-absorbing dust that is most likely disk- shaped (can't be verified from the side view in this image):

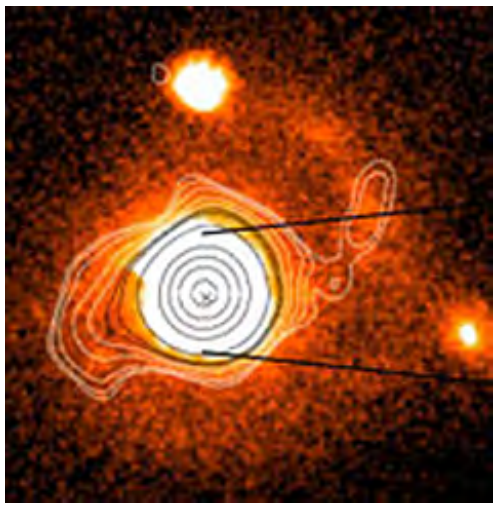
An Earth-based telescope has captured this view of a nearby star (nicknamed the "flying Saucer"), again with a girdling disk of dust (and an as yet unexplained anomaly in that the upper half of the image is redder than the bluer lower half):

Examination of these gas and dust clouds by HST has led to the discovery of small clumps or knots of organized gas-dust enrichment within the protoplanetary disks called *Propylids* found in the neighborhood of a parent star. This may be a more advanced stage of materials concentration that results in a new star with an envelope of gas-dust suited to accretion that produces planets. Three such Propylids are evident in this image of the nebula associated with the Orion galactic cluster.

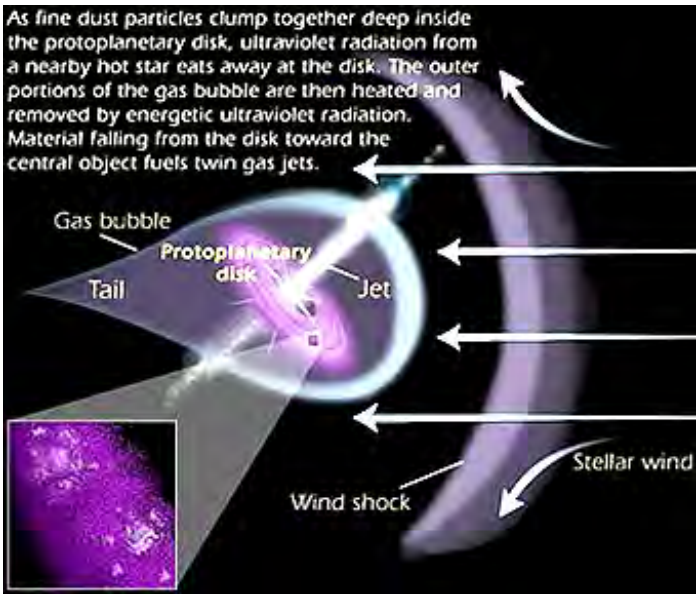
Other individuals, at least one possibly formed as recently as 100,000 years ago, were found during the Orion study (see page 20-2 for a view of the entire young nebula in the Orion group); look like this in closer views:



Propylids are vulnerable to being destroyed by UV radiation from massive, young nearby stars. It is surprising therefore that many Propylids (some shown below) have survived in the Carina Nebula, which has numerous UV-emitting stars. Other factors must be involved



Development of planetary cores, from which full larger planets then form, is a race against time. Examine this model:



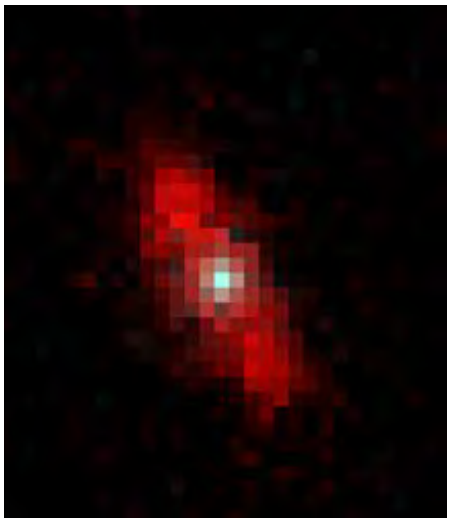
The major threat to protoplanet formation is the stellar (solar) wind coming from the parent star. UV radiation is also able to break apart the smaller particles. Wind and UV radiation are capable

of pulling apart small particles. These particles in the early stages are tiny bits of dust (solid; solids with an ice coating; ice) which are charged electrostatically. These will collide from time to time and may stick. If not disrupted by the stellar wind the now larger particles can again bump with others. By the stage in which some particle conglomerations have reached baseball size, they can resist stellar wind forces and survive to grow ever larger through collisions.

Thus, as the gas and dust cloud forms around a growing star, particles of solids begin to clump and some survive disrupting actions. However, much of the gas and dust may be pushed continuously away by the wind and radiation, so that the amount of material available to form planets generally diminishes over time. Evidence suggests that most Propylids are blown away before a planet grows large enough to survive, implying that the planet formation process may be less efficient and common than had been thought during the last decade. If the planetary cores do build up fast enough, they will survive the expulsion of the bulk of the gas/dust. This phase of planet formation occurs typically in a time frame of just 100,000 years or so; it is estimated that 90% of such clouds are dissipated before significant planetary cores can form. Planet accretion leading to survival is estimated to take up to 10,000,000 years.

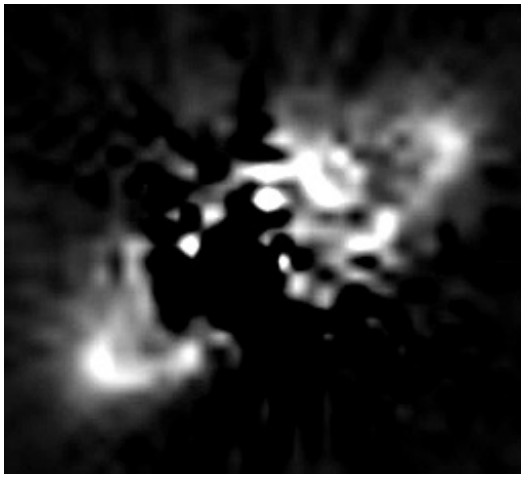
Most galaxies began during the first half of the Universe and contained a large number of massive stars that formed early in galactic histories. These galaxies have continuously been evolving through the eons as their stars synthesized elements (see page 20-7) and then dispersed them by the Supernova process. New, mostly Main Sequence stars, chemically enriched with elements of higher atomic number, have continued to form well into younger times up through the present. The smaller stars have lasted much longer and are probably the preferred sizes suited to planetary formation and survival. Even today stars are developing from the gases contained in remaining nebular material, so new planets could still be forming.

A telescope observation, reported in April, 1998, records the sighting (through the Keck II telescope on Mauna Kea, Hawaii) of what is interpreted to be another "solar" system around star HR 4796 (about 220 light years away). This image (at a lower resolution in which individual pixels stand out), taken in the IR, shows this central star (yellow white) surrounded by a lenticular (in an oblique view), flattened disk of gases and solid matter (glowing hot [reds] in the Infrared):



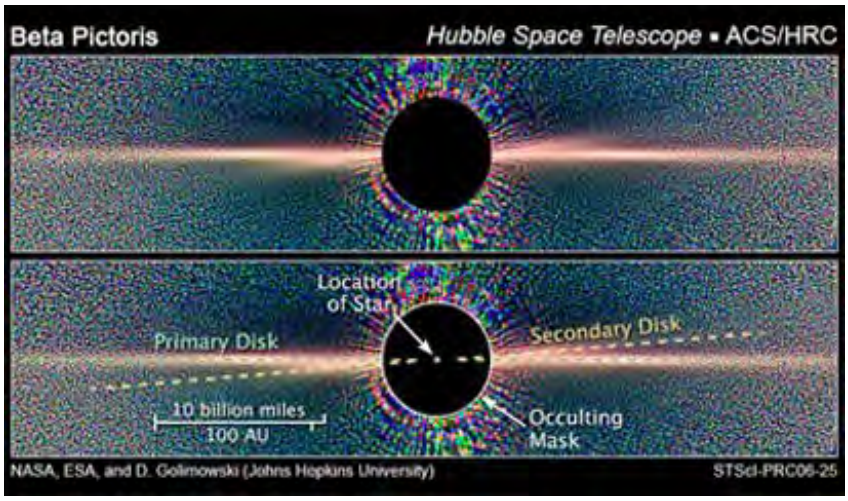
The diameter of the lens is about 200 A.U. No evidence of individual planets can be made out but the discoverers judge this feature, which has caused quite a stir of interest, to be an emerging planetary system in a "young adult" stage of development. It will certainly be a target for more detailed HST observations.

More recently, the Hubble Space Telescope has imaged star HR4796A, in our galaxy, which shows both a disk and irregular dust and gas clouds. This disk is interpreted to be in a more advanced (mature) stage of development than the protodisks shown above. Although not discernible, there may be planets already in the evolving gas/dust cloud which is made visible by the star's light.

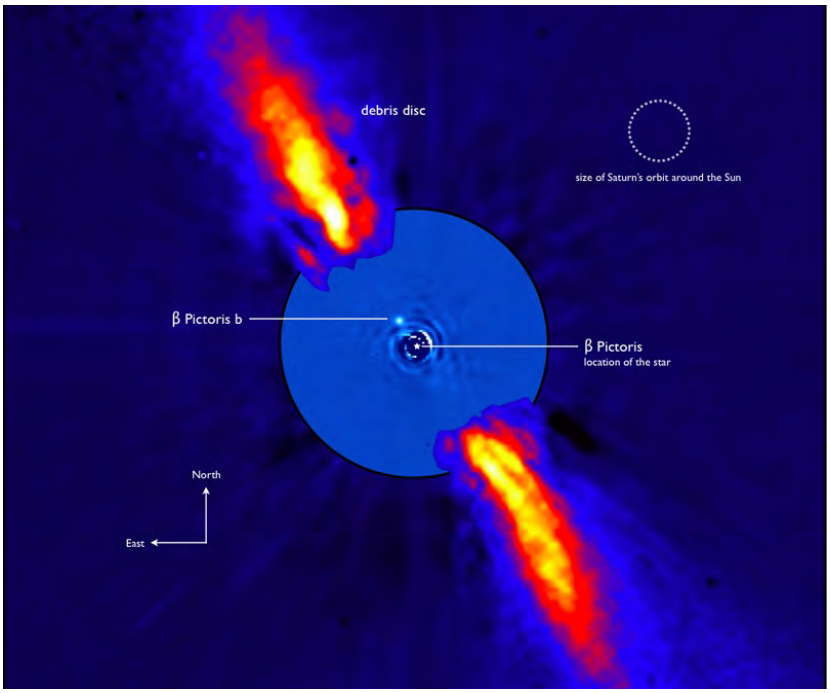


This next Spitzer Space Telescope image shows HD6105, a collection of gas and dust that appears to be forming planets. Its central star has been blacked out in this image.

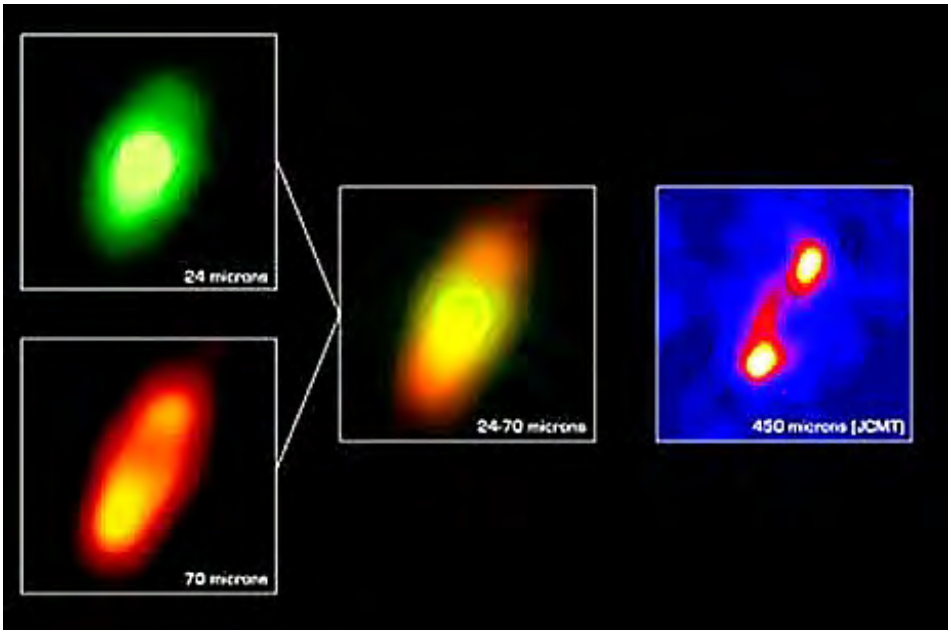
Another Hubble achievement is the imaging of a prominent disk around Beta Pictoris (63 l.y. away) using the coronagraph to block out that star's surface-emanating light. There is a more obscure, but real, partial second disk developed about 8° offset from the primary disk plane. At least one planet (roughly Jupiter-sized) was also indirectly found in this system.

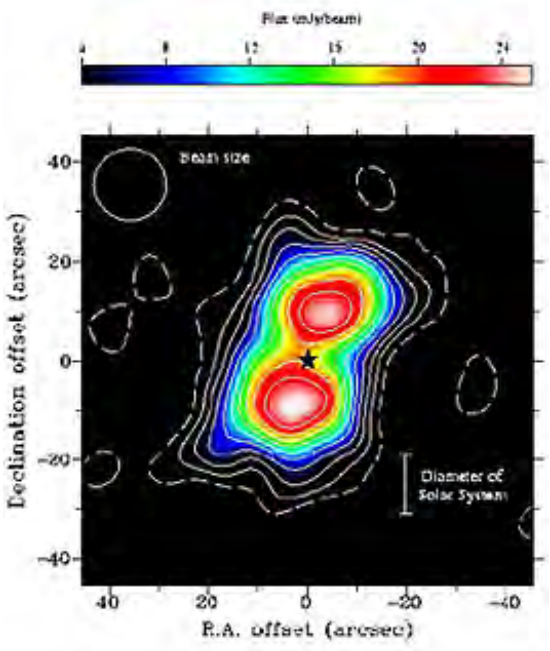


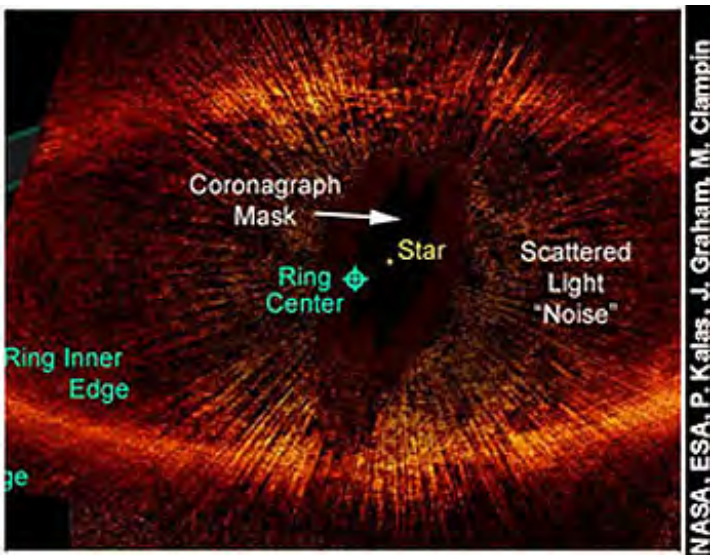
This next image of Beta Pictoris, made by a European Southern Observatory IR telescope, has actually imaged a large planet inside the dust disk. The planet has been given the provisional name Beta Pictoris b.



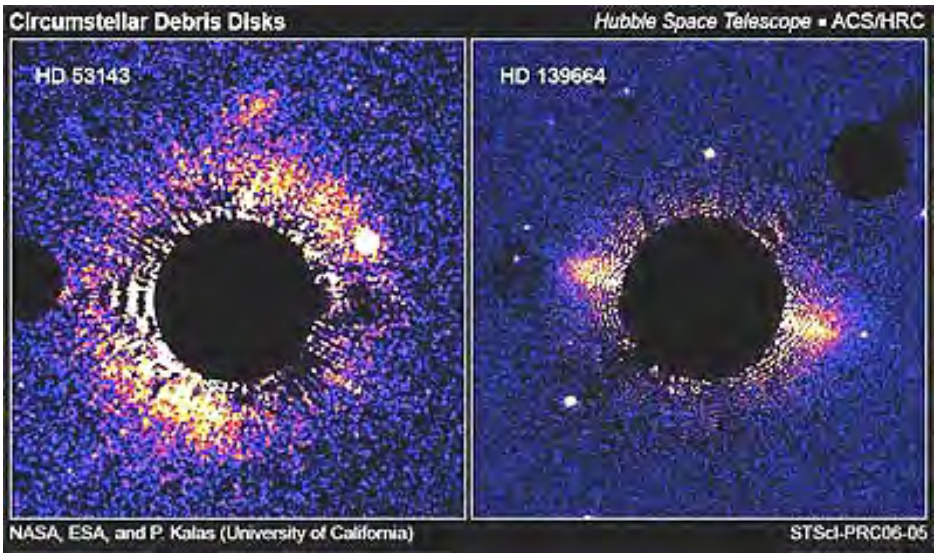
A very well-developed debris disk occurs around the young (200 million years) star Fomalhaut, the 17th brightest star in the sky (southern celestial hemisphere), just 25 light years away. The disk, about 150 A.U (20 billion kilometers) from its off-center star, is analogous to the Kuiper Belt of debris around our Sun. It appears to consist of icy dust. Time-motion studies indicates that there may already be one or more planets that are influencing the large objects in the disk, much like small moons perturb Saturn's rings. Three illustrations showing the Fomalhaut disk, the first with the Spitzer Space Telescope, the second made with IRAS, and the third through the Hubble Space Telescope, appear here:



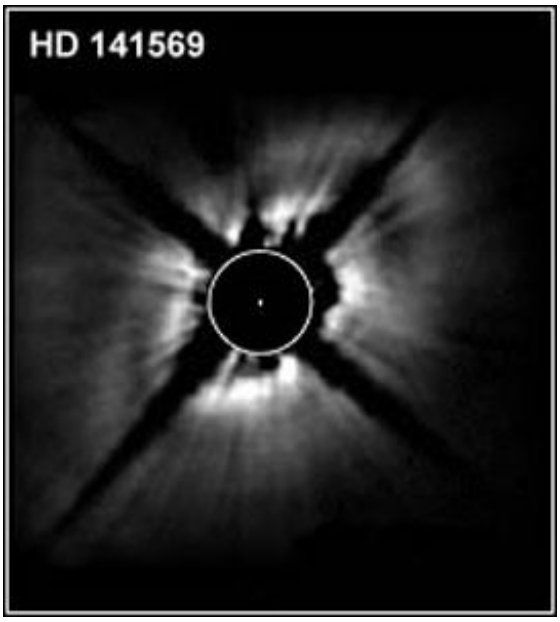




The HST has also imaged two other distinctive debris disks around a central star, as shown in the next illustration. The disk on the left is caused by reflected star light off myriads of small particles; the star, blocked out, is a Red Dwarf. On the right is another disk seen at nearly normal to a blocked out Red Dwarf 88 l.y. from Earth that here appears orange because of a different combination of spectral bands used to create the image. This star-disk system may be as young as 250 million years since its protostar began to burn.

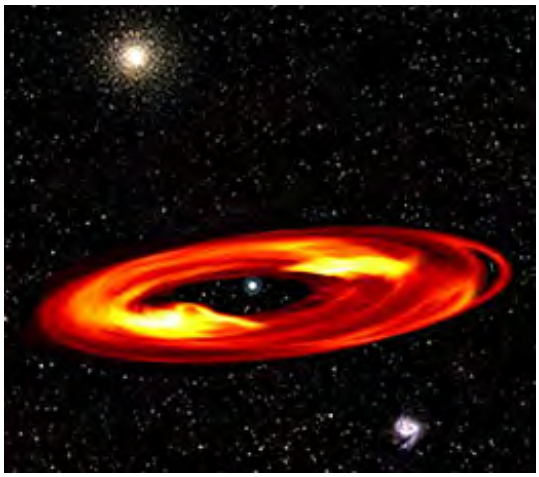


Theory indicates that, in the earlier stages of planetary formation, some number of broad rings should develop at various distances around the central star. One or more of these would appear as torus like glowing collections of dust and gases. At least two stars with this feature were imaged by HST and reported in January of 1999. Here is one of the star-ring systems:



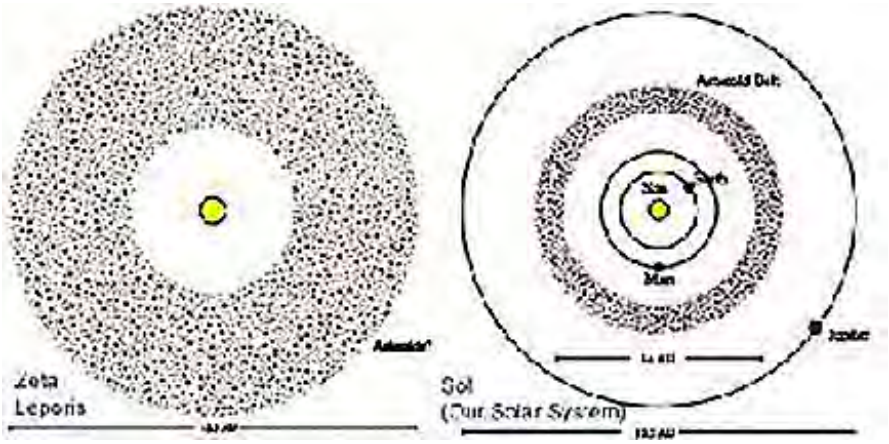
Visible is a bright ring at some considerable distance out from the tiny parent star (white dot) and a more diffuse, darker mass extending beyond, both features occupying a flattened disc. In this instance, there are no rings close in (analogous to the regions occupied by the inner planets of the Solar System). The white circle is added by the astronomers to mark the boundary within which no visible planetary disk matter has been detected; the broad black cross (X) is an optical artifact. This star is about 350 light years from Earth.

A recent image, made from data detected at 1.3 mm by a French radio telescope, may have caught the formation of two large clots of matter likely to eventually contract into giant planets. These occur in the ring around the central star Vega, 25 light years away (in the Constellation Lyra). Here is an image based on observations made by D. Wilner and D. Aguilar of Harvard's Smithsonian Center for Astrophysics. (Note: this image has been enhanced artificially as an artist's rendition.)



C. Chen and M. Jura, Univ. of California-Berkeley have detected (monitoring Infrared radiation through the Keck telescope on Mauna Kea) a ring or disk of dust which seems to contain asteroid-like bodies around zeta Leporis, a star some 70 l.y. from Earth that is twice the mass of Sol and 15 times its luminosity. The ring, first detected by IRAS, is much closer (~5-12 A.U.) to its parent star than the distances found for other recently observed or inferred planetary bodies around stars. Although imagery of the star, HR 1998, does not directly reveal these bodies, their

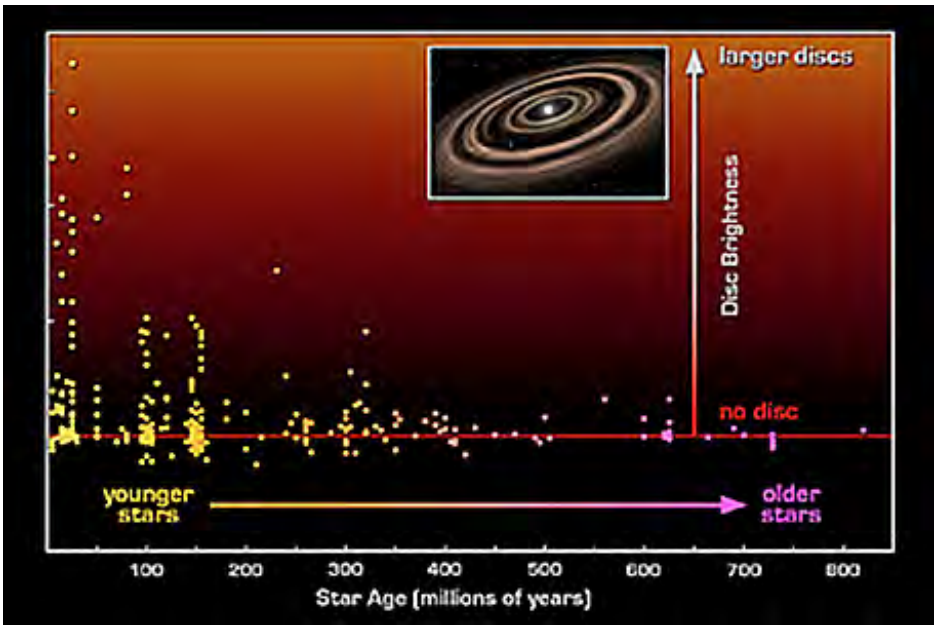
presence is inferred from their average temperature of 350°K (77°C or 170°F). From the data they accumulated, Chen and Jura have produced a plot of the asteroidal ring around zeta Leporis, with a comparison of the Sun's asteroidal belt also displayed:



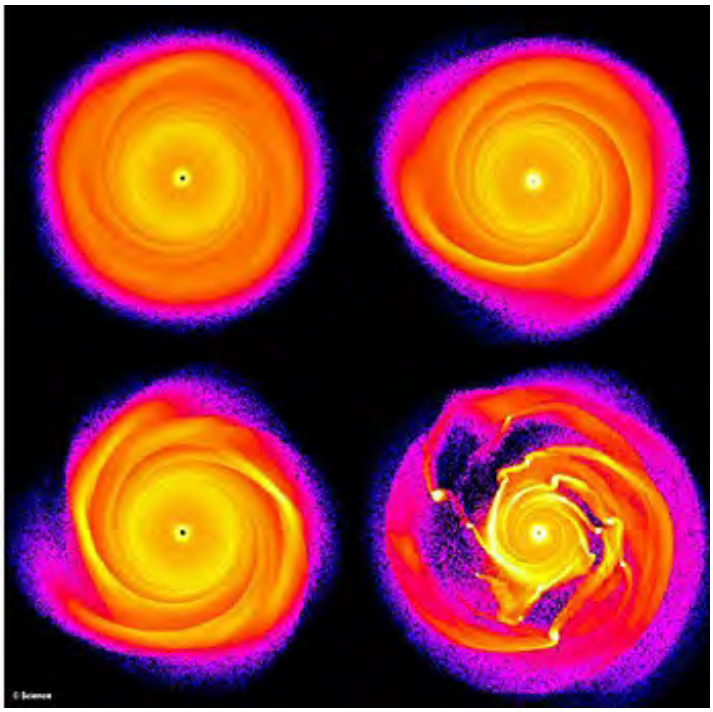
Chen and Jura propose this ring to be the precursor of eventual formation, by collision of asteroidal bodies, of rocky planets analogous to those of the Solar System. These bodies, form from smaller particles (dust) condensed from the gas-particle cloud associated with the forming star. Much of that dust can move inward towards the star by a process called Poynting-Robertson drag. This is caused by radiation from the parent star being absorbed and re-radiated differentially, leading to a Doppler effect (here, the energy of emission in the direction of dust motion is at shorter wavelengths [more energetic] and thus by retro-action slows the particles) that promotes drift of the dust towards the star.

There is evidence in our own Solar System that dust is commonplace and widespread. Astronomers can have their observations slightly diminished by the phenomenon called "Gegenschein" (German for "counter shine"). This has been attributed to dust in intersolar space beyond Earth. Here is a photo of this effect:

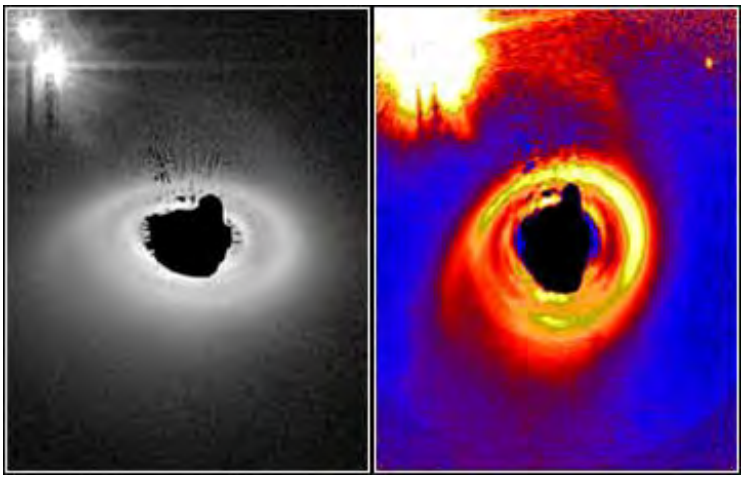
From the above, one necessary step in planetary formation is the development of a protoplanetary disk around a suitable star (or a binary star pair; a significant fraction of the planets found so far are [surprisingly] associated with binary or even ternary stars, thus increasing the likelihood of planetary systems being ubiquitous since more than half of galactic stars are of the binary mode). An inventory of detectable disks in nearby neighborhoods of the Milky Way found disks around 236 stars; the instrument used was the Spitzer Space Telescope whose Infrared sensors are adept at measuring hot gas and particles. As evident in the graph below, most of these disks are associated with younger stars but the distribution includes even older (>800 million years) stars:



The following is a generally accepted model (called "core accretion") for establishing a planetary system: A nebula is subject to gravitational irregularities and other perturbations that cause free-fall collapse to numerous clots around which surrounding gases and particulates usually adopt a disklike form. Over time, the disk tends to organize in spiral arms of gas and matter, which increasingly become disorganized by clotting (discussed below). Consider this generalized sequence:



Another example of these extrasolar planet-star formation gas/dust disks appears below; read the caption for details.



The influence of gravity, which builds up progressively as planetesimal clots enlarge, is the prime driving force promoting both planet and star formation. In some instances, shock waves from a Supernova can cause interstellar matter to initiate collapse and compress into protostars and debris orbiting them. Matter is also redistributed along magnetic field lines by magnetohydrodyamic processes. The main phases of planetary formation extend over about 10-20 x 10^6 years but it may require up to 10^8 years to progress from the early infall to the late T Tauri stage of a protostar's development. While a particular clot is organizing, the materials tend to redistribute such that Hydrogen and much of the lighter elements flow towards a growing center to accumulate in a gravitationally balanced sphere, the star. Under one set of conditions, instabilities lead to a double (binary) star pair. As protostars form, the rotating gases and dust particles collect in a spinning disk around each center and eventually organize by accretion into planets. The same process, with variants, works on single stars. The time frame for the above model suggests planets will appear within one hundred million years or less after the nebular gas and dust have begun to behave as a unit in space.

If our Solar System is the norm, inner planets should be rocky, with thin or absent atmosphere (lost from insufficient gravitational ability to retain the gases or by being swept away by the solar wind). Outer planets should have rocky cores and be less susceptible to loss of gases, so that their increased mass serves to gather in still more gases. However, the discovery that giant planets can lie quite close to their parent stars places this assumption of size distribution with distance into question.

Alan Boss of the Carnegie Institution of Washington has argued that the outer gas planets Uranus and Neptune have much less gas than would be expected from conventional planetary system models. He claims that large quantities of gas were driven away in the earlier history of these planets by UV radiation from nearby stars in the local cluster. It is reasonable to expect that stellar windw, UV radiation, and other "forces" from neighboring stars might affect planetary history but his hypothesis remains in dispute.

Two recent hypotheses are adding new twists to the above concepts. First, in addition to or in place of core accretion, another mechanism called "disc instability" may play an important, perhaps key role, in planetary inception. This is related to gravitational irregularities that can cause rather rapid accumulation of materials in the proto-planetary disc. Earlier-formed planets can contribute by setting up further instabilities. A second idea holds that planets can move inward or even outward in a form of migration or "wandering" so that their orbits change both in relative distance to their parent star(s) and in eccentricity.

But for the present, astronomers continue to build and refine their models on the much easier-to-make observations at the astronomically short distances within the Solar System. Like other stars, the Sun (whose diameter is 1,392,000 km [870,000 miles]) is an end-product of gravitationally-driven condensation and collapse of Hydrogen/Helium gases and associated matter (both solid and gaseous) consisting of other elements and compounds that once made up a diffuse (density ~ 1000 atoms/cc) nebula. Probably many stars were generated in the

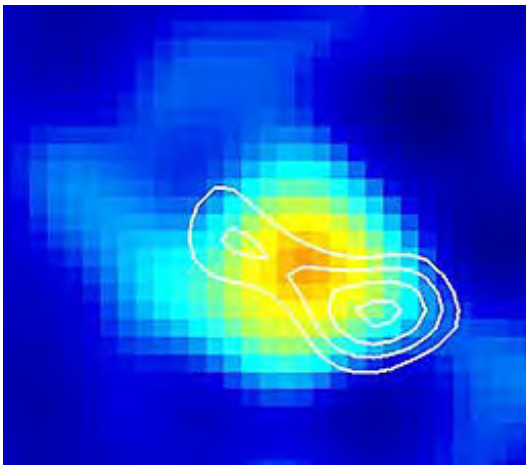
timeframe of a few hundred million years from this particular "cloud".

The protosun built up from centripetal, gravity-induced infall of nebular substances towards one of the concentration centers in the nebula. The bulk of the gases enters the resulting star itself along with much of the other materials, leaving an enveloping residue of matter enriched in Si, C, O (and H), N, Ca, Mg, Fe, Ti, Al, Na, K, and S (most organized into compounds, particularly silicates, that can be sampled by recovery of iron and stony meteorites - representing fragments of comets and broken protoplanets that are swept up onto Earth). This material, bound by gravity to the Sun but free to move inertially in encircling orbits, remained distributed in the space making up the Solar System. This system of particles rather rapidly organized into a disc-like shape whose present radius is about 100 A.U. (about 9300 million miles). The disc rotated slowly (counterclockwise relative to a viewpoint above the north celestial pole [which passes through Polaris, the North Star]), its motion influenced by external gravitational effects from nearby stars.

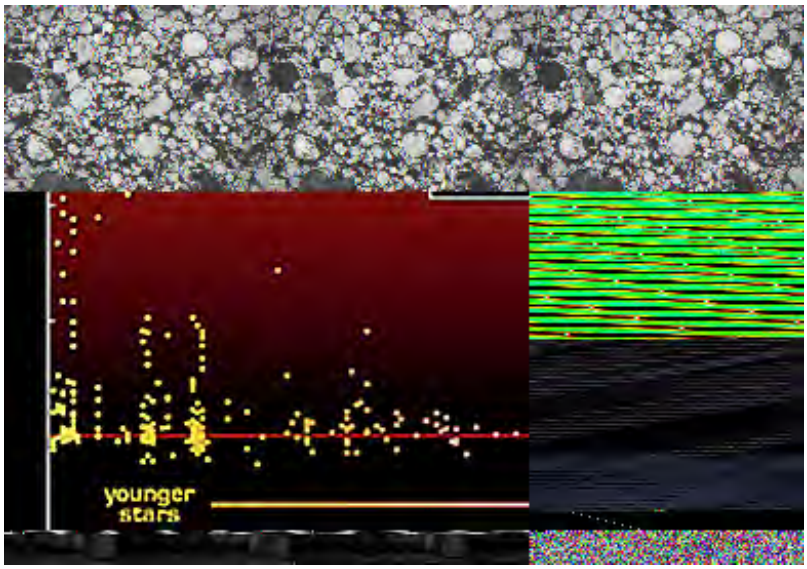
As this rotation got underway, and thereafter, the solar (stellar) magnetic field churned up the dust and gases (descriptively compared to the action of an "eggbeater" in a thin batter) causing them to collect into clots much smaller than the Sun that underwent various degrees of condensation. This field also expels and guides this material into jets that carry matter out to great distances, as seen here in this Hubble Space Telescope view of a jet ejecting from another star in our galaxy:

Both jets and irregular nebular patches (e.g., the Horsehead and Eagle nebulas shown above) contain not only gases but significant amounts of dust. The dust is very small and consists of three types: 1) core-mantle elliptical particles, typically 0.3 to 0.5 microns in long dimension, with a silicate interior coated by icy forms of gases; 2) carbonaceous particles (~0.005 microns), and 3) open frothy clots called PAH dust (polycyclic aromatic hydrocarbons) (~0.002 microns). Shock waves and radiation can strip off the ice mantle leaving grains that are incorporated into coalescing bodies that form the prototypes which accrete into the planetesimals from which asteroids or planets then build up. Ultraviolet radiation can modify the organics into more complex molecular forms. In this way, organic molecules are introduced from space onto planetary surface and, if conditions are right, can eventually serve as viable ingredients for the inception of living things (see below).

The possible role of shock waves in planetary formation is now the subject of considerable study. Evidence for a shock wave that develops as material falls towards a nearby protostar against its remaining gas/dust cloud has been observed at L1157, in which the present cloud is about 20 times the Solar System diameter. As this cloud proceeds to infall into the newborn star, it organizes into a disk and produces shock waves that may clump dust together, as described in the next paragraph. Here is this cloud:



For the Solar System, shock waves and intense radiation acted on the dust such that some of it melted into tiny droplets which chill into *chondrules*. These spherical bodies then were caught up with remaining dust to produce the primitive small solid bodies (fluffy "rockballs") that populated much of the heliosphere surrounding the Sun. Today we can analyze samples of these accreted bodies as *meteorites* which are small pieces of larger bodies torn loose from asteroids and put in orbits that eventually reach Earth. (You can review some basic knowledge of meteorites by clicking to page 19-2.) Most infallen meteorites are ordinary chondrites that, in thin section, appear much like this sample from the Tieschitz meteorite:



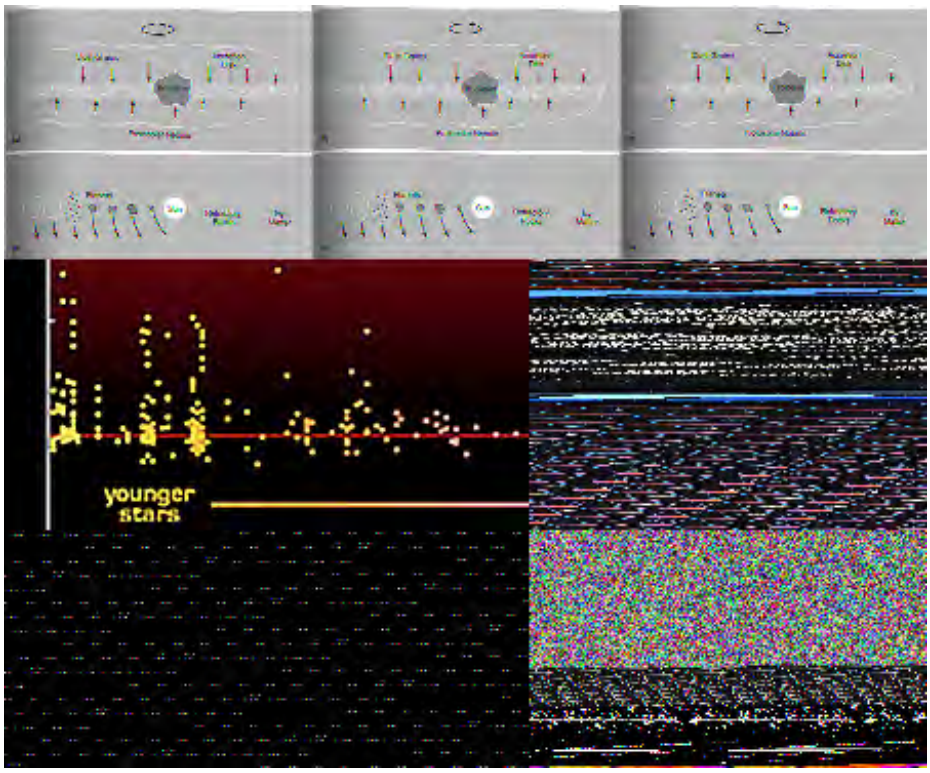
The most primitive meteorites, called carbonaceous chondrites, are enriched in Carbon and contain water. Other meteorites are iron-rich (some with > 90% metallic Iron), and may have once been the interiors of planetary bodies since disrupted. The chondrules themselves generally show a very limited size range, suggesting that ones larger than these fell back into the Sun through gravitational pull whereas smaller ones were swept away into interstellar space through expulsion by shock waves and solar wind.

Magnetically-driven eddies within the gas/dust cloud helped to impart additional angular momentum to the larger condensed rotating objects beyond the spherical Sun (which possesses only 0.55% of this momentum even though it contains 99.87% of the total mass of the system). These objects now remain in orbits around the Sun in positions that remain stable because of the counterbalance between centrifugal forces related to angular momentum and inward-directed gravitational pull from the Sun.

What are the currently favored models for planet formation? Two general models (mainly for formation of large planets with thick gas atmospheres) - Accretion and Gas Collapse - are popular now, and both may have operated. These models are shown in these two panel sequences:

Both models start with a gas/particulate cloud denser than its surroundings. In many instances, this cloud is a remnant of a supernova explosion. This is the case for the Solar System nebula, as evidenced by the higher concentrations of heavier elements - the normal end-product of the destruction of a star in which these elements have been synthesized.

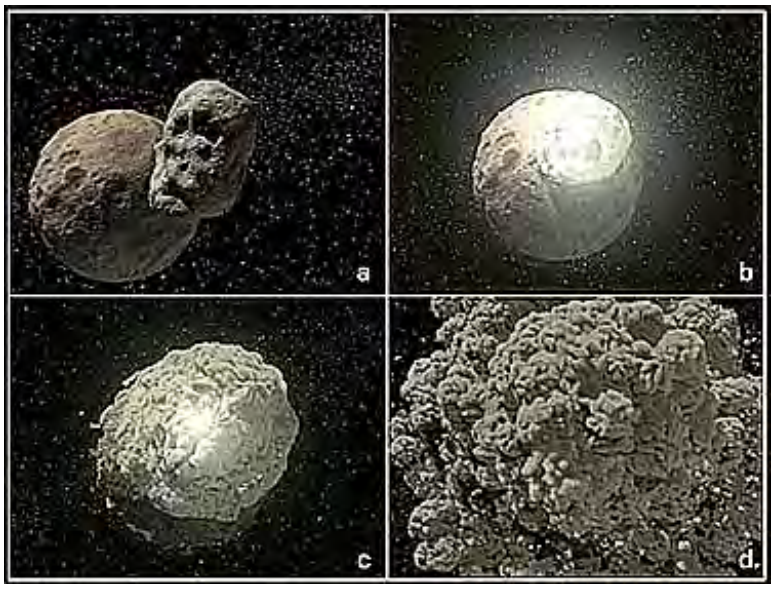
For the Accretion model, as the formative process operated, local instabilities in the nebula tied to the Sun caused the chondrule-laden rockballs within turbulent zones to cluster and further aggregate into objects ranging from meter-size up to *planetesimal* dimensions (tens to a few hundred kilometers, typified [perhaps coincidentally] by asteroid proportions).



From J. Silk, The Big Bang, 2nd Ed., © 1989. Reproduced by permission of W.H. Freeman Co., New York

During this growth stage, smaller planetesimals tended to break apart repeatedly from mutual collisions while larger ones survived by attracting most of the smaller ones gravitationally, growing by *accretion* as new matter impacted on their surfaces. Once started, "runaway" growth ensues so that many planetesimals combine into bodies that eventually enlarge into fullblown planets. The bulk of the matter beyond the Sun was swept into the planets and their satellites, although some remains in comets and cosmic dust. Mercury, and some Outer Planet satellites are preserved remnants of this later stage in planetary growth, as indicated by their heavily cratered surfaces that were never destroyed by subsequent processes such as erosion. In contrast, the Moon appears to have built up by re-aggregation of debris hurled into space as ejecta from a giant impact on Earth soon after our planet formed. Once collected into a sphere (which probably melted), the lunar surface continued to be bombarded with its own remnants as well as asteroids and other space debris. Its oldest craters are hundred of millions of years younger than the time at which the debris reassembled, melted and formed the lunar sphere; at least some of its larger basins are somewhat older.

The formation of the Moon by collision between two separate but large planetary bodies likely was not a unique event in Solar System history. During the earlier stages of accretion more planets than now exist probably formed. Their number was reduced by a collision which could have caused both bodies to be disrupted into particles of varying size in a "collision cloud" which then reassembled into a single planet whose mass was approximately that of the two earlier bodies (some matter was lost to space). The asteroid belt between Mars and Jupiter may be an example of collisional debris that failed to reorganize into what would have been the fifth rocky planet.



As mentioned above, in our Solar System, the four inner planets (the *Terrestrial Group*) are largely rocky (silicates, oxides, and some free iron; three with atmospheres) and the outer four (*Giant Group*) are mostly gases with possible rock cores. These Giants developed large enough cores to attract and capture significant fractions of the nebular gases dispersed in the accretion disk.

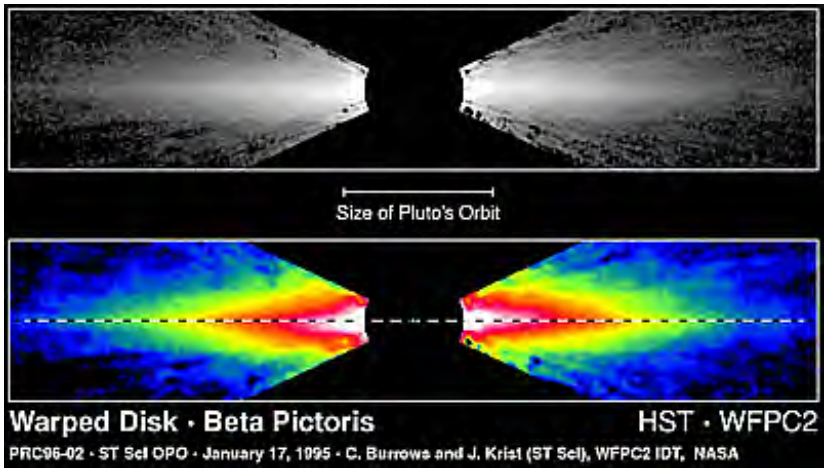
Analysis of Argon, Nitrogen, and other gases in Jupiter indicates their amounts are such that this Giant must have formed under very cold conditions; if further work bears this out, Solar System scientists may adopt, as one plausible explanation, an origin of Jupiter (and perhaps the other Giants) at much greater initial distances from the Sun with these having since moved significantly closer through orbital contraction or decay. The Dwarf planet Pluto, the smallest and, at times, farthest out (its elliptical orbit periodically brings it within that of Neptune), appears to be made up of rock and ice and may be a captured satellite of Neptune.

Theoreticians differ as to the exact methods and sequence in which the planets accumulated after the condensation and planetesimal phases. Timing is a critical aspect of the formation history. One version - the *equilibrium condensation* model - considers condensation to happen early and quickly, in a few million years, with the observed sunward zoning of higher temperature minerals and greater densities in the rocky inner planets both being consequences of the increasing temperature profile inward across the solar nebula. Accretion was stretched out over 100 million years or so. The *heterogeneous accretion* model holds condensation and buildup of planetesimals to proceed simultaneously over a few tens of millions of years. Neither model adequately explains the fact that both high and low temperature minerals aggregate together in the inner planets to provide materials capable of generating the atmospheric gases released from these planets. The models also do not fully account for the strong preferential concentration of Iron and other siderophile ("iron-loving") elements in the inner, terrestrial planets. One solution is to add (by impact) low temperature material to the growing protoplanets carried in along eccentric orbits from asteroidal and Giant planet regions. This material is then homogenized during the total melting assumed for each inner planet early in its evolution. This melting is the consequence of heat deposited from accretionary impacts, from gravitational contraction, and from release during radioactive decay. As cooling ensues, materials are redistributed during the general differentiation that carries heavy metals and compounds towards the center and allows light materials to "float" upwards towards the surface.

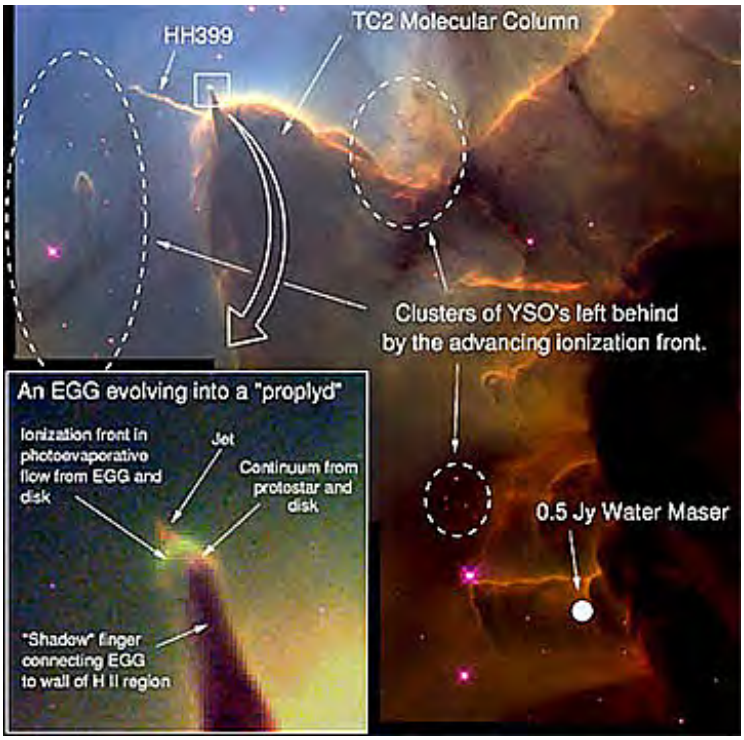
Less is known about the long-term evolutionary history of planets and their eventual demise (destruction). Extrapolating from our Solar System with its two major types of planets - Rocky and Gaseous - and the variety of surfaces on them and their satellites, it is evident that a great range of sizes, compositions, and surficial states can be expected among the millions of planets that many believe exist in the Universe. In the Solar System its complement of planets have survived essentially intact (possible exception: the asteroid belt) since the Sun itself organized

some 4.5 to 5 billion years ago). The Sun is expected to burn out its fuel in another 5 billion years, when it expands rapidly into a Red Giant. The outward surge of its gaseous envelope should consume many - maybe all - of the named planets as well as other solar material. This is probably the usual mechanism of most planetary destruction - consumption by Red Giant expansion or by Novae or Supernovae (see top of page 20-6). Another possibility: gravitational pull brings the planets into their parent stars. Generally, planetary systems around massive stars, if indeed these do form, will be short-lived as those stars themselves do not last billions of years (thus, such stars are not likely to harbor life since not enough time elapses to permit development by evolution [see below]). Smaller stars, such as G types, are much more favorable bodies for fostering life on any planets that may revolve about them, owing to their longer spans of existence.

There is a second form of dust around stars that has been produced *after* they were formed and well along their trail of evolution. This is described in the April, 2004 edition of Scientific American in an article by Davod Ardila entitled "The Hidden Members of Planetary Systems". (Ardila points out that not all stars with this kind of dust necessarily have associated planets). The dust is of two types: 1) micron-sized particles that are analogous to the dust in the solar inner planet belt that gives rise to Zodaical light at sunset; 2) dust of a range of sizes that exists further out (beyond Jupiter and the Kuiper belt for our system but for some stars the disk extends out to notably greater distances). The zodaical dust is produced by release from comets and grinding of asteroidal-sized bodies that collide and abrade over time. The larger particles tend to spiral into the parent star, the smaller are pushed away from the star by radiation pressure. Over time, the amount of dust will diminish. But some of the dust may be incorporated in planets within this circumstellar debris cloud, already formed or yet to form. The temperatures associated with dust belts and clouds varies, so that telescope sensors will pick up measurable EM radiation at different wavelengths. One of the best examples of a huge dust disk is found around the star Beta Pictoris, 63 light years from Earth. The disc extends out about 1100 A.U. (about a 460 billion kilometers in diameter). There is a suggestion of one or two planets within the disk. The Visible light HST image below shows the symmetrical disk (lower image is colored to indicate density differences); the black center is due to screening out the star itself using a coronagraph accessory on the telescope.



Recently, another plausible model for the origin of planets has been reported by Dr. Jeff Hester and colleagues at Arizona State University. The figure below is relevant:



As with the other models, clouds of Hydrogen gas and silicate dust are needed. Shown here is the Trifid nebula. Within it are now being formed a range of embryonic stars which include besides those of Earth-size, more massive stars that explode. The astronomers studying this "nursery" of stars point out that massive stars can produce the isotope of Iron Fe^{60} whose half life is about 1.5 million years. Its stable daughter product Ni^{60} has been found in meteorites, whose parent sources presumably formed along with the stars like the Sun. This implies that the Sun was born out of a gas-debris cloud that had been enriched by radioactive Fe^{60} from one or more exploding (Supernovae) stars in its neighborhood. These stars were much more massive than the Sun. The nebulae like Trifid, e.g., Eagle and Orion looked at earlier in this Section, are enriched in HII (doubly ionized Hydrogen). As shock waves produce ionization of the Hydrogen, YSOs (Young Stellar Objects) will form at various sizes. In the ongoing process of star formation, EGGs (Evaporating Gaseous Globules) develop and evolve into associated propylids. The propylids later shed some of their material leaving stars on the Main Sequence of sizes similar to the Sun.

Little has been said about life on planets on this page - this will now be reviewed on the next page, 20-12. However, a recent article by Beer, King, Livio, and Pringle in the Monthly Notices of the Royal Astronomical Society has put forth an argument that life must be rare. This is deduced from the observation of the ~200 planetary systems so far discovered. Nearly all of these apparently have only Giant gas ball planets that are much closer to their central star that Jupiter through Neptune in our Solar System. If this turns out to be the case then rocky planets are scarce - those closer to their stars are prone to having much of their gas envelopes blown away by stellar wind and high temperatures proximate to the star. The flaw (and saving grace, for those who want life discovered elsewhere) in the argument is that none of the current methods of planet detection are capable of finding smaller planets but are biased towards locating big gas balls.

While most planets are believed to have rocky cores, stellar wind and explosion processes tend to blow off gases from smaller inner planets. The essential conditions needed for organic molecules to develop are water (preferably in liquid form, but life in steam or ice is believed possible), an appropriate temperature range, some semblance of a favorable atmosphere (but anaerobic or Oxygen-free environments on Earth can contain life), and the appropriate ingredients (C, H, O, N, and P; an Si life system, instead of C, is theoretically possible). We shall see on page 20-12a that the Drake equation provides a mathematical means of estimating the opportunity for organic molecules to form on planets in some fraction of the star systems.

Likewise, the likelihood for life to occur on planets seems to follow the Goldilocks dictum (page 20-11a): "not too hot, not too cold, just right".

Having postulated that planets are probably rather commonplace in the Universe, let us study on the next 3 pages the types of and conditions for life having formed on Earth by processes becoming ever better understood. After that we will return to the topic of planetary systems, now within the framework of life as we know it on Earth and suspect it on other planets outside the Solar System.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Nature and Origin of Life on Planetary Bodies (Continued)

After the review of the nature of organic molecules and of life itself, and our diversion with Evolution, we turn our attention next to evidence of organic molecules in interplanetary space (including comets and asteroids) and, at a much larger spatial level, in galaxies and interstellar clouds. From both direct sample analysis to direct remote sensing (with observations in the UV, Infrared, and Radio [telescope] wavelengths providing much of the detections of these molecules), a fair amount is now known about organic species beyond Earth in the solar, stellar, galactic, and intergalactic regions of space.

Regarding the Solar System, some information has already been introduced in Section 19, especially on pages dealing with Europa, Titan, and asteroids and comets. Overall, there seems to be a wide abundance of organic molecules, but so far none can be unequivocally equated with molecules that are or were once alive. The best evidence of the diversity of organic molecules is found in certain meteorites, mainly those classed as Carbonaceous Chondrites (CC). One in particular has proved a "standard" for direct proof of a broad variety of molecules. This is the Murchison CC which struck in 1969 north of Melbourne, Australia and was found within a day, allowing rapid sampling of the 100 kg collected that minimized contamination. Here is a piece of Murchison:

Murchison has been found to contain Carbon (as graphite) amd more than 70 extraterrestrial amino acids and several other classes of compounds including carboxylic acids, hydroxy carboxylic acids, sulphonic and phosphonic acids, aliphatic, aromatic and polar hydroCarbons, fullerenes, heterocycles as well as kerogen, Carbonyl compounds, alcohols, amines and amides (in all, about 250 hydrocarbon species). Most of the amino acids do not have terrestrial counterparts but 8 of the 20, principally glycine, that make up proteins were in the group; no RNA or DNA has yet been found in meteorites or anywhere else in the Universe.

Peculiar structures in Murchison are provocative in their resemblance to primitive microfossils (compare with ALH84001 on page 19-13). Consider this slice of Murchison:

The same year that Murchison was discovered, a brilliant fireball over the state of Chihuahua in Mexico led to the discovered of hundreds of fragments of a carbonaceous chondrite. Known as the Allende meteorite (named from a nearby town), this is by far the most (in weight) of any known carbonaceous chondrite, thus providing abundant material for scientific study. It contained 16 species of amino acids (some not found on Earth).

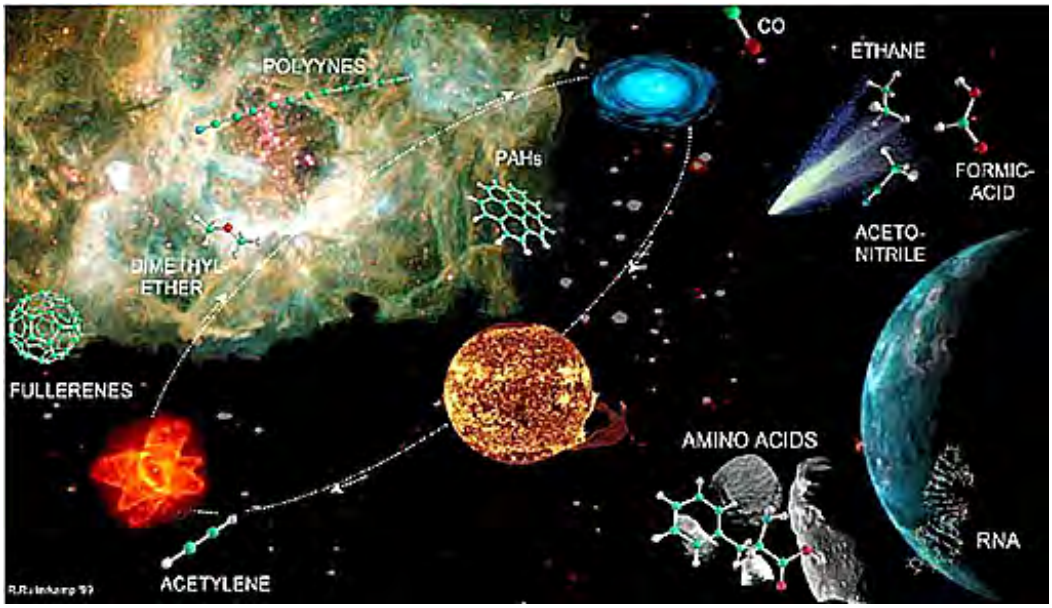
Murchison and Allende (and other meteorites) have a distinctive characteristic which bears on the origin of terrestrial life. Earth's amino acids have a "left-handed" chirality (the type of symmetry that is expressed as one holds his/her two hands together). It could just as well have had a "right-handed" symmetry. The vast majority of proteins in a terrestrial environment (including living things) also have left-handed amino acids (sugars, by contrast, possess right-handed symmetry). This strongly implies that the beginnings of life some 3.5+ billion years ago came about by synthesis of organic molecules that obtained their amino acids from extraterrestrial bodies such as meteorites, in which right-handed molecules have been found.

Other CCs have similar but fewer organic species. The Murray CC has simple sugar and sugar alcohol molecules. Some astrobiologists believe these CCs are cometary nuclei; others think they may have been part of the asteroid belt. Spectra obtained for Halley's comet show presence of Carbon dioxide, ammonia, and (with some uncertainty) purines and pyrimidines - the bases in RNA-DNA nucleotides. Water, HCN, formic acid, methane, and ethanol have been observed in the Solar System. The bottom line: the ingredients that were used in the Miller-Urey experiment are all present in the Solar System beyond Earth. If actual life has not developed in other parts of the Solar System, then at least those ingredients needed to produce life in Earth's

early oceans could have been introduced by infall of comets, asteroids, and meteorites. (If life itself is extraterrestrial, this notion is embodied in the term "panspermia".)

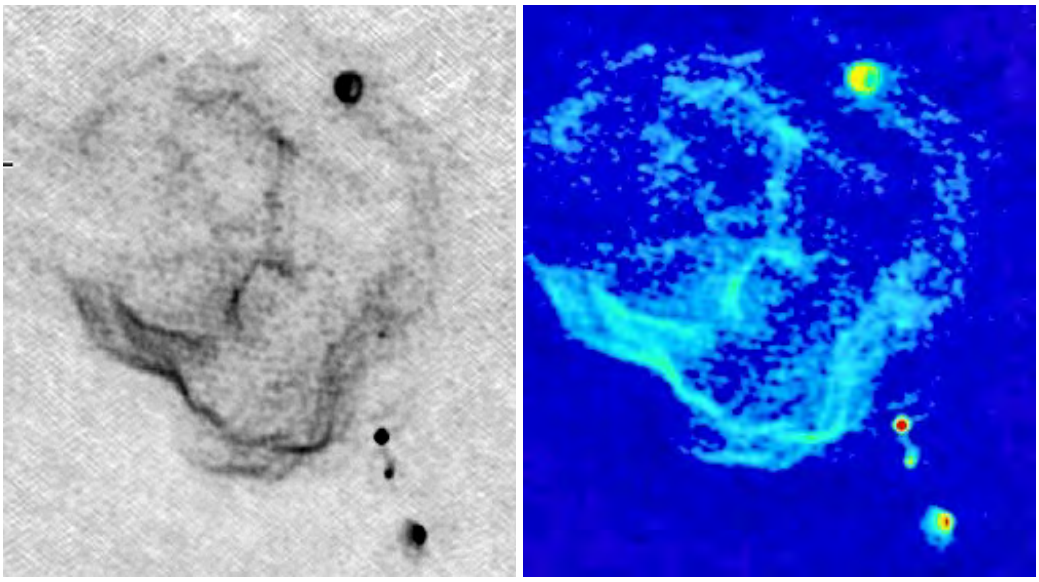
Where once the Miller-Urey experiment had provided apparently definitive insight into the origin of life on Earth as a purely terrestrial event, the discoveries in the carbonaceous chondrites offer a tantallizing alternative - at least some of the primitive building blocks of life may have first originated in extraterrestrial environments. Those who favor this also point out that the CCs contain water. This suggests that the water in the Earth's surficial locations - mainly the oceans - could also have an extraterrestrial origin. (But degassing of the melted Earth, an earlier hypothesis, remains viable as a source of much/most(?) of the primordial water, provided that didn't escape into space.)

While organic molecules in the Solar System remain of high interest to exobiologists, their presence beyond our star's spatial confines is also of considerable significance, since this would demonstrate the universality of conditions that produce both organic molecules and possible organisms. Telescope data-gathering is proving effective in verifying the presence of organic molecules in protplanetary disks, around certain stars, and, most frequently, in stellar molecular clouds. This (rather fanciful) diagram suggests the variety that has been confirmed or may be expected to be discovered in future observations:



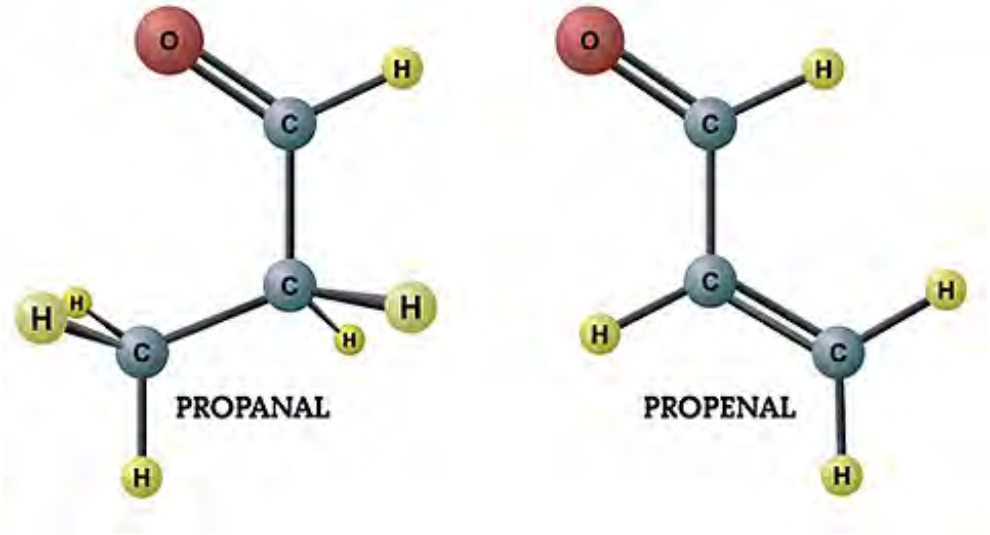
The most effective instrument in searching for signatures of organic molecules in this, and other, molecular clouds is the radio telescope, such as this 12 meter dish at the Kitt Peak Observatory:

Probably the most studied interstellar molecular cloud so far is Sagittarius B2, about 26000 l.y. from Earth, near the galactic center. Here is an optical telescope image (as a negative) and a Radio telescope view of this discrete object:

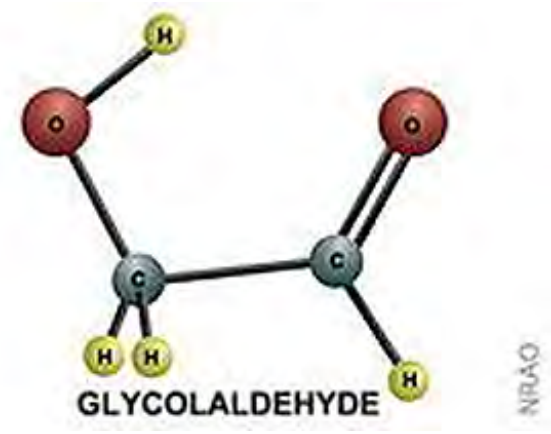


The average temperature of such a cloud is about 30° K. This seems quite low for active formation of organic molecules, implying that these formed (probably elsewhere) under much warmer conditions before being incorporated into the molecular cloud.

Propanal and propenal are two distinctive molecules found in molecular clouds, such as Sagittarius B2.



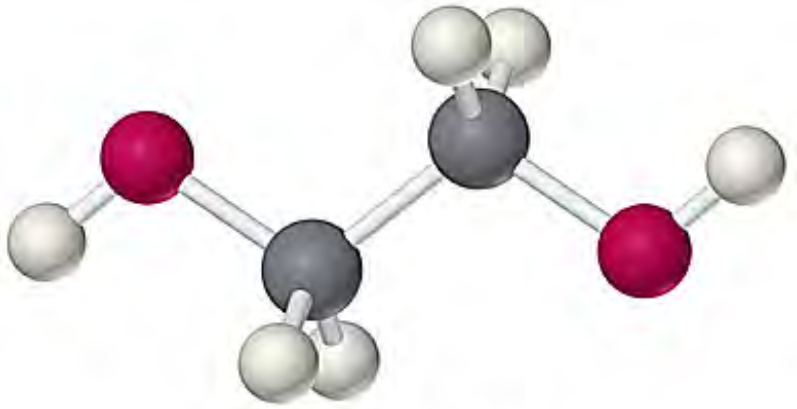
Recently, a two Carbon sugar molecule, glycoaldehyde, was discovered by radio astronomy data analysis of Sagittarius B:



The significance of this finding is that glycoaldehyde can react with propenal to form ribose, a molecule that is a central constituent of RNA.

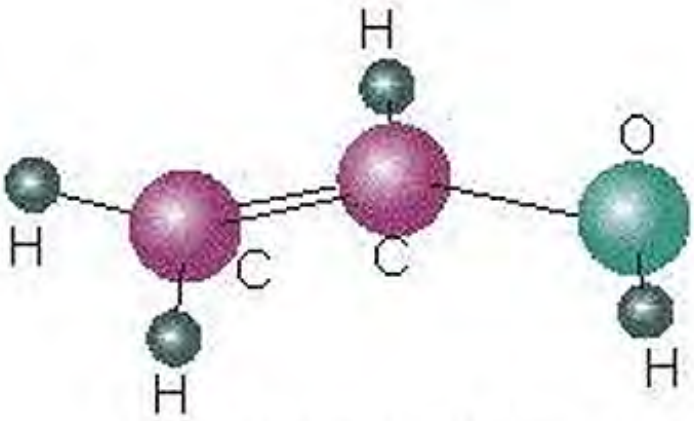
The list of identified organic molecules discovered in molecular clouds, protoplanetary discs, and other source presently exceeds 140. We will consider several examples below. But, if you want to have a fuller itemization, check out this Web site: National Radio Astronomy Observatory (includes data on a comet analysis). From this site, one can visit a Primer on the spectroscopy of organic molecules in molecular clouds and stars by H.A. Wootten of the University of Virginia.

Among the more frequent constituents important to organic molecular species are water, ammonia, acetylene, methane, and ethanol. Several species or groups are particularly interesting. One is ethylene glycol, which we know as the main constituent in one type of anti-freeze used in autos. Its formula, $C_2H_6O_4$ is shown in this ball model format:



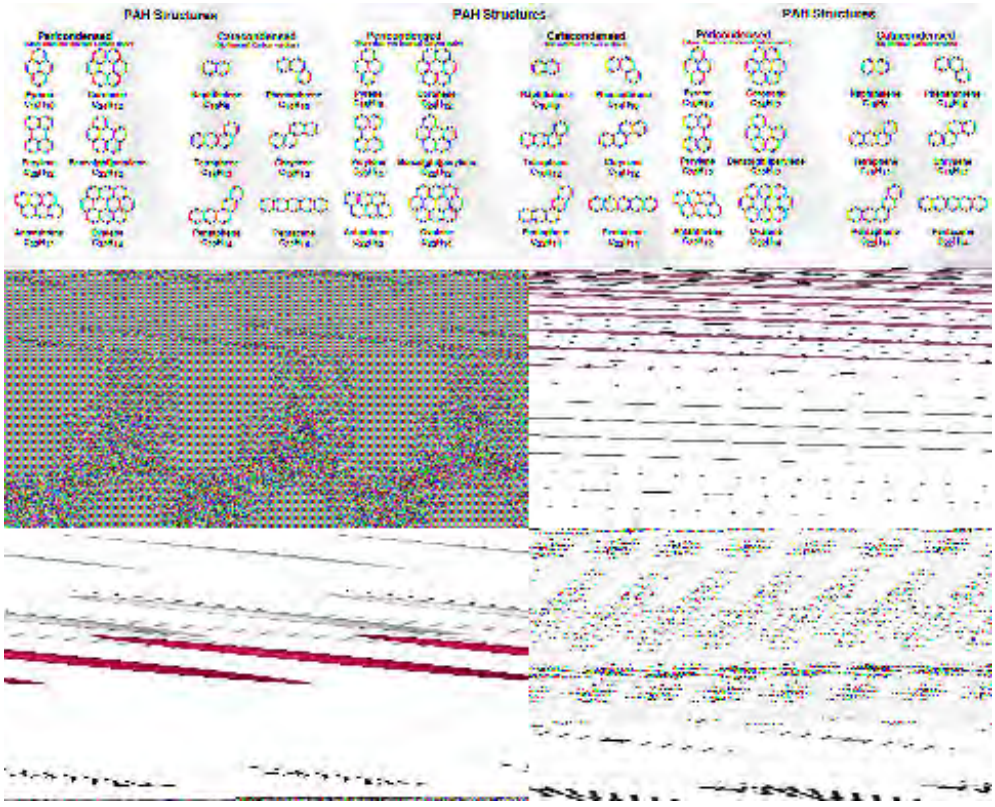
Ethylene glycol

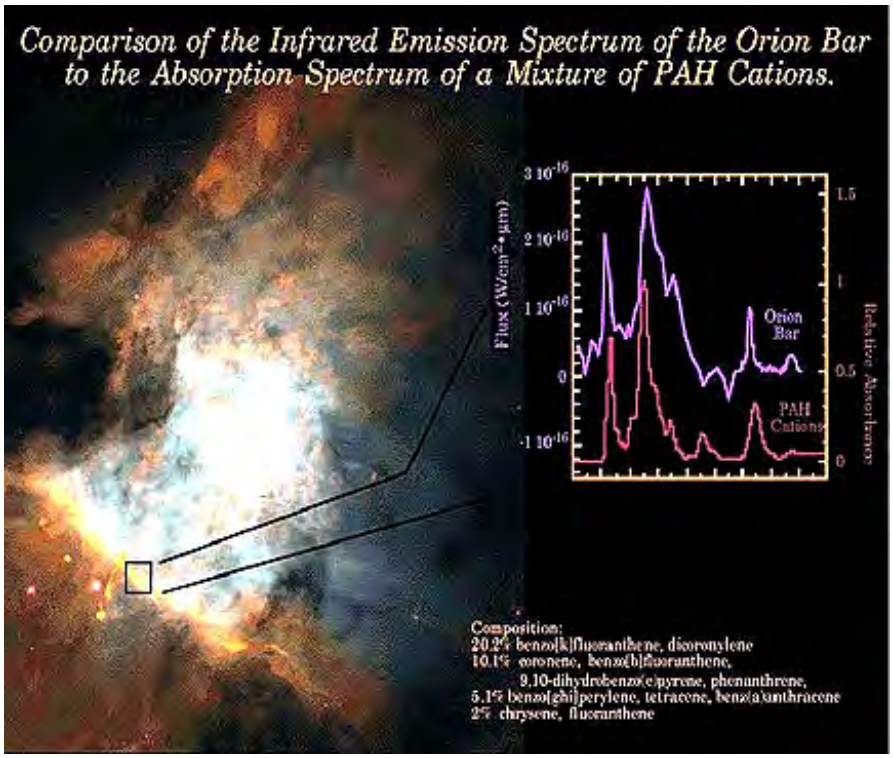
Vinyl alcohol is another common molecule:



Vinyl alcohol (syn)

On Earth, Polycyclic Aromatic Hydrocarbons, or PAHs, comprise the largest superclass of organic molecules. Most are built as ring structures (e.g., benzene). PAHs have been found in measurements made by the Spitzer Space Telescope in distant galaxies, some as old as 10 billion years. Here is a set of structural diagrams of some common PAHs that have been found in interstellar molecular clouds (lower illustration).





From the foregoing, one vital conclusion can be drawn: If indeed there is life elsewhere in the Universe, it is highly likely that most forms will be primitive (compared with Earth), prokaryotic, and small (microbial, requiring high magnification tools [e.g., electron microscope] to examine if samples are ever recovered). This seems probable for Mars, for Titan, and for Europa in our Solar System. Despite this likelihood that life forms will on many planetary bodies (those that can support them) be small and not advanced in the evolutionary chain, we as humans retain a hope - almost a compelling desire - that intellectual forms exist elsewhere.

Driven by this last incentive, Science's attempts to determine whether life can exist throughout the Universe generally have followed the same approach as planetologists use to study the other planets of the Solar System - we use our knowledge of Earth to provide an information base and a methodology for comparative study of these planets. Thus Earth is the *reference standard*. Earth had the right conditions that permitted life to develop some 4 billion years ago (either spontaneously from chemicals derived from gases expelled from its interior and/or from organic materials brought in by meteorites, etc from elsewhere in the Solar System (panspermia concept). In general, in a planetary system the most critical factors are size of the central star (too big restricts the time available for evolution to work), the star's composition (a multi-generation star containing Carbon in its makeup and in the gas/dust cloud from which it is born), the size of the planet (large enough to retain an atmosphere; small enough not to develop a very thick atmosphere with poisonous gases) and distance from the stellar parent (so that thermal energy is available but temperature range falls above freezing of water but below its conversion to steam). The optimal conditions for life to appear seem to be a hydrosphere, an atmosphere, a critical range of temperatures (the "Goldilocks dictum" - not too hot, not too cold, just right), and appropriate incredients.

With the background gained from the above paragraphs, let us return to our more general speculations on life throughout the Universe, using chemical and other non-biological evidence: As hinted at in previous paragraphs, an underlying and perhaps critical condition most pertinent to possibilities for life elsewhere in the Universe is associated with the concept of the *metallicity* of the star groups. Metallicity defines the proportion or percentage of chemical elements with atomic numbers above 2 in the total mix of elemental and molecular gases and dust available for star formation. Smaller stars with metallicities between 60 and 200% of the Sun's value are favored. These stars comprise about 20% of the total in a galaxy. Other, more negative, factors include the frequency of Supernovae (which over time build up the supply of the "metal" atoms),

excessive radiation (continuous or in bursts), and the distribution and numbers of objects capable of destroying protoplanets by impact. Giant planets seem less likely to foster conditions that would aid life's establishment. They conclude that 1) most stars don't have planets and 2) complex life is rare even on those stars with planets. While they don't propose that Earth is unique, they do caution that the statistical distribution from their GHZ and CHZ models support the notion that it may prove very hard to find evidence of life of any kind either in the Milky Way or (even more so) in more distant galaxies.

If Earth defines the standard state, from which there can be no major deviations if life is to form and survive, then life-supporting planets are constrained to vary in their physical and chemical properties only within a narrow range. A planet must have accessible and reactive Carbon capable of polymerizing (some have postulated an alternative element, silicon, as the keystone in complex life-sustaining molecules but no such compounds have been successfully synthesized and made to function like Carbon life). It appears (but is not totally certain) that water is also essential. If so, in an envIron like Earth, this limits the range of temperatures at which life *originates* to water's freezing and vaporization (boiling) point values of 0 to 100 °C (these values are changed somewhat by prevailing pressures). However, life, once formed, has been found to exist at temperatures that are higher and lower, but generally in the presence of liquid water; witness, the "smokers", which host symbiotically specialized life, that eject superheated water and steam on active divergent ridges on the ocean floor. Atmospheres of Oxygen and Nitrogen favor many forms of life; planets that are either airless or contain hostile chemicals such as methane and sulphur compounds tend to suppress life.

If such conditions, and their ranges, are indeed limiting factors, then only a few, if any, planets in a given planetary system are properly suited to the origin, development, and persistence of living creatures. Thus, while billions of planets are potentially existent universally today, only a small fraction are suited to supporting life. These will be confined to those at a distance from their star where temperatures are in appropriate ranges to allow **water** in a suitable state (not chemically bound or heated so that all has evaporated and escaped to space). They will have proper sizes to maintain fostering atmospheres. Their chemistry will allow production of molecules (most likely Carbon-based) that can, over time, evolve into organic ones of sufficient complexity to merit the status of "living".

Water has been detected in the Solar System mainly on Earth, on Mars, on icy satellites, and in comets. A search for this compound beyond the Solar System has finally met with success. Astronomers have now detected water around CW Leonus, a Carbon-rich star in its waning life, some 500 l.y. from Earth. This water is believed to now be vapor derived from billions of comets about the star, as it rapidly releases its heat during an explosive phase.

The astronomer Carl Sagan addressed this question of how remote sensing similar to what we have described in this Tutorial could be used to detect life on another planet. His reasoning was based on how such sensors actually detect life on Earth (at the time of his writing an article for Scientific American in 1994, resolution of civilian space images generally could not clearly identify the works of Man, so he confined his surmise to looking at non-human indicators). He pointed out that a Landsat TM type sensor orbiting another planet could detect water, Oxygen, Carbon dioxide, chlorophyll absorption, and the bright response of vegetation in the Near-IR. Today, a high resolution hyperspectral sensor could do these things much better.

Now, to a topic of great popular interest (especially as it ties in with our fascination with UFOs). One of the prime motivations that has stemmed from space exploration continues to the the **S**earch for **E**xtra**T**errestrial **I**ntelligence (SETI) - which is a much higher goal than simply seeking evidence that lower levels of life exist elsewhere. This has so far been something of an *ad hoc* effort by a small number of dedicated astronomers who have had limited support from private sources. Now, NASA and other large organizations have become involved and a more concerted and systematic hunt for advanced life forms is being funded.

In fact, the hunt has turned *serious* in that significant funding is now being poured into the

search. The best hope still remains to seek out radio signals that have some non-random and intelligible pattern. In Carl Sagan's "Contact", a book adapted into a movie, the heroine, an astronomer (played convincingly by Jodie Foster), was using an array of radio wave receivers in New Mexico that actually has been functional for several decades. SETIites are now awaiting the full deployments of a much larger array, up to 350 dishes, each 6.1 meters in diameter, at the Hot Creek Radio Observatory site 470 km (290 miles) northeast of San Francisco. This is known as the Allen Telescope Array (ATA), which will produce a Very Long Baseline configuration and will initially monitor the centimeter wavelength region. The first 42 dishes are now being emplaced. Here are an artist's drawing of an individual dish and what the full array may look like:

Astronomers have now selected the first five stars to be monitored by the ATA, as listed here:

* beta CVn, similar to the Sun, 26 light years from Earth, in the Constellation Canis Venatui (Hound Dog).

* HD 10307, another solar analogue about 42 light-years away. It has almost the same mass, temperature and metallicity of the Sun. It also has a benign companion star.

* HD 211415, about half the metal content of Sun and a bit cooler, this star is in just a little farther away than HD 10307.

* 18 Sco, a popular target for proposed planet searches. The star, in the constellation Scorpio, is almost an identical twin to the Sun.

* 51 Pegasus. Already famous. In 1995, Swiss astronomers reported they had detected the first planet beyond our Solar System in orbit around 51 Pegasus. An American team soon verified the finding of the Jupiter-like object and the rush to find more extra-solar planets was on.

On the drawing boards, but with the possibility of postponement or cancellation owing to budget problems, is TPF (the Terrestrial Planet Finder), an advanced and specialized space telescope. Its top five candidates for the planetary search are:

* Epsilon Indi A: Turnbull's top TPF mission choice; this star is only about one-tenth as bright as the Sun and about 11.8 light-years away in the constellation Indus.

* Epsilon Eridani: This star is a bit smaller and cooler than our Sun; it is located about 10.5 light-years away in the constellation Eridanus.

* Omicron2 Eridani: A yellow-orange star about 16 light-years away that is roughly the same age as our Sun.

* Alpha Centauri B: This triple star system is located just 4.35 light-years away and one of the Sun's closest stellar neighbors.

* Tau Ceti: This star is a G-class star and is in the same brightness category as the Sun. Despite being relatively metal-poor, it is long-lived enough for complex life forms to evolve.

The goals of this mission are to answer these questions:

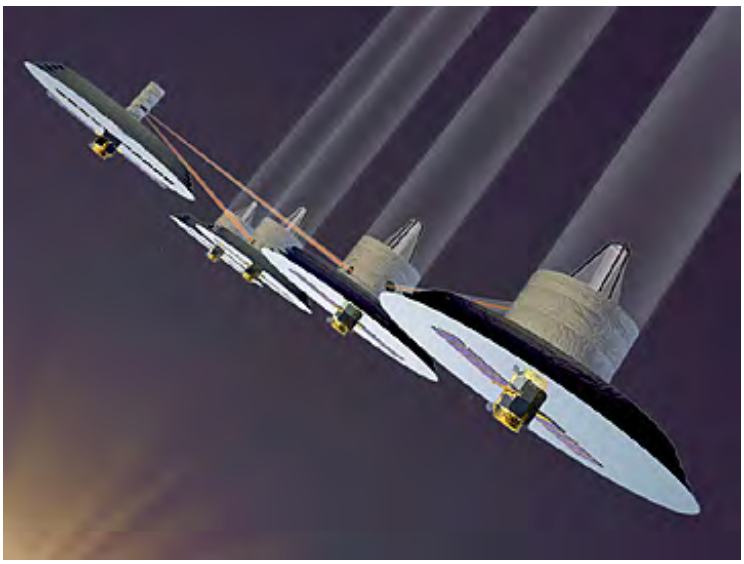
* Are there Earth-like planets in the "habitable zones" around their parent stars where the surface temperature is capable of supporting liquid water over a range of surface pressures?

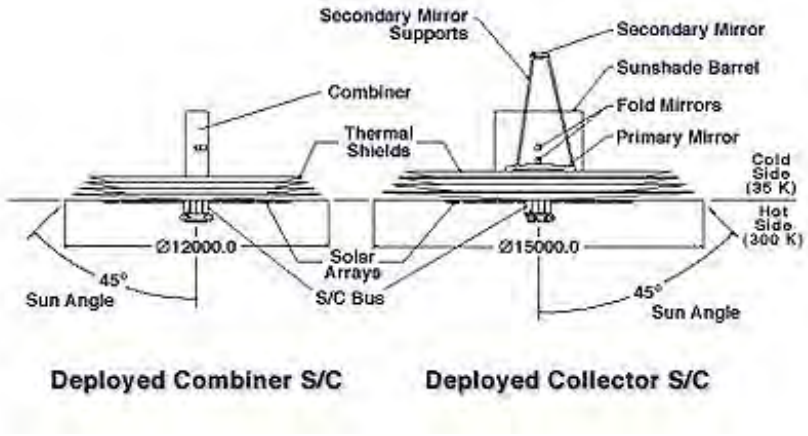
* What are the compositions of the atmospheres of terrestrial planets orbiting nearby stars? Is water, Carbon monoxide, or Carbon dioxide present?

* Are there atmospheric components or conditions attributable to primitive life, such as ozone or molecular Oxygen, seen in the Earth's atmosphere?

* How do planets form out of disks of solid and gaseous material around young stars?

The TPF at present will consist of four individual spacecraft that remain near each other as they 'formation fly'. They will sense in the Infrared and use Interferometry to obtain the signals. These two illustrations indicate the concept:





Writer's comment: Finding life elsewhere in the Universe ranks high among the most important tasks (opportunities) open to society. One hopes that Congress and/or the general public will clamor for more missions like this (which at the moment is facing the "axe") and opt for less capital spent in foreign martial involvements.

The ways in which SETI now seeks evidence for intelligent life are diverse and innovative. This is a very daunting and intriguing subject that could warrant considerable coverage space on this page but reluctantly must be confined to a synopsis of a few key ideas. However, we choose to guide you to several sites on the Internet for many of the omitted details. The starting point is the SETI site itself. The SETI Institute is currently being directed by Dr. Frank Drake, the originator of what is now known as Drake's Equation. Here is another Internet site that discusses in some depth that equation: Drake Equation.

For the record, we now state the Drake Equation (in its "dimensional analysis" format) and add some comments on values used in its terms (you can choose your own set of value in (2) above to see how changes affect the outcome):

$$N = R\, f_p\, n_e\, f_l\, f_i\, f_c\, L$$

where,

N = the number of communicating civilizations in the Universe.

R = the rate of formation of stars of types around which planets can form

n_e = the number of Earth-like bodies in the planetary system

f_l = the fraction of planets with life

f_i = the fraction with intelligent life

f_c = the fraction that has developed interstellar communication systems

L = the lifetime (span) of civilizations (up to extinction)

Of course, none of these parameters has yet been fixed with reasonably certainty, so that many values (and ranges thereof) have been proposed. One common set (but still provisional) has R = 10; f_p = 0.5; n_e = 0.2; f_l = 0.2; f_i = 0.2; f_c = 0.2, and L = 50000 yrs (source: Article "Why ET hasn't Called", by M. Shermer; Scientific American, August 2002). For our galaxy, this set of values give N = 400 civilizations.

Each of these term inputs is easily challenged. For f_p, any number chosen would depend on such variables as the total number of stars in a galaxy and the type of star suited to planetary

formation (usually limited to G and K types) and its percentage of the total population. Recent estimates center around 10% (factor = 0.1). At the 2001 Annual Meeting of the American Astronomical Society, Dr. J. Bally of the Univ. of Colorado presented an argument which concludes that only about 5% of the stars in the Universe are capable of producing (surviving) planetary systems. Massive stars will blow away the gas and dust needed for planets to form. Binary or ternary star systems, which are the most common arrangement, also are unfavorable for planetary growth, especially since one of the pair or trio is likely to be a Giant, and matter is collected as jets owing to attraction. He concludes that planets need to form soon after a star is born in order to have a reasonable chance for survival.

Plausible arguments can be made for values/ranges of each of the others. For example, f_l will be sensitive to such related variables as the presence and importance of water, the nature of the evolved atmosphere, and the time needed for higher order life forms to evolve (relative to planet age; rate can vary as these others assume values other than that of our Earth). The possibility of non-Carbon-based life forms must be factored in. In Shermer's article, he calls attention to the great uncertainty of L, the time over which any intelligences will persist before extinction. Pessimists feel that this number can be small. For Earth, civilized life is only about 5000 years old (based on the time at which agriculture and earliest urbanization become practiced). With the advent of the atomic bomb, these doomsayers consider that total mass destruction may be likely. On the other hand, one can conceptualize human society as overcoming its self-destruction tendencies and lasting (into the future) for millions of years (the upper limit then may relate either to a catastrophic impact or the stage in which the Sun burns out and expands to envelop the inner planets).

Still another factor that is not stated in the Drake Equation (or commonly discussed online) is the nature and strength of the communication signal. Present-day radio waves are probably too weak to have much effect in all but the nearest part of our galaxy. Higher energy waves (such as Gamma radiation) would be more powerful but we on Earth have not yet devised suitable transmitters that efficiently send out these rays. And any signal transmitted must move away in many directions from it planetary (spherical) source; if only directional beams are emitted, the number of planetary receivers (made by other intelligent recipients) in the right position to pick the signals will be greatly reduced. But, directional beams (those using laser light are especially promising) have one advantage - they remain concentrated over a small angular volume. As we on Earth continue to find many more planets (most would be in our Galaxy at this stage of our detection capabilities), we can systematically send signals to these with a good chance to intercept them in the narrow field of view encompassing our signals. Likewise, any intelligent and technically capable life on any one(s) of these may have by now spotted our Solar System and have, or will, send signals to us. What the "message" should be, owing to uncertainties of language recognition and translation, is probably dictated by the need to transmit something of a universal nature: sending intermittent signals (sound or light) that consist of a series of prime numbers is a favorite suggestion. Since mathematics has a unifying generality to it - all intelligent beings should find some of the same fundamental theorems and expressions - the message we receive (or send) is most likely to be in that format (spoken/written anguage is ruled out because each one on Earth has a unique set of meanings attached to words that therefore precludes universality). Or, the message could be made up of musical notes (example: "Close Encounters of the Third Kind").

Start with hypothetical observers at two points A and B not then in contact in early spacetime. Over expansion time, their light cones would eventually intersect, allowing each to see (at time t_1) other parts of the Universe in common but not yet one another. At a later time, beyond t_2 ("now") in the future, the horizons of A and B (boundaries of the two light cones) will finally intersect, allowing each to peer back into the past history of the other.

All in all, we are still a "far cry" removed from having any reasonable estimate of probabilities of other civilizations actually extant. Shermer develops arguments that produce numbers as low as 2 to 3 for our galaxy (if really strong signals can reach Earth from other galaxies, this number

would then greatly rise). T.R. McDonough of the Planetary Society arrives at 4000 as a reasonable "guess" for the number of planets in our galaxy; Carl Sagan during an optimistic moment, came up with 1,000,000. Today, no one is arriving at zero, so those seeking ET can remain hopeful, even optimistic.

In October 2001, the Scientific American magazine carried a review article (pages 61-67) by G. Gonzalez, D. Brownlee, and P. D. Ward bearing the provocative title of "Refuges for Life in a Hostile Universe". We will not paraphrase its many intriguing statements and conclusions but urge you to track the article down and read through it. Its bottom line is that there appears to be only a narrow range of conditions around stars and within galaxies that would likely harbor life (organic matter; not necessarily intelligent). They define CHZ and GHZ as Circumstellar and Galactic Habitable Zones respectively. In systems like our Sun's this is confined to a narrow inner zone. In spiral galaxies the GHZ is beyond the inner bulge, halo, and thick disk that consist mostly of old stars; the more favorable region is an annulus about midway from the center to the edge defined by the spiral arms, in that part known as the *thin disk*.

A splendid review of the origin and occurrence of life beyond Earth is the 1999 book by Paul Davies, *The Fifth Miracle: The Search for the Origin and Meaning of Life*, Simon & Schuster.

Whether the evolutionary mechanisms found to operate on Earth - namely, change into diverse and usually more complex forms in response to changing conditions, aided by genetic processes and natural selection - lead to intelligent life elsewhere can only so far be the subject of speculation (devoid of any direct evidence). But, again, considering the large number of favorable planets - almost certainly in the millions - spread throughout the billions of galaxies, it would not be surprising, and is almost to be expected, that organisms with consciousness and other aspects of intelligence will someday be communicated with, thus supporting the thesis of Universal life. It is provocative to conjecture whether these "alien" thinkers have some insight into the concept of an Intelligent Designer (Creator or God) and whether they believe, as most here still do, in the special gift of that God of the "soul" destined to persist in some form of immortality.

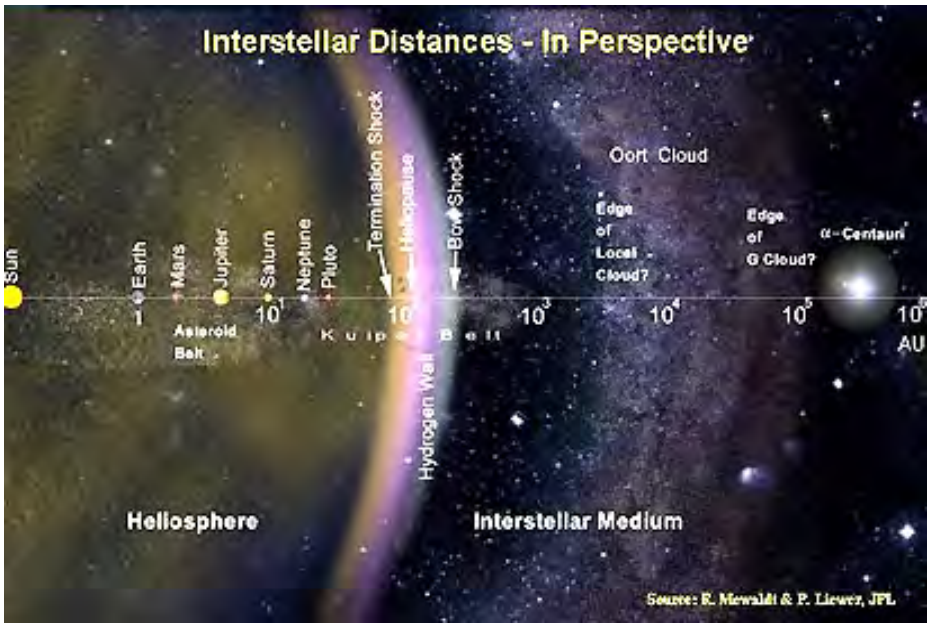
Left unsaid up to now is the questions "What kind of intelligence?" and "How intelligent?" Will the aliens be significantly smarter than humans presently are? The answer is probably YES - for two obvious reasons: 1) 'Homo' has been around for only about 2 million years, and has only really developed a sophisticated intelligence in the last few thousand (or hundred!) years, so it is sensibly likely that evolution culminating in intelligence has proceeded farther on many planets; and 2) since getting to Earth from planets light years away requires some very advanced techniques not yet discovered and developed by terrestrial scientists, any aliens who do arrive or visit would be much more complexly evolved mentally.

This is perhaps a propitious moment to interject the writer's (NMS) personal biases about the likelihood of extraterrestrial life and, in particular, intelligent life. I, and others, believe, with almost absolute certainty, that life exists elsewhere in the galaxy, and even more certainly, in most other galaxies. I suspect that some of that life has a form or degree of intelligence. My argument (much like ones put forth by experts in this field) for this view is largely statistical. With so many billions of stars and billions of galaxies, the probability that planets are the norm and on at least some the conditions favoring life - evolved to varying degrees - are in place is high enough to conclude safely that life exists. The implications for this are enormous, and were part of my speculation near the bottom of page 20-10: The scientific model for a Universe (or a Multiverse) contains a strong likelihood that evolution in the broadest sense has produced creatures that include intelligent beings; all of this is purely natural, the consequence of the inevitable occurrence of life-forming environments in a diverse Universe. The metaphysical model differs in a fundamental way: it holds that a Creator or Intelligent Designer either directly made intelligent life in a miraculous way or decided to produce that life by a master plan based on evolutionary principles. But, it seems ridiculous to have produced humans only on one of the likely billions of planets existant, with all other stars/galaxies being there just for our amusement or to give astronomers something to do. If God exists, I have come to believe that the grand scheme the

Creator Being imposed on the Universe (Multiverse) was to set into the creation model the inherent predisposition for life to have developed on a vast myriad of planets. In simple terms: God intended for intelligent life to be widespread universally and not just on a single planet - Earth. **This idea, and related topics, is explored more fully on the next page.**

Notice that we have progressed well down this page without dwelling in detail on the favorite topic among those - scientists and laypersons - who speculate on the possibilities and ramifications of intelligent "alien" life in our galaxy, and by reasonable inference, in most other galaxies: the reality of whether we have <u>really</u> been visited by these creatures in their spaceships (UFO's) at times in the past, and a corollary, whether we on Earth will ever have the means to visit other planets by some means of space travel. In light of present knowledge, any extraterrestrials will almost certainly not originate in any other Solar System planet but would reach us from beyond - well into outer space. We, in turn, must gain experience in space travel by first journeying to one or more of our neighboring planets. Probably in the lifetime of some who read these words this will happen. But extending this travel to other stars - **interstellar travel** within the Milky Way - may not happen in this timeframe, although 100, 500, 1000, a million years hence, the advances in science and technology should bring this about. But, complications and limitations must be overcome.

By far the biggest problem in interstellar travel is <u>distance</u>. A simple example illustrates the difficulty. The nearest star is Proxima Centauri, 4.2 light years away (actually, Alpha Centauri, 0.1 l.y. farther away, is a better choice owing to its size), or about 42 trillion kilometers (26 trillion miles) from Earth, would be a reasonable first target. To help you visualize these distances, look at this diagram (the distances are in Astronomical Units [A.U.])



If a manned spacecraft were to leave the Solar System at the same velocity that Pioneer 10 had when it actually did escape the system, namely, 60000 km/hr (37000 mph), it would take approximately 80000 years to reach one of these two stars. (And, the same time to return to Earth, unless a one-way trip is the choice, then at least double that total.) This obviously would be impractical under the psychology of today's thinking. The solution should be obvious: Earthlings must build a spacecraft that contains all those materials and provisions needed to sustain life for eons. Chief among these would be foodstuffs, water, Oxygen and other essentials. Then, even if human life spans can be extended, the strategy would still have to be: to continually recreate people on board through breeding, so that those who arrive at Alpha Centauri (hopefully, we will discover planets there; none have been detected as yet) will be many generations down the time path of continuing life of the future. Trips to other more distant stars may be more enticing if these show evidence of planets that can sustain life. Surely, from the billions of humans on Earth when this time of launch comes there may be volunteers who are agreeable to setting forth on this voyage; if and when we develop the appropriate technology,

the travelers may need to consent to being placed in some kind of suspended animation (a body freeze technique is commonly proposed to put living creatures into hibernation). Such a trip is likely to fulfill at least one of four motivations: 1) either the innate need for man to *explore*; (2) and/or a desire to establish contact and exchange knowledge with other *civilizations*; (3) and/or a compelling need to *survive* if Earth should threatened to become uninhabitable); (4) and/or a decision to *colonize* another (uninhabited) planet with suitable living conditions for organisms, or if intelligent beings are found, to *settle* with them.

However, most readers of this Tutorial would judge the long duration travel scenario (with or without hibernation) to be rather undesirable even if altruistic. Is there any alternative. Yes, if we can find new means of propulsion through space at much greater speeds than presently achievable. For instance, let the inertial velocity reach 0.2 the speed of light 'c' (60000 km/sec). If that occurs soon after launch, the spacecraft should reach Alpha Centauri in 5 x 4.3 l.y., or 21 years (earthtime), or 42+ years roundtrip. Theoretically, this transit time can be greatly shortened if the spacecraft attains a velocity near light speed. However, relativistic effects will come into play (see Preface to this Section), including differential aging between space travelers and those remaining on Earth. Thus, time dilation would make the high speed trip appear to those on Earth to have taken longer than the 42 years plus adjustments for being somewhat under light speed.

There is another problem that may even supercede the time constraint on long distance travel. This is the radiation - cosmic rays and other particles - that pervade space. Calculations based on studies of humans in near Earth missions and of objects in space then recovered indicate that unless the spacecraft is properly sealed within a protective radiation-shielding material, the humans on board would receive lethal doses after traveling for several years. This might even apply to trips as short as Mars and back. The quantity of protective material is large, adding so much weight to the spacecraft that known and existing propulsion systems fall way short of being powerful enough to launch the craft into the long journey. Technology must find acceptable means to overcome this problem. The solution is not yet in sight but future discoveries hopefully will neutralize the radiation hazards.

There are many other factors required for success and safety to take into consideration. Perhaps at the top of that list is to find propulsion systems that can reach these high speeds (significant fractions of the speed of light); to get to such speeds requires huge expenditures of energy. Presently, no energy supply is known, but some sound proposals for possibilities already have surfaced: ion engines, anti-matter engines, controlled nuclear processes, gravitational "slings", laser beams pushing on light sails, and various other innovative but speculative mechanisms for powerful propulsion systems. Quantum-minded thinkers can conjure up schemes that depend on "wormholes" and "quantum tunneling", "timewarps", and the like. A long shot depends on the proof of existence of the hypothesized "tachyons", particles that travel at faster than the speed of light; if real and accessible, some technique would be needed to harness them as aids to propulsion.

Whatever propulsion system is eventually proven feasible, one requisite is that it be one that travels with the spacecraft (instead of a "one-shot" push at the outset) so that there would always be a means for course corrections, maneuvering, handling the unforseen, possible landing, and eventual return to Earth if that is a mission requirement. If we base our predictions for space travel eventuality on the huge advances in science and technology over the last two centuries - now seeming to occur as though growing exponentially - it is reasonable to expect that the possibility of interstellar travel will turn into reality in the not too distant future (say, in this millenium). If that is achieved, then the counterpoint argument is that aliens "out there" may be ahead of us and have in fact visited Earth - if we are judged to be interesting enough.

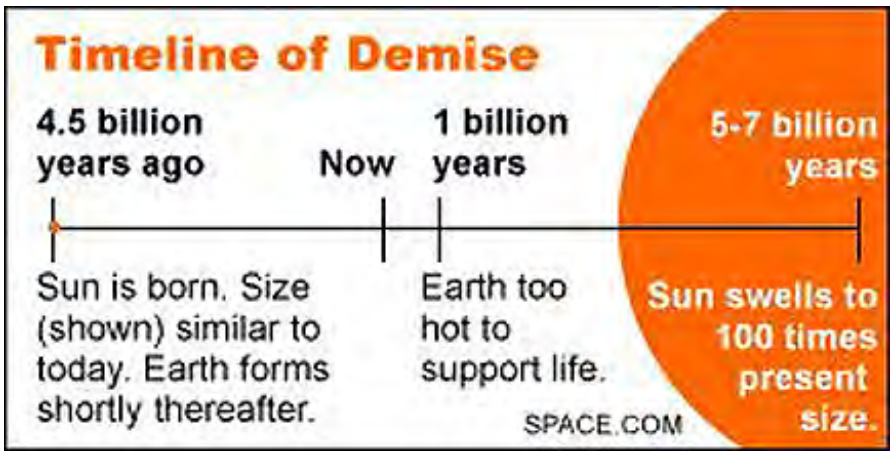
So, is there any "bottom line" to this speculation on aliens. There certainly is no consensus among scientists. The writer has a strong bias: Earth has never been visited by alien intelligents. Why do I say that? Because the distances between inhabited planets (as probabilities imply they exist) are just too great for space travel. Propulsion systems near the speed of light

probably are unattainable. Sending creatures who reproduce, and thus arrival occurs generations later, is probably not appealing to the erstwhile travelers. Instead, the aliens would more likely send unmanned robotic spacecraft and/or would send signals into space to communicate. But, personal appearances by the aliens are (so far) just "figments" of human imaginations.

To sum up the hard realities of space travel: 1) present and foreseeable technologies are still far short of making such trips plausible, safe and worthwhile 2) in time (probably many centuries), humans may learn enough to engage in such endeavors; 3) almost certainly, a manned trip will be preceded by unmanned spacecraft to prove the workability of the technology; 4) if mankind survives itself (or evolves into some form of superintelligence that can control the various threats to its self-destruction (wars; envIronmental nihilism) on Earth, trips to other stars seems someday to be inevitable; and 5) in the meantime, we should continue to inventory suitable planets and search for intelligent life elsewhere (SETI), so as to develop the incentive for undertaking travel to candidate planets; we should also hunt for any evidence that Earth has previously been visited (stated in the Fermi Paradox: If aliens have already come to Earth, as might be expected since statistically there could well be many more advanced civilizations than those on Earth today if intelligent life is widespread in the Universe, then why haven't we found any valid signs of their visit?)

An excellent article for the layperson on space travel feasibility, simply titled *Star Trek*, by W.S. Weed, appeared in the August 2003 issue of *Discovery Magazine*. It is well worth reading to realize both potential and problems in visiting even the nearby stars. Its gist is summed here: This is the list of the 5 propulsion systems reviewed in the article: Atomic Rockets; Nuclear Fusion: Antimatter; Laser Sail; and Fusion Ramjet. Within several decades one or more of these systems, and probably others yet conceived, may become practical working modes that can propel space travelers at speeds from 0.1 to perhaps 0.8 that of light. Most of the systems discussed would require that humans onboard would have to spend from 40 to 100 years to reach stars close-by. The various problems they would face: food/water; air; deletorious gravitational effects; radiation threats, and, not the least, psychological adjustments are all solvable but require considerable refinement from present-day capabilities. Of course, the age problem (not helped if the speed of light is not approached) can be handled by onboard breeding and birth, i.e, resorting to multiple generations.

Perhaps this is the point to introduce a note of caution. Based on some current trends, we ourselves may do intelligent life in, or a super asteroid, might extinguish us. That is one motivation for continuing to seek means to escape the confines of our Solar System and seek out new planetary systems around other stars. There is another compelling reason, purely natural in scope, suggested by this illustration:



Thus, the Earth and its siblings are doomed ultimately by the exhaustion of fuel in the Sun which will lead to a catastrophic (if life still exists then) event associated with the Red Giant phase (mentioned on page 20-5a). Even before the next 5 billion years, severe warming may wipe out life if countermeasures haven't been found in the next billion years. Or, the ultimate Doomsday

scenario: Man, or some successive genus or new type of intelligence, destroys "himself" through acts of his own (common example - total nuclear holocaust). A frightening yet plausible mode of destruction is for humans to develop thinking machines that may become "smarter" that people and develop the capability to subordinate their makers or even gain a control that leads to the demise of the flesh-and-blood race. (This is already happening at levels below us humans as we are driving lower forms of life into premature extinction). Yet, somehow I fail to feel much personal concern at this time, as I happily believe that humans will mature enough in the future to assure their control over their survival and to become good stewards of other life.

In closing Section 20, contemplation of Cosmology, and its astronomical substructure, is a truly humbling experience for one's brain. The wonder is: that there exists on one tiny point in a humongous Universe something called the "conscious" and that the human mind (yours, for instance) can conceive of, and begin to understand, the truths of this Universe's attributes and history that are continually being discovered and refined. Finally, it is both astonishing and reassuring to realize that all beings - be they people or animals/plants or inanimate matter - are remarkably cosmological in nature: All the atoms in our bodies were once contained in stars or interstellar space; our parts in a sense are variously billions of years old and their atomic constituents, even after multiple dispersions and reassemblies, will last at least as long as the present Universe - estimated to continue for perhaps 50 b.y or more. In one way, then, our essences will have achieved some kind of plausible Immortality in view of the many incarnations that preceded our current atomic arrangement and are yet to happen. But, from the humble side, perhaps "We are not alone" and certainly we are not at the center of the known Universe; our importance is *sui generis* and our rank among the Universe's populations is probably just average.

Despite this humbling thought, you, the reader, and me, the writer, and all of humanity possess a fundamental trait in common: We wonder about the meaning of it all, and especially what has caused mankind and all of the Universe(s) to have come into existence. This inclination - which comes under the heading of Teleology (referring to prime causes and purposes) is not the normal purview of Science. Instead, the speculations are the mainstays of Philosophy and Metaphysics. The writer has long engaged in trying to relate Science to the Ultimate Meaning of it all. I have added a page (Next button below or page 20-12b) to this Section which documents much of my thinking, to which you are invited to visit next.

Additional Readings and Web Sites

This Section on Cosmology has doubtless been heavy going for most readers. Some may be inspired to wish to learn more. Refer to the Preface link accessed from page 20-1 for the references to books consulted by the writer in preparing this overview. We want to steer you to a Web site that allows you to seek more pictures and textual information about most of the topics that have been covered or touched upon so far in this Appendix. NASA Goddard astronomers have put together a Web site that features many previous Astronomy Pictures of the Day (APOD). In the Search box, you simply type in a topic, e.g., young stars; supernova; black holes; spiral galaxies, planet, etc. and if the category is there a running text with links to subtopics and pictures will be delivered. This can be an adventure. Another web site is Wipedia's Physical Cosmology. The next site gives a Brief history of Cosmology . Both history and metaphysics are found at this PBS site. Finally, a series of useful topical links are embedded in this AIP site.

Primary Author: Nicholas M. Short, Sr.

Some Philosophical Implications concerning the Cause and Purpose of the Universe

THE TOPICS TREATED ON THIS PAGE ARE STRICTLY THE THOUGHTS OF THE TUTORIAL'S PRINCIPAL WRITER (NICHOLAS M. SHORT) AND IN NO WAY ARE ENDORSED OFFICIALLY BY NASA.

All that remains in our Tour through Astronomy and Cosmology is to see how the Universe may have come to be and to weigh the possibilities of life - especially in some intelligent form(s) - throughout the Universe. That is treated below, along with metaphysical speculations about not only ***how*** but ***why*** Universe(s) exist(s) in the first place, and, more particularly, musings on the role *intelligent* beings might play in the physical and (if it exists) spiritual aspects of this existence,assuming the Universe has some absolute reality. We have on previous pages left largely unsaid this most fundamental of questions: Can a Universe (or Multiverse) come into existence on its own or does it (them) require some external agent, and how did that agent (one version being "God") itself arise without some preceding cause? And its corollary: Can there be an Uncaused Cause? This page will explore this conundrum, but don't expect any pat answers.

Topics such as Multiverses represent excursions of the human mind well into the abstract. Contemplating the idea of a Multiverse seems to carry one into a state of purely fanciful thinking. Can a concept as "far out" as Multiverses have any claim to be part of "reality". The answer is unequivocally YES in the sense of such a "thing" (the catch word that allows anything to be a valid idea) being POSSIBLE. But that in itself doesn't prove the reality of the "thing conceived".

If one accepts the duality of human existence - towit, there are **physical**, material aspects (what Science describes) together with **abstract**, spiritual aspects (including consciousness, ideals, morality, etc.) that reside together in human nature - then we can justify adding the ideas set forth in the remainder of this page as worthy of speculative consideration. We will thus attempt to erect a bridge between Physical (Concrete Things) and Metaphysical (Abstract Things).

To begin: As a comment on the above ideas, from a metaphysical perspective, the argument that quantum mechanical principles are sufficient to explain creation of a Universe *ex nihilo* (from seemingly nothing) by postulating *virtual particles* that follow some mechanism converting them to "degrees of universe" has one serious flaw. These particles, even though they may not survive for any extended moment because the vast majority are quickly annihilated, still are **something** in the realm of existence. Therefore, they have a reality and are subject to principles of causality (that is, every effect must have a cause except - to solve the dilemma of getting started - an uncaused cause). So too is the existence of a vacuum which serves as the matrix for virtual particles. Thus, the age-old doctrine (traceable to the Greeks) that "only nothing can come from nothing" remains applicable - some independent agent would seem to be needed to "create" the vacuum and the potentiality of virtual particles. Of course, a conundrum arises at once: **what is the origin (cause) of the agent**. Theologians who cite the agent to be what is called "God" or "The Intelligent Designer" are forced to postulate a special - and probably unique - property of the Designer being uncaused.

Various philosphers, theologians, and some scientists have attempted to reconcile what many believe to be a presumptive basic conflict between Science and Religion, namely, is the Universe purely natural and self-generated or was it created and designed by a pre-existing Intellect. The writer has his own thoughts on this most profound of subjects. They

are summarized in three premises: First, for the sake of argument I postulate that it is conceivable, even plausible, that there is in actuality some kind of Creator. Second, let's accept that essentially all that has been discovered, described, and tested by Science is valid to the extent now known (future scientific studies will likely modify some current concepts, discard others, and find new laws and principles). Then, third - the key resolution and settlement to all the controversy - is that <u>both the religionists and the materialists are right!!</u> Under this model, what human thinking has done, in Science, in Sociology, in Theology, in any field attempting this explanatory act, is just to unravel the way in which a supernatural creator-being set into motion (thereby establishing reality) all that we perceive in our created Universe. Thus, Science is discovering not only the natural "modis operandi" of the Universe's origin and development but also the way in which it starts and evolves according to a blueprint established by an independent agent (God; Deus; Yahweh, Allah, Ishvara, Krishna, many others; or, from an atheistic viewpoint, some physical entity, beit a force or pure energy). Theologians have long argued between the concept of Theism - the god-being takes an interactive and continuing role in the affairs of humans and all of creation - and the concept of Deism - the god-being established the principles by which the Universe operates, set it into motion, and since then, at least on Earth, only occasionally interacts with humans, thus does not guide or direct the world on a daily basis.

(As an interpolated "aside": Einstein was a deist, but not quite an agnostic. He was partial to the philosophy of the 17th century thinker, Baruch Spinoza. Spinoza developed a form of pantheism that held that God and Nature were one and the same. To learn more, go to the Internet and google (or Yahoo or Bing) for "god vs science Albert Einstein" and for "Spinoza".)

Another very important - fundamental - issue can be raised that has a profound implication for the nature and history of the Universe. It is this point that Intelligent Design advocates stress: there are a number (at least 47) fundamental parameters involved in the Universe's inception and development that, if changed even a small amount, would have so affected the process that the Universe could not have developed the way it is or, for that matter, even come into existence.

This touches upon the argument of "fine tuning", related to but distinct from the Anthropic Principle (the Universe is meaningful and knowable because Man exists as a conscious being capable of deciphering this meaning). At the time of the Big Bang and thereafter, certain constants and parameters (those fundamental to Physics) involved in the process, and the mathematical equations that delineate the initiating process and all that follows in the natural order, had to have very specific (precise) values. Otherwise, if these numbers deviated notably from their actual values, and sometimes by just a little, the organization and evolution of the Universe we observe would not have been possible, or, if it could exist, would be so different that it might never have fostered the conditions that led to organic life, here on Earth and probably elsewhere in the Cosmos.

Many of these parameters are listed here:

Evidence For Design In The Universe

From Limits for the Universe by Hugh Ross, Ph.D.

1. **Gravitational coupling constant.** If larger: No stars less than 1.4 solar masses, hence short stellar lifespans. If smaller: No stars more than 0.8 solar masses, hence no heavy element production

2. **Strong nuclear force coupling constant.** If larger: No Hydrogen; nuclei essential for life are unstable. If smaller: No elements other than Hydrogen.

3. **Weak nuclear force coupling constant.** If larger: All Hydrogen is converted to helium in the

Big Bang, hence too much heavy elements. If smaller: No Helium produced from Big Bang, hence not enough heavy elements.

4. **Electromagnetic coupling constant**. If larger: No chemical bonding; elements more massive than Bboron are unstable to fission. If smaller: No chemical bonding.

5. **Ratio of protons to electrons formation** If larger: electromagnetism dominates gravity preventing galaxy, star, and planet formation. If smaller: Electromagnetism dominates gravity preventing galaxy, star, and planet formation.

6. **Ratio of electron to proton mass**. If larger: No chemical bonding. If smaller: No chemical bonding.

7. **Expansion rate of the universe**. If larger: No galaxy formation. If smaller: Universe collapses prior to star formation.

8. **Entropy level of universe**. If larger: No star condensation within the proto-galaxies If smaller: No proto-galaxy formation

9. **Mass density of the universe**. If larger: Too much Deuterium from Big Bang, hence stars burn too rapidly If smaller: No Helium from Big Bang, hence not enough heavy elements

10. **Age of the universe**. If older: No solar-type stars in a stable burning phase in the right part of the galaxy. If younger: Solar-type stars in a stable burning phase would not yet have formed.

11. **Initial uniformity of radiation**. If smoother: Stars, star clusters, and galaxies would not have formed. If coarser: Universe by now would be mostly Black Holes and empty space.

12. **Average distance between stars**. If larger: Heavy element density too thin for rocky planet production, If smaller: Planetary orbits become destabilized.

13. **Solar luminosity**. If increases too soon: Runaway Greenhouse effect. If increases too late: Frozen oceans.

14. **Fine structure constant***. If larger: No stars more than 0.7 solar masses. If smaller: No stars less then 1.8 solar masses.

15. **Decay rate of the proton**. If greater: Life would be exterminated by the release of radiation. If smaller: Insufficient matter in the Universe for life.

16. **C^{12} to O^{16} energy level ratio**. If larger: Insufficient Oxygen If smaller: Insufficient Carbon.

17. **Decay rate of Be^{8}**. If slower: Heavy element fusion would generate catastrophic explosions in all the stars. If faster: No element production beyond Beryllium and, hence, no life chemistry possible.

18. **Mass difference between the neutron and the proton**. If greater: Protons would decay before stable nuclei could form If smaller: Protons would decay before stable nuclei could form.

19. **Initial excess of nucleons over anti-nucleons**. If greater: Too much radiation for planets to form. If smaller: Not enough matter for galaxies or stars to form.

20. **Galaxy type**. If too elliptical: Star formation ceases before sufficient heavy element buildup for life chemistry. If too irregular: Radiation exposure on occasion is too severe and/or heavy elements for life chemistry are not available.

21. **Parent star distance from center of galaxy**. If farther: Quantity of heavy elements would be insufficient to make rocky planets If closer: Stellar density and radiation would be too great.

22. **Number of stars in the planetary system**. If more than one: Tidal interactions would disrupt planetary orbits. If less than one: Heat produced would be insufficient for life.

23. **Parent star birth date**. If more recent: Star wuld not yet have reached stable burning phase. If less recent: Stellar system would not yet contain enough heavy elements.

24. **Parent star mass**. If greater: Luminosity would change too fast; star would burn too rapidly. If less: Range of distances appropriate for life would be too narrow; tidal forces would disrupt the rotational period for a planet of the right distance; UV radiation would be inadequate for plants to make sugars and Oxygen.

25. **Parent star age**. If older: Luminosity of star would change too quickly. If younger: Luminosity of star would change too quickly.

26. **Parent star color**. If redder: Photosynthetic response would be insufficient. If bluer: Photosynthetic response would be insufficient.

27. **Supernovae eruptions**. If too close: Life on the planet would be exterminated. If too far: Not enough heavy element ashes for the formation of rocky planets. If too infrequent: Not enough heavy element ashes for the formation of rocky planets. If too frequent: Life on the planet would be exterminated.

28. **White/Dwarf binaries**. If too few: Insufficient fluorine produced for life chemistry to proceed. If too many: Disruption of planetary orbits from stellar density; life on the planet would be exterminated.

29. **Surface gravity (escape velocity)**. If stronger: Atmosphere would retain too much ammonia and methane. If weaker: Planet's atmosphere would lose too much water.

30. **Distance from parent star**. If farther: Planet would be too cool for a stable water cycle. If closer: Planet would be too warm for a stable water cycle.

31. **Inclination of orbit**. If too great: Temperature differences on the planet would be too extreme.

32. **Orbital eccentricity**. If too great: Seasonal temperature differences would be too extreme.

33. **Axial tilt**. If greater: Surface temperature differences would be too great. If less: Surface temperature differences would be too great.

34. **Rotation period**. If longer: Diurnal temperature differences would be too great. If shorter: Atmospheric wind velocities would be too great.

35. **Gravitational interaction with a moon**. If greater: Tidal effects on the oceans, atmosphere, and rotational period would be too severe. If less: Orbital obliquity changes would cause climatic instabilities.

36. **Magnetic field**. If stronger: Electromagnetic storms would be too severe. If weaker: Inadequate protection from hard steller radiation.

37. **Thickness of crust**. If thicker: Too much Oxygen would be transferred from the atmosphere to the crust. If thinner: Volcanic and tectonic activity would be too great.

38. **Albedo (ratio of reflected light to total amount falling on surface)**. If greater: Runaway ice age would develop. If less: Runaway Greenhouse effect would develop.

39. **Oxygen to Nitrogen ratio in atmosphere**. If larger: Advanced life functions would proceed too quickly. If smaller: Advanced life functions would proceed too slowly.

40. **Carbon dioxide level in atmosphere**. If greater: Runaway greenhouse effect would develop. If less: Plants would not be able to maintain efficient photosynthesis.

41 **Water vapor level in atmosphere**. If greater: Runaway Greenhouse effect would develop. If less: Rainfall would be too meager for advanced life on the land.

42. **Ozone level in atmosphere**. If greater: Surface temperatures would be too low. If less Surface temperatures would be too high; there would be too much uv radiation at the surface.

43. **Atmospheric electric discharge rate**. If greater: Too much fire destruction would occur. If less: Too little nitrogen would be fixed in the atmosphere.

44. **Oxygen quantity in atmosphere**. If greater: Plants and hydrocarbons would burn up too easily. If less: Advanced animals would have too little to breathe.

45. **Oceans to continents ratio**. If greater: Diversity and complexity of life-forms would be limited. If smaller: diversity and complexity of life-forms would be limited.

46. **Soil mineralization**. If too nutrient poor: diversity and complexity of life-forms would be limited. If too nutrient rich: Diversity and complexity of life-forms would be limited.

47. **Seismic activity**. If greater: Too many life-forms would be destroyed. If less: Nutrients on ocean floors (from river runoff) would not be recycled to the continents through tectonic uplift.

(To give a balanced opinion of these contants, we should mention that some scientists have challenged the supposition that these parameters are not necessarily constant over the long time the Universe has existed.)

The use of "greater" or "less" (or equivalents) in connection with the fundamental parameters listed above is judged by the writer to imply not slight amounts but significant changes for at least some of the 47 fundamental parameter items shown. For example, if Oxygen varies by, say, 2%, life forms probably could have evolved; continental/ocean ratios have varied considerably in the past without derailing the advance of life. Also, many scientists disagree, play down, and debate the validity of some (perhaps many) of the above items in the list. Certain ones have not yet been proved adequately. Some items are redundant or cross-correlated. Many seem to be valid, with acceptable certainty, only for Earth (so far) where observations and experience are based on scientific studies. Finally, it may be that some of the above values could vary by large amounts if certain other values also change in a compensatory way that permits development of planets with different parameters while still allowing for appearance and evolution of life, although no one has supported this premise satisfactorily.

Note that some of the items in the above parameter listing apply mainly to the Universe at the cosmological level but others are applicable primarily to the Earth and life thereon and by inference other inhabited planets. (The conditions that allow life at various levels have more flexibility - can vary over wider limits - but still are subject to narrower ranges.)

Nevertheless, both the nontheistic scientists (those that argue for a purely natural origin of the Universe and its evolution - both cosmic and organic) without the need for an external intelligent cause or designer (what is usually meant by "God") and theistic scientists (those who favor a designing intelligence that at the least set the Universe in motion and assigned it the parameters stated above) have developed intriguing arguments for their viewpoints. The first group believes the Universe to be self-originating and self-controlled, perhaps accidental (governed by chance) rather than arbitrarily preplanned, and subject to principles of "Natural Selection" (in which the controlling parameters may have changed as it evolves). Thus the parameters may not have always been constant. The second group cites the complexity of the Universe and particularly of the human mind as very, very unlikely to originate by chance alone; some external pre-existing being must have guided its inception and development. For that

group, the parameters did not just "happen" to be right, since the probability that they all assumed the limited (almost immutable) values assigned by "good luck" is extraordinarily small. In their view, the parameters are truly immutably, having been established as they are now from the outset of the Big Bang.

Who is "right" is so far not settled. Each may possess some truth or measure of reality. Perhaps both are right to varying degrees. Or neither has yet found out the real nature of the Cosmos, and may never. But human nature seeks a responsible deity or at least a mind-force that seemingly best explains our existence. And, we are "hooked" on the notion of Causality - everything must have a cause. Ultimately, the optimal argument for Science - irrespective of a Designer involvement - to account for the realities it is discovering may rest on teleological arguments. Hopefully, those would provide answers to the basic question of "WHY", namely, do purposes reside in the natural world as driving or controlling principles. Or is the apparent fact that we exist within a material framework of laws of physics sufficient to explain the naturalistic model without recourse to an external agent. In this latter alternative, the notion of consciousness, ideas, ideals, spirituality, and soul are just properties of the brain as a physical host device. Still, human nature being what it is, most people, including many scientists, yearn for a sense of purpose that was imposed by a superintelligent creator.

Most of the above musings reflect the writer's personal thinking. I have three remaining thoughts of moment: 1) Science considers only what can be sought and explained in the **physical Universe**; it cannot observe, measure, and test the abstractions inherent in the concept of God, although some of its methodology is adaptable to metaphysical inquiry; 2) Science is yet to embark on a mission to try to explain the **purpose** of the Universe; it probably is not equipped to do so; but the question ultimately needs to be addressed, as for example if the Universe came into existence out of a quantum fluctuation in the vacuum energy (as many cosmologists now believe), what brought this energy itself into existence; and 3) why would the Creator/Designer have produced a **huge** Universe solely to place thinking beings (humans) on one small planet near an average star in just one galaxy among billions; it seems far more sensible to postulate that intelligent life exists in abundance in the multitudes of galaxies throughout the Universe; this "we are one among many" hypothesis suggests to me that the reason for the vast Universe as a place where life is widespread is that an omniscient Creator intended thinking creatures to be everywhere, and thus there are lots of everywheres.

The question of purpose for the Universe itself has challenged philosophers and theologicans for more than two millenia. An interesting website considers this subject. Access it at The Templeton Foundation website.

For the curious, these paperback books by Paul Davies offer valuable insights into both scientific and metaphysical aspects of Cosmology: *God and the New Physics*, *The Fifth Miracle*, and *The Mind of God* [especially Ch. 2]., Touchstone Books, Simon & Schuster, Inc.; this author considers the question of life elsewhere in the Universe in *Are We Alone: Philosophical Implications of the Discovery of Extraterrestrial Life*, Basic Books, 1995. Dr. Davies (now at Arizona State University) has followed up these books with an extremely insightful and provocative article in the September 2003 issue of *The Atlantic Monthly* entitled "E.T. and God" - highly recommended for its synoptic overview.

A classic review of the role of Science in understanding the need for a God is *Science and Creation*, by John Polkinghorne, New Science Library, 1989. A book that concentrates on relating cosmological discoveries to teachings associated with the Christian God is *Beyond the Cosmos* by Hugh Ross, Oxford Press, 1996. A balanced and thorough review that considers how religious beliefs and observations of the physical Universe are not necessarily incompatible is *When Science meets Religion* by Ian Barbour, Harper, 1999. A general overview of arguments contrary to the exclusively natural and spontaneous Universe such as described in this Section is *A Case against Accident and Self-Organization*, by Dean L. Overman, Rowman & Littlefield. One of the best read by the writer (NMS)that supports the existence of an Intelligent Creator within the framework of Science is *The God Hypothesis: Discovering Design in our "Just Right" Goldilocks*

Universe, by Michael Corey, Roman & Littlefield, 1999. A recent book that takes just the opposite view - summarizing the thoughts of nontheistic scientists on the capability of a self-created Universe and thus dispensing with the need for a Creator - is *The God Delusion*, by Richard Dawkins, Houghton Mifflin, 2006.

Dawkins' book relies mostly on arguments other than cosmological to demonstrate his inferred likelihood that no Creator is necessary. Corey's book is the most complete and broad treatise relying on cosmological arguments to deduce that a Creator must have acted to initiate the Big Bang and to program in all the conditions that favored the eventual appearance of life (he tends to be very repetitious in his presentation). He has two main theses: 1) There are many constants (including those above) that are both immutable (have acted throughout the Universe since its beginning, without significant change) and fine-tuned (if their values change even a little, the Universe as it is could not come into existence, or if formed, would not survive with conditions favoring life), and 2) humans, as far as we now know, are the most advanced and complex beings within the material Universe - and, although they have taken a long time to appear, have evolved under a very narrow range of circumstances that permit them to now contemplate the kinds of topics treated on this page.

Two books that summarize the arguments for Intelligent Design as an alternative to naturalistic metaphysics are *Darwin's Black Box*, by Michael Behe, Simon & Schuster, and *Intelligent Design*, by William Dembski, Intravarsity Press.

In the March, 2002 issue of the magazine *First Things*, which deals with topics in philosophy, theology, and the social order, an outstanding review of the possibilities of life elsewhere in the Universe and its implications for humankind, written by Fred Heeren, and entitled *Home Alone in the Universe?*, provides comprehensive and provocative insights into the question of how discovery of intelligent life beyond the Solar System would affect and change the outlook of Earth's inhabitants towards their place and role in the Universe. Highly recommended!

An interesting article dealing with the Anthropic Principle applied to cosmological ideas by Victor J. Stenger is found on the Internet. In this essay, Stenger argues against a recent resurgence in reconciliation between Science and Religion as still fraught with false premises. He concludes that a purely natural Universe is a very real possibility but while discounting the idea of Intelligent Design, he does not rule it out based on some overriding proof of its falsity but includes it in the list of possibilities. We strongly urge you to read this article and ponder its consequences. Then, you might wish to consider his article on The other side of time.

Another point of view that elevates Science almost to the status of a modern religion for our time and the future is presented in this Internet site by the aforementioned Lonely Planets website by David Grinspoon. Like many scientists now practicing, he believes that formal religions (those relying on revelations rather tban rigorous testing) might be trending towards the superfluous. Or, at the least they are in need of major restructuring and reconciliation with the World System of Science and Natural Philosophy.

The above ideas are admittedly speculative. Some were taken from a letter the writer had published in a local newspaper in 2000. For those interested in the thoughts expressed in that letter, click on this LETTER link

Primary Author: Nicholas M. Short, Sr.

Life in the Universe: II. Origin and History of Life on Earth

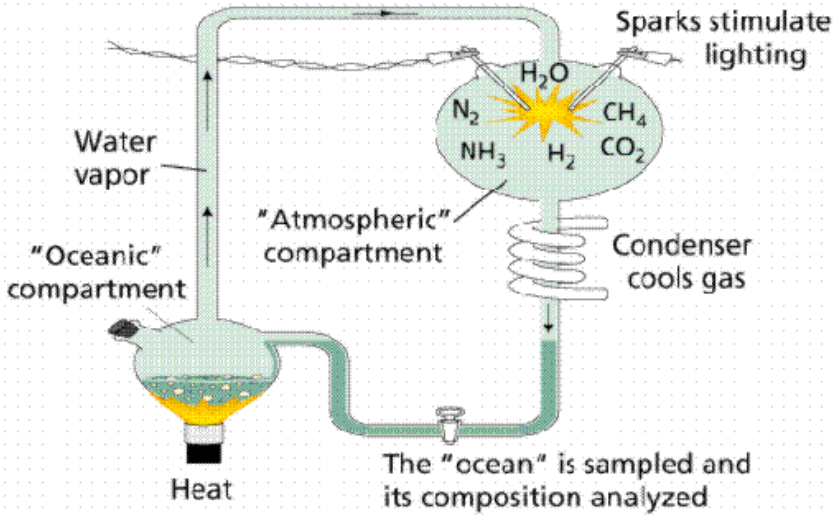
Now, with this background in Biochemistry we move on to facts and speculations about the origin and development of life on Earth (and by inference, elsewhere in the Universe). This general topic is intimately tied to the theory or concept of Evolution. For our purposes, an understanding of how Evolution works, or even that it is a real process, is not necessary to work through this page. But the ideas behind Evolution are important knowledge, the gist of which is worth your time to become familiar with. So, as an option, before proceeding you are invited to peruse the next page dealing with this subject, accessed by clicking on this highlighted Evolution word; otherwise, defer until the end of this page. On the present page we will consider the progression of life through geologic time but only the evolution of humans will be examined from that perspective.

There is general consensus that the first organic ingredients that became plentiful in oceans ond other energy-rich envlronments associated with water were amino acids and later nucleic acids. RNA probably organized before DNA, and was needed to synthesize the proteins. Both are important in the production of nucleic acids. These constituents had to develop before they could organize into symple prokaryotic cells. The first organisms - bacteria - were single-cells that containing mostly water enclosed in a membrane, with RNA strands and ribosomes as the principal internal organelles.

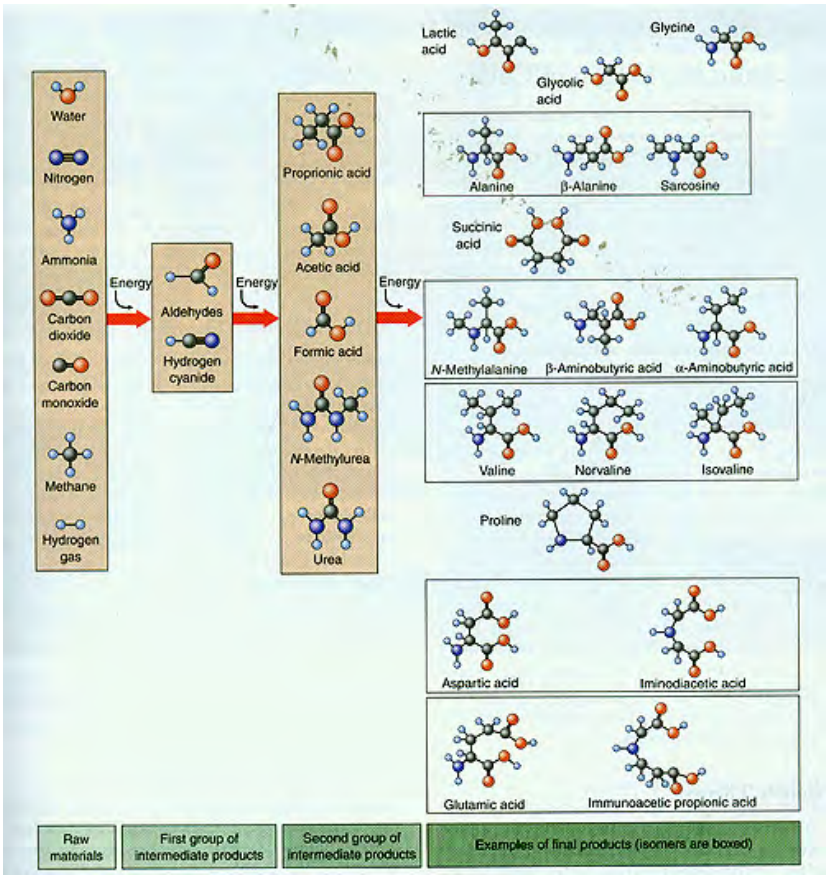
Evidence is abuilding about the Earth's atmospheric history and its relevance to the appearance of life. The most primitive atmosphere contained Nitrogen, free Hydrogen, some Carbon dioxide, and no Oxygen. In the first two billion years or so, the Sun's energy output was 70-80% of today's radiant release. To keep the early oceans from freezing, some mechanism was needed to maintain proper temperatures. Carbon dioxide - the main Greenhouse gas - could perform part of that function, but its quantity was probably too low (based on the low amounts of $FeCO_3$ or Siderite in the geologic record of early times; likewise, calcium and magnesium Carbonates [limestones] were noticeably rare).

Researchers now think that early on Carbon reacted with Hydrogen atoms to form methane (CH_4) which was more efficient than CO_2 in absorbing outgoing thermal radiation. A class of living microorganisms - methanogens - could have arisen and flourished for millions of years. The Archaebacteria, the oldest microfossils, were methanogen organisms. As these proliferated, they expelled methane in the atmosphere of the time until that gas reacted with Hydrogen and other constituents to form a "smog" (similar to that on Saturn's Titan) that built up. This in turn would absorb incoming solar irradiation and lead to a reversal of temperatures to the extent that cooling brought about an Ice Age ("Snowball Earth") some 2.3 billion years ago. Thereafter, methanogens never regained their importance as Oxygen slowly built up in the changing atmosphere. This build-up probably resulted from the onset of photosynthesis which produces Oxygen as an end product of the reaction between CO_2 and H_2O to produce glucose ($C_6H_{12}O_6$) plus Oxygen.

There are still questions about how life actually began. The key components - proteins, RNA and DNA - had to precede living cells. Speculation still continues over the mechanisms and circumstances by which the components were first produced. Although not definitively accepted as the actual scenario, an experiment in 1953 by a graduate student, Stanley L. Miller at the University of Chicago, under the tutelage of Nobel Prize winner Harold Urey, is regarded as one of the classic scientific efforts ever in the field of biology. Here is a diagram that depicts the experimental set-up:



Miller and Urey produced a primeval "micro-ocean" in one chamber. Heating it expelled water vapor into a second chamber containing the gases they thought might have existed after the molten Earth cooled to a crust and primitive ocean/atmosphere at a time much hotter than the present. Into the top flask, the water vapor-gases mix was subjected to frequent electrical sparks (to simulate lightning). As days went on, and the condensed mix was sampled and analyzed, sequences of organic molecules were synthesized, as shown here:



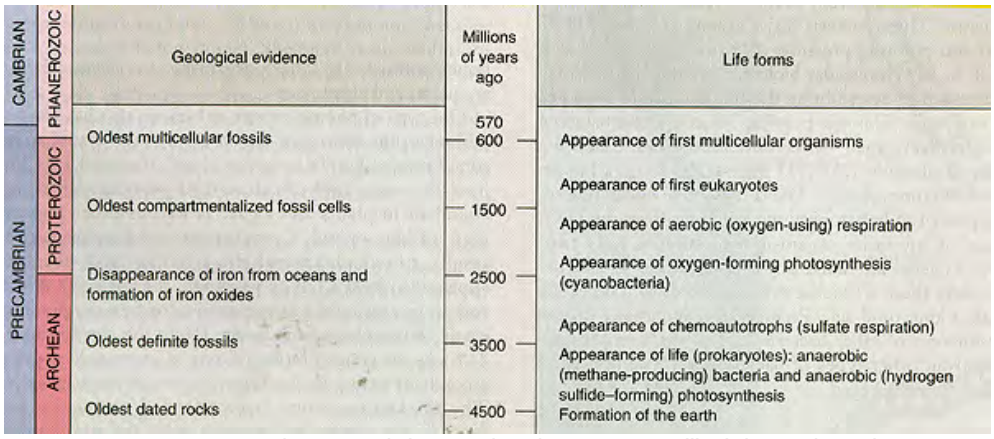
From Raven & Johnson, *Biology*, 6th Ed., McGraw-Hill Higher Education.

Variations of this famed experiment have produced still other organic molecules. The key conclusion it points to is that a reducing, hot atmosphere with compositions similar to the one they used could have generated some of the basic ingredients that later organized into life. Other sources of energy have been proposed. The conditions that prevailed then were probably like those we assign the word "*extremophile*" to. One plausible alternative is the hot waters around the deep-sea "black smokers" found around oceanic spreading ridges. Carbon escaping from a primitive Earth's mantle would react with other subsurface constituents, especially those in the water, to yield building-block molecules (life today is found around the smokers - apparently produced there - but today's water contains more Oxygen than in primitive earth envIronments [in fact, Oxygen tends to destroy these simpler molecules]). Another view, growing in acceptance, holds that at least some organic molecules were added to Earth after its general melting during bombardment by asteroids/comet (see further comments two pages hence). These extraterrestrial bodies are known to contain various amino acids. An experiment at Lawrence Livermore Laboratory in which a group of amino acids were held in a material subject

to high speed impact (from a gun) yielded the surprising result that these acids formed peptide chains, the building blocks of proteins. Thus, life on Earth could have begun internally and/or externally.

On Earth, as stated above the first life was unicellular (microbial, including an abundance of bacteria), followed much later by unicellular plant life which eventually acquired the ability to photosynthesize Carbon compounds using solar energy into monosaccharide carbohydrates, releasing Oxygen as a by-product (a build-up of Oxygen leads to formation of upper atmosphere ozone which, in turn, protects life below from destructive UV radiation). Energy sources that favor life are solar radiation, terrestrial heat, and change of state heat (nuclear decay which supplies much of Earth's heat from the interior may also provide radiation that could synthesize certain organic molecules under the right conditions). (A fourth possibility is gravitational [tidal] energy which might produce life-developing conditions; future exploration of Europa will test this mechanism by seeking life beneath its icy crust). The presence of water and a suitable atmosphere (life on Earth began in a reducing atmosphere but with photosynthesis, Oxygen has increased.

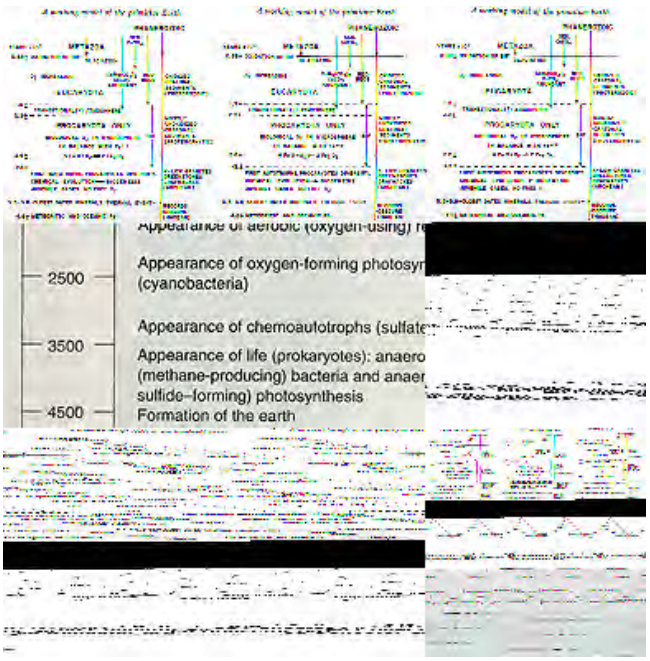
Thus, the frame of reference of any investigations of life elsewhere in the Universe continues to reside in the extensive studies of organic chemistry and biology of organisms dominating the only known place where life's existence is confirmed: our planet. Life on Earth began at least 3.5-3.8 billion years ago (a more precise time has yet to be established, since there is dispute as to the validity of proposed life forms found in ancient rocks of differing ages). Since then the history of life has been increasing complexity and diversity and adaptation of ever more (and changing) envIronments. This chart summarizes this history:



From Raven & Johnson, *Biology*, 6th Ed., McGraw-Hill Higher Education.

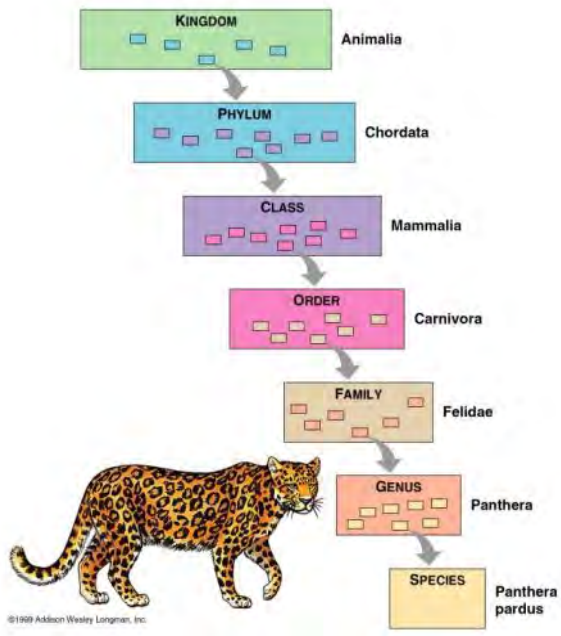
Note: The older span of time in the Archean, from about 3.8 to 4.6 billion years ago, is nowadays given the name "Hadean".

To re-enforce this brief synopsis of the history of life on Earth, we reproduce here this diagram that was placed on the first page of the survey of the basics of Geology shown on page 2-2.

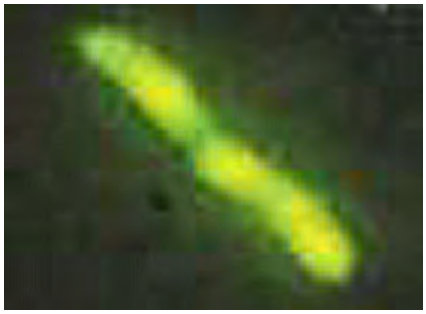


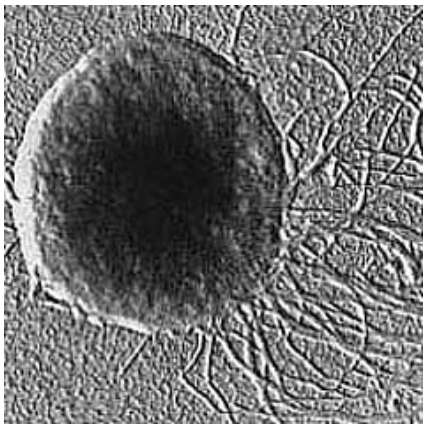
As we shall see on this and the next page, the diversity and complexity of life forms has

progressed from simple one-celled plants through a wide range of multi-celled plants and animals. The term "evolution" applies to this. In order to examine and relate this progression, it is necessary to classify and clarify lineages. This is embodied in the term **taxonomy**. In 1735 Linnaeus first proposed a system of classification that has remained in use as the standard. This diagram summarizes the hierarchy of this system:

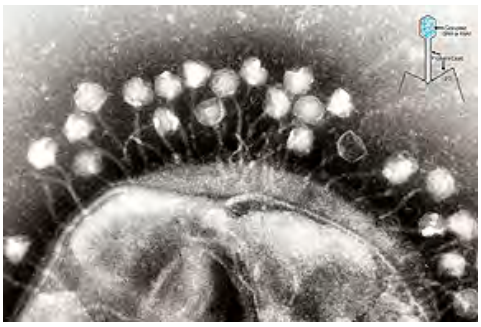


Let's now examine some of the life forms themselves, starting with the most primitive. As described above, almost certainly the first living bodies were microscopic in size, being single-celled. Bacteria were the dominant, perhaps the only, major life forms. Below are two modern day examples:





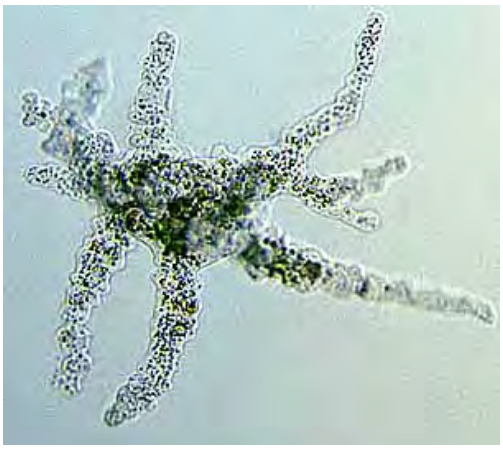
Bacteria are very abundant, as well as primitive. But the record for numbers - at least on Earth - are the viruses known as bacteriophages, which are attached to individual bacteria:



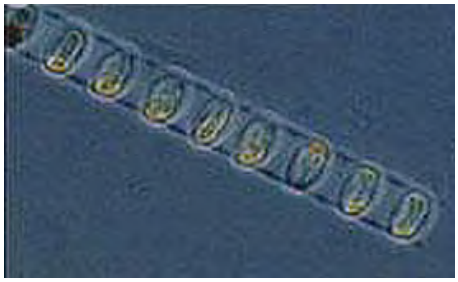
A more advanced modern unicellular animal is the Protist *Paramecium*, with dual nuclei and numerous tiny "whips" (cilia) for locomotion:

Protists first appeared about 900 million years ago.

Familiar to many is the *amoeba*. This Protist moves by extending parts of its cell as "pseudopods" in certain directions and then pulls the remainder of the cell towards one or more of these protuberances. This photomicrograph shows the Proteus species of Amoeba:

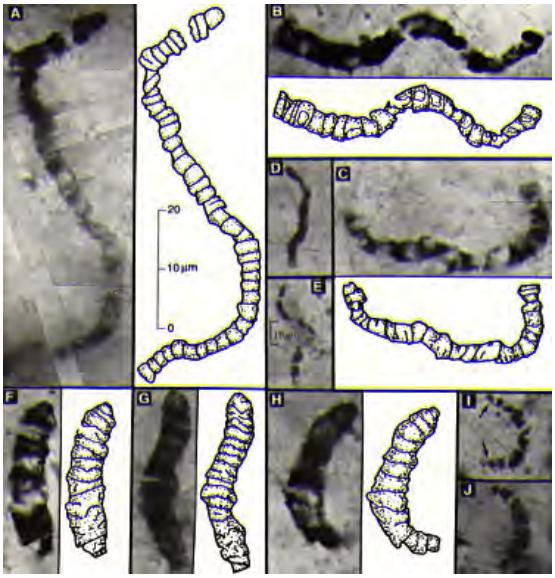


Claims of planktonic microbial life (animals and plants living at or near water surface; free-floating; largely microscopic; utilize photosynthesis in autotrophic or heterotrophic assimilation of foodstuff; in the oceans and lakes planktons are at the base of the food chain) as old as 3.8 b.y. in rocks from Greenland have been made. Here is a modern phytoplankton (microscopic plant)

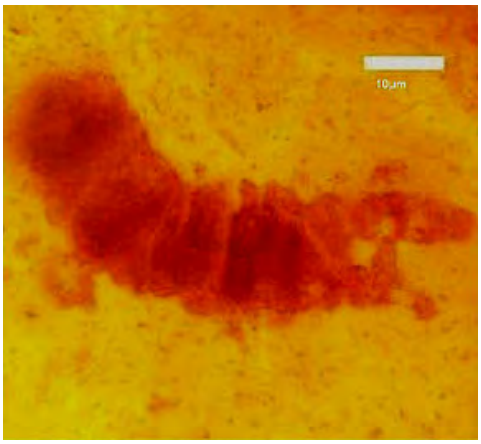


Generally accepted evidence of the oldest microfossils, cyanobacterial life, found in the 3.465 b.y. Apex Formation of Australia, has been published by J. Wm. Schopf (UCLA) and others. These remain the oldest life of any kind known on EarthIn the field, the Apex Formation is a chert unit that appears thusly:

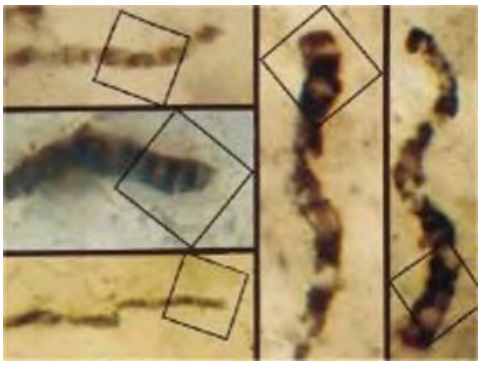
This next illustration is justly famous as the depiction of the oldest life form known, from the Apex Chert.



The type specimen from the Apex Chert is a cyanobacterium shown here after staining the sample orange.

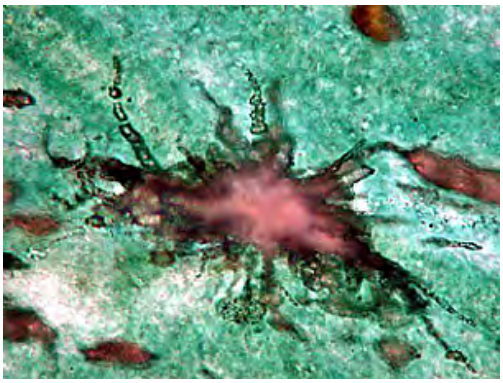


Here is another set of photos illustrating what has been discovered at the Apex locality:

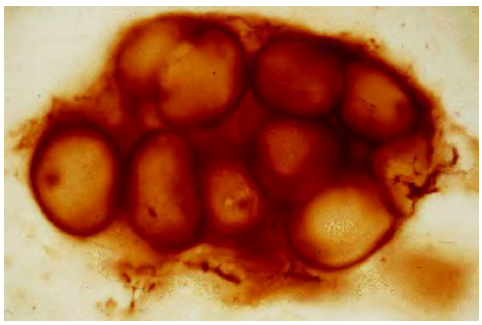


Microbial life has now been found in the rocks from the Barberton Formation in South Africa, of

3.4 billion year age. The example shown here is in a rock that was emplaced as a glassy lava before crystallizing. It is postulated that this life form actually "fed" on the rock material itself (now recrystallized into a basaltic type).

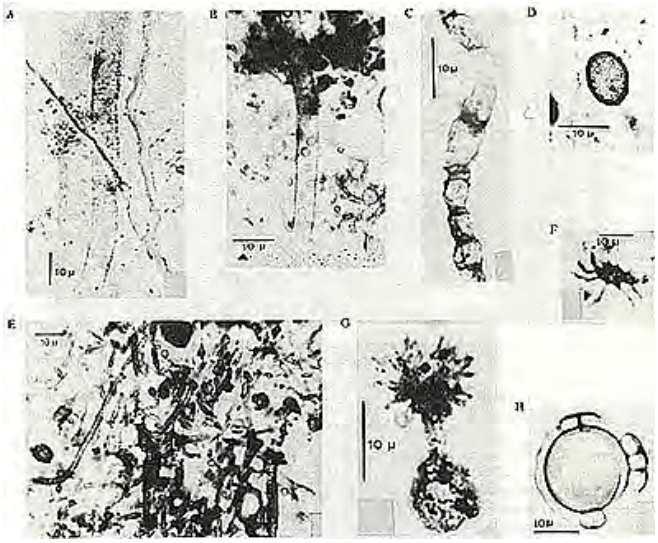


One of the prevalent life forms (perhaps as far back as 3+ billion years) falls in the general category of *cyanobacteria* (also known as blue-green algae). Fossil examples from two different ages are shown here:



Cyanobacteria were dominant for at least 2 billion years and some forms still exist today. They produce large amounts of Oxygen by photosynthesis (using sunlight to convert CO_2 and H_2O to simple sugar and free Oxygen. They played a key role in the transition of the Earth's atmosphere from reducing to a gradual buildup of Oxygen. One of the sedimentary rock types supposedly influenced by bacteria is the *Banded Iron Formation* (BIF) which occurs worldwide; it forms rich iron ore in Minnesota and Michigan. For an extended period, Iron in water envIronments grabbed the free Oxygen, slowing the buildup of that gas but in time the Iron was depleted and Oxygen then accumulated more rapidly. Here is an example of this BIF rock in which the red is rich in hematite:

Another famous locality containing a variety of ancient life forms is the Gunflint Formation (1.9 b.y. in age) in Minnesota and southern Canada. These are examples of microfossils found in rocks made up of chert from this unit:



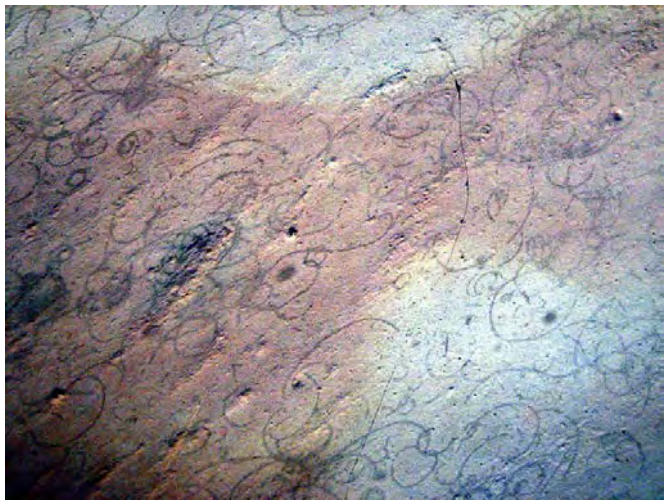
Other life forms were fungi and algae. The oldest and most famous of the larger fossils are the stromatolites of Western Australia. Stromatolites are mounds of prokaryotic algae and cyanobacteria. Modern stromatolites occur today along the Australian coast.

These bear resemblance to excavated ancient stromatolites found around Marble Bar in Western Australia dated at 3.45 billion years:

In cross-section these stromatolites have a conspicuous curved layering.

Stromatolites are also found in the Gunflint Formation, described above. Here is an outcrop on Lake Superior:

The first eukaryotic life forms may be as old as 2 billion years ago. Multicellular algae were common by about 1.2 b.y. ago; wormlike creatures (possibly animals) had appeared by then. Best known among the Eucaryotic algae is Grypania spiralis, found in ancient rocks in Michigan and in Australia. This fossil is preserved because it formed simple shells:



Life continued in this primitive state until about a billion and a half years ago when photosynthesis became a common process that helped plant life flourish and released Oxygen. The first multicelled plant is similar to modern brown algae; pieces of this have been found in Chinese rocks dating to about 1.88 billion years ago.

In the Proterozoic (2.5 billion years to 540 million years), life forms slowly changed, with eukaryotic (nucleus) single-celled phylla becoming more diverse. Protozoa (primitive forms) were joined by Metazoa (more advanced) about 1 billion years ago. The first precursors to animals were ancestors to the sponges. Here is an example:

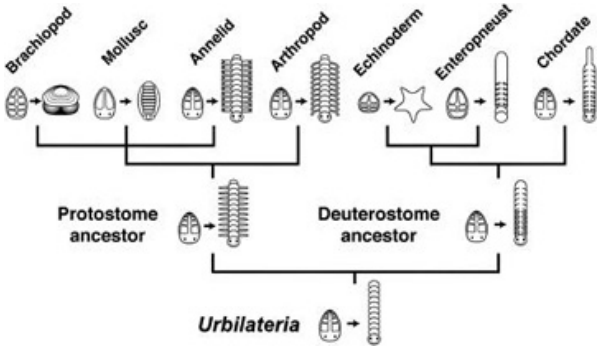
The oldest multicellular animal yet discovered is a sponge(like) fossil from rocks in Oman. These date at approximately 710 million years. No actual fossil form was recovered (sponges only have soft parts) but instead a distinctive chemical was found in the carbonate rocks. The telltale sign present in the rocks is a fatty chemical called 24-isopropylcholestane, which scientists have found only in the skeletal structures of demosponges, the most common member of the sponge family

Claims of a unicellular fossil trace in Australian rocks that are about 1.2 billion years old are still in dispute as the evidence could be the result of inorganic activity. Thiomargarita is the name given to markings in Chinese rocks about 600 million years old that have been interpreted as eggs or embryos, and by others as bacteria

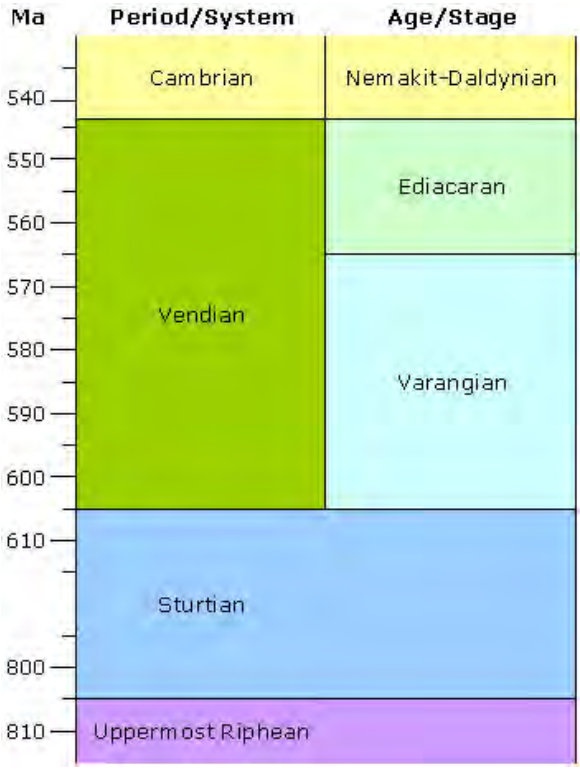
The greatest explosion of life in earth history took place about 600 to 500 million years (late Proterozoic into Cambrian) ago with the appearance of distinctive and diverse animal forms. One significant marker was the first appearance of animals with *bilateral* symmentry at least 600

million years ago. These animals were very small, generally microscopic, and usually found in shales; the samples had to be thin-sectioned by guesswork to find the tiny objects which were preserved soft parts. Representative of these is Vernanimalcula, found in China:

Some taxonomists place the appearance of bilateralism even earlier - one school maintains about a billion years ago. The common ancestor to all animals has been named Unilateria (no such fossils have been found). A general scheme for the evolutionary chain of animals follows this lineage:

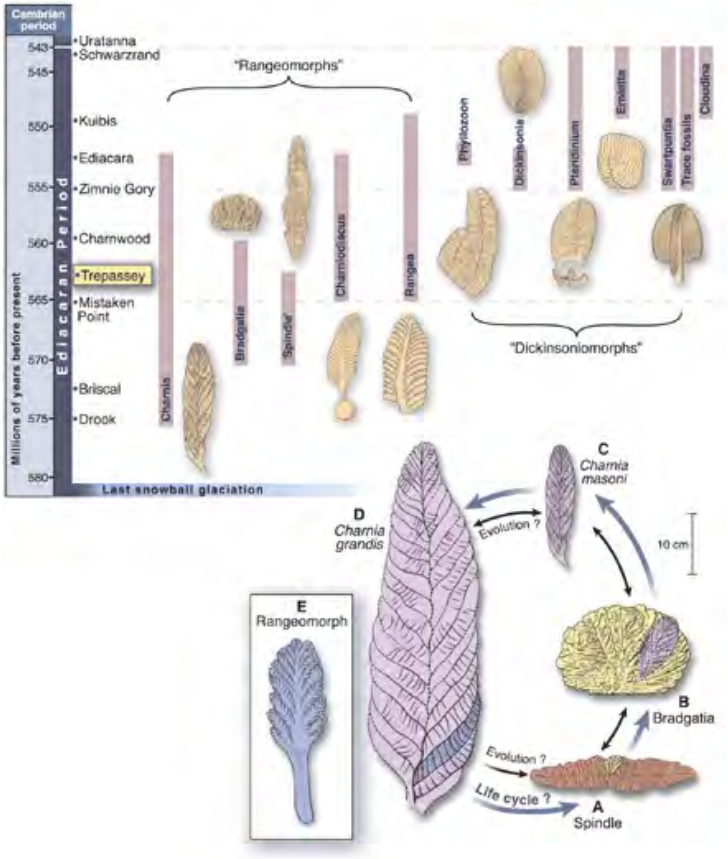


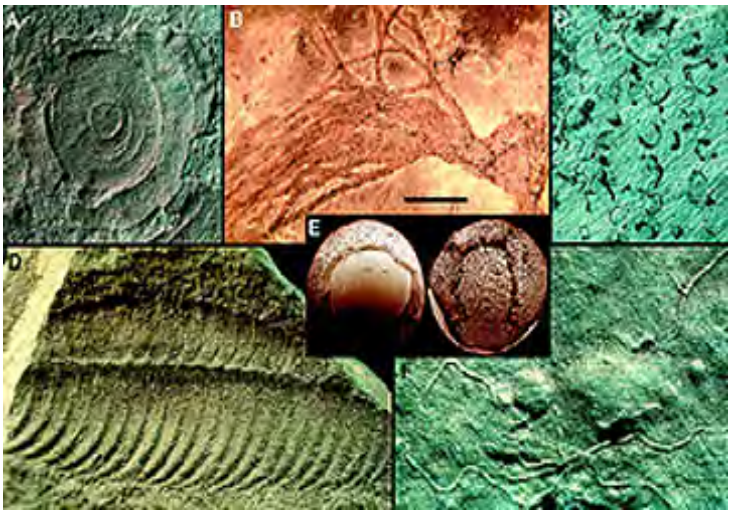
A significant upswing in the numbers and diversity of life marks the last 200 million years of the Proterozoic (often referred to as the "Neoproterozoic"). This is a time-stratigraphic chart that introduces Periods that will now be discussed:



A general view of the life in the time frame from about 605 to 542 million years ago (the Vendian), is found at this New Zealand site which concentrates on the Ediacaran epoch; it mentions Australian and other geographic localities where the assemblages have been found. The fossil life is represented entirely by creatures with soft parts only. It is suggested that these may be ancestral to later phylla observed at the beginning of the Paleozoic. Below is a chart

presenting typical Ediacaran fauna, followed by an artist's depiction of life on the sea floor at that time, and beneath that is a layout of some actual fossils:





Two of the most common guide fossils to Vendian life are shown here as actual specimens.

By the opening of the Cambrian, many forms of invertebrate life had developed (mainly for protection) external carapaces or coverings that, after death, survive as fossil shells. A survey of Cambrian stratigraphy indicates a dramatic increase in diverse marine life forms, so much so that this abundance of living creatures has been referred to as "The Cambrian Explosion".

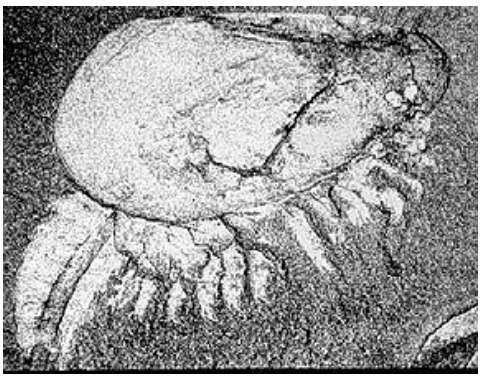
Around 535 m.y. ago, in the early Cambrian, a soft parts assemblage of fossil forms has been found in the Chengjiang Formation in Yunnan Province of southern China. Here are typical fossils;

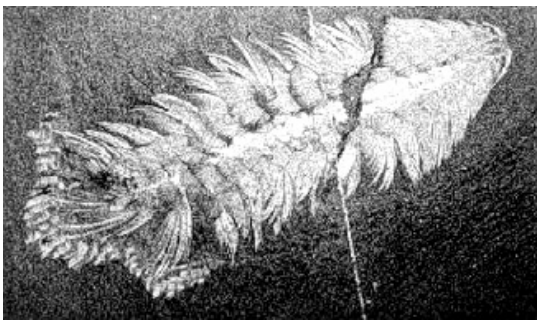
The life forms with hard parts make their first appearance in the Middle Cambrian. By far the best locality where both hard and soft parts are well preserved and displayed is in the Burgess black shale of British Columbia (Prof. Charles Walcott is famed for his pioneering studies of this assemblage) deposited about 520 m.y. ago . This sudden burst of evolution may have been tied to Oxygen reaching threshholds near the present day levels. A good review of the types of fossils found there (MacKenzie Mountains) is given by the Peabody Museum at Yale University. Here is an outcrop of this black shale near Mt. Burgess.

This is an artist's conception of typical animal and plant life in the shallow sea in which the Burgess shale was deposited:

Some of the typical life forms in the Burgess shale, each fascinating in its own right, are shown below:

Two animals similar to Trilobites.

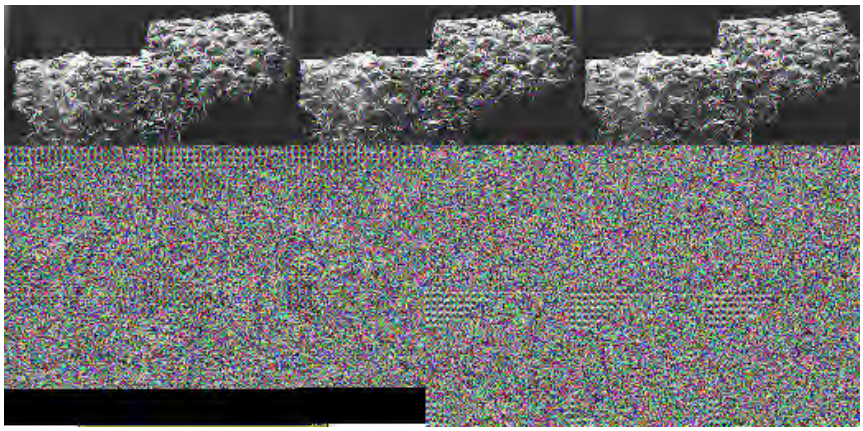




An excellent review of early life on Earth is available at this web site maintained by the University of Munster (unfortunately, many of the links no longer are active).

A second peak time in the abundance of shell-surviving life forms was in the Upper Ordovician (by this time also, the first larger vertebrates, fossil fish, had appeared). Below are two illustrations: the first, an artist' conception of marine invertebrate life in the late Ordovician; the second, a typical slab of Ordovician limestone (from Indiana) containing the fossil types listed in its caption:

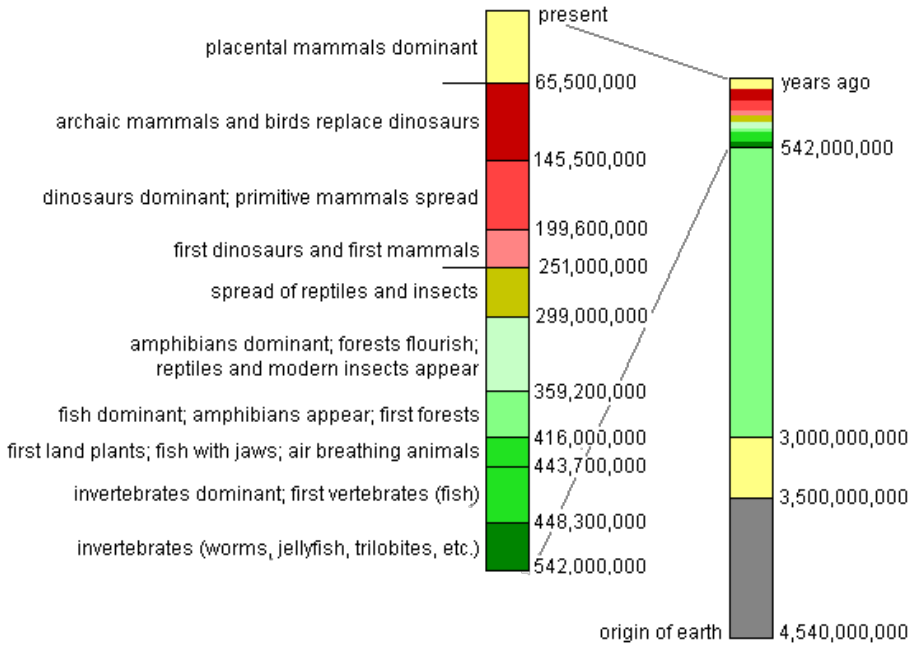
The oldest known vertebrate life may be a tiny fish (8 cm in length) called Anatoleptis, of very late Cambrian age (510 million years; younger specimens (470 m/y.) have been found in Scandanavian rocks. The photo below is a microscope view of scales from this fish, whose remains have been found in Wyoming and other parts of North America:



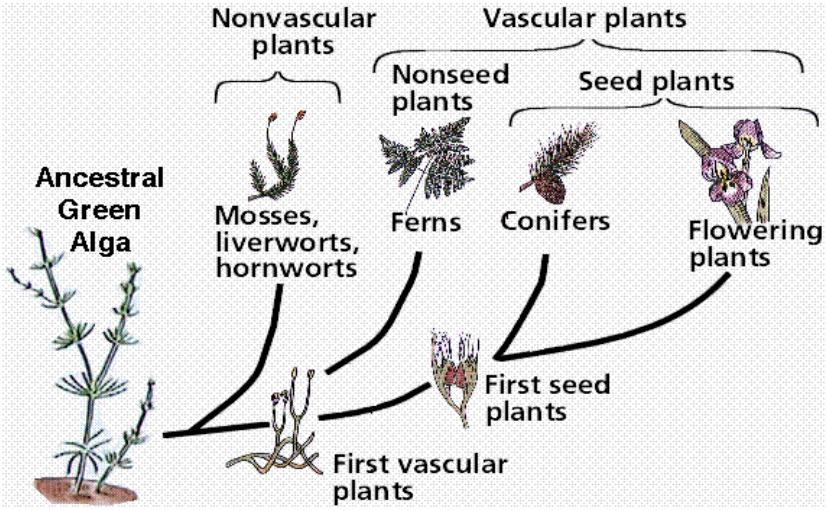
Almost nothing is known about life dwelling mainly on land. A small millipede found in rocks in Scotland, dated at 428 million years ago, is a candidate for earliest terrestrial life:

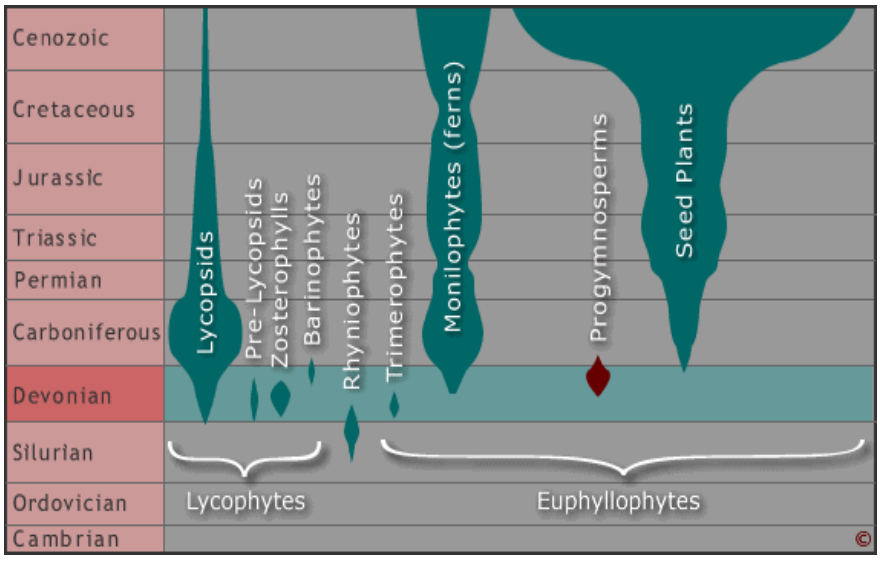
There have been several extended time spans (but occuring mainly within single Periods) when major spurts in evolution have witnessed appearances of whole new classes of animal life forms. The Cambrian Explosion described above is the prime example. In the Devonian, the first amphibians (land-dwelling animals), insects, and land trees have been found in the fossil record. A "missing link" (375 million years old) between fish and amphibians was found on Ellesmere Island in the Canadian Arctic. Known as *Tiktaalik roseae*, it has a crocodilelike head, was 3 meters long, had fins with articulation that foretold jointed legs, and probably could wiggle like a seal on land when not in the sea. Here it is:

We have alluded to the huge increase in animal life since the "Cambrian Explosion" witnessed in the Burgess Shale. This is a good place in the narrative to summarize animal life since the end of the Precambrian.



Although the most ancient life was a mix of single celled plants and animals, we have not touched upon the more advanced forms of life since the Precambrian. A major and vital evolutionary spurt was when marine plant life adapted to living on the land. This began about 480 million years ago. These next two diagrams show the history of land plants:



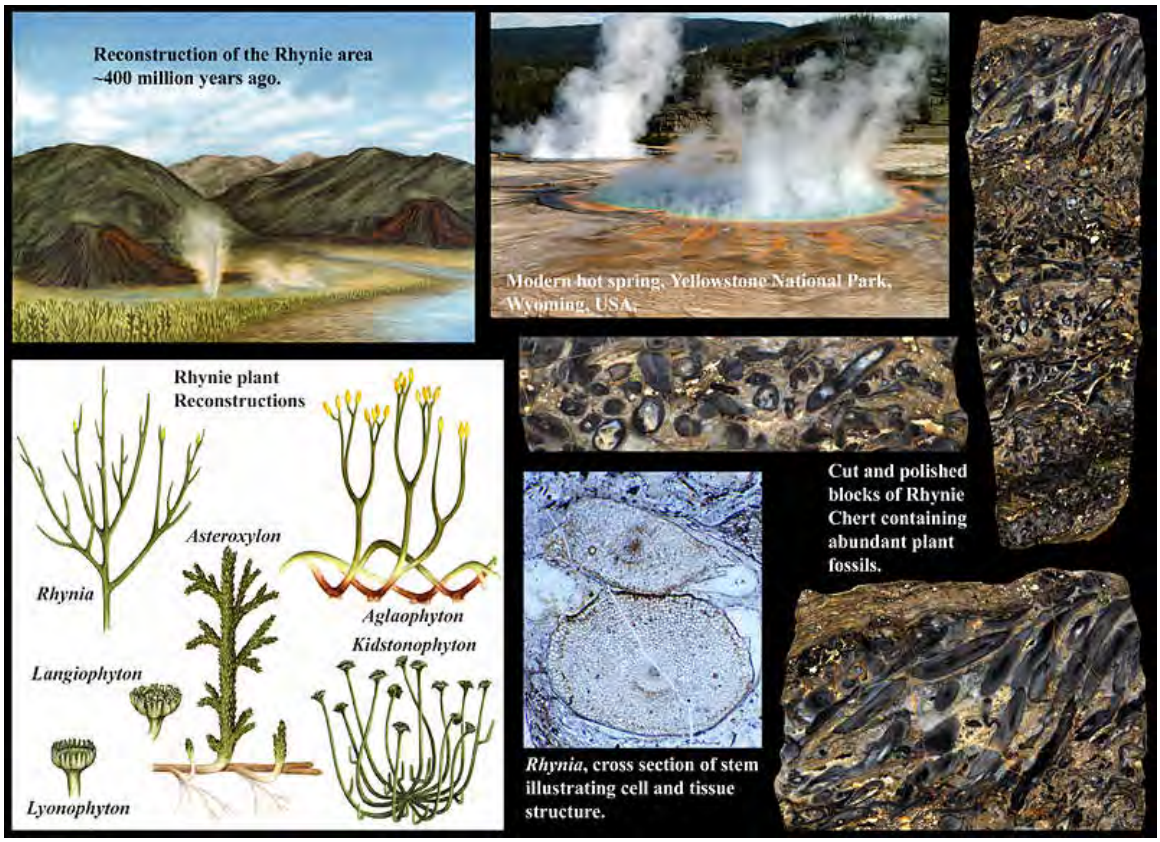


The illustration below show modern brown algae which were likely progenitors of the land plants.

By the Devonian land plants, which had started as ferns, club mosses, horsetails, and liverworts (all of these still exist), had established vascular systems, acquired the ability to root in the soil, and were able to propagate using seeds. One of the first giant trees is Archaeopteris, shown in this drawing:

A typical plant in this early stage of evolution is Cooksonia, shown here as a fossil frong:

The best known locality for early land plant fossils is the Rhynie chert deposits of Scotland:



This drawing below is a reconstruction of a typical Devonian forest:

The Carboniferous forests are among the most diverse in geologic history. Their decay in swamps has produced coal beds world wide.

Among the giant plants in the Carboniferous forest shown in this panel were Cordaites, an early relative of conifers; Calamites, a bushy horsetail; Medullosa,a seed fern (a plant with seeds and fern-like leaves); Psaronius, a tree fern; and Paralycopodites and Lepidophloios, lycopsids (scaly, pole-like trees with cones). Lepidophloios could grow to 40 m (132 ft), but most of today's lycopsids, known as quillworts and club mosses, grow only a few centimeters high.

Illustration by Mary Parrish © Smithsonian Institution

The illustration below shows a cluster of leaves of Pennsylvanian age:

The subject of land plants has been glossed over in the above paragraphs. For those seeking more information, consult the Review of plants and the Earliest land plants web sites.

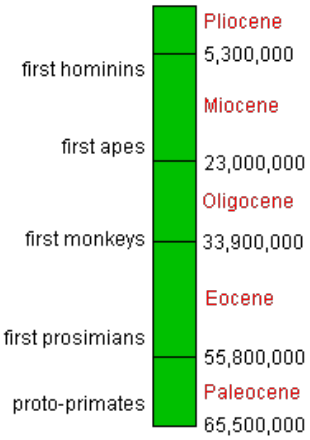
As mentioned in Section 18 and elsewhere, some of the major advances or disappearances in life forms are being attributed to worldwide effects of large impacts. But other causes of these compressed spurts or declines are probably involved in other cases: extreme volcanism; dramatic climate changes; continental splitting, are suggested. These may be manifestations of the "Punctuated Equilibrium" mode of evolutions proposed by Stephen Jay Gould and others.

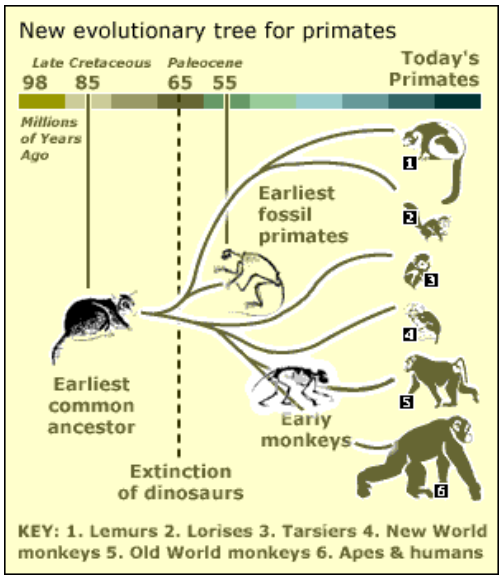
A rather fanciful panorama of life forms (Kingdoms, Phylla, Families, etc.) from the Late Precambrian to the Present is shown in this mural that is found on the Humboldt State University campus. On the left, pre-Paleozoic animals with soft and hard parts give way to the Invertebrates of the Early Paleozoic, the first fish (Ordovician; sharks in the Devonian), then Amphibians that appeared at about the same time (Mississipian) that land plants took root, with the first reptiles and dinosaurs near the end of the Paleozoic (about the time the first extensive forests spread in the Pennsylvanian), reptiles and small mammals in the Mesozoic, along with the first birds, and finally a dominance of mammals, flowering plants, and widespread forests in the Cenozoic.

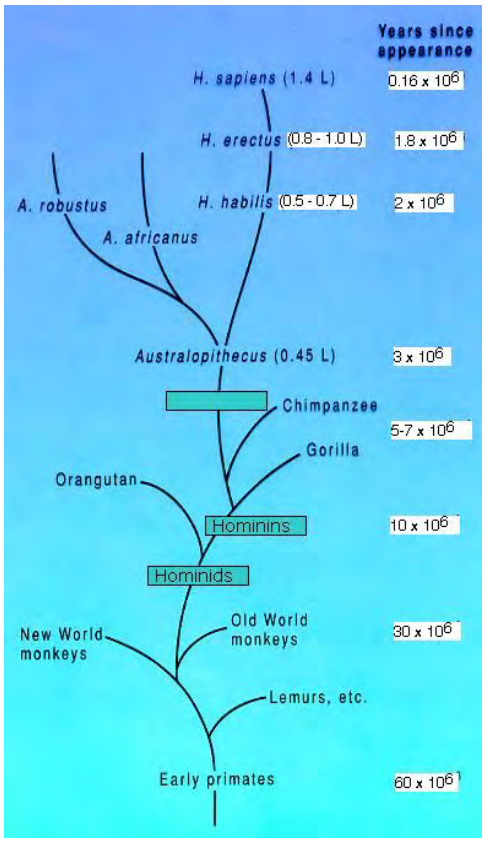
To see entire panel, scroll bottom bar to the right

Missing from the far right of this panel is the story of the hominids, which includes today's mankind. Ancestors to the hominids have been found as far back as 4.5 million years (older animals that may be links are still controversial). This subject is far too involved for any extended treatment in this Tutorial, but the following synopsis touches upon many of the key ideas. The study of ancient precursors to modern humans, is called Paleoanthropology; see this Wikipedia site for a review of mankind's genealogy.

A good, quick overview that carries back to the beginnings of life 4 billion years ago but gives some emphasis to the appearance of humans is found in the Wikipedia Timeline of Human Evolution webpage. Before embarking on the remainder of this page, you should benefit from looking at the list of hominids and their precursors at this Wikipedia list of hominids. These three diagrams are also helpful. The first is a chart that gives a general time line through the Cenozoic and into the end of the Mesozoic; the second shows one version of primate evolution; the third a variation of this with hominids included and broadly spelled out:







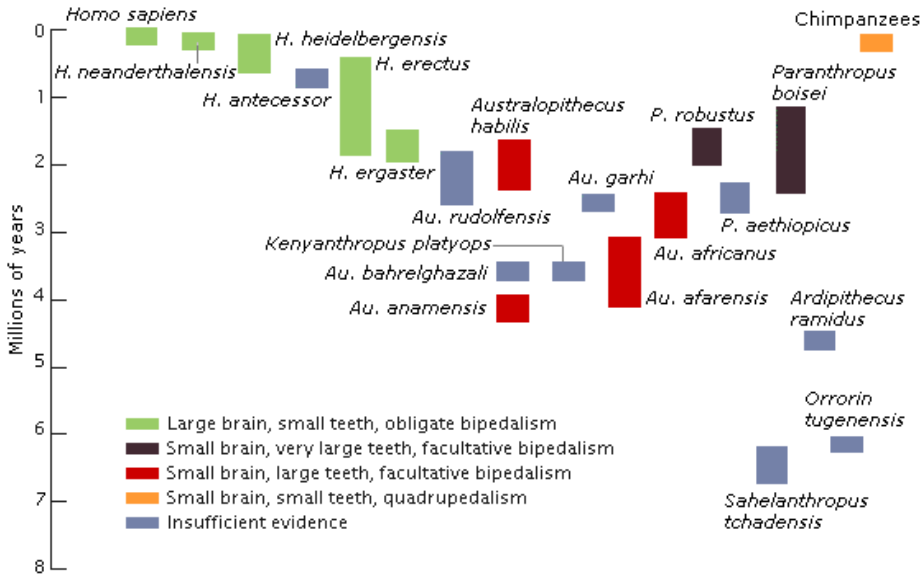
Classification of the Primates show two broad groupings: 1. The Prosimians, made up of Lemurs, and 2. The Anthropoids, that includes Old and New World monkeys, the Apes, the Chipanzees, and Humans. Of the Anthropoids, there are two major divergences over the last 30 million years. One line includes Old World monkeys, the other contains several branches of which one is the hominids. While the closest tie between Man and other similar animals is said by many to be with the Gorilla or other Apes, the Chimpanzees are also related (98% of the thousands of DNA

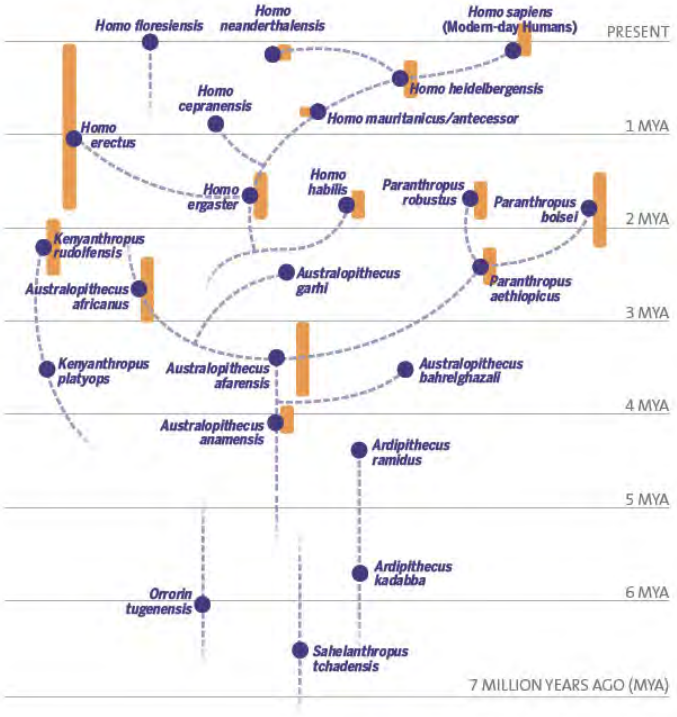
genes in the genomes of Man and Chimp are identical and located in the same relative positions).

Much publicity has been given to an announcement in May of 2009 of what is called by some as the "missing link" ancestor of both Primate Groups. A nearly complete primate fossil was found in the 1980s in the Messel shale quarry near Frankfurt, Germany. The rocks there are 47 million years in age, placing the many fossil types at Messel in the Eocene.The primate's formal name is Darwinius messelae, but it is called "Ida" (after the daughter of the scientist who led the recent research on it). These two illustrations are pertinent:

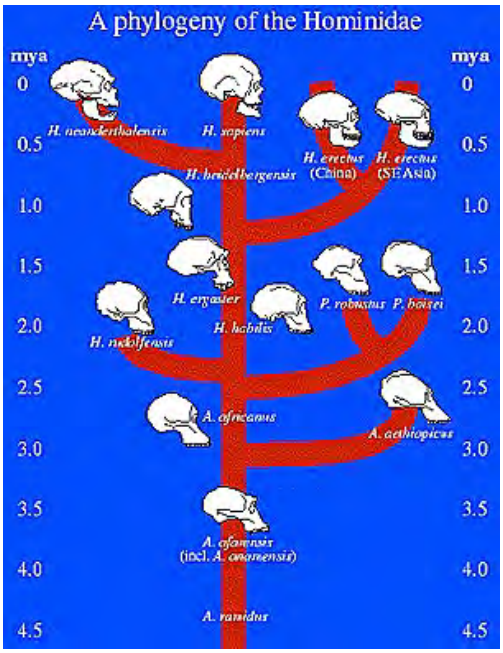
The fossil appears to be a transitional link between the Prosimians and the Anthropoids. It has some lemurlike qualities but lacks certain features of that group while showing some bones that link it to the Anthropoids. Of particular import is the opposable thumb in the hand; this is what is characteristic of primates among mammals and a trait that facilitated the ability of humans to evolve into civilized Man.

We will now concentrate on human evolution, as determined mainly from the fossil record. As a generality, the genus "Homo" is associated with modern humans and direct ancestors whereas the genus "Australopithecus" and others refer to pre-humans. Two diagrams are introduced here. The first emphasizes the time spans during which the various genera lived; the second shows one model (there are competing versions) of the lineage, as a family tree, of the anthropoids back to about 7 million years.

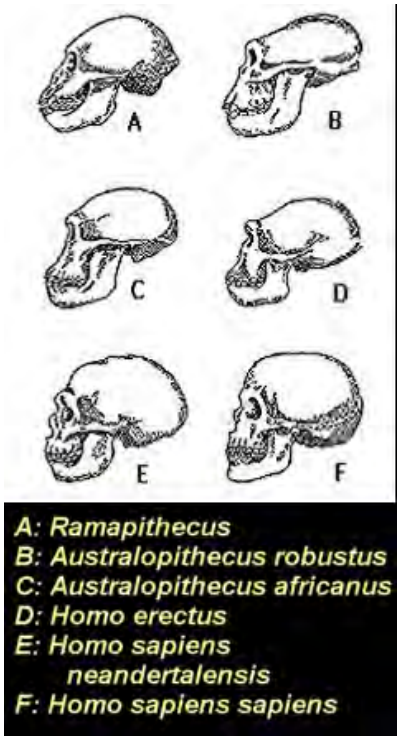




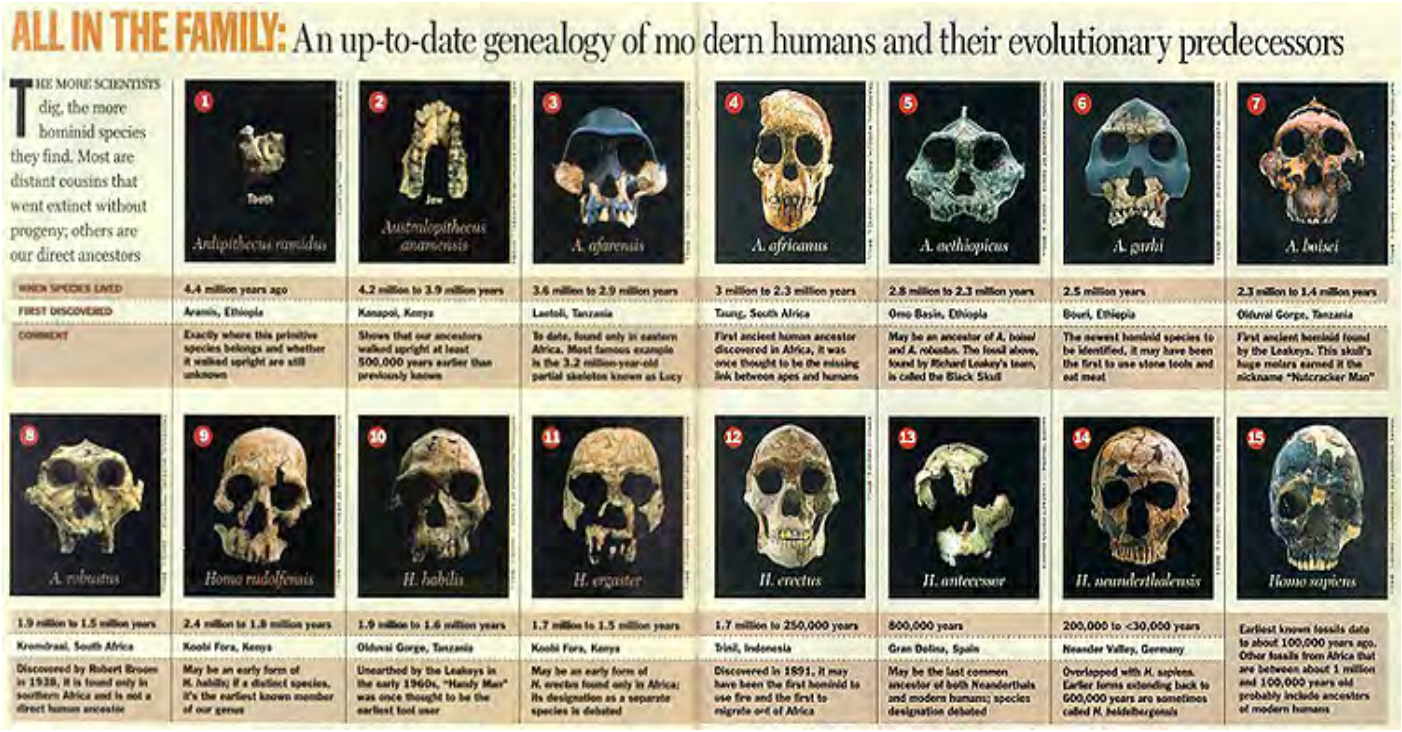
Part of this progression can be displayed in terms of an evolutionary pattern that depends on the principal evidence - interpretation of skeletal remains, in this case reconstructed heads.



The diagram below shows the skull differences in a different rendition:

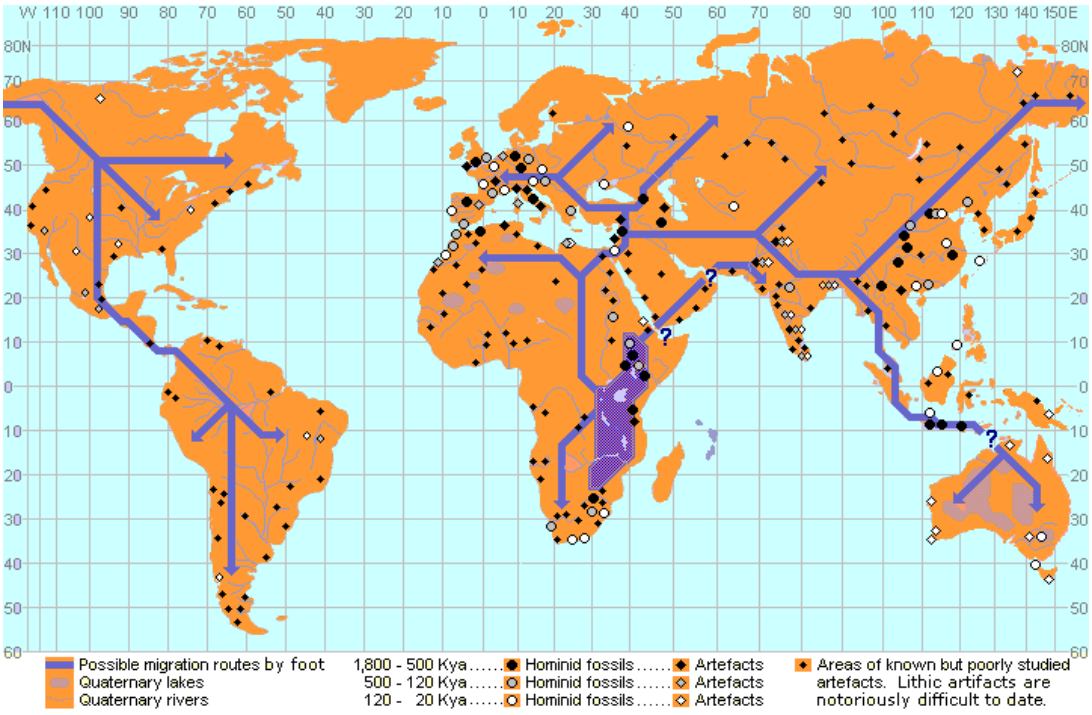


This next diagram was taken off the Internet; it is a reproduction of an August 1999 feature article in Time Magazine. It is worth examining before we add comments:

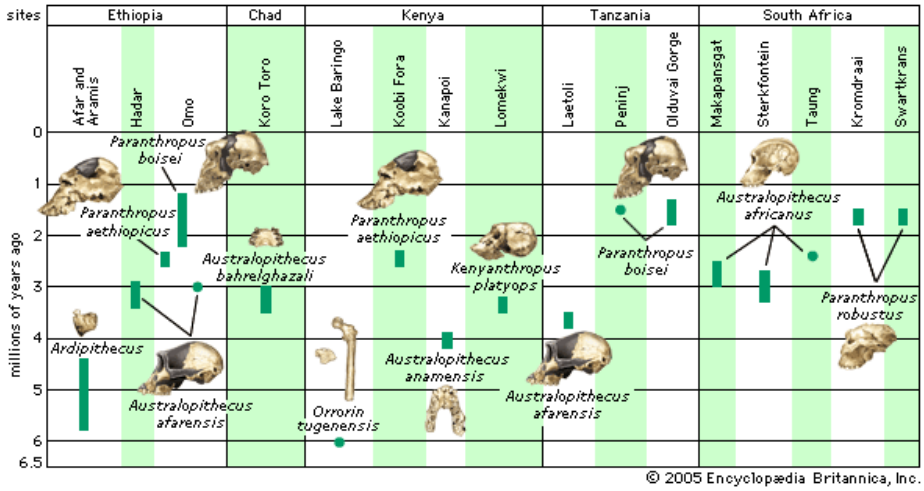


Fossil bones of various hominids have been found on all Old World continents, as evident in this map:

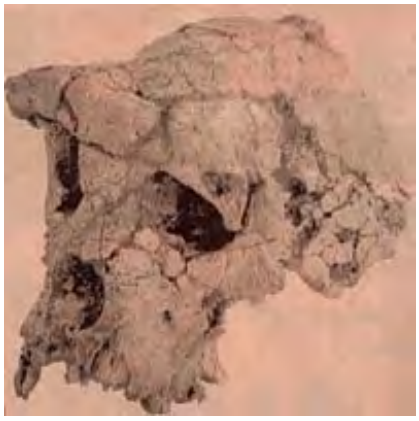
This next map incorporates the above map but shows the general pathway to the New World (N. and S. America). The blue lines are one of several proposed pathways from central Africa to the other continents (discussed below):



To most Americans, the Leakey family are the best known paleoanthropologists. Mary and Louis Leakey (and now their son, Richard Leakey) established the first lineage for the ancestors to genus Homo with their discoveries in Olduvai Gorge in Tanzania. They along, with other scientists, have visited several African countries, with usually more than one locality in each country, with finds that are pictured in this diagram:

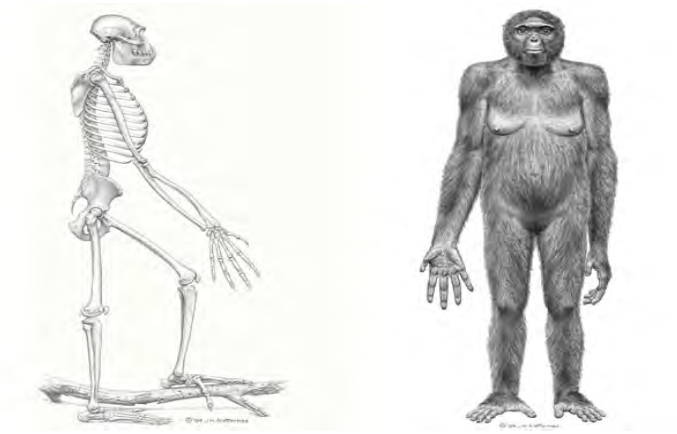


The "holy grail" for paleoanthropologists is the so-called "missing link" - the skeleton that can be considered the "starting point" for the hominids. This will likely be somewhere in Africa. Several candidates have now been found, but the experts do not agree on whether these are actually ancestral or are some offshoot in the evolutionary pathway back to the first creature than can be definitively identified as the starting point. One such candidate is named Sahelanthropus tschadensis, and is about 6 to 7 million years old.



From this, anthropologists can envision the appearance of Sahelanthropus, as shown below (this version is hairless but the creature likely was hairy like an ape):

The term "hominid" goes beyond what we picture as humans - it is applied to two-legged animals that are **bipedal** (walk upright most of the time), which includes some of the apes. New fossil finds (generally parts of skulls and leg bones) in recent years suggest that two species are important progenitors to human lineage: Orrorin tugenensis (~6 million years ago) and Ardipithecus ramidus (4.5 m.y). Ardipithecus ramidus was discovered in Ethiopia in 1994 and is now considered to be the best documented precursor to the main line of hominids. Evidence from skeletons about half complete suggest that "Ardi" walked upright but also moved about trees. The next two illustrations show first the best example of a A. ramidus skull, next a reconstructed full skull and then a reconstructed complete skeleton from which a drawing of Ardipithecus ramidus has been produced:

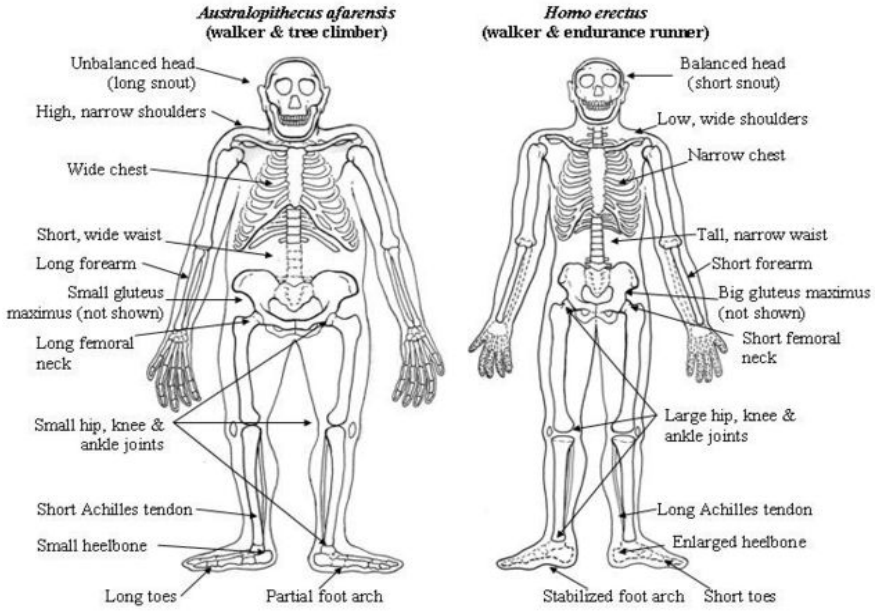


An informative discussion about these early hominids is found on this Wikipedia website.

The genus Australopithecus first appeared about 4.2 m.y. ago. The australopithecines are generally accepted as the progenitors of the hominids. A direct ancestor of the human race is generally agreed to be Australopithecus afarensis. The most famous of all paleoanthropological finds is "Lucy", of this species, found in northern Ethiopia (by Donald Johanson and his colleagues) in volcanic deposits dated at 3.2 million years. Here is her partial skeleton:

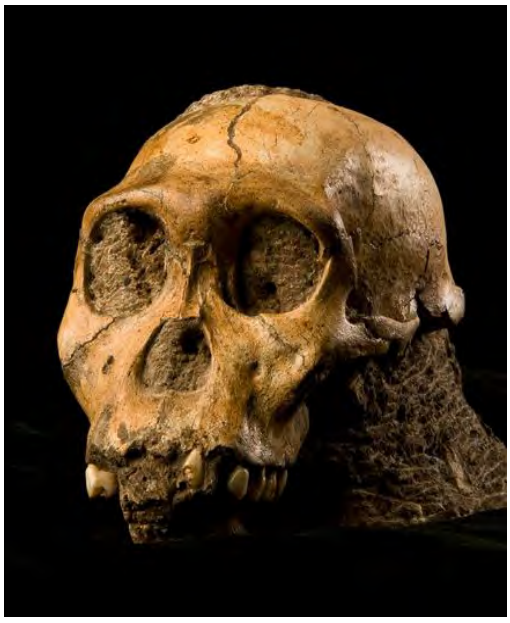
The skull of Australopithecus afarensis is distinctive, in part because of its rather close resemblance to genus Homo:

No complete Australopithecus skeleton has yet been found, but parts from different individuals help to fill the gaps. This illustration shows a reconstructed complete Australopithecus skeleton compared with modern Homo sapiens:



The vast majority of australopithecines and earlier hominids are known almost entirely from fossil remains found in Africa. Skeletal remains older than 100,000 years are rare anywhere else. This seems to argue for Africa as the ancestral home of those animals related to Man and to early contemporaries of Homo sapiens. There is some evidence for migration out of Africa of several genera but only Homo erectus so far can be cited as a species that left Africa before the last 100,000 years.

The latest find in the search for ancestors is Australopithecus sediba. Two partial skeletons of this new species were found in South Africa. They date between 1.7 and 1.9 million years ago. They have some characteristics of the genus Homo, so that their discovers claim that they may be transitional to that group. Others dispute this. The growing opinion among paleoanthropologists is that there may be no true "missing link" in the lineage tree. Instead there has been evolutionary development of various species among the several homonid genera (in parallel rather than series).



The genus Homo goes back to just over 2.3 million years; this branch evolved from the australopithecines (A. in the diagram) that extend back at least 4+ million years. Here is a skull of Homo habilis:

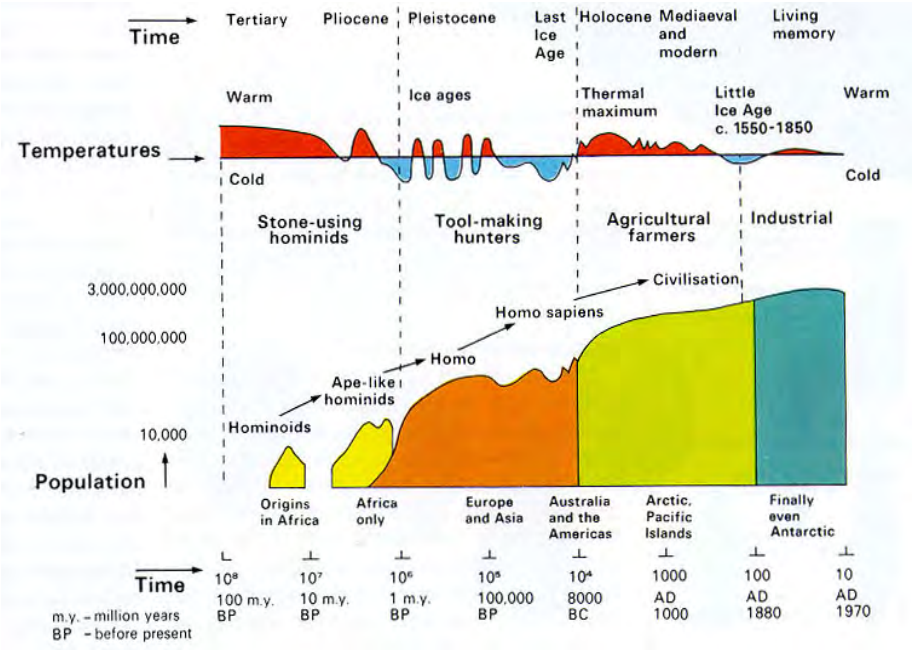
This is an artist's view of Homo habilis:

Homo habilis used crude tools - stones for breaking bones, and possibly cutting tools made by chipping flint. There is debate as to whether H. habilis was a hunter (killed animals for food) or a scavenger (fed off of kills done by other animals). The species lived mostly on the ground (in the African savannah [grasslands]) but took to the trees for safety.

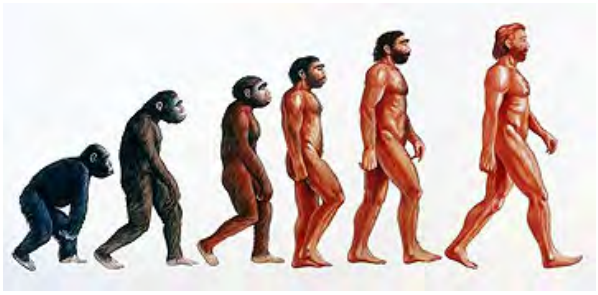
Homo erectus is presently held to be the immediate ancestor to modern Man. Homo erectus, who lived from about 1.8 million to 300000 years ago, was tall (more slender, and often above 6 feet in height). Homo erectus may have been the first homonid to develop a crude language. Bone remains of H. erectus have been found in Europe and Asia (Peking Man), even to Indonesia (Java Man). Just where H. erectus originated is not yet established; Africa is one possibility. Here is an artist's rendering of a group of H. erectus primitives hunting in the jungle:

The last few paragraphs have introduced a lot of new information. A good overview/review is found at this website:Hominid species.

Climate played a role in the evolution and migration of hominids. Warm periods, and especially ice ages, were prime factors. This chart bears on that concept:



What would become modern humans - Homo sapiens - appeared about 200,000 years ago. Prior to that a form called archaic Homo sapiens co-existed with later H. erectus. H. sapiens probably evolved from H. heidelbergensis (as did H. neanderthalis). We show below a famous painting that represents the major "players" in Man's ancestry. Read the caption for identity of each individual:

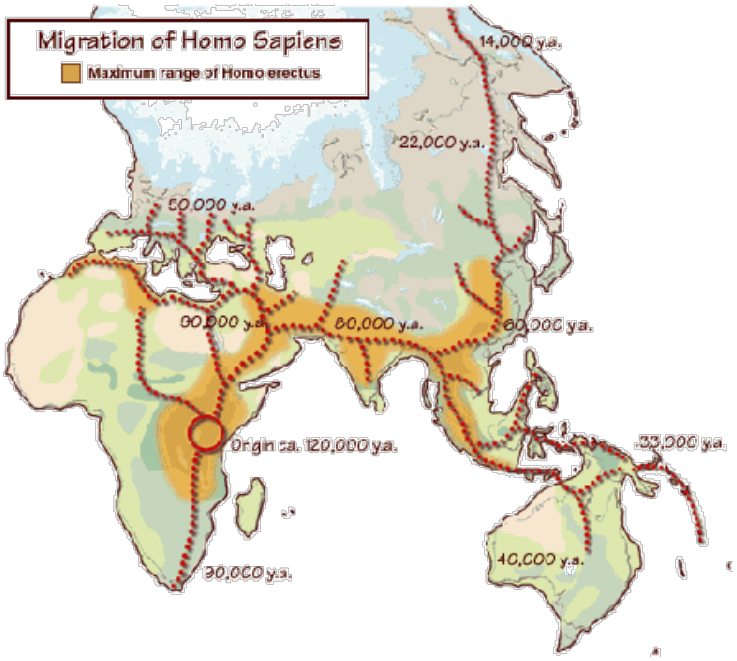


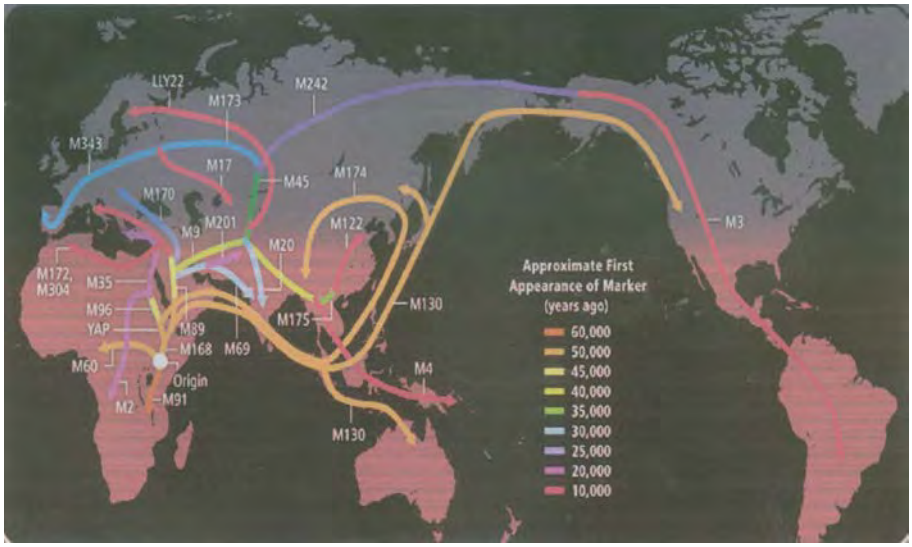
There are two schools of thought about the geographic origins of H. sapiens. One, the Multiregional group (sometimes referred to as "Eden"), considers that various strains of H. sapiens developed from Homo erectus in several regions of Asia and perhaps Europe. In this view, Homo neandertalensis (which settled mainly in Europe and the Near East) began in the western part of Asia and moved into Europe. If the Multiregional hypothesis proves valid, one or more archaic H. sapiens are the direct result. The second school is the Out-of-Africa group. Currently, this is now the more favored hypothesis. A good brief summary of the two hypotheses has been put on the Internet by Paraminder Dhillon.

In the last decade, mitochondrial DNA analysis (identifying the genes and breaking the organizational code to establish their sequence;) has led to strong evidence that H. sapiens originated in the savannah of eastern Africa. Tracing the genetics back in time has identified a baseline DNA case represented by a single individual, whimsically named "Eve" by paleobiogeneticists. Populations were small, not well intermixed, and adapted to warm climates. Based on evolutionary rates, Eve may be as old as 200000 years, but since no fossils of that age have yet been found, this progenitor may in fact be younger.

Thus, it appears that Homo sapiens began in Africa, from the Homo erectus progenitors that had settled there, and remained on the continent for about 150000 years. Some 40 genetic units lived there but were largely isolated from one another. One group is represented today by the "Bushmen" but most of the other units have died out or dispersed. There is some recent evidence that suggests H. sapiens almost died out, probably because of severe drought.

Most of the migration out of Africa has taken place over the last 60000 years. The sparsity of fossil remains has made it difficult to be precise in tracing the dispersion of Homo sapiens. These two diagrams are among several the writer found on the Internet.





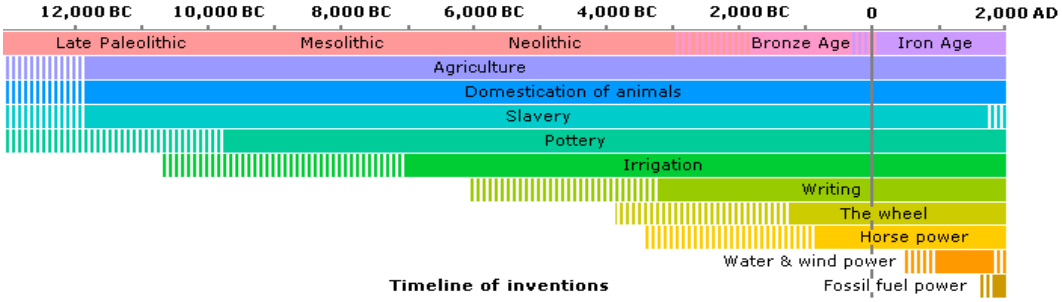
Mitochondrial DNA evidence has allowed tracing of these African units both within that continent and more recent dispersion. About 60000 years ago, some of the African units left that continent (again, drought is suspected as the cause of a need to find more find and amenable environments) and entered the Middle East and western Asia. These spread further reaching eastern Asia and Australia about 50000 years ago. Some clans moved westward into Europe about 35000 years ago (where Cro-Magnon [of cave-painting fame] co-existed for a while with the Neanderthals. (The Neanderthals, stocky and robust, were at the time better adapted to live in a climate then dominated by glacial conditions; the mental picture of a "stupid" subhuman developed when an anthropologist released [about a hundred years ago] a picture of a Neanderthal that portrayed this species as a brutish caveman-like cousin to the apes has been debunked by DNA studies (the brain of Neaderthals is actually about 20% larger than H. sapiens); the smarter Cro-Magnons may have been responsible for the demise of the Neanderthals, either by interbreeding or more likely by necine force.) Others moved across the Bering Strait into North and then South America, with notable migration into these continents about 15000 years ago (there is evidence of even earlier Man in parts of western N. America). During these last 60000 years of Exodus, genetic variations gave rise to the races and strains we observe today. A major factor in development of the races was the need to adapt to changing environments that contained different exposure rates to solar Ultraviolet rays; this UV is moderated by Vitamin D that controls the adaptation of body pigments through mutations. The original race within H. sapiens is closely related to the Negro of today.

Most readers of this Tutorial know a bit about the co-existence of Neanderhals and Cro-Magnons. The latter is one of the groups of Early Modern Humans (EMH) that dispersed worldwide. The Neanderthals were once considered a subspecies of EMH but most paleoanthropologists now consider them to be a separate species (see entry on Neanderthals at this Wikipedia site). Comparisons of the skeletons of the two hominids show distinct differences:

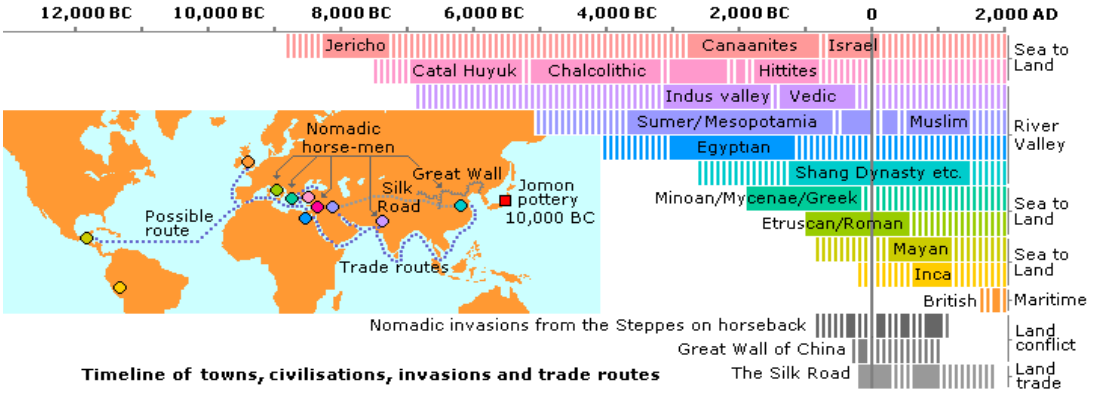
The difference between a Neanderthal and a 21st Century Caucasian man is suggested by this intriguing image:

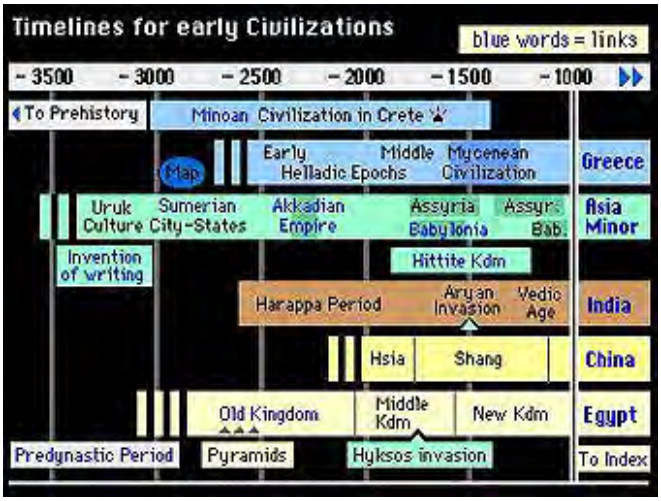
Cave wall painting is often cited as the clear sign of a distinct culture (for a summary of this art, move to this Wikepedia site. The earliest attempts at art may be markings in caves on the Nullarbor plains of South Australia. Wall paintings in southern European caves may be as old as 35000 years. This famous example, dated at 16000 years ago, is on a wall in a cave near Lascaux in southern France:

Anthropologists classify hominid pre-history using these terms: Stone age (subdivided into Paleolithic - a span over beginning more than a million years in which several Homo species used crude rock tools; Mesolithic - about 12000 to 7000 years ago; and Neolithic - 7000 to about 5500), the Bronze Age (5500 to 3200 years ago), and the Iron Age (starting about 3200 years ago). There is no evidence for when language - the use of spoken sounds to communicate ideas - began but it may have been earlier than 25000 years ago. Crop cultivations may have started about 12000 years ago.

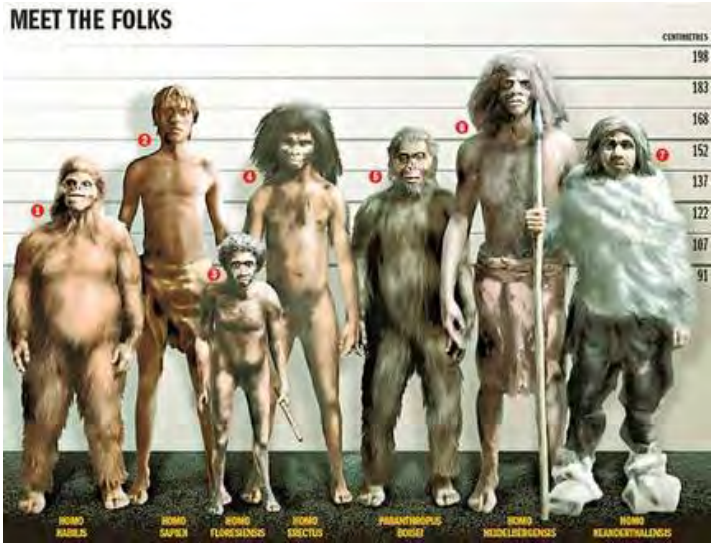


Early civilizations (defined in a narrow sense as having established large groups of Homo sapiens with distinctive cultures that are recorded as written histories) began in Mesopotamia, the Nile valley of Egypt, the Indus Valley region of modern Pakistan, in the Huang He (Yellow River) valley of China, and on the island of Crete in the Aegean Sea. These peoples built cities, created writing systems, learned to make pottery and use metals, domesticated animals, and created complex social structures with class systems. Two complimentary sets of timelines for these civilizations are:





We close this page with a rather interesting and humorous depiction of most of the accepted members of the hominids, shown side by side:



1 HOMO HABILIS ~ NICKNAME: Handyman LIVED: 2.4 to 1.6 million years ago HABITAT: Tropical Africa DIET: Omnivorous – nuts, seeds, tubers, fruits, some meat

2 HOMO SAPIEN ~ NICKNAME: Human LIVED: 200,000 years ago to present HABITAT: All DIET: Omnivorous - meat, vegetables, tubers, nuts, pizza, sushi

3 HOMO FLORESIENSIS ~ NICKNAME: Hobbit LIVED: 95,000 to 13,000 years ago HABITAT: Flores, Indonesia (tropical) DIET: Omnivorous - meat included pygmy stegodon, giant rat

4 HOMO ERECTUS ~ NICKNAME: Erectus LIVED: 1.8 million years to 100,000 years ago HABITAT: Tropical to temperate - Africa, Asia, Europe DIET: Omnivorous - meat, tubers, fruits, nuts

5 PARANTHROPUS BOISEI ~ NICKNAME: Nutcracker man LIVED: 2.3 to 1.4 million years ago HABITAT: Tropical Africa DIET: Omnivorous - nuts, seeds, leaves, tubers, fruits, maybe some meat

6 HOMO HEIDELBERGENSIS ~ NICKNAME: Goliath LIVED: 700,000 to 300,000 years ago HABITAT: Temperate and tropical, Africa and Europe DIET: Omnivorous - meat, vegetables, tubers, nuts

7 HOMO NEANDERTHALENSIS ~ NICKNAME: Neanderthal LIVED: 250,000 to 30,000 years ago HABITAT: Europe and Western Asia DIET: Relied heavily on meat, such as bison, deer and musk ox

Just in case you want to learn even more about human evolution, we provide this Internet link to the Wikipedia web site that summarizes the topic.

Now that humans have evolved into creatures with some sophistication and inventiveness, have they the ability and power to "create" life? In 2010, a partial answer to this provocative question has emerged. A group of bioscientists, led by Drs. Craig Venter, Hamilton Smith, and colleagues announced that they have produced a new species of bacterium by building up a genome using genes severed from a known bacterium *Mycoplasma genitalium* and then adding other essential parts from other living matter. While it is not the same as creation *ex nihilo*, i.e., starting from scratch, it is an important step in the human endeavor to "create life".

After this overview of the nature of life and its evolutionary progress, we need to expand our thoughts on how life may actually have originated in space, beyond just the Earth. This page is continued as page 20-12a, reached through the Next button.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Writer's Note: The reader of this Tutorial may wonder, especially after finishing this page, what its subject matter - Evolution - has to do with Remote Sensing. The answer: probably nothing, at least not directly (it does tie in with the preceding pages on life, which stem from the coverage of Cosmology that does depend on remote sensing techniques). But, the topic is INTERESTING as such, and may be worthy of your consideration.

Evolution

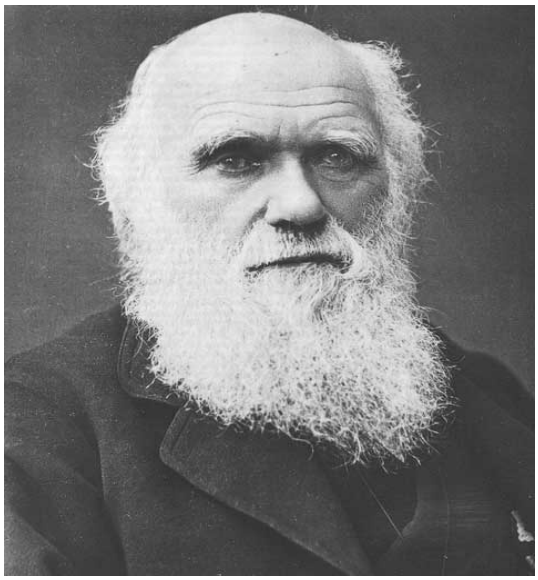
Our aim on this page is not to provide a comprehensive survey of all that is inherent and implied in the term "Evolution" but mainly to provide an indication of how it is evidenced on Earth in the records of the past - namely, in the rocks that make up geologic history. To do this, we will examine the evolutionary changes associated with one faunal type - animals that fly. When the writer embarked on preparing this page, this task seemed fairly straightforward. But as material was downloaded off the Internet, the complexity of the topic became increasingly apparent and the writer remains unsure of parts of this page. There is still much controversy over the details and the conclusions among the scientists studying the fossil records.

The vast majority of scientists, as well as a substantial fraction of the general public, have now accepted the basic principles of Darwinian evolution (and most subsequent refinements) as valid and verified (by the so-called "scientific method" of seeking truth). The position adopted by the writer (NMS) is to likewise accept these concepts as a given. No attempt is made to reconcile Evolution with its antithesis, Creationism, or to any of the models which are offshoots of the "Young Earth" hypothesis.

To get a quick overview of Evolution, we strongly suggest you read through this Wikipedia website that provides an overview of Evolution, as used in modern science. Our treatment of Evolution will be very limited on this page, to just a few ideas, so to gain a fuller understanding this Wikipedia website is strongly recommended. There are approximately 160 million citations on Google when "Evolution" is the search word. One of the best sources of information and links is supplied by the University of California-Berkeley website on Evolution.

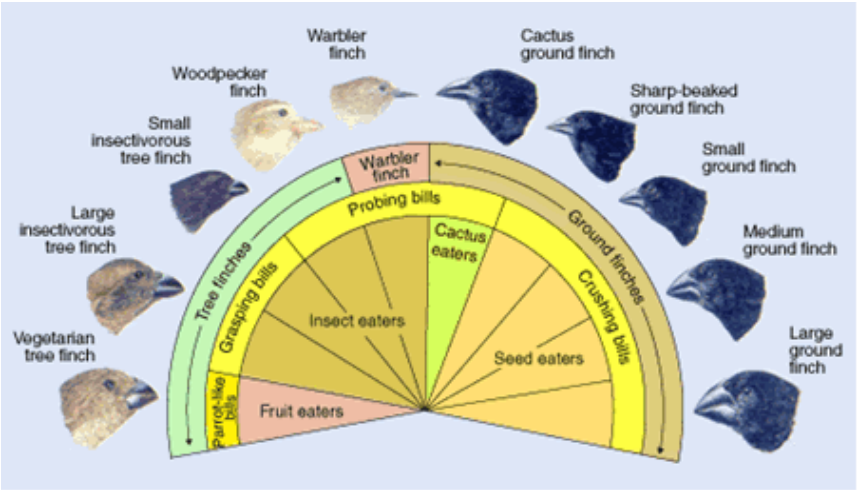
We will define "Evolution" by directly quoting the first sentence in the second Wikipedia entry: **In *biology*, evolution is *change* in the *genetic* material of a *population* of organisms from *one generation to the next*** (my words added: "in response to variations and shifts in environmental [adaptive] conditions"). The key words in this definition are italicized.

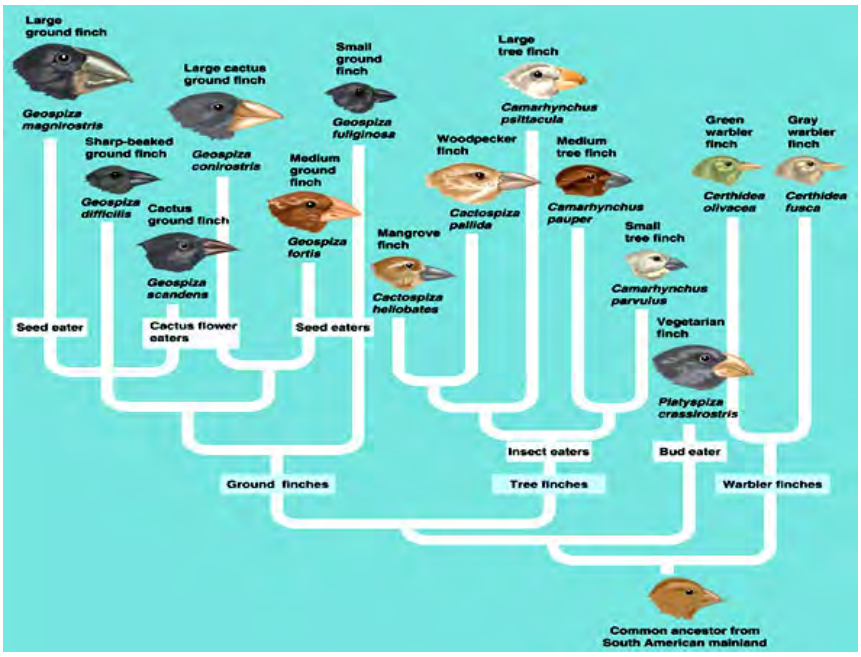
To the layman, the word "Evolution" is closely associated with one man, the Englishman Charles Darwin, who lived in the 19th Century. Nowadays, he is usually pictured as he appeared near the end of his life, complete with a scholarly white beard:



It behooves you to learn more about him by clicking on this Wikipedia website.

Darwin began to accrue the ideas underlying Evolution during what many believe to be the most famous scientific sea voyage in history, the sailing into the Pacific of the H.M.S. Beagle, a research vessel, from 1831 to 1836 with Darwin onboard as the ship's naturalist. In particular, the stopover at the Galapagos Islands west of Ecuador was to prove decisive in the eventual formulation of Evolution. Here he observed notable variations in the characteristics or traits in population among various animals from island to island. This led to the idea of adaptive radiation, best exemplified by interrelated finch (a bird) species. The most significant variable was the bird's beak, which was modified to optimize its use in the specific environment of a particular island. In time, Darwin realized that there were evolutionary lineages among these species. Here are two diagrams that speak to the changes involved:





Darwin began to devise his theory of Evolution during the voyage but, although he gave papers on the subject over the next 25 years, he did not publish the basic tenets of Natural Selection until 1859 (in part driven by the realization that a colleague, Alfred Russel Wallace, had reached similar conclusions). On November 24, 1859 (just over 150 years ago), the book that resulted, "The Origin of Species", became the most famous publication in the 19th Century. Its influence on mankind's - and especially scientists - thinking and philosophy cannot be underestimated. The Harvard biologist Ernst Mayr, possibly the greatest evolutionary theorist since Darwin, has stated: "It would be difficult to refute the claim that the Darwinian revolution was the greatest of all intellectual revolutions in the history of mankind." The Harvard paleontologist and historian of science Stephen Jay Gould called the theory of evolution one of the half dozen most important ideas in the entire history of Western thought. Below is a first edition copy (in green cover) of the book itself (this one sold for $110,000 to a collector) and a later edition needed because the book actually sold well in the 19th century.

The key concept espoused by Darwin is summed up in the phrase "survival of the fittest". Behind this deceptively simple phrase is this fundamental concept: A group of animals (or plants) represented by a species will have normal, natural variabilities in their physical properties, e.g, such diverse entities as hair color, lung capacity, ability to withstand cold. Some of these

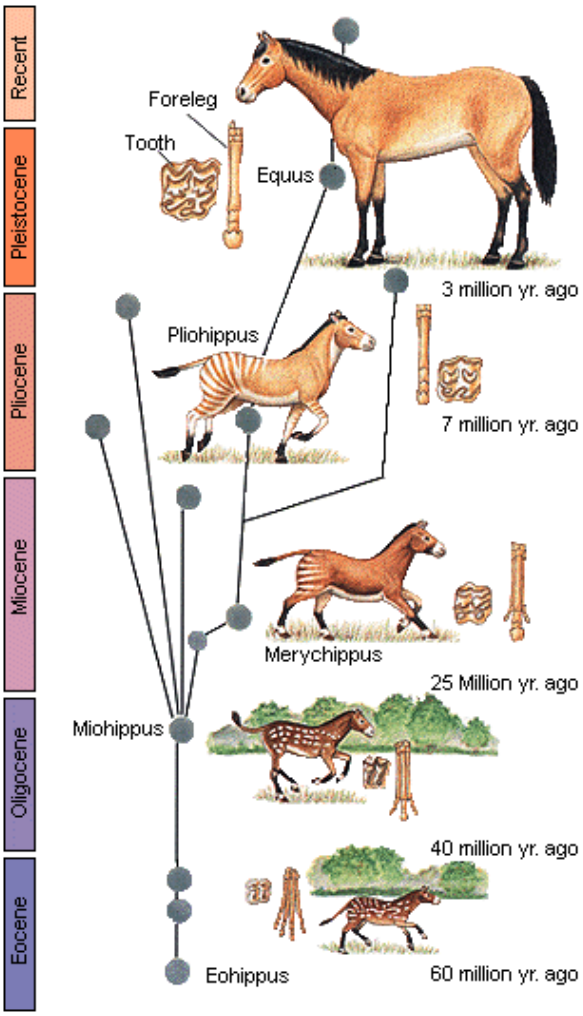
properties can be essential to each individual insofar as surviving any stress in the environment. As an example, choose the metabolic response to excessive heat. Some of the animals are less able than others to tolerate such heat. Many of these die during a hot spell. In so doing, they may have been unable or lacked the opportunity to mate and thus prolong their strain with offspring. They are "less fit" by virtue of this property (trait) weakness. Therefore, statistically they are less efficient in propagating the species, and are more likely to be eradicated, with this unfavorable property not being passed on to potential progeny. The survivors are those with better adapted critical properties (other situations can stress other properties as well, such as the ability to process available food, which some can do better than others). The overall effect is to improve the characteristics (properties) for those that do survive when the environment changes.

Natural Selection isn't the whole story. Biologists now know that chance mutuations of genes in chromosomes play a role too. This must occur within a population that is undergoing change over time. The effects are gradual, for many changing species a matter of thousands and even millions of years. Although the mutation of one species can be rapid (in terms of geologic time), as has been proved by scientific studies in the laboratory, it *seems* in many instances to be comparatively slow. Thus, when the history of a group of related animals is followed in the geologic record, new species often give the impression of appearing "suddenly", and many species disappear abruptly. This is in part illusory, owing to the fact that the record is commonly discontinuous, and conditions for the preservation of a species as a fossil are frequently not favorable. This gives rise to the "missing link" phenomenon - the gradual transition involved as one species evolves into another(s) is simply not preserved.

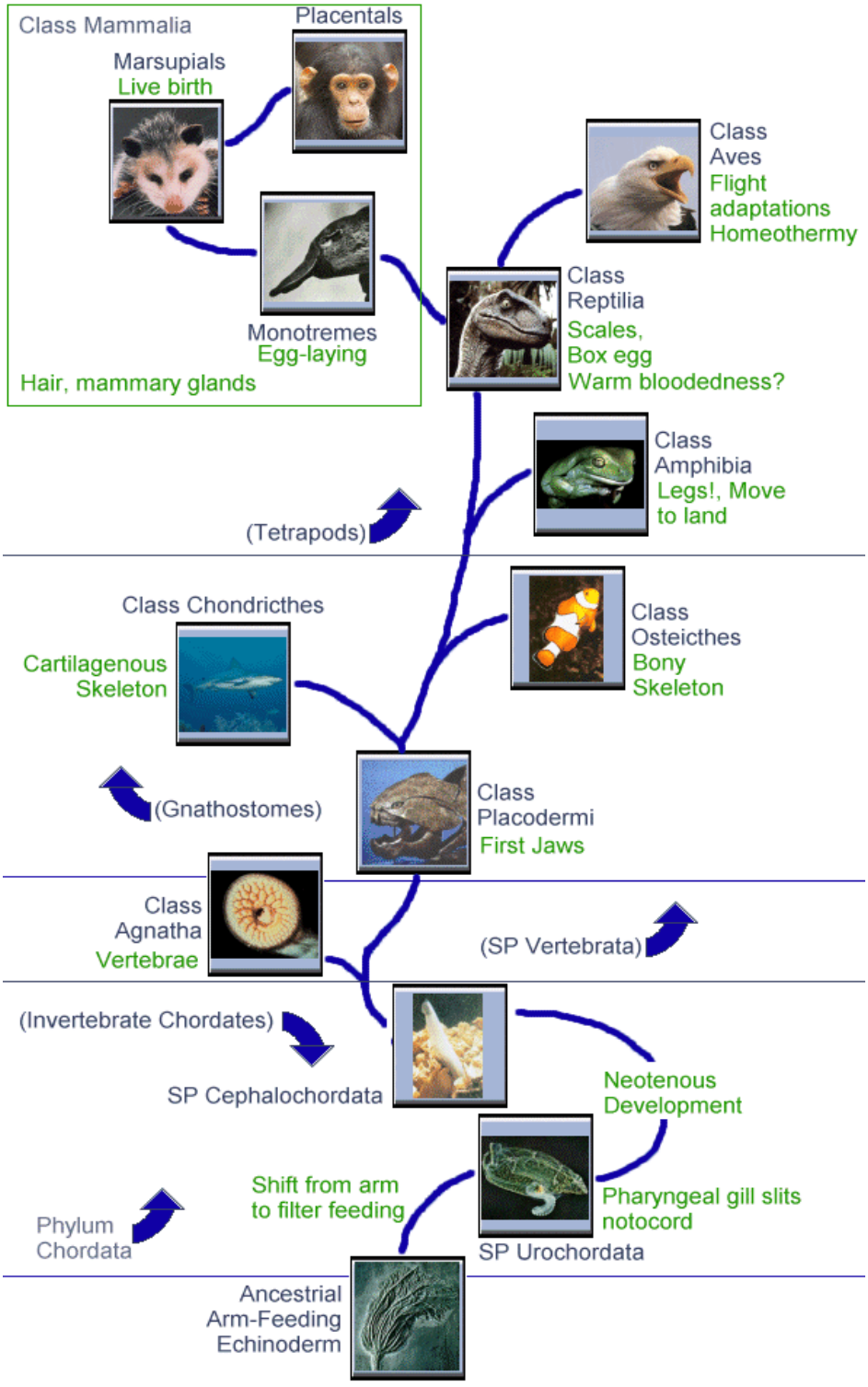
Thus, in Darwin's day, the evidence for evolution was largely macroscopic - that is, from observations of genera and species that occur in different strata of the sedimentary geologic record. But the change from one species into a new one is inferential rather than directly evident from a continuum. This raises the question: at what point in time is a new species said to first come into existence? How (and by what criteria) can this new species be fixed as really new, i.e., when in the transitional stage does one specify that a new species has emerged? (Stephen Gould's Punctuated Equilibrium model suggests that the changes can occur over relatively short time periods in which conditions, such as in local or regional environments, can undergo large changes that affect the resident populations forcing significant adaptations. Evolution proceeds therefore in "spurts".)

Darwinian evolution has been aided to some extent by the ideas put forth by the monk Gregory Mendel's hypothesis on heredity, as reviewed in this Wikipedia website. Mendel's ideas were based largely on observations. But in the 20th, and the present, Century genetics has placed the ideas of inheritance, and more broadly, those forming the foundation of Evolution on a firm molecular basis. Simply summarized, the variations in the genetic makeup of organisms, leading to the recognition of DNA and RNA as the prime factors in life functionality and reproduction (see the previous two pages), which have been determined from microbiology, comprise the essence of the mechanisms for Evolution as the process that has determined the diversity and comlexity of life on Earth.

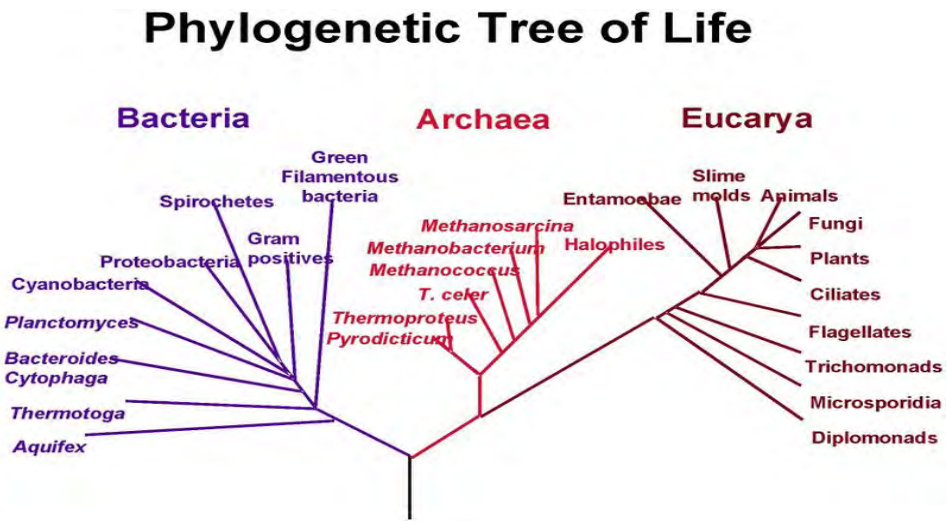
The "proof" of Evolution was, prior to the DNA studies, usually cited as found in the fossil records. One of the best lines of evidence has been the pattern of changes observed from the early Cenozoic to the Present for the horse, a mammal whose bones have been preserved in sedimentary continental rock deposits. The record is abundant. This diagram summarizes the main representatives of horse evolution from the Eocene to modern times:



The horse, an animal with a vertebrate, is part of the phylum Chordata. Chordates can be traced back through the Ordovician, but their origin from invertebrates remains uncertain. One hypothesis treats them as evolving from the Echinoderms, but this "leap back" is controversial. Here is one family tree for the Chordata, which begins with the Echinoderms:

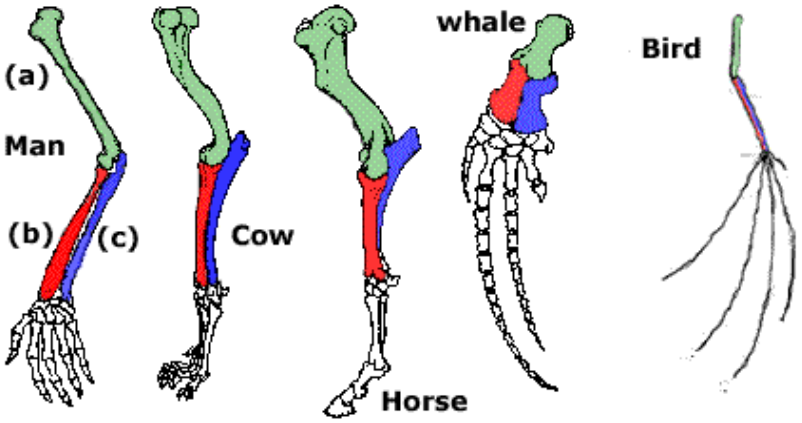


Once the general outlines of Evolution had become mainstream in scientific thought, various schemes relating the different Kingdoms and Phylla were developed (a large family tree type diagram was shown midway down page 20-12. Attempts have been made to simplify the classification by producing a broad phylogeny. Here is one version, similar to one that appeared earlier on page 20-12, which places life into three broad groups (ignore the individual genus names, given as examples), but is moot about a common ancestry:

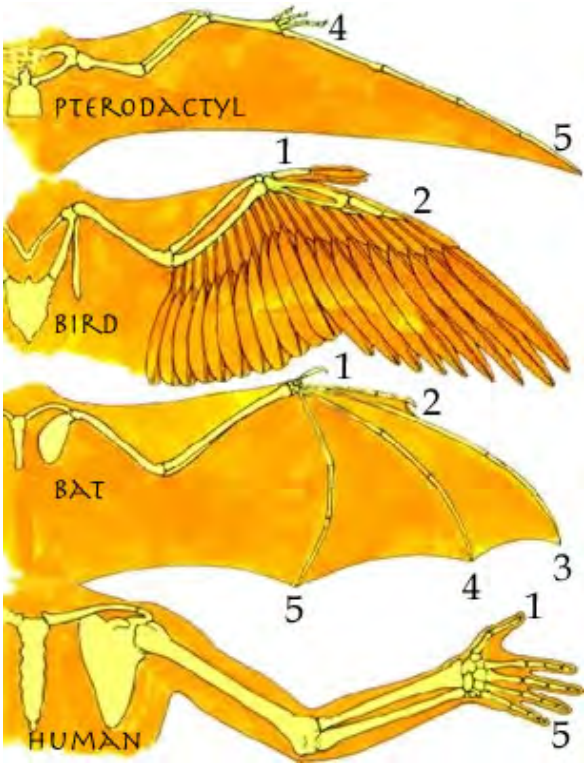


The above paragraphs just skim through a few of the intrinsic concepts that underlie Evolution. Once again, we urge you to expand your knowledge, if need be, of this great discovery by Darwin (and Wallace) by reading through the Wikipedia entry on this subject. For the remainder of the page, we choose to dwell upon a very illuminating example of Evolution in practice by looking at the (evolutionary) history of animals that have developed the ability to fly.

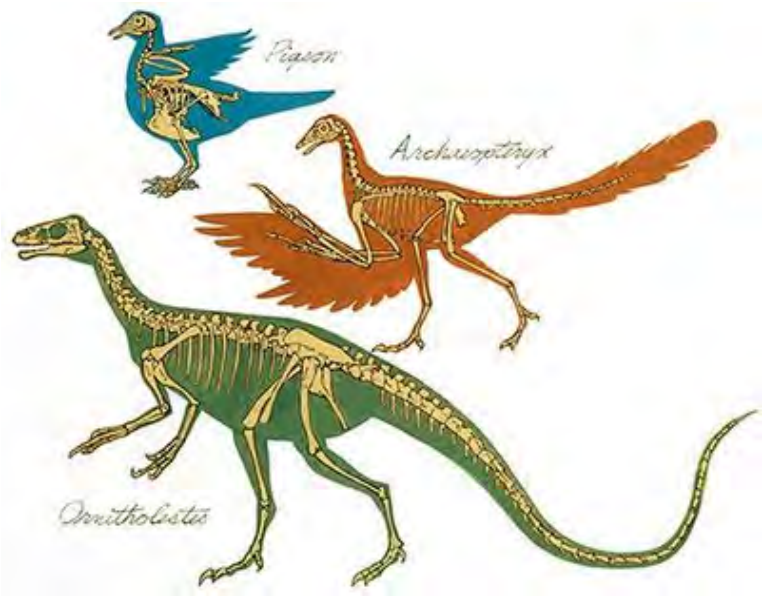
Flight is one of the properties ascribed to many different "things". Birds, bats, tree seeds, and airplanes are all said to "fly". But they are quite different entities. This ability to perform the same function in grossly different entities is described by the term "analogous". But when the property responsible for some form of locomotion has a common origin or ancestry, the term "homologous" applies. The skeletal limbs of five common animals shown below are all homologous. In the diagram, three bones are shared in common: a)Humerus; b)Radius; c) Ulna.



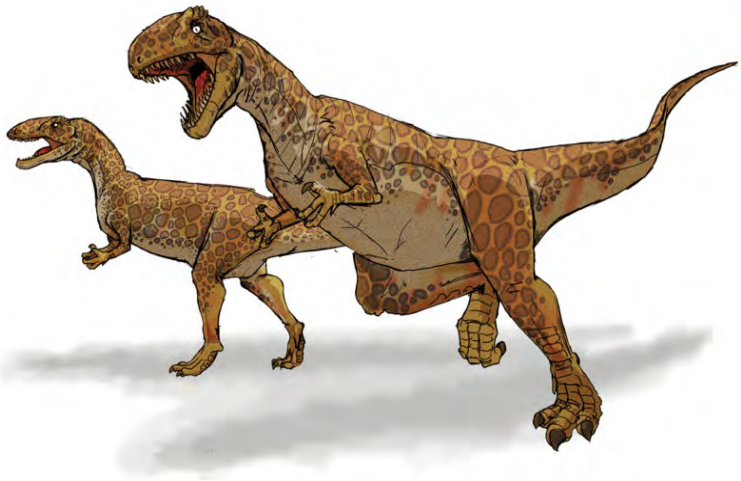
Another example of a homologous appendage on dissimilar animals is the "wing". In this diagram, the commonality of bone structure in a reptile (pterodactyl), a bird, and two mammals - a bat and a man (people fly with hang gliders, and then there is Icarus in Greek mythology) - is evident.



The gross structures of a modern bird (here, a chicken), and ancient bird (Archaeopterix) and an ornithiscian dinosaur show broad similarities (are homologous):



The dinosaurs are a famed and diverse group of animals that include the progenitors of the birds and a related group of flying reptiles, the Pterosaurs, that we will consider as evolutionary representatives of flying creatures. Before taking up flight as such, it will be instructive to examine the dinosaurs and related animals that dominated the Mesozoic Era (250 to 65 million years ago). The first time the word "dinosaur" (from Greek words that mean "terrible" "lizard") was used dates to 1842, as introduced by Sir Richard Owen of England. At that time he published his studies of bones found by three separate groups. The first was Sir William Buckland who had described the bones of Megalosaurus (named later) found in an English quarry between 1815 and 1824 (part of a Megalosaurus femur was found in 1676 but was not recognized then as a dinosaur fragment) after being advised by the visiting French scientist Georges Cuvier that it was a giant "something". Here is the "type fossil" of Megalosaurus and a reconstruction of the animal's appearance:

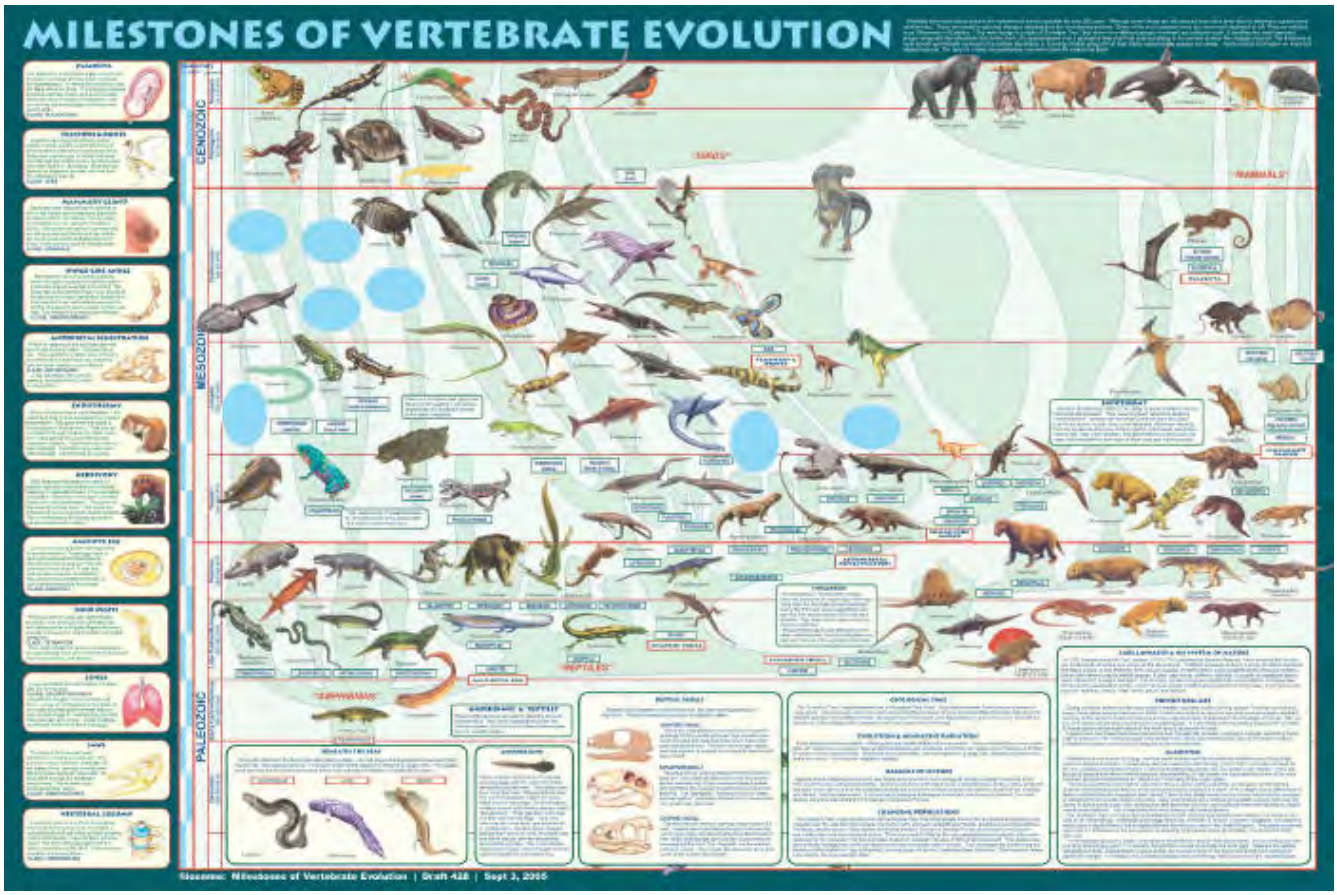


Other bones accessible to Owen had been discovered by Gideon and Mary Ann Mantell near Lewes, England. These were eventually named Iguanadon. Here is one of their finds and a reconstruction:

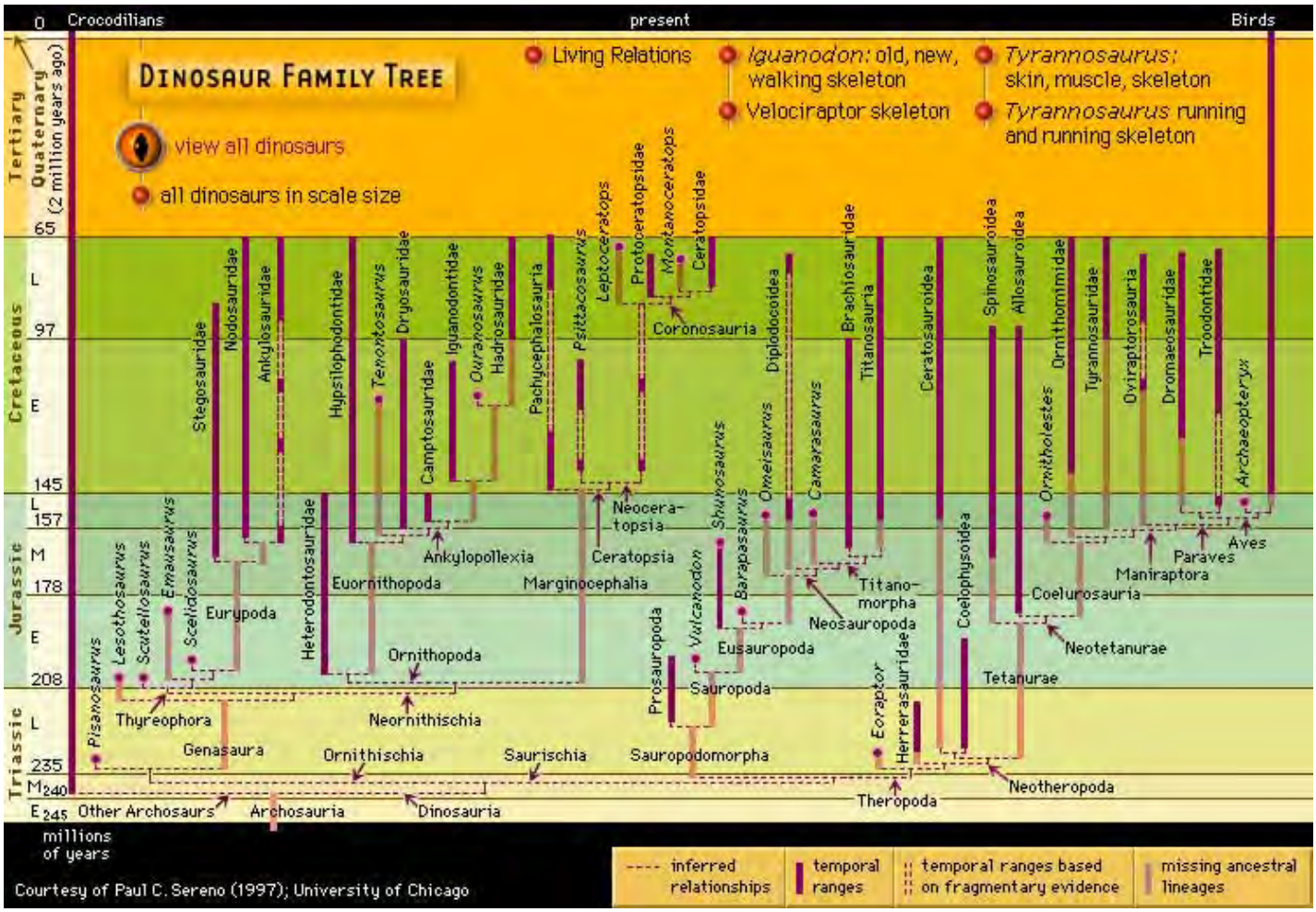
Interest in dinosaurs and other similar life forms increased rapidly in the 19th Century. That interest has burgeoned ever since (witness the popularity of "Jurassic Park"). The Class Reptilia

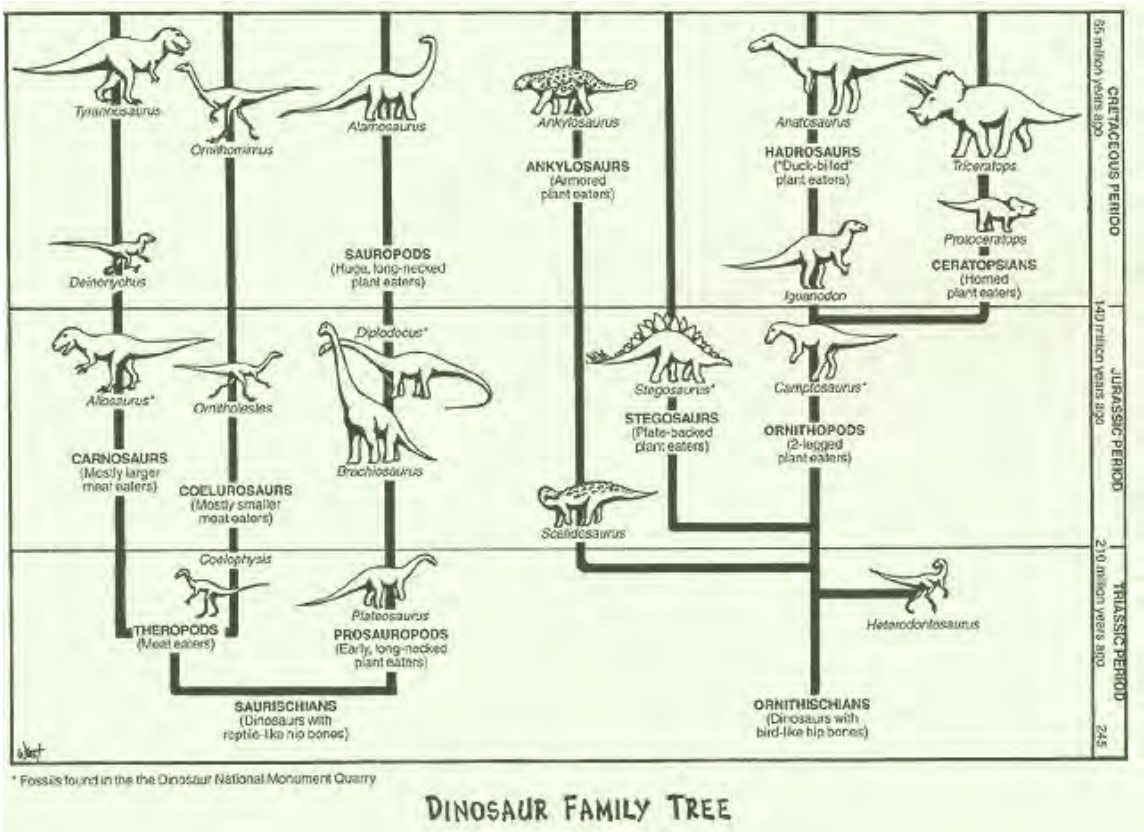
has proven to be quite diversified and is intimately tied to "flying animals". The reptiles are part of the Phylum Chordata which traces back to the Ordovician. Reptiles emerged from the Amphibians which first appeared in the Devonian. The dinosaurs dominated the land in the Mesozoic. As mentioned in Section 18, the dinosaurs disappeared abruptly at the close of the Cretaceous; their vanishing was probably due to a combination of events - climate change caused by volcanism in India and elsewhere, evolutionary "old age" as they grew too large to adapt, and the 'coup de grace', a major impact in Mexico and perhaps at other locations.

Although the script is not readable in this Internet version, this chart hints at the diversity of the Chordata (the Vertebrates).

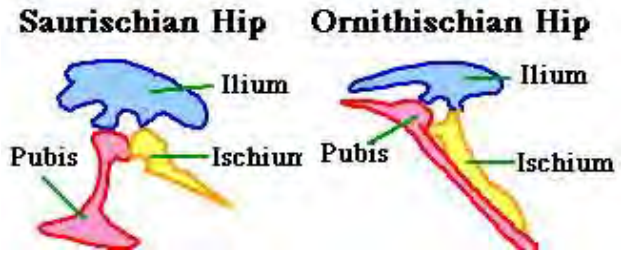


These two diagrams concentrate on the classification of dinosaurs.

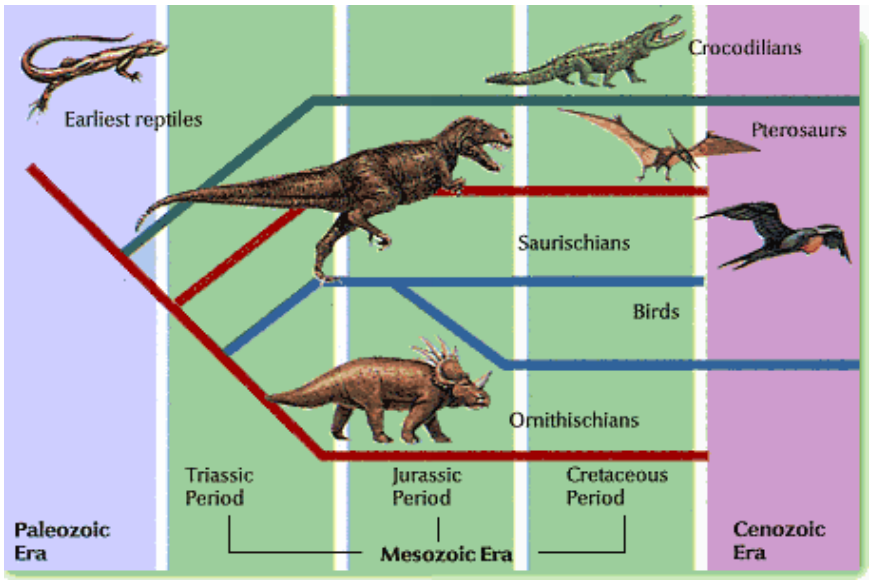




The dinosaurs are split into two main orders: the Ornithiscians and the Saurischians. Oddly, the Ornithiscians do NOT include the birds (studied by ornithologists); the flying members split from this group are the Pterosaurs. The birds have been evolved from small dinosaurs in the Saurischian order. The two orders are separated largely on the basis of hip structures:

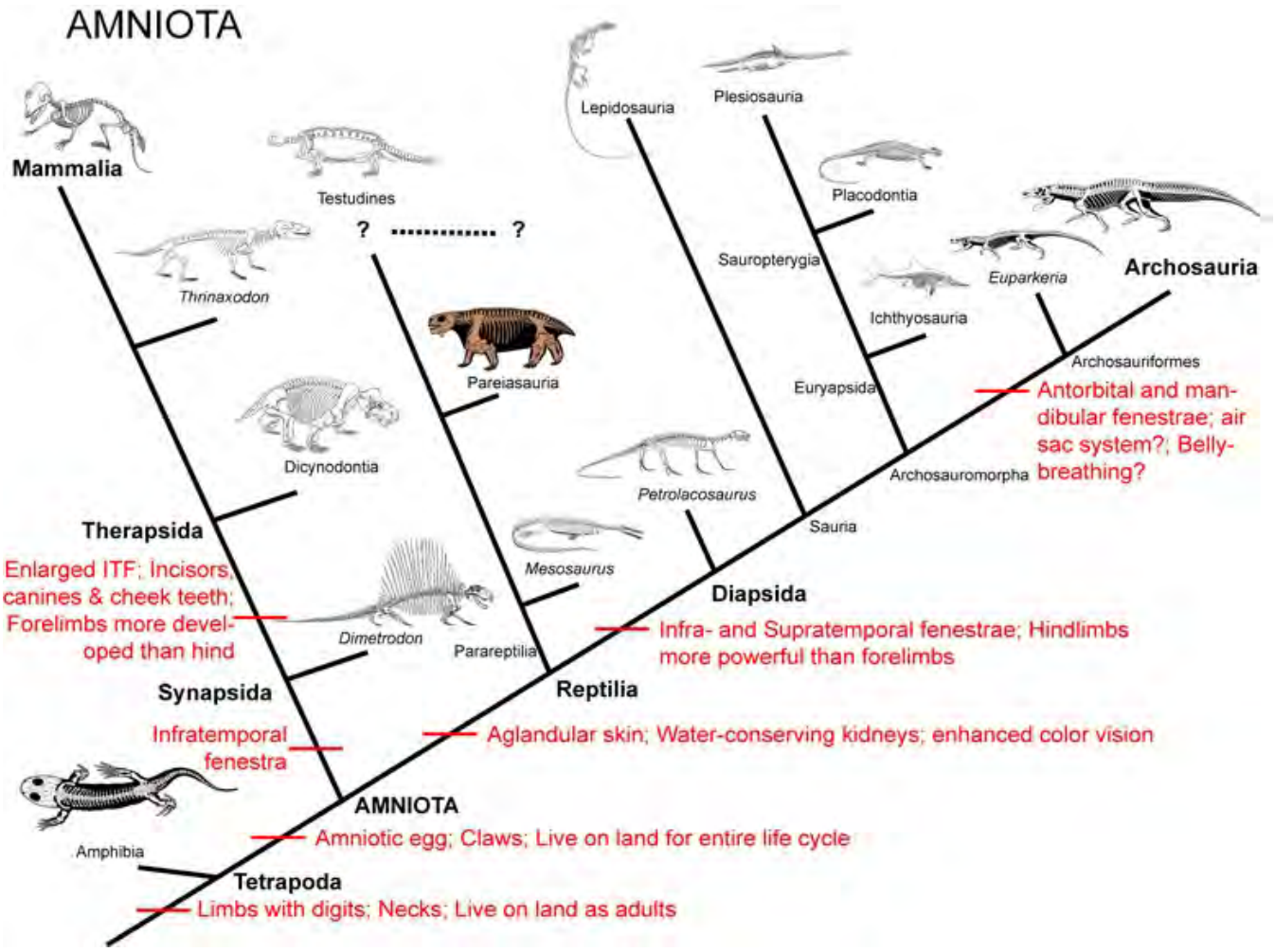


This generalized diagram helps to distinguish the main groupings of Mesozoic reptiles:

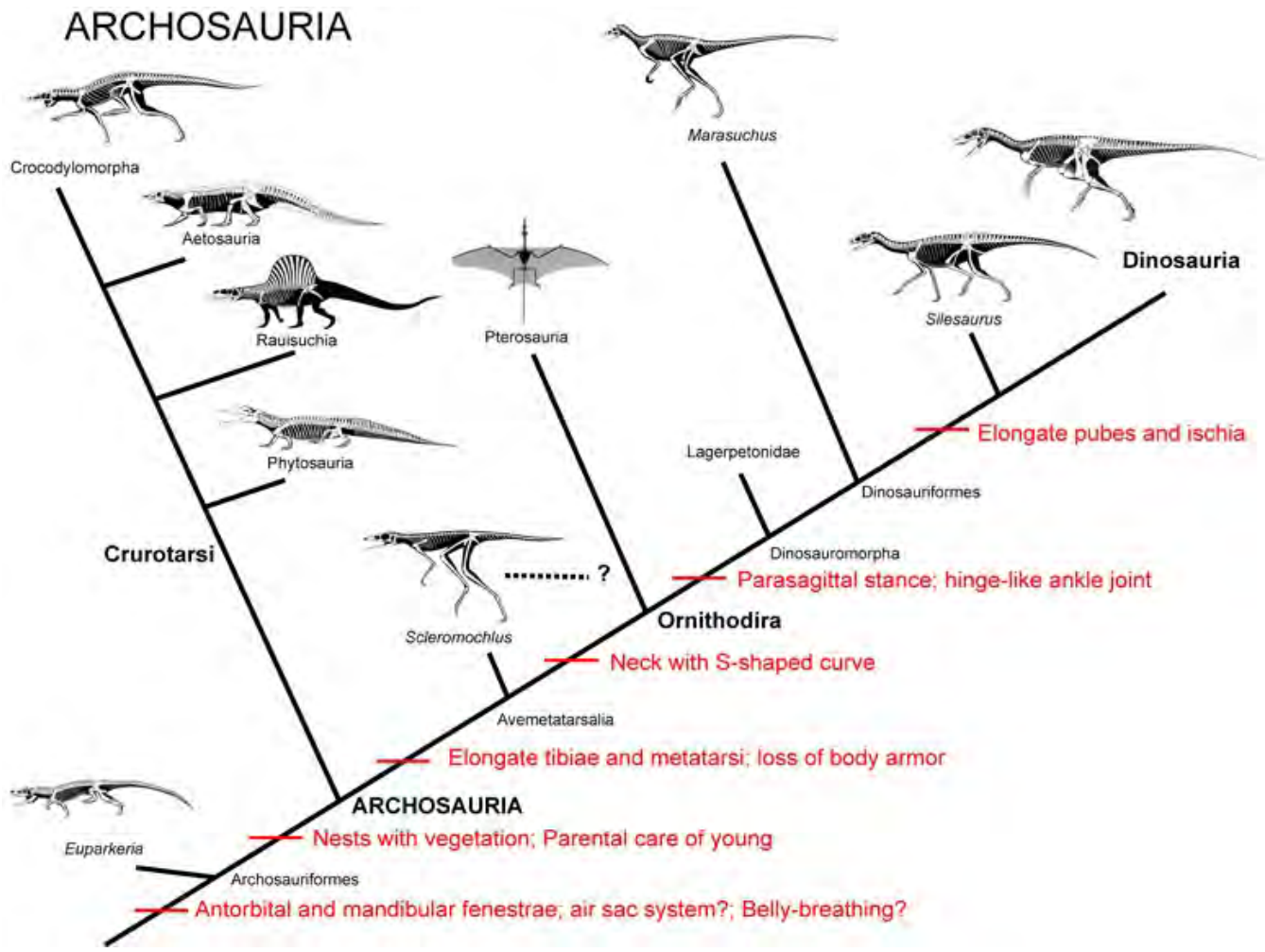


The linear branches in this diagram are an example of a cladogram. The field of cladistics refers to the "genealogy" of various interrelated organisms, showing the sequence of ancestors and descendants. Cladograms have been prepared by Dr. Thomas Holtz of the Geology Dept. of the University of Maryland for use in his courses on the dinosaurs and other evolutionary animals. Visit his website at the University of Maryland. We have downloaded six of his cladograms that describe, in sequence, the details of the evolution of animals that include pterosaurs and birds, to be discussed later on this page. They are detailed; just try to get the overall picture:

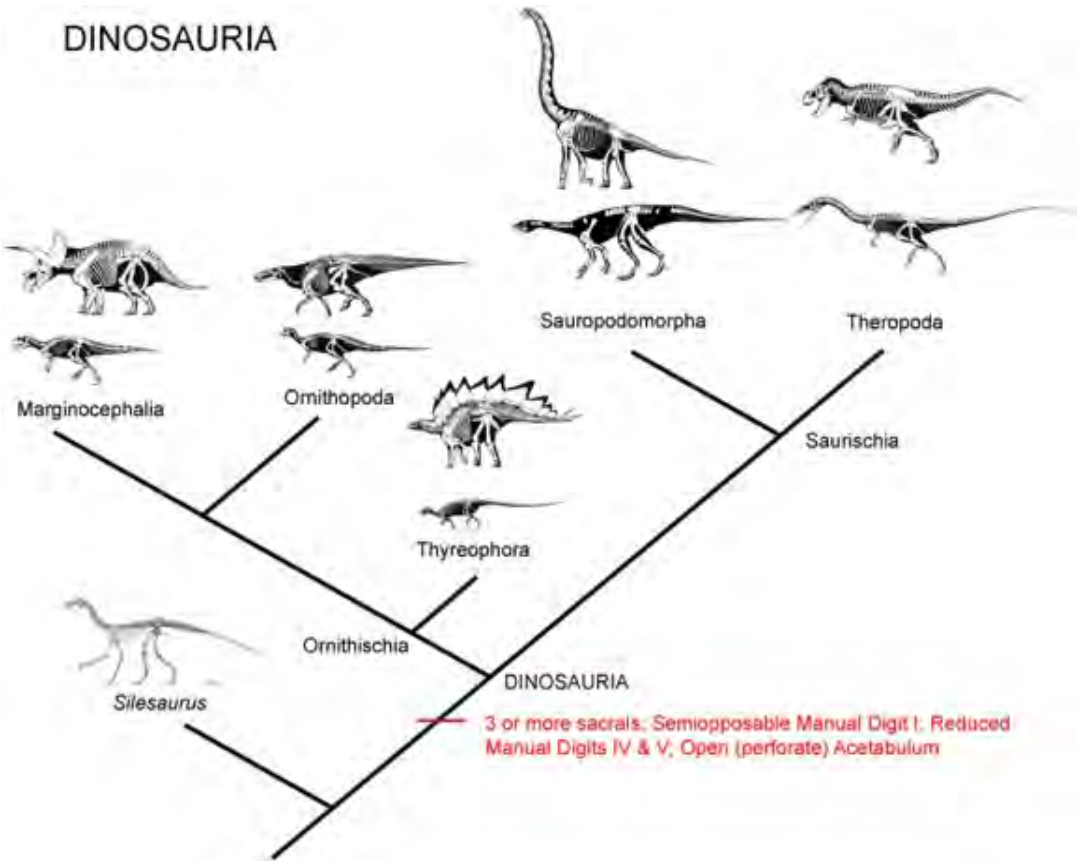
The Amniota refer to a wide diversity of animals that all developed to live on the land (but also in the sea and the air). They include the mammals, the birds, and the reptiles. In common, they utilize the amniotic egg as a means of sexual reproduction.

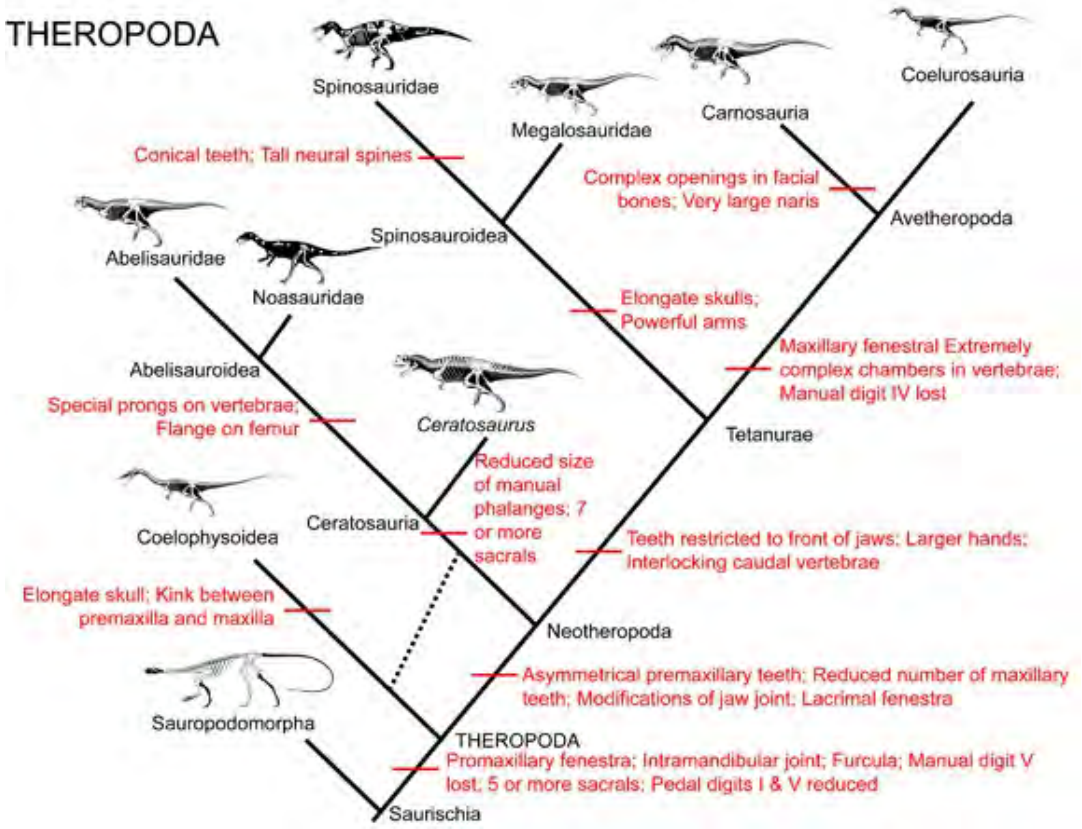


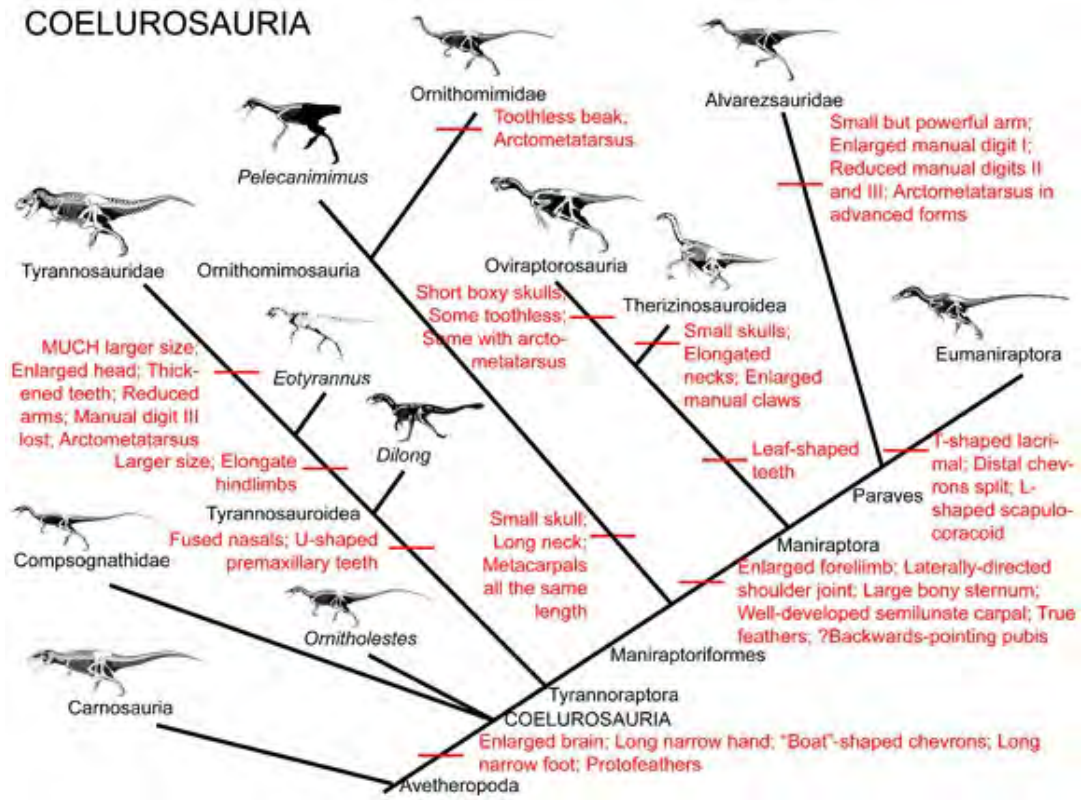
The Archiosaurs are the late Permian to early Triassic progenitors to the dinosaurs, the pterosaurs, the crocodiles, and the birds.



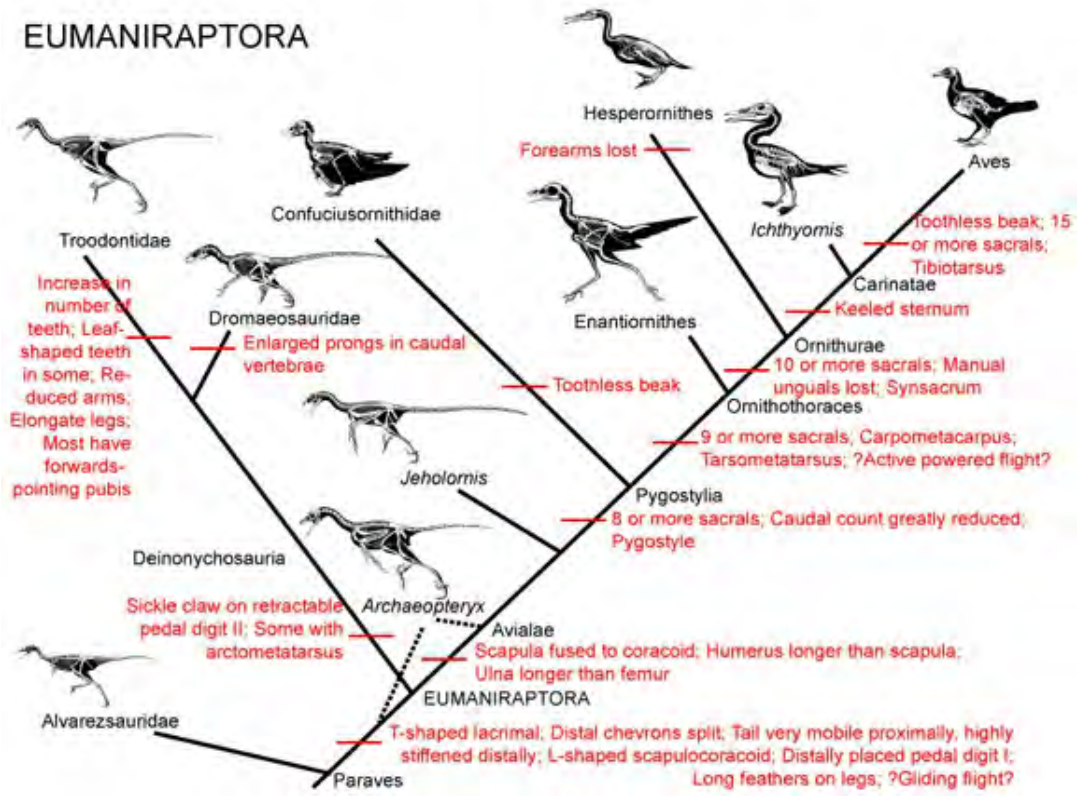
The next four cladograms proceed through the general dinosaurs (note how the Ornithischia and Saurischia are split), the Theropods, and then the Coelurosaurids.







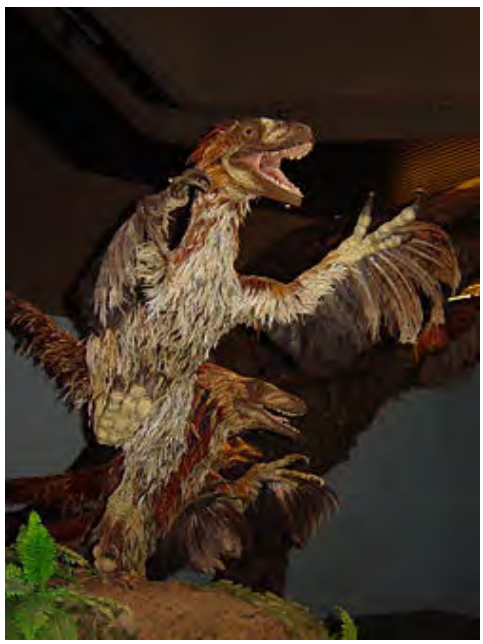
Finally, we reach the cladogram that includes the birds as well as the Velociraptors.:



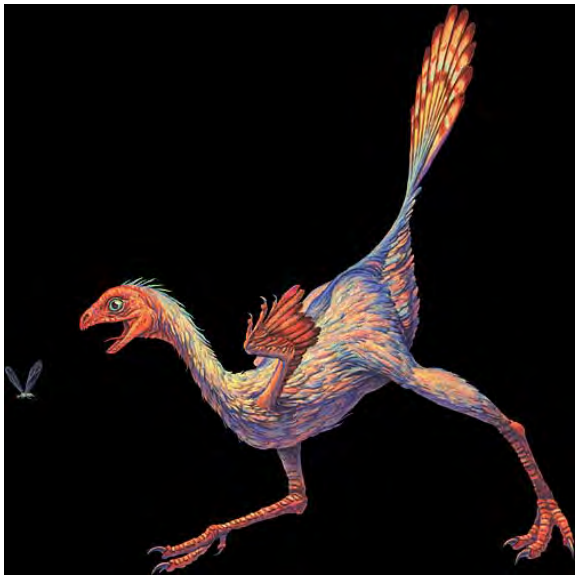
Note that the Pterosaurs are include early on in the Archosauria cladogram but the birds (Aves) do not appear until the final Eumaniraptora cladogram

There are thus two broad, and rather distantly related, groupings of winged reptiles. They possess winglike membranes in common but do not possess feathers in common. The Pterosaurs were probably without feathers.

Some dinosaurs are now known to have developed feathers but were not flight capable; Deinonychus, a velociraptor that lived about 115 million years ago, is an example:



Dinosaur feathers may be more common than once suspected. The first (oldest) dinosaur whose fossil remains show faint feathers is Sinosauropteryx. A later genus is Caudipteryx:



Another may be Anchiornus, who flight status is still conjectural.

By now you may have perceived a major distinction between the two groups of reptilian descendants that lived in the Mesozoic. The Pterosaurs are quite different from the Aves or primordial birds. Pterosaurs first evolved in the Triassic; birds first appeared in the Jurassic. We shall hereafter treat the evolutionary history of these two groups separately.

The Pterosaurs are an Order in the subclass Archosauria, part of the Class Reptilia. They have been further subdivided into the suborders Rhamphorhynoidea and Pterodactyloidea. The "Rhamphs" appeared first, in the late Triassic and died out at the end of the Jurassic. They are

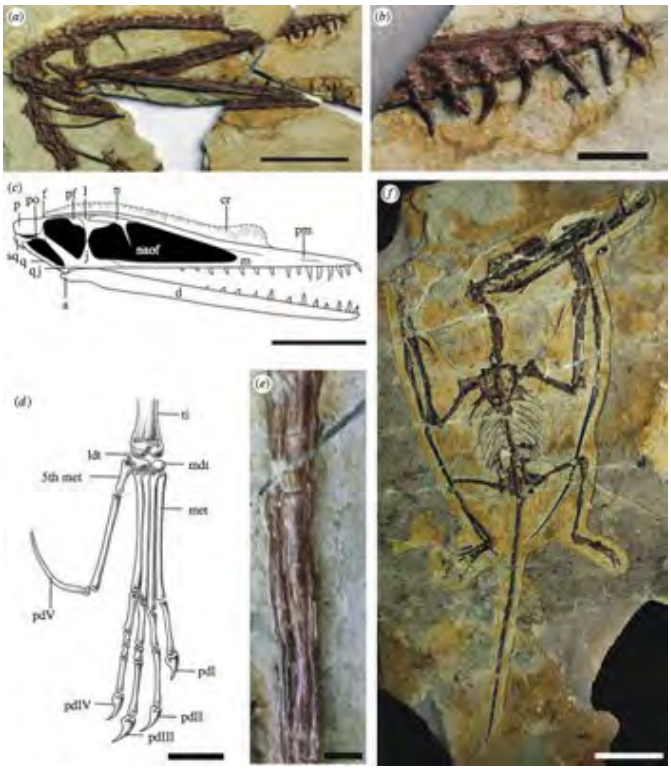
generally small and have long tails. The "Dacts" emerged in the Jurassic and died out along with the dinosaurs (the Pterosaurs are NOT themselves dinosaurs). Their size can range from small to very large; they have short tails. All Pterosaurs have membraneous wings but do not seem to have had feathers. They apparently glided rather than flapped and took off from perchs.

One of the first Rhamphs is Dimorphodon, which lived 213-203 m.y. ago.

Campylognathoides spanned an age from 200 to 186 m.y. ago; its wing span was less than 2 meters.

Rhamphorhynchus is the type example of the "Rhamphs". It had an average wing span of 1.75 m.

A recent entry into the evolution of the Pterosaurs is Darwinopteris. Twenty individual fossil remains have been found in China. Darwinopteris is considered to be transitional (it lived 160 million years ago) between the "Rhamphs" and the "Dacts". It has the tail of the former but the head of the latter.



The best known of the "Dacts" is Pterodactylus. Pterodactylus kochi is cited as the prototype, having been discovered in the famed fossil locality of the Solenhofen quarry of Bavaria in Germany. It is contemporaneous with the most famous of early birds, Archaeopteryx; both are 150 million years old.

Some "Dact" skeletons seem to have two sets of limbs, one functioning as legs, the other as claws. Dsungaripterus weii is sometimes depicted as having two sets of legs, to which membranes are attached to one. Consider these two reconstructions.

The next four artists' reconstructions show various Pterodactyls, given by their genus names. In the caption of each is the wing span and the age during which the pterosaur lived.

Genus Tapejara:

Genus Ornithocheirus:

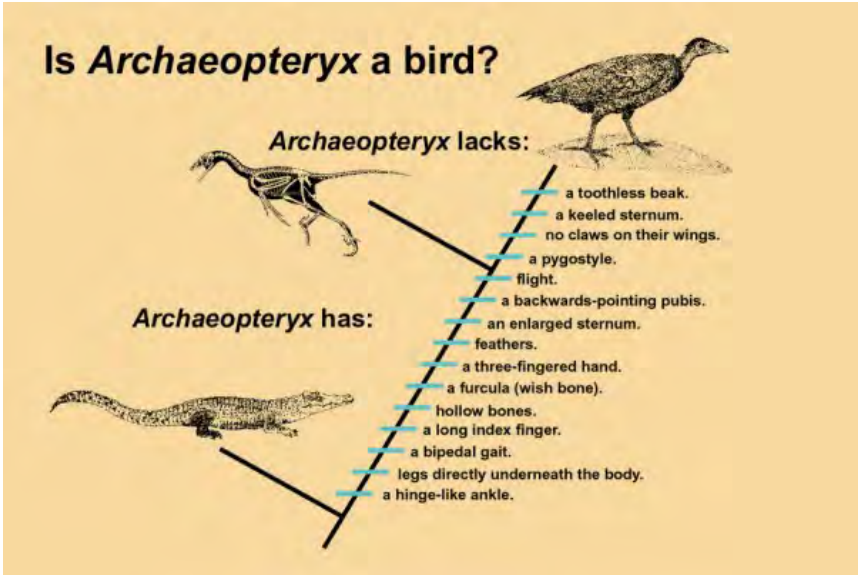
Genus Pteranodon:

Genus Quetzalcoatus;

Starting in the Jurassic, as we know for sure (but probably earlier), the first true birds appeared. These had feathers, well-developed wings beyond the membrane type, and gradually developed the ability to do more than just glide from tree to tree (attaining sustained free flight). The truly **true** first bird as such has not been identified as yet with any certainty. Various candidates have been proposed. One is Microraptor of the family Dromaeosauridae. Here is its skeleton and (a rather fanciful) reconstruction of it in flight.

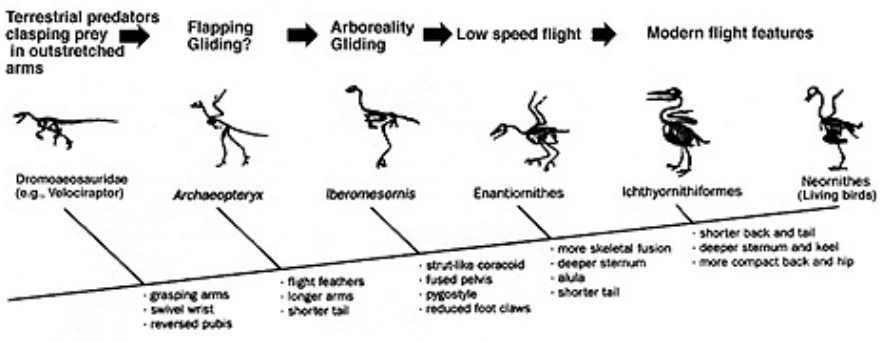
Another candidate is Sinosauropteryx, shown earlier on this page. But, for the time being, paleontologists generally agree that the oldest bird whose identity as such is known for certain is the famous Archaeopterix, the first skeleton of which was found in the Solenhofen quarry in the mid 1800s. Its age is 150 million years. Here is the type skeleton and a reconstruction:

Here are some pros and cons concerning its nature as a true bird.

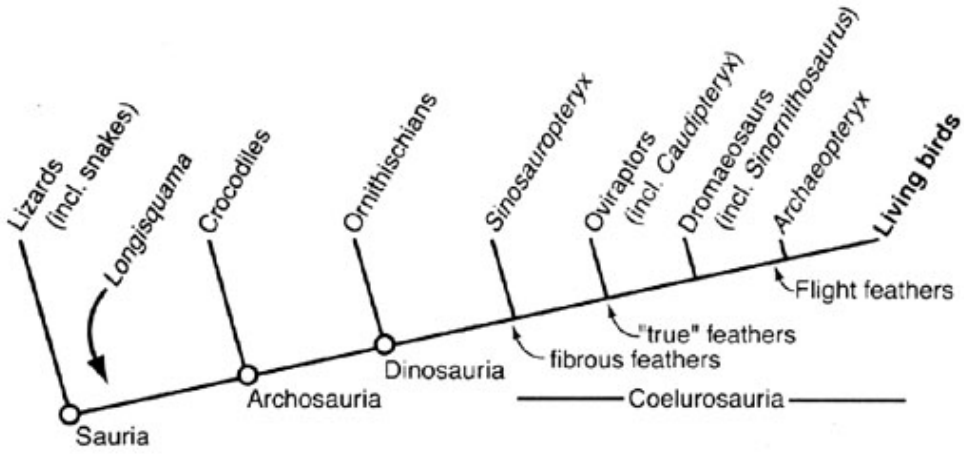


Since its discovery a number of younger Mesozoic bird remains have been found.

For the birds, the development of free flying, as is customary today, was gradual. This is one proposed scheme:



Various cladograms describing avian evolution in the Mesozoic into the Cenozoic have been put forth. Here is one example:

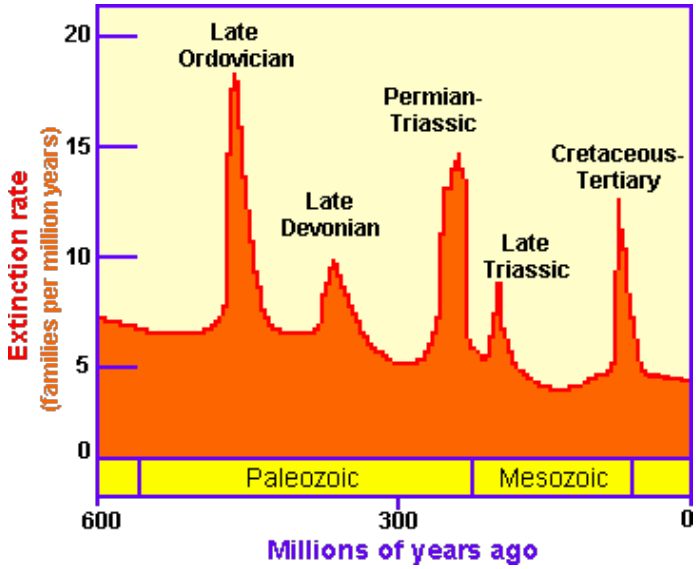


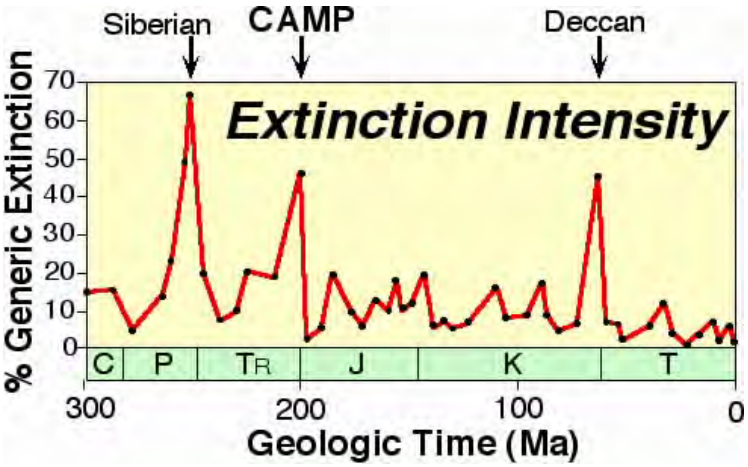
Two of the best known Mesozoic birds are Confuciusornis and Hesperonithiformes:

Birds proliferated and diversified throughout the Cenozoic, until today more than 60 billion individuals in about 15000 species exist. The Scissor-tailed Flycatcher, a common sight in Oklahoma, is a modern example that shows some similarities with the early birds.

For more insights into the origin and evolution of birds, check these Internet sites:Origin of birds and bird evolution

The birds are considered by many evolutionists to be the modern survivors of the dinosaur lineage. The dinosaurs themselves are, of course, now extinct. Extinction is one of the hallmarks of the evolutionary record. Many species, families, orders, etc. disappear from the geologic record. Several times in the distant past there have been mass extinctions that kill off significant percentages of all life forms (plants and animals, including those that inhabit the oceans). The next two diagrams show this evolutionary history.

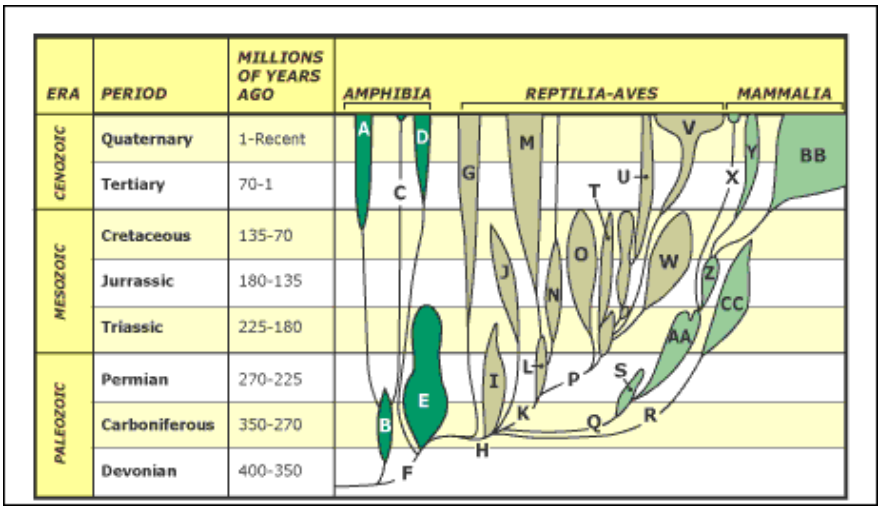




Before we close this page, we will just mention the Class Mammalia, which includes humans. The mammals first evolved at the end of the Paleozoic, thrived in the background in the Mesozoic, and emerged as the dominant animal life forms after the extinction of the dinosaurs. Their evolutionary history is too vast and complicated to warrant even cursory discussion on this page. A good overview is afforded by this Wikipedia" web site. You can gain an insight in the broadest terms into the cladistic history of the Mammals by looking at the diagram in that site listed under

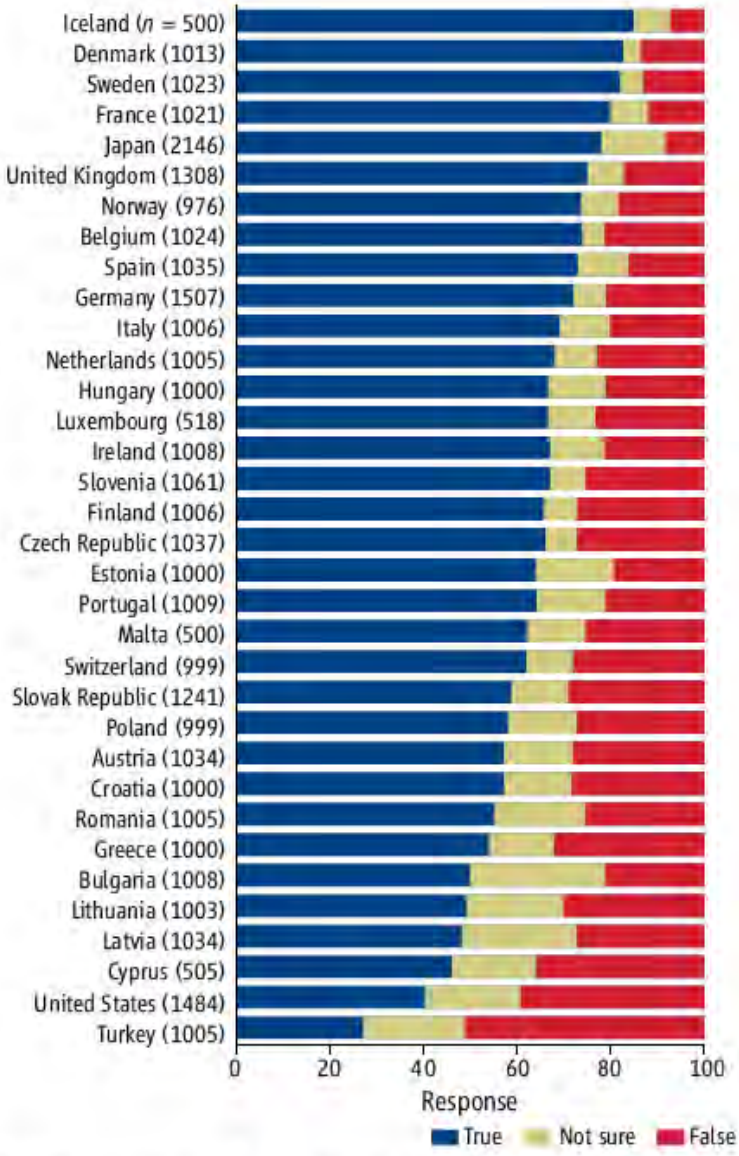
the heading of "The Ancestry of Mammals". Suffice to say here that progenitors of the Mammals extend back into the Late Carboniferous and true Mammals first appeared in the Triassic.

We will end the page with a summary diagram taken from the Internet. It is an attempt to display the phylogeny of the vertebrates, starting with amphibians, but without showing details of any cross-links between groups:



Phylogenetic origins of various groups of vertebrates. A) Urodela, B) Lepospondylii, C) Apoda, D) Anura, E) Labyrinthodontia, F) Apisidospondylii, G) Chelonia, H) Anapsida, I) Cotylosauria, J) Eurapsida, K) Diapsida, L) Eosuchia, M) Squamata, N) Rhyncocephalia, O) Ornithischia, P)Thecodontia, Q) Synapsida, R) Parapsida, S) Pelycosauria, T) Pterosauria, U) Crocodilia, V) Aves, W) Saurischia, X) Prototheria, Y) Metatheria, Z) Pantotheria, AA) Therapsida, BB) Eutheria, CC) Ichthyosauria.

So, after perusing this page, where do YOU stand on this most controversial of science subjects? If you are an American, the chances are good that you reject the basic tenets of Evolution. The acceptance/rejection ratio varies with country; for example, most Europeans accept Evolution. Check this oft-cited survey, shown in bar-graphical format:



Public acceptance of evolution in 34 countries, 2005.

Evolution remains one of the most challenged and debated ideas ever put forth by intellectuals. Four groups of people in the United States (and to a similar extent everywhere else) have emerged in the 21st century with decided views about evolution; At one extreme 1) the Creationists; at the extreme opposite 2) the Scientists; a middle ground 3) the Intelligent Designers, and 4) a large segment of the population that just don't know enough to have formed a defensible position, and probably don't care to know.

There are those who favor much of the knowledge embodied in the meaning behind "Evolution" but who admit that the concepts involved have some problems and objections. One is, simply put, how life can form from inorganic chemicals. Another is the ultimate ancestor idea: do all the branches of life indeed work back to a few simple forms? A third dispute is the "missing link" argument: While many animals, and to a lesser extent, plants, seem to have some continuity in the fossil record but important transitions are often lacking. A fourth objection is how to explain the "sudden" (geologically short time) appearance of advanced forms in the Cambrian explosion of life. There are others, as you are encouraged to seek out on your own through the Internet.

But Evolution in its general form seems to have become accepted - with modifications, clarifications, and addition - by the scientific community. No doubt, both in the foreseeable future, and beyond centuries hence, its concepts and tenets will be refined. But to overthrow it would lead to the collapse of the scientific systems that themselves have evolved from Evolution.

Primary Author: Nicholas M. Short, Sr.

Life in the Universe: I. Background in Biology

Living matter (organics) has cropped up as a subject in several Sections of this Tutorial. Some readers may desire to build up a basic background in Biology. Using the same mini-tutorial approach as was implemented elsewhere in this Tutorial for Geology and Meteorology, and in this Astronomy/Cosmology Section, the writer (NMS) has produced this and the next page. I have relied on one of my most valuable books, **Biology**, by Peter Raven and George Johnson (McGraw-Hill). In addition, I have drawn heavily on material located on the Internet. Internet tutorial on this subject found by the writer seems to be An Online Biology Book by M.J. Farabee. And here is another Internet site that specifically focuses on paleolife topics: Paleobiology (this site is actually a part of the abovementioned tutorial on Biology, prepared by M.J. Farabee of Estrella Mountain Community College of Avondale, AZ). We also recommend this paperback book: *Life Evolving: Molecule, Mind, and Meaning*, by Nobel Laureate Christian de Duve, Oxford Press. Two other web sites worth a visit are: (1) a thorough review of almost all aspects of Biology prepared by John Kimball and the more specialized Dolan DNA Learning Center. For a thorough immersion into Earth's early life history, we recommend Andrew Knoll's *Life on a Young Planet: The First Three Billion Years of Evolution on Earth*, Princeton Univ. Press, Nov. 2003. There is also a helpful Web site that treats developments in the field of Astrobiology (also called Exobiology) hosted by the Astrobiology Branch at Ames Research Center

A proviso:While the main goal of this and the next page is to provide a limited background for the reader on those aspects of Biology that deal with the origin and history of life, the effort will seem sketchy in that many topics that comprise the essence of the subject are treated either cursorily or have of necessity been omitted. The overall treatment of Biology requires considerable time and space (the writer's textbook has more than 900 pages), far more than can be devoted in this Section. But we will examine in some detail the biochemistry of life on this page, a topic of value in reviewing the story of life on Earth on the next page.

The capstone of this Section on Cosmology must surely be a consideration of the most provocative and fascinating Quest in the history of human life: the attempts to determine whether life of any kind - but specifically intelligent life (*Pyschozoic* is a term created to generalize such a stage of life on exoplanets) - exists elsewhere in the Universe. Philosophically, many on Earth hope that we are unique - thinking beings that are the pinnacle and teleological goal of a Creator's act. Scientifically, most cosmologists, biologists, etc. are coming around to the firm conviction that life does indeed exist elsewhere - throughout the Universe. This is a logical conclusion, since a huge Universe with just one tiny inhabited body on which conscious creatures exist strikes most scientists, and a growing number of philosophers, as extremely unlikely, and, from a practical sense, even a foolish, wasteful action by any Creator (this viewpoint is touched upon again later on this page).

We remind you at the outset of the excellent 2003 book on the subject of life in the Universe, David Grinspoon's *Lonely Planets: The Natural Philosophy of Alien Life*, cited on page 19-2. A few comments about the history of this idea, extracted from his book, are briefly treated before we begin with the review below:

Grinspoon points out that mankind has been speculating on life beyond the Earth for more than two millenia. The Epicureans of the late period of Greek philosophy before Rome took over that part of the Mediterranean believed that living creatures with intellects lived on one or more of

the planets and possibly the stars. Aristotle and Plato, however, argued against multiple worlds. Some early Christian theologians developed ideas that allowed for thinking life within the observed Cosmos (not quite a Universe as we know it today, but a realm that perhaps extended beyond the "spheres" that contained the Moon and Planets; see page 19-2). Similar speculations affected the medieval thinkers. But until Renaissance times, the vast majority held life to be unique to Earth. Copernicus, Kepler, and Galileo gave thought to the possibility of life elsewhere but never did seriously conjecture on the possiblities. Although not well-known to the public and even many scientists, the Italian Giordano Bruno (a Dominican friar *fried* at the stake by the Church for his radical beliefs) by 1600 had conceptualized a Cosmos filled with multiple Suns and their planets on which life was widespread.

At the dawn of the Age of Enlightenment, a treatise advocating a "plurality of worlds" was published in 1686 by Bernard le Bovier de Fontenelle. This work had a strong influence on thinkers of the day. In the first half of the 18th Century several similar and provocative books by European natural theologians followed. The German philosopher Immanuel Kant was much influenced by these writings and came up with a precursor to modern ideas for the formation of the Solar System. He espoused a much wider distribution of life within our Solar System and probably elsewhere. Laplace, who modified Kant's models, also took the view that life was established beyond Earth. The 19th Century saw similar and varied views favoring the "universality" of life. In the early 20th Century, Percival Lowell popularized this notion with his claims that "canals" existed on Mars. As the dawn of the Space Age arrived in the 1950s, many scientists and much of the general public retained the view that life was likely to be found in other parts of our solar System. The first spacecraft to fly by, orbit, and land on Mars tended to dampen this enthusiasm. But flying saucers and movies about ET and Close Encounters have pumped up the hopes of the Common Man that in time life will be found beyond the Earth.

From an anthropocentric outlook, the importance in understanding planetary formation mechanisms and history is the assumption (not yet a clearcut fact) that planets possessing certain appropriate conditions are the harbors of life. Life, it is believed by Earth dwellers who can think, may well be the most complex and advanced feature in the Universe, based on the presumption that it has evolved into a state resulting in lifeforms that perceive beyond sensing, analyze through reason, and evaluate most other aspects of known existence. Life, under this viewpoint, is the quintessential achievement in the evolution of the Universe to date. Whether life on Earth stands at the pinnacle, or somewhere below, has yet to be established - statistically, it is most likely that somewhere in the Universe even more highly developed living creatures, with superior intellects, exist today or have in the past. (The ideas just enunciated are closely associated with the modern doctrine called *humanism*).

The expectation that some life exists elsewhere in the Universe will depend on the nature of and conditions for life itself. Life can be defined by properties that are both chemical and functional. Paul Davies, in his excellent book *Other Worlds*, cites seven essential prerequisites for life to originate, survive, and flourish:

1. There must be an adequate supply of the elements that comprise organic matter - Carbon, Oxygen, Hydrogen, Phosphorus, Sulphur, Calcium, and other elements.

2. There must be little or no risk of contamination by poisonous chemicals (mainly in the atmosphere and oceans), such as ammonia or methane.

3. The climatic temperature must remain within the narrow range of 5 to 40 degrees Centigrade, which is a mere 2% of the temperature range found within the Solar System as a whole.

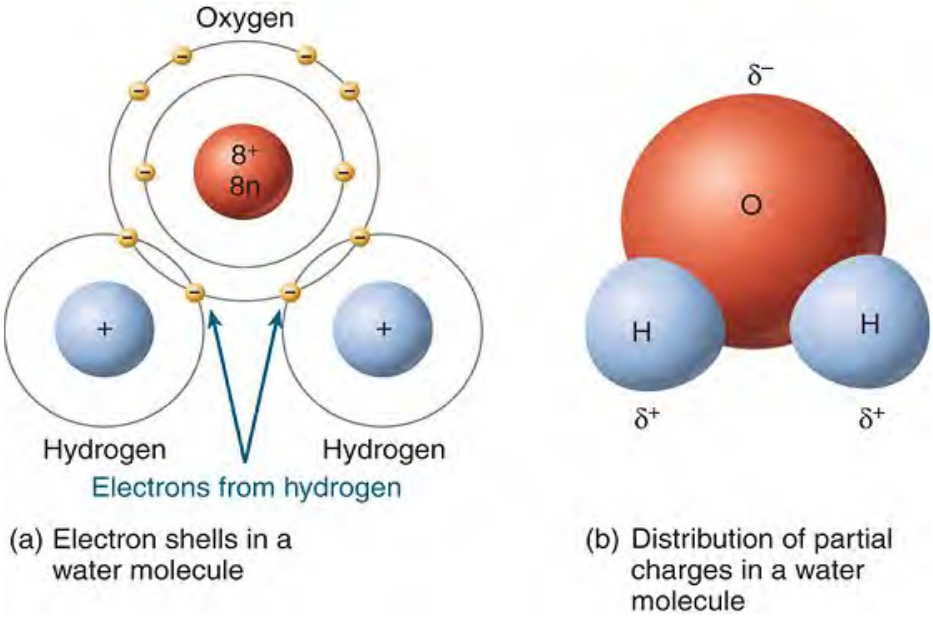
4. A stable supply of available energy must power living matter; for Earth this is primarily the Sun; internal heat sources, such as the "smokers" vents on the seafloor may also have been involved.

5. Gravity must be strong enough to keep the atmosphere from escaping into space, but it must be weak enough to enable life to move freely within the surficial envIronment.

6. A protective screen must exist to filter out the Sun's harmful Ultraviolet rays, which for Earth is the delicate layer of ozone in the upper atmosphere.

7. A magnetic field must exist in order to prevent cosmic subatomic particles that can damage or kill life from impinging on the Earth.

Missing from the Davies list, but crucial, is the **presence of water** (whether this is a universal condition or just applies to Earth is not yet established). Water is one of the most versatile and essential substances known on Earth and in and on most planets (both those of the Solar System and around other stars). Water is in essence an Oxygen atom with two embedded protons (from the Hydrogen). Here is a structural representation:



(a) Electron shells in a water molecule

(b) Distribution of partial charges in a water molecule

In this configuration the H_2O acts as a polar molecule, with one end being positively charged and the other negatively charged. In this sense, water acts both as an acid and a base (but with its pH of 7, it is considered neutral). Water can perform many functions, providing itself as a molecule or as a source of H and O ions. Among functions named as terms are its role as a solvent (breaks down the chemical structures of other substances - especially those held together by ionic bonds), hydrolysis (water splits into H^+ and OH^-, which react with various substances, especially organic molecules), hydration (water incorporated without change into the crystal structure of a substance), and redox (reduction; oxidation) reactions. Water is the most abundant molecule at the Earth's surface. It is known to humans in its three states or phases: solid (ice), liquid (water), and gas (steam). Water makes up about 70% (by weight) of the human body.

Judging on what we know conclusively from the one sample available to earthlings - namely, life on Earth itself as the only confirmed example in the Solar System - the essential chemical incredients are **Carbon**, the crucial element in organic molecules (of which proteins are the fundamental component) of great complexity and variety that are the basis of life, together with Hydrogen, Oxygen (some of which is combined as water which dominates the soft parts of human and many other organisms), Nitrogen, Phosphorus, Calcium (mainly in hard parts) and Sulphur, and to a lesser degree other elements as important functionaries, such as Iron, Magnesium, Chromium, etc. The amazing thing about this assemblage of critical elements is that they all at times in the past resided *in stars* and much of the Hydrogen itself can be traced back to the first minute of the Big Bang. You and I, as humans, are truly star people - our heritage is cosmic in that our ingredients are either primordial- or stellar-derived.

As a quick synopsis on the nature of life, here is a simple list of the "The Characteristics (Properties) of Life", adapted from one put online by the Department of Zoology, Oklahoma State University.

**Organized structures: composed of heterogenous chemicals - in units of "cells"*

**Metabolism: chemical transformations that either break down molecules to release energy (catabolism) or use energy to build up molecules (anabolism)*

**Homeostasis: which maintains internal conditions separated from an outside environment*

**Growth and Development: conversion of materials from the envIronment into components of organism*

**Regulation: coordination of the organism's internal functions, including transportation of materials needed to function*

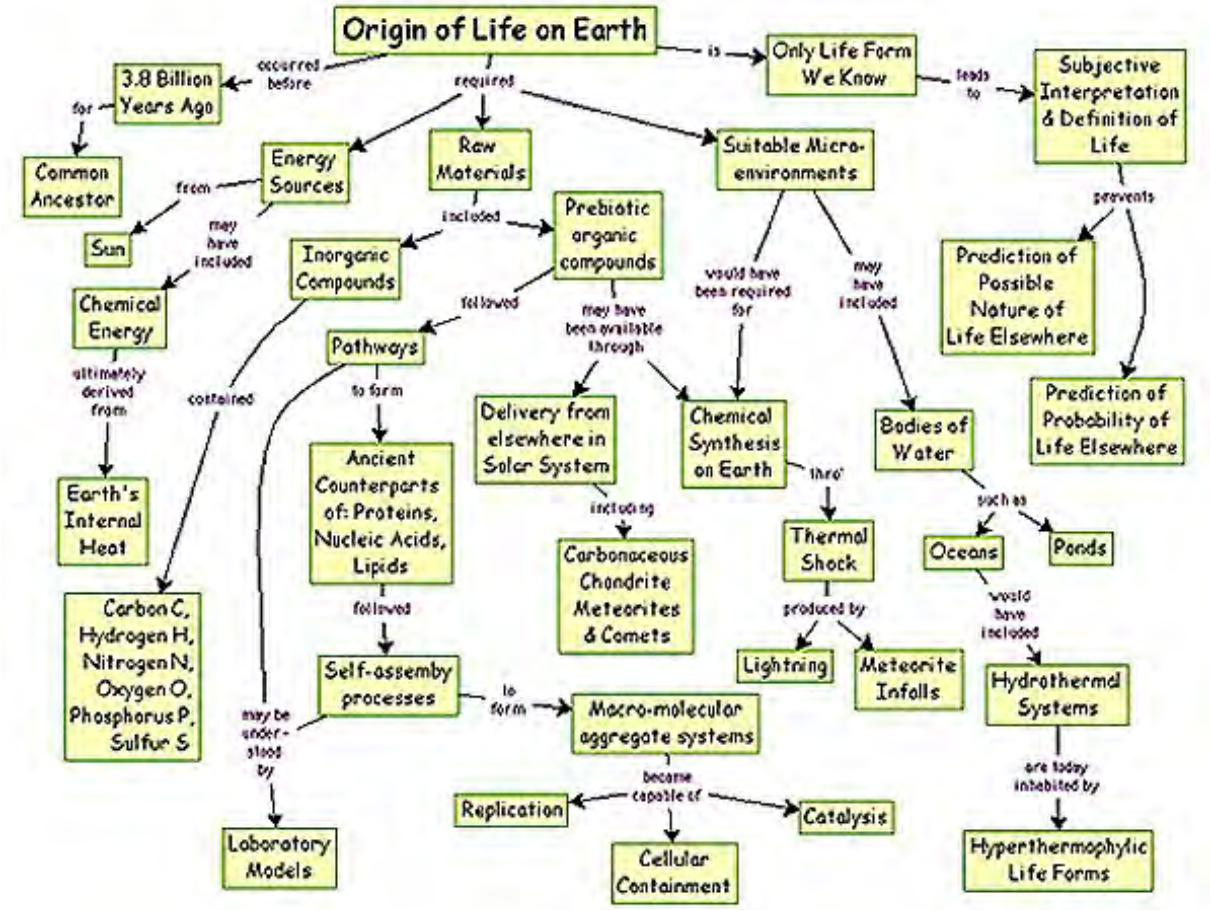
**Sensitivity: reaction to select stimuli, physiologically and/or behaviorally*

**Reproduction: making copies of individuals via the mechanism of genetic transfer: sections of DNA molecules that contain instructions for organization and metabolism*

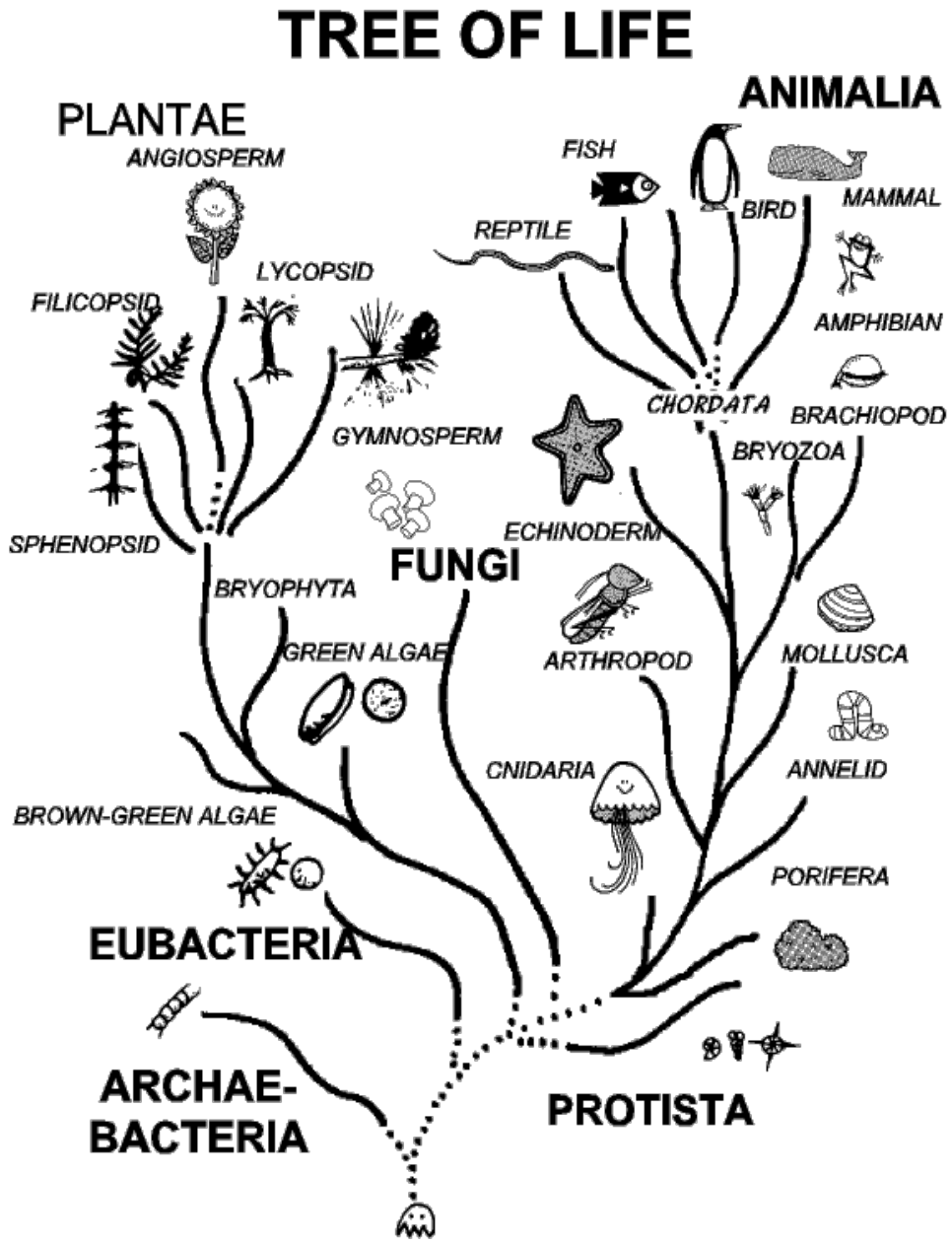
**Evolution: change in characteristics of individuals, resulting from mutation and natural selection - these result in adaptations; Heredity is the outcome.*

Thus, the principal functional manifestations of life (based on our studies of this phenomenon on Earth - our only sample so far) are, to reiterate what was listed above: cellular-organization; reproducibility; growth cycle and dependence on nutrition; metabolism (in higher forms) respiration (in some types); (usually) movement of some kind; propensity to evolutionary modification, and, for vegetative types, utilization of photosynthesis. Intelligent life, furthermore, is marked by consciousness, reasoning, abstraction, reliance on memory, communication, and awareness of time and other essentials of existence; free will and "soul" are properties of a more metaphysical nature and harder to prove as realities.

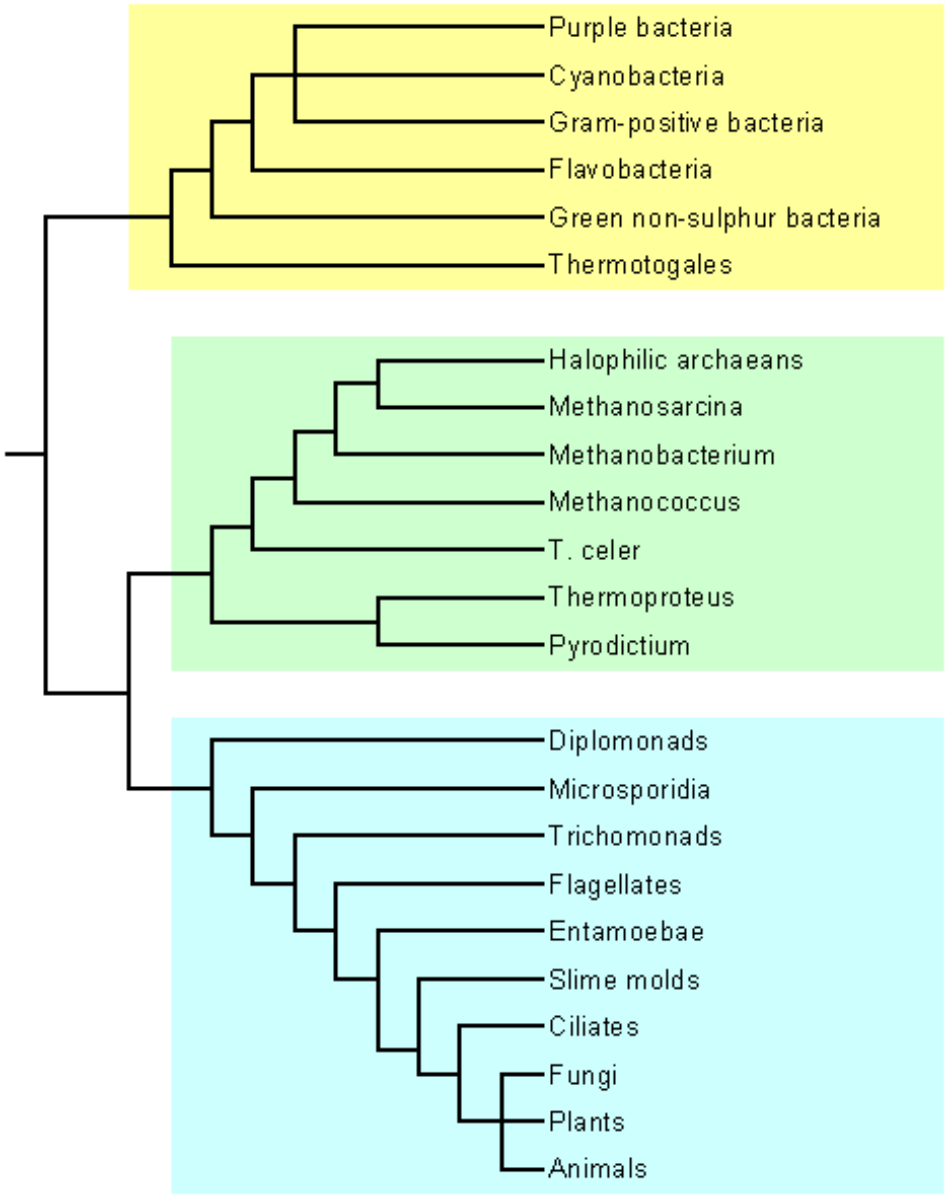
This next diagram was taken from the Internet without any indication of its source nor any explanation of its content. It is put here without any comment, treating it as a "talking point" relating to some of the questions that are relevant to the nature and origin of life on Earth. The implications within the diagram should also pertain to life elsewhere in the Universe. Draw your own ideas and conclusions.



Life falls into two broad categories: prokaryotic (no cell nucleus) and eukaryotic (nucleus). Life also can be categorized as single celled or multicelled and as autotrophic (obtains nutrients from inorganic sources) or heterotrophic (nutrients obtained by "feeding" on other organic sources). Life is classified by a taxonomic system (first espoused by Linnaeus) of hierarchical catetories - from the highest level (broadest number of constituents) to the lowest (most limited or particular). This ranges through **Kingdom; Phylum; Class; Order; Family; Genus; Species; (Subspecies)**. As an example, consider a honey bee: its species name is *mellifera*, its genus is *Apis*, it belongs to te Family *Apidae*, which is part of the order *Hymenoptera*, a nenber if the Class *Insecta*, that falls within the Phylum *Arthropoda*, in the Kingdom *Animalia*. Various proposed subdivisions of life at the Kingdom level are used: a common one is the six-kingdom system proposed by Woese: **Eubacteria**; *Archaebacteria* (some use Archaea) - both being largely single celled organisms; **Protista** (eukaryotic; both uni- and multi- cellular), which include algae, foraminifera, radiolaria, and diatoms; (the next three are all eukaryotic and multicellular) **Fungi** (yeasts; mushrooms); **Plantae**, with nonvascular plants (mosses et al) and vascular plants (ferns, conifers, angiosperms [flowering plants]); and **Animalia**,, with 14 phyla, to us the most important of which is **Chordata**, that includes amphibians, fish, reptiles, birds, and mammals. The evolutionary "Tree of Life" has followed this generalized pattern (alternate schemes have been proposed):

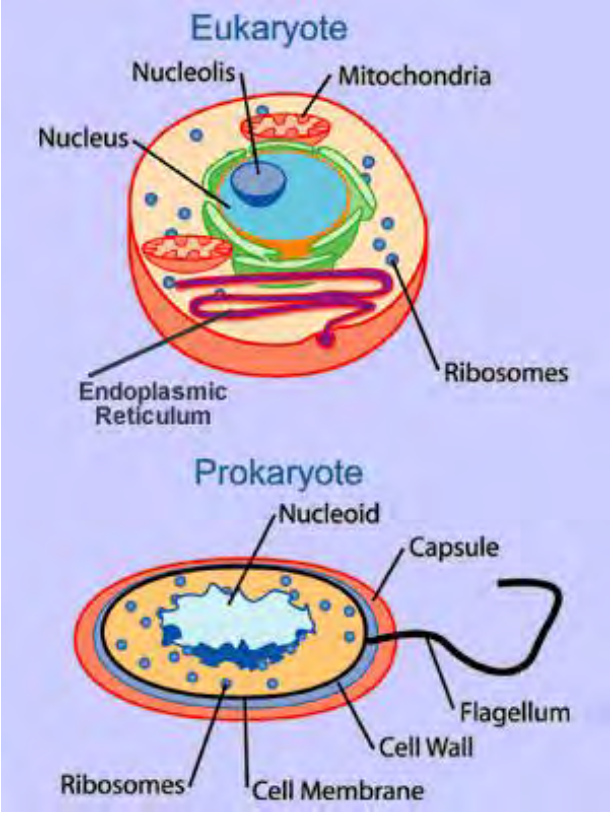


Life is often classified in three broad groupings. Here is such a version of this Tree of Life, with yellow denoting Bacteria, green assigned to Archaea, and blue representing the Eukarya, is (this type of plot of evolutionary trends is known as a cladistic diagram):

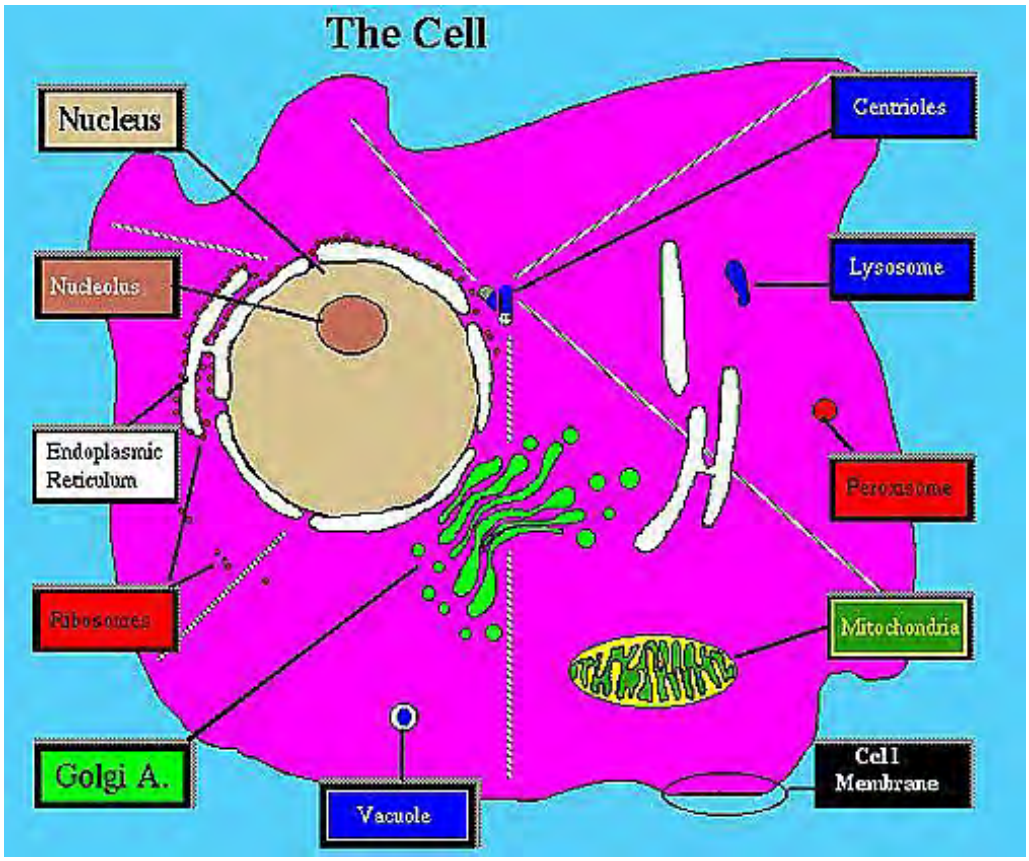


This is probably a good place on this page to present an overview about cells. For a fuller treatment go to the Wikipedia web site on Cell

Two fundamental cell types exist: Prokaryotic (no nucleus) and Eukaryotic (nucleus). Both single-celled and multi-celled organisms exist among each type. Both types contain DNA and ribosomes. Prokaryotic cells typically are 1 to 10µm in size; Eukaryotic cells can be 100 µm wide; larger cells are known. This figure depicts the two cell types:



A Eukaryote cell is much more complex, as suggested by this generalized diagram showing its makeup:



Each of these components (most are lumped under the term "organelles") will be concisely defined:

The **Cell Wall** (called Cell Membrane in animals), in bacteria is usually made of peptilogycan and in eukaryotes cellulose or chitin. Its functions are to enclose the cell interior, protect the cell, and aid in transfer of material in and out.

The **Nucleus**, commonly spherical and enclosed by a double membrane, contains the chromosomes (gene assemblages of DNA).

The **Nucleolus** is host to genes for rRNA synthesis.

Cytoplasmis the jelly-like matrix that surrounds the nucleus of a cell and is bounded by the cell membrane. It includes the organelles of the cell as well as the sugars, amino acids, and proteins that the cell uses for growth and reproduction.

Ribosomes are protein-RNA complexes that are sites of protein synthesis.

Vacuoles are open sacs available for digestion or storage of waste products; may contain degraded protein, can become water-filled.

Gogli Apparatus consists of stacks of vesicles (openings) in which proteins made in the cell are prepared for export from the cell

Lysosomes are vesicles, derived from Gogli A., containing digestive enzymes that attack defunct organelles and other cell debris.

Centrioles are specialized organelles that produce microtubules that influence cell shape, move chromosomes during division, and aid in developing cilia and flagella.

Peroxisome use enzymes to remove superfluous electrons and Hydrogen atoms; Hydrogen peroxide is a by-product.

Mitochondria are double-membrane organelles that provide "power" from the cell by oxidative metabolism.

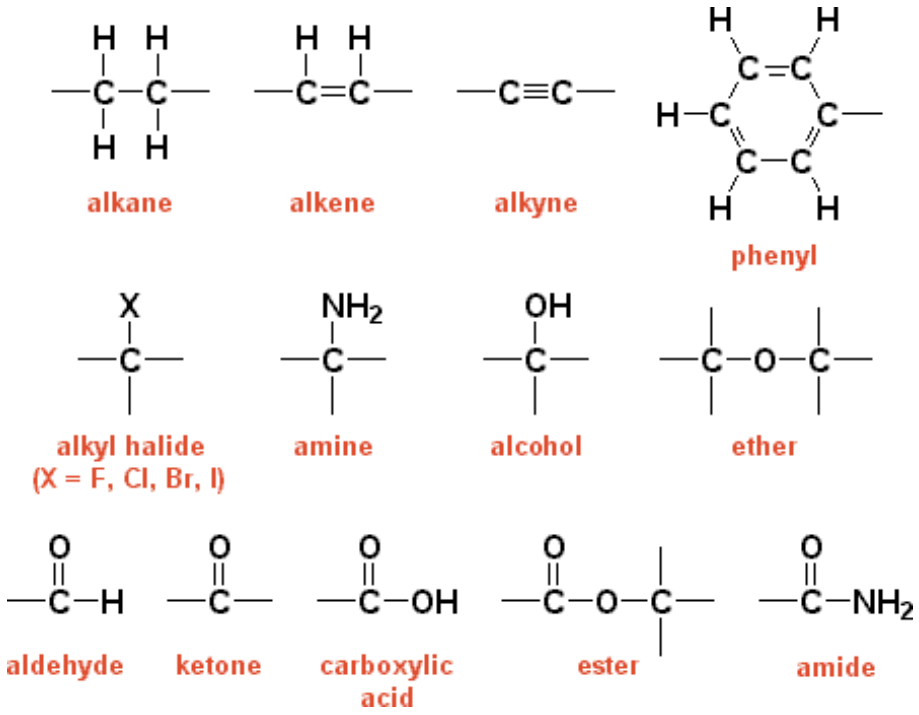
Endoplasmic Reticulum serve as networked membranes that aid in making vesibles; also assist in synthesizing proteins and lipids.

Not shown in diagram: **Chloroplasts**, which control photosynthesis in plants and **Chromosomes**, which are long DNA threads that host hereditary information.

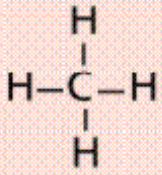
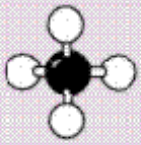
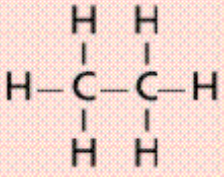
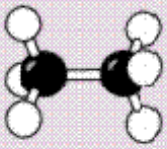
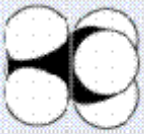
These definitions refer repeatedly to DNA and RNA. These are two nucleic acids which will be discussed in detail further down this page.

Understanding how life is organized into cells (through the science of Biology) requires some core knowledge of Organic Chemistry and its subfield Biochemistry. Only the most rudimentary ideas can be covered on the relevant sections of this page. Most of the illustrations were taken from the Online Biology Book, cited above, prepared by M.J. Farabee. He attributes on his site most of the illustrations we will use to Purves et al.; *Life: The Science of Biology*; Sinauer and Assoc. Publishers and M. Freeman & Company.

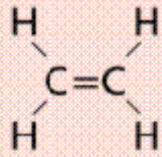
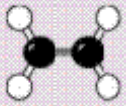
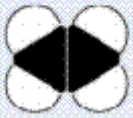
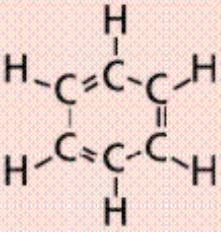
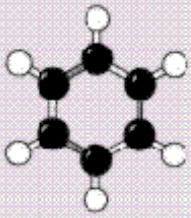
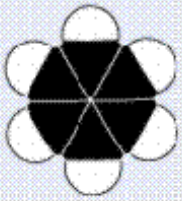
Organic substances are combinations of Carbon with other elements (most common are Hydrogen, Oxygen, Nitrogen, Sulphur, and Phosphorus). Bonding with the Carbon is of the covalent type - two atoms share one or more valence electrons. There are literally tens of thousands of organic compounds. But most of these can be grouped in systematic ways. One is the so-called *functional group* in which there is a basic unit built around carbon that can be combined with other elements or radicals. This diagram gives some examples as an overview (other, more specific diagrams will appear later on this page):



Organic molecules in their simplest form constitute the Hydrocarbons which are built from Carbon atoms (which have 4 electrons in their outer shell and can accept 4 more to complete the shell. This can start from the simplest Hydrocarbon (CH_4, methane, and built up in chains or rings, as indicated in the diagrams below:

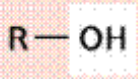
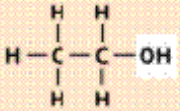
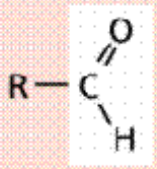
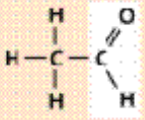
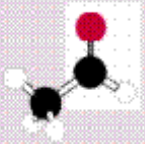
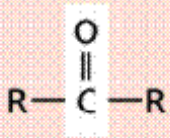
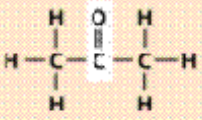
Compound (molecular formula)	Structural formula	Ball-and-stick model	Space-filling model
Methane CH_4			
Ethane C_2H_6			

From Farabee/Purves et al - see above citations

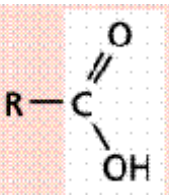
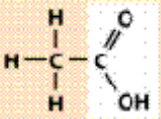
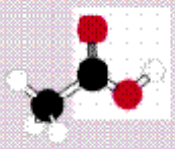
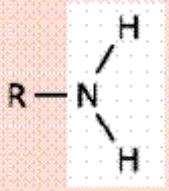
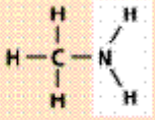
Compound (molecular formula)	Structural formula	Ball-and-stick model	Space-filling model
Ethylene (Ethene) C_2H_4			
Benzene C_6H_6			

From Farabee/Purves et al - see above citations

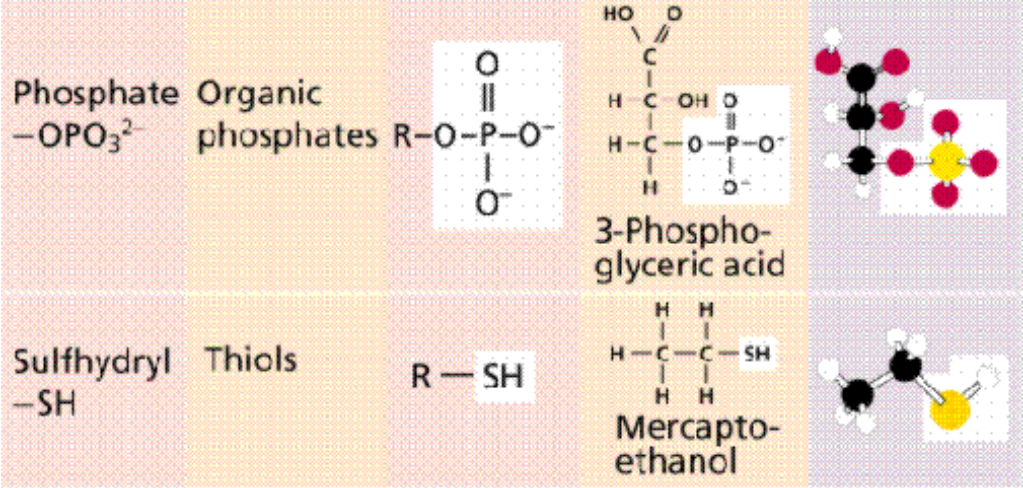
A wide range of more complex organic molecules can develop from adding various molecular functional groups that include N, O, P, S and other elements to open positions in the Carbon shell or to Hydrogens. These are the principal units:

Functional group	Class of compounds	Structural formula	Example	Ball-and-stick model
Hydroxyl –OH	Alcohols		 Ethanol	
Carbonyl –CHO	Aldehydes		 Acetaldehyde	
Carbonyl >CO	Ketones		 Acetone	

From Farabee/Purves et al - see above citations

Carboxyl –COOH	Carboxylic acids		 Acetic acid	
Amino $-NH_2$	Amines		 Methylamine	

From Farabee/Purves et al - see above citations



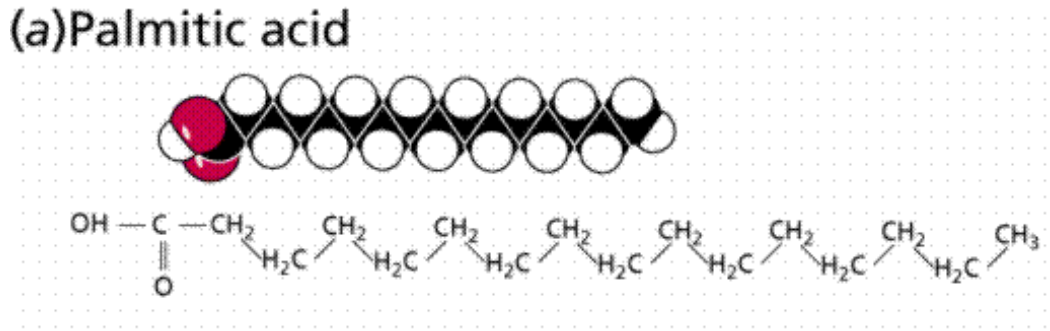
From Farabee/Purves et al - see above citations

These molecular varieties fall into groups such as saturated Hydrocarbons with single H-C bonds (Alkanes), double bonds (Alkenes), and triple bonds (Alyknes); ring structures are represented by Aromatic Hydrocarbons. Isomers are organic molecules with the same chemical formula but different atomic arrangements. Among the derivative organic molecules (Hydrocarbons with parts replaced by functional groups) are Alchohols (--OH functional group), Ether (--O--), Aldehyde (--CHO), Ketone (--CO--), and Carboxylic Acid (--CO_3H).

Biochemistry is a vast subfield of the more general Organic Chemistry. While the subject is complex and detailed (see links above), a few general ideas developed around key illustrations taken from M.J. Farabee's site (cited above), and one from Raven/Johnson are introduced here:

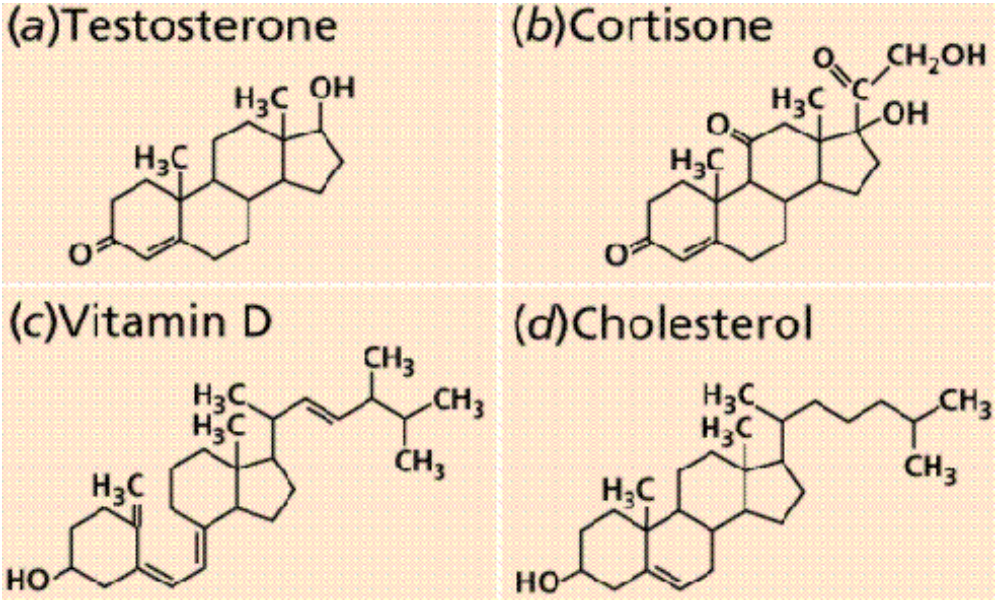
There are four major groups of organic molecules that also are the fundamental categories in living matter: Lipids; Carbohydrates; Proteins; Nucleic Acids (RNA; DNA). We will describe the first two in limited detail.

Lipids include fats, fatty acids, certain oils, waxes, terpenes, and steroids (one example being cholesterol). They consist of polymers of CH_2 and CH_3. Glycerol, a typical fatty acid, has the formula: $HOCH_2CH(OH)CH_2OH$. Palmitic Acid has this structural arrangement:



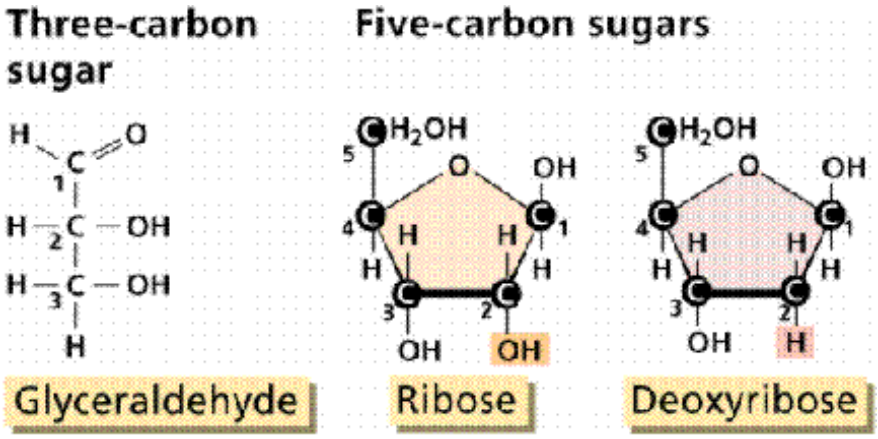
From Farabee/Purves et al - see above citations

The molecular structures of some common steroids are depicted below:



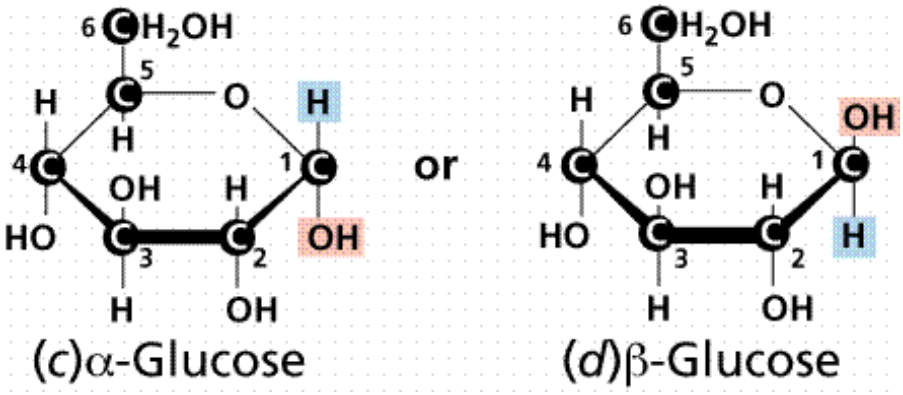
From Farabee/Purves et al - see above citations

The second major group, the **Carbohydrates**, also known as Saccharides, includes sugars, starches, glycogens, and cellulose. The basic formulaic unit is: CH_2O. One important group, the Monosaccharides, includes ribose and deoxyribose, ring structures made up of pentagonal (5 Carbon) sites - these are two fundamental components of RNA and DNA. Example:

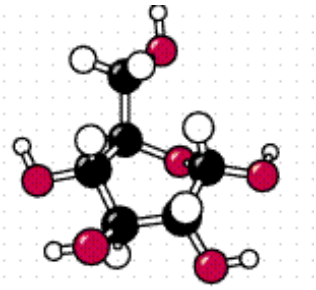


From Farabee/Purves et al - see above citations

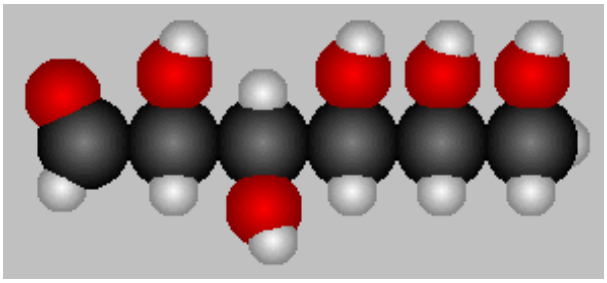
Glucose is said to be the most abundant biochemical molecule within terrestrial life: Its formula is $C_6H_{12}O_6$. Here are several views of glucose; at the top are two isomers, below is a 3-D stick and ball representation, and at the bottom is a side view:



From Farabee/Purves et al - see above citations

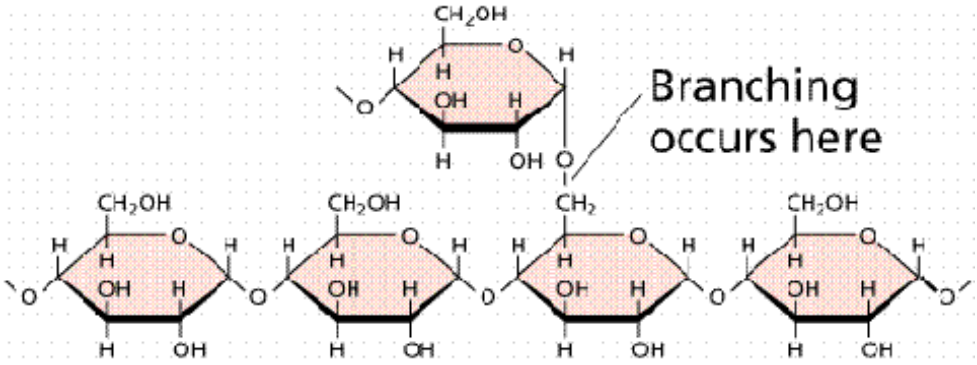


From Farabee/Purves et al - see above citations



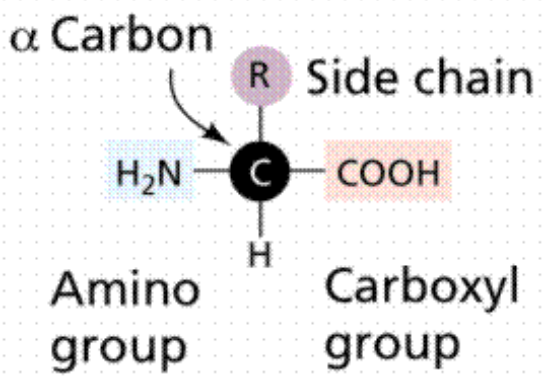
From Farabee/Purves et al - see above citations

Structures comprised of two joined rings are Dissarcharides (e.g. Sucrose and Lactose). Chains of rings make up Polysaccharides" These include starches (which store food energy in plants), glycogen (energy storage in animals) and cellusose (important cell wall component in plants). A starch molecule and



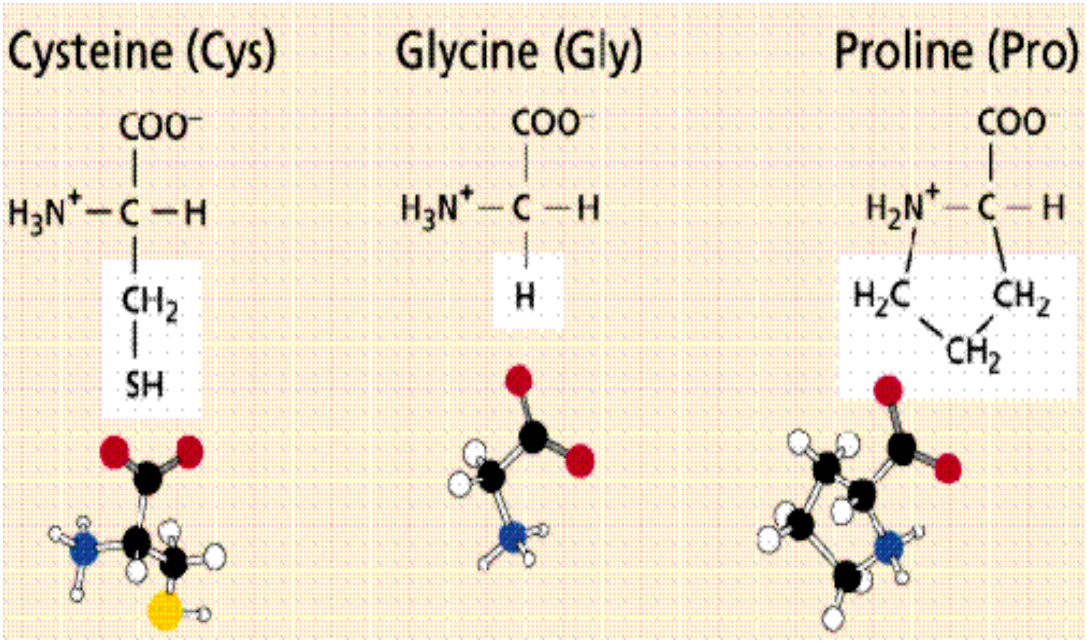
From Farabee/Purves et al - see above citations

Proteins, the most abundant constituents of organic matter, are built from linked individual amino acids. There are 100s of such acids but only 20 occur in proteins involving human tissue. The basic protein unit consists of a central Carbon, an amino group, a carboxyl group, and a side group, labeled R, that can consist of a variety of functional groups; the general molecule looks like this:



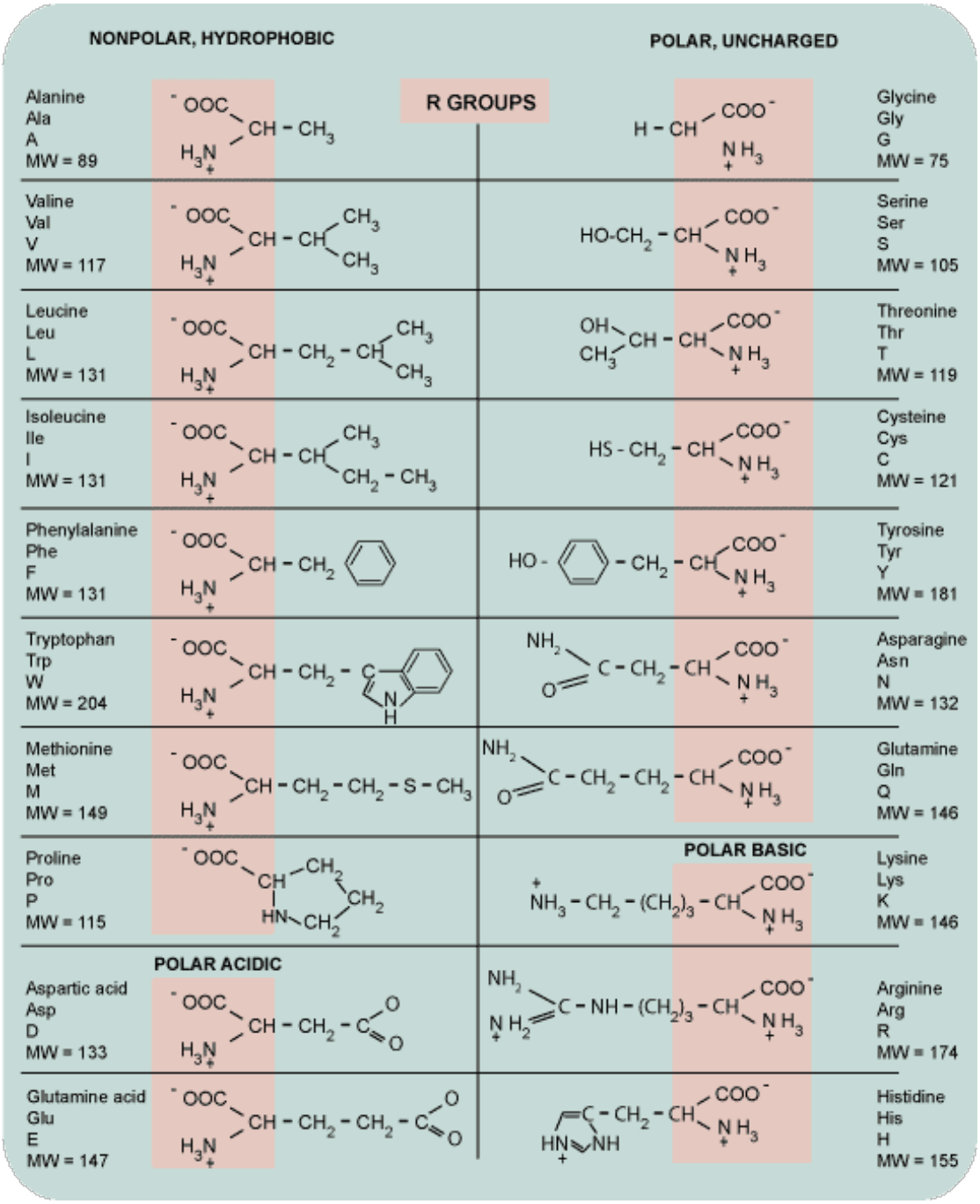
From Farabee/Purves et al - see above citations

Three examples of individual amino acid types are shown below:

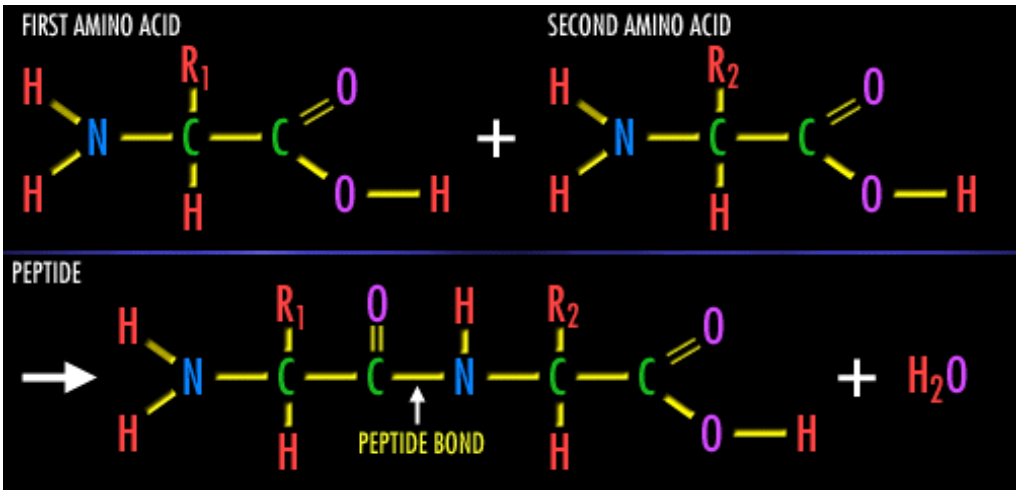


From Farabee/Purves et al - see above citations

The 20 amino acids are shown structurally in terms of different side groups in this chart:

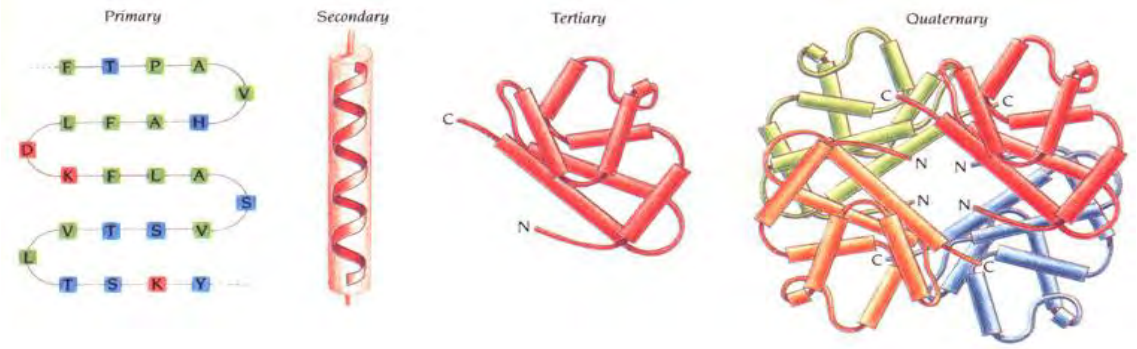


All 20 are α-amino acids, meaning that the amino group is always bounded to a Carbon atom. Various combinations of the 20 can link (bond) with one another (amino group to carboxyl group) to form polymers. The linkage is termed peptide bonding. Oligopeptides link only a few amino acids; chains of 100s of amino acids (polypeptides) are common:

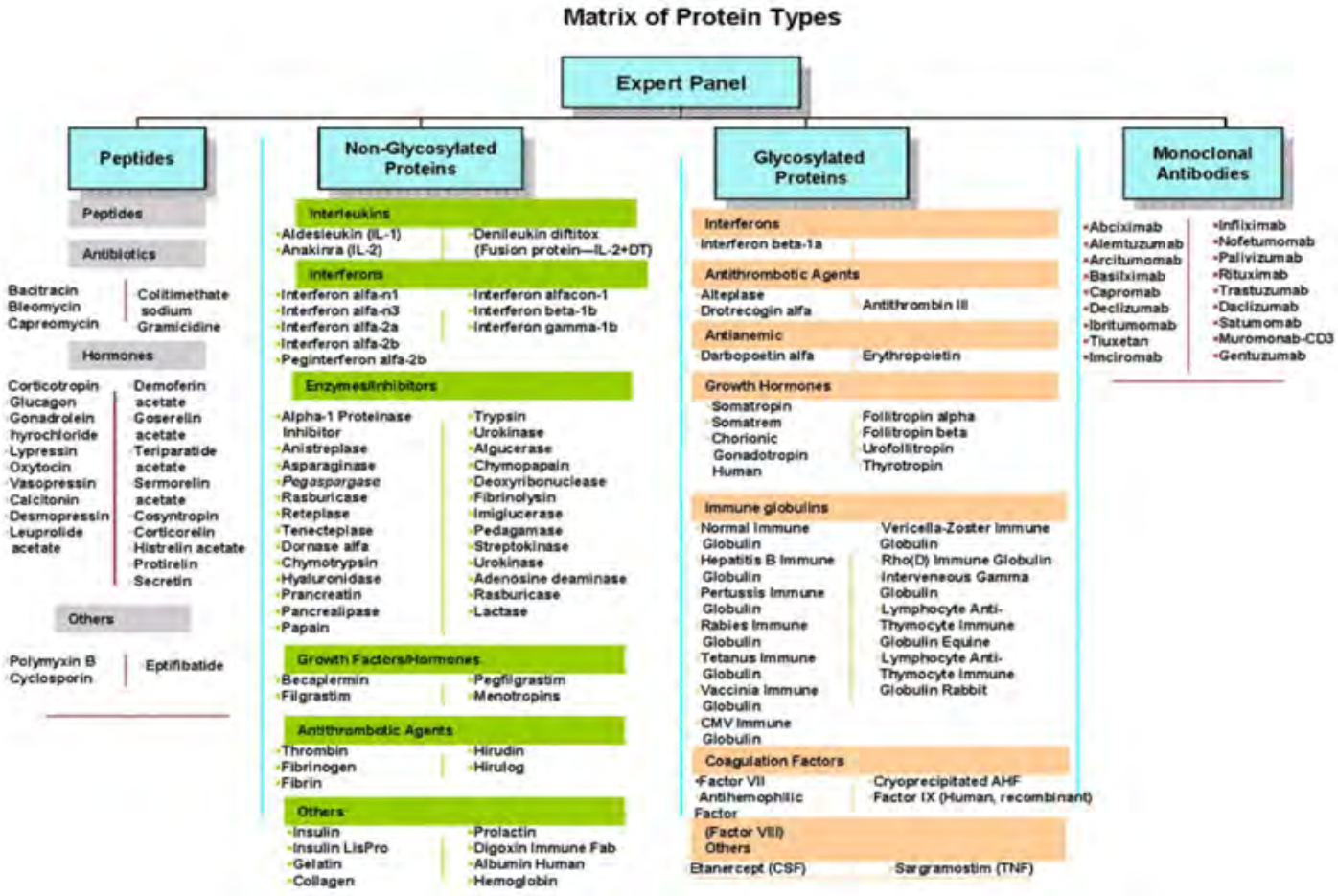


From Farabee/Purves et al - see above citations

Protein synthesis is a major goal in biochemistry. This is proving difficult because getting the amino acids in the right order is hard, unwanted reactions among side chains is common, and peptidization gives off energy which can decompose the desired end product. Protein structure can be simple chains (primary) or helical or pleated (secondary) or complexly folded (tertiary and quaternary structures).

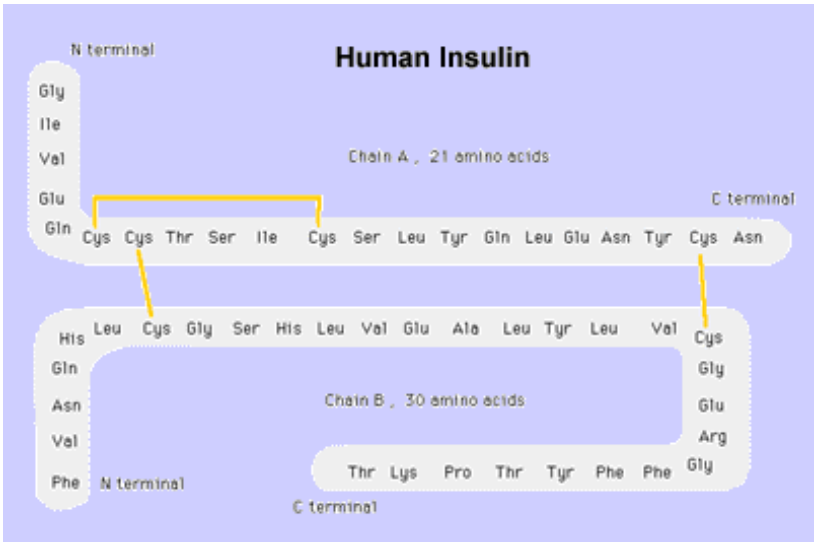


Proteins are the *dominant molecular types in cells*. Specific proteins (composition and/or structure) can specialize into Enzymes (that promote change and formation of new organic material by catalysis (bring about reactions without themselves being changed), Hormones (perform important physiological functions), Structural proteins (hair, skin, etc.), Transport proteins (carry material across membranes); Antibodies (involved in immune systems), and Muscle proteins (actin and myosin in fibrous forms). Enzymes are particularly important since they are involved in most biochemical reactions and can be very efficient; many enzymes work by attaching to the molecule they affect, which serves as the substrate responsive to their action. The list of important proteins is long and diverse. This table names many that are important:

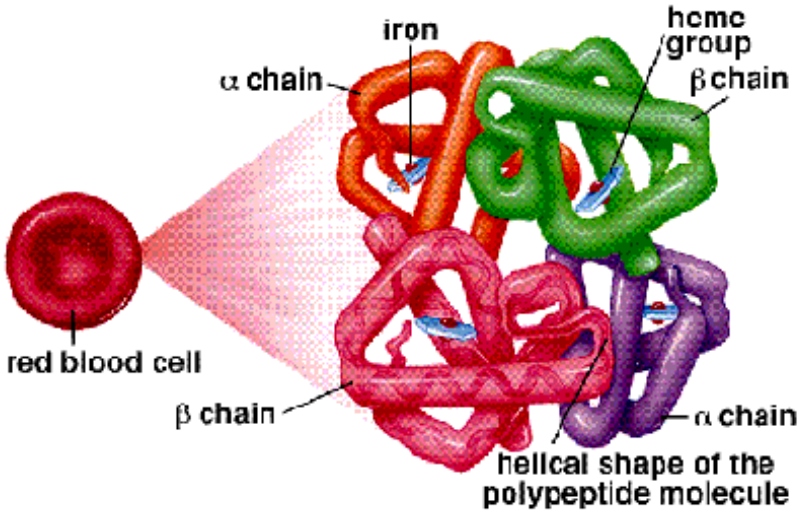


An example of how sequences of amino acids give rise to a specific protein molecule is given by

insulin, used by mammals to extract energy from sugars:



One of the most important proteins is hemoglobin, the main constituent of a red blood cell found in the circulatory systems of many mammals, including Man. As seen in the illustration below, hemoglobins consist of 4 folded units, 2 α-globin twisted chains and 2 β-globin chains, each harboring a heme (haem) group (blue disk) that contains Iron. The four hemes allow loose bonding of Oxygen to the Iron, which is carried in the blood and released where needed to promote oxidation reactions; this globular molecule also can carry CO_2 for removal during lung exhalation:



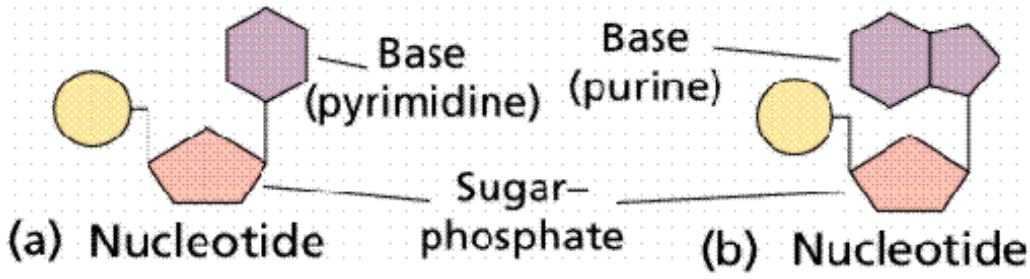
From Farabee/Purves et al - see above citations

Proteins have many uses in animals. For the human body, these are some of the more common functions: 1) Enzymes - control metabolism; 2) Immuniglobulins - provide antibody defense against foreign cell invaders; 3) Globins - carry O_2 and CO_2 in blood and muscle; 4) Transporters - controls osmosis; 5) Fibers - affect cartilage, nails, hair; 6) Muscles - control movement of body parts; 7) Albumin - osmosis in the blood; 8) Hormones - affect blood glucose; water retention.

The topic of how new proteins are generated is a vital one in general Biology. We must defer an answer to this until you have grasped the basics of how Nucleic Acids function, after which we will return to the protein production topic near the bottom of this page.

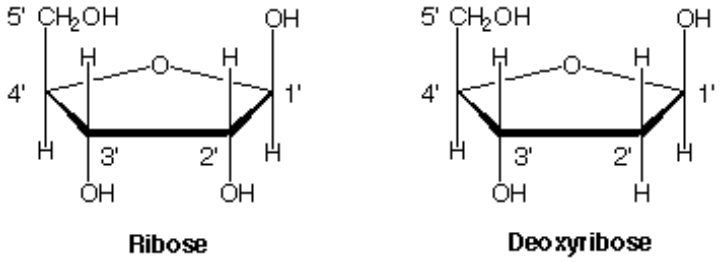
The fourth group, **Nucleic Acids**, are of fundamental essence to life in that they contain the biochemical molecules that hold the blueprints for making and copying cells. RNA and DNA are called the information-bearing cells that determine the nature of any given organism, from simple bacteria to humans. DNA is found within all cells, both prokaryotic and eukaryotic. The basic unit

is called a *Nucleotide*, consisting as shown below of a monosaccharide 5-Carbon sugar (pink), one of 5 Nitrogenous bases (purple), and a phosphate group (yellow).

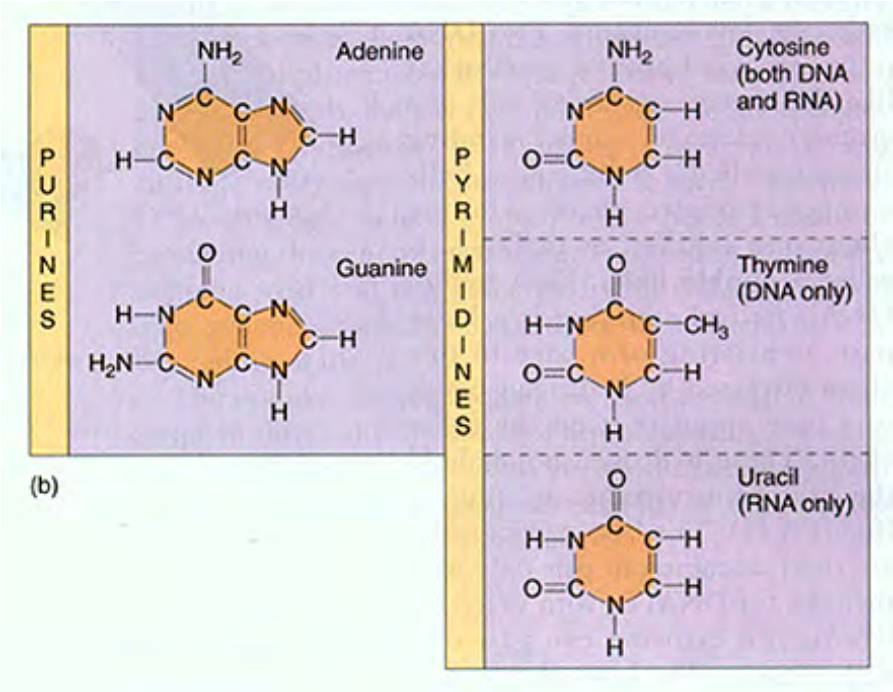


From Farabee/Purves et al - see above citations

The 5-carbon sugar molecule comes in two forms (ribose and dioxyribose). These are nearly identical, the only difference being that OH (hydroxyl)is the radical attached to Carbon-2 in the ribose and O (Oxygen) is attached to Carbon-2 in the dioxyribose molecule.

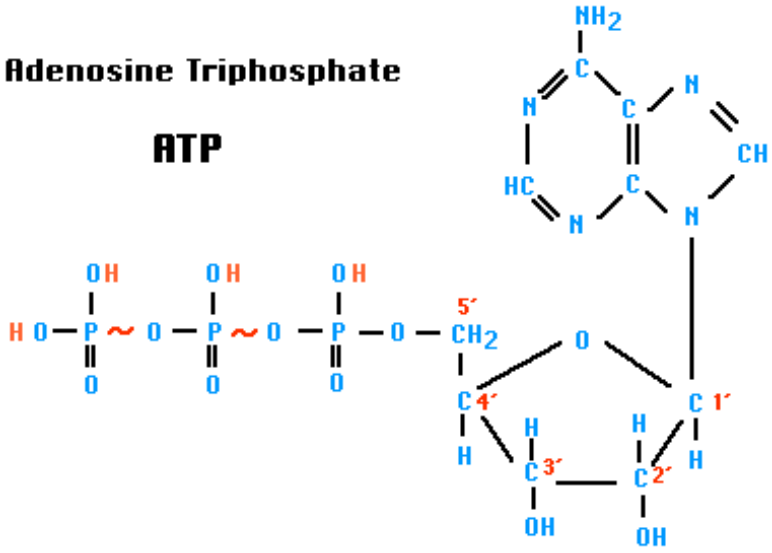


The structure and composition of the five bases (usually identified by their first letters: **A**denine; **C**ytosine; **G**uanine; **T**hymine, and **U**racil) is given by this diagram; note that they are organized into two group - the single ring Pyrimidines and the double ring Purines.



From Raven and Johnson, Biology, 6th Ed., McGraw-Hill, Inc

The bases can be part of important organic molecules involved in various life-sustaining processes. For example, Adenine is a component of ATP (Adenosine Triphosphate) (see below) that has several crital functions including metabolism (energy-supplying reactions) and the formation of DNA. Here is the structural formula for ATP:



However, the trilogy of the 5-carbon sugar molecule, a phosphate radical, and four of the five Nitrogenous bases organized together makes up the most fundamental organic units in all life forms. The two units are Ribonucleic acid (RNA) and Dioxyribonucleic acid (DNA). These are compared in this diagram:

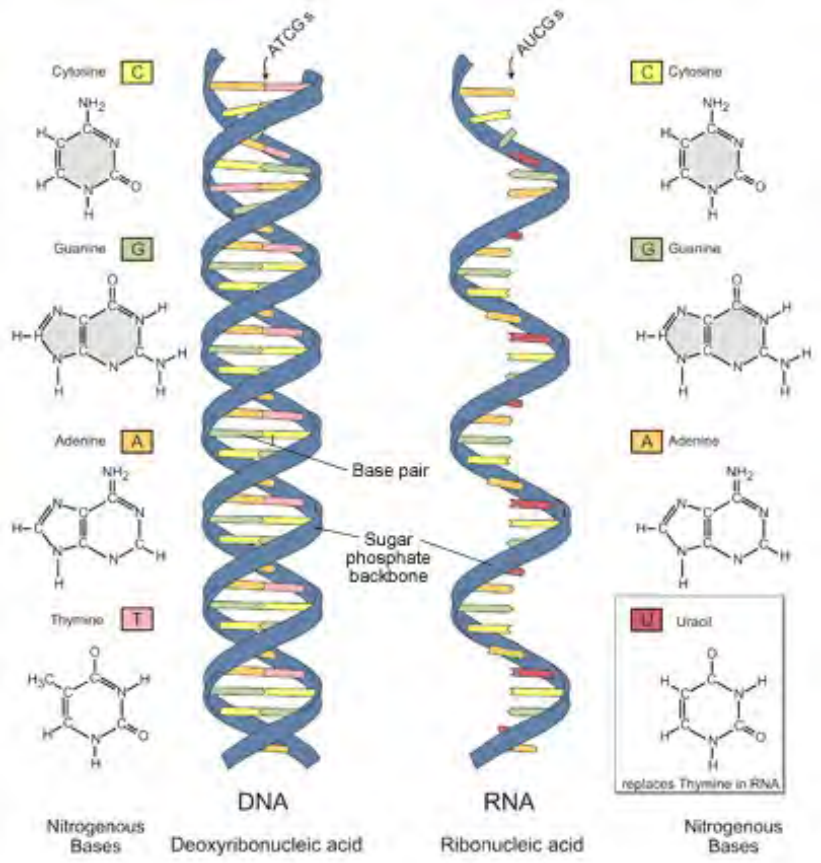
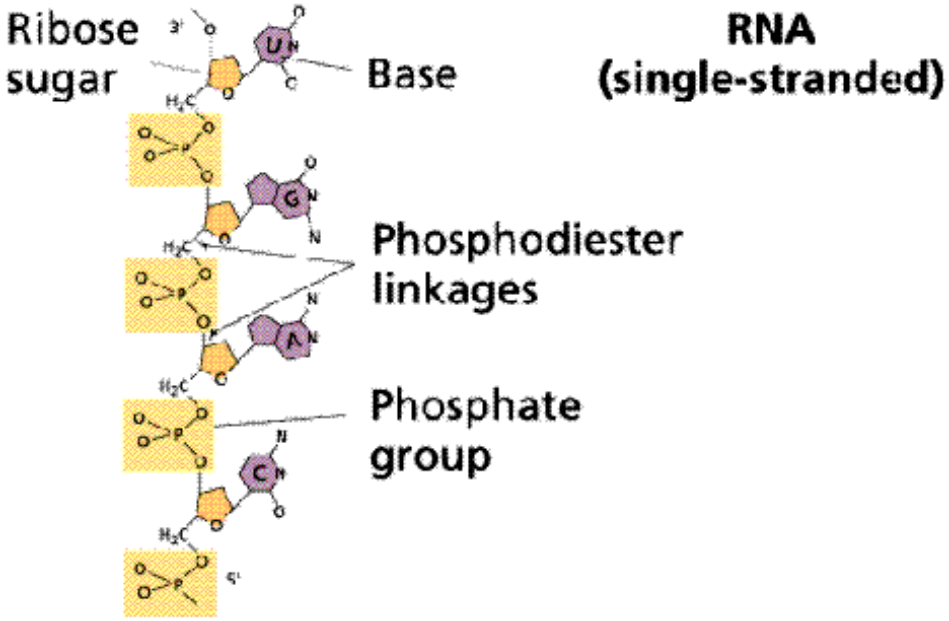


Image adapted from: National Human Genome Research Institute.

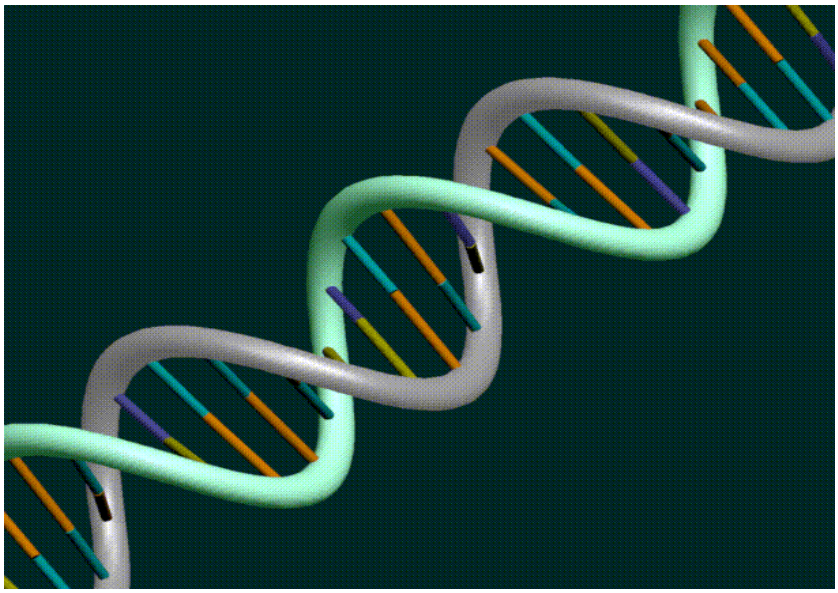
RNA occurs in single strands in which the sugars are linked by the phosphate (PO_4) phosphodiester bond and the bases lie on the other side of the chain. The four bases present are A, C, G, and U. Multiple chains of nucleotides make up the nucleic acid.



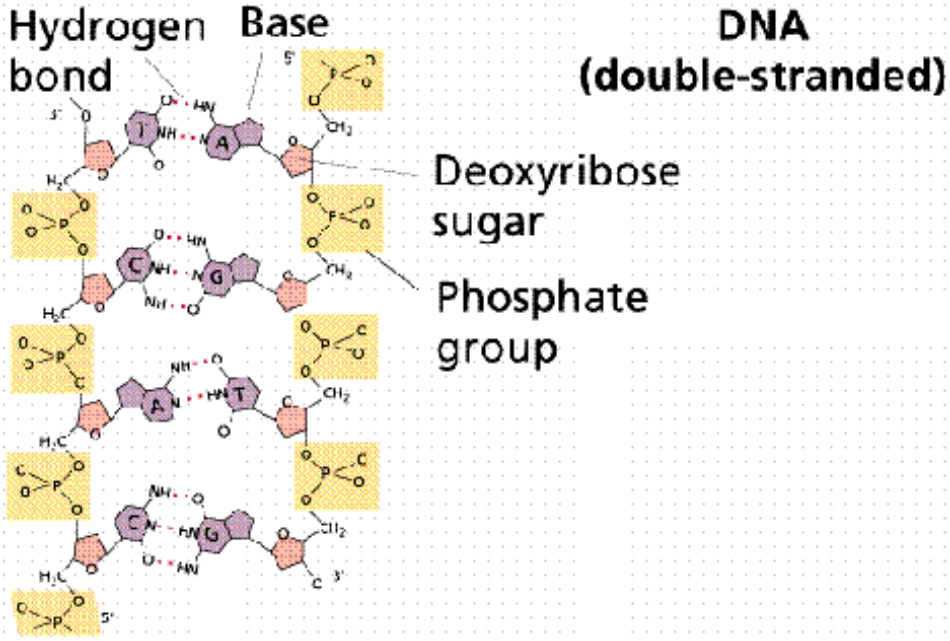
From Farabee/Purves et al - see above citations

The main roles of RNA are in its intimate involvement in the production of proteins, and its intermediary action in duplicating DNA during cell division and growth.

DNA, arguably the most famous organic molecule of all, consists of two chains (sometimes called "backbones") side by side that are coiled into spiral shapes. This diagram shows the general pattern of their structure:

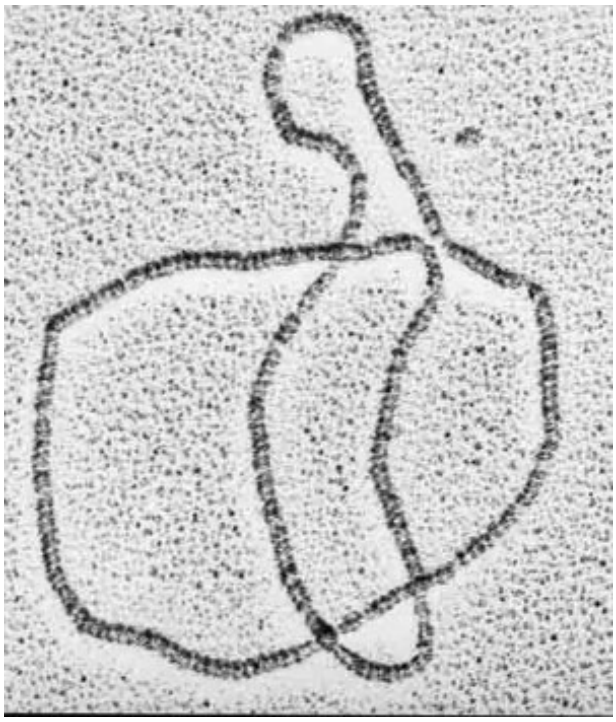


The bases A, C, G, and T (note that T replaces the U present in RNA) produce a link (the straight lines in the above diagram) between the two chains that tie them together. Only certain pairings are allowed: A - T and G - C. These couple by Hydrogen bonds. Here is a schematic of this arrangement.

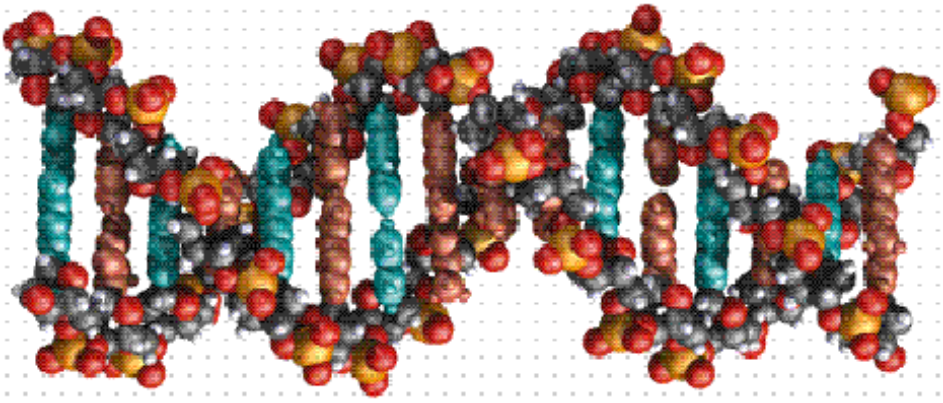


From Farabee/Purves et al - see above citations

A DNA molecule is very long (a few meters) but extremely thin (narrow; measured in nanometers). Here is an electron microscope photo of a DNA strand:

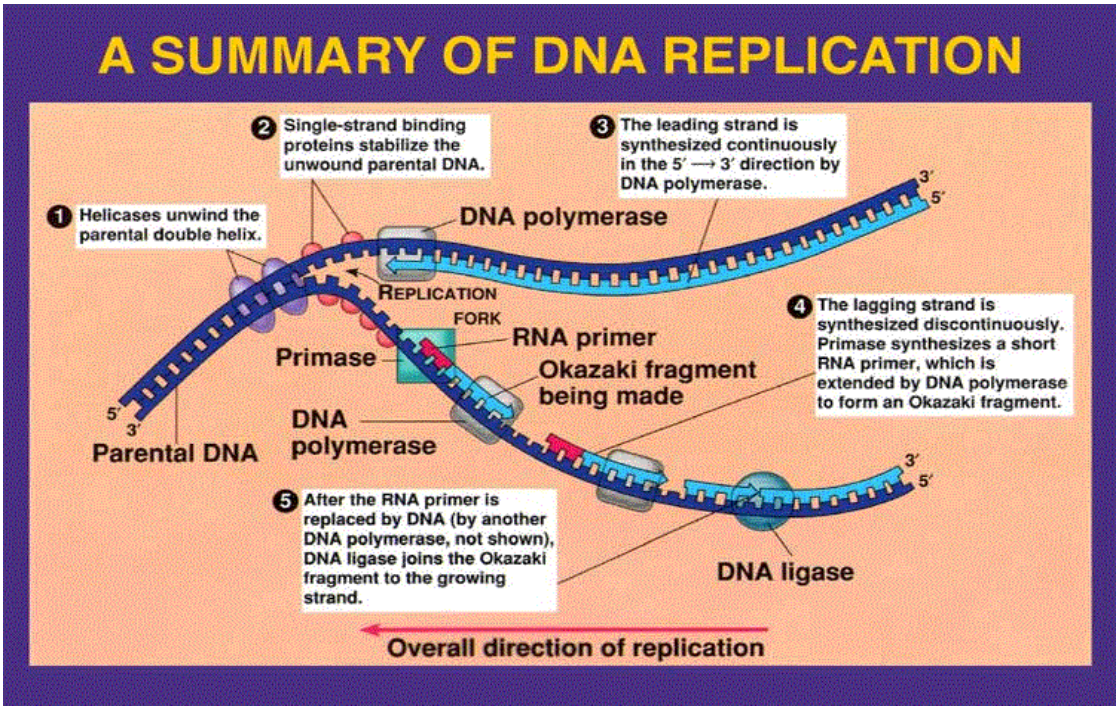


In 1953 Francis Crick and James Watson discovered through analysis of X-ray crystallography photos (done earlier by Rosalind Franklin and Maurice Wilkins - all four living in England) that the double-chained DNA was composed of two spiraling chains arranged in the famed appellation "*Double Helix*"; this discovery won Crick Watson, and Wilkins the Nobel Prize in Biology - Franklin unfortunately was dead by then). Here is how it looks as rendered by colored balls representing the sugar, phosphate, and base components:



From Farabee/Purves et al - see above citations

A process called "**replication**" involves duplicating specific DNA molecules. Rather than a lengthy discussion of this topic here, we refer you to the relevant Wikipedia web site. However, you should look at this next diagram (found commonly on the Internet) to gain a quick overview:



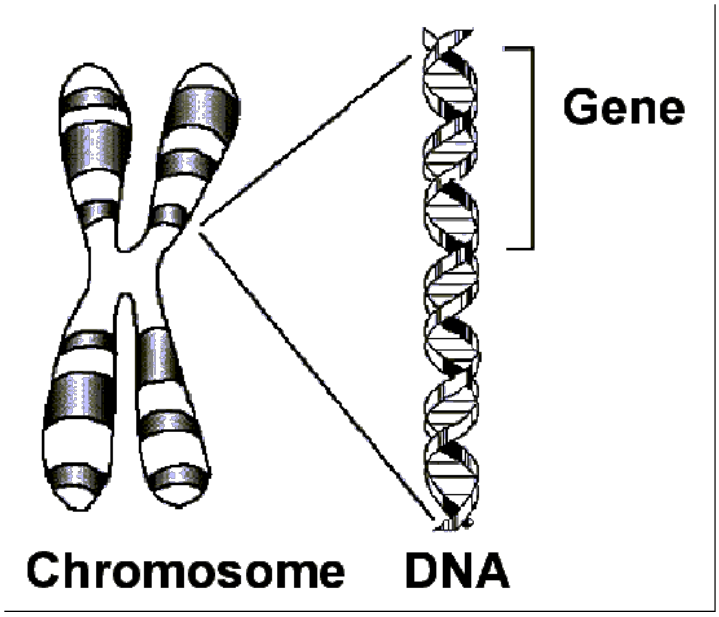
During replication mistakes in reproducing a gene pairing constitute the biochemical explanation for the **mutations** that Charles Darwin and his compatriot Alfred Wallace cited as the basic cause for natural selection. That is the cornerstone of the Theory of Evolution - this holds that individuals in a species that have the best adapted means for survival in their envIronment(s) will live longer and therefore have the better likelihood of passing their specific gene makeup to offspring in the gene pool ("Survival of the fittest"). These "good" genes thus persist in the population, and enrich it. Occasional mutations over thousands of generations in an organism's reproductive history lead to gradual changes until differences are great enough to warrant designation as new species. More fundamental and long-ranging changes lead to generic, familial, and higher level variants that become new types of animals or plants.

The importance of RNA and DNA reside in their roles in making new cells and in determining the nature/function of a cell by imparting correct instructions for that in the process. DNA codes the hereditary information needed to reproduce an organism and also is involved with RNA in producing new protein cells. This subject, at the heart of concepts of genetics (genes, chromosomes, and genomes), is far too voluminous (knowledge-intensive) to cover on this page (see the Biology Tutorial Internet sites for the details). Suffice to say that the RNA and DNA strands may be very long (macromolecules consisting of 1000s of nucleotides). For DNA, the sequence of A-T amd G-C, arranged in groups of 3 (codons), can, at such lengths, encode a

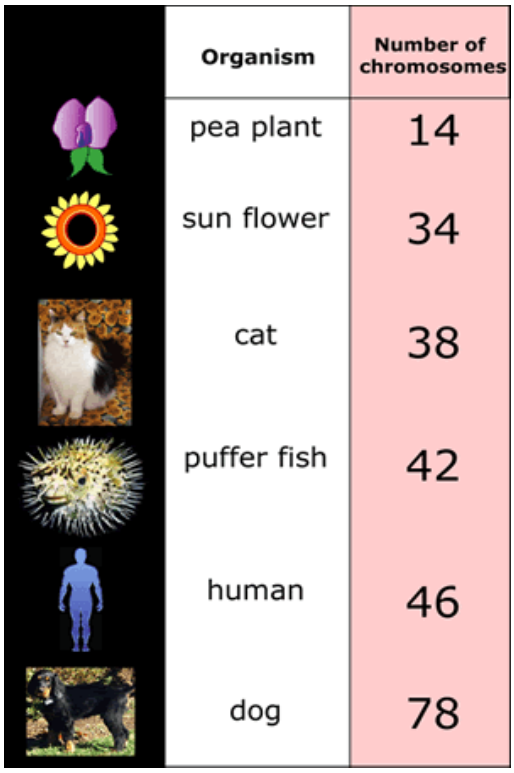
huge number of combinations. These lead to differences in a cell's nature (for example, different proteins are produced by RNA groupings that in turn are produced from DNA control), and when they are set in some fixed pattern of the sequenced pairings, a specific gene is determined. A ***gene*** is the fundamental unit of a (usually very long) sequence of nucleotides involved in DNA and RNA molecules. A string of varying genes make up a linear or circular chain of DNA and proteins that comprise ***chromosomes***, the organic molecule that determine sex, heredity and the production of proteins. The full complement of various genes in some particular pattern establishes the ***genome*** that uniquely specifies a given organism.

A whole Tutorial could be devoted to describing and explaining genes and chromosomes. Worthy as these subjects are, we will divert you to these two Wikipedia sites, the first on genes and the second on chromosomes. Here we will make a few comments using illustrations with brief write-ups.

First, this general diagram which shows a paired set of individual chromosomes, with black and white bands representing genes, each of which contains thousands of DNA strands (one being shown as expanded) each one of which can consist of a base pair CG or AT.

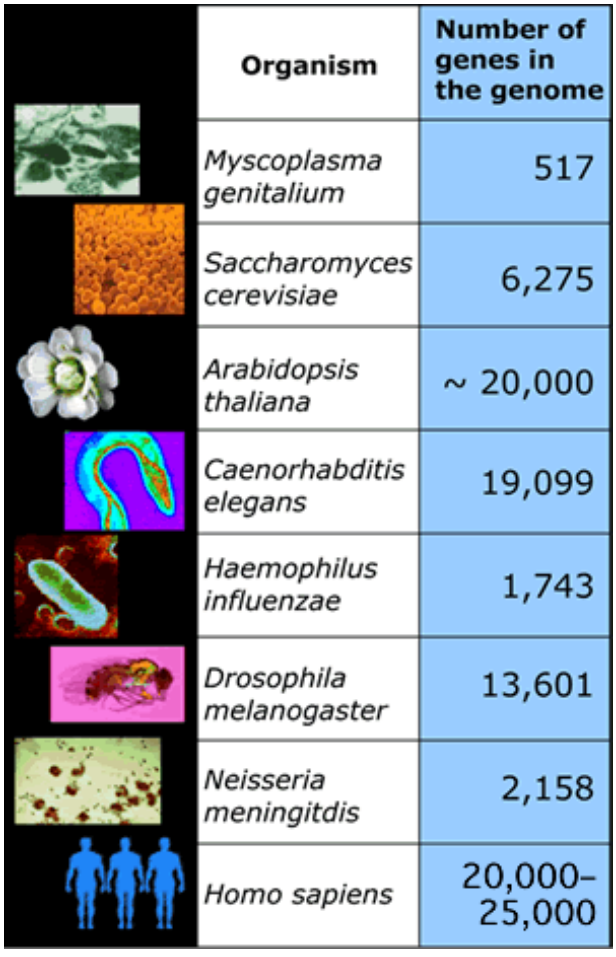


The number of pairs of chromosomes differs for different living organisms.



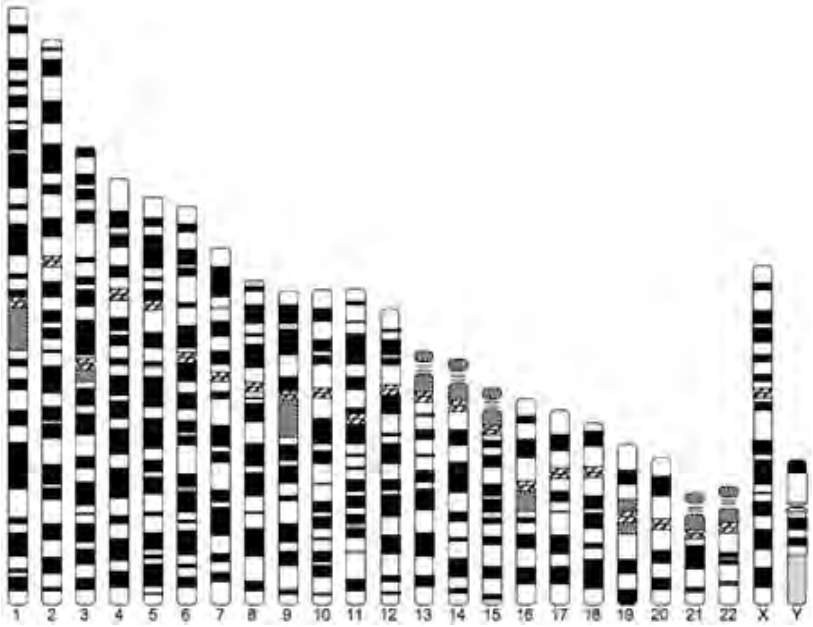
Organism	Number of chromosomes
pea plant	14
sun flower	34
cat	38
puffer fish	42
human	46
dog	78

The genes contain the basic information that defines the species and characterizes its individual traits. Note the number of genes estimated for the 46 chromosomes of the human species.

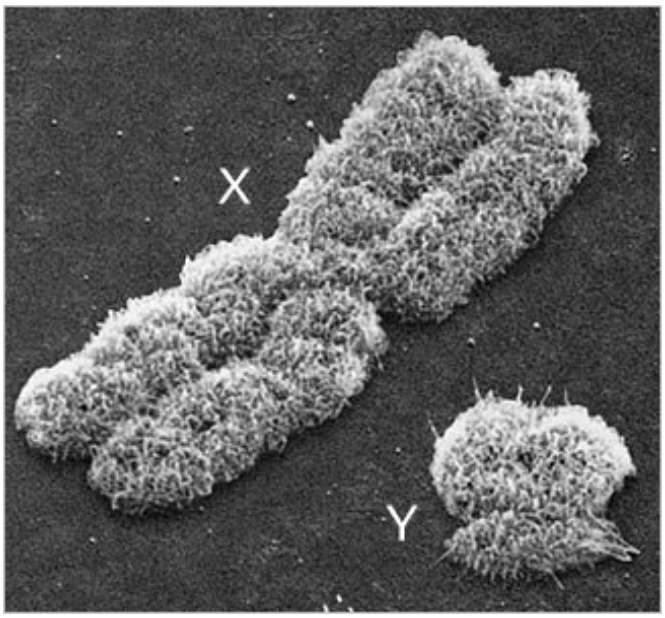


Organism	Number of genes in the genome
Myscoplasma genitalium	517
Saccharomyces cerevisiae	6,275
Arabidopsis thaliana	~ 20,000
Caenorhabditis elegans	19,099
Haemophilus influenzae	1,743
Drosophila melanogaster	13,601
Neisseria meningitdis	2,158
Homo sapiens	20,000–25,000

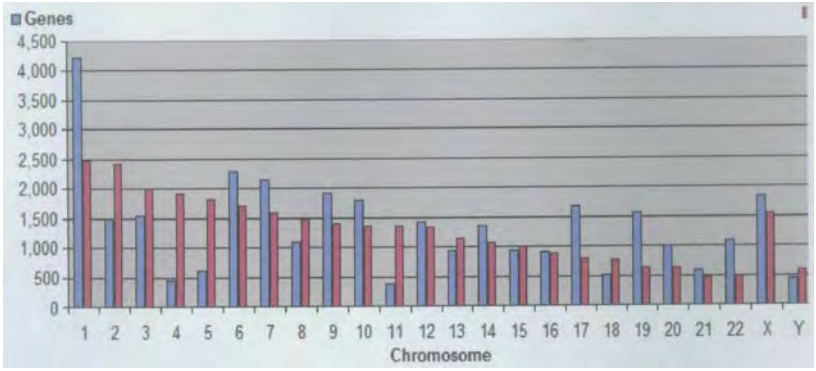
The human chromosome assemblage is shown in this general diagram; the black and white markings are meant only to convey the idea that various genes of differing makeup are present in each chromosome:



In the human gene assemblage there are 46 (23 times 2) chromosomes (long strands of DNA containing all the genetic information [the number varies among species]). Two of these are the sex chromosomes; the remainder are "autosomes" in which each chromosome has its own individual assemblage of genes that controls some aspect of the genetic makeup (other than the sex) of the organism (22 on the haploid strand of a human). The female human has two X chromosomes; the male has a paired X and Y chromosome. Here is a photo of X and Y chromosomes as imaged in an electron microscope:



The number of genes varies from one numbered chromosome to the next (left side of the next diagram). The total number of bases is shown on ther right (unfortunately, the numbers have been cropped off, but each horizontal black line is a jump of 50000 from the preceding line [starting at 0 at the bottom and ending at 450000 at the top]); most but not all the bases have a specific sequence (the coding in the DNA that determines the characteristics of the species). Some of the genes are dedicated to producing proteins instead of species traits.

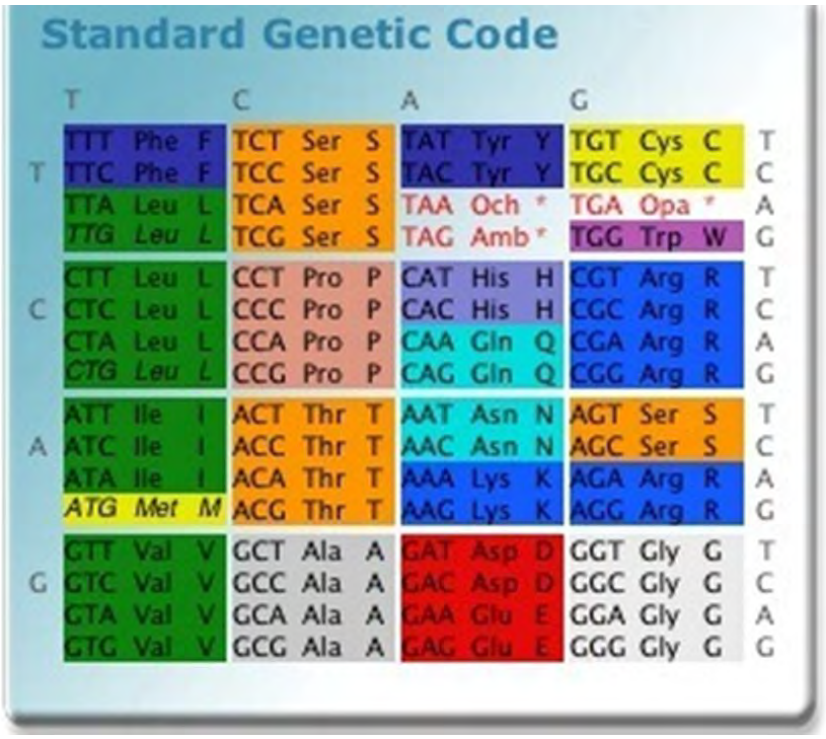


It is interesting to realize that many species share some of the same genes. The sequence may vary. In humans, almost 99% of the genes are the same in any two individuals that can be compared but small differences are distinctive (this fact allows DNA analysis to be used in tracking down an individual suspected of a crime).

The sequence of bases, which defines the genetic code, is divided into triplets called codons. The standard genetic code for RNA consists of 4^3 or 64 combinations. This is the resulting Codon Table:

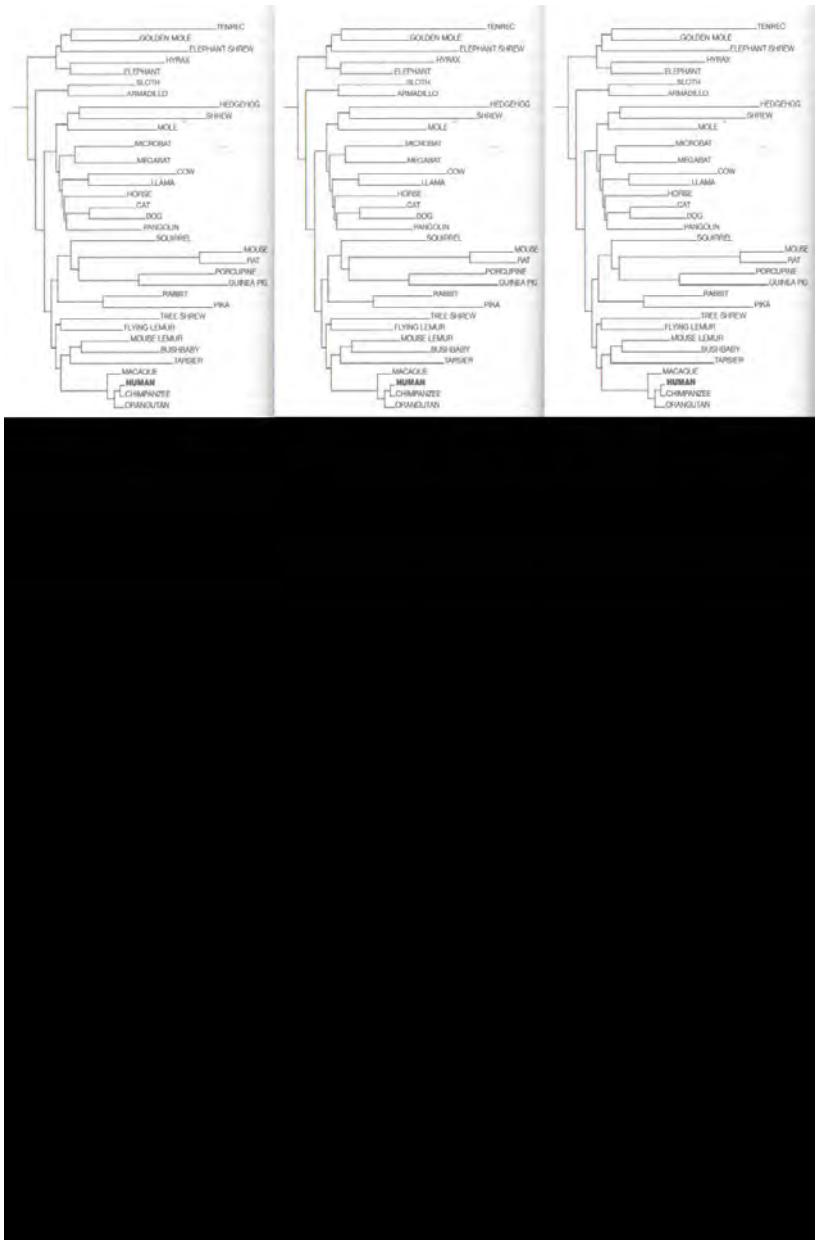
Note how it is organized. The first letter in a Codon is taken from the left side. The second letter is selected according to the Column involved, as indicated in the top row. The third letter is chosen from the right side, in the sequence U, C, A, G. Observe that names have been given to different codons.

The same idea applies to this Codon Table set up for DNA. The difference is that U is now replaced by T:



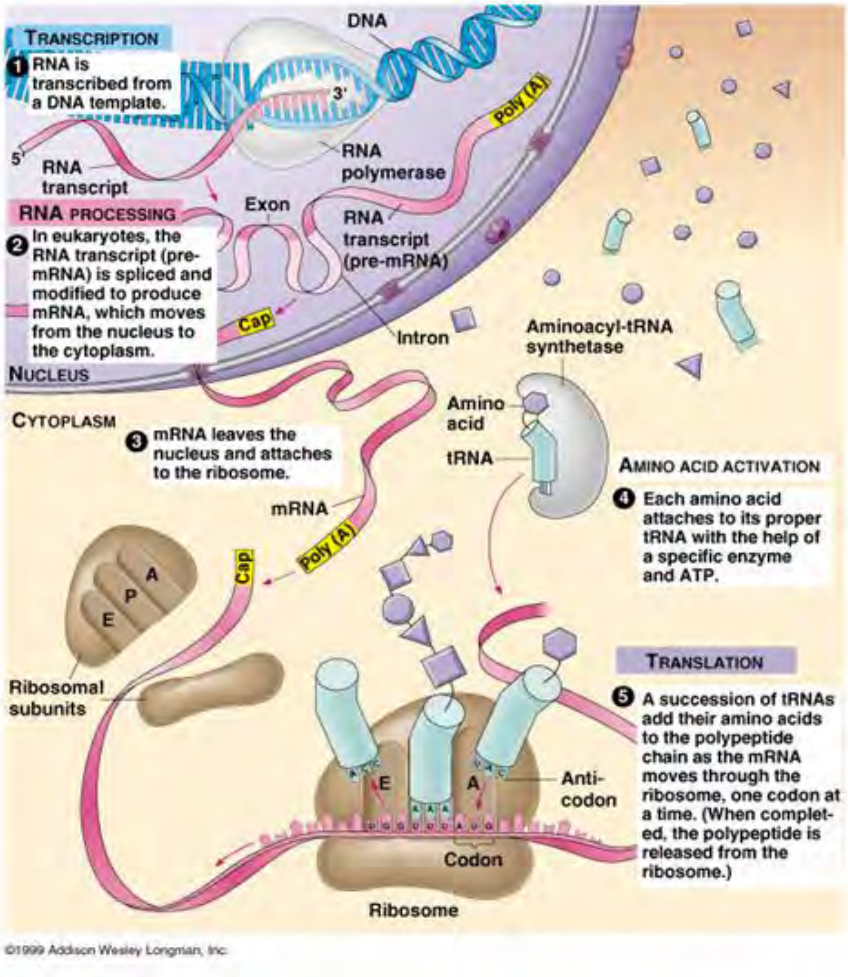
Thus, in a chromosome, the genes consist of various sequential combinations of codons that can number in the thousands

Working out the genomes of individual species and genera has proved a bonanza for reconstructing the patterns of evolution, both within and between species. This cladistic analysis has led to various "trees of life" that present the relationships among the animals or plants being compared. Here is an interesting one found on page 128 in "The Language of God" by Francis S. Collins:

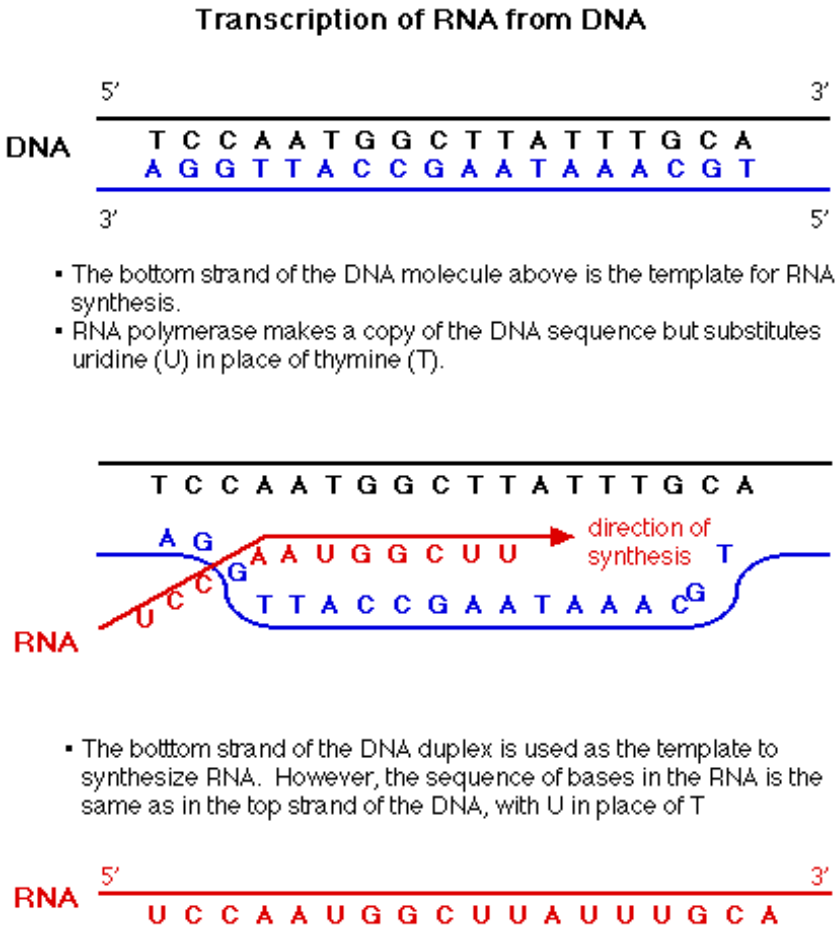


Another important topic that is just too involved to be treated on this review page is Mendelian genetics. Instead, we refer you to this Wikipedia web site. We also have omitted any discussion of cell division and organism reproduction, both sexual and asexual, two important but very involved topics. Again, we steer you to another Wikipedia web site for a general overview, which has links to other relevant terms such as gametes, alleles, mitosis, meiosis. You may also wish to learn these terms: haploid, diploid, genotype and phenotype, zygote, homozygous and heterozygous, blastula and gastrula (most of these can be looked up directly in an Internet Search engine).

Armed with this knowledge of DNA, RNA, and genetics, we are equipped to examine the roles they play in producing or synthesizing new organic molecules. To illustrate this, we will show how the most abundant and versatile molecules - the proteins - are generated. Start with this Overview diagram:

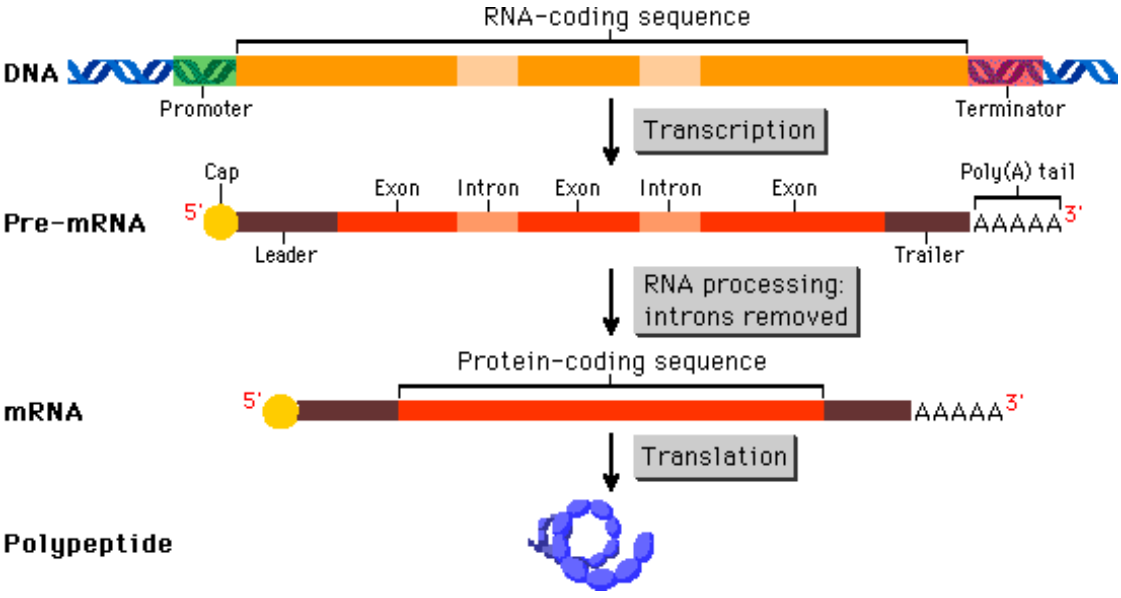


Deceptively simple in this representation, the process is actually complex. In the first step, **Transcription** (TrScrip), the RNA complement in one strand is synthesized by enzymes into a messenger segment, m-RNA, that is a reversed copy of the original DNA strand, but with T (thymine) replaced by U (uracil).

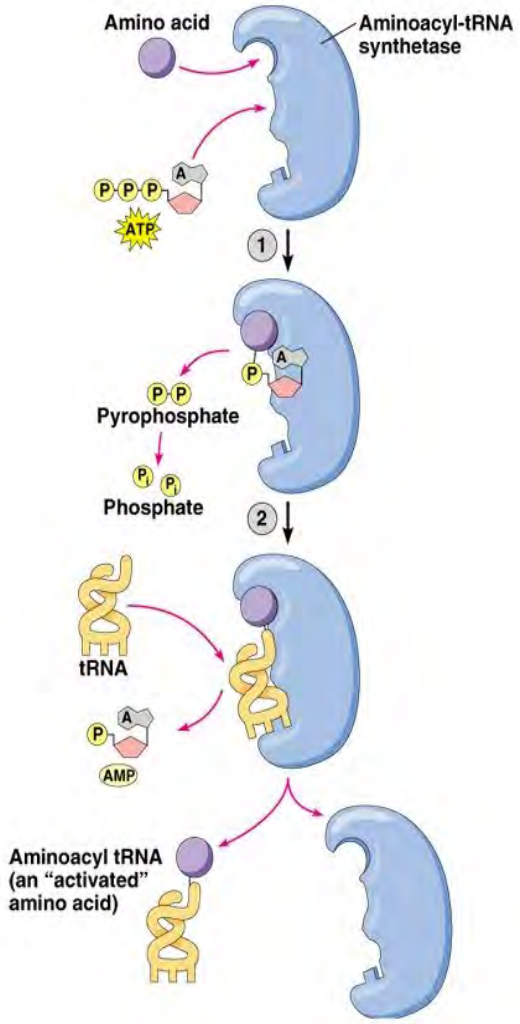


The m-RNA serves as a template for natural protein synthesis within a ribosome (see cell diagram below. which also contains r-RNA (**ribosomal RBA**) This ribosome assemblage participates in the manufacture of a protein by **Translation**) (TrLate) using t-RNA cells (**transfer-RNA** containing various amino acids that bind to the m-RNA by Hydrogen bonding. The process continues as the polypeptide protein builds until addition of a 'chain termination' group signals the particular protein with its diagnostic codification is completed. The specific protein composition depends on the sequences involved in the synthesis. The production of mRNA occurs within the nucleus; the mRNA migrates into the cytoplasm and seeks out ribosomes where the next step - Translation - takes place.

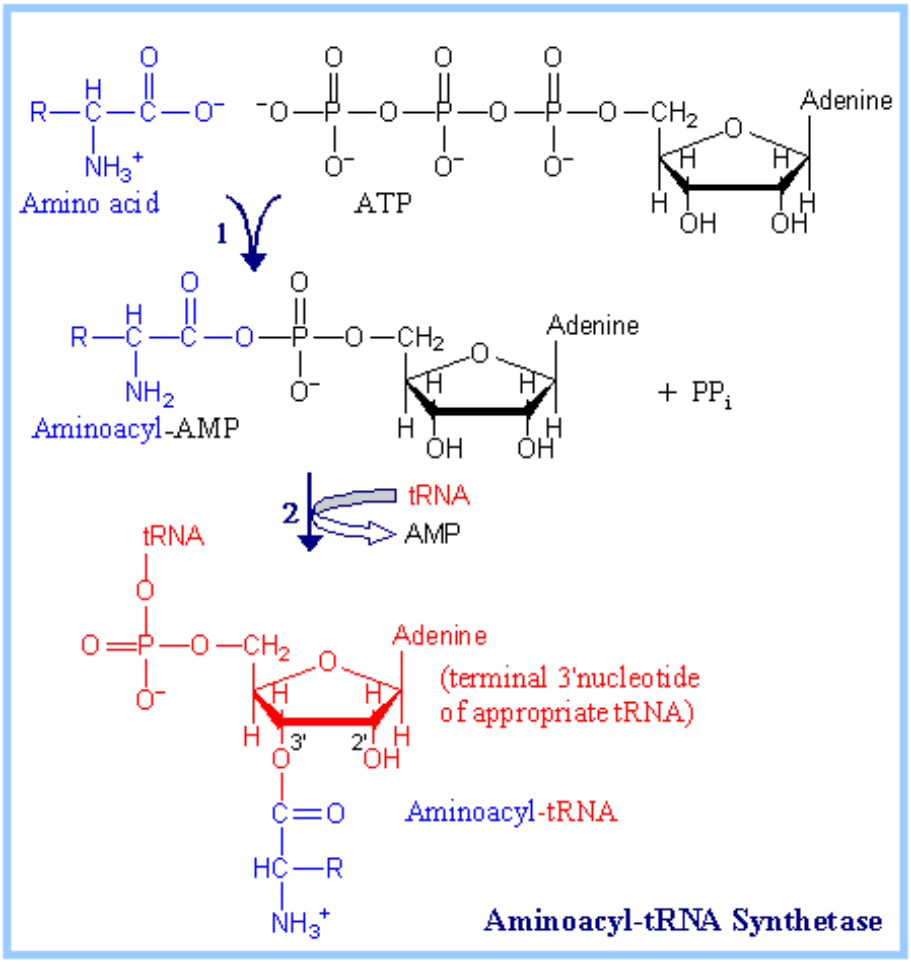
In the Transcription phase, segments, called Introns, of the first resulting RNA are not used in building the mRNA. Only the Exons are spliced together in sequence. Note that there are Promotor (initiator) and Terminator segments.



There is a key intermediate step in Translation. The cell in which the synthesis occurs contains various amino acids. The mRNA must be matched up and bound with its appropriate amino acid. That amino acid is specified by its codon. An enzyme molecule called aminoacyl-tRNA synthetase plays the critical role by serving as a template for preparing a paired tRNA-amino acid that will then be used in the final phase of protein synthesis. Check this figure:

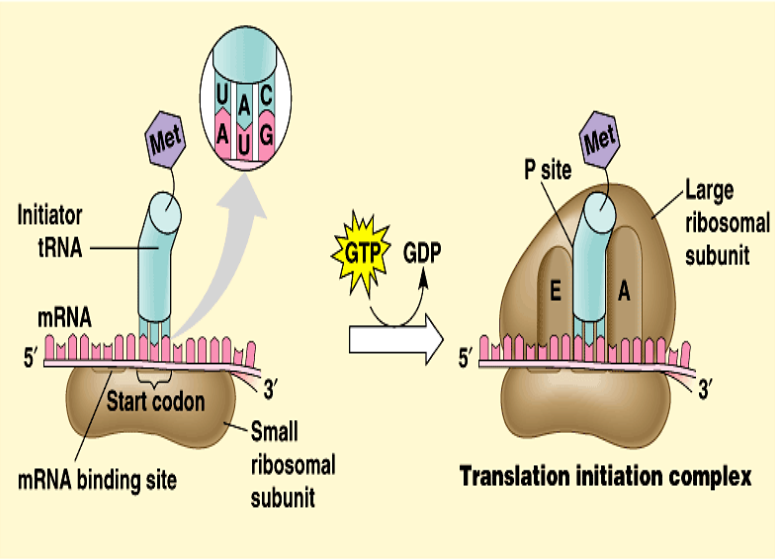


There are two sites on the molecule that are involved. One is so configured that it can accept (bind) only one specific amino acid. The other site accepts the tRNA. The two are brought together through the action of adenosine triphosphate (ATP) as a catalyst. The pair is then released. The same process takes place with other amino acids, each with their site specific aminoacyl-tRNA synthetase molecule. This diagram shows the steps involved in terms of structural formulae:

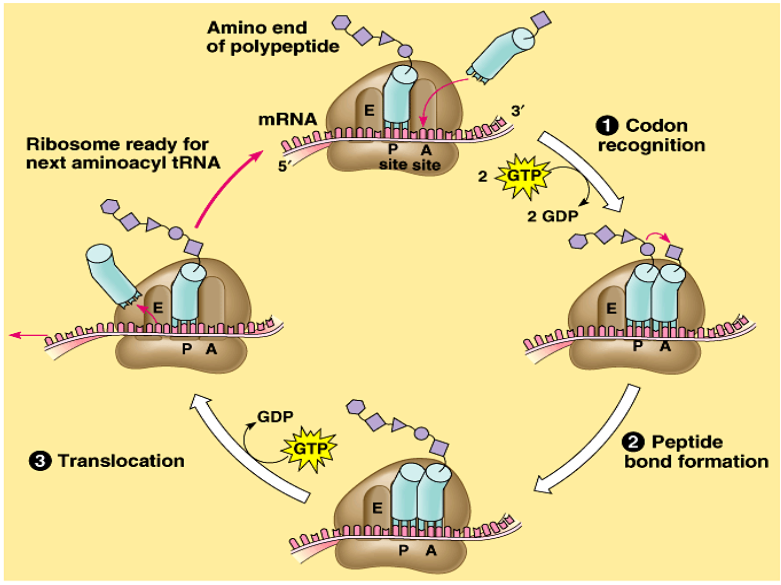


The organic chemistry of the process is complicated and will not be further summarized here. Consult this NIH web site for more details.

The final phase of protein synthesis now enters the picture. This involves the interaction within a ribosome that uses the various tRNA-amino acid pairs to form a polypeptide protein molecule. The pairs are added sequentially according to the master template. The process is involved and will not be discussed here in any detail. Perhaps these two figures will help to visualize what happens.



Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.



Copyright © Pearson Education, Inc., publishing as Benjamin Cummings.

Because of the importance of this topic - protein production - the writer has extracted excerpts from two web sites. The information is bounded by two green lines. Examine these paragraphs if you wish, otherwise skip to the final material below the second green line

---

The process of synthesizing a protein from an mRNA template is known as translation. The mRNA is loaded onto the ribosome and is read three nucleotides at a time by matching each codon to its base pairing anticodon located on a transfer RNA molecule (tRNA), which carries the amino acid corresponding to the codon it recognizes. The amino acid is joined by its carboxyl group to the 3' OH of the tRNA by an ester bond. The enzyme aminoacyl tRNA synthetase "charges" the tRNA molecules with the correct amino acids. The growing polypeptide is often termed the nascent chain. Proteins are always biosynthesized from N-terminus to C-terminus.[11]

Activation of amino acids is carried out by a two step process catalyzed by aminoacyl-tRNA synthetases. Each tRNA, and the amino acid it carries, are recognized by individual aminoacyl-tRNA synthetases. This means there exists at least 20 different aminoacyl-tRNA synthetases, there are actually at least 21 since the initiator met-tRNA of both prokaryotes and eukaryotes is distinct from non-initiator met-tRNAs. Activation of amino acids requires energy in the form of ATP and occurs in a two step reaction catalyzed by the aminoacyl-tRNA synthetases. First the enzyme attaches the amino acid to the a-phosphate of ATP with the concomitant release of pyrophosphate. This is termed an aminoacyl-adenylate intermediate. In the second step the enzyme catalyzes transfer of the amino acid to either the 2'– or 3'–OH of the ribose portion of the 3'-terminal adenosine residue of the tRNA generating the activated aminoacyl-tRNA. Although these reaction are freely reversible, the forward reaction is favored by the coupled hydrolysis of PPi.

Accurate recognition of the correct amino acid as well as the correct tRNA is different for each aminoacyl-tRNA synthetase. Since the different amino acids have different R groups, the enzyme for each amino acid has a different binding pocket for its specific amino acid. It is not the anticodon that determines the tRNA utilized by the synthetases. Although the exact mechanism is not known for all synthetases, it is likely to be a combination of the presence of specific modified bases and the secondary structure of the tRNA that is correctly recognized by the synthetases.

It is absolutely necessary that the discrimination of correct amino acid and correct tRNA be made by a given synthetase prior to release of the aminoacyl-tRNA from the enzyme. Once the product is released there is no further way to proof-read whether a given tRNA is coupled to its corresponding tRNA. Erroneous coupling would lead to the wrong amino acid being incorporated into the polypeptide since the discrimination of amino acid during protein synthesis comes from

the recognition of the anticodon of a tRNA by the codon of the mRNA and not by recognition of the amino acid.

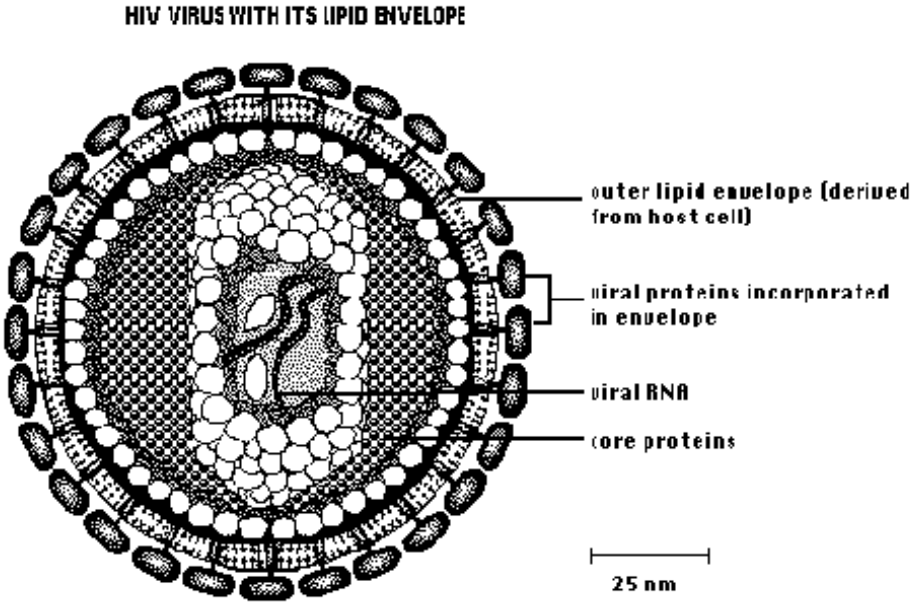
Most microorganisms and plants can biosynthesize all 20 standard amino acids, while animals (including humans) must obtain some of the amino acids from their diet. The amino acids that an organism cannot synthesize on its own are referred to as essential amino acids. Key enzymes that synthesize certain amino acids are not present in animals — such as aspartokinase, which catalyzes the first step in the synthesis of lysine, methionine, and threonine from aspartate. If amino acids are present in the environment, microorganisms can conserve energy by taking up the amino acids from their surroundings and downregulating their biosynthetic pathways.

In animals, amino acids are obtained through the consumption of foods containing protein. Ingested proteins are broken down through digestion, which typically involves denaturation of the protein through exposure to acid and hydrolysis by enzymes called proteases. Some ingested amino acids are used for protein biosynthesis, while others are converted to glucose through gluconeogenesis, or fed into the citric acid cycle. This use of protein as a fuel is particularly important under starvation conditions as it allows the body's own proteins to be used to support life, particularly those found in muscle.

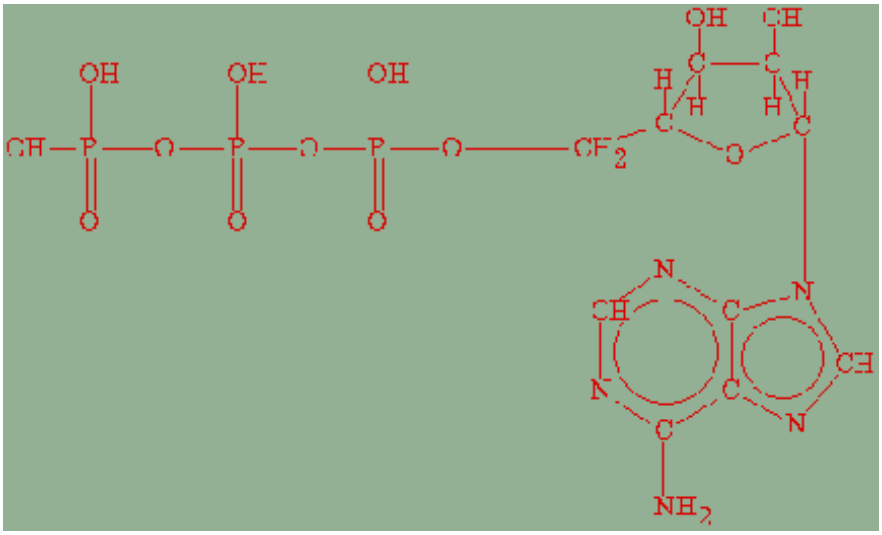
Humans can produce 10 of the 20 amino acids. The others must be supplied in the food. Failure to obtain enough of even 1 of the 10 essential amino acids, those that we cannot make, results in degradation of the body's proteins—muscle and so forth—to obtain the one amino acid that is needed. Unlike fat and starch, the human body does not store excess amino acids for later use —the amino acids must be in the food every day. The 10 amino acids that we can produce are alanine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, proline, serine and tyrosine. Tyrosine is produced from phenylalanine, so if the diet is deficient in phenylalanine, tyrosine will be required as well. The essential amino acids are arginine (required for the young, but not for adults), histidine, isoleucine, leucine, lysine, methionine, phenylalanine, threonine, tryptophan, and valine. These amino acids are required in the diet.

Now we shall finish this page with several special topics:

Mentioned in passing are **viruses**. These are strands or segments of nucleic acids encased in a lipid and/or protein coating. They are not living matter although they can reproduce in cells. Although some viruses are beneficial, most disrupt chromosomes, and can insert themselves into DNA or RNA, causing infections. The best known virus today is HIV, which through AIDS destroys the body's ability to overcome disease and cancer. These belong to the class of organic materials known as **antigens**. In animals, the antigens that threaten their well-being are attacked by **antibodies**, generated by immune response systems (e.g., lymphocytes)

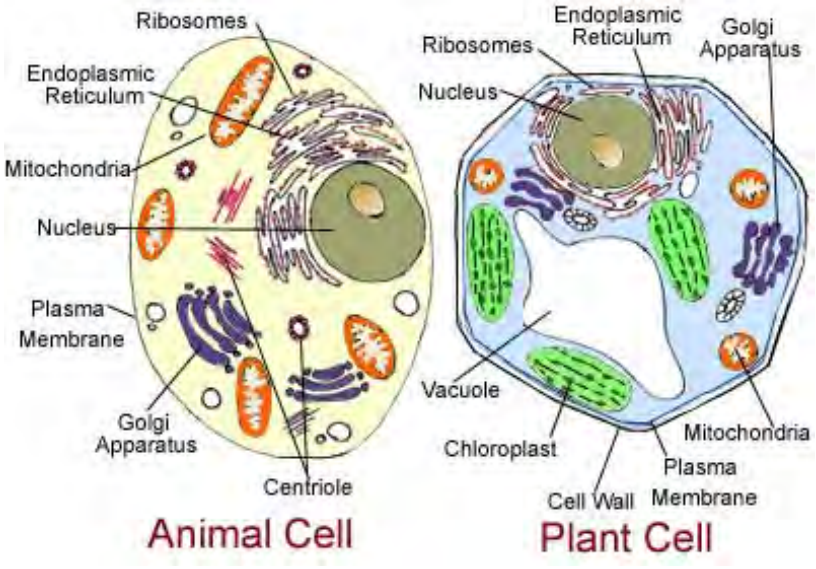


A brief comment about how living organisms derive energy needed to function, usually through **metabolism**, which describes several possible chemical reactions. A general way is any process that releases energy when bonds are broken. In plants, photosynthesis is an endothermic reaction involving interaction of sunlight with chloroplasts. In animals, respiration and fermentation release energy when appropriate organic molecules are catabolized (degraded or broken down). About half that energy is stored in a group of molecules, chief of which is ATP (Adenosine TriPhosphate) synthesized naturally from ribose sugar, an adenine base, and phosphate molecules. Here is its structural formula:



From Farabee/Purves et al - see above citations

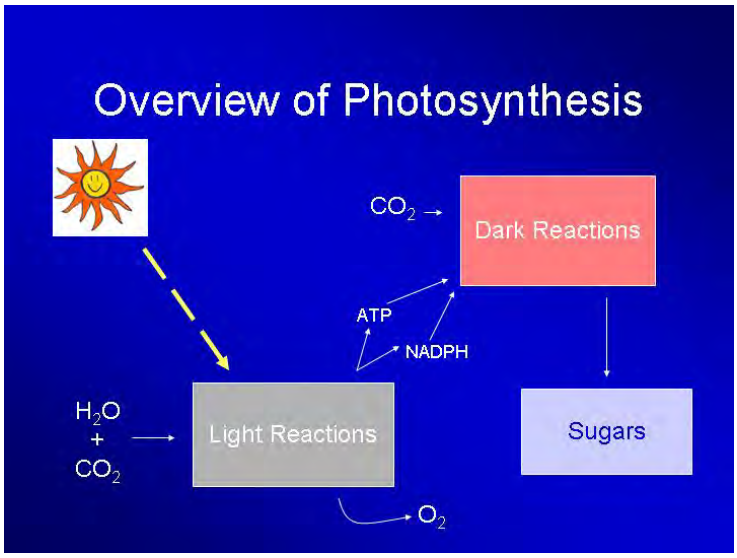
At the beginning of this page we pointed out that there are three fundamental life forms: Bacteria, Plants, and Animals. Bacteria cells are, as was shown, fairly simple. Plant and animal cells are more complex, as summarized by these diagrams:



Plants are very important in that they are the largest source of "food" which provides both the material and the energy needed to create and sustain life. The basic process involves photosynthesis in which the energy sources is (usually) the Sun. Solar photons (hv) energize a reaction between CO_2 and H_2O to yield glucose and Oxygen, according to this formula:

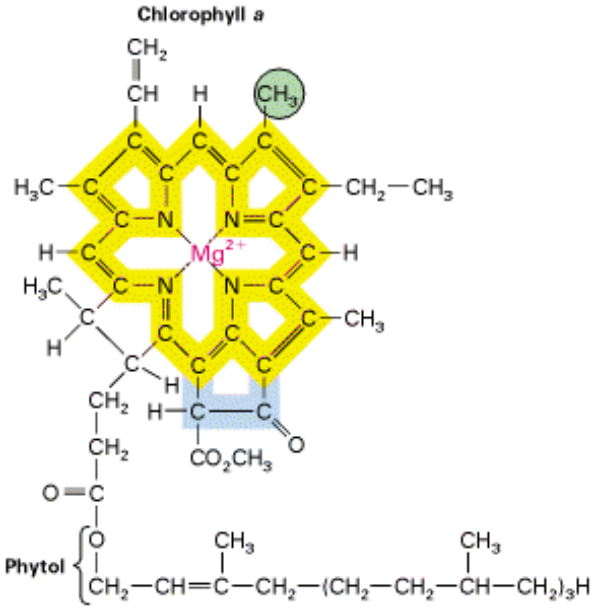
$$\mathbf{6CO_2 + 12\ H_2O + hv \text{ ---> } C_6H_{12}O_6 + 6O_2 + 6\ H_2O}$$

This diagram describes the photosynthesis process:

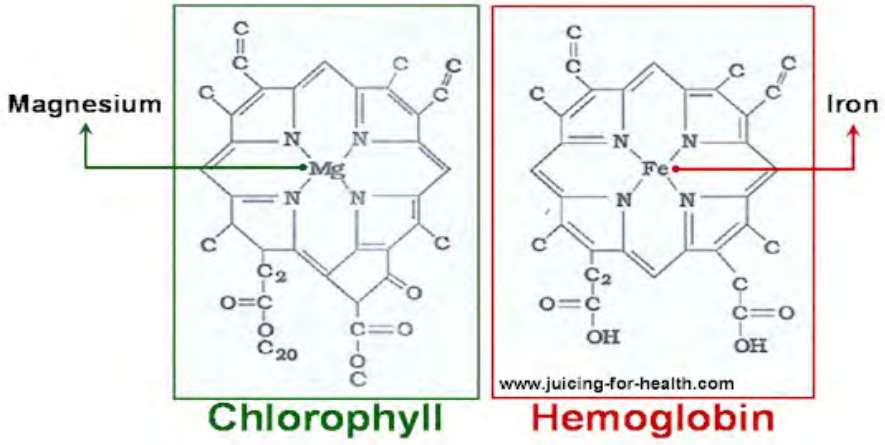


Oxygenic photosynthesis is carried out in two steps, 1) the light reactions which use pigments to capture solar energy and 2) the dark reactions which use the energy from the light reactions to fix atmospherically derived carbon (CO_2) into organic carbon (sugars). It should be noted that the final product is not actually glucose as depicted in the equation above,we just don't usually discuss the last step when we talk about photosynthesis. Glucose does result from photosynthesis, but it usually present in polymeric form either as sucrose (a dimer) or starch (a polymer).

The organic molecule Chlorophyll is the essential agent in plant photosynthesis. It consists of a porphyrin ring and a hydrophobic phytol tail, as shown in this structural formula.



The process is centered within the chloroplasts. It is interesting to note the structural similarities between chlorophyll and hemoglobin, a molecule found in higher order animals.



Hemoglobin is involved in one of the most vital processes in the human body (and in other higher animals) - respiration. Oxygen from the air attaches to hemoglobin carried in the blood and reacts with various molecules to catabolically break them down, releasing energy. Aerobic oxidation is expressed in this formula:

$$C_6H_{12}O_6\ (aq) + 6\ O_2\ (g) \rightarrow 6\ CO_2\ (g) + 6\ H_2O\ (l)$$

Some animals and plants can obtain energy without the use of Oxygen. This is an anaerobic process.

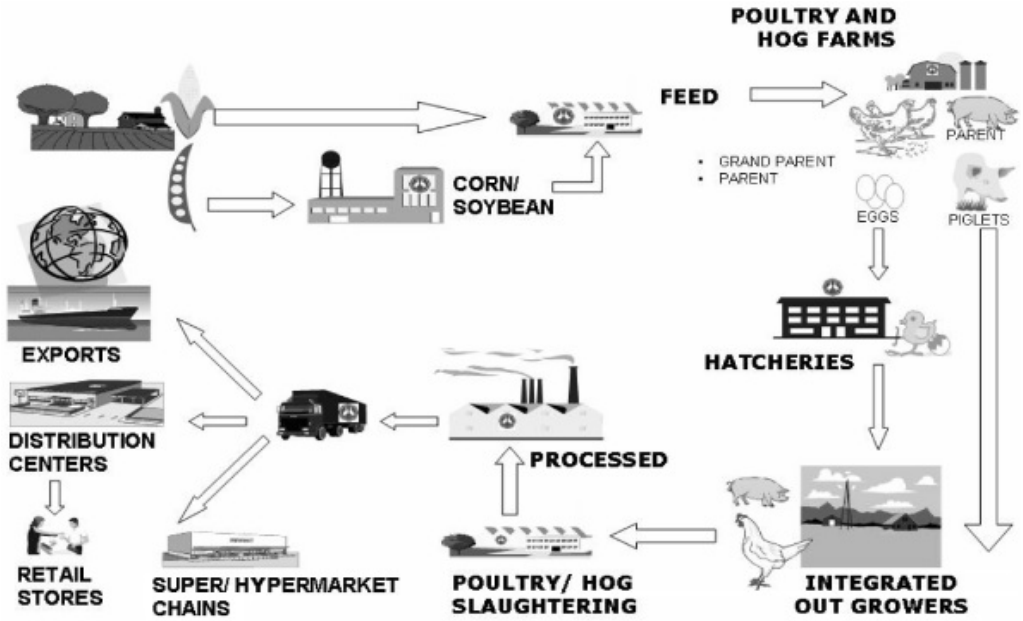
Plants and animals interact with each other as sources of life-sustaining **food**. Plants form the base of the food chain. Next are the herbivores - animals that consume only plants. At the top are the omnivores - animals that eat both plants and other animals. Food chains can also be described in terms of producers and consumers, predators and prey, or trophic levels. The sequence of consumption will vary with the different ecosystems involved. Here is one example:

Sample Food Chains

Trophic Level	Grassland Biome	Pond Biome	Ocean Biome
Primary Producer	grass	algae	phytoplankton
Primary Consumer	grasshopper	mosquito larva	zooplankton
Secondary Consumer	rat	dragonfly larva	fish
Tertiary Consumer	snake	fish	seal
Quaternary Consumer	hawk	raccoon	white shark

©EnchantedLearning.com

Twenty-first century humankind has a sophisticated food chain system in terms of its infrastructure. Consider this example which starts with a farm (upper left) and shows food production as illustrated by chickens (it could be cows or pigs or crops) as it moves to its final destination - people - through modern food distribution centers.



Now, armed with this general background in Biology/Biochemistry, we will proceed to the next page, which covers the origin and history of life (primarily on Earth) and the tenets of Evolution.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

ASTRONOMY AND COSMOLOGY:

The Description, Origin, and Development of the Universe

DISCLAIMER

At the outset of this Section the writer (N.M. Short) wishes to let the reader know that I am not an astronomer or cosmologist by profession. But as a geologist I do have enough background in Science to have taught myself nearly everything you will encounter in this Section. There are probably errors or misconceptions - if you discover any, I'd appreciate your contacting me through my email.

Preparing this Section has been a revelation - almost a life-changing experience. It has given me a profound philosophic comprehension of the most intimate nature of existence itself. I am now constantly revising my understanding of what all beings, life, and myself are all about. Night after night, as I prepare for sleep, I keep asking myself two deep questions: How do we fit into this vast Universe, and why is it there at all? As I approach my end (at this writing 82 and in poor health), I am discovering renewed hope that answers will be found in an afterlife. For those who read through the Section, I trust you will re-examine your outlook based on realizing your place in the grand scheme of things.

More than 10 years after I began this Section, I purchased a Cosmology DVD course available from The Teaching Company, which specializes in providing college level courses covering almost all areas of knowledge. The course is conducted by Dr. Mark Whittle, a Professor at the University of Virginia. It is an extraordinary offering which thoroughly and clearly surveys the entire field. It is my hope to be able to incorporate ideas and material from this course into this Section, thereby revising it by modifying and expanding the content.

Note: 1) Most of the pages in this Section are image-intensive, so that the large number of illustrations can lead to a lengthy download time for those using modems connected to telephone lines; 2) Some parts or ideas presented in this Section may seem repetitious, i.e., are stated more than once; some of this reiteration is deliberate - much of the topics covered tend to be complex and unfamiliar to the non-specialist reader (those who are not astronomers, cosmologists, physicists), so that repeating is a helpful aid in reminding one of these previously developed ideas and tying them (making them relevant) to the other subjects where they later appear.

There will be no individual page summaries in Section 20 which deals with Cosmology: The Origin, Composition, Structure, Development, and History of the Universe (or Universes, if there are more than one). This is largely because of the complexity and wide range of ideas on each page: this does not lend itself easily to synopsize. The reader instead should work through the knowledge imparted on each page without the aid of a preview or reduction to a simplified digest. If the field is new to you, several readings of this Section may be needed to facilitate mastery of this ultimate subject: the Origin of Everything. Also, if a novice, you should profit from working

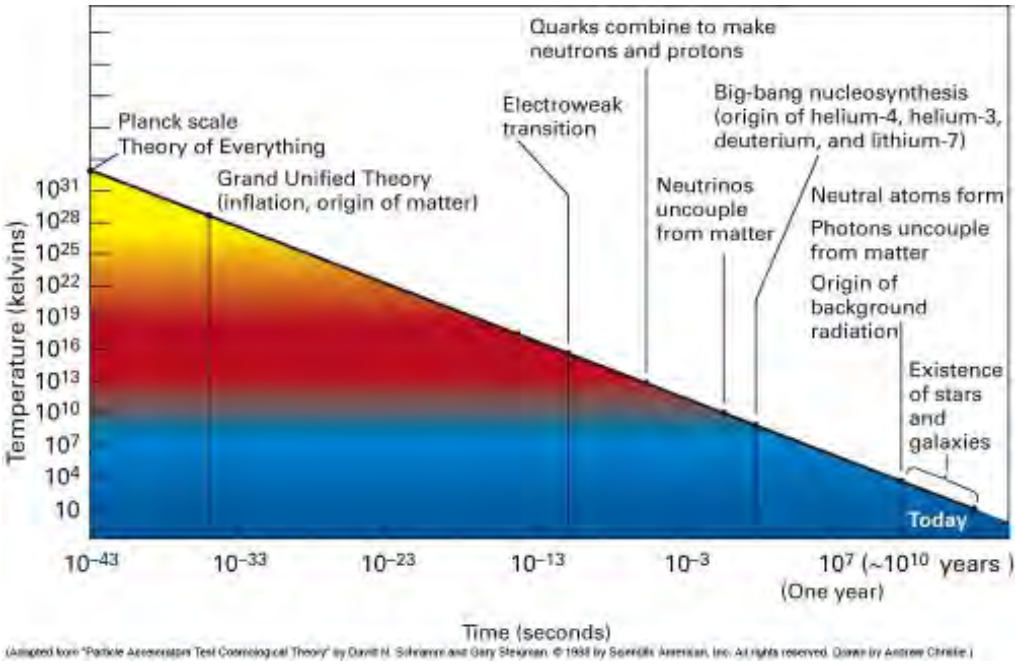
through the excellent online "textbook" in Astronomy prepared by Dr. J. Schombert at the University of Oregon, which have been referenced in the Preface . In keeping with the Overview and the 20 Sections that have followed, every illustration will be accompanied by a synoptic caption.Despite this absence of page summaries, we will attempt to abridge the overall ideas underlying Astronomy and Cosmology in this ***synopsis*** *shown in green:*

ASTRONOMY deals mainly with the description of the objects, materials, structure, and distribution of everything that appears to exist beyond the Earth itself. COSMOLOGY deals with the origin, development, and future expectations of/for the Universe Thus, Astronomy is mostly concerned with the "what" and "where" whereas Cosmology is more concerned on the "how" and "why". Both focus on the subjects of stars, galaxies, interstellar/intergalactic gases, and the void called "space", but from differing perspectives.

Astronomy as an observing "science" traces its roots to early civilizations such as the pre-Christian era Babylonians, Egyptians, Greeks, and Chinese and the Mayans and Aztecs in the New World. Star groupings, the constellations, were established and became involved in myths that suggested deity controls of how the World (i.e., the Universe) is able to function. ,.

Ideas of an Earth-centered Universe began in early times, with both myths and theological explanations for the meaning and cause(s) of the physical (natural) World (including and beyond the Earth) gradually being supplanted by scientifically-based observations. Key ideas that provide this basis include the postulates by such Greek philosophers as Pythagoras, Euxodus, and Aristotle (the latter proposed the Earth as the center of the Universe, with the Moon, Sun, and stars being embedded in crystal spheres that rotated around Earth) and the later (ca. 140 BCE) Ptolemaic description of epicyclic "heavenly" motions; these persisted largely as philosophical musings until the advent of Copernicus in the 16th Century CE who posited the heliocentric theory for the Solar System (but that had been suggested - and discounted - by Aristarchus in 280 BCE), followed by important contributions from Tycho Brahe (detailed measurements of motions of celestial bodies) and Johannes Kepler (Laws of planetary movements) soon thereafter. Galileo was the first to use the telecope for astronomical observations; his observations confirmed Copernicus's revolutionary idea that the Earth was not the center of the Universe nor the Solar System. Isaac Newton provided the foundation for the movements of stars and planets with his Laws of Gravity and Motion. William Herschel in the late 1700s CE provided the first proof that the Milky Way, in which the Sun is located. is an "Island Universe", namely a huge cluster of stars comprising a galaxy; Herschel surmised that other such galaxies must exist. This led to the beginning of the modern era of Cosmology stemming for work by Edwin Hubble and others in the 1920s, who showed that there were many galaxies beyond the Milky Way and some of these were located at huge distances from Earth. Hubble also confirmed the idea of expansion from a Big Bang that had been put forth by the Abbe Lemaitre in the early 1920s. Of great importance to cosmologists was the new framework for a proper understanding of the Universe and of the laws of physics that affect it that was established by Albert Einstein's theories of Special and General Relativity (see Preface).

Before the beginning of this (there may be more than one) finite Universe there was (at least for those within it) no time nor space, no energy (in the discrete forms we know) nor matter - at least in the sense that we perceive these fundamental "qualities". What may have existed is some as yet undefined quantum state in an endless void in which fluctuations in the "emptiness" - the void - led to extremely fleeting "particles" containing the essence to grow into a Universe. Essentially all such evanescent moments ended with the disappearance of these particles (still unidentified entities: energetic virtual (potential non-yet existing particles which may be those constituting the Dark energy that dominates the Universe [see this page and Pages 20-8 and 20-9] have been proposed). But the potential was there for one such particle to spring to existence at a moment (the singularity) that witnessed "creation of space and matter" from when our Universe sprang. This diagram may help as you work through the next three paragraphs (variations will appear later on this page):

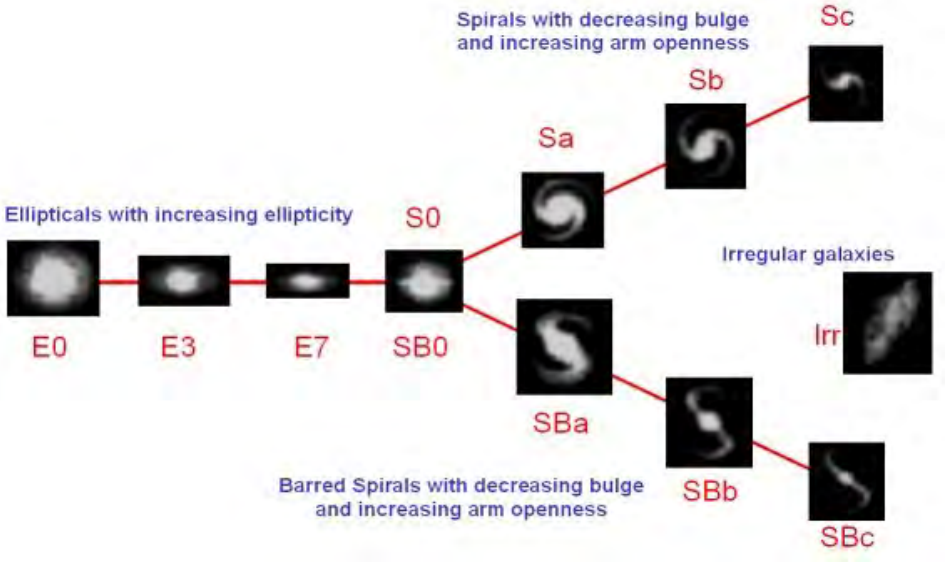


The singularity particle was so unstable that it "exploded" into what is known colloquially as the "Big Bang". That took place some 13.7 billion (+/- 200 million) years ago (see page 20-9 for a discussion as to how this value has been reached and has changed several times in recent years). What actually happened is not a true explosion but a process that is described as the creation of "space" which has been ***expanding*** *ever since. The first minute of Universe time was the critical stage leading to the state of the Universe we observe today. We can trace theoretically events during the minute back to 10^{-43} sec(onds) - an instant known as the Planck time - when the Universe was infinitesimally small. (Experimentally, astrophysicists can actually reconstruct the sequence and verify the essential physics of the Universe's early conditions back to 10^{-12} seconds and to particle sizes as small as 10^{-17} meters; better yet, a significant number (most?) of the particles and forces [and fields through which they interact] have now been defined and all but a few actually found and identified under laboratory conditions.) Initially, the fundamental forces (strong; weak; electromagnetic; gravity) were unified (as may be explained through one new theory in physics called "superstrings"). But, they quickly separated systematically into the individual four prime forces. Although expansion was rapid, at about 10^{-35} seconds, there was a one-time only extreme acceleration of this minute Universe through a process called Inflation. Inflation may be responsible for the likelihood that the Universe is much larger than the 13.7 billion light year limit of the observable Universe imposed by the speed of light.*

Thereafter, in this first minute as expansion continued and the proto-Universe cooled to lower energy levels, the fermions (matter), controlled by the appropriate bosons (force), began to organize into the protons and neutrons (both composed of quarks), electrons, mesons, neutrinos, and others of the myriads of particles continually being discovered in high energy accelerator experiment in physics labs.

As the first minute ended, some particles began to associate with others (while probably all the anti-matter that should have been created was destroyed). In the first few minutes, particles began to organize into nuclei that were part of a plasma state in which the mix included electrons, photons, neutrinos and others. In the next 380,000 years or so, this particle-radiation state witnessed the beginnings of organization into atoms, mostly of Hydrogen and some helium. After that time the Universe became "transparent" so that communication through photon (light) radiation was possible between segments of the Universe close enough to exchange information at the speed of light. The Universe was almost completely homogeneous and isotropic on a grand scale but locally tiny fluctuations in the state of matter (mostly H and He), as appear in the irregularities in the Cosmic Background Radiation (greatly cooled Big Bang "afterglow' that pervades the Universe and marks its observable edge), led to gravitational clumping (into nebulas) that grew simply because these slight increases in density continued to increase the organization through the force of gravitational attraction. From this eventually, in the first billion

years, stars began to form and to arrange in clusters called galaxies. These adopt specific shapes, such as spiral, elliptical, or irregular. This diagram, a classification of forms first put forth by Edwin Hubble, shows the range of shapes (note: it is not an evolutionary chain):



A star is defined as a massive, spherical astronomical body that is undergoing (or has undergone) burning of nuclear fuels (initially Hydrogen and, if hot enough, Helium); as it evolves elements of greater atomic number are consumed as well) so as to release energy in large amounts of both luminous and non-luminous radiation (over a wide range of the EM spectrum); stars eventually change significantly in mass, size, and luminous output with some finally surviving only as very dense cores (neutron stars) of minimal luminosity. Stars burn their Hydrogen at high temperatures, during which (depending on their size) they convert this fuel to heavier elements (largest ones can produce elements up to iron in the Periodic Table). Large stars die out rapidly (a few hundred thousand to one or more billion years); small stars can persist for times that are comparable to the total life of the Universe. During their stable lifetimes, the stars hold together by a fine balance between inward contraction under gravity, involving internal heating up, and the outward pressure of the radiation produced by nuclear processes. Many stars can explode as supernovae. Various types of stars evolve over time through distinct pathways; among these are Red Giants; White Dwarfs; Brown Dwarfs, and Neutron Stars. Black Holes are another, perhaps widespread, constituent of space. As a star forms out of nebular material - gases mainly of some Hydrogen and helium, and other elements in various forms, including particulate dust), some of this material not drawn into the growing star may collect in clots that would form planetary bodies - rocks and gas balls - similar to those making up our Solar System.

The composition of the Universe has only recently been determined fairly precisely. Ordinary matter, making up the stars, galaxies, gas/dust clouds and a small fraction of the so-called empty space, accounts for about 4% (most of that is Hydrogen and some Helium). The rest is present in Dark Matter (undetected directly by any technique so far; includes the WIMPS and MACHOs discussed on page 20-9) which seemingly increases around galaxies, and makes up about 23% of a Universe and Dark Energy (about 73%), tied to a still mysterious force that seems to act like the anti-gravity force first postulated by Einstein,which he called the Cosmological Constant, and is the prime candidate for causing the recent observation that the Universe now is once again expanding after slowing down for the first seven or so billion years.

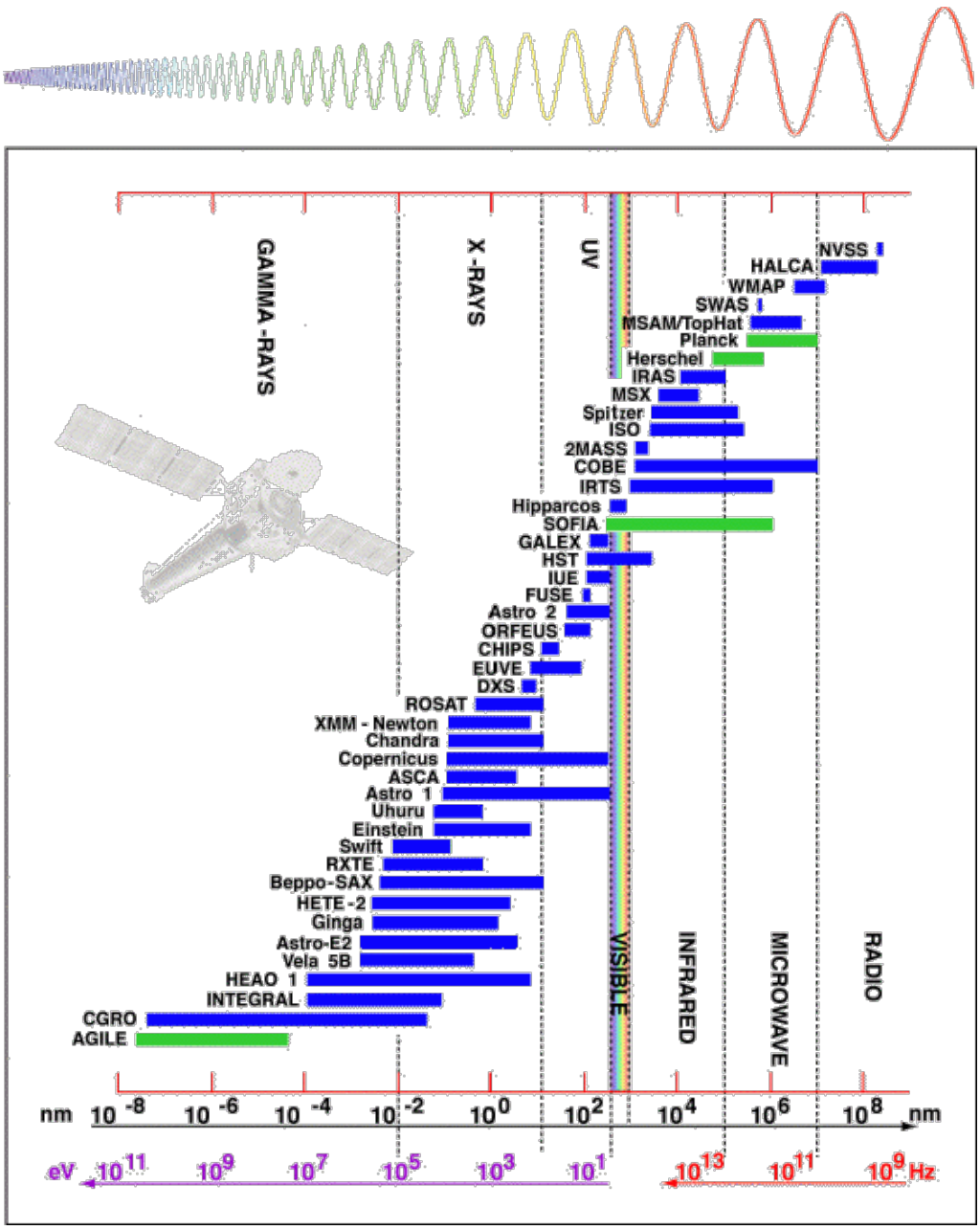
The fate of the Universe depends ultimately on how much mass (and its convertible form, energy) it has. If that number is high the Universe's expansion may slow down and eventually reverse (contract) so that all matter and energy collect again at a superdense point which may undergo another Big Bang. Or the matter/energy is insufficient to slow expansion and the Universe enlarges forever. The shape of the Universe will depend on the nature of the expansion; recent evidence indicates that it may be "flat". At large scales the Universe is subject to the laws of Relativity (but equally as important is the role of matter/energy at the smallest - quantum - scales). Recent information favors endless expansion and the possibility that the rate of expansion is now

increasing.

Add to all of this the theoretical (quantum-driven) possibility that there may be multiple universes, unable to communicate with one another, with new ones forming at various times and perhaps old ones dying in some way. The mind boggles at this point. But even more amazing is the realization that there is something we humans recognize as "mind" - our most valuable property and objectively the most powerful entity so far discovered in the Universe. Our minds have identified the ways in which planets form, including those suited to hosting living creatures, and the very nature of life itself.

Humankind has in the last 400 years, and especially the last 50 years, developed the skills and the will to explore our Universe. We now obtain data of great explanatory/interpretive value using telescopes that gather in radiation from all parts of the EM spectrum. Thus, there are now specialized observing systems that sample in the gamma-ray, x-ray, ultraviolet, visible, near and far infrared, and radio wavelength regions of the spectrum. Astronomy is probably the prime user of nearly all segments of that spectrum, as it gathers its information almost exclusively by remote sensing methods.

The big advance during the last 50 years has been to place astronomical telescopes into space, in orbits above the Earth's atmosphere. The most famed of these is the Hubble Space Telescope (HST) which has dazzled astronomers, other scientists, and the world public with its abundance of extraordinary images. There have been other great space observatories; we mention most of the best known in this Overview: the Compton Gamma-Ray Observatory (CGRO); the Chandra X-Ray Telescope; XMM-Newton; Extreme Ultraviolet Explorer (EUVE); the International Ultraviolet Explorer (IUE); the Far Ultraviolet Explorer (FUSE); the Galaxy Evolution Explorer (GALEX); the Infrared Astronomical Satellite (IRAS); the Infrared Space Observatory (ISO); the Space InfraRed Telescope Facity (SIRTF; renamed the Spitzer Space Telescope. Only two radio telescopes have yet been orbited but plans are underway for more. A pictorial overview of the major space observatories is presented in this illustration:



Some of these observatories are considered in this Section; others are not. If curious about the latter, check any out through Google or Yahoo. Note: those with green bars were scheduled for launch after this chart was prepared. Links to these and other (unnamed here) astronomical observatories are given on this OAOS web site.

THIS IS A GOOD PLACE, IN THIS SYNOPSIS, TO MENTION SOME OF THE MAJOR INDIVIDUALS OF THE 20TH CENTURY AND THEIR CONTRIBUTIONS TO COSMOLOGY (MORE DETAILS WILL FOLLOW ON THIS AND LATER PAGES):

***MAX PLANCK**: THE ORIGINATOR OF SOME OF THE IDEAS THAT LED TO QUANTUM PHYSICS.*

***VESTO SLIPHER**: DISCOVERED THE "RED SHIFT" OF STELLAR SPECTRA, INDICATING GALAXIES WERE MOVING AWAY FROM THE EARTH AS AN OBSERVING PLATFORM.*

***ALBERT EINSTEIN**: THE "GIANT INTELLECT" WHOSE CONCEPTS OF RELATIVITY CHANGED PHYSICS AND RECAST OUR UNDERSTANDING OF THE UNIVERSE; STATED THE NOTION OF SPACETIME AND DEVISED NEW CONCEPT OF GRAVITY; HE BELIEVED IN A STEADY STATE UNIVERSE.*

***WILLEM DE SITTER**:FROM HIS SOLUTION OF EINSTEIN'S GENERAL THEORY OF RELATIVITY EQUATIONS, CONCLUDED THE UNIVERSE WAS EXPANDING.*

***ALEXANDER FRIEDMANN**: THE RUSSIAN MATHEMATICIANT WHO CONFIRMEDED THE POSSIBILITY OF AN EXPANDING, FINITE UNIVERSE.*

***GEORGES LEMAITRE**: THE BELGIAN PRIEST WHO CONCEIVED OF THE UNIVERSE'S EXPANSION FROM A VERY SMALL VOLUME (THE PRIMORIDAL ATOM; SIZE ROUGHLY THAT OF THE SOLAR SYSTEM) THAT "EXPLODED" AT THE BEGINNING OF TIME (SINGULARITY)".*

***EDWIN HUBBLE**: THE ASTRONOMER WHO DISCOVERED GALAXIES BEYOND THE MILKY WAY AND PRESENTED EVIDENCE FOR EXPANSION.*

***GEORGE GAMOW**: THE PHYSICIST WHO EXPLAINED HOW STARS FORM AND BURN THEIR HYDROGEN FUEL.*

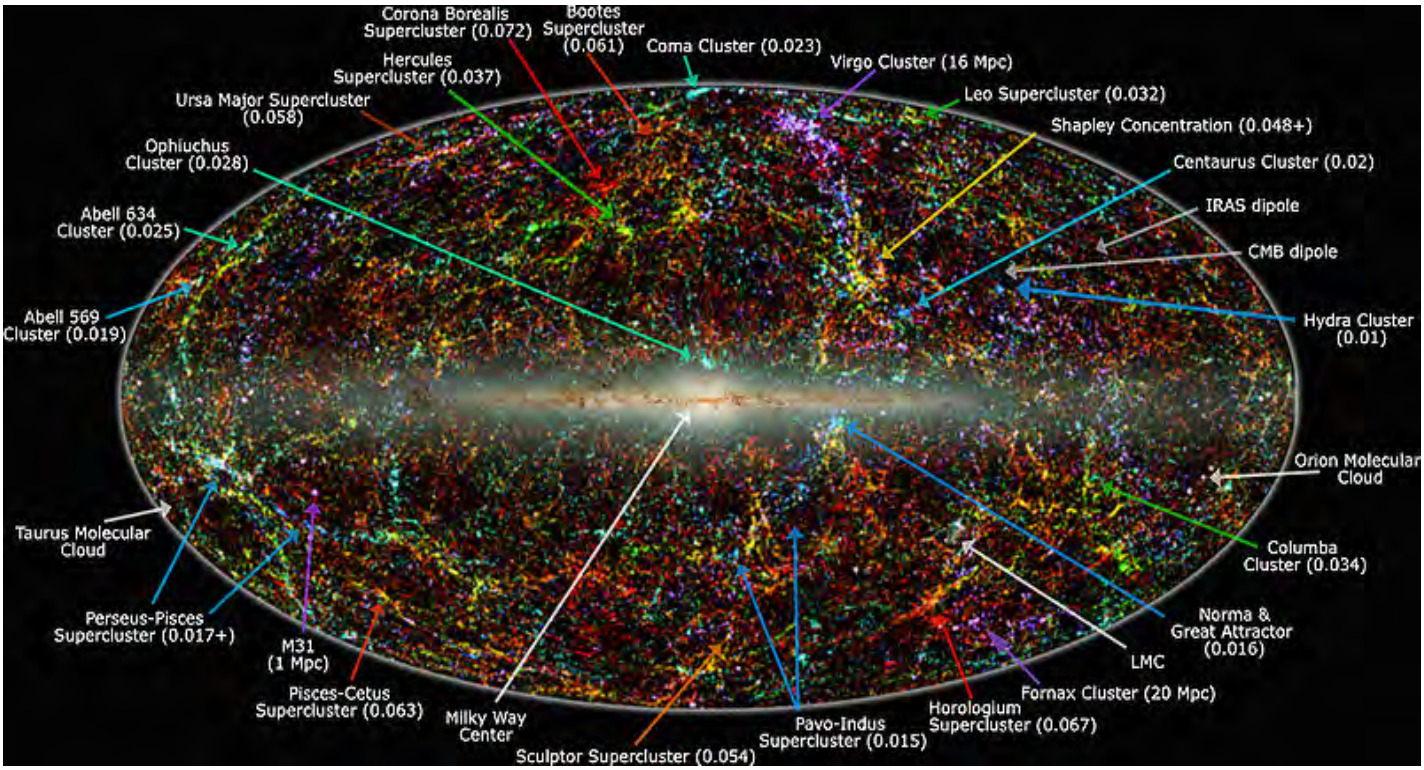
***FRED HOYLE**: THE ASTRONOMER WHO EXPLAINED HOW ELEMENTS HEAVIER THAN HELIUM ARE PRODUCED BY FUSION IN STARS; HE ALSO COINED THE TERM "BIG BANG" (AS A DERISION OF EXPANSION CONCEPTS) AND CHAMPIONED A STEADY STATE UNIVERSE.*

***BRANDON CARTER**: THE PHYSICIST WHO PROPOSED THE MODERN CONCEPT OF "THE ANTHROPIC PRINCIPLE" - THE UNIVERSE HAS JUST THE RIGHT SET OF PROPERTIES TO ALLOW LIFE TO DEVELOP WITHIN IT AT SOME EVOLUTIONARY STAGE (AND REQUIRES INTELLIGENT LIFE TO REALIZE THAT IT EXISTS).*

***ALAN GUTH**: THE COSMOLOGIST WHO PROPOSED THE IDEA OF THE BRIEF SUPEREXPANSION KNOWN AS INFLATION THAT BEST ACCOUNTS FOR THE UNIVERSE'S SIZE AND PROPERTIES.*

***ARNO PENSIAS** AND **ROBERT WILSON**: THE TWO ENGINEERS WHO FORTUITOUSLY DISCOVERED THE COSMIC BACKGROUND RADIATION (BEING SOUGHT THEN BY **ROBERT DICKE**'S GROUP AT PRINCETON UNIVERSITY); PREDICTED BY EINSTEIN, THIS RADIATION IS STRONG EVIDENCE FOR THE BIG BANG AND IT HELPS TO ESTABLISH THE "TRUE" AGE OF THE UNIVERSE (IN TERMS OF EARTH YEARS).*

We close this Overview summary with a map that shows the main features of the Observed Universe; other kinds of maps are possible and several will appear later in the Section.



The first page consists of:

A GENERAL OVERVIEW OF THE COSMOS:

With regards to:

The Big Bang; The Overall Characteristics and Structure of the Universe; The First Few Minutes; The Nature and Origin of Matter; The Early Eras; The Subsequent History of Expansion

Before we enter this long page, the writer (NMS) would like to present his own definition of a word - perhaps setting a precedent. Many readers are familiar with the famed astronomer Carl Sagan's TV series on the "Cosmos". But it is not easy to find a good working definition of that term that all agree on. Typing in the word on Google led to many entries not related to astronomy. The Wikipedia entry gave this information (reproduced here as two italicized parts extracted from that Website):

*In physical cosmology, the term **cosmos** is often used in a technical way, referring to a particular space-time continuum within the (postulated) multiverse. Our particular cosmos is generally capitalized as the Cosmos. The philosopher Ken Wilber uses the term "kosmos" to refer to all of manifest existence, including various realms of consciousness.*

Using this for support, we will define this use of "Cosmos" in this way: Everything that can be conceived to exist in a real and physical way which includes all that is within our Universe, any other Universe (Multiverse concept) and any of the vacuum containing virtual particles that fills the (non)space between the (possibly infinite number of) multiverses. Thus, when we refer specifically to the "Universe", we will mean this (our) Universe; when we refer to the Cosmos, we are taking into account the speculation by scientists, and by science fiction writers, that there can be more than one Universe, even

though at present there is no way to prove the actual existence of anything beyond the limits of our observations through telescopes. So far then, the notion of any physical reality beyond these limits is conceptual rather than factual.

While we are redefining terms, lets consider "space" itself. The Space Program in its broadest connotation refers to monitoring activities connected with Earth, to the exploration of the Solar System, and to the observations of the farthest reaches of the Cosmos (at least to the detectable photons making up the Cosmic Background Radiation and to the first stars [or more properly the first galaxies since individual stars at great distances cannot be resolved]). Space in a "geometric" sense is easily visualized as that which includes all detectable cosmic objects (stars, galaxies, gas clouds, etc.) and everything between any two such objects. The "between" however is not empty. Even if nothing can be detected by present means, quantum theory holds that such space contains vacuum energy and even virtual particles which can "pop" in and out of active existence. In this sense, space is that which resides within the Cosmos regardless of whether there are any markers that establish the "between". It would not be surprising if such 'space' is eventually shown to be infinite.

For now, all we can talk about from direct contact is the Observable Universe. The subject of the Observable Universe will be explored several times in this Section. For those anxious for a preview, check out this Wikipedia website.

Introduction

Before beginning this Section, we urge you to read through a page called the Preface (once there, hit your BACK button on the browser you use to return to this page). The Preface contains four major topics: 1) the role of remote sensing in astronomy; 2) some suitable references for additional information; and basic principles of 3) Relativity, and 4) Quantum Physics. The Preface contains a list of some very readable books and a number of Internet links to reviews or tutorials on Astronomy/Cosmology. Also, most of the illustrations in this Section were made from images and data acquired by spaceborne Observatories. A brief description of those Observatories is given on this Wikipedia website. Most of the ground-based Observatories are listed in this Caltech site.

As we did in Section 19, we begin with this statement: Astronomy and Cosmology depend almost entirely on remote sensing technology (mainly telescopes with various sensors) to gather the data and mold these into information about every thing in space beyond our Solar System.

Cosmologists - those who study the origin, structure, composition, space-time relations, and evolution of the astronomical Universe (and the possibility of a Cosmos as defined above) - generally agree that the **Universe** had a finite beginning and that it is expanding at a steady rate so that any two points (e.g., galaxies) move away from each other at speeds proportional to their separation. (The expansion of space has been referred to as the Hubble Flow, to honor Edwin Hubble who first verified the expansion). This beginning is commonly referred to as the **Big Bang**, which is not an explosion in the sense of, say, the detonation of dynamite but is an "explosion" of space itself as a continuing expansion accompanied at the outset by the creation and release of all energy and matter now occupying the ever growing Universe. (The Big Bang received its descriptive name as a disparaging comment from the astronomer Fred Hoyle, who advocated instead an infinitely large Universe of constant matter density [requiring continuous creation of new particles to maintain the density even as the Universe expanded within its infinite limits] as described in his [now rejected] Steady State model [developed in consort with Hermann Bondi and Thomas Gold]. This model also infers its Universe to have always existed [no creation event] and will exist largely unchanged [except for its expansion] forever; variants of this and other models have been put forth, as described on page 20-9).

As of 1990 the time of the Big Bang had been placed between 12 and 16 Ga ago (Ga = 1 billion

years [b.y.]) ; the current best estimate (derived from observations made by the Hubble telescope and WMAP [a cosmic background radiation satellite]) lies close to 14 Ga (13.7 Ga is now recognized as the most accurate value [see page 20-9]). This is derived by measuring the time needed for light to have traveled from the observable outer limit of the Universe to Earth in terms of *light years* *, which can be converted to distances. In a sense, the term "light year" has a dual meaning. Thus, when the value of 2000 light years is stated for a star or galaxy, one could think in terms of distance: the entity is 2 x 10^3 x (3600 x 24 x 365.4 [the number of seconds in a year] x 2.998.... x 10^8 m/sec (see first footnote *), approximately 11.8 quintillion kilometers, away from the Earth as the observing platform. Or, one might think in terms of age: relativistically, we see the entity as it was 2000 years ago when the light first left it; cosmically we always look back in time when observing stars and galaxies. Both distance and age are valid connotations.

At this outset, let us define the term "Universe". The (this; ours) Universe will be specified as everything that lies spatially within the outermost limit of matter and energy that has participated in the expansion of Space since the moment of the Big Bang. In this definition, **the Universe** (the one we live in; in principle, there may be other Universes [see page 20-10]) is finite in both space and time (note: it had a beginning and seemingly will last in some state for many billions of years to come [possibly infinitely]). This Universe is said to be **homogeneous** and **isotropic**. Homogeneity means that the entities involved are the same in all locations. Isotropy means that the entities are the same in all directions. These terms imply uniformity at some scales - generally large (cosmic). Thus, the Universe would appear much the same at any point within it. If we were to observe the Universe around us from a planet in some other galaxy, we would see generally the same set of physical conditions and the same general appearance and distribution of other galaxies elsewhere in the Universe as we now actually do from Earth. This must be modified by the scale of observation. The Universe shows apparent inhomogeneities, such as clumping of energy and clustering of galaxies, in regions that are less than about 200 million light years in size. But at larger scales the Universe approaches a more uniform or smooth status. (A broader meaning often applied to "Universe" holds it to include all that can be conceived to exist either physically and/or metaphysically; but as stated above we prefer using the term "Cosmos" for this idea.)

The *Cosmological Principle*, which is deducible from the postulates of homogeneity and isotropy, states that the Universe will look the same no matter where the observer is located within it. A corollary of this states that there is no real center for the Universe. But an observer at any location may think he/she is at the center. That notion as applied to Earth dwellers persisted until the 16th Century when Nicolaus Copernicus presented arguments that negated the geocentric view favored by philosophers and theologians and replaced it with the heliocentric view (the Sun is the center for the planets). (Galileo got into deep trouble with the Catholic Church for his support of Copernican centricity). The Sun was dismissed as a candidate for the Universe's center when, first, its place in the Milky Way was determined to be about a third of the way out from our galaxy's center, and then the galaxy itself was shown by Edwin Hubble in 1923 to be just one of many both nearby and far away from the Sun's galaxy.

The physical conditions that guaranteed the present Universe must have burst into existence almost instantaneously. During the first minute of the Universe's history, many of the fundamental principles of both Quantum Physics (or, as applied to this situation, Quantum Cosmology) and Relativity - the two greatest scientific discoveries of the 20th Century (see Preface, accessed by link above) - played key roles in setting up the special conditions of this Universe that have been uncovered and defined in the 20th Century. Quantum processes were a vital governing factor during the buildup and modifications of the particles and subparticles that arose in the earliest stages. Likewise, Relativity influenced the space-time growth of the Cosmos from the very start.

In the most widely accepted current model of the Universe, there is no starting place or time in the conventional sense of human experience. *Space*** , as now defined and constrained by the outer limits of the observable Universe, did not yet exist (see below); also, *sequential events*, embedded in a temporal continuum, had not begun. The observable Universe is just the visible or

detectable part extending to that part of the Universe where objects or sources of radiation have sent signals traveling at the speed of light over an elapsed time not greater (usually somewhat less) than the time (age) of the start of expansion. Most cosmologists now feel with some confidence that there is something real and physical beyond the observable Universe (be it the unseen parts of our Universe or some other Universe(s) but it is too far away for light to have had enough time to reach Earth's ground or orbiting telescopes). That observed part plus the unobserved part together make up the **Cosmos**.

Everything that exists physically is included in the Cosmos. (One can debate whether things "spiritual" are only the thought processes that have a physical basis, or do these things really exist independently.) As this Section unfolds, you will come to realize that there is a hierarchy that deals with the physical entities within the Universe, arranged (in part) by a progression of decreasing sizes. That hierarchy, in its simplest form, is:

?? --> THE ABSOLUTE VACUUM (THE COSMOS) --> OUR UNIVERSE (PERHAPS OTHER UNIVERSES)-->THE INTERGALACTIC MEDIUM (VOIDS CONTAINING GAS, PHOTONS, AND OTHER PARTICLES) --> GALACTIC CLUSTERS --> GALAXIES --> INTRAGALACTIC GASES --> STAR CLUSTERS --> STARS --> PLANETARY SYSTEMS --> PLANETS --> SATELLITES (MOONS) --> FIRST ORDER SURFICIAL AND INTERIOR FEATURES ON PLANETS --> LOCAL FEATURES --> INDIVIDUAL ENTITIES (FOR EXAMPLE, HUMANS) --> ATOMS AND MOLECULES --> SUBATOMIC PARTICLES --> SUPERSTRINGS -->?

A Broad View of the Universe's Organization and Evolution

The mysterious Absolute Vacuum (the writer's term) will be considered later in this Section (suffice to say now that it a rather abstract concept that considers the possibility of a dimensionless emptiness that stretches to infinity; time also is eternal, having no real beginning or end). The initiating event which started our Universe from out of that Vacuum, referred to as the **Big Bang** (**BB**), began at a point so small that the notion of spatial three-dimensions [3-D] has no conceptual meaning. The event sprang from some sort of quantum state of still-being-defined nature that marks the inception of space/time (thus, without a preceding "where/when"; philosophically "uncaused"), from which all that was to become the Universe can be mentally envisioned to have been concentrated. This *singularity* is described as a state that is not quite a point (dimensionless) condition which has extreme curvature and before which there was no "yesterday". The singularity is the first event in Universe history, so that it connotes a beginning-of-time aspect in addition to its beginning-of-space implication. At the very beginning, its physical nature transcends the laws of physics (including relativity); these laws break down, i.e., do not apply, but almost immediately came into existence. This extremely small point condition nevertheless contained all the energy within the eventual Universe. This singularity energy is measured as a temperature that reached to billions of degrees centigrade. The density of the point at the moment of singularity was extremely high - far greater than that characteristic of Black Holes.

At the very beginning of this (our) Universe, multidimensional space and time came into being and began to take on physical characteristics. But at the cosmic scale, these two fundamental properties must, according to Special Relativity, comprise the 4-dimensional spacetime Universe (see Preface for a definition of spacetime) we now observe (according to some theories discussed below and on page 20-10, additional dimensions are possible). The exact nature (concept) of time is still not fully understood and is subject to continuing debate (for an excellent review of time, read *About Time: Einstein's Unfinished Revolution* by Paul Davies, 1995); also consult his Web site on "What happened before the Big Bang" at this site (the host site contains many interesting and provocative articles; click on Albert Einstein within the page that comes up to get to the parent site). There is, of course, the conventional time of everyday experience on Earth (years, days, seconds, etc.), measured fairly precisely by atomic clocks (e.g., the pulsating beat of a cesium atom, used to define the 'second') and less so by mechanical timepieces or crystal watches. There are the redefining ideas of time consequent upon Special Relativity, in

which the perception of time units proceeds faster or slower depending on frames of reference moving at different relative velocities. There is the notion of "eternity" in which time just is - has no specific beginning or ending.

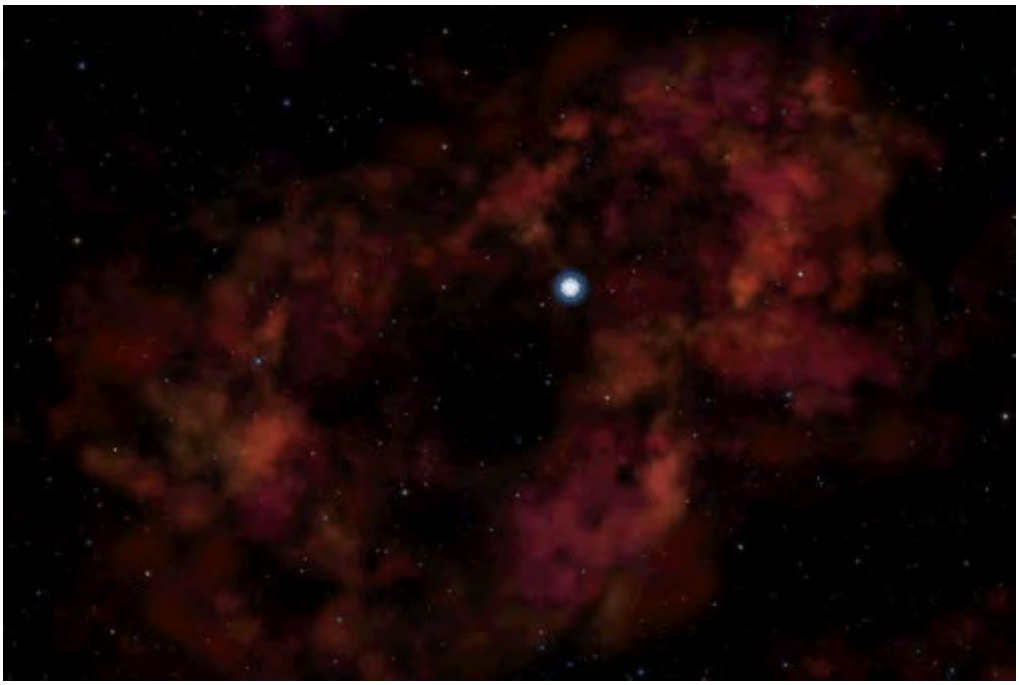
But, all these measures and concepts are difficult to extrapolate to that nebulous temporal state (if real) which was before the singularity at which our Universe came into being. But, time had to separate at that instant and become measurable in terms we have set forth to use its property of steady progression of a temporal nature. If nothing existed prior to the singularity event, then scientists presently have no means to determine and measure the nature of the time that was involved as a prior state. If ours is not the only Universe (see the discussion of multiverses on page 20-10), and other Universes existed before the one we observe, then time in some way can be pushed backward to their inceptions. One possibility is an infinite number of Universes in *time* and *space*, with no end points for starts and finishes (read Paul Davies' book for the philosophical as well as physical implications of time, and the still unresolved dilemmas in specifying the meaning of time).

For our purposes in studying the Cosmology of the one known Universe, we will assume time started at the moment the Universe sprang into existence. Arbitrarily, we postulate that time is immutable (a second at the beginning is of the same duration as a second is defined by today); there are models that postulate variable time values but we will ignore these. We accept the subsequent progression of time as being comprehensible in the units we define for Earth living. Thus, the Universe, under this proposition, can be dated as to its age in years - the year is an arbitrary unit, being the present day time involved in the Earth's complete revolution around the Sun.

At the very beginning, the fundamental energy within the singularity point may have been (or been related to) **gravitational energy** that controlled the nature of what existed at the singularity moment. An alternative driver now being investigated is some form of **repulsive energy** (similar to that once proposed by Albert Einstein as his 'Cosmological Constant'; but with a different numerical value) such as Quintessence (see page 20-10) which may prove to be related to the "Dark Energy" (page 20-9) that seemingly dominates the present Universe. (It is now customary to think of gravity as a positive effect and this repulsive energy as countering gravity as a negative effect [it may be, or be equivalent to, Dark Energy].) At the instant of singularity, the initial energy (some of which was about to become matter) was compressed into a state of extremely high density (density = mass or amount of matter [or its energy equivalent] per specific [unit] volume), estimated to be about 10^{90} kg/cc (kilograms per cubic centimeter) and extraordinary temperatures, perhaps in excess of 10^{32} °K (K = Kelvin = 273 + °C [C = degrees Centigrade]). (Note: the term "vacuum density" has been used in reference to this pre-Big Bang state; the density in this case refers to energy [which is a surrogate for mass according to $E = mc^2$]; this vacuum density is said to be very large.) Both high values are without any counterpart in the presently observed Universe; particle accelerators are not yet close to reproducing these ultrahigh temperatures. As you will see below, certain forms of matter came from the pure energy released during the first fraction of a second of the Universe's history. The famed Einstein equation $E = mc^2$ accounts for the fact that under the right conditions, energy can convert to matter, and *vice-versa*.

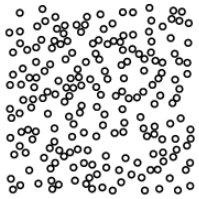
At the instant of the Big Bang's singularity, the particle (whatever its nature; the term "particle" also refers more generally to any of the fundamental entities such as protons, photons, muons, etc. that constitute matter and energy [see below on this page]) proved exceptionally unstable and proceeded to "come apart" by experiencing something that has been likened to an "explosion", which goes under the popular name of the "Big Bang" (**BB**). In TV shows that have an astronomy theme, such as seen on the History and Discovery channels, the depiction of the BB resembles a detonation or explosion (in the shows you usually hear a banging noise; this is meaningless since the BB moves into a true vacuum, which cannot support sound), and the terms have been applied (incorrectly) to the event. There is this fundamental difference: *In a conventional explosion, every thing involved is hurled outward from the point of initiation as an advancing and enlarging front that moves into existing space, leaving the volume between the*

front and the point devoid of the explosive debris; this volume increases in size as the debris progresses outward. To depict this visually, look at this animation of the explosion around the star Eta Carinae (you must be on the Internet):

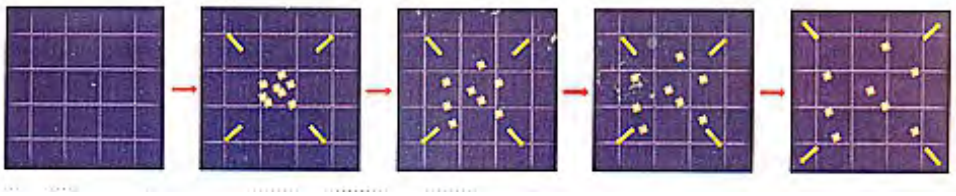


In the Big Bang, there is no real hurling away of the material released; instead all the material (from eventual galaxies down to subatomic particles) simply ***expands*** around the singularity point ***creating its own space as it enlarges***. The explosion is described as "not into space" but "*of space*". No center (in space today) can be specified for the BB since all points in the new finite but growing space simply draw apart more or less equally ***as space stretches*** under conditions in which pressure and density remain uniform and isotropic everywhere.

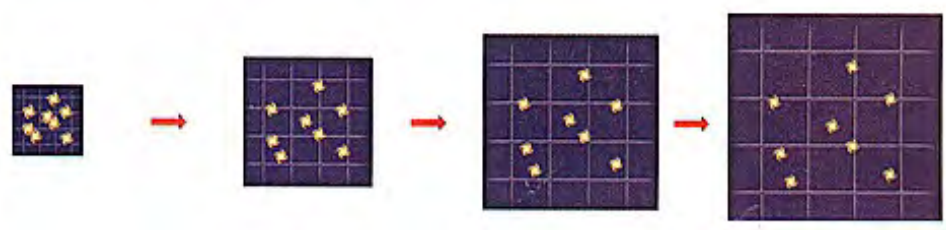
Several Internet sites actually have movies that simulate the expansion. This one works on the writer's computer; it should work on yours if you have the right software program installed:



This difference in behavior between conventional and BB is clarified in the two illustrations (read their captions for more information) below taken from (Six) Misconceptions about the Big Bang, by C.H. Lineweaver and T.M. Davis, *Scientific American*, March 2005, cited again on pages 20-8 - 20-10.



A Conventional Explosion



The Big Bang Expansion

ILLUSTRATIONS COURTESY OF ALFRED T. KAMAJIAN

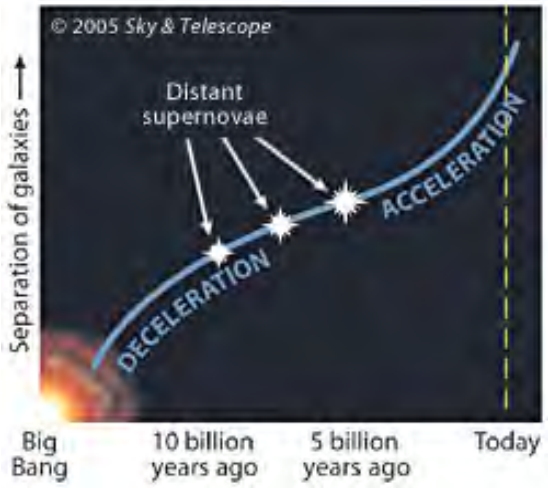
Intuition suggests that the dots in the above illustration would also be expanding (enlarging). These yellow dots represent galaxies. But in fact they remain roughly the same size during the expansion and move as a unit. The star distances within the galaxies stay about the same. This is due to the strong interstellar gravitation that holds a galaxy, once formed, to a near constant size. Over time as galaxies spread apart from expansion their mutual interactive gravitational

attraction weakens and this may disturb their shapes and shift the stars; but the distances between stars in a galaxy remain more or less the same (i.e., are not affected by the general space expansion) owing to the countering effects of gravity.

Note that in the diagram, the expansion seems to start from a specific point. But for the Big Bang one cannot speak of a "there" in reference to the singularity point because the space that characterizes our Universe did not start to form until the moment of its beginning. It is difficult to think of any "there" since no dimensional frame of reference can be specified. At the outset of "creation" the singularity was made up of pure energy of some kind (in a "virtual" state within the false vacuum). What might have preceded this moment at which the Universe springs into being and how the singularity point came to be (become) remains speculative; theoreticians in the Sciences have proposed inventive, although somewhat abstract, solutions but the alternative and traditional views of philosophers (metaphysicians) are still taken seriously by many in the scientific community. This last idea is treated again near the bottom of Page 20-11.

Expansion is continuing through the present and into the future in part because the inertial effects (evident in the observed recessional motions of galaxies, etc.) imposed at the initial push still influence how space grows and, now it is believed, in part due to the continuing action of the above-mentioned repulsive energy. After the freeing of gravity from the other fundamental forces (see below), it has since been acting on all particles, from those grouped collectively into stars and intragalactic hydrogen/helium clouds making up the galaxies to individual nucleons, photons, etc. - thus at macro- to micro-scales. Gravity therefore exacts one controlling influence on the rate of expansion, serving to slow it down. But, this rate should be decreasing over time because gravity between the Universe's constituents weakens as expansion forces them further apart (Newtonian inverse square of distance r effect [$1/r^2$]). As we shall elaborate later, recent evidence suggests that there are also anti-gravity forces (enabled by the repulsive energy of presently uncertain nature) that act to overcome the restraining effects of gravity; these forces seek to increase the expansion rate and over time push matter apart in a general dispersion. It is now believed that these forces are becoming greater than the countering force of gravity that eventually would have reversed expansion and caused a general collapse.

Gravity and the Kinetic Energy of outward expansion together constitute the total energy released from the Big Bang. By convention, the Kinetic Energy is taken as "positive" and Gravitational Energy as "negative". The two major energies comprise the Total Energy of Expansion. Thus; KE + GE = TE. Evidence favors a TE > 1, so that the Universe is likely to expand forever. The history of the expansion has been one of three successive stages: rapid deceleration, modest deceleration, and then exponential acceleration. This last step results from the increasing influence of Dark Energy (which maintains a constant density). Matter is thus said to be "falling outward" in the expanding sphere that comprises the observable Universe. This illustration generalizes the expansion history of the Universe:



This diagram also has implicit reference to the above three major stages in the Universe's evolution: 1) initial acceleration accompanied by the dominance of radiation in the first few

hundred thousand years of expansion; 2) the major role of matter as a factor in gravity over the next 7 billion or so years; and 3) the increasing importance of Dark Energy (or some other driving force) as the cause of re-acceleration since then.

Some readers may wish to acquire a broader insight into the topic of Universe expansion that describes a simplied Model, using an enlarging balloon as an analogy for the spacetime **expansion** of the Universe that has continued after the first eras of the Big Bang. This and related subjects are considered in more detail on pages 20-8, 20-9, and 20-10. But if you want to acquire a better understanding of the nature of Universe expansion before proceeding on this page, you can access a relevant review now on the separate page 20-1a. Check especially the paragraph in red at the bottom of page 20-1a, which alludes to models other than the Big Bang that could cause expansion and different sets of fundamental cosmological parameters.

This is an appropriate point to insert comments about the concept of the Instanton. This is an alternative version of the notion of the Singularity event described in previous paragraphs. It is a different version of what happens just prior to the Big Bang. The Instanton is a condition that derives from Yang-Mills Gauge theory which is a part of what is known as Quantum Chromodynamics (QCD). We will not delve further into Instantons (a rather difficult to follow summary is given at this Wikipedia website). Cosmologists such as Stephen Hawkings and Neil Turok have adapted Instanton theory to the conceptualizing of what was before and led up to the Big Bang, or any of the competing ideas for the Universe's inception. In a nutshell, they envision a process by which a quantum fluctuation in the vacuum or void prior to the initiation of the Big Bang led to the appearance of energy by a quantum tunneling process. Their "Pea Instanton", which had such high temperatures and pressures that it had to "explode", was created in this way. Rather than pursue this topic further here, we refer you to the Cambridge University link at the bottom of the Preface and to the links on this additional Web site: J.T. Wong.

Many scientists believe that what may have "existed" prior to the Universe was a quantum state (in a sense, analogous to the condition of "potency" in ancient Greek philosophy) which influenced a true vacuum (no matter whatsoever) that somehow possessed a high level of energy (of unknown nature but not, however, as photon radiation). As will become evident in this Section, astrophysicists now believe that no total vacuum exists anywhere -- all of space (the Cosmos) contains some form of energy having as yet an undetermined quantum nature. Countless quantum fluctuations (which in quantum theory are said *not* to depend on [obey] metaphysical cause/effect controls and are *not* subject to time ordering) in this vacuum energy density produced sets of virtual particles and anti-particles (analogs to positrons, the positively-charged equivalent of an electron; neutrons and anti-neutrons, etc) that could (according to quantum theory) into existence out if the Cosmos for very brief moments but then nearly all were annihilated. The nature of these particles is currently not well known but they may relate to the so-called Dark Energy that dominates the Universe.

Rarely, annihilation did not occur (as would be consistent with the probabilistic nature of Quantum Physics), so that a particle could grow and trigger a 'phase transition' that led to the Big Bang event from whence all that entails our Universe - with its internal matter, energy, space, and time - came into being. In this quantum model, particles could either be destroyed by interacting with antiparticles or could emerge from the vacuum from time to time and survive, leading to mulitple universes that, as far as we know theoretically, cannot have any direct contact. If so, the number of unconnected Universes may be very large, or very small if the success rate of a particle conversion to a Universe birth is near infinitely low frequency of occurrence. Though no one yet has offered any proof of a multiverse Cosmos, the likelihood is that the vast majority of virtual particles do not explode into individual Universes, but statistically some do; each Universe may have its own set of parameters and laws of physics and these conditions may never be "right enough" to foster life.

This non-contact status is one example of prohibition by relativistic limits, in which information

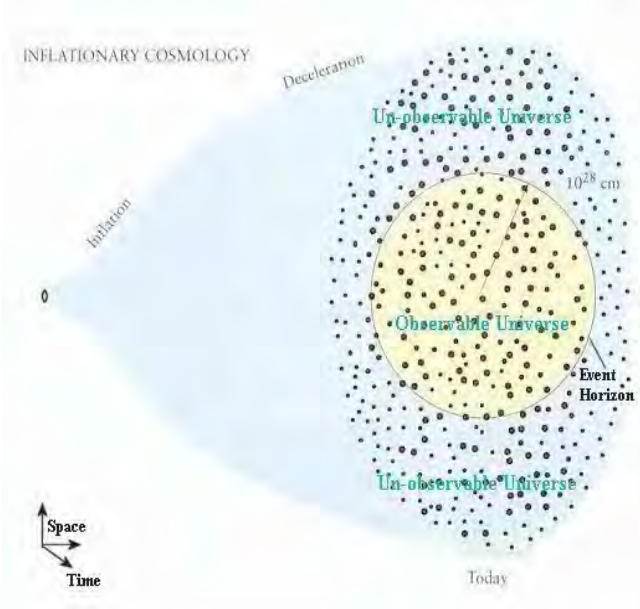
travelling at the speed of light cannot reach us from beyond the horizon of our own observable universe. For our Universe, the concept of the *Cosmological Horizon* refers to the boundary or outer limits of the Universe that we can establish contact with, i.e., the farthest extent of the observable Universe that can be seen through the best telescopes. This is approximated by the currently observed farthest galaxies that formed in the first billion years of time in our Universe's history. This Horizon is also conceptualized as the surface dividing spacetime (which includes all locatable 4-dimensional points) into what we can see and measure from what is hidden and unobservable. The observable therefore must lie within our *Light Cone*, an imaginary surface that encloses all possible paths of light reaching us since the beginning of time. (The fifth illustration below is an example). Check page 20-10 for further discussion of these ideas.

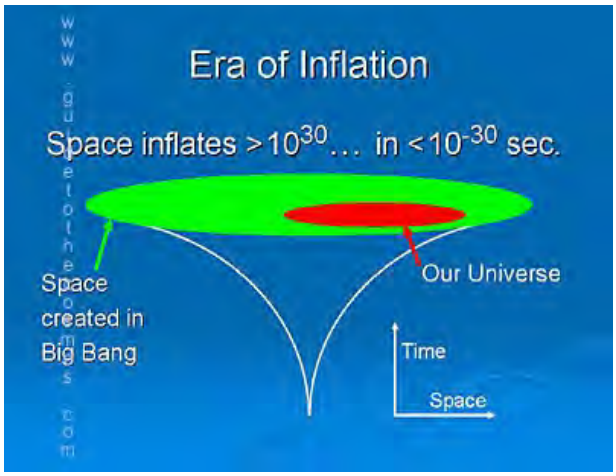
If the Universe is about 14 billion years old, then light leaving just formed protogalaxies near the observable spatial limit (outer horizon) of the Universe departed some 13+ billion years ago but this radiation is only now reaching us, since it had to traverse across a Universe that was expanding (ever increasing the distances from Earth to the outer edges) and drawing the protogalaxies away from us. Scientists actually have detected cosmic background radiation (CBR), the "afterglow" (see below and page 20-9), which pervades the entire Universe. Its first confirmable appearance was only about 380,000 years after the BB, at the time when detectable radiation could penetrate ion clouds that blocked its escape. This appearance of CBR is the present limit to the farthest lookback time involved, i.e., the extent to which we can peer into the past to find the earliest discernible event; nothing that occurred between the BB and the 300,000 years following it up to the first detection of CBR can be directly detected or measured.

The Hubble Space Telescope has now seen faint galaxies that are close to the cosmological horizon. Light from these left them about 13 billion years ago. At that time all the earliest galaxies were much closer to each other. Over the next 13 billion years - to the present - the Universe has been expanding, so that the light initially emitted "way back then" has had to traverse an ever enlarging distance which we perceive as the observable Universe.

The ultimate size of the **observable Universe** - a term we have applied several times earlier on this page - is still an open question. It is also a very complicated question, as you would deduce if you read this Wikipedia website that contemplates the possibilities of Universes of different sizes, depending on the definitions and assumptions used.

Cosmologists are inclined to cite as one specifiable size some specific *observable* Universe which is the subset of a possibly much larger Universe that lies beyond the event horizon (limit of the spacetime distribution within which the earliest light has traveled to our planet); in this conception, what is being seen from Earth is just that part of a still larger assemblage of galaxies from which light has had enough time to reach our telescopes since the Big Bang 13.7 billion years ago. We see outward in all directions to those galaxies at the limit, as they were in their earliest appearances (they formed about 0.5 to 1 billion years after the Big Bang), and they appear much the same no matter what direction we look at them. Thus we can imagine a sphere of galaxies whose radius is at least that of the first galaxies, e.g., let us say 13 billion light years away (to the first recordable event horizon). The diameter of the sphere of the observable Universe is thus 27.4 billion light years. For this sphere, in our frame of reference, we perceive ourselves as being at the center. But someone observing from a planet in a galaxy elsewhere would see the same thing (the sphere thus seems centered at that planet). In this conception the presumption is that there are many billions of galaxies situated beyond the limit of detection that has been set by the time since the Big Bang over which light has traveled at its apparently constant speed. Perhaps these diagrams will help visualize this:





In these diagrams the regions beyond the event horizon (which defines the observable Universe) were formed at the same time as those within. The extent of galaxies beyond the observable limit can be even much larger than shown in the diagrams but may(?) be finite (see two paragraphs down). Just how large is still unknown, is still conjectural, and is still dependent on which model or theory is being used. Two values show up in an Internet search of "the size of the Universe": a radius of 42 billion light years and a radius of 78 billion light years. Explanations of how these and other values have been reached are obtusely documented (as an example, see the 8th entry in Ned Wright's FAQs) and reasons for the mechanisms that lead to sizes beyond the observable Universe are glossed over. The most common argument states that the greater proposed radii are the consequence of Inflation (see next paragraph and again elsewhere on this page) which carried the earliest products of the Big Bang to distances much farther than the 13.7 billion light year limit imposed by observation.

Thus, there seems to be a paradox here. How can there be more galaxies outside the observable sphere? Inflation, which occurs almost at the beginning of the Big Bang, increases the rate (which has varied in the past) of expansion of the Universe by a very large factor (one proposed value = 10^{50}; higher and lower rates have been proposed for various models). This is much greater than the speed of light (this does not violate the Einsteinian tenet that **radiation within the Universe cannot go faster than light speed**, which applies to movements of photons *within* space, whereas it can be argued that **space *itself* can move faster the speed of light**). Thus, there are a multitude of galaxies and other matter/energy outside of that spherical portion of space that can be observed that are part of the vast segment of megaspace (all space out to the farthest extent of the Universe) produced by the Big Bang + Inflation; we can't see them simply because they are too far for light from them to have had time to reach us since the beginning.

The concept of Inflation, coupled with the illustrations above, seems to be a straightforward and easily grasped clue to envisioning how there can be matter and energy - perhaps even as galaxies - beyond the 13.7 billion light year horizon. But there is a competing relativistic explanation that leads directly to the extension of the Universe's limits or size to values quoted above, to such numbers as 42 and 78 billion light years. The writer (NMS) first encountered this explanation in Lecture 8 of the Cosmology DVD, presented by Dr. Mark Whittle, that was cited at the beginning of this page. A search of all the main Internet references and several Cosmology textbooks failed to find anything comparable to his review.

Whittle begins that lecture by defining four terms: 1) d_{emit} = the distance light has traveled (to us) since it set out; 2) d_{now} = the distance that the galaxy actually is from us today; 3) d_{LT} = the distance that a photon of light has traveled since it first left its source until now: and 4) the lookback time $t_{LT} = (t_{now} - t_{emit}$, which is numerically equal to d_{LT} but in units of years. These distances always have this interrelation: $d_{emit} < d_{LT} < d_{now}$.

Two other terms are used in his calculations: 1) Scale factor S(t), which ranges from 0 (at the Big Bang) and 1 (now), and describes the extent of expansion at any stage; and 2) the so-called redshift of electromagnetic radiation (light) *(defined as the displacement of spectral lines of elements such as H and Fe [whose wavelengths have reference values determined in their rest state in laboratories on Earth] towards longer wavelengths owing mainly to the outward expansion with progressively increasing velocities of the Universe*; discussed below and on page 20-9) expressed as a variant known as the Redshift Stretch Factor (RSF). RSF = $d_{now}/d_{emit} = \lambda_{now}/\lambda_{emit}$, which ranges from 1 to infinity. (The RSF has a value of 1000 associated with the Cosmic Background Radiation that is detectable about 400000 years after the Big Bang). (Note: redshifts took on large values during the first part of the Universe's expansion, especially during the first minute of the BB; this is quantified on page 20-9.)

It is not practical to try to reconstruct the details of his calculations as presented in the lecture. The specification of d_{emit}, which is just any time one chooses to start the journey of a photon (light) from a galaxy at any time between the beginning and now, and d_{LT}, which is synonomous with the st(age) of the light source (a galaxy specified to be seen today as it was, say, 8 billion years ago is said to be 8 billion light years away), are readily understandable. But when d_{now} for this galaxy is stated as a number greater than 13.7 billion light years, the situation seems bewildering and counterintuitive.

Whittle gives two examples with concrete values that specify the various distances. In the first, he starts with a galaxy whose RSF = 3. This means that the Universe was 3 times smaller, so $S(t_{emit})$ was 1/3. Referring to the appropriate S(t) curve, t_{emit} is found to be 10.5 billion years ago, so that d_{LT} is 10.5 billion light years. But since then, the Universe has been expanding so that the distance that light travels at a constant velocity keeps enlarging. This means that the present day (now) distance is greater than d_{LT}. To determine this one must resort to calculus and integrate the distance of many small, ever changing increments over time. For a small interval, this distance = $c\Delta t$; over the full time of stretching (expansion) from then to now, the relevant S(t) must be used in the integration of the varying $c\Delta t$. The calculus formula for this is:

$$\int c/S(t)dt$$

(Note: The html editor program used for this Tutorial has no character for the integral symbol, which was here extracted from the Internet, nor does it allow the formula to its right to be placed on the same line; so, in the above equation just mentally place the integral symbol next to the formula.)

For this set of parameters, the integrated d_{now} becomes 15 billion light years. Using the same approach, and specifying the Redshift Stretch Factor as RSF = 1000, d_{LT} comes out as 13.7

billion light years, d_{now} is calculated to be 46 billion light years, and d_{emit} is found to be 26 million light years (a reasonable estimate for the very early Universe's limits).

These seem to be strange numbers, and the concept is hard to visualize. The writer (NMS) came up empty after seeking to find a dynamic illustration (or even a set of sequential static pictures) on the Internet. So, I have decided to make my own - which is only an approximation. Consider these:

At t(0)

O O

At t(1)

O <-- **O**

At t(2)

O <----- **O**

At t(3)

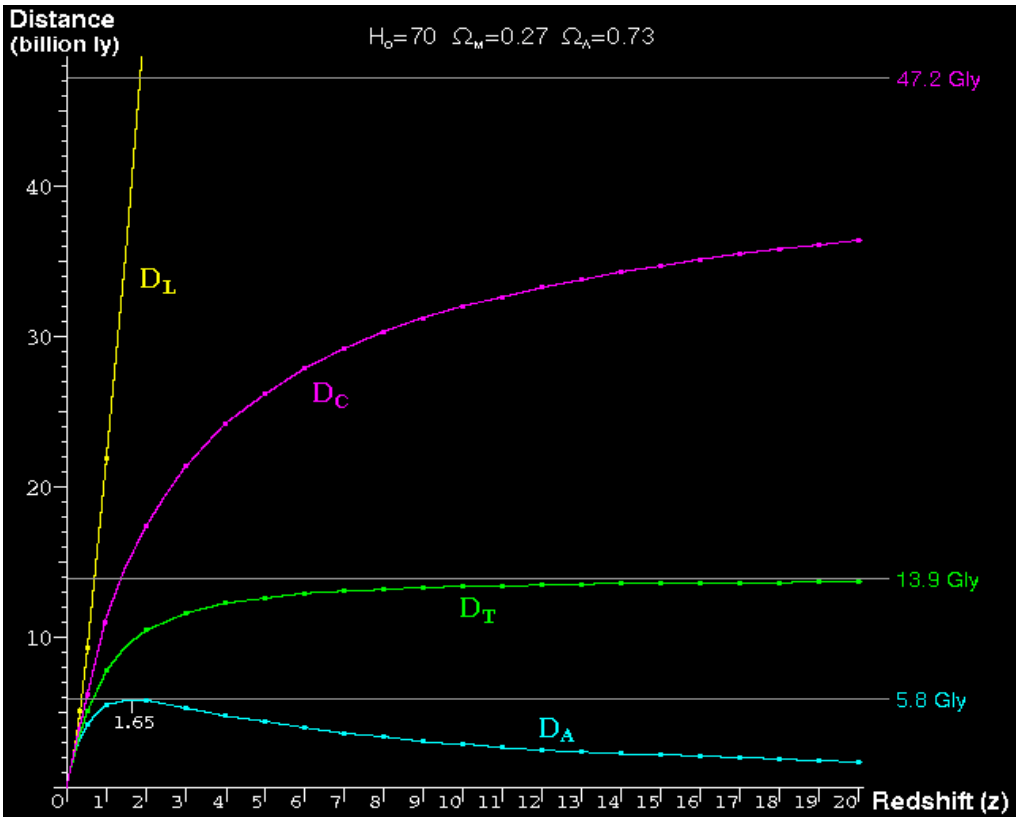
O <--------- **O**

At t(4)

O<---------- **O**

At t(0), just after the Big Bang, two entities (let's make one **O**, on the right, what would later become a now distant galaxy and the other **O** being our galaxy containing what eventually becomes our planet Earth) are adjacent and light (represented by an arrow, whose length represents the distance traveled from the distant galaxy in the time interval t(n)) has just begun its travel. At t(1), the entities are expanded some distance apart at an amount governed by the Hubble expansion rate, but light has traveled (whose d_{LT} is shown by the number of dashes in the <--) a lesser distance determined by its constant velocity and the time involved than this expansion distance. At t(2), the separation distance of the **O**s is now greater, but note that the light travel distance is itself now more but still increasingly less than the amount of **O** separation. This holds for t(3) and any other intervening intervals. At t(4), the light has now reached the left **O**, which is Earth, and its distance (d_{LT} or 13.7 billion light years) is considerably less than the gap between the two **O**s. The increasing disparity between the **O** separations in successive times and the light travel distances is accounted for by the non-constant rate of expansion versus the constant rate of light propagation.

The question of the size of the Universe is not well covered on the Internet. One web site that proved informative is recommended: Distance Scale of the Universe. The figure below is taken from that site. Check the site for an explanation of the four distance parameters.

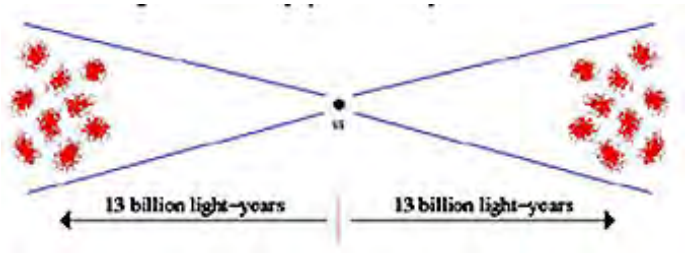


The writer (NMS) has tried to find a description of what may lie beyond the 13.7 billion light year limit for the observable Universe. An Internet search found hundreds of citations - mostly blogs - which range from inconclusive to gibberish. I do not have an answer but this seems plausible: If there is organized matter (from nebulae to possible protogalaxies) beyond the observable Universe, then it should be similar to the earliest galaxies observed so far; or it might just be an extension of the Cosmic Background Radiation which can be traced to the "edge" of the observed Universe.

As is true for most earthbound illustrations that try to depict relativity and spacetime, this is an imperfect diagram. But it may give you some insight beyond what the exposition of the Whittle lecture calculations provided. Let's return to a consideration of plausible models for our, and perhaps other, Universe(s).

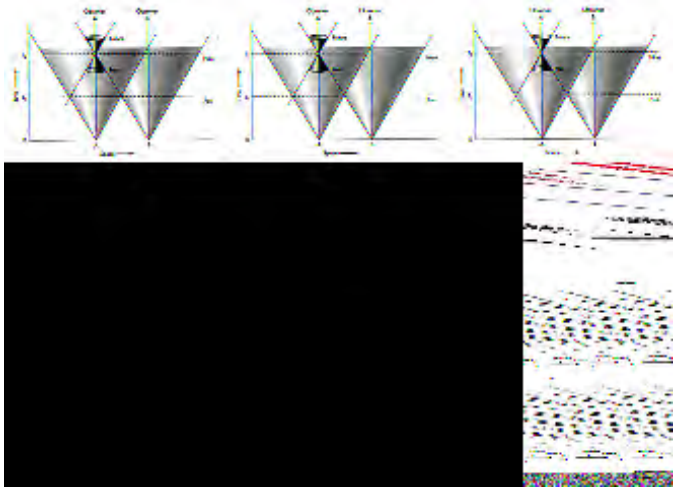
In one school of thought megaspace (the Cosmos - all possible space, not just that within the observable space of our limited Universe) is infinite. In another view, space is finite but much of it is beyond our detection. (What is outside of this megaspace is still conjectural). There is a variant of this (see page 20-10) embodied in the concept of Multiverses. In one model, each Universe can be likened to an expanding bubble and the bubbles may not be in contact (but in principle could interact) - the space between them itself also expands. Multiverses may be finite or infinite. Here is a pictorial example:

A corollary: In the Standard Model for the Big Bang, there have been and are parts of the Universe which cannot directly influence each other because there hasn't been enough time for light from any one part to have reached some others. Thus, the 'horizon' relative to Earth as the observing point (but any other position in the Universe is equally as valid an observing point) refers to the spatial or time limit that demarcates between what we can establish contact with in any part of the Universe and what lies beyond. This means that if an observer at one point in the observable Universe (as a sphere) sent a message shortly after the Big Bang to an observer at an antipodal point, there hasn't been enough time for the message to be received. This figure illustrates an extreme example of parts that cannot mutually communicate:



This gives rise to a seeming paradox that is implicit in the "Horizon problem". Simply stated: how can these isolated regions have very similar properties (such as similar densities of Dark Matter, Cosmic Background Radiation, and numbers of galaxies) if they are not in contact. This appears to violate the fundamental principle of universal causality, which holds that during expansion all parts of the Universe would need to have been in communication (by light transfer or other means of exchanging energy) so that the fundamental principles of physics would have ample causal opportunity to influence each other. This is seemingly necessary if at a gross scale the Universe is to maintain uniformity (the essence of the Cosmological Principle which postulates broad homogeneity and isotropism within the Universe as a whole). One explanation that accounts for the causality needed to obey this principle is given below in the subsection dealing with Inflation.

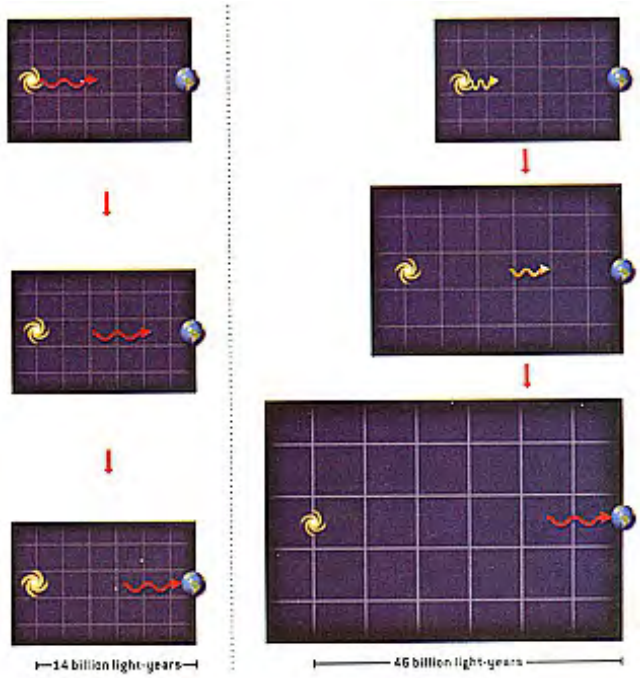
Nevertheless the isolation of regions of the Universe from one another is a real fact, as evident in the above illustration. And, specifically there were situations whereby some parts of the Universe were not in causal contact shortly after the Big Bang, and thus not visible to one another during early cosmic history, but will eventually, as expansion proceeds, become known to each other. Consider this next diagram based on spacetime light cones (see Preface):



From J. Silk, The Big Bang, 2nd Ed., © 1989. Reproduced by permission of W.H. Freeman Co., New York

Start with hypothetical observers at two points A and B not then in contact in early spacetime. Over expansion time, their light cones would eventually intersect, allowing each to see (at time t_1) other parts of the Universe in common but not yet one another. At a later time, beyond t_2 ("now") in the future, the horizons of A and B (boundaries of the two light cones) will finally intersect, allowing each to peer back into the past history of the other.

These intriguing ideas just discussed actually don't tell the full story. One model of Universe expansion arrives at a Universe whose farthest opposite points are now about 42 billion light years apart. Check this diagram:



From: Misconceptions about the Big Bang, by C.H. Lineweaver and T.M. Davis, *Scientific American*, March 2005
ILLUSTRATIONS COURTESY OF ALFRED T. KAMAJIAN

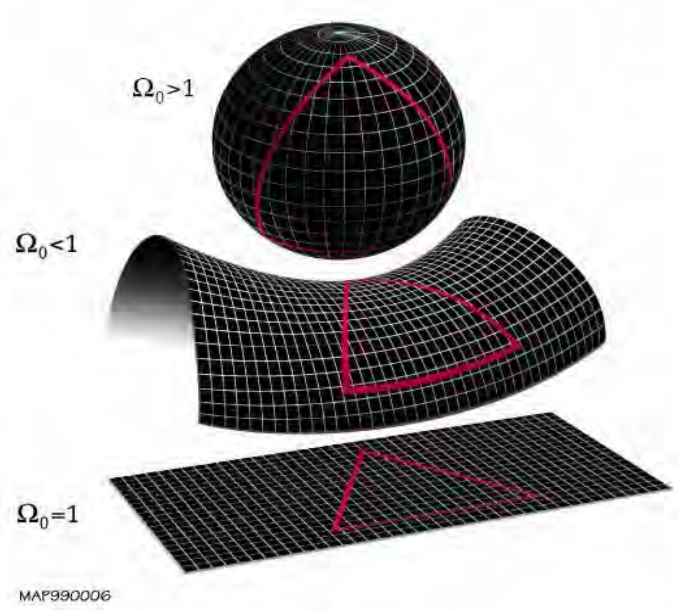
On the left side is the model that equates the radius of the observable Universe with the age (rounded off here at 14 b.y.). This just assumes a "static" condition in which the age is determined by light speed alone. Light from the yellow spiral galaxy (which is the most distant from Earth) in the top panel left 14 b.y ago and arrives now in the bottom panel. But, in actuality, during that 14 billion years, the Universe has continued to expand. If one assumes a expansion factor of three during the 14 b.y. time interval, the situation is as pictured on the right. The light

emitted from the galaxy has in 14 b.y. had to travel an ever enlarging space, so that **today** the galaxy is at least 42 billion years distant from Earth. Since we don't know where Earth really is in this finite Universe, it is currently impossible to determine the actual farthest points in opposing directions [on the sphere model, or on the flat model. But this diagram is important in indicating the Universe is really larger than the 28 billion light year dimensions cited above. How much larger is still speculative: Since we haven't any direct information about the extent of galaxies beyond the observable Horizon, we cannot specify a known size; several proposed models arrive at different dimensions including those greater than 42 billion light years.

Perchance at this point you may be confused a bit by these "heady" concepts. Some insight and a fresh look might result by checking these Expansion of Space, Gary Felder, Wikipedia and Prof. Seligman Internet sites.

Commenting further on the Universe's geometry: One view holds the present Universe to be finite but without boundaries. Its temporal character is such that it had a discrete beginning but will keep on existing and growing into the infinite future (unless there is sufficient [as yet undiscovered] mass to provide gravitational forces that slow the expansion and eventually cause contraction [collapse]). A much different model considers the Universe to be infinite in time and space - it always was and always will be (philosophically, these can be tied to concepts that equate God as an "intellectual presence" distributed throughout this naturalistic Universe)

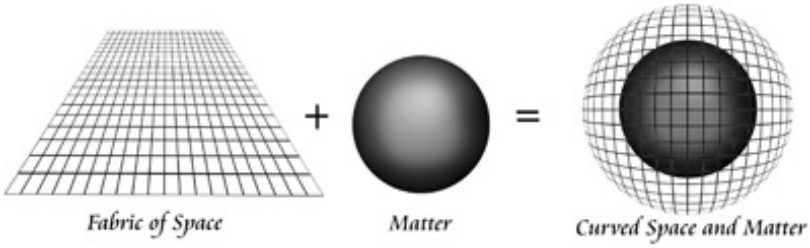
Models for the Universe's shape hold it to be analogous to spherical, hyperbolic, or flat. A parameter called critical density (Ω) determines the shape (page 20-10). This diagram illustrates the three types:

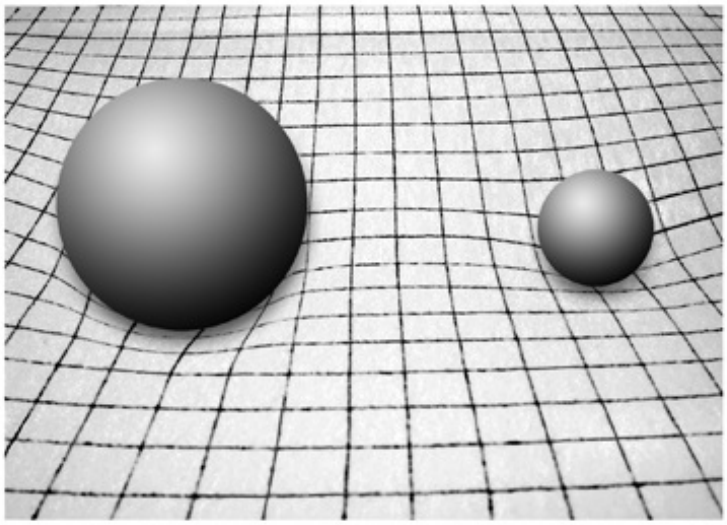


In addition to specifying the Universe's shape, cosmologists seek to know whether it is open or closed, whether it is presently decelerating or accelerating, and whether it is infinite or finite in time and space - these topics are treated in detail on pages 20-8, 20-9, and 20-10. For now, lets preview these topics by saying that the **prevailing view** is that the Universe is **flat**, is **open**, and is **accelerating its expansion**.

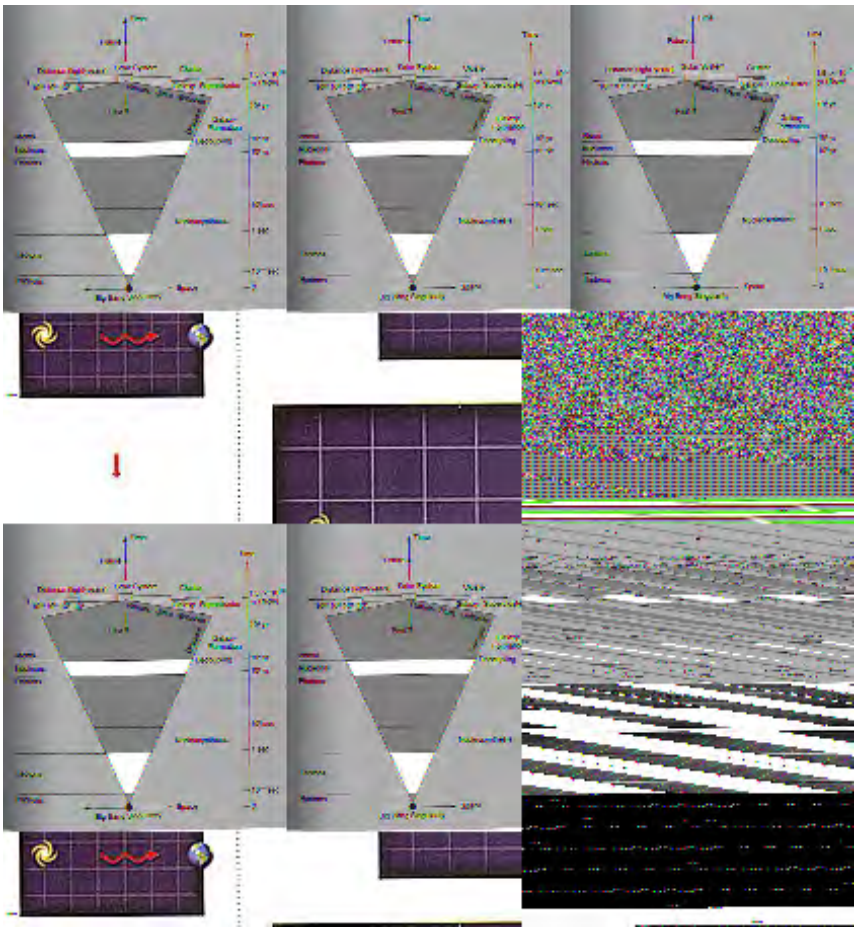
Einstein, in particular, showed that any three-dimensional expansion must also consider the effects of the fourth dimension - time - to account for the behaviour of light traveling great distances in a vast "volume" (without known boundaries) making up what we conceive of as "space". He also deduced that space (within the Universe) must be curved (and light and other radiation will therefore follow curved paths as the shortest distance between widely separated points) and would, in his view, expand dynamically in a 4-dimensional spherical geometry (a spacetime dimensionality). (Einstein, at least in his early thinking, also considered the Universe to be finite and eternal; he did not take a firm stand on its overall shape.) As was considered in the Preface, the internal curvature of space was deduced by Einstein to be the consequence of the

interaction between matter (responsible for gravity) and the "fabric" of space. These two diagrams depict this:





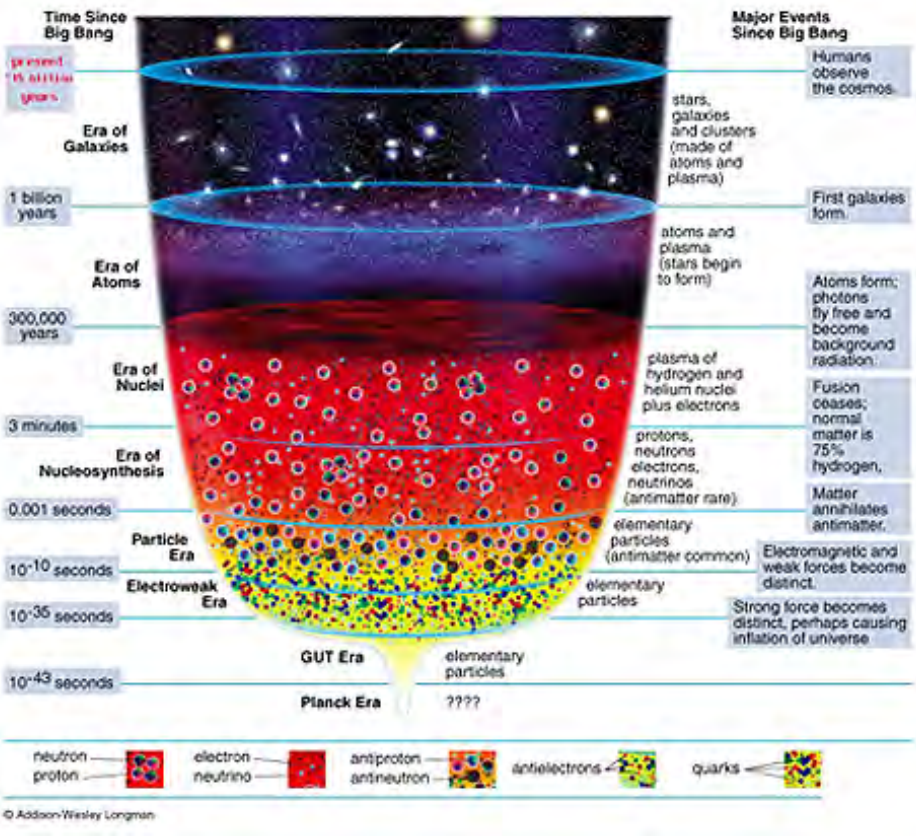
The next figure is a spacetime diagram that summarizes the history of the expanding and evolving Universe in terms of the general or Standard **Big Bang** (BB) model for its inception:



From J. Silk, The Big Bang, 2nd Ed., © 1989. Reproduced by permission of W.H. Freeman Co., New York

Each major step in this time history of the creation and development of the physical Universe will be reviewed in some detail later on this page. For now, the diagram lets us extract this sequence: 1) By the end of the first millionth of a second, hadrons (quarks that make up baryons [including protons and neutrons] and mesons) had formed; 2) in the next interval of time up to 1 second leptons (electrons, muons, and neutrinos) came into existence; 3) over the next 1000 or so seconds nucleosynthesis of mostly Hydrogen, some Deuterium, priordial Helium, and a tiny amount of Lithium occurred, that is, their nuclei started to form; 4) as expansion continued over the next several hundred thousand years, the particles in this young Universe would remain invisible to any backward-in-time-looking detector because of "opaqueness" imposed on the photons by the free electrons (not yet coupled with atomic nuclei) that interact with the photon "fog"; 5) thereafter, the nucleons consisting of H and He protons and neutrons began to combine with electrons in the process of decoupling; 6) during the post-decoupling stage in the first million years the now stable atoms of Hydrogen and Helium started to clump together into gaseous clouds and then stars to form the first galaxies.

A variation of this figure with additional information appears below:



The Big Bang as an expansion theory traces its roots to ideas proposed by A. Friedmann in 1922 to counter ideas attendant to Albert Einstein's Theory of General Relativity, from which that titan had (erroneously) derived a model of a static, non-expanding, eternal universe (he eventually abandoned this model as evidence for expansion was repeatedly verified and he realized his General Relativity proved very germane to the expansion models). This fundamental equation, which introduces the Scale Factor R (see page 20-8), can take several forms, one of which is (see caption for units):

$$\left(\frac{1}{R}\frac{dR}{dt}\right)^2 = \frac{8\pi G}{3}\rho - \frac{kc^2}{R^2}$$

Multiplying each term by R^2 yields this equation which expresses the rate of change of the cosmic Scale Factor R with time:

$$(dR/dt)^2 = (8 \Pi G)/3 \rho R^2 - kc^2$$

The Abbe George Lemaitre (a Belgian Catholic priest - a Jesuit) in 1927, as an outgrowth of his Ph.D. thesis at MIT, set forth another expansion model that started with his proposed "Primeval (or Primordial) Atom", a hot, dense, very small object. (Lemaitre is credited with moving the idea of expansion into the mainstream of cosmology, but as indicated above, A. Friedmann had devised an expansion model 5 years earlier) The nature of a Big Bang was refined and embellished by the team of G. Gamow, R. Alpher, and R. Hermann and by others in the 1930s; their calculations showed that Hydrogen and some Helium are the dominant atomic species (with minor amounts of Lithium) that could form from their model of the Big Bang and its consequences. (The heavier elements up to Iron are produced by nucleosynthesis in the hot interiors of stars, as first demonstrated by Fred Hoyle and colleagues at the University of Cambridge.) Confirming evidence for expansion came from Edwin Hubble in the late 1920s. (Hubble's other major achievement was to prove the existence of galaxies outside of the Milky Way; he developed methods for determining distances to Andromeda and other nearby clusters of stars that were greater than the dimensions of the Milky Way.) The Big Bang can be mentally related to the above-mentioned singularity event by *imagining* that the expansion is run in reverse (like playing a film backwards): all materials that now appear as though moving outward (as space itself expands) would, if reversed in direction, then appear to ultimately converge on a "point of origin".

As described later in this Section (page 20-9), the BB concept drew its principal support from the observations by Edwin Hubble and others on radiation redshifts associated with the distribution of galaxy velocities. These redshifts (changes in the frequency [a decrease] of the EM radiation from excited atoms, resulting in relativistic increases in wavelengths owing to accelerations analogous to the Doppler effect [which causes a drop in pitch of a train whistle as it recedes from the listener; see page 20-9]) rise in value as light and other radiation from galaxies comes from ever farther positions in the expanding Universe. Those galaxies with higher redshifts are also ones that display as we see them now younger conditions - thus, we see them as they were in the earlier stages of cosmic time; being farther away it has taken longer for emitted light to get from the starting point to detectors at Earth.

The italicized segments below were taken from the University of Virginia's website on Cosmology:

The Doppler Redshift results from the relative motion of the light emitting object and the observer. If the source of light is moving away from you then the wavelength of the light is stretched out, i.e., the light is shifted towards the red. These effects, individually called the blueshift, and the redshift are together known as doppler shifts. The shift in the wavelength is given by a simple formula

(Observed wavelength - Rest wavelength)/(Rest wavelength) = (v/c)

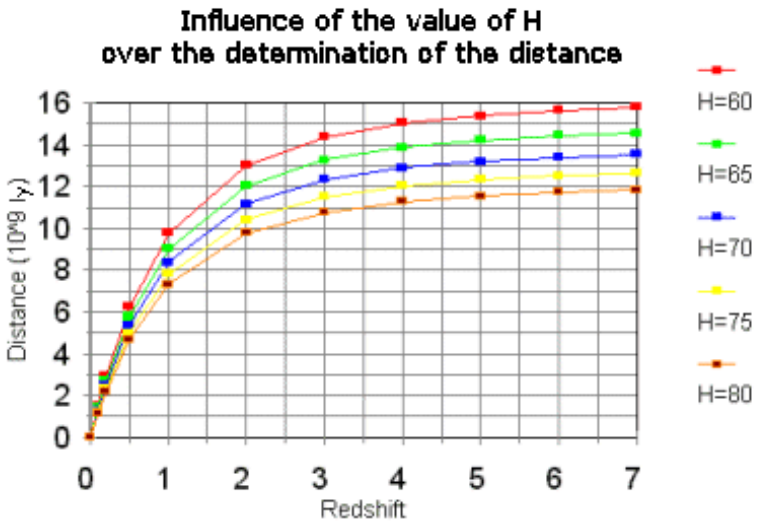
so long as the velocity v is much less than the speed of light. A relativistic doppler formula is required when velocity is comparable to the speed of light.

The Cosmological Redshift is a redshift caused by the expansion of space. The wavelength of light increases as it traverses the expanding universe between its point of emission and its point of detection by the same amount that space has expanded during the crossing time.

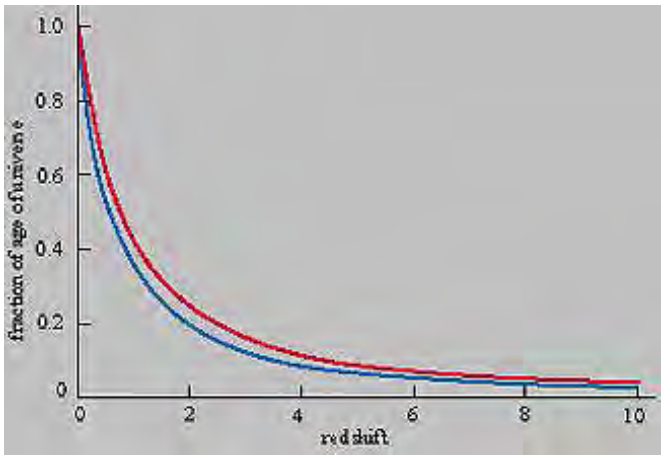
The Gravitational Redshift is a shift in the frequency of a photon to lower energy as it climbs out of a gravitational field.

The general pattern of redshift change with distance (which in this diagram is given as the ages of the galaxies examined in terms of how long it has taken light from each to reach the Earth

[thus those farthest away are shown as the youngest) follows this plot (shown for four values of the Hubble Constant H (see two paragraphs below), of which 72 is the current most favored value) is shown in this plot:



The exponential shape of this curve is carried over to the next plot which is a generalized representation of the cosmological redshift versus time since the Big Bang (set at 0 [13.7 b.y. ago], with the present age being 1). So far, the largest redshift actually measured is about 10, for a faint galaxy (observed by the Hubble Space Telescope) that may be as old as 13.4 billion years. /p>



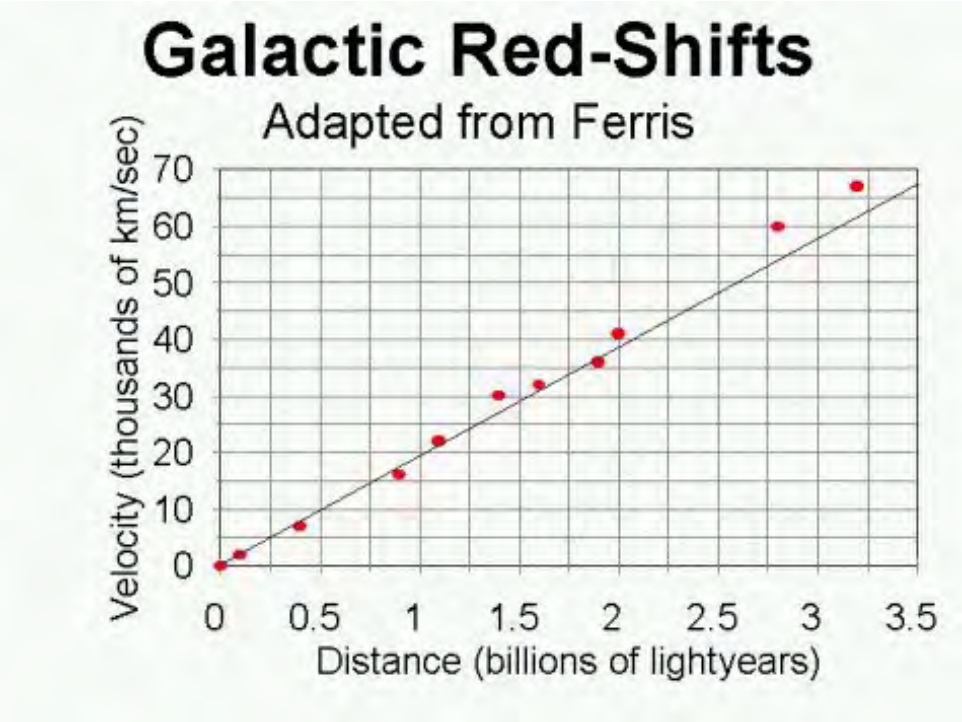
The Universe has been enlarging ever since this first abrupt Big Bang, with space itself doing the *expanding*, and galaxies *drawing apart*, so that the size of the knowable part of this vast collection of galaxies, stars, gases, and dust is now measured in billions of light years (representing the distances reached by the fastest moving material [near the speed of light] since the moment of the Big Bang [~14 billion years ago]). This age or time since inception is determined from the Hubble Constant H (which may change its value) which is derived from the slope of a plot of distance (to stellar or galactic sources of light) versus the velocity of each source (see page 20-9).

The Hubble Constant H is a fundamental cosmological value that determines the rate of expansion of the Universe. It is a part of this key equation:

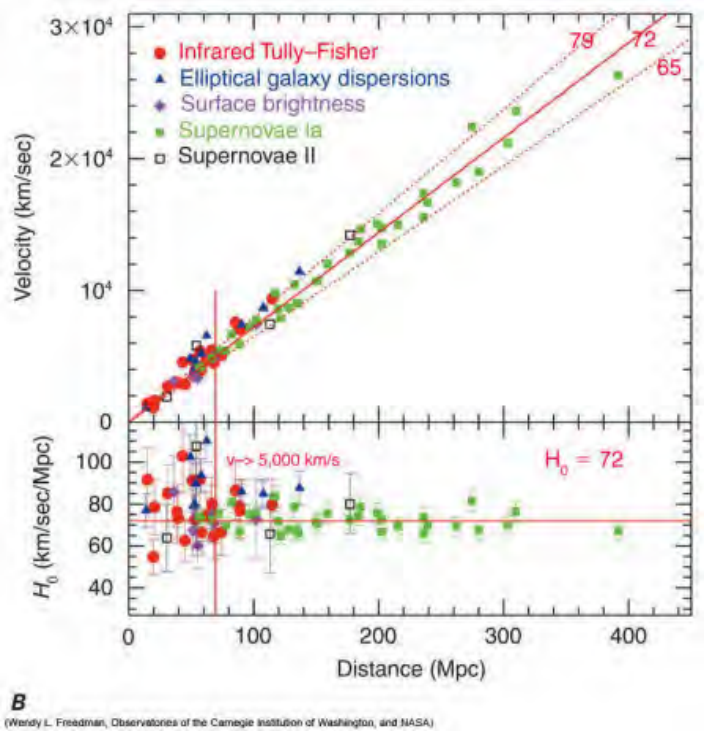
v = Hd

This, the Hubble equation, implies that the velocity of any point in expanding space (such as the location of a galaxy) has some current fixed value. Of greater import, the Hubble expansion directly leads to this conclusion: The galaxies are moving away from Earth at recessional

velocities that increase systematically with distance from our planet (with corresponding increases in redshift). This is shown in this diagram, for galaxies up to a few billion light years away:



This plot shows that the speed of recession increases as one progresses outward (towards the outer limits of the observed Universe) by an amount derived from the value of H. This makes sense in that if all points began, at the Big Bang, from the same point at the moment of singularity and have now spread apart by expansion, the outermost points (earliest stars and galaxies) must have moved the fastest and those at the full range of distances along a line of sight going back to the point of observation are moving at progressively lesser velocities. In this way, one can say that everything is expanding, at a rate determined by the value of H. That value has now been determined to an accuracy of +/- 10% and is given as: 71 km/sec/Megaparsec or 21.5 km/sec/million light years. This diagram shows how H (given as H_0) is determined as the slope of a straight line plot, using distances determined by different methods (page 20-9):



The Hubble constant affords a measure of the age of the Universe, as will be developed on page 20-9. As a quick preview, consider this: Replace velocity in the above equation with d/t. The equation then becomes: $d/t = Hd$. Divide both sides by d and invert t, so that: $t = 1/H_0$. An age of 13.7 billion years is the current best estimate.

The Hubble Constant also is applied to determining how fast galaxies at the observable Universe horizon are traveling. Starting with 71 km/sec/Megaparsec as the rate (and assuming this rate to be constant [it may have been slower in the past and is now faster], so this is an average) of expansion, the distance to a 13.7 billion light year object (if it could be seen) is about 4200 Megaparsecs (about 1.3×10^{23} km or 9×10^{22} miles). At 100 Megaparsecs, the velocity of a receding galaxy (as the observed object) is ~7000 km/sec (from a plot of velocity versus distance for an H of 71 km/sec/parsec). The velocity at that distance would then be 7000 x 42 = 294000 km sec - almost the speed of light

Aside from quantum speculation, nothing is really known about the state of the Universe-to-be just prior to the initiation of the Big Bang (a moment known as the **Planck Epoch)**. The Laws and the 20 or so fundamental parameters or factors that control the observed behavior of all that *is* knowable in the Universe today become the prevailing reality at the instant of the Big Bang, but Science cannot as yet account for the "why" of their particular formulation and values, i.e., what controls their specifics and could they have come into existence spontaneously without any external originator, the "Creator" or "Designer". Among these conditions that had to be "fine-tuned" just right is this partial, but very significant list: homogeneity and isotropy of the Universe (the Cosmological Principle); relative amount of matter and anti-matter; the H/He and H/deuterium ratios; the neutron/proton ratio; the degree of chaos at the outset; the balance between nuclear attraction and electric repulsion; the optimal strength of gravity; the decay history of initial particles; the total number of neutrinos produced early on; the eventual mass density which affects the Critical Density; the specific (but varying) rates of expansion after the Big Bang; the delicate balance between Temperature and Pressure, both during the first moments, and much later during star formation; the ability within stars to produce carbon - essential to life; and much more. (See also another list at the bottom of page 20-11a.)

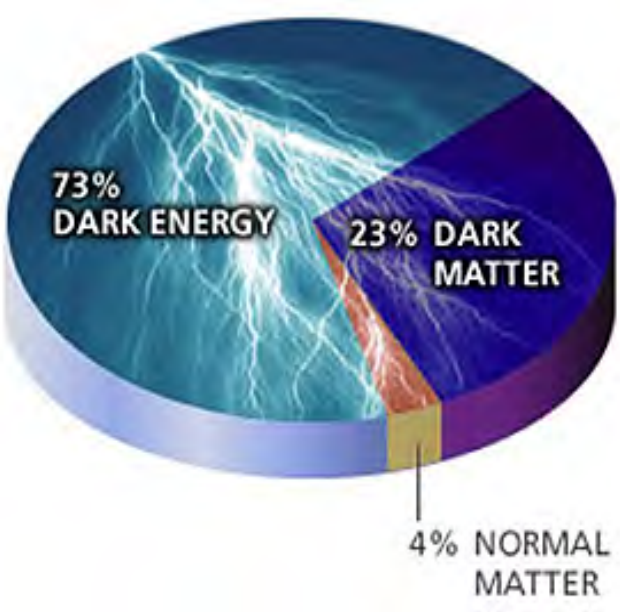
Some of these are interdependent but the important point is that **if the observed values of these parameters/factors were to differ by small to moderate amounts, the Universe that we live in could almost certainly not have led to conditions that eventually fostered intelligent life capable of evolving during the history of the Universe as we know it**. Also presumably necessary: beings that can attest to the Universe's existence and properties by making observations and deductions that lead to knowledge of the Universe. This requires the eventual appearance of "conscious reasoning" at least at the level conducted by humans on Earth, and perhaps also human-like creatures existing elsewhere in the Universe, - this concept is one of the tenets in what is referred to as the "Anthropic Principle" (page 20-11).

Moving on from the overview of the Universe afforded by the above paragraphs, we consider next the high temperature physics that pertains to what occurred in the vital first minute of the Big Bang.

The First Minute of the Big Bang

The Overall Composition of the Universe

Events in the first few fractions of a second, and subsequent time to the end of the first few minutes, of the Universe's existence have determined the nature and compositional makeup of the subsequent observed Universe of today. At present, that Universe consists of three prime components (in percentages determined from Cosmic Microwave Radiation data; WMAP Observatory; page 20-9), as shown in this figure:



Thus, ordinary matter (Fermions) is only 4% of all particles in the Universe. Although the two "Darks" comprise the remaining 96%, not much concrete knowledge has accrued about the nature of either Dark Matter or Dark Energy. Both presumably exist but proof of this is still lacking. Dark Matter is inferred to keep galaxies from flying apart. Dark Energy is inferred to be causing the renewed expansion of the Universe. Much effort is underway to prove their reality and determine their essential characteristics. Particle Accelerators are probing matter at high energies to gain a better understanding of their possible physical makeup. Dedicated cosmological observatories of the future may provide important insights.

This chart further subdivides the compositional makeup of the Universe:

COMPOSITION OF THE UNIVERSE

MATERIAL	REPRESENTATIVE PARTICLES	TYPICAL PARTICLE MASS OR ENERGY (ELECTRON VOLTS)	NUMBER OF PARTICLES IN OBSERVED UNIVERSE	PROBABLE CONTRIBUTION TO MASS OF UNIVERSE	SAMPLE EVIDENCE
Ordinary ("baryonic") matter	Protons, electrons	10^6 to 10^9	10^{78}	5%	Direct observation, inference from element abundances
Radiation	Cosmic microwave background photons	10^{-4}	10^{87}	0.005%	Microwave telescope observations
Hot dark matter	Neutrinos	< 1	10^{87}	0.3%	Neutrino measurements, inference from cosmic structure
Cold dark matter	Supersymmetric particles?	10^{11}	10^{77}	25%	Inference from galaxy dynamics
Dark energy	"Scalar" particles?	10^{-33} (assuming dark energy comprises particles)	10^{118}	70%	Supernova observations of accelerated cosmic expansion

Both matter and energy are said to consist of particles, such as atoms, electrons, etc. for the former and photons and bosons for the latter. There are an estimated billion photons for every atom in the Universe. The total mass of the Universe - consisting of matter particles and photon energy particles converted to mass ($E = mc^2$) is approximately 10^{53} kg. If elemental matter (dominated by hydrogen) is expressed as the number of atoms uniformly dispersed (redistributed from galaxies and stars) per cubic meter, the average density of atomic matter is (estimates vary) between about 2 and 6 atoms of H per m^3.

The Initialization of the Big Bang

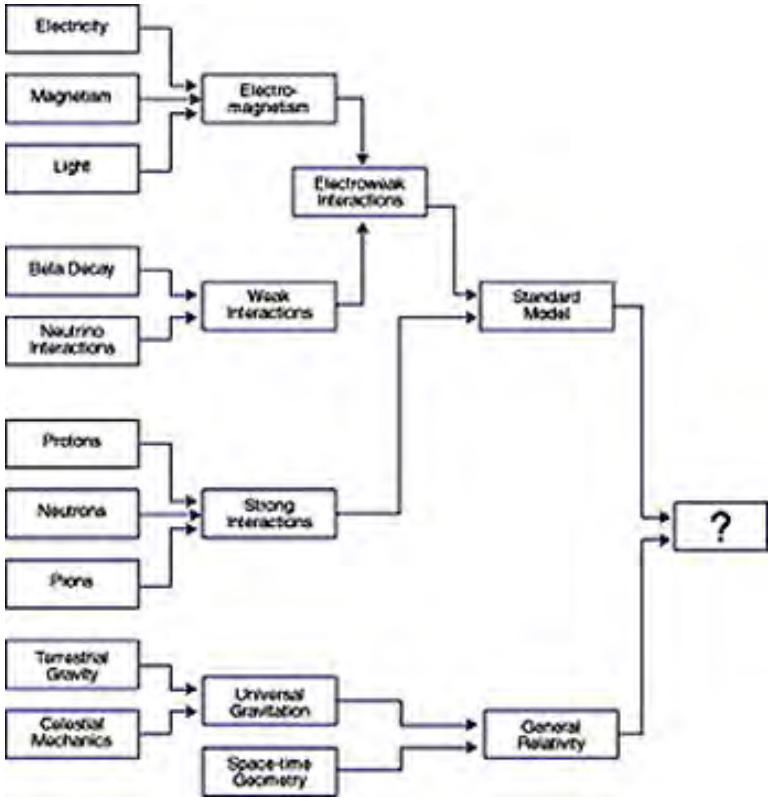
At the instantaneous moment of the Universe's conception, gravity, matter, and energy all co-existed in some incredibly concentrated form (but capable of supporting *fields* of action) that cannot be adequately duplicated or defined by experiment on Earth since it requires energy at levels of at a minimum 10^{19} GeV (Giga-electron volts; "Giga" refers to a billion; one electron-volt is the energy acquired by a single electron when accelerated through a potential drop of one volt; 1 eV = 1.602 x 10^{-12} ergs). A value of 10^{19} GeV is vastly greater than currently obtainable on Earth by any controllable process (presently, the upper limit obtained experimentally in high energy physics labs (with their large particle accelerators and colliders is ~10^3 GeV). Best postulates consider the Big Bang (whatever its origin) at its singularity instant to be governed by principles underlying quantum mechanics, have maximum order (zero entropy [see page 20-8]), and be multidimensional (i.e., greater than the four dimensions - three spatial and one in time - that emerged at the start of spacetime as the Big Bang got underway). Quantum theory does not rule out discrete "things" (some form of energy or matter) to have existed prior to the inception of the Planck Epoch; on the other hand, this existence is not required or necessary. But, as implied above and discussed in detail on page 20-10, "fluctuations" within possible energy fields in a pre-Universe quantum state (an abstract but potentially real condition that runs counter to philosophical notions of "being") may have been the triggering factor that started the BB.

This theory allows cosmologists to begin the Universe at a parameter called the Planck time , given as 10^{-43} seconds (what happened or existed at even earlier time - the Planck Era - is not currently knowable [but can be inferred] with the principles of physics developed so far). At that instant, the Universe which sprang forth must have been at least as small at 1.6 x 10^{-35} meters - the Planck length (about the same size as a string in superstring theory [see below]). At the initiation of the Big Bang, the four *fundamental forces* (gravity, and the strong [nuclear], weak [radioactivity], and electromagnetic [radiation] forces, referred to collectively as the Superforce) that held the Universe together co-existed (were unified) momentarily until about 10^{-36} sec) in a special physical state - GUT; see below - that obeyed the conditions imposed by one meaning of the term Symmetry***. During this fraction of a second interval, gravity then was as strong as the other forces. Its tendency to hold the singularity point together had to be overcome by the force that activated the Big Bang. The onset of fundamental force separation may have been tied to the force driving Inflation (see below).

(As a momentary aside: The Planck length, although exceedingly small, is nevertheless an allusion to dimensions. Dimension in turn implies space. It refers to the size of the Universe at the inception of the Planck Epoch. It is calculated in the framework of quantum cosmology using the speed of light c, the Planck constant h, and the Universal Gravitational Constant G as inputs. The Big Bang did two really important things 1) it released pent-up energy to power the expansion and creation of matter, and 2) it made the starting point for space into which particles emerged as energy cooled and these started to separate. From the first moment until now the main theme of the Univere's history has been the expansion of space and the redistribution and repositioning of its constituent particles.)

But gravity thereafter rapidly decreased in relative strength so that today at the atomic scale it is 2 x 10^{-39} weaker than the electrical force between a proton and an electron (according to one recent theory, gravity remained strong until about 10^{-19} seconds). However, since the forces between protons (positive) and electrons (negative) are neutralized (balanced) in ordinary matter, the now much weaker gravitational force is the primary one that persists and acts to hold together collective macro-matter (at scales larger than atoms, specifically those bodies at rest or in motion subject to and described by Newton's Laws; includes those aspects of movements of planets, stars, and galaxies that can be treated non-relativistically). And gravity has the fortunate property of acting over very long distances (decreasing as the inverse square law: $1/r^2$). Although we think of gravity as the most pervasive force acting within the Universe, there is growing evidence that some form of gravity-like force also resides within an atom's nucleus but extends its effects over very short (atomic scale) distances.

The non-gravity forces that separated from the gravitational force are described by the still developing Grand Unified Theory or **GUT**, which seeks to explain how they co-existed. The GUT itself is a subset of the Theory of Everything (**TOE**) which, when it is finally worked out, will specify a single force or condition (or, metaphysically, a state of *Being*) that describes the situation at the very inception of the Universe. Thus, TOE unites the gravity field with the quantum fields associated with the other forces that emerged as separate entities almost instantaneously at the start of the Big Bang. The TOE speculates on what may have existed or happened prior to the Big Bang, based on both quantum principles and belief that some other type of [pre-Bang] physics yet to be developed governed the pre-Universe void. At the Planck time, the four united fundamental forces make up the Unified Epoch. The flow chart below specifies the major components of each of the forces as they are assumed to exist after the first minute of the Big Bang; the arrows proceed timewise to more fundamental states at the very beginning. When unified at the outset of the Big Bang, they are presumed to exist in a state shown by the **?**, whose nature and properties are still being explored theoretically; at present this condition cannot be produced experimentally because of the huge energies (way beyond present capabilities in laboratories) involved; note that the involvement of Inflation is not considered.



The history of these forces during the first second of the Big Bang is discussed in more detail later on this page.

One model, now gaining some favor, based on Superstring theory (discussed near the bottom of this page) contends that at the first moment of the Big Bang (at the 10^{-43} sec arbitrary starting point), the Universe-to-be consisted of 10 dimensions. As the process of the Universe's birth starts, six of those dimensions collapse (but presently exist on microscales as small as 10^{-32} centimeters) and the remaining four (three spatial; one time) enlarged to the Universe of today.

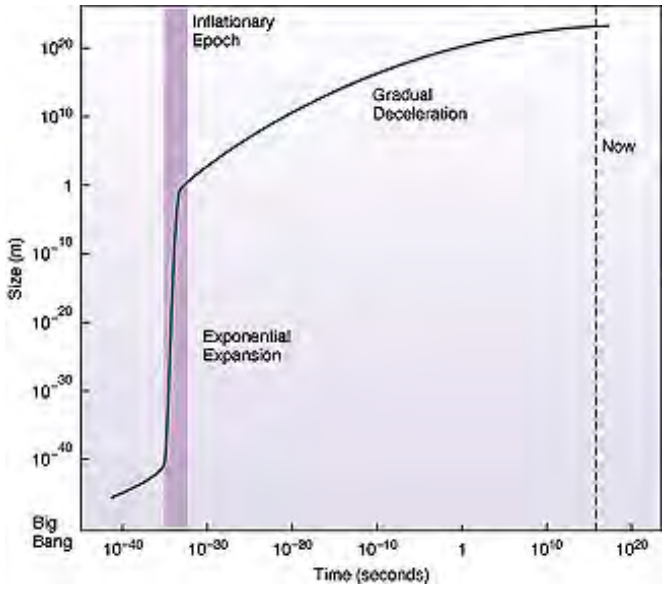
The behavior of these forces in the earliest moments of the Big Bang was critical to the construction and development of the Universe as we perceive it today. Gravity in particular controls the ultimate fate of the Universe's expansion (see below) and formation of stars and galactic clusters. (According to Einsteinian Relativity, gravity, which we intuitively perceive as attractive forces between masses, is a fundamental geometric property of spacetime that depends closely on the curvature of space, such that concentrations of matter can "bend" space itself; Einstein and others have predicted the existence of gravitational waves that

interact with matter; see the Preface for additional treatment). For all its importance, it is surprising that gravity is by far the weakest of the four primary forces. Its role in keeping macro-matter together and controlling how celestial bodies maintain their orbits is just that it becomes the strong, action-at-a-distance force left whenever the other forces are electrically neutral and have influence only out to very short distances.

The Inflationary Hot Big Bang Stage

The first fundamental change during the Big Bang is known as the GUT transition, occuring at 10^{-36} seconds, when the strong and electroweak forces began to separate. Between 10^{-36} and 10^{-33} sec (a minuscule but vital interval of time - about a billionth of a trillionth of a trillionth earth seconds - referred to as the **Inflationary Stage**), there is evidence of a critical mechanism that explains certain fundamental properties of the Universe. Inflation was first proposed by Alan Guth, then at Princeton University (now, at MIT), to account for some aspects of the Universe's growth [see below] that faced serious difficulties in the Standard Model. This has come to be known as the Inflationary Model for the Big Bang. The Inflation theory holds that the nascent and still minute Universe underwent a major *phase transition* (probably thermodynamic; water is a commonly cited example of such transition, as when it goes from liquid to solid [ice] or liquid to gas [steam]) in which repulsion forces caused a huge exponential increase in the rate of expansion of space. In fact, during the Inflation stage, the Special Relativity restriction which states that nothing can move faster than the speed of light was violated such that whatever was released during the Big Bang drew apart at rates greater than light speed. Through this brief moment of Inflation the micro-Universe grew from an infinitesimal size estimated to be about 10^{-28} meters (but still potentially containing all the matter and energy [extremely dense] that was to become the Universe as it is now) to that of a grapefruit or perhaps even a pumpkin (an upper size limit is given as a meter). This is an expansion factor that has not yet been agreed upon by cosmologists. A commonly cited value is 10^{78}, but estimates in the literature for the growth in this minute span of inflation range from about 10^{30} to as high as 10^{90}. To get a sense of the effect of such a factor, consider this analogy: the expansion is equivalent to increasing the size of the proton ($\sim 10^{-13}$ cm) to roughly the size of a sphere 10,000,000 times the Solar System's diameter (arbitrarily, taken as the distance from the Sun to the far orbital position of Pluto, or $\sim 5.9 \times 10^{9}$ km).

This extraordinary growth determined the eventual spatial curvature of the present Universe (in the most "popular" model, tending towards [or perhaps equal to] "flat"). It also accounts for the overall uniformity found in the Universe today because initial irregularities would be stretched so thin by the magnitude of inflation that they would act as though "invisible". This next diagram illustrates the extreme growth of the incipient Universe during the Inflationary moment (both horizontal and vertical scales are in powers of ten); in the version shown, the Big Bang expansion is shown as decelerating over time but a vital modification which restores acceleration expansion to the Universe's progression is discussed on page page 20-10.



From Astronomica.org

At the inception of Inflation, the temperature was ~ 10^{28} °K. Within this inflationary period, temperatures dropped drastically to ~10^{12} ° K by 10^{-10} sec. During this critical moment, the (preordained ??) physical conditions that led to the present Universe were established. The driving force behind this huge "leap" in size (which has happened at this extreme rate only once in Universe history)is postulated by some as a momentary state of gravity as a repulsive (negative) force (possibly the vacuum energy that is equivalent to Einstein's once-defunct Cosmological Constant but in a new form). Forces related to fundamental particles such the **Higgs boson** or the postulated "**Inflaton**" (see above) may have powered this tremendous expansion. Either particle is associated with a field (spatial region over which the force is said to operate; the force normally diminishes progressively with distance to the particle).

Still another hypothesis is that the energy was derived during the separation of gravitational force from the remaining three forces (see third diagram below). This may have released a huge amount of energy capable of bringing about the extreme repulsion that marks the brief moments of inflation (see paragraphs on page 20-10 that describe Einstein's Cosmological Constant which depends on a similar repulsive energy related to an as yet undiscovered but apparently real Dark Energy). (Recent discoveries indicate that the Universe is now undergoing a second but relatively much slower rate of accelerating expansion that has turned around the post Big Bang gravitationally-mandated deceleration, beginning at some [still undetermined] stage [probably prior to the last 7 billion years] of the Universe's growth; again, see page 20-10.)

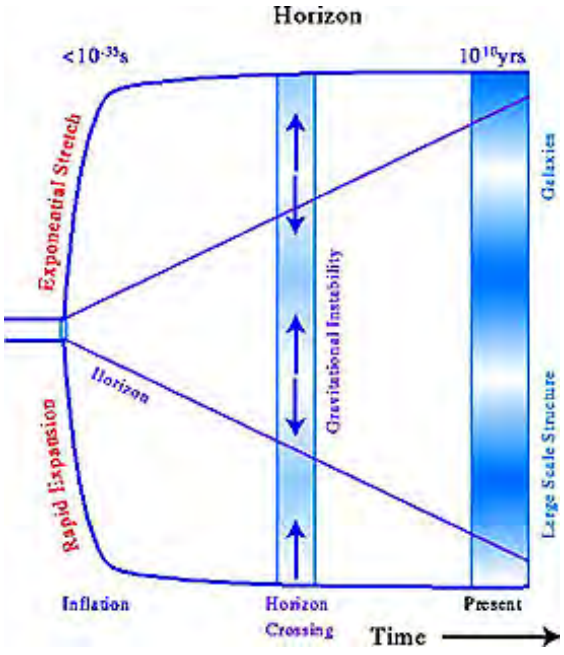
However, as of 2008 a newer explanation as to what powered the expansion has become ascendant. Since no matter yet existed, the cause of the Inflation state is postulated to be a quantum fluctuation in an extremely dense vacuum. The source of the vacuum energy that drove Inflation has not been precisely identified but theory ascribes this to potential energy within the quantum field of the Inflaton. During this very brief inflationary period, the "empty" surrounding Cosmos (the "beyond", where only energy existed but into which virtual particles come and go but usually fail to initiate any surviving Big Bang event) is entered by the activated particle at a rate greater than the speed of light. Specifically, a metastable state called the *false vacuum* - devoid of matter *per se* but containing some kind of energy - underwent a decay or phase change by quantum processes to a momentary energy density that produces the negative pressure capable of powering the inflation. Inflation continues until the false vacuum potential (which starts out as positive when its associated density field is zero), which initiated the expansion, drops to zero.

During inflation, as gravity began to act independently, *gravitational waves* were produced that had a critical bearing on the minute but vital variations in distribution of temperatures (and matter) in the subsequent history of the Universe as we know it. As time proceeded, gravity then

reverted to the attractive force that took over control of further expansion. The Universe thereafter has a positive density field that varies in space and time. Inflation probably ended at 10^{-34} sec, after which the vacuum energy transitioned into real energy. Thereafter, particles began to appear as the Standard Hot Big Bang per se commenced.

Advantages of the Inflationary model are that it sets the stage for the "creation" of matter, it accounts for the apparent "flatness" of the Universe's shape, and it helps to explain its large-scale homogeneity and isotropy (smoothness). Before the Inflation began this uniformity condition (homogeneity) existed, with the initial conditions in causal contact, and was subsequently "frozen in" to the Universe by the rapidity of inflationary expansion. Thus, prior to the moment of Inflation, all parts of the incipient Universe were in communication with one another and their properties were coordinated. But, as Inflation proceeded, in which the components moved apart at a much faster rate that exceeded the speed of light, this communication was lost momentarily, whereas the previous uniformity was largely maintained during this super-expansion. Since then the components of the Universe (e.g., galaxies) in some parts of the Universe have been regaining communications with other parts. However, the model suggests that during inflation, energy may not have been perfectly uniformly distributed, producing narrow zones of greater concentration called "cosmic strings". These, during the following slower expansion, served as the irregularities which eventually led to concentrations of matter that localized into the early Universe structure around which the first galaxies formed. These irregularities may have been quantum fluctuations. The slight departures from homogeneity also show up in the variations detected in the Cosmic Background Radiation (CBR, see page 20-9).

Inflation also seems to solve the above-mentioned "horizon problem" (recall that in one meaning *horizon* refers to regions of the Universe that are limited in their interactions [causal contact between regions] by the distances photons can travel at light speed during the interval of time in which a cosmological phenomenon is being considered). This problem is illustrated by this diagram:

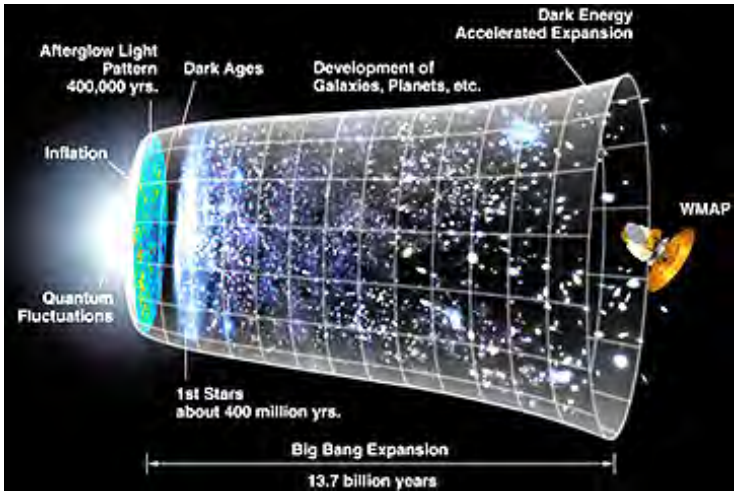


In this diagram parts of the Universe seem to lie outside these horizon limits (in purple). Such distant parts are not now in contact with one another (do not exchange light signals) and would seem causally independent. The Inflation model gets around this by 1) assuming these and all parts were in contact in that miniscule fraction of the Universe's first second before Inflation, and thus 2) had inherited, or "locked in" the co-ordinating physics underlying the Universe's operations that subsequently preserved general uniformity as the Universe went through its huge inflationary expansion.

Inflation also has been invoked to account for the possibility of matter and energy, perhaps even

galaxies, being distributed beyond the 13.7 billion year horizon that limits what telescopes could potentially observe. Since the very earliest constituents of the first moments of the Big Bang were traveling faster than light speed, some of these constituents were flung beyond what cosmologists are restricted to in their observations.

A good summary of the essence and history of Inflation is at a Web site prepared by John Gribbin. This next diagram serves to show (conceptually) how inflation affected all that followed; it also previews much of what will be discussed later on this page.. Several variants of this diagram will appear throughout Section 20. Note the very large increase in the Universe's size shown on the left, followed by expansion at a decreasing rate, and then slow but steady reversal leading to an increasing rate (which occurs when the vacuum energy [negative or dark energy] of the Universe begins to exceed the positive gravitational energy), shown on the right.



Although theoretical calculations and certain experiments seem to be confirming the essential points in the Inflation model, not every cosmoscientist has come to accept this innovative explanation of the earliest moments of the Universe and the consequences of its subsequent history that inflation seems to predict. In the past few years, some have turned their attention to alternate models. Most striking in its departure is the Varying speed of Light (VSL) model first espoused by Dr. Joao Magueijo in 1995, who later joined forces with Dr. Andreas Albrecht when they collaborated at the Imperial College in London. The essence of VSL is that during roughly the same time in the first BB second that Inflation would have operated, at this earliest moment the intense energy being released would cause the speed of light to be greater than today's value. That speed, ever decreasing, would then converge on the now constant value today, thus meeting Einstein's fundamental posit that this speed is constant. Magueijo and Albrecht have calculated that this phenomenon of rapidly dropping speed in these early instances can produce most of the same outcomes that the spatial expansion of Inflation leads to. Initially largely rejected by his colleagues, recent observations of possible light speed changes in the post BB Universe, if confirmed, have refocused attention on VSL. Like Inflation, VSL remains hard to prove since its essential characteristics occur under physical conditions that are still near-impossible to duplicate experimentally.

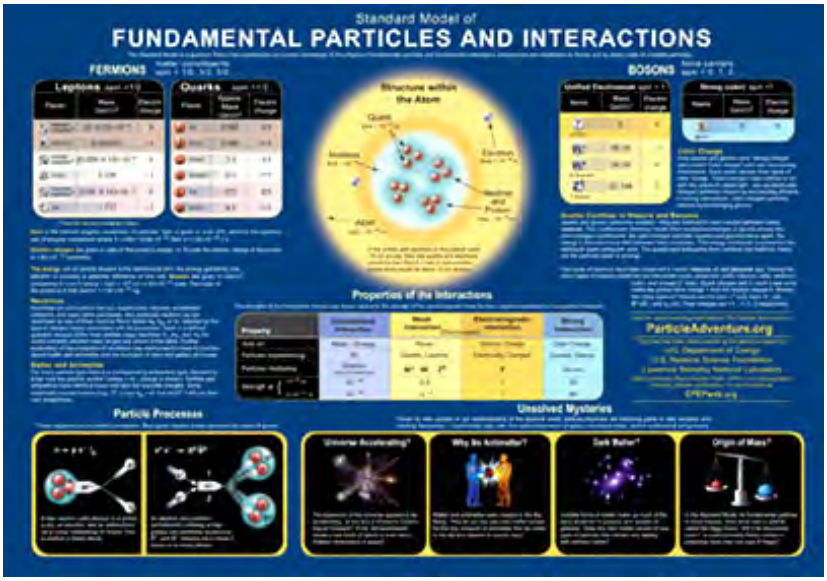
The Subsequent Standard Hot Big Bang Stage; the First Particles

After the Inflation interval, the behavior of the expanding Universe in the first minute reverts to the Standard Model. Matter was created during this time; this matter makes up two classes - the **Fermions** (*Hadrons* and *Leptons*), and **Bosons** - force particles that interact with the Fermions.

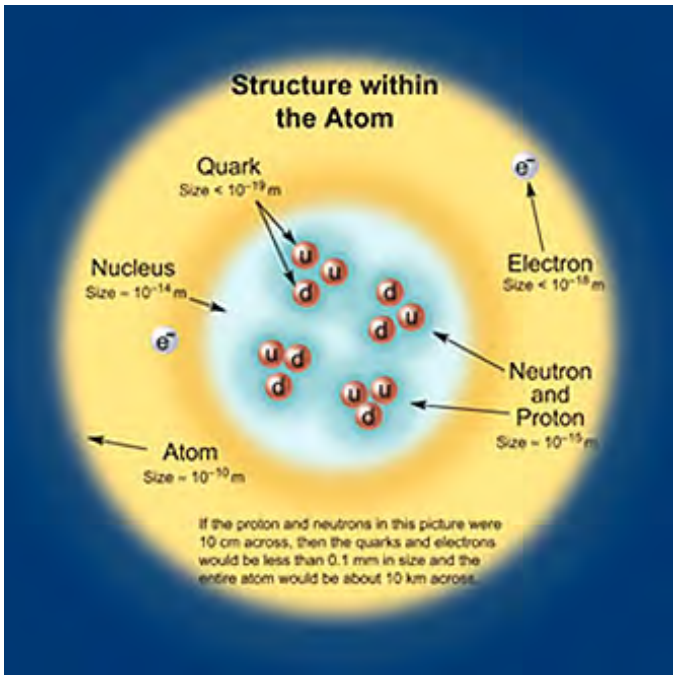
The Hadrons are made up of smaller particles called quarks (which themselves may consist of even smaller particle called superstrings [discussed near the end of this page]). From 10^{-34} sec to 10^{-5} sec the first quarks (see below) co-existed with particles called gluons. A condition called Supersymmetry (discussed near bottom of this page) prevailed but underwent change; dark

matter particles (page 20-9) may have been created in this interval.

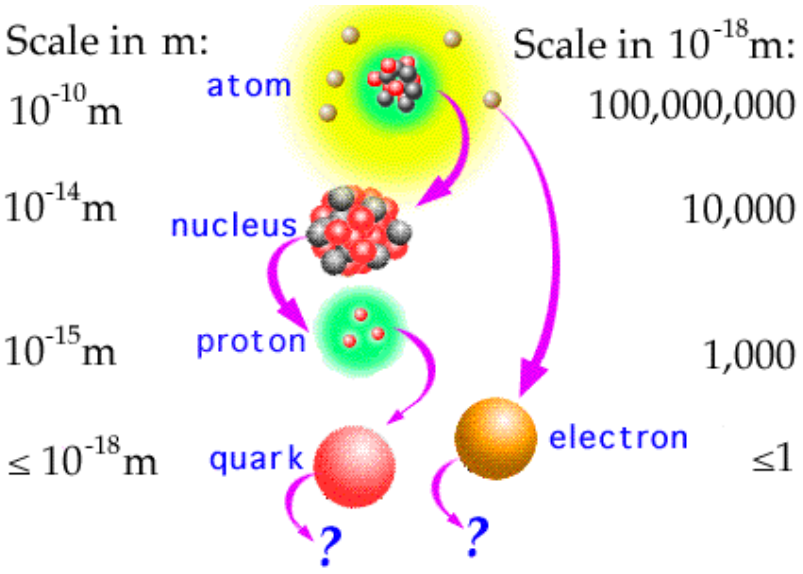
Before we continue to unfold the first minute, in the next paragraphs, we need to discuss the nature of matter and the interactions between matter and energy. One major source of information on the Fundamental Particles making up matter, and of several illustrations below, is found at the Contemporary Physics Education Project website. So that you can recognize it, the coordinating chart that comes up after you click on this link is reproduced here, even though unreadable (working with the website version can provide a satisfactory size):



These particles participate in the makeup of what is called an **atom**. The structure of an atom is shown next; discussion of the particles making up atoms then follows:



As we shall shortly note, atoms are composed of particles whose sizes range widely, as summarized in this scale diagram:



As described above, during the first fraction of a second following the Planck moment incredible events unfolded in rapid succession that led to release of kinetic energy that powered the Universe's development and created the initial stages of radiation. Matter was formed as condensed energy (an $E = mc^2$ transformation)(in the first minute some of the matter decayed back into radiation, releasing neutrinos and other particles). These primitive components of matter (or more properly, *mass*[****]) rapidly organized into a myriad of elementary particles. Examine these diagrams:

FERMIONS matter constituents spin = 1/2, 3/2, 5/2, ...

Leptons spin =1/2

Flavor	Mass GeV/c²	Electric charge
ν_L lightest neutrino*	$(0–0.13)\times10^{-9}$	0
e electron	0.000511	−1
ν_M middle neutrino*	$(0.009–0.13)\times10^{-9}$	0
μ muon	0.106	−1
ν_H heaviest neutrino*	$(0.04–0.14)\times10^{-9}$	0
τ tau	1.777	−1

Quarks spin =1/2

Flavor	Approx. Mass GeV/c²	Electric charge
u up	0.002	2/3
d down	0.005	−1/3
c charm	1.3	2/3
s strange	0.1	−1/3
t top	173	2/3
b bottom	4.2	−1/3

Baryons qqq and Antibaryons $\bar{q}\bar{q}\bar{q}$

Baryons are fermionic hadrons.

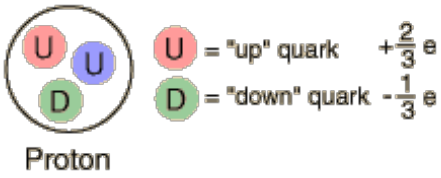
These are a few of the many types of baryons.

Symbol	Name	Quark content	Electric charge	Mass GeV/c²	Spin
p	proton	uud	1	0.938	1/2
$\bar{p}$	antiproton	$\bar{u}\bar{u}\bar{d}$	−1	0.938	1/2
n	neutron	udd	0	0.940	1/2
Λ	lambda	uds	0	1.116	1/2
Ω^-	omega	sss	−1	1.672	3/2

Consider this further information:

I) the ***FERMIONS***: all particles with quantum spins of 1/2 of odd whole numbers such as 1, 3, 5 (includes protons, electrons, neutrons); they all obey the Pauli Exclusion Principle which states that no two different particles can have the same values of the four quantum numbers QN; these are [1], Principal QN "n"; [2] Azimuthal QN "l"; [3] Magnetic QN "m"; [4] Spin QN 's"). Fermions can be divided into subgroups:

1) The heavier **Hadrons** (minute particles, consisting of certain **QUARK** combinations held together by *Gluons* (massless force particles) permitting strong interactions within atomic nuclei), further subdivided into (a) the **Baryons** (combinations of three quarks [see 4th paragraph below on this page] that include the familiar protons and neutrons (each about 10^{-13} cm in size [compared with diameters on the order of 10^{-8} cm for the classical Bohr atom]) and (b) the **Mesons** (short-lived heavier particles) families. The two most familiar Baryons are the proton and the neutron, made from up and down quarks (the other 4 quarks do not occur in Hadrons but have been discovered as products released during particle accelerator experiments.) These diagrams depict the quark makeup of a proton and a neutron:

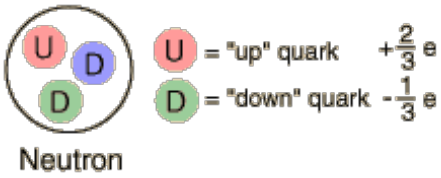


$m_p = 1836.15\ m_e$

Mass = 1.6726×10^{-27} kg

$= 938.27231\ MeV/c^2$

$= 1.00727647\ u$



$m_p = 1838.68\ m_e$

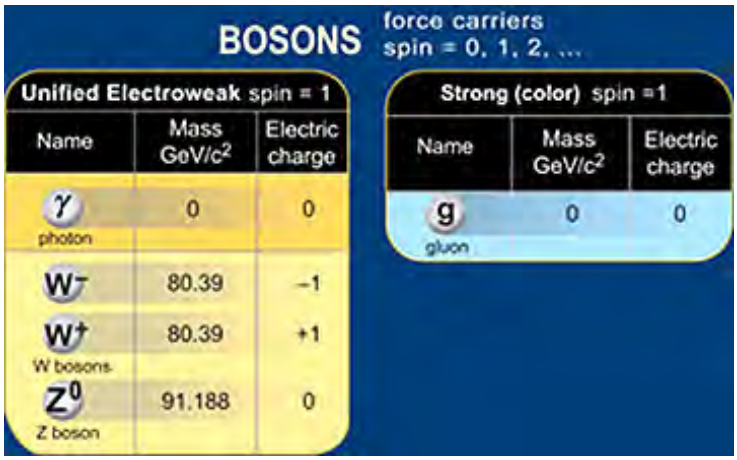
Mass = 1.6749×10^{-27} kg

$= 939.5656\ MeV/c^2$

$= 1.0086647\ u$

2) The **Leptons**, even tinier discrete particles that are weakly interacting. They are represented by electrons, tauons, muons, and three types of neutrinos (electron-neutrino; tau-neutrino; muon-neutrino; the discovery of the latter two imply that the neutrino may have a small mass, and if proved could account for some of the missing matter in the Universe talked about later in this Section.

II) ***BOSONS***, the force carrying messenger particles; these have unit [1] spins. Best known of the bosons are the 1) **photons** (which have zero rest mass) that are quanta ***** of radiant energy responsible for electromagnetic (EM) forces which travel at light speed as oscillatory (sinusoidal) waves and 2) the **gluons** that bind the nucleus by mitigating against the strong repelling forces therein. A boson that theory says exists, but as yet has not been "found" is 3) the **graviton**, which transfers the force of gravity (also, at the speed of light). This chart summarizes bosons:



Much of the above information is summarized in the chart below. This classification of particles and their interactions is an integral part of the Standard Model for the ways in which matter is

put together, which applies to any Big Bang scenario (without the refinements of Inflation) that leads to a broadly homogeneous, isotropic large-scale Universe and is an acceptable summary of what is verifiably known now about the origin of matter and energy (with the caveat that the model is subject to continual modification or revision).

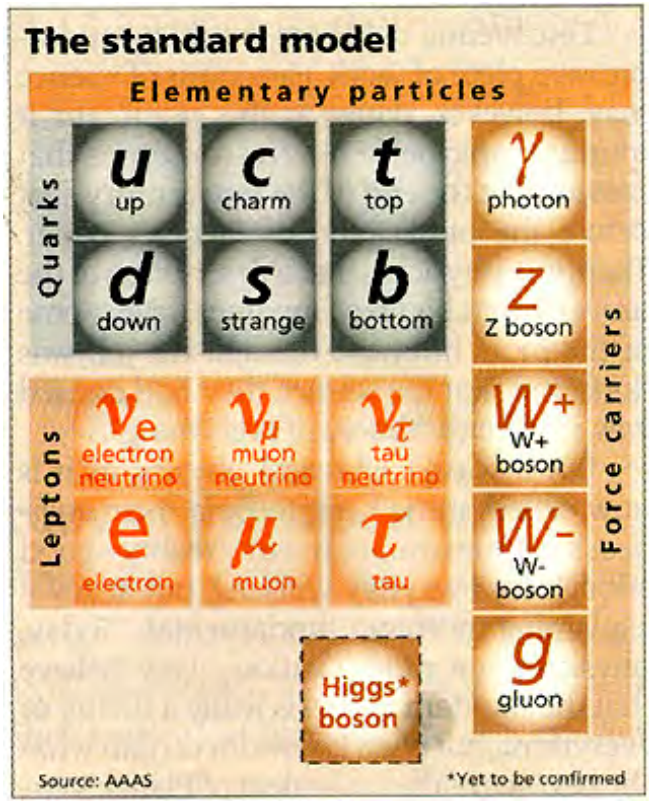
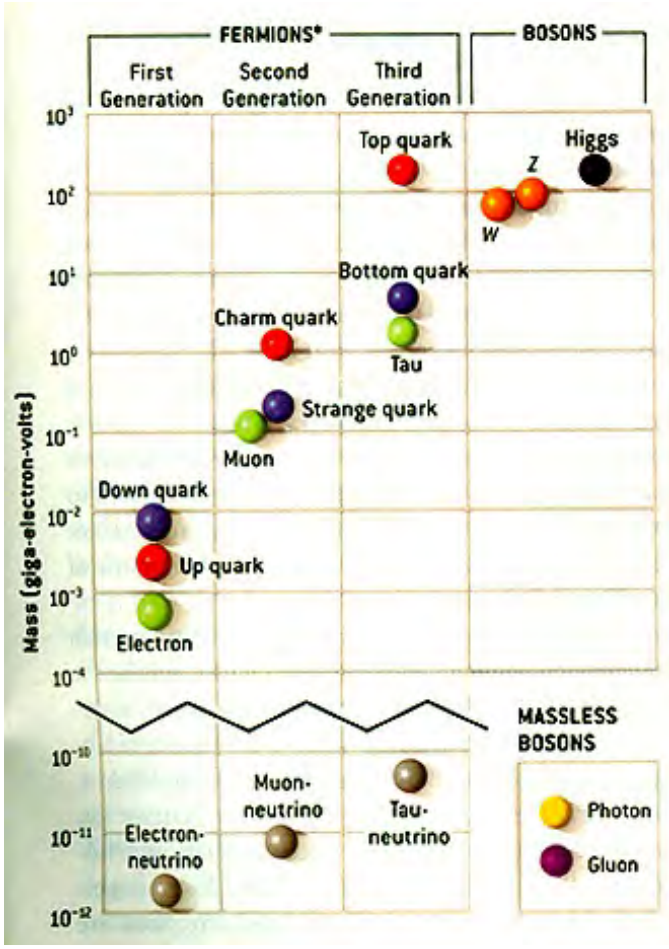


Illustration produced by AAAS, taken from The Economist, Oct. 7-12, 2000, p. 96

In this classification, the major entities are the three families: Quarks (gray), making up the Fermions, and including Baryons; Leptons (orange), and Bosons (brown). The quark particles have generally been discovered and proved to exist from high energy physics experiments using particle accelerators.

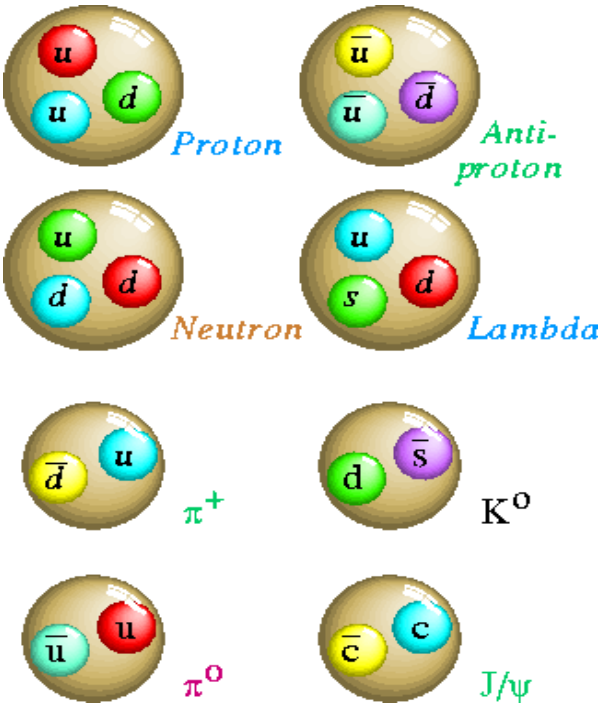
A variant of this classification which arranges the mass and force particles according to measured or estimated mass of each type of particle is shown below. In this version, Leptons are considered as classified within a more general category of Fermions. The chart emphasizes the growing belief that mass itself is governed by the relative contribution from the Higgs Boson (see footnote 4 (****). The different masses arise from different interactions of fundamental particles with the Higgs field (of which there may be different sets with different values and properties),



From *The Dawn of Physics Beyond the Standard Model*, by Gordon Kane, Scientific American, June 2003

Quarks were the first (sub)particles to form during the early moments of the first minute. The nomenclature for the 6 quarks are descriptive terms for convenience and carry no special physical significance. There are six types or "flavors" (Up, Strange, etc.), each subject to variants or "colors"; various combinations of quarks give rise to the different nucleons). Quarks have a baryon number of +1/3, charge numbers of +2/3(up) and -1/3(down), and a spin quantum number of 1/2. The two Baryons familiar to most are made of three Quarks: the *proton* consists of two up (each +2/3) and one down quark (-1/3) for a net charge of 1; the *neutron* two down and one up quark, for a net charge of 0 (zero). *Mesons* contain only two Quarks.

Quarks also can have a reverse sign, thus they can organize into anti-protons and anti-neutrons. Other combinations of Quarks lead to more exotic particles; one group includes mesons, which include members such as the pion Π^-, consisting of an anti-up quark (-u) and a (d) quark and the kaon K^+ made up of a (u) and an (-s) quark. Some of the particles made from quarks apparently exist only when created in particle accelerators. The next diagram and the chart below it document such particles (anti-quarks have a - on top of the letter:



MESON = $q\bar{q}$		quarks	electric charge
π^+	pion	$u\bar{d}$	+1
K^-	kaon	$S\bar{u}$	-1
K^0	kaon	$d\bar{s}$	0
ρ^+	rho	$u\bar{d}$	+1
D^+	D	$c\bar{d}$	+1
η_c	eta-c	$c\bar{c}$	0

According to a Higgs field-based model, protons and neutrons derive their mass from the energy of motion of the quarks and gluons flying within the Higgs field (which, according to theory, pervades the Cosmos) around the nucleus.

The Leptons have much smaller masses and are single particles (not containing the quark subparticles). They are not influenced by the strong nuclear force but can interact through the weak nuclear force. Three of the Leptons (upper row) are neutrinos which have extraordinary penetrating power (one can pass through the entire Earth without interacting or changing); once thought to be massless, evidence now suggests a very small mass.

The force particles (Bosons) are involved with the individual fundamental forces mentioned above. For example, the gluon holds the nucleus of baryons together; Z and W Bosons control the weak nuclear force; photons are the force carriers that are associated with electromagnetic radiation; gravitons transmit the force of gravity. The Higgs Boson has not yet actually been proved to exist (but from theory is considered almost certainly to be real); recent experiments in a European supercollider may have witnessed a few genuine Higgs particles but confirmation will likely await several new supercolliders capable of much higher energies due to come on line before the end of the first decade in 2000. The Higgs Boson is considered to be the force particle that accounts for mass in the fundamental particles that have that property.

Little has been said so far about neutrinos. But these are particles of great importance. They reside in the nucleus and are released during both radioactive decay and nuclear fusion. They are one of the four fundamental particles. They are without electric charge (with an energy of 0.001 eV) (unique in this respect; neutrons, also chargeless, actually consist of a combined

proton [+] and electron [-]). Once thought to be massless, much like the photon, experiments have now shown that they travel at less than light speed and hence possess some minute mass (any particle traveling at the speed of light is without mass, i.e., all its mass is converted to energy). They are super-abundant (~100 million of them for every atom in the Universe); more than a trillion pass through your body at any one second. They are ubiquitous: they are a main constituent of cosmic rays; they are produced during the Sun's persistent fusion; they even form by collisions within the atmosphere. Yet, they pass through the full Earth almost unimpeded, rarely interacting with nuclei.

Neutrinos were first proposed in 1936 by Wolfgang Pauli of Quantum Mechanics fame. But their existence was first confirmed in 1956 through an elaborate experiment by Dr. Ray Davis who built a detector consisting of a large vat of cleaning fluid set deep within a South Dakota gold mine. Extremely rare collisions with chlorine in the fluid produced argon. Davis consistently came up with a tiny number (3 argon atoms in a set time span) that was at odds with the model proposed by Dr. John Bahcall of Princeon based on his model of thermonuclear fusion in the Sun, who detected 10 per unit time. The discrepancy was not resolved for decades until it was realized that the neutrinos, which had the three "flavors" mentioned above, actually underwent a transition process (called oscillation) of the original ν_e or electron neutrino (the only flavor produced in solar fusion) into tau and muon neutrinos, all three in about equal proportions. Since the Davis detector only could count electron neutrinos, the discrepancy was explained. A subsequent detector, a deuterium water vat in a Sudbury, Ontario nickel mine, has proved capable of counting all three flavors, so that both Davis and Bahcall were right insofar as their predictions and counting difference are now explained.

The importance of neutrinos is still not fully understood. They are essential in accounting for energy involved in nuclear processes, so that the Law of Conservation of Energy is sustained. Their role in the Big Bang, and before it, is still being investigated. They may have pre-existed the BB and may be one of the virtual particles. There is growing evidence that a type called the "sterile neutrino" (that is not influenced by the weak force) is involved (and may be the main constituent) in what is known as Dark Matter (see page 20-9). They also are possibly important in determining the imbalance between matter and anti-matter. They are important in the high temperature processes of the initial minutes of the Big Bang because they are factors in some of the possible reactions, especially in the formation of helium, and thus helped to determine the relative abundances of H, He, Li, and Be - those elements that mark the initial composition of the material Universe. The detection of neutrinos (which cannot be measured directly by instruments since they have no charge) is one of the great "detective stories" successes of Physics. Suffice to say that neutrinos, like photons, pervade the Universe from the beginning to the present. After the first second of the Big Bang, they effectively decoupled from other particles and presumably co-exist with the photon Cosmic Microwave Background Radiation as Cosmic Neutrino Background.

Mention is made at this point of a particle, the Tachyon, whose existence has never been proved. Tachyons are theoretical quantum particles which can travel faster that the speed of light (several other conceptual conditions allow for circumstances in which light speed can be exceeded. Their attractiveness is that, should they exist, time travel is made more efficient. To learn more, visit this Wikipedia website.

The Standard Model is itself a great achievement in Astrophysics. When examined rigorously, it is now considered only an approximation to full reality in subatomic physics. It fails, for example, to **explain and integrate gravity**. Attempts to revise it are leading to what is called the Supersymmetric Standard Model (SSM), mentioned again below. Theoreticians believe that gravity must have its own Boson which they have named the *Graviton*. Although it most likely exists in some form, its actuality has yet to be proved. It has not yet been found during any of the current particle accelerator experiments (which are also looking for the Higgs Boson, but none now operating may be powerful enough).

Now, returning to the events of the first minute: The history (pattern) of force dissociation during

the first second is depicted in this illustration:

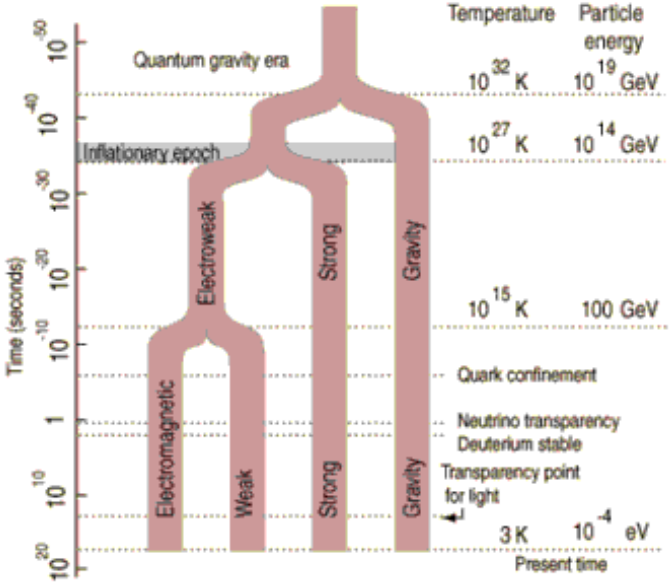
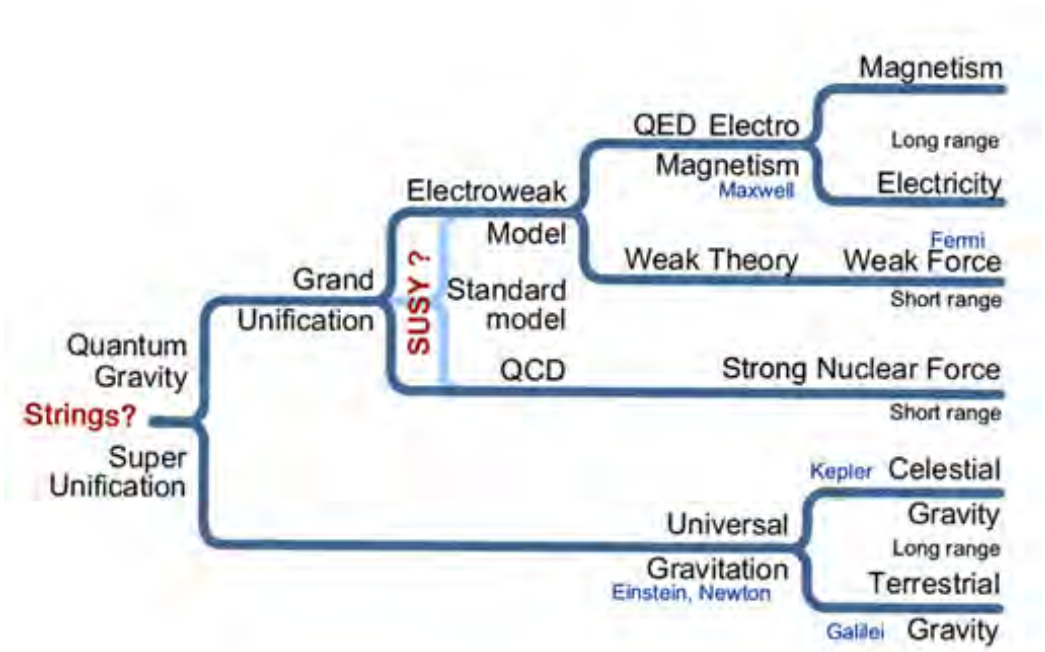


Figure 1: Schematic depiction of the decoupling of the unified force into our current 4 fundamental forces (Credit: Federation of American Scientists).

This next diagram says much the same thing but breaks the forces down into more detail.



By ~10^{-35} sec there was a fundamental *symmetry break* that brought on a split between the GUT forces and the other fundamental force known as gravity, dependent on the Graviton. (Symmetry is a general term that applies to quantities that remain invariant under specified transformations.) At 10^{-35} second there was a further split of non-gravitational forces into the Strong and the Electroweak (combination of Weak [responsible for radioactivity] and Electromagnetic) forces; the Electroweak pairing then separated into today's EM and Weak forces at about 10^{-10} sec. From 10^{-35} to 10^{-6} sec, matter consisted of the subatomic quarks (**Quark Era)**, and their binding particles, the gluons, present but not yet involved in producing nucleons (protons, neutrons). Temperatures were still too high (10^{28} °K) to foster quark organization into these nucleons. By the start of this interval, at the time when energy levels

dropped to about 10^{-16} GeV, the GUT state underwent dissociation into the Strong nuclear force (binding nuclei) and the Electroweak force. At about 10^{-9} sec, by which time temperatures had fallen to ~10^{15} K, the Weak nuclear force (involved in radioactive decay) and the Electromagnetic (EM) force (associated with photon radiation) separated and began to operate independently. Then, by 10^{-5} seconds, the six fundamental Quarks had organized in combinations of 2 or 3 into hadrons during the brief **Hadron Era.**. This era lasted only until about 10^{-4} seconds. There was almost total proton-antiproton and neutron-antineutron annihilation, releasing huge amounts of photons. Surviving protons formed by this time remained stable but some neutrons produced later experienced decay into protons and electrons. This Era was followed at 10^{-4} seconds, lasting up to one second or so, by the emergence of electrons, neutrinos and other Leptons (**Lepton Era**) which persisted until about 10 seconds. Thus, prior to 10^{-5} seconds, quarks had formed almost exclusively, but by the end of the first second of time they were greatly reduced in number as free (unorganized) particles, even as Hadrons, Leptons (especially neutrinos) and photons (the particle carriers of electromagnetic energy) were becoming the dominant products. Until about 10 seconds there was also some electron-positron annihilation. As electrons emerged, some reacted with protons to form neutrons, releasing neutrinos. From this point on, the ratio of Baryons to photons is 1 to a billion (a similar number holds for the ratio of Baryons to neutrinos).

As indicated above, from the GUT stage onward, both matter and antimatter were being created (baryogenesis) simultaneously. By 10^{-4} sec both quark particles and antiparticles (with opposite charges, e.g., at the Lepton level an anti-electron or positron would have a + charge) that had earlier coexisted had now interacted by mutual annihilation. Neutrinos and antineutrinos released by proton-electron reactions also experienced this destruction. So, at this moment only a residue of elementary particles survived - (almost?) all antiparticles apparently were completely wiped out leaving only some of the numerically larger amounts of particles. This annihilation left the particles that make up the Universe today in slight excess over the completely depleted antiparticles (since the Big Bang quantum theory allows antiparticles to be produced in the vacuum of space but these are eventually destroyed by encounters with particles).

Annihilation is an extremely efficient process for generating the profuse amount of energy released when positrons and electrons meet - destruction of a pair produces 10^6 electron volts. During the annihilation phase, a great quantity of high energy gamma ray radiation and other energetic photons resulting from the matter-antimatter interactions came to dominate the particles in the incipient Universe. These photons have survived as the Cosmic Background Radiation.

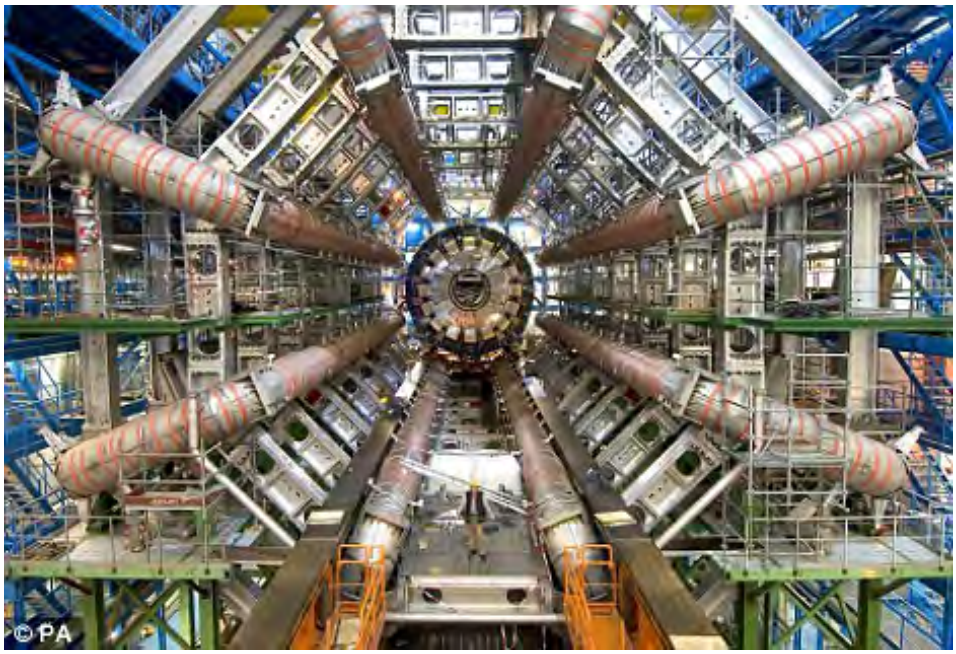
By 10^{-3} seconds, the temperature had now dropped to 10^{14} K and the proto-Universe had a diameter roughly the size of our present Solar System. In the next few seconds, temperatures dropped below a level where further antiparticle production took place in abundance. The particles making up the Universe today represent the excess over the very few surviving antiparticles. Most of the latter would have concentrated in near empty space outside any cluster of matter (the stars, galaxies, gas clouds, etc.) - if antiparticles still co-exist in significant amounts in conjunction with the particles we deal with on Earth or in the denser cosmic world, the effects of destruction might be detectable; no evidence that this is going on to a noticeable degree has been found.

At the one second stage, the Universe had already expanded ****** to a diameter of about 1 to 10 light years even as its density had decreased to ~10 kg/cc (kilograms per cubic centimeter), and its temperature had dropped to about 10^{10} K. By this time all the fundamental particles (essential matter) now in the Universe had be created, largely from the vast quantities of photons (energy "fuel") released during the first second. As of the first minute, about 1 free neutron existed for every seven protons, although all of these neutrons would eventually combine with protons in isotopes and heavier elements. The general excess of protons persisted, making those Hydrogen atom nuclei then and still the most prominent atomic species

in the Universe.

Much of what is known about events, conditions, and sequences during the first minute of the Universe has been surmised from theoretical hypotheses and calculations together with high energy experiments. One focus is on the strong force between Quarks, examined using the theory of Quantum Chromodynamics (QCD). Experimental verification of Quark behavior requires building of Particle Accelerators or Colliders that are major undertakings in terms of costs and technical "know-how". In February, 2000 an announcement from CERN in Geneva claims to have reproduced conditions equivalent to the first ten microseconds (10^{-6} sec) of the Big Bang. In effect, a "mini-Big Bang" is generated that lasts for much less than a trillionth of a second. To achieve this, the SPS-CERN Accelerator hurls Lead atoms (heavy) in a beam that strikes Lead or Gold targets at tremendous velocities. Momentarily, temperatures at the collision point reach 100,000 times that of the Sun's interior (~1.5 billion °C), at which the physicists interpreting the experiment postulated production of a plasma (electrically-charged "gas") emanating from the contact zone that is composed, for a very brief instant, of Quarks and Gluons. These quickly recombine into protons, neutrons, and electrons as the heated material dissipates. Other more recent experiments have carried conditions back to 10^{-14} second. Energies comparable to those extant during the first moments (earlier than 10^{-14} sec) are so great that no appropriate experimental setup is feasible for the foreseeable future, and may never be attainable in physics labs on Earth.

New colliders, generating at least 10 times more energy, are coming on line since 2000, so that relevant new experiments will likely confirm the theoretical models that describe the history of the later part of the first minute. Now active is the Relativistic Heavy Ion Collider (RHIC) operating at the Brookhaven Laboratory on Long Island. Using high speed Gold nuclei that are driven in opposite directions, collisions now duplicate the conditions representing the first microsecond. At that moment it now appears that the quark-gluon "soup" was liquidlike rather than a gas plasma. Some parameters at this moment are: Energy released = 20000 GeV; Temperature = approaching 6 trillion °C; Pressure at impact = 10^{30} atmospheres; Duration of event = 10^{-23} sec. In 2008, the Large Hadron Collider came online, operated briefly, but had to shut down for an indefinite period because of a peripheral hardware malfunction involving a few of its magnets. An idea of its size is given by the photo below. More about the LHC, the world's largest, is found at this Wikipedia website.



There is one fundamental topic that we've mentioned only in passing on this page: the two main categories of constituents of the physical Universe - Dark Matter (23%) and Dark Energy (73%) - which together with Ordinary Matter (4%) comprise all that is present in the observable Universe. Suffice to say here that the current distribution, percentage-wise, of all these

categories was established during the first minute of the BB. The treatment of the two "Dark" types is deferred until page 20-9 and 20-10 in order to develop other cosmological concepts needed to plumb the meaning and implications of the categories.

We close this part of the page by commenting on some other topics in Big Bang expansion. Newer models treating aspects of the physics and mechanisms of expansion during the first fraction of a second of the Big Bang have been proposed (see below) and the theory behind each is currently being tested experimentally. We will cite and briefly describe three of the most intriguing at the moment, but will forego any in-depth explanation:

1) *Primordial Chaos*: which postulates that in the earliest stages of the Big Bang the distribution and behavior of matter and energy in the incipient Universe was notably disordered and inhomogeneous, irregular, and turbulent, with variations in temperature and other scalar (non-directional) properties, anisostropic expansion rates, and other disturbances in the initial conditions within various parts of the rapidly changing microverse (a variant, called the Mixmaster model, considers the expansion to oscillate into a few momentary contractions at the outset). As the Universe grew both during Inflation and afterwards, these irregularities were smoothed out, leading to the gross isotropy of the present Universe; one version assumes a cold rather than very hot initial state;

2) *Supersymmetry*: a symmetry property which states that for every Fermion (quantum spin of 1/2) there must be a corresponding force-carrying Boson (quantum spin of 1), called a *sparticle* of the appropriate kind; likewise each Boson has a corresponding Fermion sparticle; thus, in this model the number of particles that exist is doubled; the concept predicts that there must be some subatomic particles still to be discovered if this pairing is valid); these supersymmetric particles are heavier than normal particles and may be a major constituent of Dark Matter (page 20-9). The Supersymmetry Standard Model (SSM) is proving to be a plausible means for simplifying the broken symmetry problems that beset the Standard Model;

3) *Extra Dimensions*: Such as those associated with Superstring theory; (discussed near the bottom of this page).

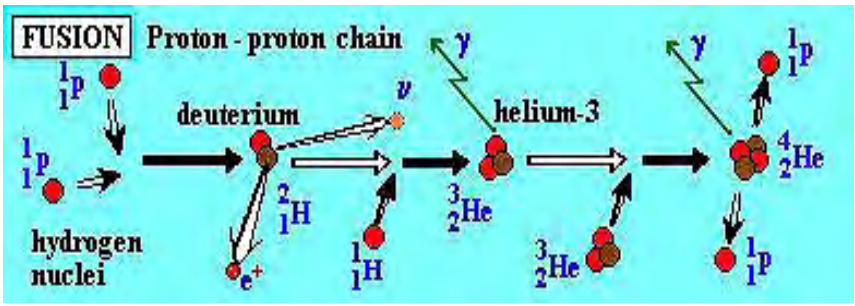
The Evolutionary Eras after the First Minute

For the first few hundred thousand years following the Big Bang, radiation was predominant over matter (Radiation Era). Later, as the Universe cooled matter was ascendant, becoming increasingly important (Matter Era).

The extremely hot, dense "soup" of matter and energy that began in the first minute is often described as the "primeval fireball". It has been likened to something akin to a *thermonuclear fusion* event, yielding a detonation-like release of energy on a grandiose scale that is just hinted at by a Hydrogen bomb's explosion. (This is how the Big Bang is usually depicted, as a reddish flame, on TV shows about the Universe. This is erroneous because Hydrogen atoms did not exist as such in the earliest Universe - only Hydrogen nuclei. The energy release would not be visible [such radiation is characteristic of much lower temperature processes] but the fireball "glow" would radiate at very short wavelengths [gamma rays among them]). This so-called invisible fireball (i.e., consisting of shortwave energy outside the visible spectrum) cooled as the Universe expanded. As stated earlier, the fireball's existence has been perpetuated as the Cosmic Background Radiation, which is the Universe-filling remnant of the initial (and small) 'fireball'.

Over the next 10 to 100 seconds after the first minute, during the first stage of the **Nucleosynthesis Epoch**, the predominant process was the production of stable nuclei (nucleons) of Hydrogen and helium. Some of the protons (p^+) and electrons (e^-) that survived initial annihilation combined to produce new neutrons (n) by weak force interactions, which added to the supply of remaining hadronic neutrons. During this stage, at first the dominant atomic nucleus was just a single proton (Hydrogen of A=1). Temperatures during this phase

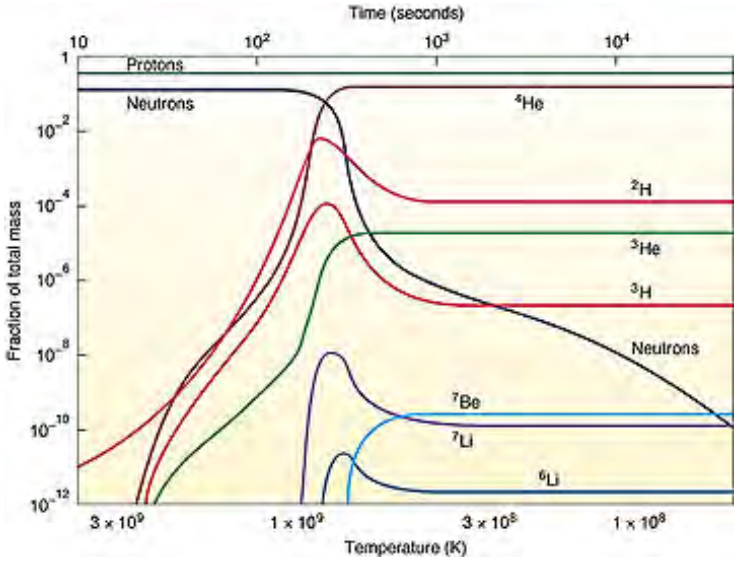
were a few billion degrees Kelvin. The basic fusion processes that formed Hydrogen and Helium isotopes are shown in this diagram:



As temperatures dropped below 10^9 °K (at ~ 3 minutes), some of the neutrons started combining with available protons (Hydrogen nuclei) to form Deuterons (heavy Hydrogen or H^2 nuclei) plus Gamma (γ) rays (resulting from the conservation of the binding energy released in the reaction). When a neutron is captured at lower temperatures, the assemblage is a Deuterium atom. Today there is ~1 such H^2 atom per 30000 Hydrogen atoms - the survival ratio; since deuterium is not produced in most stars, the deuterium we find on Earth (in heavy water molecules) is thought to be a remnant from the first seconds of the Big Bang. The amount detected provides a good theoretical control on the nuclear processes acting during the early Big Bang. A much smaller fraction of the Deuterium can capture a second neutron to form the more unstable H^3 or Tritium.

Reaction between a Deuteron and and a proton can produce helium (He^3). The much more abundant He^4 (two protons; two neutrons) is generated in several ways: by reactions between two Deuterons, between H^3 and a proton (rare), between He^3 and a neutron, or between two He^3 nuclei plus a released proton. Two other elements are also nucleosynthesized in this early stage in very small quantities: Lithium (Li; 3 protons; 4 neutrons): $He^4 + H^3 \rightarrow Li^7 + \gamma$ and Beryllium (Be; 4 protons + 3 neutrons): $He^4 + He^4 \rightarrow Be^8 + e^-$ (under the still high temperatures during nucleosynthesis, most of this highly unstable Be decays to Li).

The general time line for formation of these elements during primary nucleosynthesis appears in this next diagram which plots mass numbers of the primordial isotopes. For calculation purposes, the abundance of the Hydrogen proton is arbitrarily set at 1 - but it actually does not remain constant in the ensuing processes in which the other nucleons develop as temperatures drop in the relative abundances shown (the numbers on the ordinate have a hard to see lowered minus sign in front of the exponent). Note that the number of neutrons drop as these become part of the forming elements. The nucleosynthesis for He, Be, and Li was finished at less than 1000 seconds.



Formation of the elements heavier than Helium generally takes place within the stars as they burn their Hydrogen fuel (see page 20-7 for further details). Elements with higher atomic numbers (symbol = Z, whose value is the unique number of protons in the nucleus of a given element) are not produced at all during this initial nucleosynthesis because of energy barriers at Z = 5 (Boron) and Z = 8 (Oxygen); also the statistical probability of two nucleons of just the right kind meeting is quite low. This stability gap is overcome in stars by the fusion of 3 He^4 nuclei into a single C^{12} nucleus. The higher atomic number elements through Iron are created in more massive stars as they contract and experience rising temperatures by a complexity of fusion processes such as Helium nuclei capture, proton capture, and reactions between resulting highee Z nuclei themselves.

Thus, this brief era witnessed the synthesis of the primordial nuclear constituents -- ~90% Hydrogen/Deuterium and ~10% Helium by numbers of particles and 75-25% by mass -- that make up the two elements subsequently dominating the Universe, along with minute amounts of Lithium and Boron. Most Helium was produced at this early time, but younger Helium is also the product of Hydrogen burning in stars; the ratio of He/H has remained nearly constant because about as much new He is then created in star fusion as is converted to heavier elements during stellar evolution. The Hydrogen and Helium nuclei generated in this critical time span during the original nucleosynthesis later became the basic building materials for stars, which in turn are the sites of the internal stellar nucleosynthesis (fusion) that eventually spawned the elements with atomic numbers Z up to 26 (Fe or Iron); these account for the dominant elements, in terms of both mass and frequency, in the Universe (elements with Z > 26 are produced in other ways, such as neutron capture and neutron decay, that require energy input rather than release [as occurs for elements of Z < 26], as described later). (More about the creation [formation] of the heavier elements is covered on page 20-7.)

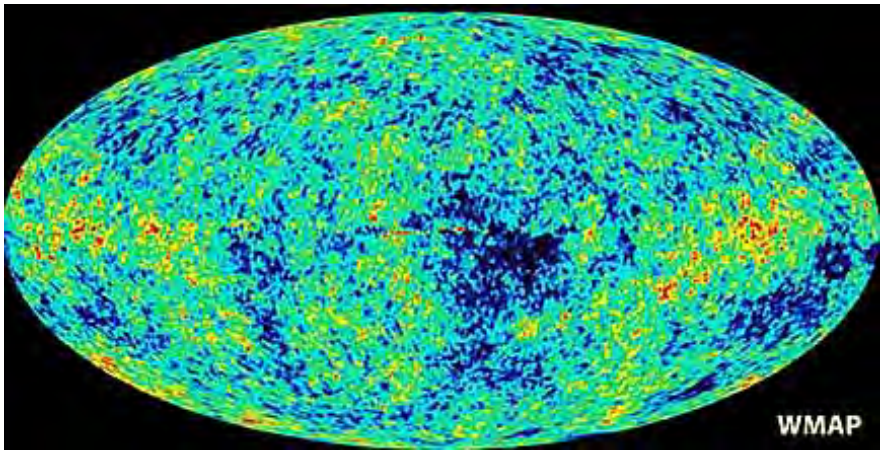
(An astounding fact, worthy of prominent insertion at this point: The vast majority of the Hydrogen atoms in your body and mine, present as Hydrogen-bearing substances, including water and various organic compounds, throughout the Earth [and extrapolated in scale up to the full content of the Universe] is primordial, that is, consists of the same individual protons that formed in the first minute of the Big Bang and then the nucleons of H during nucleosynthesis and the H atoms [single electron] soon thereafter. The additional elements in our bodies, O, C, N, Ca, Na, Mg, K, Al, Fe and others, were generated exclusively in stars, as we shall see later. We therefore consist of truly old matter, billions of years in age, and are in a sense "immortal" or "eternal" with respect to the future. Although seemingly far-fetched, some of an individual's atoms can conceivably end up in another human's body - reincarnation of sorts - as atoms released during decay may migrate into the food chain [although actual tracing of specific atoms through the transferrence is next to impossible]; or a more direct path by cannabalism is an alternative means.)

As the fireball subsided with continuing Universe expansion, the matter produced was dispersed in a still very dense "soup" of photon radiation, H and He nuclei, and free electrons. The radiation started out as predominantly gamma rays and then X-ray photon radiation along with neutrinos plus nucleons and other elementary particles. This mix of radiation, ionized H and He nuclei, and free electrons is called a plasma. With increasing expansion, the plasma progressively cooled so that the peak radiation wavelengths moved successively through longer values through the ultraviolet. The time that lasted from after the first few minutes to about 380,000 years (cosmic time, i.e., since the moment of the Big Bang) is known as the **Radiation Era** (connoting the dominance of electromagnetic radiation in the form of photons). As expansion proceeded, the mass-equivalent radiation ($E = mc^2$ equivalency) density decreased as mass density increased (today, mass density significantly exceeds radiation energy density even though the *number* of photons is much larger [in a ratio of ~1 billion photons to every Baryon]). Matter began to dominate after ~57000 years but temperatures remained too hot (above 3000 °K) for electrons to combine with nuclei. The Universe during this stage was opaque because, even with decreasing photon density, detectable radiation at any wavelengths was prevented from

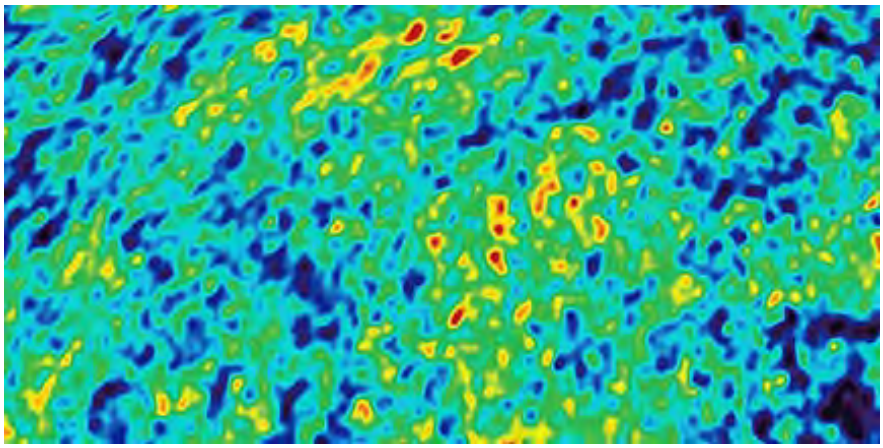
traversing or leaving the enlarging primitive Universe's confines owing to internal scattering by free electrons in the radiation "fog".

This era of opaqueness ended roughly 380,000 years after the Big Bang (some recent estimates put this termination at closer to 500,000 years after the BB) with the onset of the **Decoupling Era**, at which stage cooling had dropped below 3,000° K. When this temperature was reached, protons and Helium nuclei could start combining with electrons to form stable Hydrogen and Helium *atoms* - a process known as **Recombination**). Prior to that the photons intermixed with the H and He nuclei had been more energetic, so that they would destroy any atoms that resulted from electron capture. But, by 3,000° K the H and He could retain captured electrons. Thus, the number of free (unattached) electrons in the plasma dropped significantly so that their scattering effect on the photons diminished to the extent that the photons could have been seen as visible light "if any observer had been around to witness the radiation". (Today there are about one free proton and electron for every 100,000 atoms.) As this era began, the Universe was about 1/200th its present size. The Decoupling Era lasted until approximately the first million years of Universe history.

Since that time this glow is represented by the Cosmic Background Radiation (CBR). The general term CBR refers to the photon radiation distributed throughout the Universe at any time (including the times older than Recombination). However, it would be first detectable after the start of the Decoupling Era. This glowing radiation has continued to redshift (brought on by continuing expansion of space) from the initial short wavelengths enroute to its present-day microwave emission wavelengths, thus it is known today as the Cosmic Microwave Radiation (CMR, a synonym for the more general CBR but with the connotation of describing the current peak wavelengths that fall within the microwave region of the EM spectrum). The Cosmic Background Radiation (discussed in detail on page 20-9 and also treated effectively in this Wikipedia website) represents the residual "afterglow" (sometimes referred to as "leftover radiation") of the Big Bang - which started out as gamma radiation characteristic of very hot thermal radiation (billions of °K) that has now cooled thermodynamically (from expansion) to its present 2.73...°K (for this temperature, the radiation is detectable in the microwave region). This radiation seems nearly isotropic in all directions towards which we observe the Universe. However, it does show tiny, but extremely important deviations in temperature; the diagram below (shown again on page 20-9) is reproduced here to illustrate these variations.

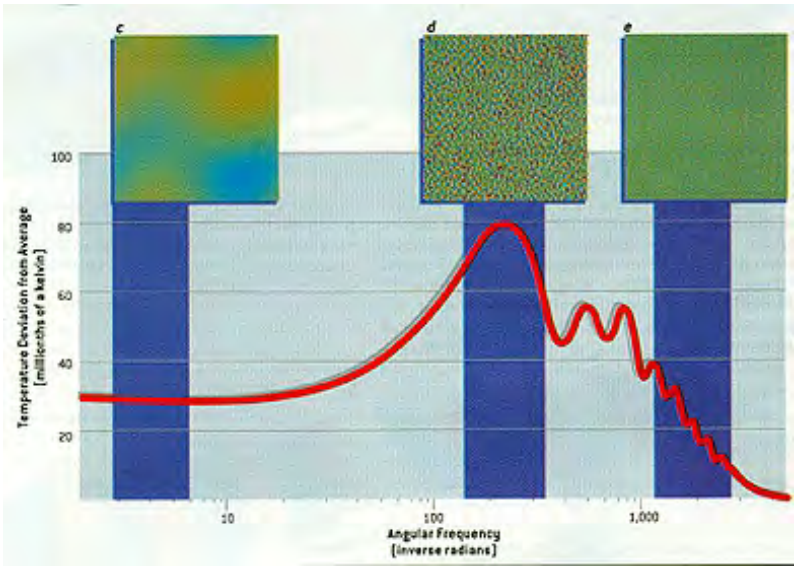


A small strip within this diagram shows in more detail the distribution of these temperature variations.



These variations become crucially important for the subsequent development of the Universe. Just how they formed is still not fully known (again, check page 20-9 for speculations). But they seem to have developed during or just after Inflation. As the inflationary moments took place and a Universe of very small but definite size resulted, variations in energy density - which led to the temperature variations - resulted when the expansion produced acoustic waves (analogous to the familiar sound waves but at much, much longer wavelengths [lower frequencies]). These waves behave like acoustic waves because they were propagated as alternating compressions and rarefactions of the particles that had come into existence and spread out after the Inflation. Like a traveling acoustic wave, those in the early expansion moved outward over times at cosmic scales. Since the medium through which they moved was comparitively tenuous (thin), their action required very high levels of energy (or power, to use the term that includes time). Interferences led to the slight differences in energy density that translate into the observed temperature differences.

Assuming then that this acoustic behavior describes the wave action in this expansion, then the spatial distribution of the temperature variations can lead to what is known as a *power spectrum*. Here is one derived from the WMAP data:



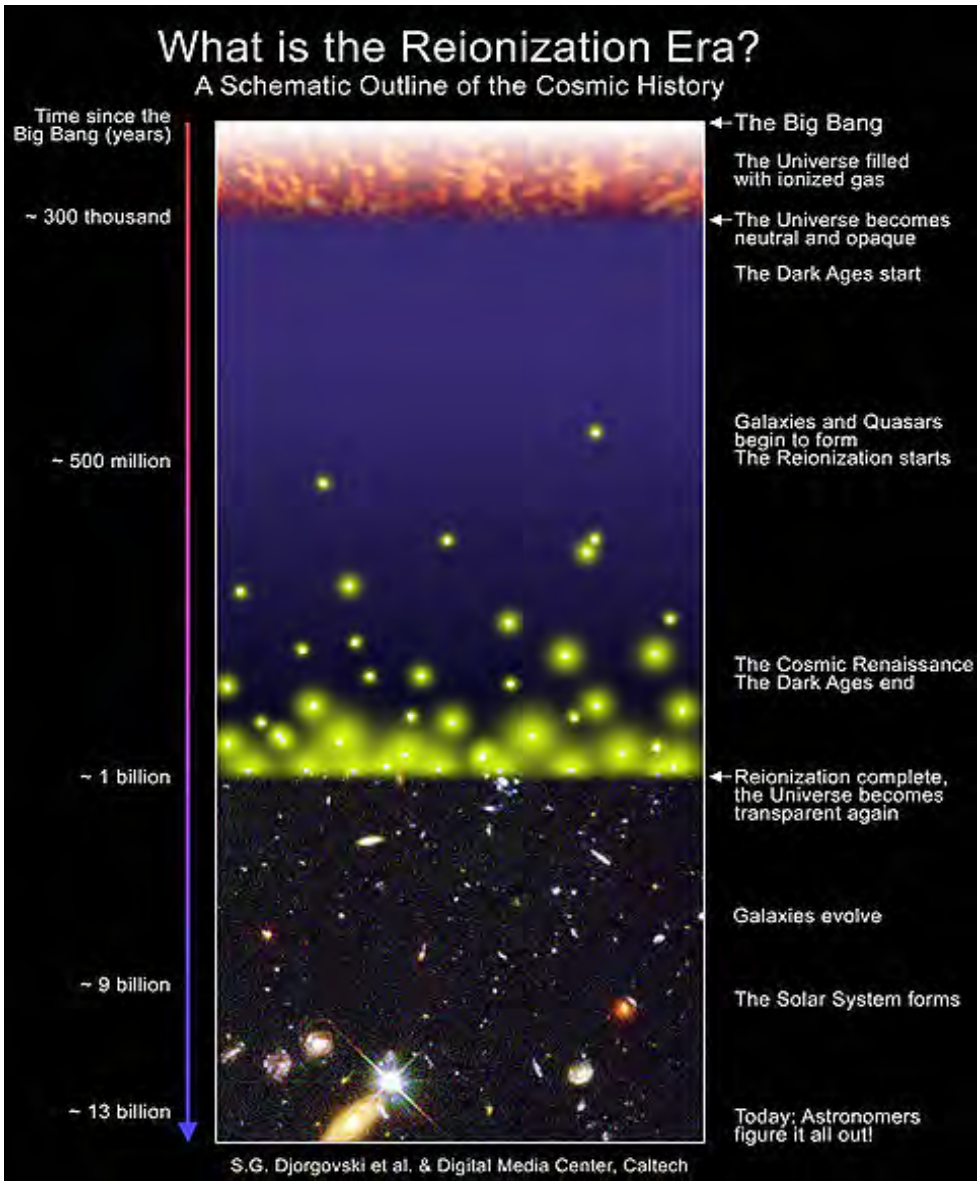
From Scientific American, February 2004

The horizontal axis plots the size of the sky (celestial sphere; essentially the imagined hemisphere above us as we look up from the horizon all about) in terms of angular frequency. The vertical axis shows variations of temperature in the millionths of a degree scale. At larger scales (c) (angular frequency equivalent to 30° but calculated in radians), the deviations from an average are moderate. At the scale of (d), about 1°, the variations are maximum. At smaller angular distances, the amount of temperature variations decreases progressively. The entire curve has the shape of a power spectrum plot. From an acoustic standpoint, the peak (d) is equivalent to the acoustic "fundamental tone"; successive peaks towards (e) are overtones.

The effect of this acoustical behavior (influenced and modified by gravity waves), up through Recombination, was to redistribute Baryons and photons produced after Inflation such that they tended to concentrate in the acoustical troughs (analogous to rarefactions in sound waves). This leads to very small, but higher temperatures in the peaks compared with lower temperatures in the troughs. Dark Matter also was of greater density in the peaks. The net result was that both energy and matter densities became spatially distributed into regions of highs and lows (but still very small variations from the overall norm) as seen in the CBR diagram (interferences may account for the non-uniform patterns). The higher density regions became the sites where the first stars and later galaxies preferentially formed. These regions were also sites of slightly greater gravitational energy; thus, they preferentially attracted more hydrogen atoms and other matter which in turn raised the gravitation state even more, bringing in still more matter and allowing gases to accumulate in huge "clouds" that in time organized into galaxies. The time interval when this occurred is narrow and critical; the material that would make up stars and galaxies had to be condensed and yet not too far apart to escape the collecting ability of gravity.

From about 5 to about 200 million years, as temperatures fell through 3000 to 600° K, the CBR photons dispersed throughout the Universe were emitting in the infrared. If any mammalian observer could have looked about the expanding Universe during this stage, this radiation would have been invisible and nothing could be envisioned since galaxies and stars had not heated up enough to be luminous. This period is referred to as the *Dark Age*, which gradually moderated until about a billion years after the BB. The atomic Hydrogen that dominated baryonic matter during this time began to clump up in denser pockets that became the small haloes (also containing more dark matter) which, as the gas built up, condensed into the first stars, a few of which may be as old as 200 million years after the BB.

As these first stars and protogalaxies began to develop, their strong outputs of electromagnetic radiation caused a **Re-ionization** (removal of electrons) of some of the Hydrogen in the still largely "empty" space. When Hydrogen protons and electrons are separate, they cannot capture (absorb) energy in the form of photons. Photons may be scattered, but scattering interactions are infrequent if the density of the plasma is low. Thus, a universe full of low density ionized hydrogen will be relatively translucent. The net effect of these new sources of ionizing radiation (in the galaxies) was that the Universe now became rather rapidly *transparent* to radiation. This allowed visible light photons to pass through interstellar space, which is an almost perfect vacuum and by itself is black, i.e., does not give off luminous self-radiation but does contain very low densities of photons and other particles (about 3 atoms per cubic meter of near-empty space). This transparency facilitates free passage from external sources of visible wavelengths within any region of the Universe. This next diagram, which incorporates the phase transition called Re-ionization, was taken from this Wikipedia web site.



Evidence for this re-ionization has been found so far not from visible light but by using UV radiation to "see" quasars (huge energy sources) that frequently formed in the first billion years. Early massive galaxies and stars produce an abundance of Ultraviolet radiation. Thus, as stars and galaxies began to form, their thermal and other energy outputs would ionize the interstellar Hydrogen, allowing their light to appear as now detectable in the visible range, so that the Universe at this stage started to show the stars as individuals and clusters. This did not happen "all at once" but gradually as galaxies formed and made their regions transparent; thus "holes" appeared intermittently in the opaque early Universe letting light from the reionizing process in galactic neighborhoods begin to spread through their surroundings as the opaqueness progressively dissipated. Ionized Hydrogen is today commonplace in intergalactic space; also present in the "vacuum" is neutral Hydrogen (detectable using its 21 cm wavelength radiation in the microwave region; this diagnostic wavelength can be longer owing to the redshift of EM frequencies that result from expansion of the Universe [page 20-9]).

Reionization is a consequence of the appearance of the first galaxies. As the first billion years ensued, conditions turned favorable for the the clustering of matter into stars (slight increases in density) that eventually gave rise to groupings that underwent organization into galaxies. As the first stars formed, especially supermassive ones, their radiation (particularly strong in the UV) began to ionize the neutral Hydrogen gas in space, forming huge "bubbles" of now ionized gas. These bubbles served as regions where more stars formed into clusters that became the initial smaller dwarf galaxies. The bubbles grew and interconnected into the larger galaxies now widely distributed. Enough neutral Hydrogen was left over to provide the material that continued to condense into stars. (For more details of this process, see Abraham Loeb, *The Dark Ages of the Universe*, in Scientific American, November 2006.)

We have mentioned "stars" several times earlier on this page (a star is defined in red in the Overview at the beginning of this page): just what is a **star**? It is indubitably **the most fundamental large single object in the Universe**. Stars began to form around sometime during the Decoupling Era. At this time, most neutral atoms in the Universe were Hydrogen. Irregularities in atom densities led to clotting of Hydrogen which continued to grow because the gravitational attraction of a clot becomes ever stronger, pulling in matter, mostly Hydrogen. Eventually, clots of Hydrogen - in spherical shapes because of the uniform pull of gravity in larger bodies - reached millions of kilometers in size. As such sizes are attained, gravity causes contraction of the bodies thereby increasing the pressures within their interiors. Temperatures increase concurrently.

At about 10,000,000 degrees Kelvin, star-interior temperatures together with the contraction pressures in the bodies can act on individual Hydrogen nuclei, which heretofore being positively charged had tended to repel each other. Under these new conditions the T-P factors are able to cause nuclei pairs to join rather than break apart. This is **fusion**, the basic process that ignites Hydrogen bombs, releasing the energy left over after the nuclear uniting that is experienced as an explosion. Helium is the result of two Hydrogen nuclei being fused. At higher temperatures and pressures elements of higher atomic numbers are produced (page 20-7). Of those elements, Carbon, Oxygen, and Nitrogen are vital as the building blocks of life; as mentioned above they all are produced within stars.

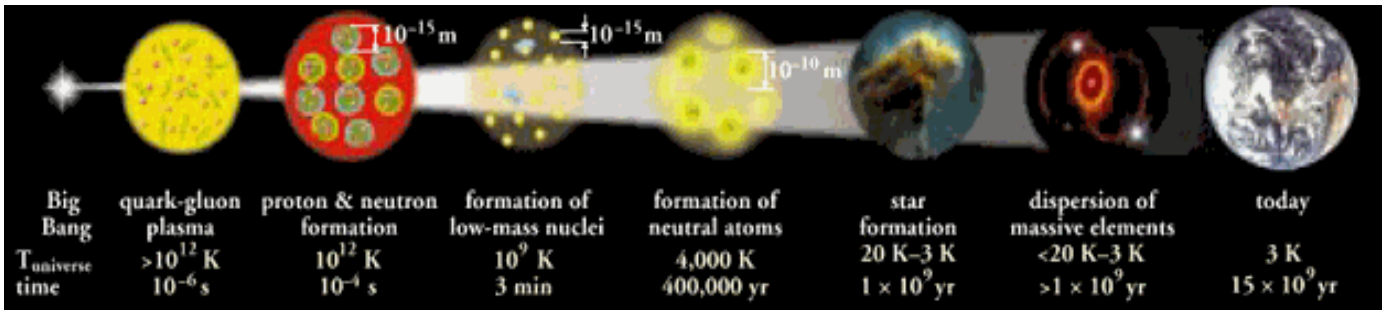
Stars are the objects that we see in **galaxies** - organized collections of stars held in a grouping determined by mutual gravitational attractions. Galaxies began to form during the first half billion years of the Universe's life. The total life of a star itself will depend on the amount of mass (mainly Hydrogen) it contains. Small stars live for many billions of years. During the stages in its lifetime in which it is stable before it has consumed its fuel, a star maintains a balance (equilibrium) between inward contraction powered by gravity and outward expansion as gases and radiation are generated from nuclear processes. Large stars die (usually violently) much earlier as they rapidly convert Hydrogen to heavier elements and use up this fuel. These big (shorter-duration lives) stars can destroy themselves by explosions termed Supernovae, which may leave behind some of the mass as Black Holes. Stars 1 to 10 times the mass of the Sun shed material with their cores becoming white dwarfs.

We will consider (next) on page 20-2 and 20-2a the nature of stars, how they form, how they produce heavier elements (covered in more detail on page 20-7), and how they end their existence. Galaxies will be treated in detail on page 20-3 and 20-3a. All we will say at this stage about galaxies is that they started out as mostly irregular in shape, then elliptical structures developed, while the familiar spiral galaxies formed later with most having organized by about 7 billion years ago, and collisions among galaxies have occurred throughout Universe time (when spiral galaxies collide elliptical types result) but were more common when the Universe was smaller (more dense).

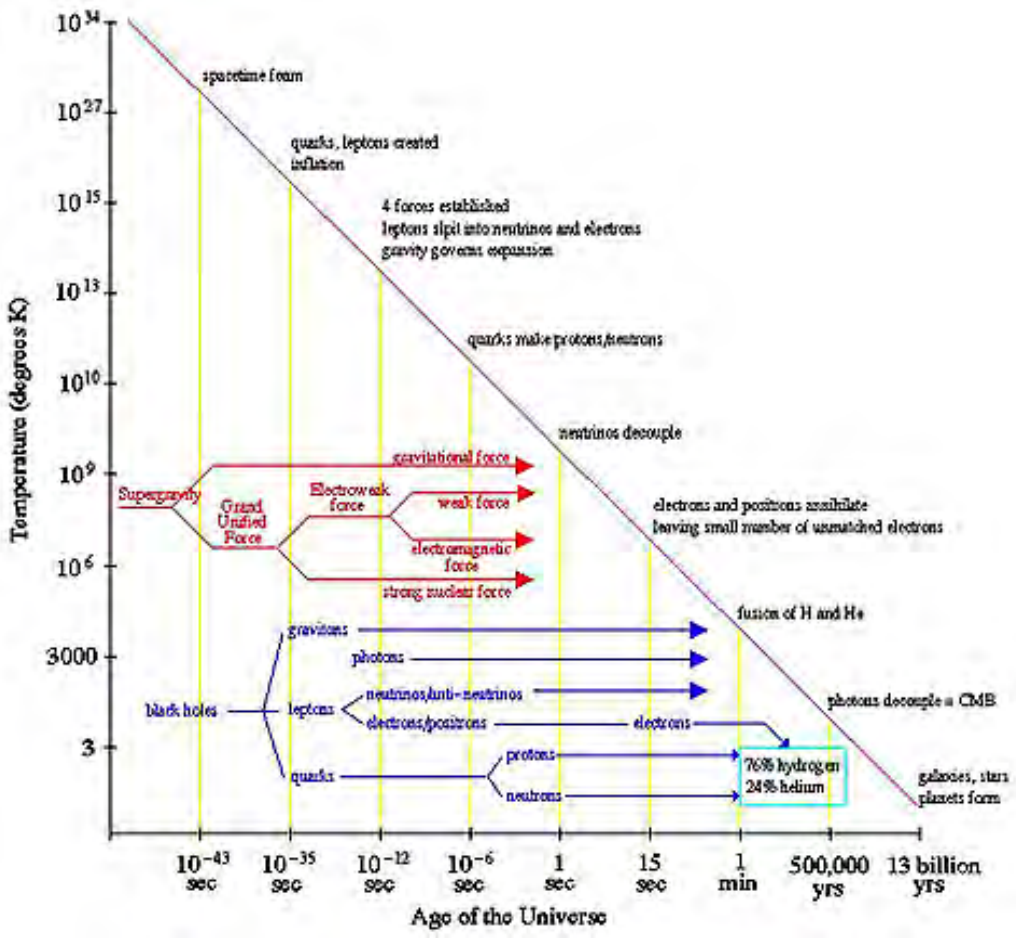
For a review of many of the ideas in this subsection, consult this Wikipedia website: "Timeline of the Big Bang".

A Recapitulation; Special Topics

Let us summarize several of the above ideas on this page, plus some others introduced in the next pages, with two diagrams. The first is a variant of the above Silk diagram for the development of the Universe after the Big Bang; it is largely self-explanatory:



The second diagram has been reproduced from one of the Websites mentioned in the Preface, the course developed by Dr. J. Schombert of the University of Oregon, labeled on his site 21st Century Science site, specifically the section entitled "The Birth of the Universe".



This diagram serves to summarize much of what has been already introduced on this page, but introduces the idea that Black Holes may have formed at the very moment of inception of matter. Black Holes (in this Section often abbreviated "B.H.") are ubiquitous objects found mostly within galaxies (but some may exist in intergalactic space) including massive ones at or near galactic centers. They are extremely dense, so much so that their extraordinarily intense gravitational pull prevents radiation from escaping them (exception: Hawking radiation) but also causes material around them to be pulled into them, commonly generating huge amounts of energy release that can be detected over the entire spectrum. (Quasars are one manifestation of this energy release.) They range in size from very small (centimeters) to sizes on planetary scales (these latter are referred to as Supermassive B.H.'s. Black Holes commmonly form from ultimate collapse of very massive stars. Black Holes play an important - perhaps critical - role in getting galaxies started and are thought to lie in the central region of most (possibly all) galaxies.

The preceding diagram gives the intergalactic temperature history of the Universe on the ordinate. For emphasis, that history is singled out in this table:

Time after B.B.	Temp.	What's Happening in the Universe?
$< 10^{-43}$ s	$> 10^{33}$ K	??? mysterious 'superforce' ???
10^{-43} s	10^{32} K	particle soup: quarks, antiquarks, photons
10^{-36} s	10^{28} K	inflationary epoch
10^{-36}...10^{-34} s	~10^{27} K	expansion slows down
10^{-12} s	10^{16} K	cooling of universe
10^{-6} s	10^{13} K	free quarks combine into protons, neutrons
0.01 s	10^{11} K	protons, neutrons, electrons, positrons, neutrinos
1 s	10^{10} K	neutrons decay into protons, electrons, neutrinos
3 min	10^{9} K	nucleosynthesis → ^{2}H, ^{4}He, ^{3}He, Li, Be, B
300,000 yr	3000 K	electrons and nuclei combine into neutral atoms
500 Myr...1 Gyr	30 K	first generation of stars form, heavy metals
13 Gyr	2.726 K	galaxies, large-scale structure in existence

The general pattern of development of the Universe since its first billion years has been one of increasing maturity of galaxies, destruction of some stars and formation of others, collisions among some galaxies, and in its second half the overall increase in expansion rate.

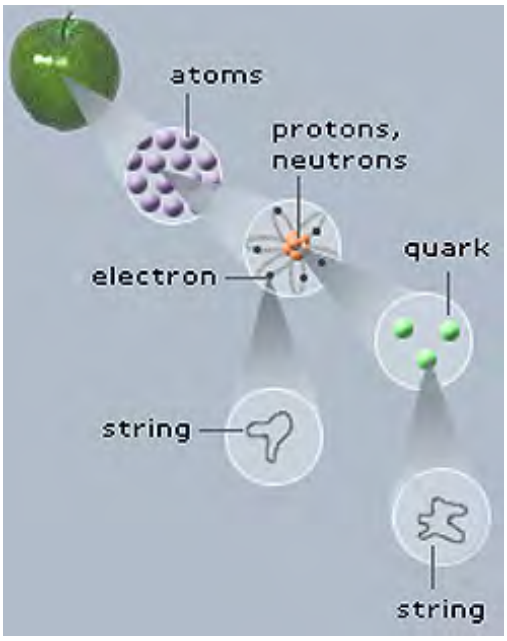
Three additional comments are appropriate here, now that the above ideas have given you a basic understanding within which they become relevant:

First, The terms "mass density" and "energy density" have appeared several times in the above paragraphs. In the initial moments of the Universe, radiation (photon) energy density was dominant. By the time temperatures had fallen to ~10000 °K, when the Universe was about 1/10000 its present size, radiation mass density (remember the $E = mc^2$ equivalency) became about equal to matter density. After the first second or so, the mass density has come to exceed radiation density, despite the aforementioned preponderance of photons over Hadrons and Leptons. In terms of numbers of particles, there are about a billion photons for every proton and electron in the Universe. In terms of matter, hydrogen is dominant: if spread uniformly throughout the observable Universe, the number of hydrogen atoms has a density of about 1 x 10^{-30} g/cu.meters. Estimates of the actual numbers of hydrogen atoms averaged out (dispersed from galaxies and stars) in space range from about 1 to 6 per cubic meters.

Second, some recent hypotheses contained in the concepts of *Hyperspace* consider the Universe at the Planck time to have consisted of 10 dimensions [other models begin with as many as 23 dimensions but these reduce to fewer dimensions owing to symmetry and other factors]. The chief advantage of this multidimensionality lies in its mathematical "elegance" which helps to simplify and unify the relevant equations of physics. As the Big Bang then commenced, this general dimensionality split into the 4 dimensions of the extant macro-Universe that underwent expansion and 6 dimensions that simultaneously collapsed into quantum space realms having dimensions of around 10^{-32} centimeters in size. This rather abstruse concept is explored in depth in the book *Hyperspace* by Michio Kaku (Anchor Books).

The third comment considers that the ultimate physical entities that make up both matter and energy may be smaller than Quarks and Leptons; these are known as ***Superstrings*** - one dimensional subparticles (of minimum lengths estimated to be 10^{-34} meters - the so-called irreducible quantum of length (given the symbol l_S) - that vibrate at different frequencies and combine in various ways (straight to looped; in bundles) to then make up the many different fundamental particles. Each species of particle has its characteristic vibrational frequency or harmonic. While a mathematical description of superstrings can be reasonably postulated, proof of their existence has yet been to be verified but theory favors their existence and they are consistent with quantum physics. Moreover, superstrings account well for some of the fundamental ideas and properties of matter, including its behavior before, at, and after the Big Bang (see page 20-11). Thus, superstrings constitute the core makeup of particles that are obvious to us as the inhabitants of 3-dimensional space. (Whether superstrings themselves are

further reducible to constituent entities or are the smallest finite entities is not known.) This diagram may help to visualize a bit of the idea.



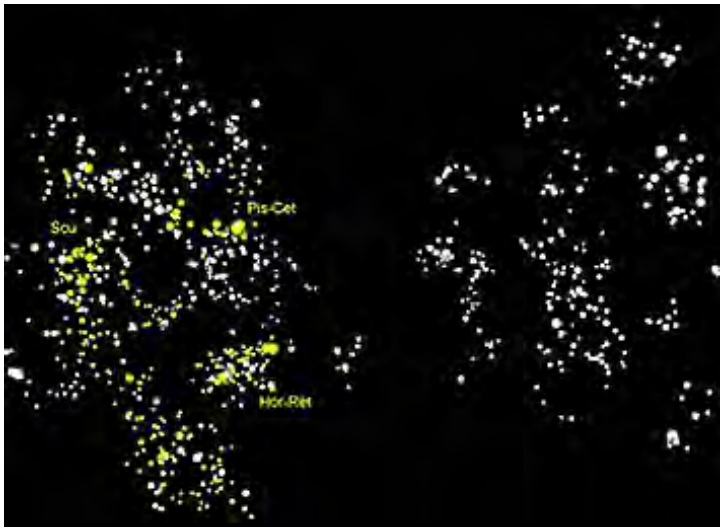
In addition to the 4th dimension and implications for the nature of time, superstrings are tied to 6 (or in some models 7) more curled dimensions whose spatial arrangement around a particle is expressed by a curvature of radius R (probably very small - in the range of l_s or somewhat larger but one recent model allows R to be up to 1 millimeter). (The idea that physical entities at subatomic sizes can have more than 3 spatial dimensions was first put forth early in the 20th Century by Theodor Kaluza [a German] and Oskar Klein [a Swede] to explain aspects of electromagnetism.) Superstrings therefore exist in hyperspace. If superstring theory proves to be valid, it will be one of the greatest achievements ever in physics. It is currently the most promising way to reconcile quantum theory and relativity. A more recent variant accounts for the graviton and contributes to an explanation of the role of gravity, the pervasive but weak force that is critical to the development and maintenance of our Universe. This is the so-called M-theory (M stands for multidimensional "membranes" (commonly spoken of as "branes" by superstring theorists). This theory postulates an 11th dimension (the membrane); when added to the dimensional mix, the result permits gravitons to fit in the general picture.

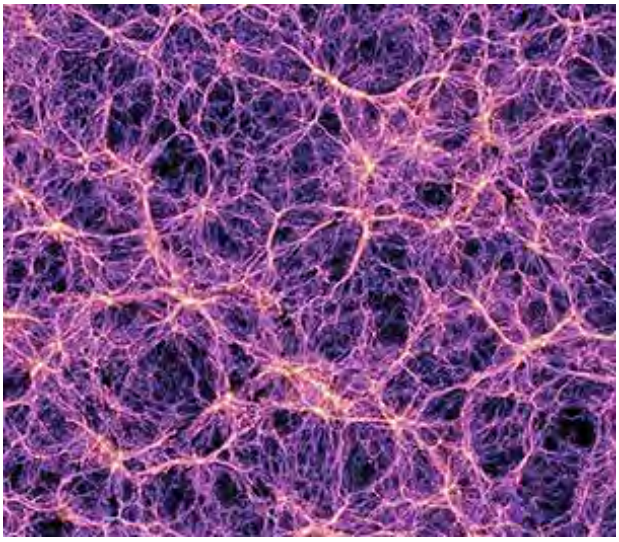
The original idea for superstrings is traced to a model proposed in 1968 by the Italian physicist Gabriele Veneziano (now at CERN) that at first ran into many difficulties, most being overcome as theoreticians began to seriously consider the concept. An outstanding review of what is up to today known or surmised about superstrings, in the context of its importance to Cosmology, has been summarized in a book (which reached best seller status) by Brian Greene, *The Elegant Universe*, 1999, W.W. Norton & Co.). Greene has apparently replaced Carl Sagan as the "guru" of Science whose personality favors an ability to popularize such hard concepts to master (his rival is Neil DeGrasse Tyson). Public Broadcasting (PBS) through its Nova program has aired a 3 hour special called "The Elegant Universe" (beautifully done!). I am attaching this PBS website address (assuming they retain it online for future times) that summarizes the fundamentals of superstring theory a la Greene. There is also a web page covering the basic concepts that purports to be the official site for a survey of Superstring Theory. And Wikipedia has several Web sites that deal with these topics; they are:Superstring theory; M-theory; Hyperspace web references; Hyperspace in Science Fiction.

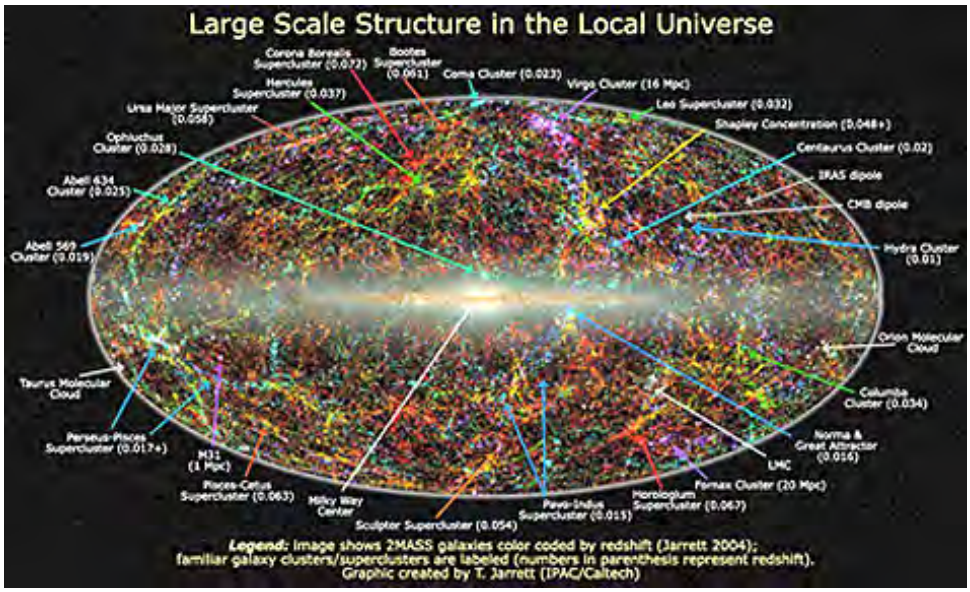
Having originated the Universe through the preferred Big Bang Model and then considering the first few hundred million years of its early history, you will now embark on a systematic review of the major aspects of Cosmology, beginning with a survey of "stars" (after a brief "detour" to examine the Hubble Space Telescope - the space observatory that revolutionized our

understanding of the Cosmos). However, it is useful now to consider succinctly the topic of "Structure of the Universe" as an Introduction to these other topics.

From an astronomical perspective, the Universe beyond Earth can be considered to have this hierarchy of major components, arranged by size: Cosmic dust and gas; Asteroids and Comets; Planets and their Moons; Stars (with the Sun as the prime model); Galaxies and Black Holes; Galaxy Clusters; Superclusters; the Cosmic Web (Filaments and Voids); the Universe itself (which embraces all its matter and energy); (and the possibility that there may be multiple Universes or Multiverses). As a preview of what will follow [mainly in the subsection on Galaxies], here are three illustrations, showing a Supercluster, a model of Filaments, and the structure of the "local" region of the Universe:







A comment before closing this page: Humans are the most highly developed assemblages of ordinary matter (this must be qualified by stating that other, more advanced thinking creatures may reside elsewhere in the Universe). But, as you saw in a diagram near the middle of this page, ordinary matter makes up only about 4% of all entities (matter and energy) in the Universe. One is tempted to claim that stars and galaxies are the principal manifestations of ordinary matter. But recent studies, utilizing data from the interactions of quasar (extreme luminous energy sources surrounding Black Holes), show that the majority of baryonic matter resides in inter- and intra-galactic space. This matter is probably unevenly distributed, with most found within the filaments just described. The matter is presumed to be composed of ionized hydrogen and some ionized heavier elements.

We have now reached the close of this first, largely overview page dealing with Cosmology. Heady stuff!! Perhaps you need to review the main ideas behind the Big Bang using another source. To do that, check this Wikipedia website that deals with the Big Bang And, in the interests of fairness (by trying to eliminate a bias favoring the Big Bang concept), the writer will here in this paragraph, give several Internet connections to websites that dispute the "correctness" of the Big Bang, i.e., critique the concept and suggest alternatives. These were found as Google sites by typing this topic: "Proof of the Big Bang". These three are typical of the pros and cons that topic elicited: Big Bang skeptic (this review traces the conversion of Richard Carrier from a negative to a positive view of the Big Bang); What Big Bang (the discreditation of the Big Bang by Alexander Shulgin); and The Big Bang is challenged (by an unidentified author). Suffice to say that the writer is among the majority of scientists who currently accept the general tenets of the Big Bang as described on this page as being "a reasonable hypothesis" that better explains the Universe than any other competing model; but aspects of the present Big Bang concept are likely to be modified, or perhaps rejected, and new variations or components of the hypothesis will be "discovered" in the future.

This has been a lengthy page. The remaining pages in this Section are also long. We will start with a diversion that describes the most important astronomical observatory ever constructed by humans - the Hubble Space Telescope. Then a page pair on stars, followed by two pages on galaxies, both topics centered on observations in visible light. A page follows that treats observations of heavenly bodies in other regions of the electromagnetic spectrum. A page dealing with special features of galaxies then ensues. After that, a page considers supernovae, pulsars, quasars, and black holes. The next page examines how the elements came to exist. The following page looks at the nature of spacetime and expansion. Evidence for the Big Bang is then treated, along with a discussion of CBR and Dark Matter/Energy. Page 20-10 looks at some recent ideas including the newly discovered idea of an accelerating Univers. Then, models for planetary formation are treated. A page is devoted to the nature of life in the Universe; this includes a "mini-course" on Biology and a history of life on Earth. The last page includes some philosophical musings about "the meaning of it all".

*A measure of cosmic distance to any object beyond our Sun is the *light year* (l.y.), defined as the distance (~ 9.46 x 10^{12} or 9,460,000,000,000 km or ~5.9 quadrilion miles) traveled by a photon moving at the speed of light (2.998.... x 10^{8} m/sec, usually rounded off and expressed as 300,000 km/sec) during a journey of 1 Earth year; another distance parameter is the *parsec*, which is the distance traversed in 3.3 l.y. Keep in mind that this is an arbitrary and anthropomorphic parameter in that the year is strictly valid only for the Solar System, and more particularly applies to just that time determined by the number of days in which Earth takes to travel one complete orbital cycle around the Sun. Another comment: The parts of the Universe now visible in terms of maximum measurable light year distances are thought to be a region within a (possibly much) larger Universe of matter and energy, with light from these portions beyond the detectable limits having not yet arrived at Earth.

** It is often difficult to find a clear definition of the term "space" in most textbooks (just look for the word in their index - it is almost always absent). We tend to think first of the "out there" that has been reached and explored by unmanned probes and by astronauts as the "space" of interest. One definition recently encountered describes space as "the dimensionality that is characterized by containing the universal gravity field". The writer (NMS) has tried to think up a more general definition. It goes like this: *Space is the totality of that entity that contains all real particles of matter/energy, both dispersed and concentrated (in star and galaxy clots), which fill, and are confined to, spatial dimensions that appear to be changing (enlarging) with time.* Anything one can conceive that lies outside this has no meaning in terms of a geometric framework but can be conceptualized by the word "**total vacuum**". This vacuum would contain no energy whatsoever (therefore, no mass), and would be devoid of spatial dimensions and time. In the quantum world a false vacuum is hypothesized as occupied by virtual particles capable of creating new matter and space if a fluctuation succeeds in making a (or perhaps many) new Universe(s).

*** Symmetry in everyday experience relates to geometric or spatial distribution of points of reference on a body that repeat systematically when the body is subjected to specific regular movements. When rotated, translated, or reversed as a reflection, the points after a certain amount of movement are repeated in their same relative positions as it returns to its initial position. For example, a cube rotated 360° around an axis passing through the centers of two opposing faces will repeat the square initially facing the observer four times [90° increments. The concept of symmetry as applied to subatomic physics has other, although related, meanings that depend on conservation laws as well as relevance to spatial patterns. In general terms, this mode of symmetry refers to any quantity that remains unchanged (invariant) during a transformation. Implied are the possibilities of particle equivalency and interchangeability (the term "shuffled" may be used to refer such shifts).

Expressed mathematically, certain fundamental equations are symmetrical if they remain unchanged after their components (terms) are shuffled or rotated. In quantum mechanics, gauge (Yang-Mills) symmetry involves invariance when the three non-gravitational forces (as a system) undergo allowable shifts in the values of the force charges. At the subatomic level in the first moments of the Big Bang, symmetry is applied to a state in which the fundamental forces and their corresponding particles are combined, interchangeable, and equivalent; during this brief time, particles can "convert" into one another, e.g., hadrons in leptons or vice versa. When this symmetry is "broken", after the GUT state, the forces and their corresponding particles become separate and distinct. The progressive breaking of symmetry during the first minute of the Big Bang has been likened (analogous) to crystallization of a magma (igneous rock) by the process of differentiation. At some temperature (range), a crystal of a mineral with a certain composition precipitates out; if it can leave the fluid magma (crystal settling), the remaining magma has changed in composition. At a lower temperature, a second mineral species crystallizes, further altering the magma composition. When the last mineral species crystallizes, at still lower temperatures, the magma is now solidified. All the minerals that crystallized remain, each with its own composition. In the Big Bang, as temperatures fall, different fundamental particles become released, altering the energy state of the initial mix, as specific temperatures are reached (and at different times) until the final result is the appearance of all these particles, which as the Universe further expands and cools become bound in specific arrangements (e.g., neutrons and protons forming H and He nuclei; later picking up electrons to convert to atoms) that ultimately reorganize in stars, galaxies, and the inter- and intra-galactic medium of near empty space.

**** The familiar term "mass" needs some explanation. The Newtonian definition simply refers to the quantity of matter, in terms of its density and bulk. The relativistic quantum concept of matter considers mass to arise whenever a particle interacts with the Higgs quantum field. When a particle such as a Quark or a W Boson reacts to the Higgs field (as yet only a postulate awaiting verification of its existence), which is held to be omnipresent in empty space, in current theory this gives rise to the Higgs particle (the Boson that fosters its interaction as a force; see other references to this elswhere on this page). In so doing, the particle acquires the property called "mass". The different particles shown in the classification table in the text on this page gain their different masses because of variations in strength of interaction. In the SuperSymmetric Standard Model, there are two kinds of Higgs fields and five species of Higgs Bosons. The quantum Higgs field differs from gravitational and electromagnetic fields in that the latter assume 'zero field strength' in their lowest energy states whereas the Higgs field is 'non-zero' at its minimum energy state. The Higgs field (if it existed before the Big Bang) therefore has a finite (negative) value in a Universe-to-be that may not yet have formed discrete mass particles; as particles are created just after the singularity event they quickly react with the Higgs field to evolve into states where mass becomes a real property. For further information, we refer you to "The Mysteries of Mass" by Gordon Kane, *Scientific American*, July 2005.

***** Energy can be said to be quantized, that is, is associated with quanta (singular, quantum) which are discrete particles having different units of energy (E) whose values are given by the Planck equation $E = hc/\lambda$ where h = Planck's constant, c = speed of light (~300,000 km/sec), and λ = the wavelength of the radiation wave for the particular energy state of the quantum being considered; the energy values vary with λ as positioned on the electromagnetic spectrum (a plot of continuously varying wavelengths).

******This extremely rapid enlargement reflects the earlier influence of Inflation with its initially very high expansion rates. Keep in mind that many of the parametric values cited in cosmological research are current estimates or approximations that may change as new data are acquired and/or depend on the particular cosmological model being used (e.g., standard versus inflationary Big Bang models). Among these, the most sought-after parameter is H, the Hubble Constant (discussed later in this review), being one of the prime goals for observations from the Hubble Space Telescope·

PREVIOUS NEXT

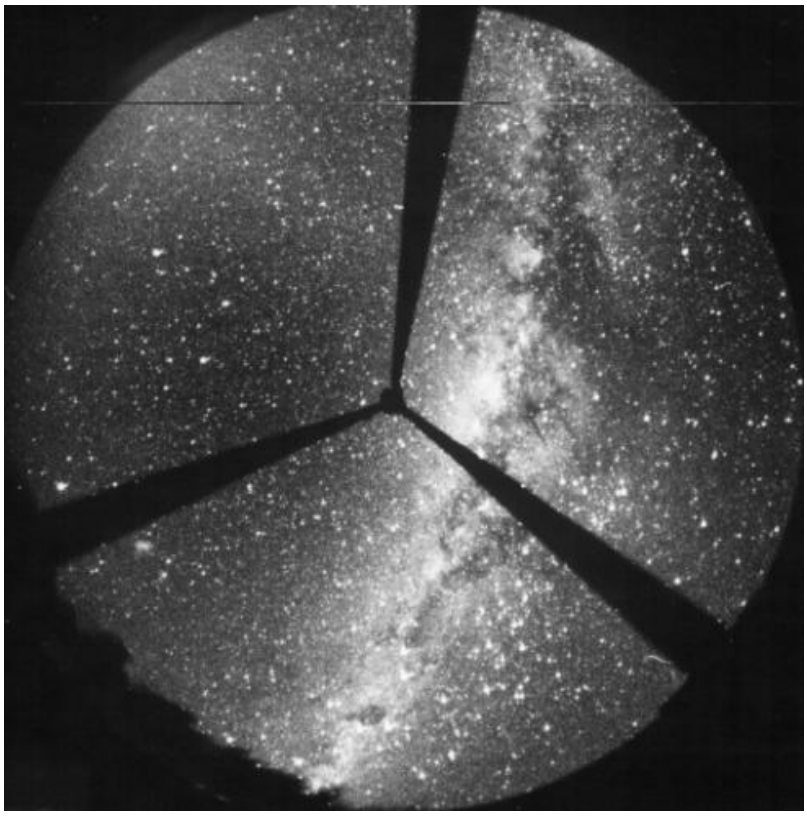
Primary Author: Nicholas M. Short, Sr.

The Nature and Evolution of Galaxies

Prior to the 1920s, astronomers considered the Universe to consist of a single huge clustering of stars that was named the *Milky Way* (M.W.). As visualized with the naked eye in a clear, moonless desert night, the M.W. appears as a band running across the celestial sphere containg about a thousand visible stars. That number increases significantly when a time-exposure photo of the Milky Way using color filters is made, such as this one:

Using his newly crafted telescope, Galileo was the first to realize that the Milky Way contains many more stars that the eye can see unaided

This excellent photo montage, made in 1926 by E. Houck and A. Goode using a blue filter and a total exposure time of 45 minutes while moving their camera in snyc with the Earth's rotation, portrays the denseness of stars in the Milky Way, implying the number of stars were in the millions.



But as the Big Bang concept took hold, it was realized that expansion rates would carry distant "stars" well beyond the M.W.'s sphere of influence. In the late 1920s, **Edwin Hubble** was the first to present strong evidence that these stars were actually other galaxies. (However, historically Immanuel Kant in the 18th Century proposed that the Milky Way band was a vast collection of stars in a stretched out band and that some of the stars visible in the telescopes of his day were indeed other Milky Ways.) Thus the Universe became much bigger and contains a myriad (billions) of galaxies that make up the visible entities filling expanding space. Still, when one spots points of light in the heavens with the naked eye, what is seen are several planets, a large number of nearby stars, and a very few galaxies close by.

A ***galaxy*** is an organized concentration or clumping of stars held together by mutual gravitational interaction in an aggregate containing millions to billions of discrete individual stellar objects grouped into specific geometric arrangements (spiral; elliptical; irregular). A feel for this huge number of stars within a galaxy is given by this set of images:

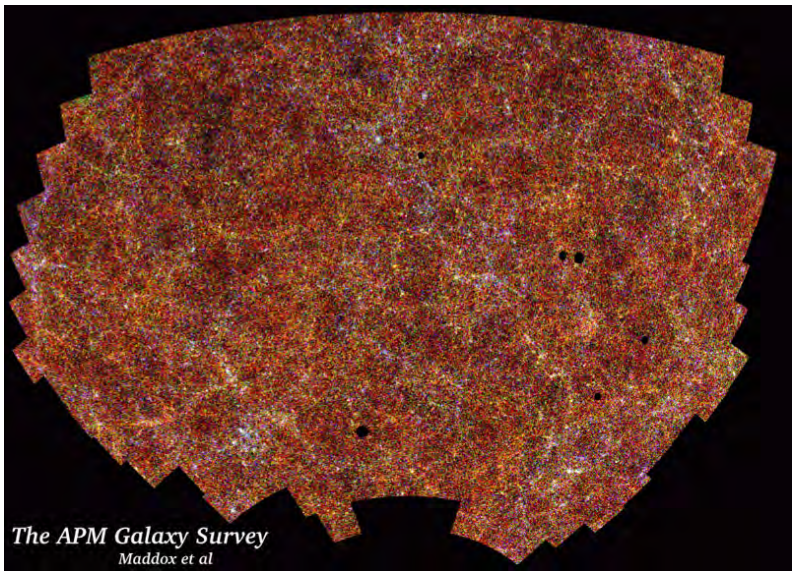
The galaxy is NGC300, which is a neighbor to the Milky Way, being about 6.5 million light years away. The image in the upper left is a telescope view. In the upper right is an ACS Hubble image of this galaxy. Careful processing has resolved the many stars in the rectangular inset from that galaxy. While they appear to have a very high density, this is in a sense an illustion because, while the picture is 2-dimensional, the field of view is 3-dimensional, with stars at various slight differences in distances along the telescope's line of sight. Actually, the distance between stars is very much greater than their diameters.

Typical maximum dimensions of a galaxy range from 80,000 to 150,000 light years in space-time diameter. The central disk of the most photogenic type - the Spiral Galaxy - is about 10000 light years in thickness. Galaxies contain huge numbers of individual stars - a common number cited is 100 billion stars, but some have less and others up to a trillion (this estimate is based on the growing knowledge of the abundance of red and brown dwarf stars). At least 10 billion - probably many more - galaxies may have developed in the observable Universe. Of course, these numbers are estimates made by sampling regions of space close to us; attempts to accurately inventory *all* galaxies and stars by some counting approach are currently not feasible, and would suffer from incompleteness owing to the probable existence of stars/galaxies beyond observable limits.

While the question is not fully settled as to whether stars must have formed before (as contrasted to "during") the first galaxies, there is a growing consensus that a group of very massive, gradually heated stars emerged before any galaxies. These stars, almost entirely Hydrogen and Helium, organized rapidly, burned for a short time (around 3 million years), underwent collapse and exploded as supernovae. Being the first "furnaces"to produce heavier (atomic weight) elements, the destroyed stars yielded materials (including carbon, calcium and oxygen) that became incorporated in the first galaxies to form. This topic - element formation in stars - was treated in more detail on page 20-7.

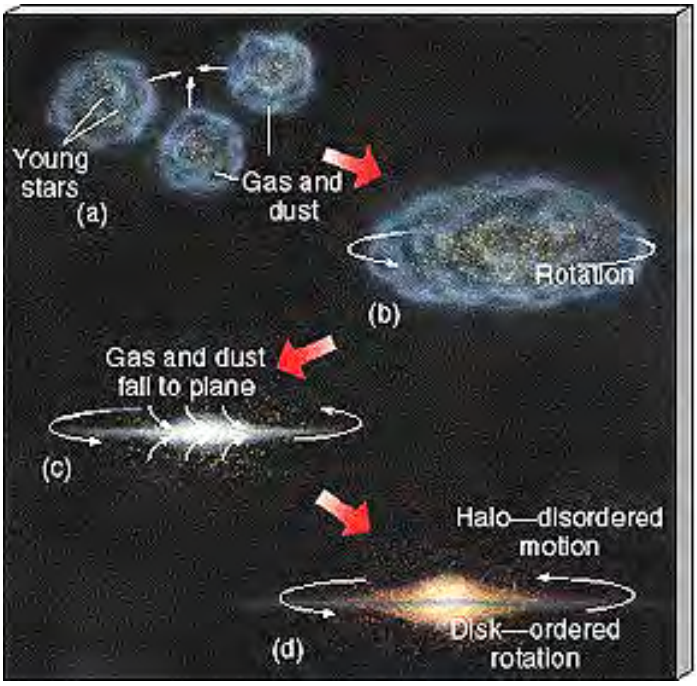
Just as there are billions of stars in a single galaxy, such as the Milky Way, there are many billions of galaxies throughout the observable Universe. One of the largest Sky surveys to date uses the APM (Automatic Plate Measurement) technique to image galaxies between Magnitude 17 to 22 - out to intermediate spatial depths. Here is a composite made by the University of Nottingham that contains at least 2 million galaxies over a 100° portion of the sky outward from the Earth's South Pole. Individual galaxies can be resolved in enlargements but in this rendition the sky seems filled with galaxies (most rendered in red) - obeying the Cosmological Principle (the

Universe is isotropic at large scales).



(A parenthetical aside: This image seems to show a crowded sky, with light sources [galaxies] almost touching. That would support Olbers Paradox: the sky should be uniformly bright if the Universe is infinite, since any ray looking outward from Earth would eventually intersect a major light source. But this is obviously not so -- the night sky is dark. The reason: the Universe is not infinite - at least the observable part; the distances between galaxies is huge with respect to individual galaxies; and, the majority of radiation reaching Earth is outside the Visible light spectral range [Cosmic Background Radiation; X-ray sources, etc.])

A synopsis of galaxy formation - for the spiral type - is evident in this simplistic model, which is generalized:



In this model, which has aspects that go back to Laplace in the 18th Century, a cloud of Hydrogen/helium gas begins to form myriads of stars. As this continues, the cloud may contract somewhat and the assemblage of stars begin to rotate around a common center. This cloud has also be called a "halo" and as such may be the precursor to the Halo that surrounds the Milky Way and other galaxies (see below). With rotation, there is a tendency for the cloud to assume a more oblate ellipsoidal shape and begin to spin. The spinning produces strings of stars in at least several distinct arms. When well developed, the stars have organized into a spiral galaxy.

Now to some more details: As the dispersing mix of primordial H and He (He comprises about 10% of the various atomic species present) atoms, photons, and other particles continued to expand (thereby progressively decreasing in density), it eventually cooled to temperatures around a few degrees Kelvin (see Cosmic Background Radiation on page 20-9). Large-scale variations (called fluctuations or seed perturbations) in mass and energy density, whose origin can be traced to the early moments of the Big Bang, occurred at random throughout the enlarging Universe. These regions where the density was greater eventually grew (as described below) into protogalaxies and then galaxies. As early as the first 100 million years (m.y.) (cosmic time; measured from the moment of the Big Bang) and perhaps as far back as just after the Decoupling Era, but especially in the first 1 to 2 billion years, protogalaxies (incipient or first stage assemblages of the Hydrogen-rich gas that evolve into galaxies) began by means of gravitational attraction to develop as denser regions throughout the expanding Universe. This process was guided by gravity-driven irregularities or *ripples* in the almost homogeneous distribution of particles in the early stage expansion of the Universe. These denser strands or pockets of matter evolved over time as stars formed and collected into fullblown galaxies, with most now observed having formed during the first four billion years. The principal hallmark of galaxies is that they consist of billions of stars (whose nature and development are described on page 20-2).

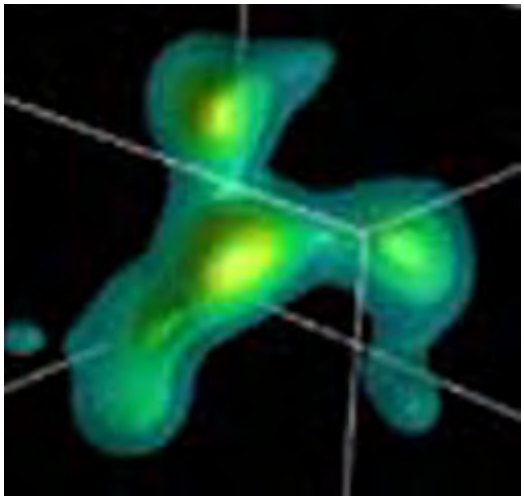
Amazingly, despite the vast number of stars in a galaxy, most of the Universe's space is nearly empty of luminous matter, making up intragalactic and, even more so, intergalactic open regions. Likewise, individual stars in a galaxy are widely separated (a scale analogy: if a star is represented by a marble just 1 centimeter in diameter, the average distance to its nearest neighbor stars is around 300 kilometers [~200 miles]). All stars together (totaled for all galaxies) comprise just about 1 part per million by size within the space dimensions calculated for the known Universe: thus in the total volume of observable space, "void" dominates and luminous objects are an exceedingly small part (far less than one might expect by looking through a telescope in which much of the field of view seems occupied by points of light [galaxies or galactic clusters], since there are huge distances between them in the direction of viewing). In terms of mass, stars likewise constitute less than 2% of the total presently calculated for the Universe.

While most galaxies are very old, some are younger and a small fraction may even have started forming in the last few hundred million years. One example (below) of an embryonic galaxy is Hubble-X , in the constellation Sagittarius (NGC6822), which is about 1.6 billion light years from Earth. Evidence based on star characteristics indicates the cloud started producing stars only about 4 million years ago, but a well-defined galactic shape is yet to emerge.

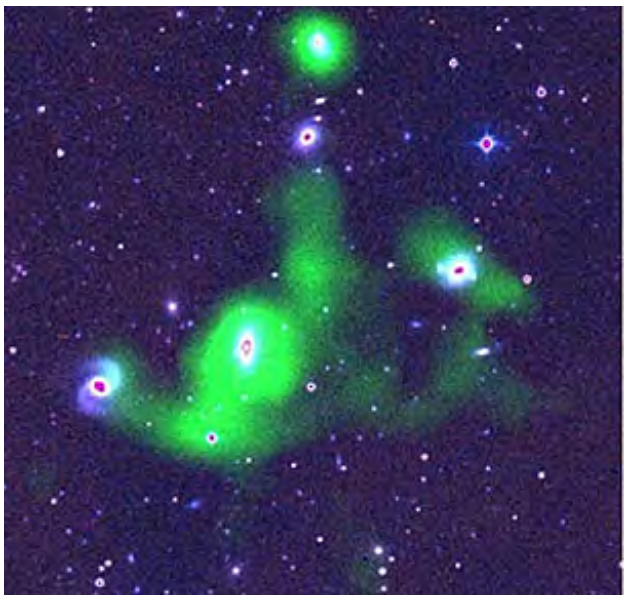
However, in general within galaxies the majority of larger stars has since expired (by supernova explosions, etc.) even as new stars (including those of masses up to 100 times that of the Sun) continually form (some recently, in Universe time) from the debris and gases remaining in the intragalactic materials that persist throughout the history of the galaxy. Other materials are

drawn in as encountered during a galaxy's travels in space.

The starting point of galaxy formation requires accumulation of Hydrogen-rich gas, with some Helium, in a great cloud (many millions of light years in dimension). This stellar nursery may have been similar to what are referred to as a "Molecular Cloud" or the large "Giant Molecular Cloud" (GMC) because much of its Hydrogen is combined as H_2 (see below). However, the typical galactic cloud would have been much larger - containing billions of stars and being at least a few million light years across - than these molecular clouds which usually contain stars only in the millions or less that are observed today in and around existing galaxies. Some of these clouds are huge. The largest found to date is nearly 200 million light years in size and consists of several lobes of gas within which galaxies appear to be forming. This rendition of its appearance through a telescope (this appears to be an artist's conception) shows its shape:



This image below, made from radio telescope data (see 20-4), shows a huge cloud of cold Hydrogen gas (green) in the Hickman Compact Group:



For this accumulation (build-up) of gases to happen there must initially be localized regions of the expanding Universe whose density is slightly greater than the generally uniform distribution of matter and photons that, most cosmologists believe, was the outcome of the processes operating during the earliest stages of Big Bang expansion. Studies of cosmic background radiation (see page 20-9) indicate these density disparities may have been as small as 1 part in one hundred thousand. The slight differences in density also give rise to slightly greater gravitational forces which act to draw material towards these local perturbations.

As more matter accrues within a growing cloud, its internal gravity continues to increase and

draw in still more gases. The molecular cloud eventually reaches a density that requires it to then undergo local clumping of gases into clots that grow into still denser concentrations to become stars (these smaller clots can exist for much of the galaxy's life but are the sites of further star formation).

The next HST image shows huge clots of gas and dust in a more advanced stage of development in which stars will eventually form *en masse* as part of a spiral or globular galaxy (see below):

Many star-forming clouds are very rich in dust, in addition to the Hydrogen gas, which make them appear as discrete dark clouds. As we shall see on page 20-7, these clouds contain various amounts of heavier elements (but still only a small fraction of the total number of Hydrogen and Helium atoms present) produced within the first stars and dispersed when these exploded as supernovae. A prime example of vast dust cloud "nurseries" from which stars are born are shown in these next images. The first shows much of the Eagle Nebula (M16), which we first depicted on page 20-2:

The great protuberances of dust-gas within the Eagle Nebula are called pillars, evident in the above image. Below that, is the trio of 3 pillars in this nebula; this image is now near the top of the list of most "spectacular" of all HST images captured so far. (See page 20-11 for three more views of this nebula.)

This assemblage of gas and dust, in which new stars have or will be formed, is not as large as some others that have been detected. This is evident in the Spitzer Space Telescope image of the W5 nebula (in the Cassiopeia constellation) in which the Eagle nebula image is shown to scale (inset) for comparison with the much larger W5 in which many nascent stars are present.

Studies of GMCs prove informative regarding the processes involved in building stars within galaxies. To some extent, they are miniature versions of the super-clouds that evolve into galaxies (a galaxy also can grow by capturing or contacting more GMCs). Fortunately there are several GMCs close (1500 light years) to Earth that serve as a "laboratory" for observing star formation processes and subgalactic growth. These are found within the famed Orion ("The Hunter") constellation that occurs near the celestial equator near the star Taurus (for location see star chart labeled "Southern Horizon, Winter" near the middle of page 20-2). Probably the most studied of all nebulae is the dominant feature within this constellation that is known as the Orion Nebula. It consists of two prominent nebulae M42 (larger) and M43 above it (in some images this appears to be a discrete entity but high resolution images show it to be continuous with M42).

This ground telescope view shows the larger M42 and smaller M43 as drawn apart, with the former including the section known as the Trapezium. Below that is an image made by the HST:

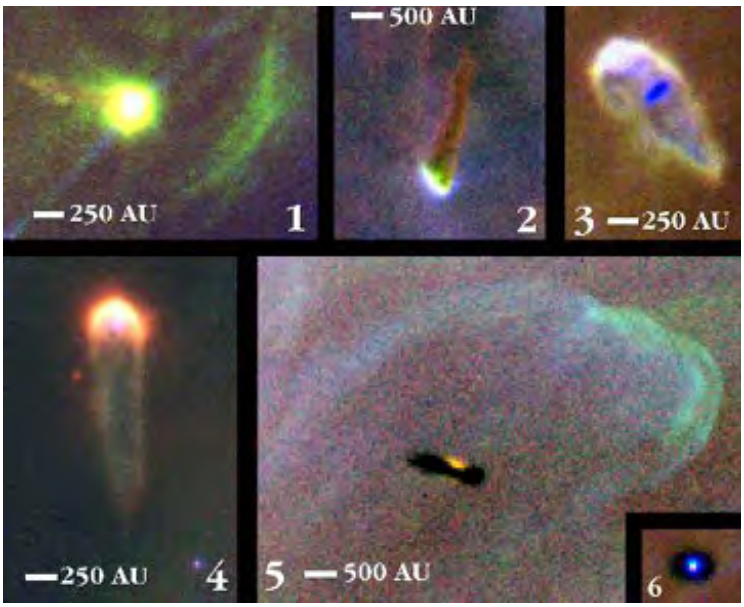
And, M43 by itself:

The interior of M42 as seen in visible light shows the clouds of Hydrogen and dust typical of GMCs; the same area in infrared light brings out the principal stars in this region.

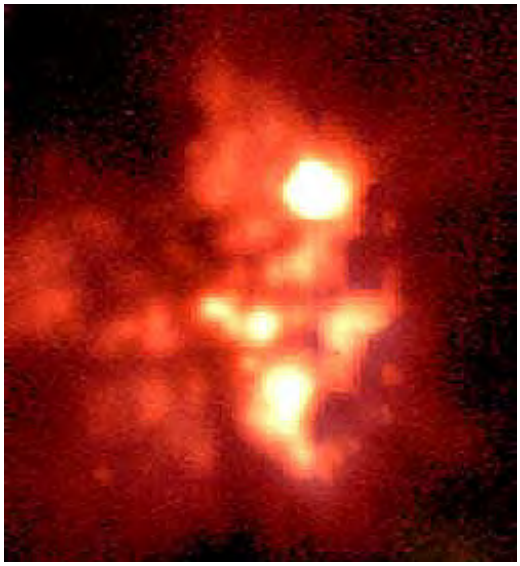
Within M42 are four very large, bright stars that comprise a geometric figure known as a trapezium. Some of the brightest large stars in the Trapezium of M42 are shown here optically:

The next view shows the clouds around the 4-star Trapezium and the stars themselves:

The Trapezium region is a major "nursery" for stars forming and evolving, as shown in this 6 panel montage:



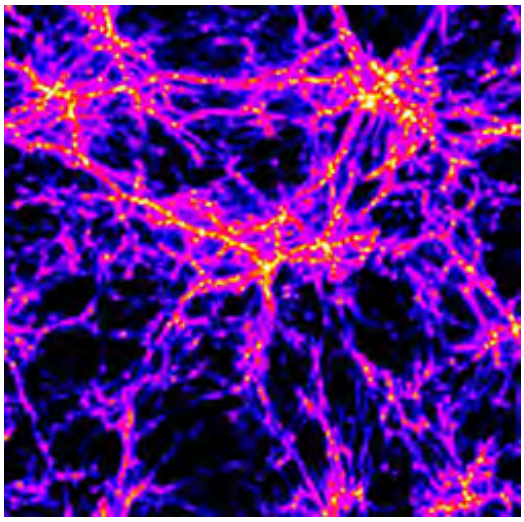
That these clouds are thermally active, especially where the clots are organizing into protostars, is evident in this ratio image made from thermal bands, as follows - 20µm/10µm - in this view of a cloud near the center of the Orion galaxy, made using the TIMMI2 (second Thermal IR MultiMode Instrument) on the 3.5 m telescope operated by the European Southern Observatory:



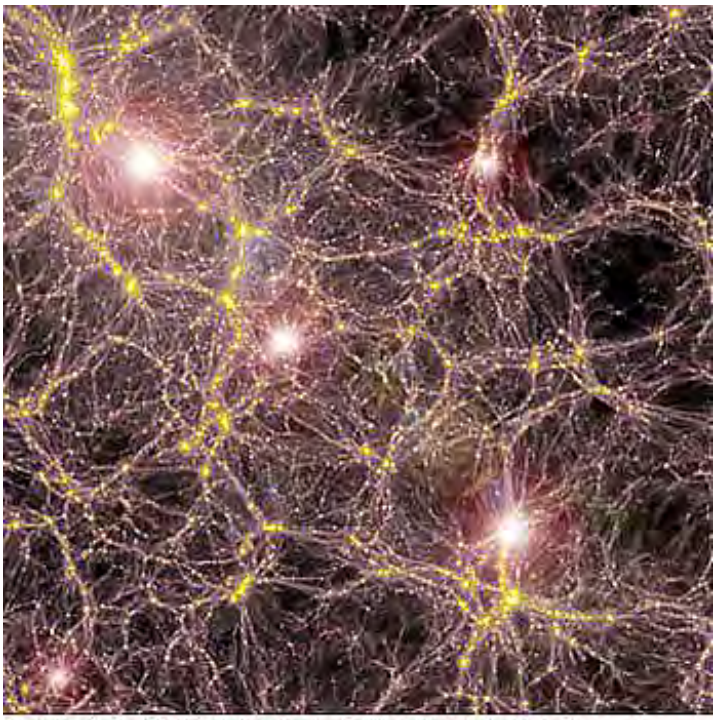
B33 is better known as Barnard's nebula which, when enlarged in this HST image, produces one of the most "popular" of images from that telescope, given the nickname of "Horsehead Nebula". Here are three views:

Returning now to more general considerations, one model ("top down") of early galaxy evolution considers a cloud to fragment into star groupings as it develops from hot dark (radiating but not luminous) gaseous matter. Another galactic model ("bottom up") begins the process with localized multi-star formation from cold dark (low levels of EM radiance) matter, with subsequent aggregation into fewer stars that grow mainly by collision (sometimes described as "cannibalism") with one another. Recent observations suggest the bottom up model describes the predominant process.

In the first billion years or so (the oldest galaxy found so far became organized about 400,000 million years after the Big Bang) of the Universe, as galaxies developed, models for their spatial configuration may have looked something like this computer-generated simulation of filaments within which gases of varying density (high = yellow; lower = blue) lead to organization into individual or clusters (see below) of galaxies. :



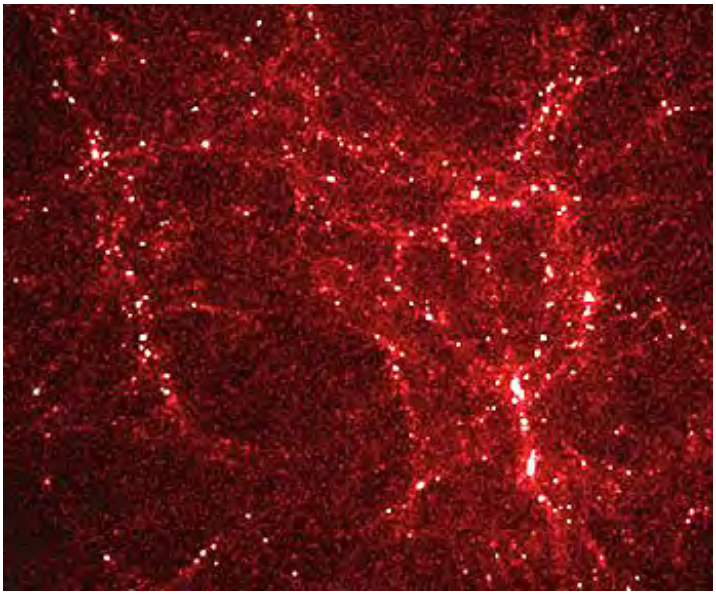
Another similated model, again highlighting filaments of Hydrogen gas, shows a similar pattern.



Christine Lafon (CfA), Volker Springel (Max Planck Institute for Astrophysics), and Lars Hemquist (CfA).

Several points made in the press release accompanying this illustration: 1) the development of filaments establishes connections between zones of higher Hydrogen concentration; 2) this pattern is in part related to the much smaller size of the Universe at the time, with greater density of Hydrogen; 3) as this stage progresses star formation is very rapid; 4) some stars grow to sizes of 200 or more times the mass of the Sun (roughly twice as large as the biggest stars observed today); 5) these stars burned under conditions that led to nuclear reactions that synthesized elements up to iron in atomic number (discussed on page 20-7), with iron itself being abundant; and 6) such massive stars rapidly exhausted their fuel and exploded violently as supernovae (page 20-6), so that as more advanced forms of galaxies evolved the stars comprising them contained varying amounts of the elements heavier than Hydrogen and helium (later stars and galaxies were even further enriched in these elements as burning-heavy element production continued to add the heavier elements to the gases and dust from which galaxies developed and more stars emerged and larger ones "died").

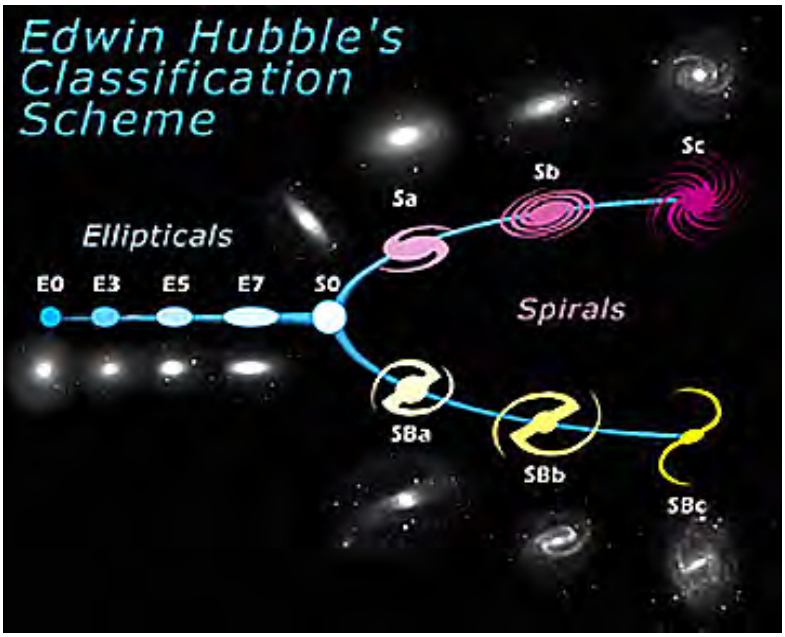
That such filaments actually exist is suggested by this HST view of a very old network of filamentous galaxies and stars in deep space.



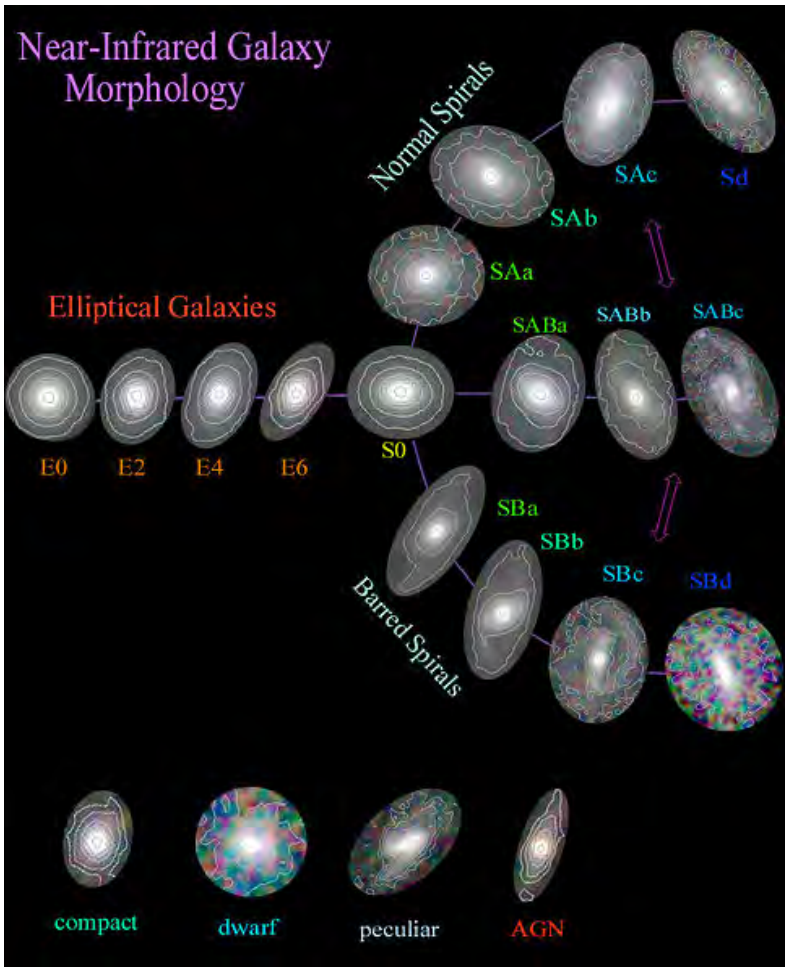
Thus, star formation that goes hand-in-hand with galaxy evolution is a general process that can

take place wherever widespread-to-local concentrations of dominantly Hydrogen gas produce clouds of matter of sufficient density to initiate gravitational contractions. Typically, only a few percent of a cloud's mass will be organized into stars.

There are four general types of galaxies, classified by their geometric shapes (morphologies) and distributions of the stars that comprise them. These are 1) Spirals (the most common), 2) Ellipticals, 3) Dwarfs, and 4) Irregular. Most astronomers add a fifth type - Lenticular - intermediate between Spirals and Ellipticals. The major forms are indicated, with their symbols, in this diagram (the Dwarfs and the Irregular or Peculiar groups are not included but are discussed below). In Hubble's time, opinion favored a left to right evolutionary trend, i.e, ellipticals may (but do not necessarily) morph into spirals. Today, whatever changes occur are from right to left. As mentioned on this page, one process involves collision of two spirals that removes the arms, builds up the central core, and leads to an elliptical.

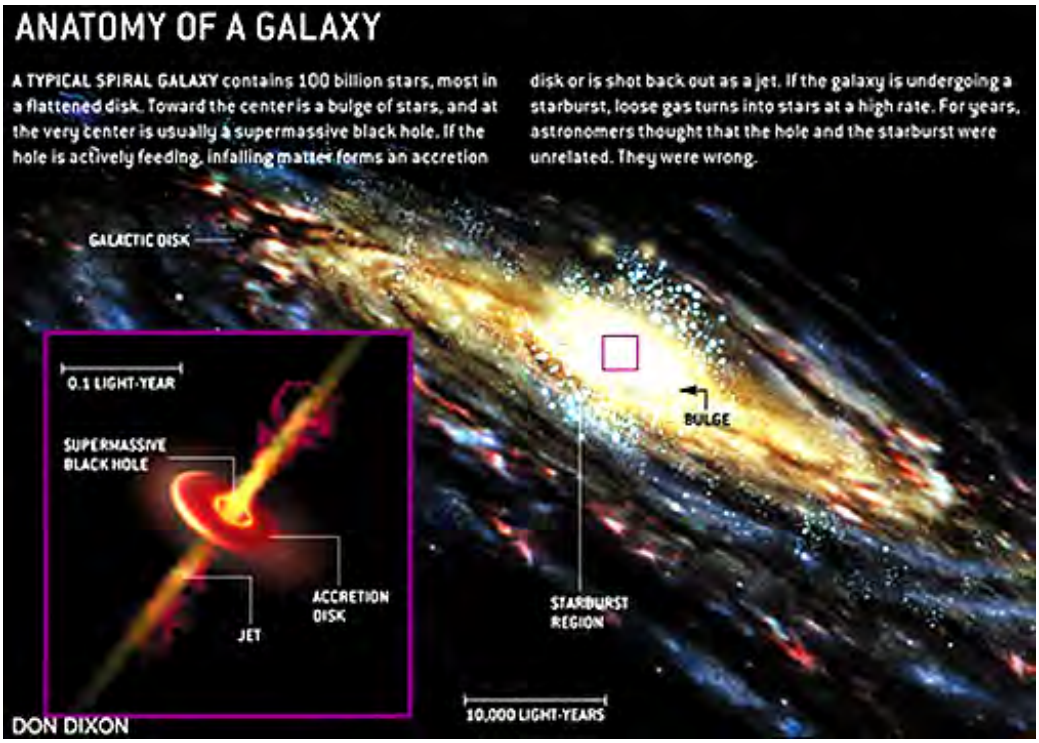


The Hubble classification has remained essentially intact to the present but has been refined to include some non-mainstream types. This is an up-to-date version:



Before looking at each type in detail, it should be mentioned that another classification scheme is currently evolving and gaining favor with some astronomers. This is based on the amount, distribution, and activity within the gas and dust that comprises the interstellar part of a galaxy. The dust behaves in a diagnostic way in infrared light - it both absorbs and emits light in those wavelengths, thus bringing out characteristics not seen in visible light. Elliptical galaxies have low dust contents. Spiral galaxies contain far more dust than the other types: Here is an HST optical image of NGC5746 which suggests some dust and a Spitzer Space Telescope (IRAS) spacecraft image of that galaxy which shows an abundance of dust, shown in red, owing to its thermal radiation in the infrared.

In our discussions of galaxies, we will stick with the standard classification based on visible light morphology. *Spiral galaxies*, which seem at present to be the dominant type, consist of stars arranged in a flattened disc wherein younger (blue) stars are strung out in several prominent spiraling arms that emanate from a central nucleus or bulge that is comprised of a denser collection of older (yellow to orange) stars. Compared to the entire Universe - with both galactic and intergalactic components filling the space - this central core is about 100 billion times the density of the Universe as a whole (this also applies to elliptical galaxies described below). Typical spiral galaxies, such as those shown below, are about 100,000 light years in diameter; disc thicknesses are less than 10,000 l.y. The disc shape results from a greater degree of collapse in one direction and a significant transfer of angular momentum to the disc arms as a effect of tidal (gravitational) interaction with nearby galaxies (clots of dark matter). Spiral galaxies slowly rotate; the galaxy containing the Sun completes one full revolution about its center in 200 million years. Stars closer to the center move faster than those further out, which contributes to the bending that makes up the spiral arms. This general diagram (artist's conception) of a spiral galaxy shows its principal parts; note the central region labeled "bulge" - this is often associated with AGNs described below:



From *The Galactic Odd Couple* by Kimberley Weaver, Scientific American, July 2003

One of the Sa types, that is characterized by a minimum of spiral arms, is this one, found in the

Draco constellation:

This HST image shows the well-organized spiral galaxy NGC4414 (NGC refers to New General Catalog, one of several systematic listings of stars and galaxies observed through telescopes), one with multiple arms in which much gas and dust still remains:

Another spiral, with prominent dust and red to blue stars in its arms, is M51, the Whirlpool galaxy.

NGC1232 is one of the most "perfect" spiral galaxies yet imaged; it lies 100 million light years from Earth. It has 6 distinct spiral arms, each separated by regions of low star density. Some consider it a twin to our Milky Way. Here it is imaged by the European Southern Observatory (ESO) telescope, using UV, Blue, and Red band images to make this color composite:

A different impression of NGC1232 is given by this ESO Visible-Infrared telescope image, which indicates that the galaxy is warmly glowing:

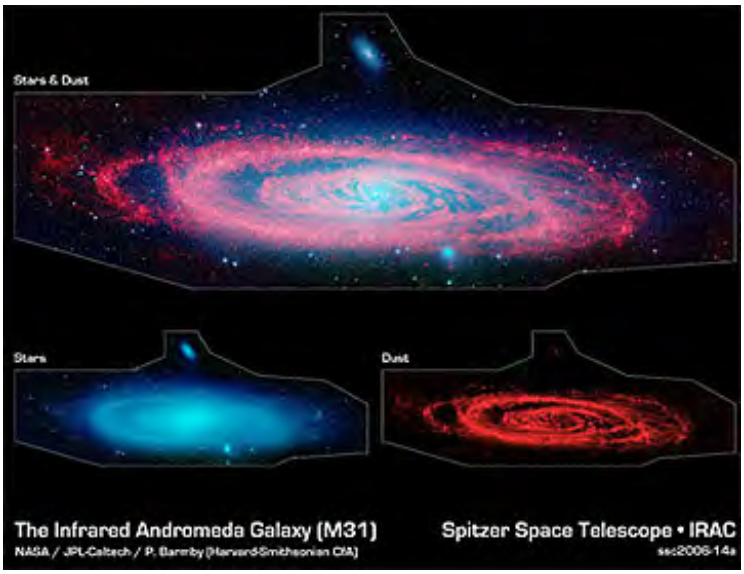
The relative "thinness" of a spiral galaxy is evident when it is oriented so as to be seen "edge-on", that is a side view looking parallel to its spiral plane. NGC4013, 55 million l.y. away, shows this perspective. Note the large amounts of cosmic dust which masks most of its stars.

The dust in the outer arms is apparent as a band in this edge-on view of the Sombrero galaxy:

Spiral galaxy M64 has an anomalous outer ring of dust. The stars in the inner arms are rotating in one direction (clockwise) whereas the outer dust is rotating in the opposite direction. One explanation for this seemingly contradictory behavior is that the dust is not part of the original galaxy but has been captured and continues its previous rotational direction; star formation is especially frequent in the shear zone between the two rotating subsystems:

One of the most famed galaxies is Andromeda (M31) - largely because it is closest to Earth (about 2.2 million light years), can be seen by the naked eye, and thus its structure is visible in telescopes used by amateur astronomers. Here is an HST image of Andromeda:

This next image of Andromeda was made by Spitzer's infrared cameras at two wavelengths. By utilizing different spectral bands, star distribution can be separated from dust. The blues represent large active stars; the reds are caused by dust (containing polycyclic aromatic hydrocarbons) that are distributed in spiral bands from which stars will later grow:



A spiral galaxy can contain up to 2-3 hundred billion individual stars; a few have even larger numbers. Around this type of galaxy are lesser numbers of stars, scattered and isolated or in globular clusters (but still well into the millions) arranged in a "halo" that extends for thousands of light years above and below the plane of the disc (see below). However, the bulk of the mass within the halo, with its important gravitational effects, is not luminous and is now presumed present as Cold Dark Matter (CDM; discussed again on page 20-9). Thus, the halo is often referred to as the Dark Halo. Its importance in galactic evolution and stability is discussed near the bottom of this page.

Spirals can develop unusual distributions of stars outside the disc; in the next example a ring has formed around NGC4650A that could be part of a second galaxy that has collided with the obvious spiral, stripping off stars from that galaxy's spiral arms.

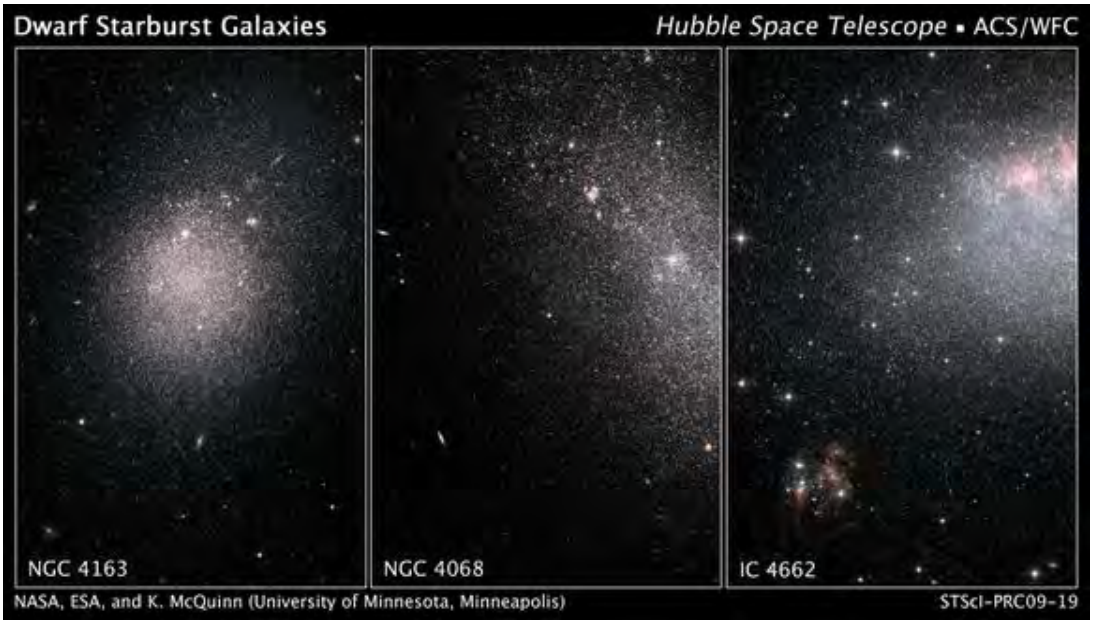
However, such protusions perpendicular to the galactic plane can show a compositional difference. In this view of galaxy M82 (see page 20-4 for additional images of this galaxy), the reddish material moving away from the plane is excited Hydrogen in much richer amounts than within the galaxy which here shows as bright blue from its myriads of stars.

A recent combination of a Hubble image and a ground-based telescope image shows the red areas in the above image of M82 are actually composed of jets of Hydrogen moving at humongous speeds of 1.6 million kilometers/hr (1 million mph)

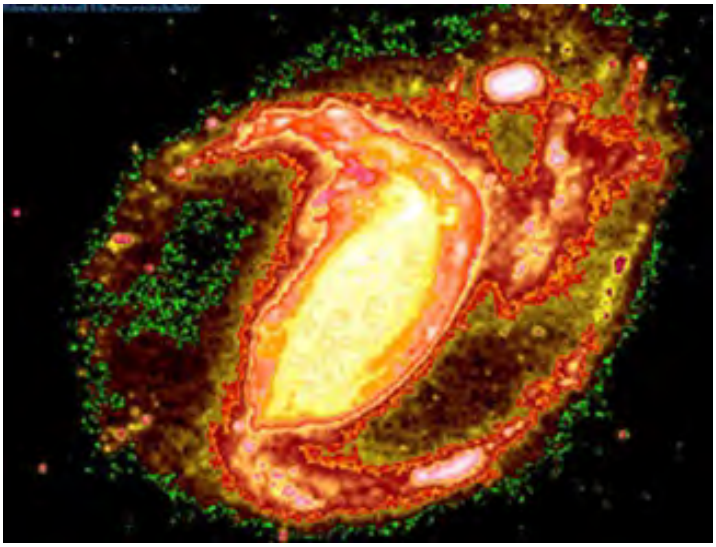
M82 (sometimes called the Cigar Galaxy because of its shape as seen head on) is the prototype

of what is known as a Starburst galaxy. Stars are being produced at much higher rates than normal (commonly this results from a collision with another galaxy [page 20-4]). In this rendition, M82 appears as a nearly continuous white which suggests much above average brightness that would ensue from ultrahigh star formation rates:

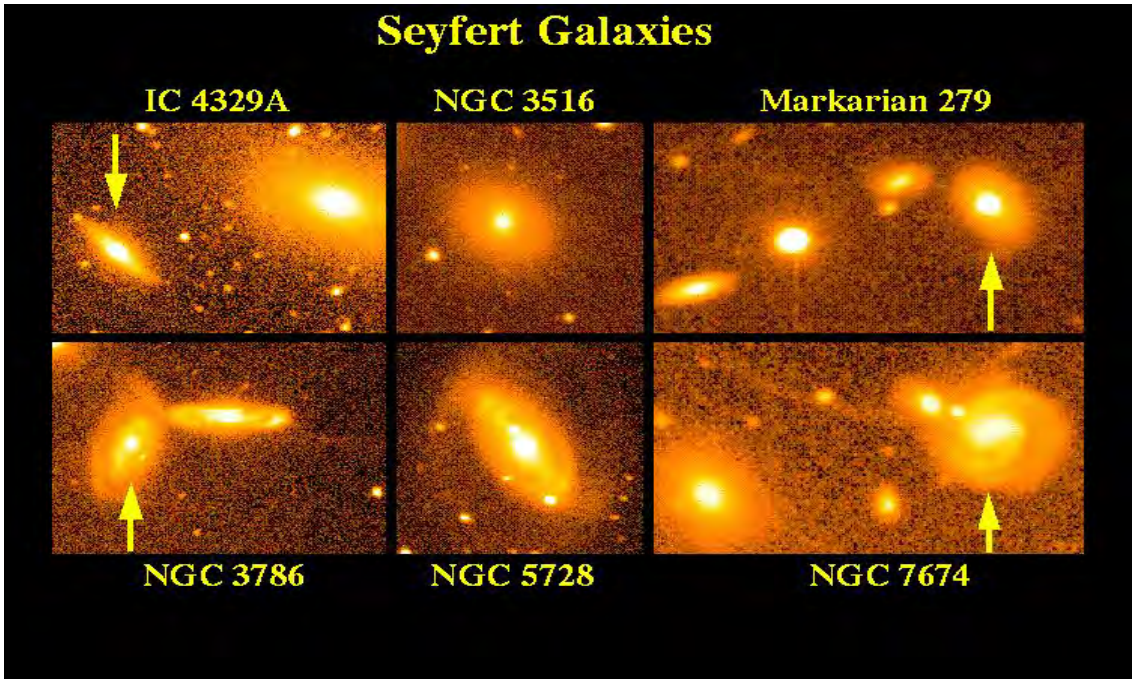
Recent studies have shown that a Starburst galaxy may experience long periods of normal star formation rates, then it undergoes much higher rates for 100 to 200 million years, followed by relative quietude. Here are three Starburst galaxies involved in this study:



Many spiral galaxies, including our Milky Way, have an increased number of stars emanating in a narrow zone directionally from their centers. These are known as *barred galaxies*. Two renditions of NGC1097, a type example of barred galaxies, are shown below. The bar effect depends to some extent on the orientation of the galaxy as viewed. The greater population of stars in the bar segment represents greater production outside the core, with the stars being drawn out as the spiral arms develop. The importance of barred galaxies is considered on the next page.



About 2% of spiral galaxies contain an especially bright central region (an AGN, see page 20-5). These, known as *Seyfert galaxies*, are marked by a notable concentration of dispersed ionized Hydrogen gas, excited to high levels luminosity, i.e., the brightness is not just from stars alone (those present tend to be blue [relatively young]).



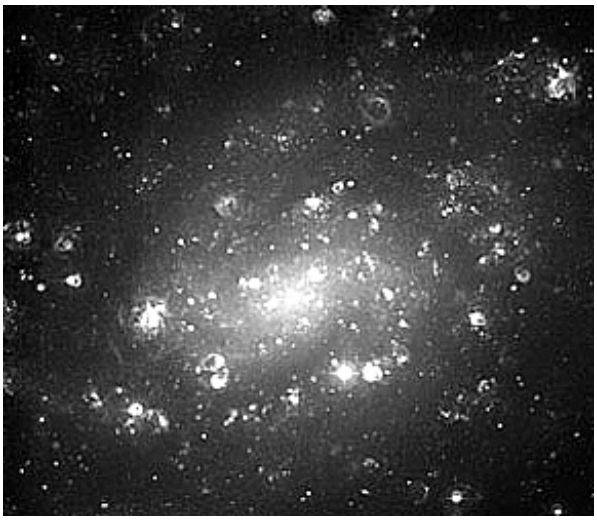
This central region emits radiation that gives rise to strong, broad spectral lines. This spectral signature is similar to, but distinguishable from, a typical quasar (see page 20-6). The cause of the glow may, as is also the case for quasars, be a Black Hole at the galaxy nucleus (there is growing evidence that Black Holes are generally present at the center of spiral galaxies). This glow probably emanates both from a much denser concentration of stars and from excited gases. The Seyfert class is one that has an Active Galactic Nucleus (AGN), whose trademark is that it is a strong radio wave source (however, most radio galaxies are elliptical). The core of an active Seyfert galaxy (in the Constellation Circinus) at a distance of 13 million light years from Earth is a very bright AGN. The greens and reds are excited states of Hydrogen gas presumably heated by radiation from the Black Hole.

A large AGN dominates this next galaxy (the Pinwheel) which contains a thick circlet of stars (a Starburst) just beyond the dense interior concentration of stars and outwardly scatterings of dispersed stars in the galactic plane but without well developed arms.

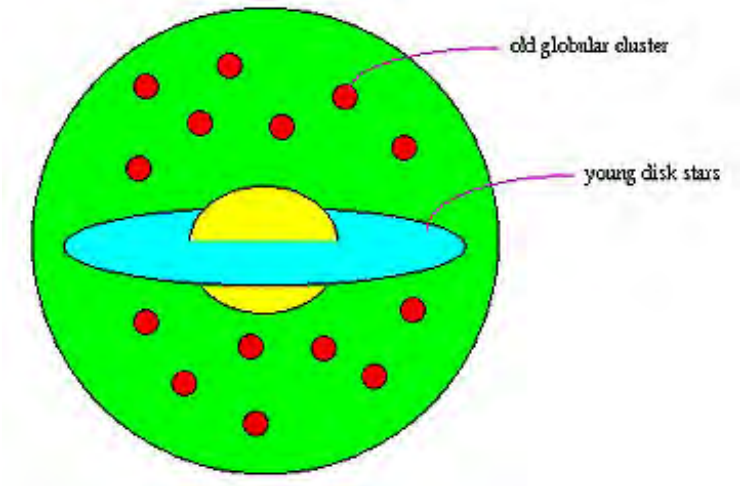
AGNs are a minority in both spiral and elliptical galaxies, but they are the source of extreme energy output. Within them almost exclusively are the quasars (see page 20-6) that are the visible manifestations of matter falling into Black Holes. There is growing evidence that supermassive B.H.'s are at the center of most (perhaps all) larger galaxies that have a bright central bulge. AGNs reached their peak around 4 billion years after the Big Bang, having taken some time to build up to the condition in which huge energy outputs result from their numbers of quasars, and are less frequent in younger galaxies. Although still not proved from observations, many astronomers believe that one or more AGN episodes took place in both elliptical and spiral galaxies at some stage(s) of their histories. The relationships between AGNs and Starbursts

(described later), and their mutual association with Black Holes, will be established on page 20-4.

As alluded to in previous paragraphs, between star groupings in the arms and central region of spiral galaxies there remains much Hydrogen gas and dust in large clots from which more stars will form later. The gas is ionized (H_{II}) and radiates at several discrete wavelengths. The Wide Field Imager (WFI) of the 2.2 meter MPG/ESO telescope at the southern hemisphere La Silla Observatory has imaged the spiral galaxy NGC300 with a filter that selectively passes ionized Hydrogen radiation, so that the stars are screened out leaving only the Hydrogen clots. As seen below, these clots are irregular in shape but widespread:

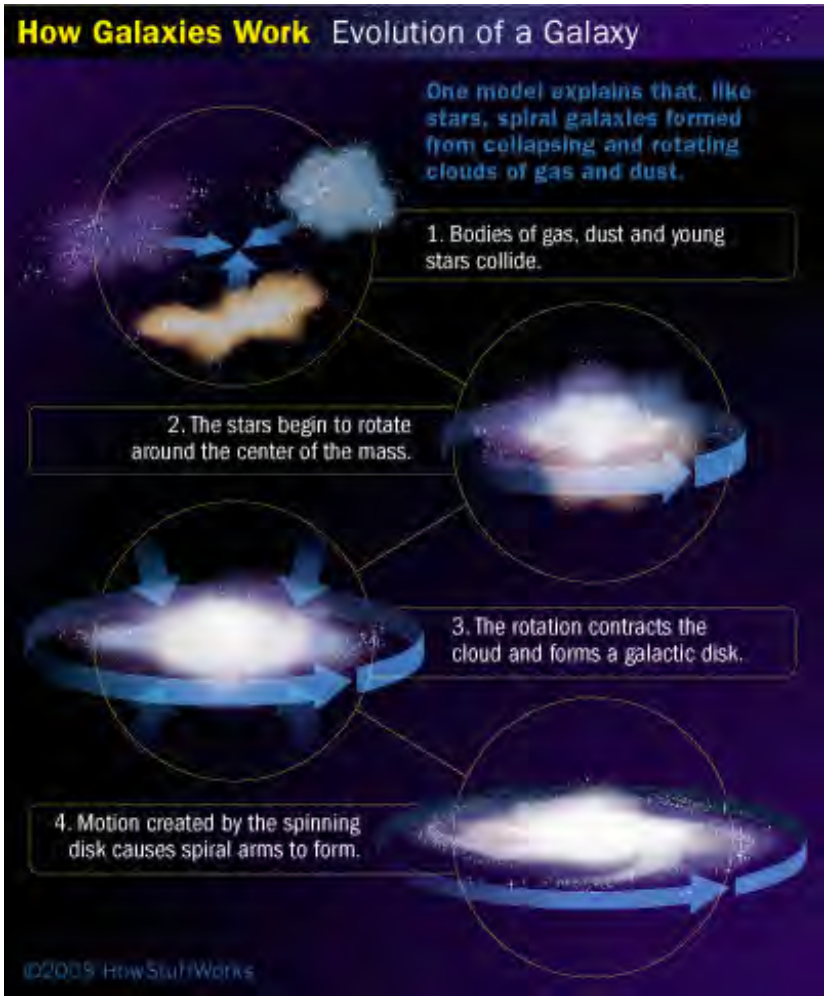


Sky surveys (especially with the Hubble Space Telescope) indicate that spiral galaxies contain a large number of individual stars, clusters, and even small satellite galaxies, and considerable Hydrogen gas and dust, dispersed in galactic space around the central disk in what is called the halo region. A halo appears to be a roughly spherical envelope that surrounds galaxies in general. Haloes develop around protogalaxies and aid in the subsequent development of each type. The density of gas and dust within the halo space is overall less than that within the central disk. Globular star clusters (see below) are the most distinct entity in this distribution. This next figure is a simple diagram of the four principal components of spiral galaxies; the green marks the halo region:



The galaxy NGC 5746 possesses a distinct gas halo, seen in this HST image as a uniform blue region. Its population of halo stars appears to be low.

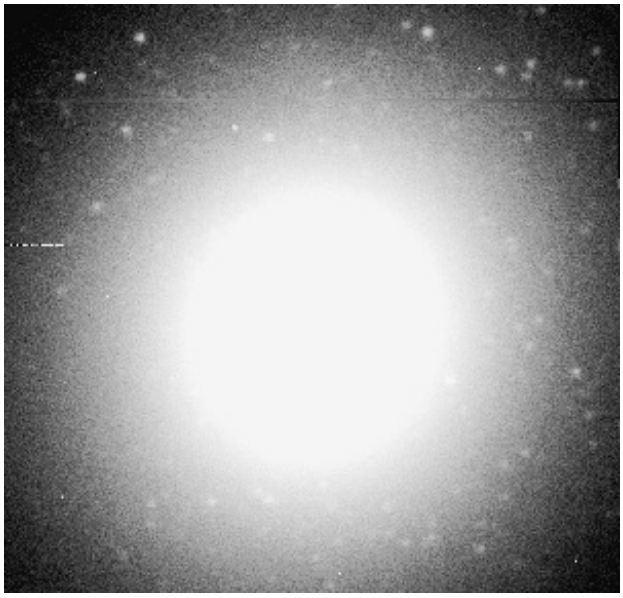
The genesis of the spiral type of galaxy is fairly well understood. It starts with gravitational action within a denser part of the intergalactic medium. Dark matter exerts a control over the resulting collapse of hydrogen gas in the protogalactic molecular cloud. This collapse tends to be asymmetric (uneven). Turbulence within aids in setting the contracting cloud into rotation. When dense enough, the gas begins to organize into individual stars. The cloud further contracts preferentially in one direction and a disc shape results from centrifugal flow within the star and gas assemblage. The angular speed of the stars rotating about a center varies with distance outward; this imparts a curvature to the pattern of stars within the disc. The stars tend to locate in streamers that become the arms of the spiral galaxy. This diagram is a simplified version of this model (there are variants involving collisions, shock waves, density waves, etc. that modify the basic idea):



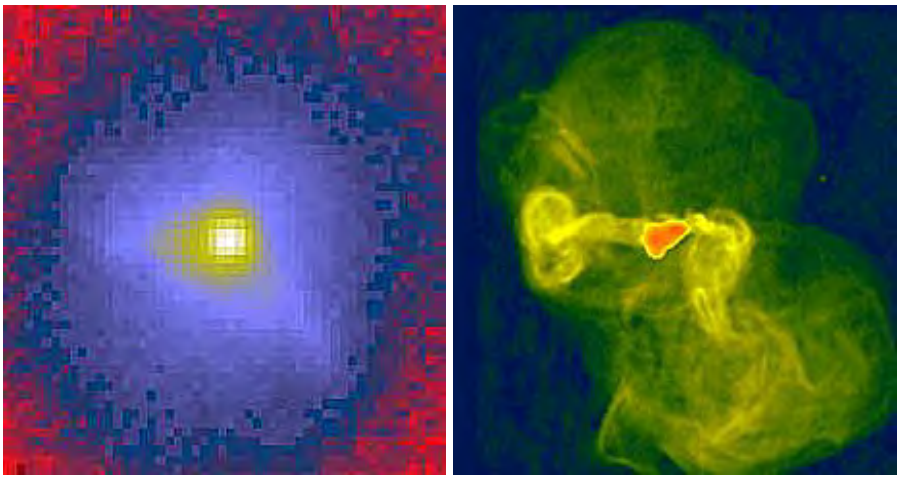
Also enclosed by a dark matter halo is the second major galactic type, the *Elliptical Galaxy,*, marked by mostly old stars (populations up to 10^{11} individuals). Ellipticals comprise about 15% of regular types. Such a galaxy is now believed to originate through collisions, tidal disruption and other interactions, between small galaxies or even large spirals. leading to merging and destruction of the spiral arms (some ellipticals may have formed in the early Universe simply by a

collapse mechanism still poorly understood). Elliptical galaxies rotate more slowly than spiral ones, so that the tendency to evolve into a flattened disc is thwarted. Elliptical (the majority are almost spherical) galaxies, generally more massive than spiral galaxies, usually occur in groups or clusters. Both Giant and Dwarf varieties are known. The typical elliptical galaxy contains a larger percentage of red stars than found in spiral galaxies (those have more blue or hotter stars than red); however, being more compact ellipticals are usually brighter than spirals. Recent observations of elliptical galaxies have found that there are still many younger blue stars. Elliptical galaxies, although more massive than spirals, contain much lower amounts of dust and are gas-poor which suggests that overall they contain a larger fraction of older stars than in the more abundant spirals. Here is a typical example; in this and many other images of ellipticals, the individual stars are not resolved - they are so densely packed that the galaxy image appears to be uniformly bright:

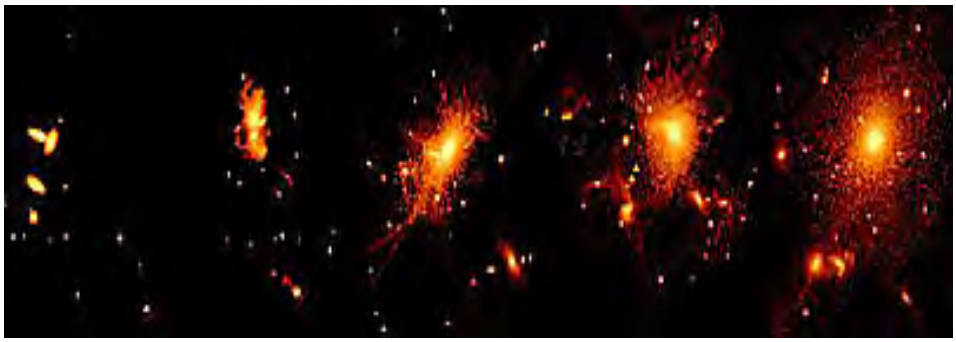
The Giant Elliptical Galaxy is probably the brightest of any category of galaxies. It can contain as many as a trillion stars. Among the best known is Messier 87 (M87) shown below as seen in visible light.



Giant Ellipticals are strong sources of radiation beyond the visible range (discussed on page 20-4). Although we are "jumping the gun" a bit, it is instructive to show M87 as an X-ray source (detected by Rosat) and as a Radio source:

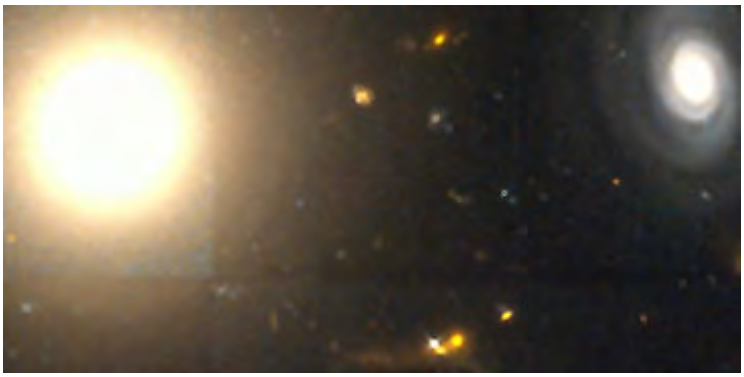


As may be the case for most elliptical galaxies, which holds that many (most) of these form by collisions (see below), the Giant Galaxy type almost certainly results from multiple elliptical galaxy collisions, as depicted in this simulation:



Both spiral and elliptical galaxies can group in clusters from a few tens to hundreds of thousands of individual galaxies. Such clustering is a direct consequence of the uneven distribution of Dark Matter into halos that developed soon after the Big Bang. The image below is a giant cluster of many hundreds of elliptical galaxies (most burned to the red stage) lying at 9 billion light years from Earth. The cluster density makes for an apparently huge single entity but is actually the glow from many close-spaced ellipticals. The image is a composite of ESA's XMM-Newton X-ray space telescope and ground imagery taken through the European Space Observatory (Vis-IR) telescope in Chili.

This HST view (within the Coma Cluster) shows a Giant Elliptical Galaxy on the left and a rather diffuse Spiral Galaxy on the right; being at similar distances the relative sizes are valid:



Elliptical galaxies tend to occur in clusters of this one type, but with a few gas-poor spiral galaxies within a cluster. Spiral galaxies are more scattered in space.

Rare among these principal galaxy types is the so-called Ringed Galaxy. This example is known as Hoag's Object, found in the constellation Serpens and situated about 600,000,000 light years from Earth. In size, its diameter is 120,000 l.y., slightly larger than the Milky Way. Its central nucleus consists of densely packed yellow (old) stars which together resemble an elliptical galaxy. The ring consists mainly of younger blue stars. In the gap in between there is a dearth of stars of either type.

The origin of Ringed Galaxies is still uncertain but a stage of redistribution after the collision of two galaxies is a plausible explanation.

A second example has brought out a bit of humor among the astronomer clique. Look at this Hubble image of Ringed Galaxy ARP 147 (on the right):

Those who viewed this image noted that the spiral galaxy on the left was shaped almost as a "1". This led one wag to rate this scene as "a perfect 10".

There is another galaxy type - the Lenticular Galaxy - that some consider deserving of its own category. Generally, most such galaxies are equivalent to the SO group at the base of the two spiral branches of the Hubble galaxy classification. In side view, a lenticular galaxy is just that - a double convex shape, much like an optical lens. When seen face on (as from the top), the SO type has no distinct or discernible individual spiral arms but in the part beyond the center (which may be a massive core but a few have very non-descript cores) individual stars are evident but distributed randomly and at various densities. When Lenticular galaxies are examined in this way, they show rudimentary spiral arms, suggesting they are an incipient spiral transitional to that class or are a degenerate spiral no longer typical of that class. Most of their stars are old (yellow) both in the core and the surroundings, which makes this type similar to elliptical galaxies - except for its pronounced disk shape. However, gas and dust seem in short supply, suggesting that little subsequent evolution is likely. Several examples of Lenticular Galaxies are shown in this next sequence; see their captions for details.

About 10% of galaxies are neither spiral nor elliptical but have what can be described as "irregular" shapes. As a type example, here is NGC 1569, about 7 million light years away:

Many of these irregular galaxies are actually two (or more) colliding galaxies (discussed in more detail on page 20-4). Some are described as "peculiar galaxies", a term coined by Halton Arp who compiled a catalog of these in 1966. A classic peculiar galaxy is the pair NGC 6621/6622:

Globular star clusters are intermediate between simple star clusters and galaxies - each is an aggregate of 100,000 to a million stars. These are much like miniature elliptical galaxies but have far fewer individual stars. Like the latter many seem to have a predominance of old stars. The largest concentration of these stars is in the interior of the cluster. Typical Globular Cluster densities are several hundred stars per cubic light-year (compared with typical densities of 0.01 to 0.1 stars per $(ly)^3$, which is the norm for most galactic space). Because of the higher densities

within clusters, the frequency of star collisions (normally a rare event) is considerably greater.

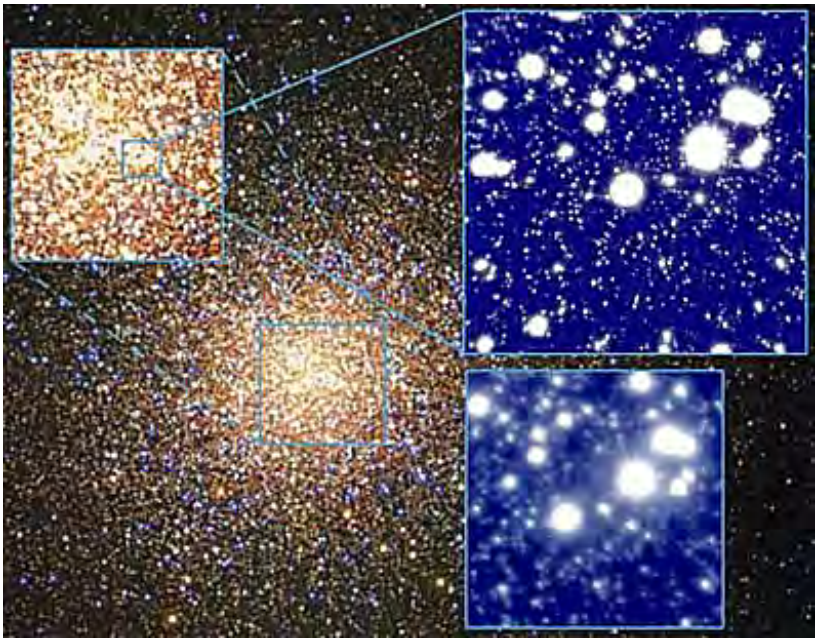
Although some clusters are found around elliptical galaxies, the globular clusters mostly occur in much larger numbers within the halos of spiral galaxies, i.e., in orbits at all angles to the galactic plane within an imaginary sphere that may be 200,000 light years or more in diameter. The Milky Way contains less than 200 such clusters; Andromeda almost 500. Most globular clusters contain old (> 7 billion years) stars. Globular clusters have proved to be a primary means of determining the ages of the oldest stars in the Universe. The halo region around a galaxy also contains millions of isolated (non-clustered) individual stars, or small groupings. Below is globular cluster NGC6093:

The largest globular cluster around the Milky Way is NGC5139, estimated to contain up to 10 million stars.

Some globular clusters are more open and contain less stars, as holds for cluster M3:

The Wide Field Camera on the Hubble Space Telescope has captured a view of just how dense are the stars in globular clusters. Here is an image of a small part of the Omega Centauri globular cluster, just outside our galaxy about 13000 light years away. At least 30000 stars appear in this segment of the cluster; most of these are similar in size and luminosity to the Sun, and some of the larger ones (yellow) are Red Giants. The cluster is 12 billion years or older in age. A large number of blue-white stars formed early on have since lost their luminosity as they converted to white dwarfs and neutron stars.

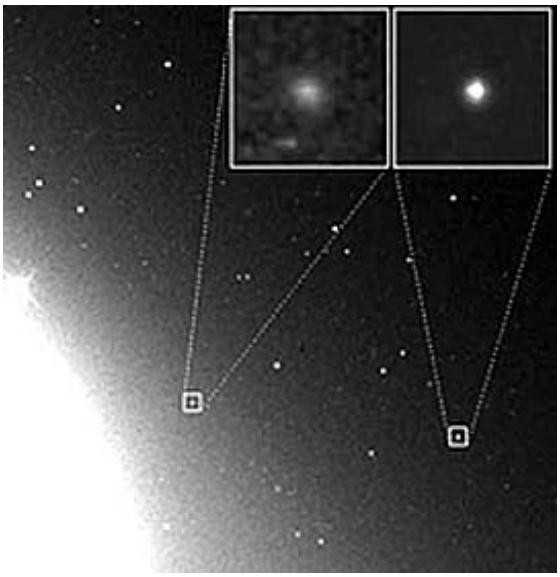
M13, the Great Hercules Globular Cluster, is imaged first in full below; the second diagram shows it again with an inset of the central region in the upper left, and two insets on the right in which individual stars are separated to give an indication of actual spacing:



Astronomers have concluded that globular clusters formed mostly during the early stages of galaxy formation (and since most galaxies appear old, clusters are ancient cosmic features - **they are old**) and then became much rarer as the Universe expanded. It is now known that clusters have been forming continuously over time from denser pockets of Hydrogen in the halo regions. Young(er) clusters have been found around galaxies a few billion light years or less from the M.W. These contain stars with concentrations of heavier elements that could only have reached those levels after many episodes of stellar explosions; thus many of their stars must be young. Star clusters have been observed near galaxies that collide, indicating that one process of formation is related to interactions between merging galaxy pairs. Thus globular clusters probably formed at maximum rates in the early Universe but intermediate age and even young clusters indicate that this component of galactic systems can develop at any time.

Occasionally, telescopes locate masses of small and large stars that are still in the process of organizing into individual globular clusters. Such is the case for this grouping that is part of the Large Magellanic Cloud (a galaxy group discussed on the next page). 30-Doradus is seen here as a color composite with red contributed by x-radiation, blue from UV radiation, and green from ionized Hydrogen gas:

Recently, another class of galaxies has been discovered. Called "fuzzy" clusters because of their appearance, those few found so far occur in the plane of a galaxy rather than well out in its halo (as do most globular clusters), are larger 50-100 l.y. across (globulars are usually 15-20 l.y.), and consist of dominantly old red stars. Here is a view made by the HST and supported by Keck Telescope observations that shows a fuzzy cluster on the upper left; a farther out globular cluster appears to its right.

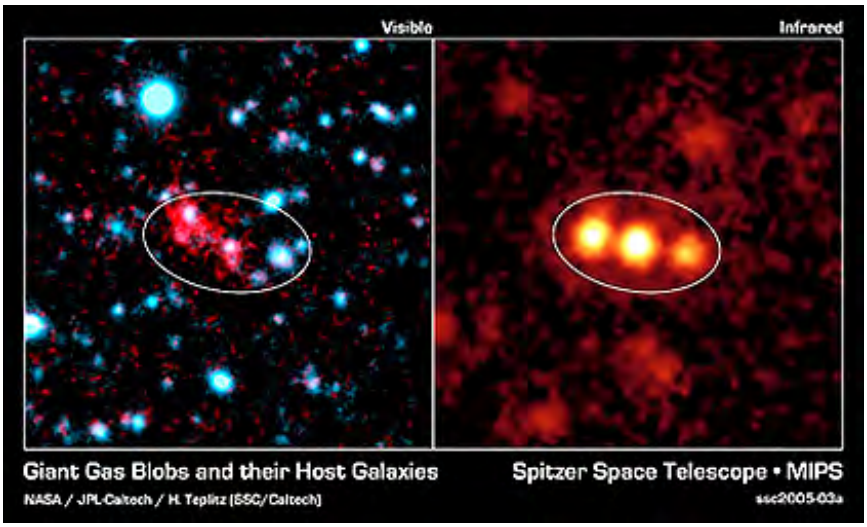


Still another category of globular clusters, long predicted, as at last been imaged and verified. This defining image is shown below; note the four small boxes:

These clusters are not associated with galaxies as is the usual case. They are isolated in intergalactic space, a fact that has led them to be called "orphan clusters". They contain up to a million stars. Although only a few have been detected so far, they likely are fairly common throughout the Universe. The favored explanation is that they were torn from parent galaxies by other galaxies and dragged into open space; alternatively, they may just be incipient clusters trying to build to full-scale galaxies.

In the Milky Way, there are star clusters (mainly in the halo) that consist mainly of old stars. There may be only a few thousand, up to 100000, individual stars. The grouping is controlled by mutual gravitational attraction. This type is called an Open Cluster. The Pleides, shown on page 2a, is one example. Here is another (M11, the Wild Duck Cluster, about 4 light years away):

The above galaxy types have in common the fact that they show some type of geometric organization. But objects in the Universe that are galaxy-size but seem to be irregular in shape have been given the colloquial name of "blobs". Recent studies of these blobs indicate that when at least some are resolved in infrared light, a number of spherical objects appear. This image pair gives a good example:



The Galaxies subsection is continued on Page 20-2a: Click on Next below.

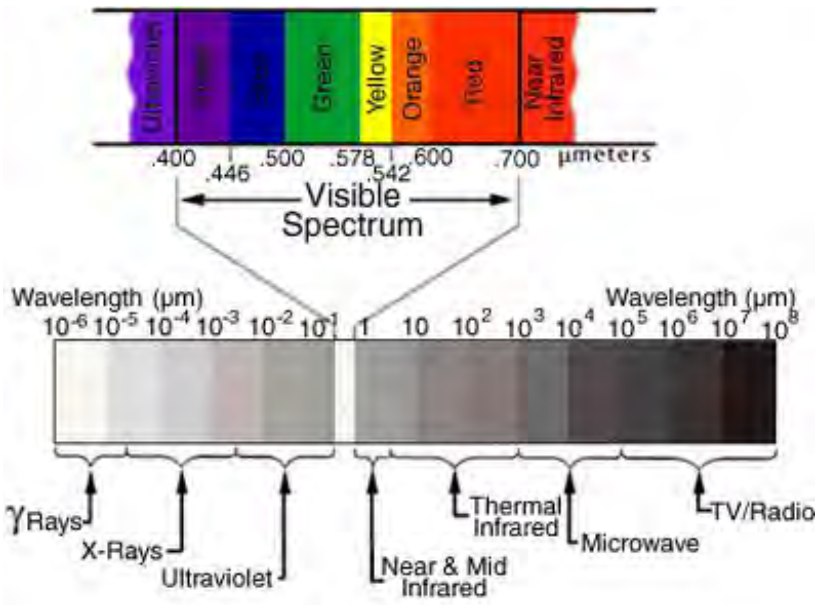
PREVIOUS rst NEXT

Primary Author: Nicholas M. Short, Sr.

Images of Galaxies and Stars outside the Visible Light Range.

Another valuable Internet site that has links to nearly all of the spaceborne telescopes and other instruments used for astronomical observations over the entire spectrum is at a site maintained by Danish Astronomers. Likewise, we repeat the Net address for space observatories that was given on page 20-2. A list, many with Web links, of both ground- and space-based telescope observatories is accessed at this Wikipedia site. Also, you may want to check out the HEASARC website, which describes NASA's High Energy Astrophysics Science Archive Research Center.

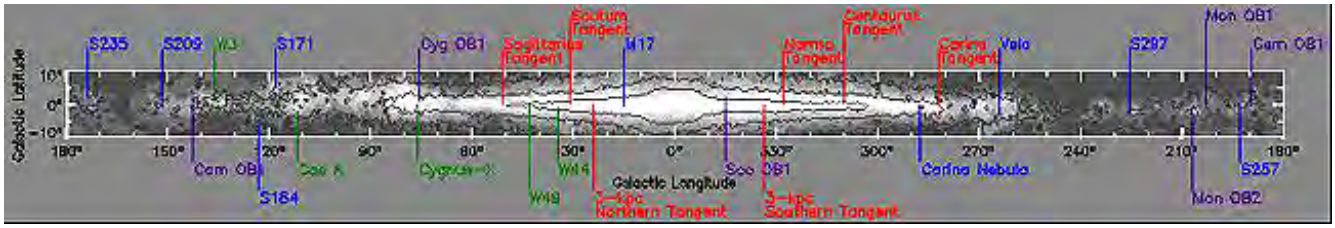
The entire EM spectrum has been utilized to study galaxies, stars, and other astronomical phenomena. As a reminder, here is still another diagram that depicts the full spectrum and its subdivisions. On this page we will examine examples of astronomical images beginning at the shortest wavelengths (Gamma Ray region) and ending at the longest (radio waves):

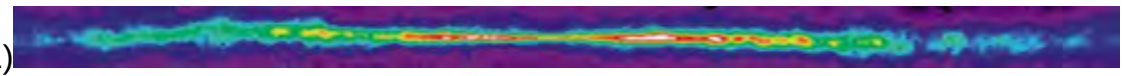


Most "star gazers" feel more comfortable looking at the luminous bodies of the Universe - stars and galaxies - as they appear in optical imagery. But, there is generally much more "illuminating" information about celestial bodies in images depicting energy distribution and intensity of radiation associated with other parts of the spectrum. **In fact, because dust often obscures phenomena within a galaxy when viewed in the visible, the ability to penetrate that dust using other wavelengths reveals many aspects or characteristics of the composition and structure of galaxies that greatly expand our knowledge of the nature of both galaxies and intergalactic space.** The usefulness of examining bodies outside the Milky Way at different wavelengths was earlier demonstrated in the multispectral images of the Crab Nebula shown near the bottom of page I-4 of the Introduction. Both ground and space observatories are operating, or will be activated later, in parts of the EM spectrum beyond the traditional visible light range.

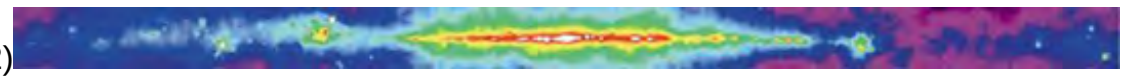
Astronomers at NASA's Goddard Space Flight Center have assembled images taken at various regions of the spectrum by instruments (ground and space based telescopes, etc.) looking at our galaxy the Milky Way, as depicted in this montage. This is not a view of the M.W. taken

externally but one looking towards its center and beyond towards its far edge; also, that part of the M.W. lying in the halo behind us is not included. The layout in each image is reconstructed as continuous; since Earth is within the galaxy, both the parts towards its center and those beyond can be seen by looking outward from various points and times from the Earth. Each image is identified by its imaging wavelength or wavelength interval of the spectrum, together with a brief description of the principal information that is associated with data collected from that region. First shown is an image of the M.W. that serves as a reference map on which useful star markers are plotted. Then, starting from the top, which is also the longest wavelengths, is a full microwave image of the M.W., with those beneath arranged by decreasing wavelengths:



(1) 

(1) **Atomic Hydrogen (1420 Mhz)**: Picks out radiation from excited neutral Hydrogen in interstellar gas and dust clouds.

(2) 

(2) **Radio Continuum (480 Mhz)**: Signal produced by fast-moving electrons; good for spotting sites of now diminished supernovae.

(3)

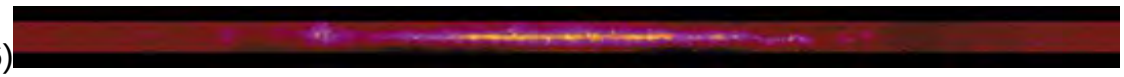
(3) **Molecular Hydrogen (115 GHz)**: Shows distribution of molecular Hydrogen associated with carbon monoxide in cold interstellar matter.

(4)

(4) **Radio Continuum (2.4-2.7 GHz)**: Caused by high energy electrons and associated warm, ionized gases.

(5)

(5) **Far-Infrared (12-100 µm)**: Radiation emanates from dust heated by stellar radiation; emphasizes active star-forming regions.

(6) 

(6)**Mid-Infrared (6.8-10.8 µm)**: Due to excitation of complex molecules in interstellar clouds and in cooler reddish stars.

(7)

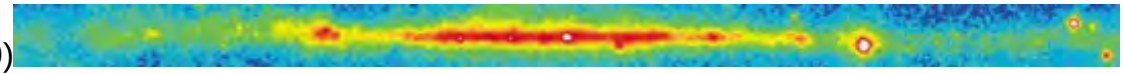
(7) **Near-Infrared (1.25-3.5 µm)**: Reveals temperatures, mainly of Giant, relatively cool stars, and shows the galactic core; dust is "transparent" in this spectral region and does not obscure many luminous features.

(8)

(8)**Visible Light (0.4-0.5 μm)**: Displays primarily nearby stars and thin ionized gas; dark areas cold.

(9)

(9) **X-rays (0.25-1.5 kiloelectron-volts)**: Reveals gases heated by shock waves from supernovae.

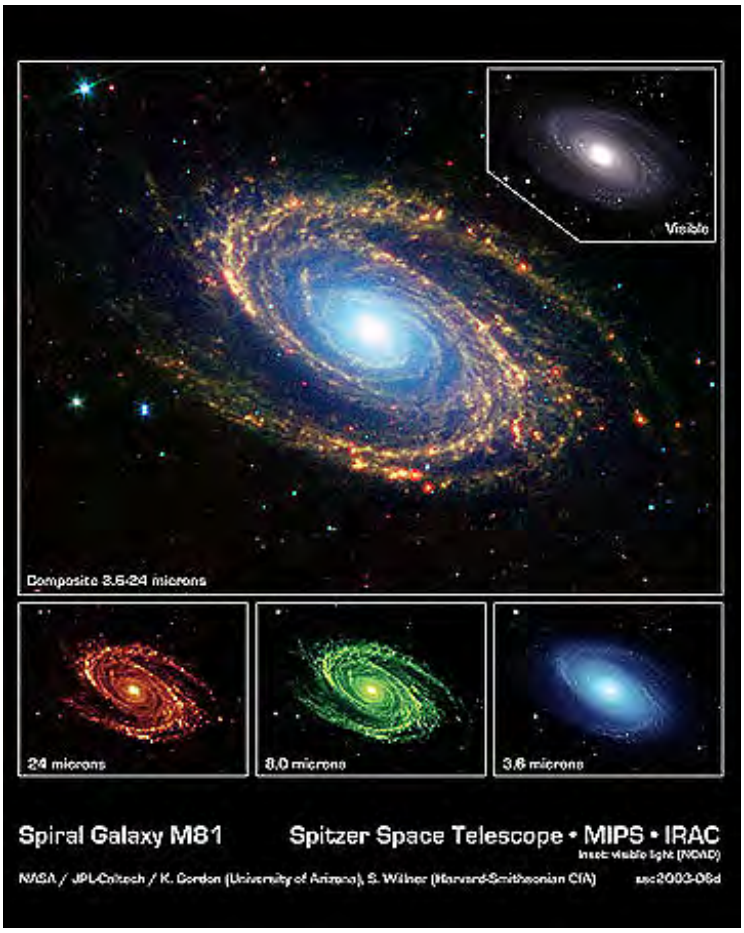
(10) 

(10) **Gamma-Rays (300 megaelectron-volts)**: Pinpoints high energy sources coming from pulsars or phenomena stemming from cosmic-rays.

This idea of imaging cosmological entities at different wavelengths can be further enforced by looking at the montage of five views of the star Centaurus A in the wavelength regions indicated on each panel.

Perhaps you noticed that one part of the EM Spectrum was omitted in the above M.W. sequence, namely, the Ultraviolet. As we shall see below, this segment has useful information.

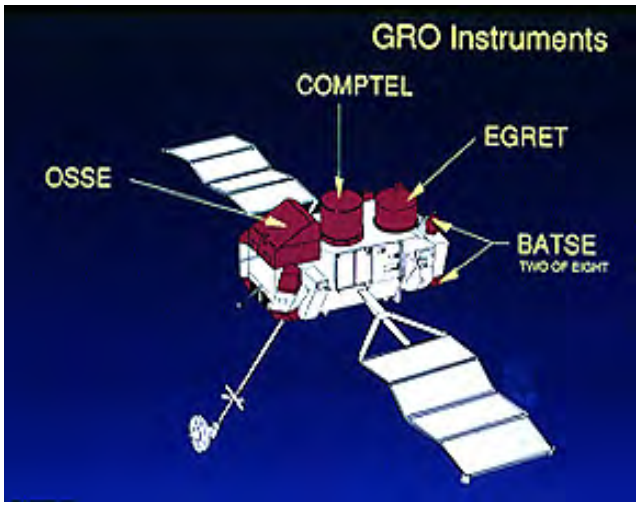
A point to be kept in mind in looking at images below, as well as on preceding and subsequent pages: Images acquired by the same or different telescopes for any of the specific regions of the EM spectrum do not necessarily look the same - some may appear notably different than others because of the way in which the image is processed and displayed (for example, different filters may be used or the image values for intensity may be rendered in color-coded levels assigned different colors). Thus, the same target in the sky, such as a specific star or galaxy, may show up with distinct differences when the image processing choice changes parameters. To illustrate, look at the different-appearing renditions that result when wavebands in discrete spectral regions, such as various parts of the infrared, are utilized, as shown by this M81 panel:



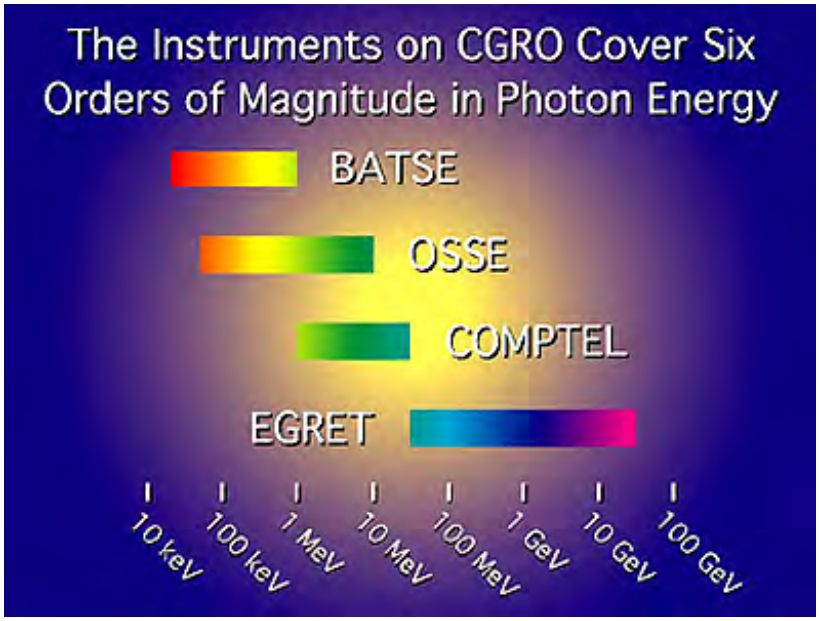
We'll start our survey of cosmic images and data sets obtained at different wavelengths by examining the phenomena obtained using the highest EM energy sources (shortest wavelengths; highest frequencies) - Gamma Rays. Gamma radiation is observable over the entire sky. It may appear as a diffuse glow or as localized (point) sources. Several modes of generation of the Gamma Rays have been considered: Black Hole influence; Neutron star attracting infalling material; supernovae; dark matter itself; collisions between antimatter and matter. One way in which this radiation is observed is in short-live *Gamma Ray bursts*.

The Soviet space program launched the International Astrophysical Observatory "Granat" in late 1989. Its seven instruments monitored stars and galaxies in the Gamma Ray and X-ray regions of the spectrum. It operated until 1998.

On April 5, 1991 NASA launched the Compton Gamma Ray Observatory (CGRO) as a complement to the HST that extends coverage into the short wavelength, high energy end of the EM spectrum. It carried four instruments that could measure radiation whose energies range from 30 MeV to 30 GeV. This huge (central part nearly the size of a school bus) sensor platform has been one of the most productive astronomical observatories orbited so far. It is shown in this artist's drawing:

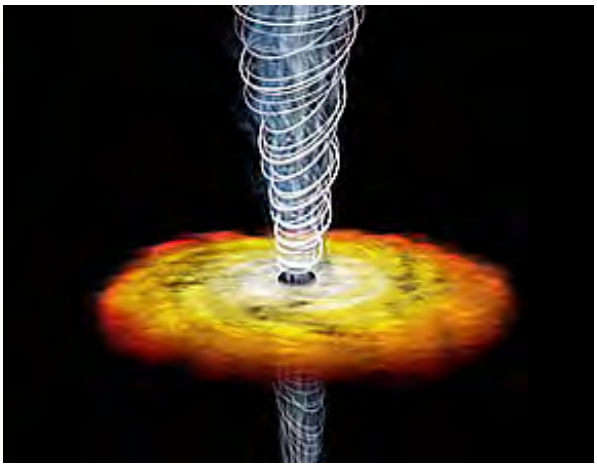


The individual range of coverage by the CGRO sensors is shown in this plot:

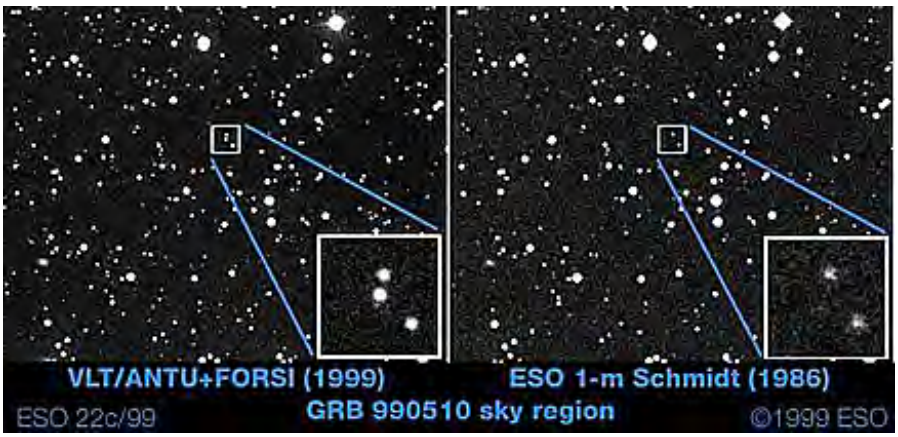


The acronyms stand for BATSE = Burst and Transient Source experiment; COMPTEL = Imaging Compton Telescope; EGRET = Energetic Gamma Ray Experiment Telescope; OSSE = Oriented Scintillation Spectral Telescope. (The CGRO was named to honor Dr. Arthur Holly Compton, an eminent physicist, Nobel Laureate, and Chancellor of Washington University in St. Louis.

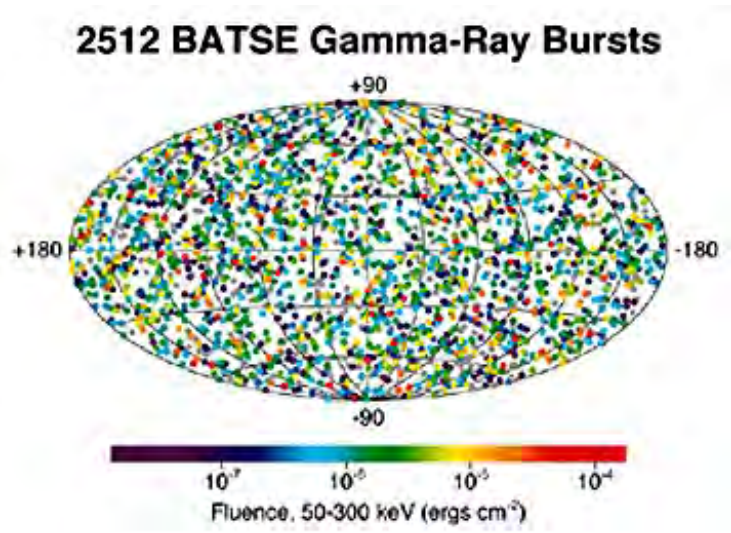
The CGRO was designed to measure radiation associated with stars and galaxies which result from high energy, usually nuclear processes. It looked particularly at supernovas, quasar and pulsar emissions, Black Hole accretions and other powerful stellar processes (next paragraph). CGRO discovered a new class of energetic objects, called *blazars*, that give off energy in the 30 MeV-30-GeV range, but actually produce detectable energy over the entire spectrum. Redshift studies (page 20-9) indicate most blazars are far from Earth and therefore quite old. While distant blazars look like bright single stars, they are actually associated with galaxies in an advanced stage of inflow of copious amounts of stars and gas/dust into supermassive Black Holes (with masses billions greater than the Sun), in so doing generating huge amounts of energy release. This means that the high luminosity correlated with high energy release persists over long-term telescope viewing. Electric and magnetic fields usually carry the luminescent materials as directional jets; thus for one to be seen from Earth our detectors must fall within the cone of a jet to been seen. That mechanism is depicted below, and beneath it is an image of the starlike blazar jet emanating from Markanian 421.



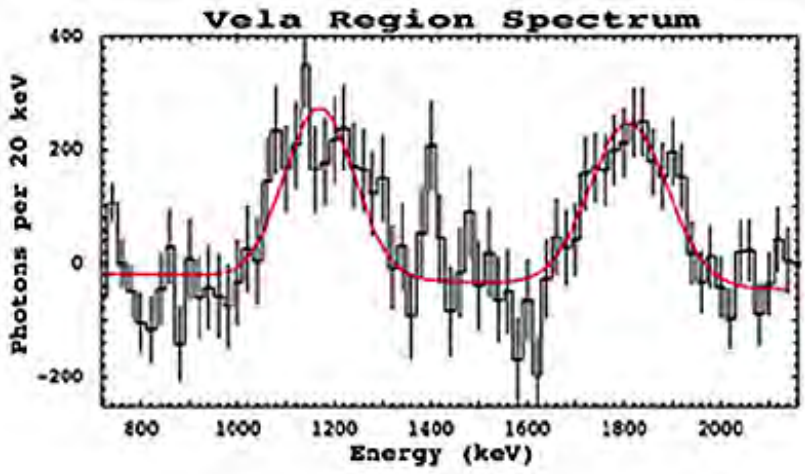
One prime astronomical target of the CGRO was to search for Gamma Ray Bursts (GRBs), which are huge releases of energy that are short-lived and variable, are widespread in the celestial sphere and occur mainly in galaxies. Here is a map of those bursts measured over time by CGRO; the local effects of the Milky Way bursts have been removed. These GRBs will be discussed in more detail at the bottom of page page 20-6. For now we will show one example of a GRB imaged by the telescope at the European Southern Observatory in South America.



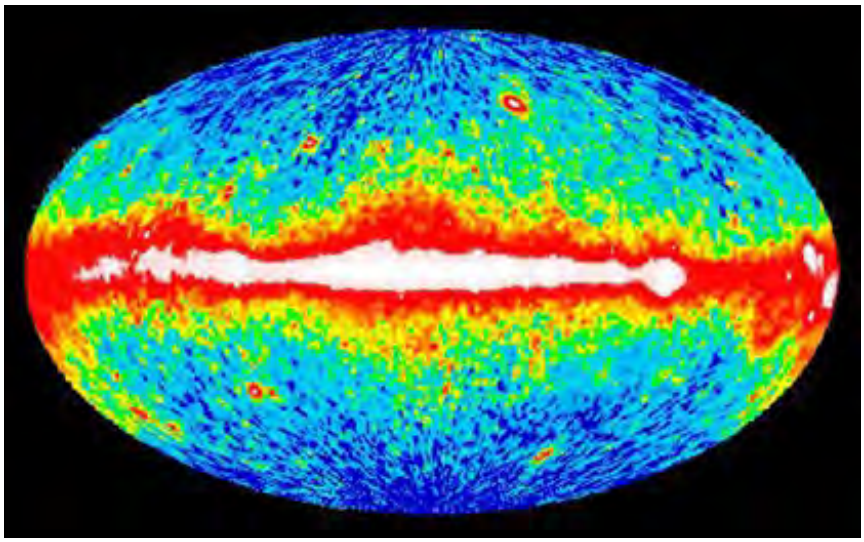
The BATSE instrument has produced the following map of GRBs across the sky:



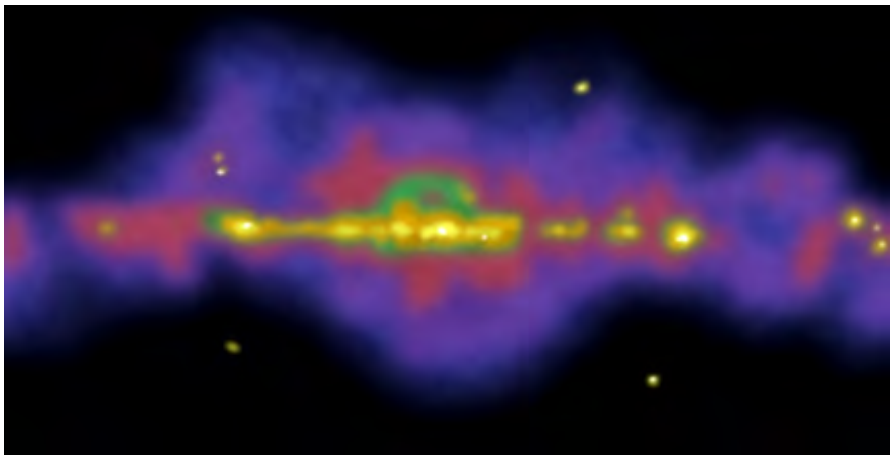
Another GRB was detected in the vicinity of the star Vela. Here is a plot of associated energy levels:



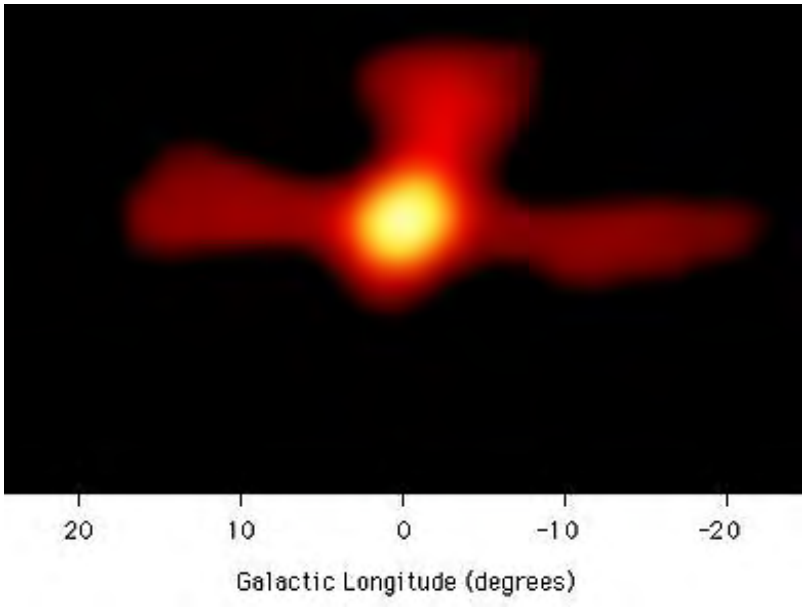
A great deal of information about Gamma Rays comes from studies within the Milky Way. The CGRO has produced this image showing a generalized Gamma Ray energy distribution over the entire M.W. disk:



More detail within this general halo is brought out by special processing, which indicates regions of strong Gamma Rays that may be pulsars or other concentrated but steady sources:



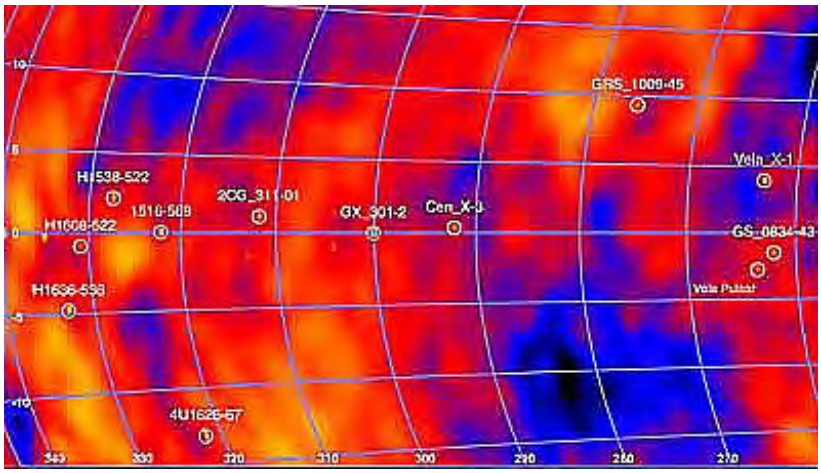
The OSSE instrument on CGRO picked up an unusual distribution of Gamma Ray energy, shown in this figure as the red glow above the Milky Way plane. It has been interpreted as a region in which antimatter (electrons are positively charged [positrons] and protons have a negative charge.) has interacted with conventional matter, releasing a huge amount of energy.



It may seem surprising that telescopes can pick up evidence of radioactivity associated with stellar or galactic material. One radioactive isotope, Al^{26}, is fairly abundant in galactic gas and dust. With a rather short half-life, when it decays it produces abundant gamma radiation. Here is a CGRO image made by the COMPTEL instrument which shows the distribution of this radiation associated with Al^{26} decay.

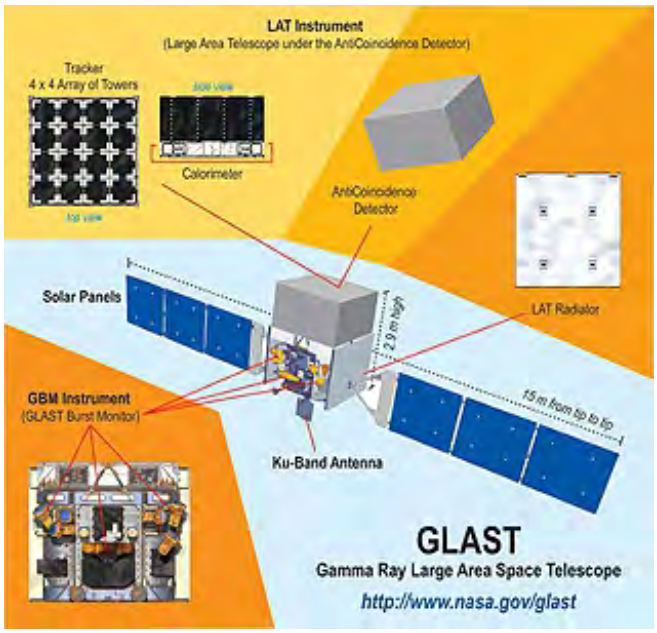
The Compton Gamma Ray Observatory was a major achievement guided by astrophysicists and operated by NASA Goddard. You can learn more about its results, with many additional images, at the CGRO site. On June 4, 2000 the CGRO was deliberatly decelerated so as to enter the atmosphere over the Pacific, as its orbital decay (adjustment fuel exhausted) meant it might fall to Earth at any time soon, possibly threatening populated areas.

An ESA satellite called Integral has made a variety of observations of the M.W.'s galactic center, where the gamma radiation is most intense. This image shows variations in intensity within the center.



Another Integral image is designed to pick out individual major sources of gamma radiation in the inner Milky Way. Some of these might be Gamma Ray bursts but most last longer and may be pulsars.

A more powerful, higher resolution Gamma Ray observor, GLAST (Gamma Ray Large Area Space Telescope), was successfully launched on June 11, 2008. (It has since been renamed the Fermi Gamma Ray Telescope.) Here is a drawing of the spacecraft with its two principal instruments. You can check its status by going to the GLAST website.



The objectives of GLAST are as follows:

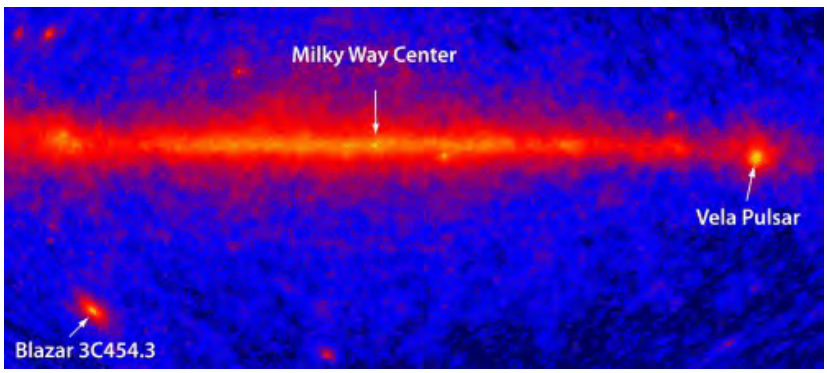
1. To understand the mechanisms of particle acceleration in AGNs, pulsars, and SNRs. This understanding is a key to solving the mysteries of the formation of jets, the extraction of rotational energy from spinning neutron stars, and the dynamics of shocks in SNRs.

2. Resolve the gamma-ray sky: unidentified sources and diffuse emission.Interstellar emission from the Milky Way and a large number of unidentified sources are prominent features of the gamma-ray sky.

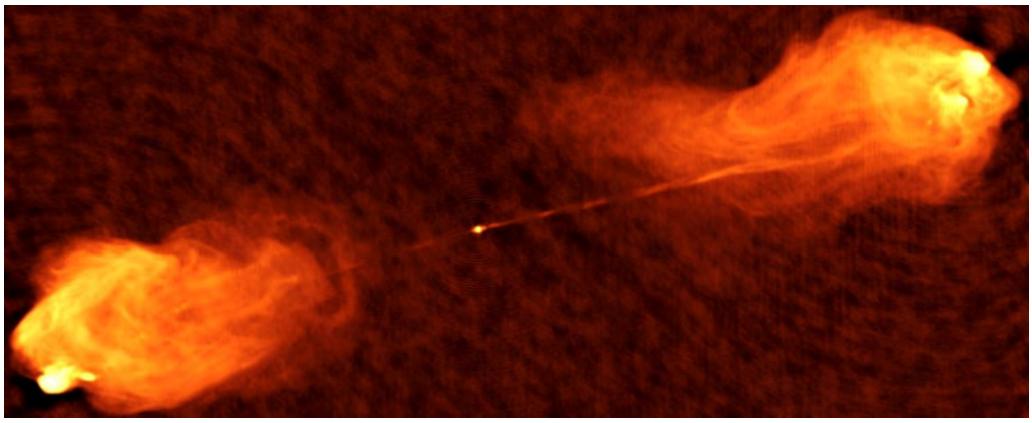
3. Determine the high-energy behavior of gamma-ray bursts and transients. Variability has long been a powerful method to decipher the workings of objects in the Universe on all scales. Variability is a central feature of the gamma-ray sky.

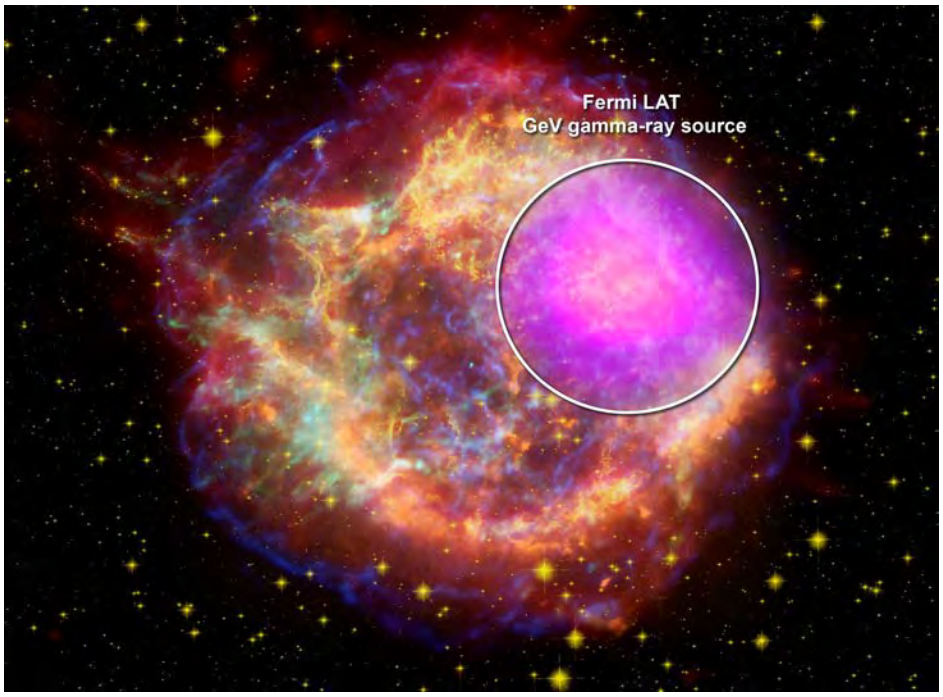
4. Probe dark matter and early Universe. Observations of gamma-ray AGN serve to probe supermassive black holes through jet formation and evolution studies, and provide constraints on the star-formation rate at early epochs through photon-photon absorption over extragalactic distances. There are also the possibilities of observing monoenergetic gamma-ray "lines" above 30 GeV from supersymmetric dark matter interaction; detecting decays of relics from the very early Universe, such as cosmic strings or evaporating primordial black holes; or even using gamma-ray bursts to detect quantum gravity effects.

Scientific results from GLAST/Fermi have been notably slow in release, although data gathering began in August. The most frequently cited result is this image showing the Gamma Ray sky associated with the Milky Way:



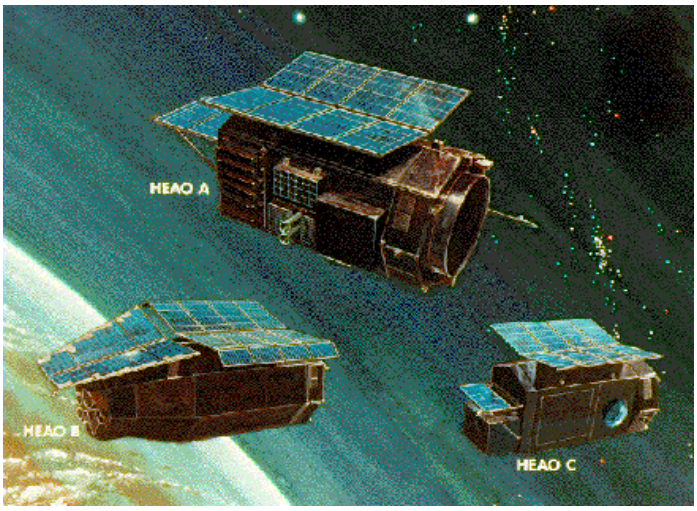
Two other images are described in their captions:



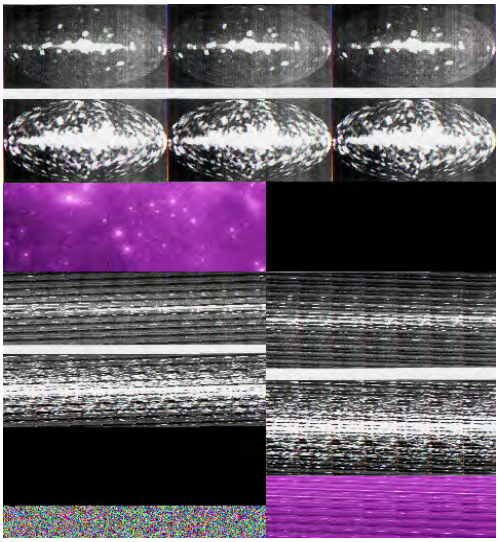


Another achievement has been to gain a better insight into the nature and origin of cosmic rays. Fermi has determined that a principal component of this radiation is energetic, high speed protons that are accelerated by magnetic fields associated with supernova bursts.

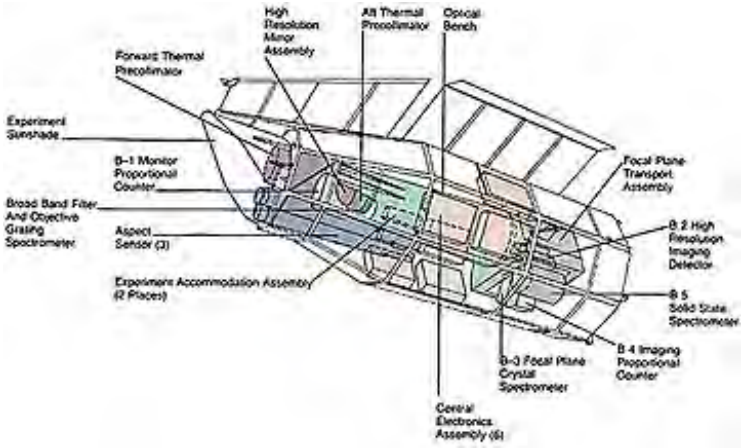
Now, let us look at high energy radiations in the X-ray region. The first focused telescopes that operated in this region were the three HEAO (High Energy Astronomical Observatory) satellites operated by Goddard Space Flight Center and flown in the 1970-1980s:



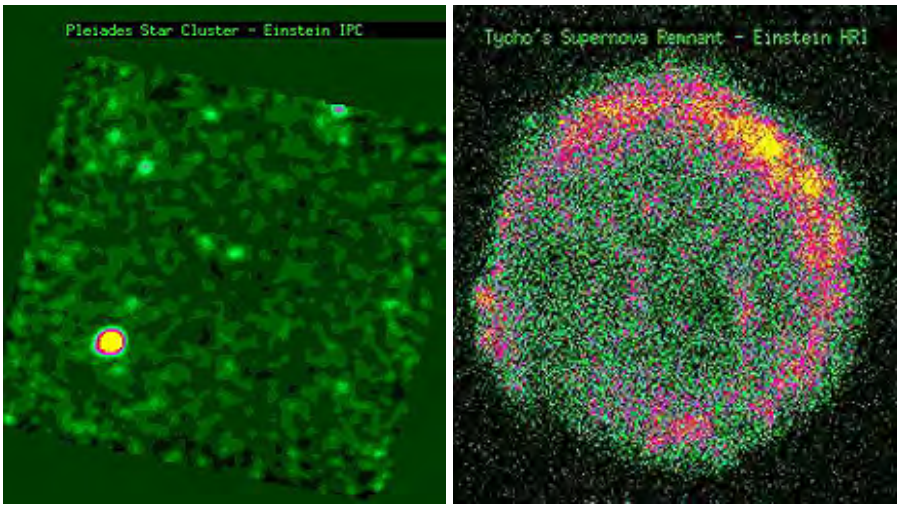
HEAO-1 found many X-ray sources within the Milky Way and also outside our galaxy (some sources have yet to be identified with specific visible sources:



HEAO-2, renamed the Einstein X-ray Observatory, gathered data between October, 1978 and April, 1981. The spacecraft had these components:



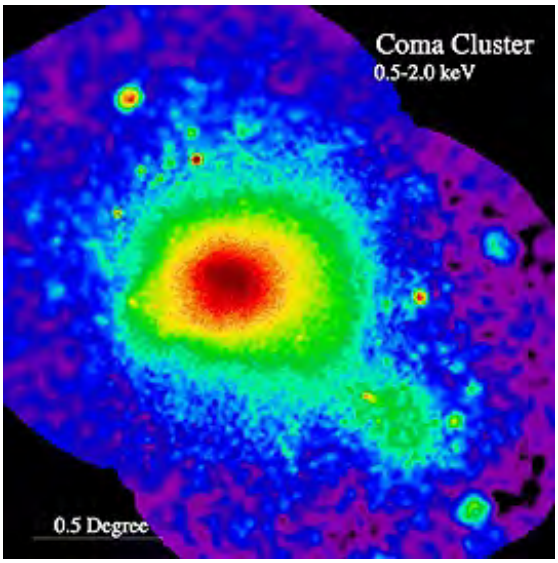
Here are two parts of the sky imaged by the Einstein X-ray Observatory:



The image below shows the spiral galaxy M83 as it appears (in a colorized rendition) optically. Against this are contours of variations in X-ray intensities (given in units of keV [kilo-electron volts]) as measured by the joint U.S.- German *Rosat* (Roentgen Satellite) launched on June 1, 1990 to monitor the entire sky. Note the close-spaced contours around the galaxy center, but several other X-ray "hot spots" are also evident.

Several mechanisms account for this X-ray generation. Most prevalent is excitation into ionized states of intragalactic gases between stars or gases between galaxies that, in the tenuous void separating the stellar bodies, are traveling at such high velocities that they represent temperatures in excess of 1,000,000 °K capable of producing strong X-ray responses.

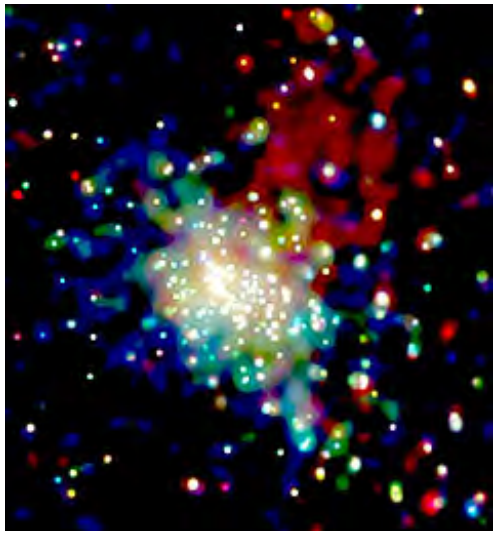
The next Rosat image portrays X-ray variations spread over the entire Coma supercluster, comprised of well over 1000 bright galaxies, located some 300,000,000 light years away. X-ray intensities vary from strong in reds to decreasingly weaker in greens to blues and purples. The interstellar gases emitting this radiation make up about 10% of the total mass of the supercluster, along with 2% more in the stars found in the individual galaxies as determined from optical measurements. The remainder of the mass is presently unaccounted for after inventories across the spectrum are related to their sources, so that the bulk of the mass is presumed associated with dark matter (see page 20-9). Thus, examining both galaxies and intergalactic regions using radiation at wavelengths both shorter and longer than the visible helps to quantify the distribution of the entire mass of the Universe.



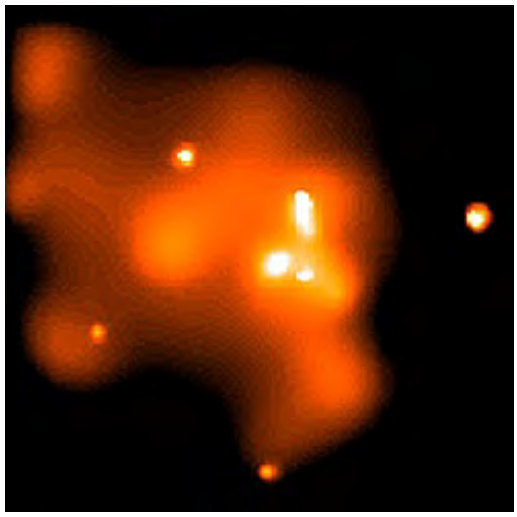
In September of 1999, NASA, guided by scientists from several nations, launched the Chandra X-ray (Telescope) (CXO). Named after the late S. Chandrasekhar, a reknown astronomer from India, Chandra is managed by the Marshall Space Center. Its length, when fully deployed, is 13.6 m (45 ft). It carries 4 sensors: a charge-coupled imaging spectrometer, a High Resolution camera, and High and Low Energy gratings.Its spatial resolution is 8 times greater than the best previous X-ray observatory and can pick out objects 20 times fainter as sources of x-radiation. Here is Chandra in space, as photographed from the Space Shuttle from which it was launched:

Its astronomical targets include quasars, supernova and other high energy-emitting objects. Here is an example of an image of a ring of X-radiation associated with the remnants of a supernova in the Constellation Tucane:

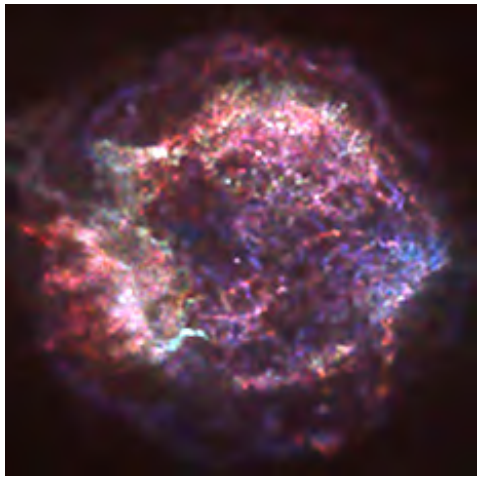
Chandra has made images of regions of more recent star formations (sometimes as bursts) in the Milky Way. This one is striking indeed.



The Milky Way galaxy has a powerful X-ray source at its center probably associated with material infall into a Black Hole, as imaged thusly:

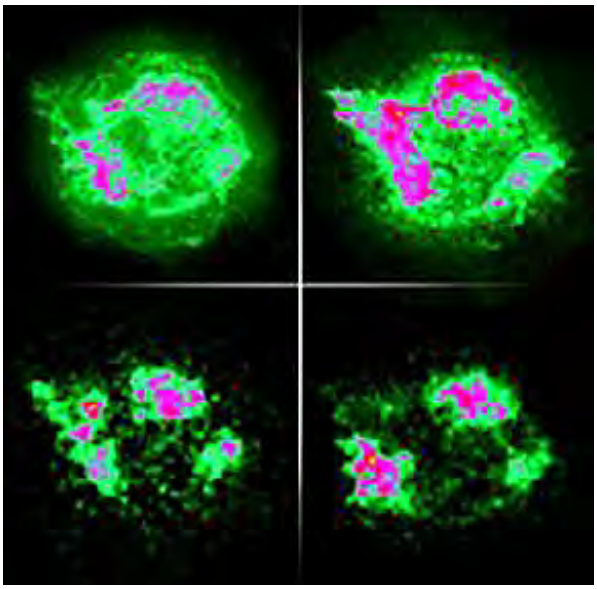


Because Chandra measures X-radiation from its targets over a range of wavelengths, individual elements which give off X-ray spectra at specific wavelengths can be detected and mapped. This has been done for the supernova Cassiopeia A (see page 20-6 for more information about supernovae). An HST optical image of this exploding star looks like this:

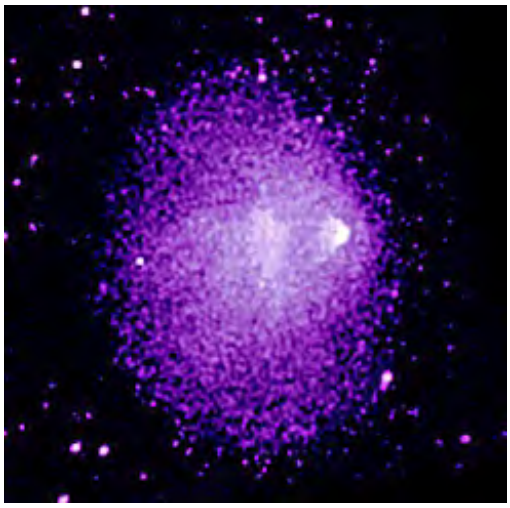


Here is a four panel set of Chandra images of Cassiopeia A. The upper left is color density map of the broad band radiation from this supernova. The upper right focuses on Silicon emission

lines; the lower left on Calcium; and the lower right on Iron. Thus Chandra is an adept tool for determining the distribution in the expelled material of various elements that were produced by nuclear burning in the star.

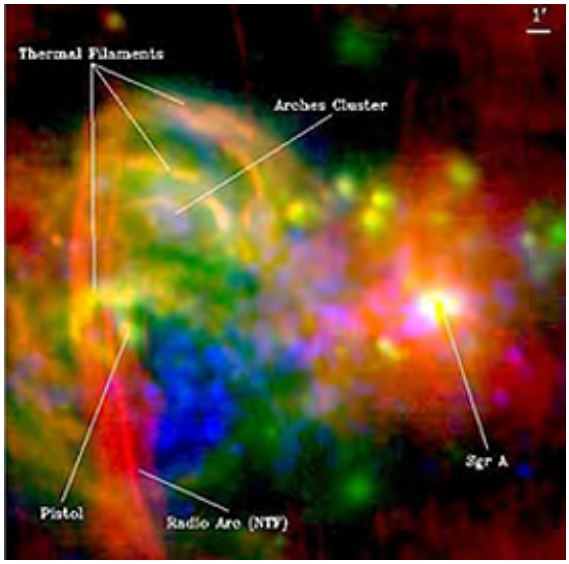


Chandra has imaged a striking example of a bow shock wave developed in a cluster of stars in a small galaxy known as IE0657-6. Gases within the cluster are extremely hot, around 100 million degrees Centigrade. The bow shock (right side in the next image) was produced from a collision with a subcluster of stars whose gases are around 70 x $10^{6\circ}$ C. In time much of the gases will be blown out from the present configuration.



Chandra has explored our Milky Way galaxy as well. This next image shows part of the central core region of the galaxy (about 400 light years wide) in which a number of very bright objects, seen in X-radiation, correspond to high energy emissions where interstellar gases are drawn into white Dwarfs, Neutron stars, and possible Black Holes, becoming continuously "ignited".

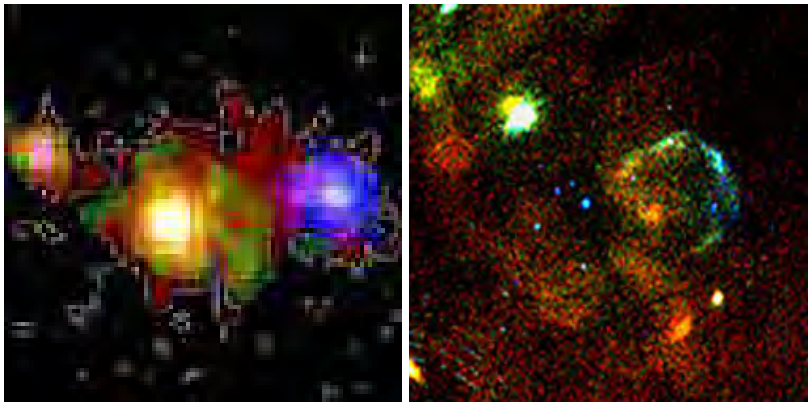
A spectacular image of part of this central region was made by the Advanced CCD Spectrometer on Chandra:



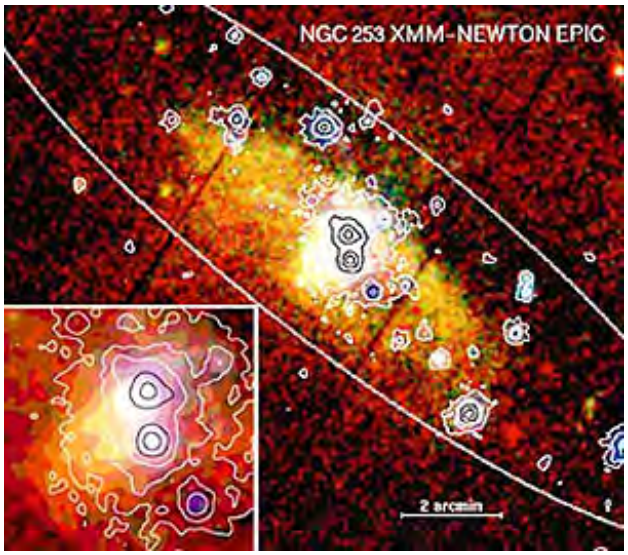
Chandra scored a first in late summer of 2003. As it monitored a galaxy in Perseus (as imaged on the left), it also detected a signal that is best interpreted as evidence of sound waves passing through the galaxy. Calculations show the sound to be in the "musical note" of B flat, but 57 octaves below the lowest octave on a standard piano. The sound waves were rendered into an image (on the right):

In December of 1999, the European Space Agency launched an even more powerful X-ray

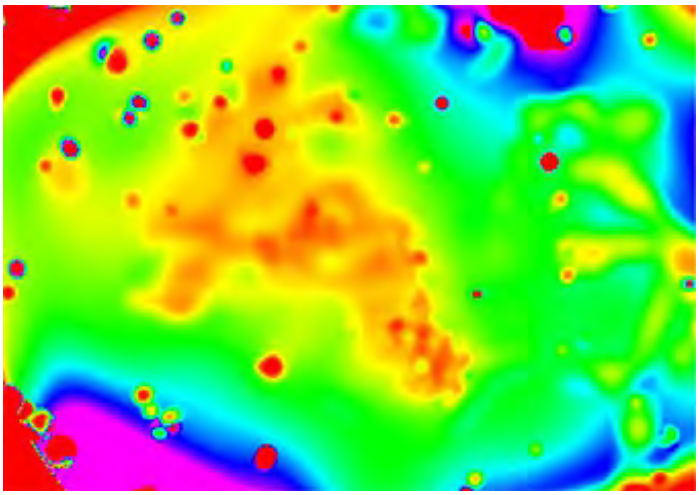
telescope known as XMM-Newton (XMM stands for X-ray Multi-Mirror). Here are two colorized images, the first showing the variations in X-ray intensities in several of the Hickson group of stars and the second showing details of a supernova explosion in the nearby Large Magellanic Cloud:



XMM-Newton has demonstrated that large X-ray energy bursts also associate with the starbursts that mark development of young stars. Here is an image of NGC253, some 8 million light years from Earth; the inset on the left is a closer look at its center.



A notable discovery made by XMM-Newton is a huge (10^9 solar masses) concentration of fast-moving (1000 km/sec) gas, shown in this image:



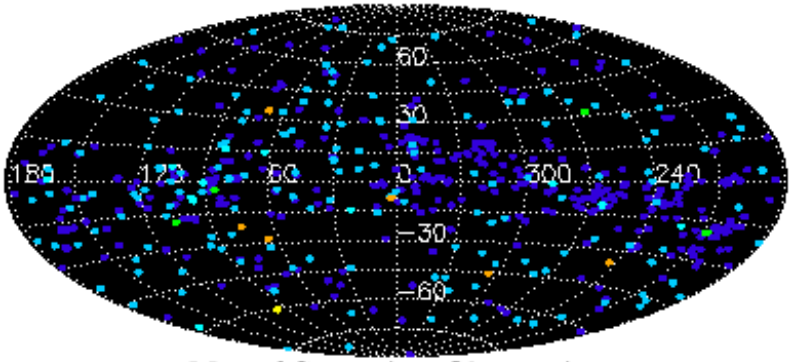
This mass of gas has been likened to a "mega-comet" (but it is not a true comet in the Solar System sense) some 5 million light years in length. The gas appears red to yellow, representing the mapping of data into entropy values. The feature is within the galaxy cluster Abell 3266. Its

discoverers believe it to be held together by Dark Matter.

XMM-Newton has taken a series of images over a period of days that can be sequenced to give a movie-like effect of the expansion and dissipation of materials during a burst. We will cite the Internet connection on which this dynamic rendition can be accessed - with the proviso that it may no longer be active. So, click on starburst to see if the "show goes on".

Still on the drawing boards (not fully funded) is the Constellation-X Observatory that will be the most sophisticated X-ray satellite yet; launch will be in 2017 or later.

Satellites began to examine the UV region of the sky with the OAO series (OAO-3 was named Copernicus) in the late 1960's. Copernicus produced maps of bright UV stars such as this:

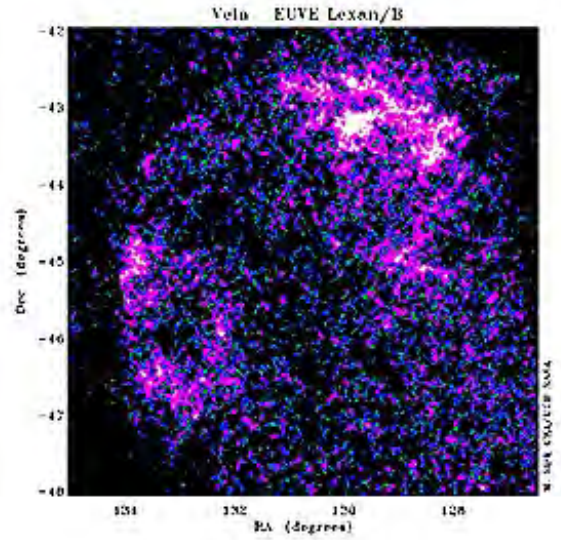


Map of Copernicus Observations

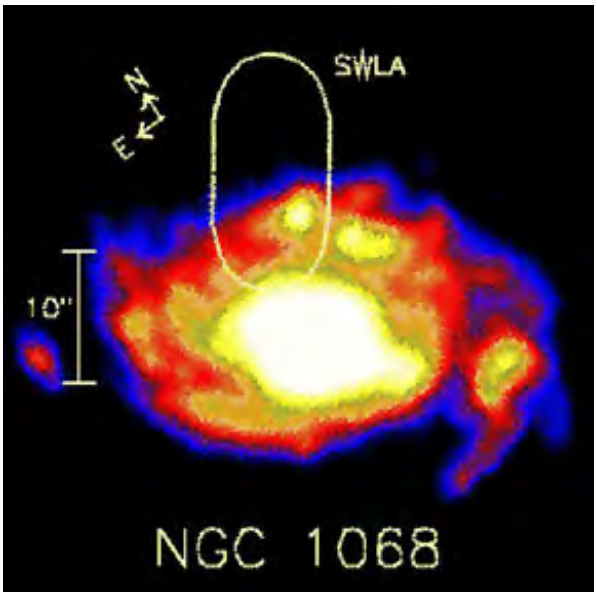
The ultraviolet (UV) region of the spectrum, from 70-2000 (0.007 - 0.2 µm) (Far) to 2000-4000 Angstroms (0.2 - 0.4 µm) (Near), has provided interesting images of stellar bodies, including the Sun. It also contains many diagnostic spectral lines helpful in determining elemental composition. In the 1970s, the Soviet space program installed UV Observatories, named Orion 1 and Orion 2, on a Salyut and Soyuz spacecraft respectively.

This next image shows the Earth as imaged by EUVE (Extreme UltraViolet Explorer, a free flying observatory launched in 1992 and operating until February, 2001; imaging from 70 to 760 Angstroms). It shows excited Helium (yellow) and Hydrogen (orange-red) in an auroral field extending well beyond the solid Earth.

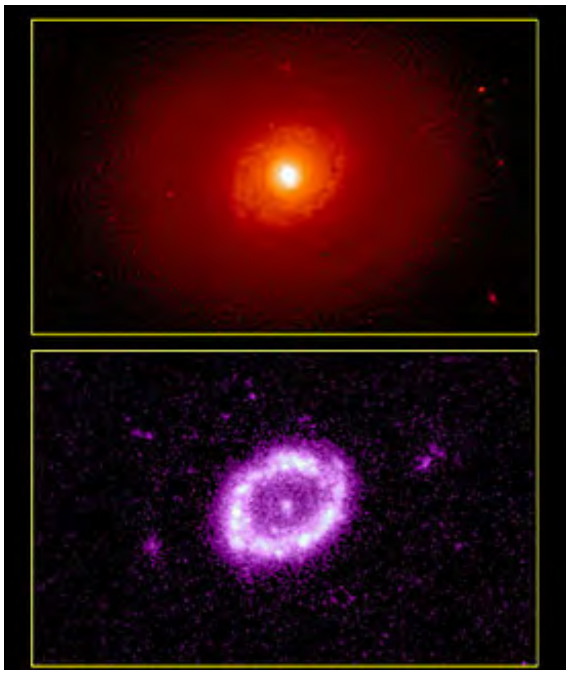
Looking outward into space, the EUVE provided this image of the Vela Supernova:



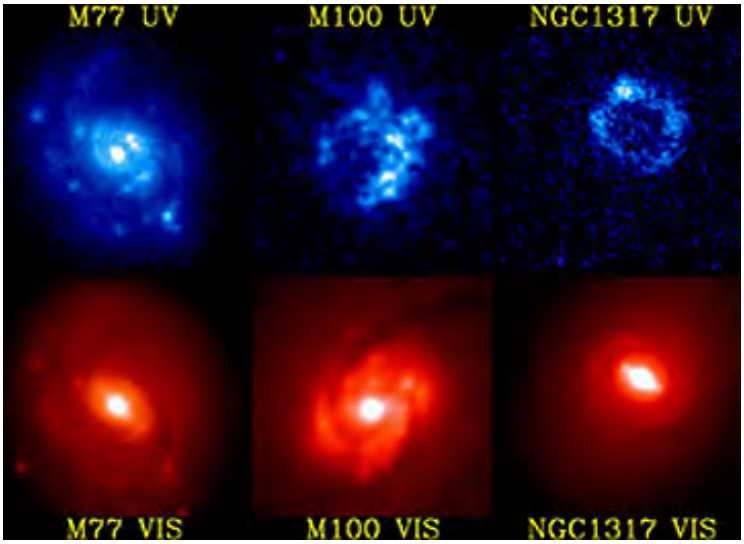
One of the first UV telescopes is the IUE (International Ultraviolet Explorer) launched jointly by ESA and NASA in 1978; it operated into 1996. This is a UV image of the galactic source NGC1680:



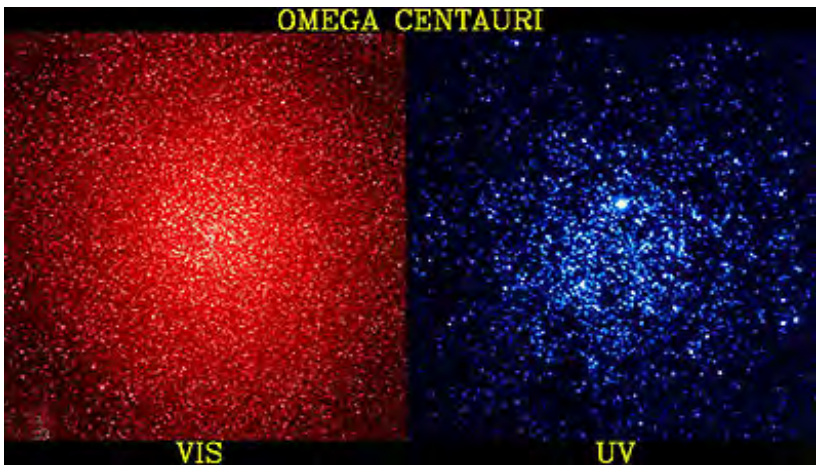
The Ultraviolet Imaging Telescope (UIT) was flown as part of Astro-1 and Astro-2 lab packages on Shuttle STS-35 and STS-67 in the mid-1990s. The telescope covers the UV range between 1200 and 3200 Angstroms. It is particularly adept at recognizing hot, young stars which give out strong UV radiation. The difference in appearance between visible and ultraviolet images is pronounced in this UIT view of the galaxy M94:



This next image shows three galaxies in UV (top) and Visible (bottom); note the structure of the spiral arms as brought out by molecular Hydrogen excitation.



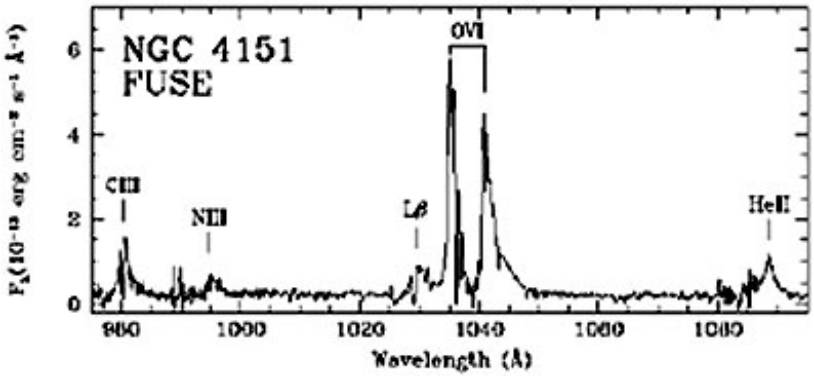
In the UIT image below, the globular cluster Omega Centauri in visible light appears to consist of mainly red to orange stars, typical of older stellar bodies. But, the UV on the right shows that there are also many younger, hotter stars.



Launched on June 24, 1999, FUSE (Far Ultraviolet Spectroscopic Explorer) gathers spectra in the interval 910 - 1180 Angstroms. The program is run out of Johns Hopkins University, with NASA,

French, and Canadian partners.

Excitation of molecular and elemental species in a star's atmosphere or a galaxy en masse in this interval provides valuable information about stellar processes. Here is a typical spectral plot obtained by FUSE from observing a galaxy.

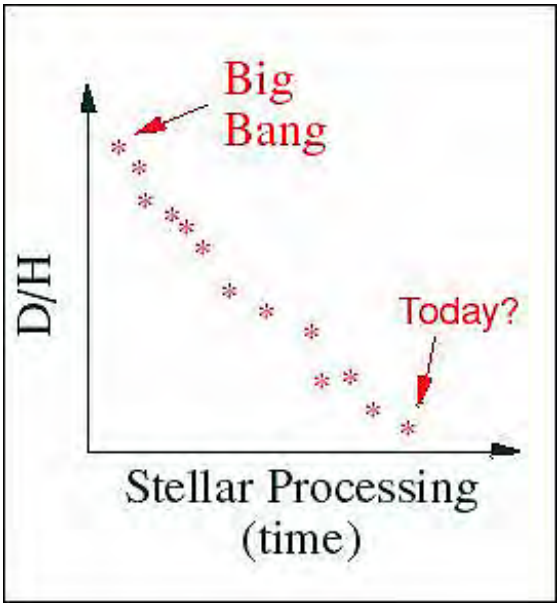


Observations through the FUSE telescope can be converted to images, such as this:

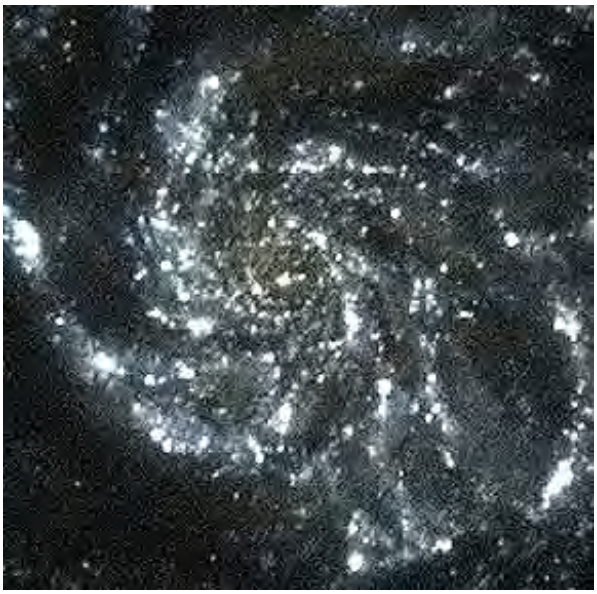
FUSE's primary goal has been to trace the history of the early Universe by monitoring the distribution of Hydrogen (H), Deuterium (D), and Helium (He) in the intergalactic medium. Deuterium was detected in this FUSE UV image of star AE Aurigae, which in visible light is shrouded by dust:

Preliminary results from FUSE indicate that Helium, formed in the first minute of the Big Bang, and then dispersed during the expansion, will prove a sensitive indicator of the inhomogenieties in the expanding Universe following the initial explosion. This is a generalized diagram of the

ratio of D to H since the Big Bang, as Deuterium is converted to He through H fusion (thus, decreasing the ratio).



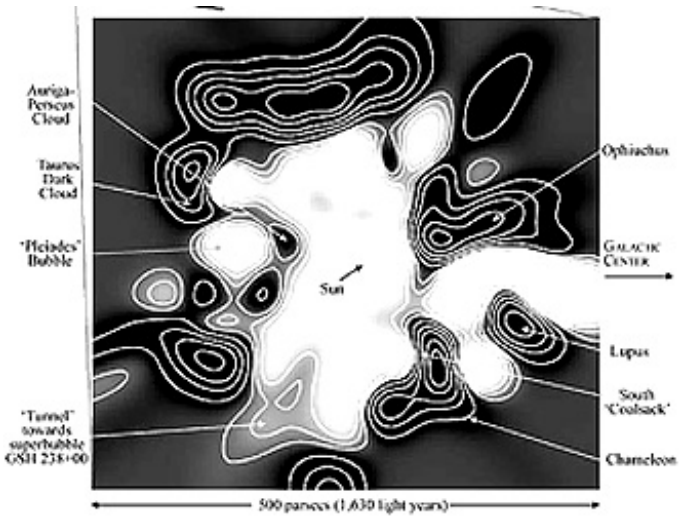
JPL has developed GALEX (Galaxy Evolution Explorer), which was launched on April 28, 2003. Designed to gather imagery in the far and near ultraviolet (FUV and NUV), it will concentrate on monitoring distant galaxies and stars (out to at least 10 billion l.y.) to determine the conditions under which they had formed in the early years of the Universe. This image of Galaxy M101 was made by combining the FUV (blue) and NUV (green) images:



Here are a trio of images of Galaxy M51, one in the UV (GALEX), the middle from an optical telescope, and the one on the right in the Near-IR (2Mass project)

A final look at a galaxy imaged by GALEX entirely in the UV: NGC 1232 is shown here as a partial false color composite made from two UV bands:

Thus, the UV is proving to be an optimum segment of the EM spectrum to study conditions in the so-called empty space which actually contains hot interstellar gas. CHIPS (Cosmic Hot Interstellar Plasma Spectrometer) is an astronomy satellite (called CHIPSat) launched on January 12, 2003. It conducts an all-sky spectroscopic survey of the diffuse background at wavelengths from 90 to 260 Å with a peak resolution of about 0.5 eV. CHIPS data is helping scientists determine the electron temperature, ionization conditions, and cooling mechanisms of the million-degree plasma believed to fill the local interstellar bubble. The majority of the luminosity from diffuse million-degree plasma is expected to emerge in the poorly-explored CHIPS band, making CHIPS data of relevance in a wide variety of Galactic and extragalactic astrophysical environments. Thus, it has measured the diffuse extreme ultraviolet glow that will better define the properties and physical processes associated with the interstellar medium. A review by Univ. of Calif-Berkeley scientists of the mission's purpose and some results is available at this UC-Berkeley site. Using CHIPS data, in part, this map of the local bubble around the Sun extending to nearby stars has been published:

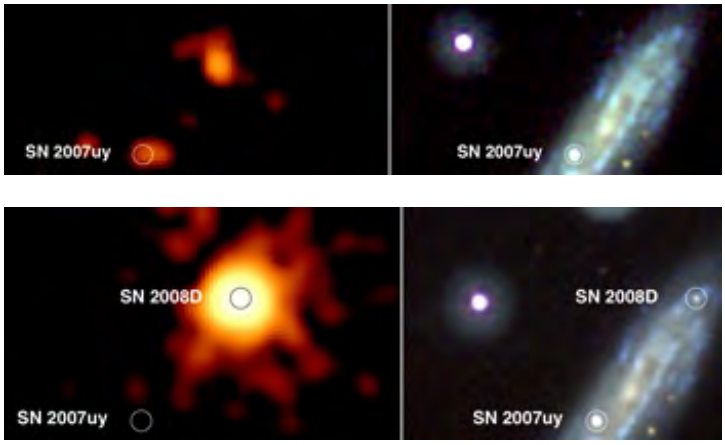


In 2004, NASA, along with several cooperating Universities and organizations, launched SWIFT, a

telescope observatory satellite whose prime mission is to search for Gamma Ray Bursts and then examine their sites in UV and Visible light. SWIFT carries a gamma-ray detector, an X-ray detector, and a detector whose operational range includes parts of the UV and Visible. Below is the first image, of the Pinwheel Galaxy, made by this third instrument:

SWIFT made this time lapse set of images of a short-lived burst from a neutron star, SGB J1550-5418, in which X-rays have excited dust surrounding it, so that a dispersed cloud of material is evident as an expanding ring:

This next pair of dual images shows one of the typical gamma-ray burst sequences sought by SWIFT. In the top image, the left view shows SN 2007uy (labelled) and a star above it that will become SN 2008D, seen on the left view of the lower image. This burst is evident in the visible image (right; lower) near the "top" of galaxy 2770.



The UV is particularly suited to spotting active star formation in galaxies. Andromeda has

numerous super-bright regions of young stars, as seen by Swift.

The UV carries into the Visible spectral range. Just beyond the Visible is the Infrared, extending from about 1 to 1000 µm. Much of the interval coincides with the thermal IR which you studied in Section 9. Hot stars are strong emitters in the IR and can be studied both as images and from their spectra. Other astronomical features amenable to IR observations include properties of accretionary disks and interstellar clouds, the structure of the H II type stars (those in an early stage of development that contain significant ionized Hydrogen in the inner part of the Hydrogen gas cloud that is the source of their nuclear fuel), and the dynamics of the Milky Way.

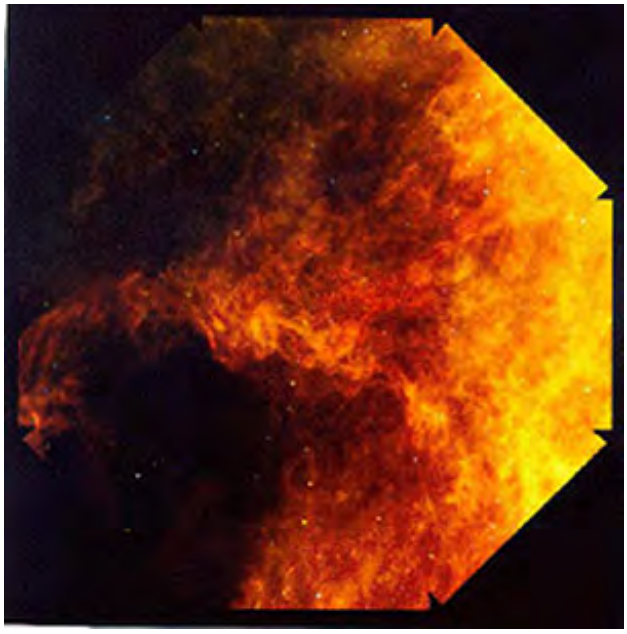
Viewing galaxies and regions of heavy dust densities in the Infrared has a distinct benefit compared with seeing the same features in the Visible. This image pair vets this statement (read its caption):

Small dark interstellar dust that obscures stars in the Visible are called Bok Globules (discovered by a Dutch astronomer of that name). They represent nebular gas and dust nearing the protostar phase (see page 20-2); such molecular Hydrogen clouds are very cold (-263°C) and generally because of their small size (about a parsec) produce only one to several stars. These globules (some of which can be nearly spherical) stand out best in images that extend into the Near IR. These two photos (acquired by ESA's New Technology Telescope) show details of a Bok Globule in Barnard 68. The left image is made from three bands in the visible; the right image consists of bands at 1.25 µm = Blue; 1.65 µm = green; and 2.16 µm = red, which renders the cloud now partially transparent so that stars behind it become visible.

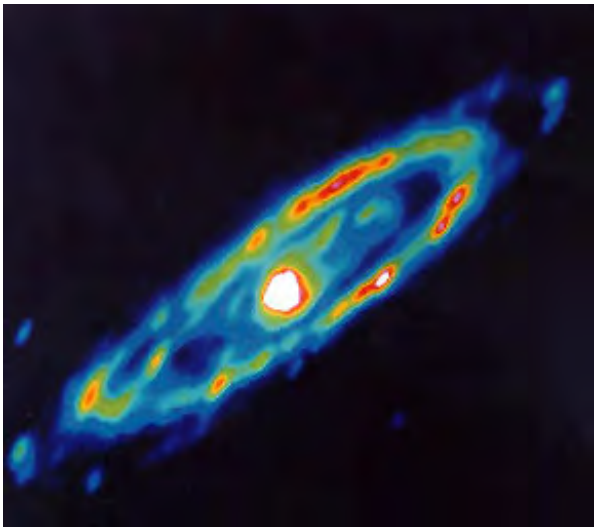
An entire galaxy (NGC2024) that is still largely shrouded by dust looks much like a visible image in this version made by the NICMOS camera on HST. The color composite consists of Blue = J band (1.6 µm); Red = K band (2.2 µm); and Green = J and K combined.

The Infrared region of the EM Spectrum has a treasuretrove of information. Hot stars are often shrouded in dust but some IR bands are "transparent" relative to that blocking material, allowing the stellar thermal source to shine through. One of the first infrared-dedicated satellites was IRAS (Infrared Astronomical Satellite) launched in January of 1983. Its sensors were tuned to the 12, 25, 60, and 100 µm IR wavelengths. Here it is at the dirt-free facility at Goddard Space Flight Center - at the time just a hundred meters from my office:

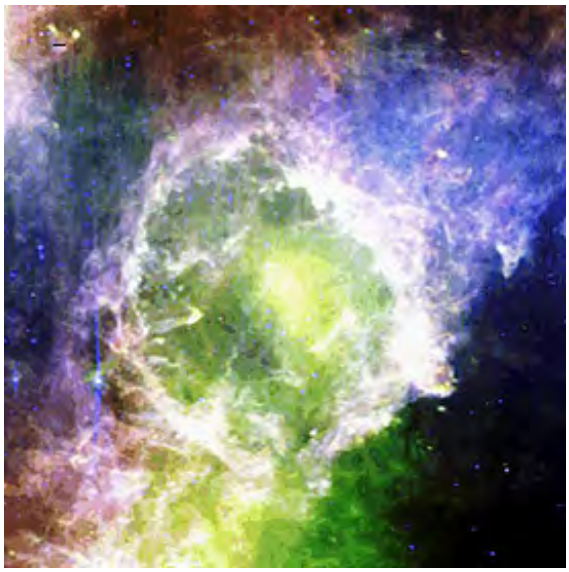
During its lifetime, IRAS discovered more than 350,000 previously undetected IR objects in the sky. This color composite of the interstellar "cirrus" clouds made up of gas and dust grains in the Milky Way that occupy a wide field centered on the North Celestial Pole is constructed from Blue = 12; Green = 60; Red = 100 µm.



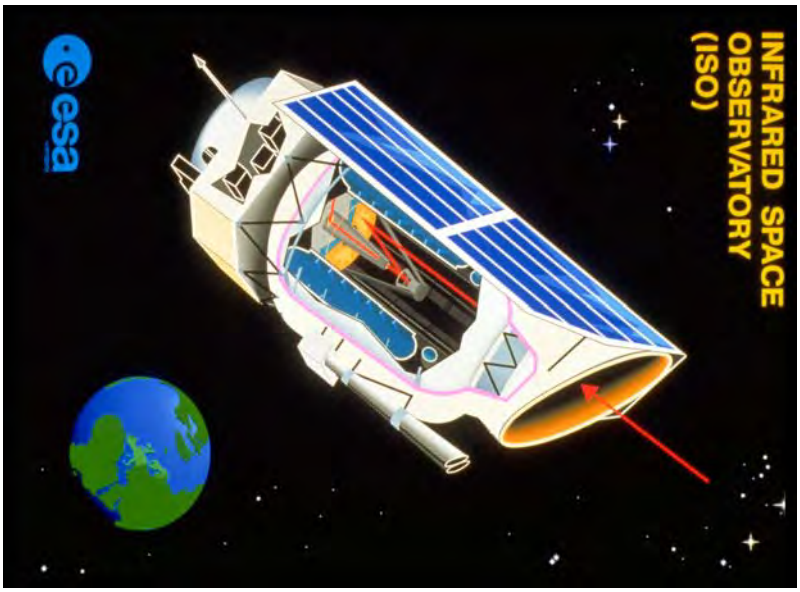
On a grander scale, look at this IRAS image of the now familiar neighbor, the Andromeda Galaxy, with color-codes indicating variations in thermal emission at 12 µm.



One of the finest IRAS images is this 25, 60, 100 µm color composite showing the nebular material around Lambda Ori:



Other IR observatories have since been placed in space. ISO, the Infrared Space Observatory, was operated by ESA from November '95 until May '98. The instruments include an IR camera, a spectrometer, and a polarimeter. The spectral range was 2.5 to 240 µm. Here is a painting of the ISO:

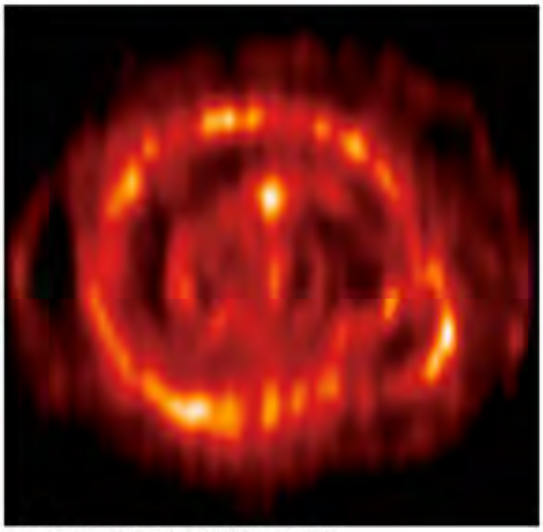


These ISO images of the familiar Andromeda galaxy are typical of this observatory's products.

Andromeda's invisible rings

M31 in the visible.

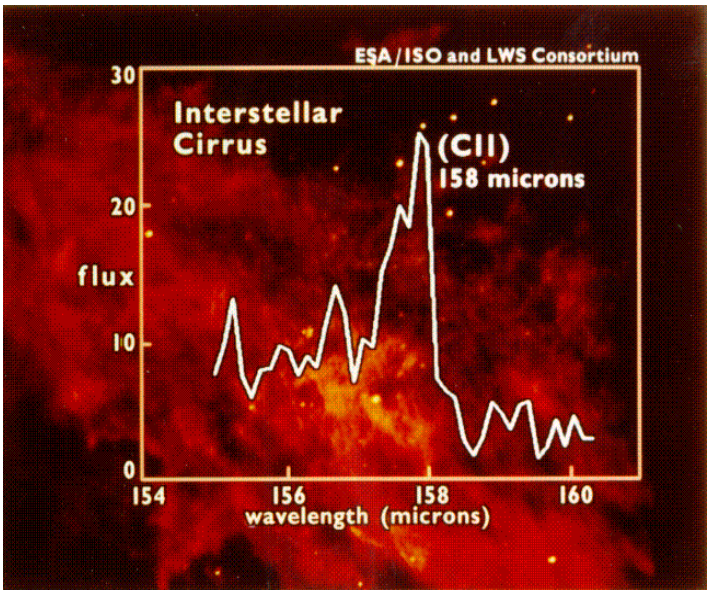
ISO map of M31.
(wavelength of 175 microns)
North is up, east is left.



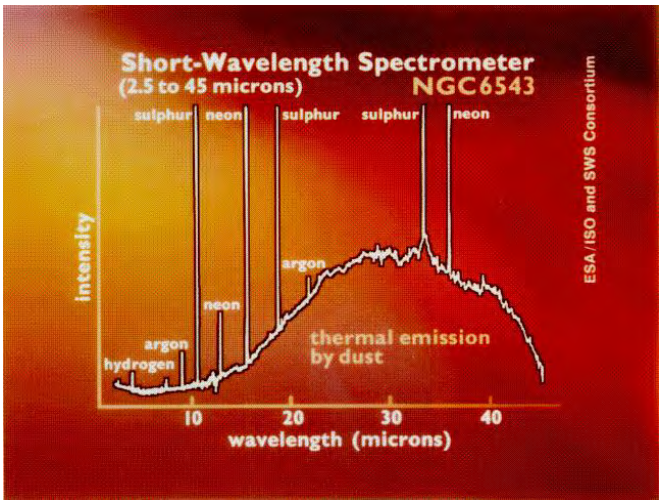
Face-on projection of the ISO map of M31.
(wavelength of 175 microns)
North-east is on the left.

Credit:
For the infrared: ESA/ISO/ISOPHOT and M. Haas et al.

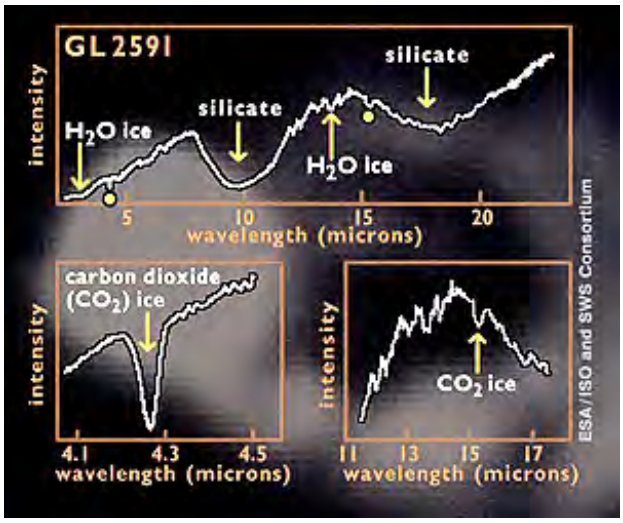
ISO can monitor the spectra of various celestial objects. This example shows the results for an interstellar cirrus-like nebula; CarbonII is a major constituent:



A IR wavelength plot of radiation received from NGC6543 shows peaks correlated with Argon, Neon, Hydrogen, and Sulphur which occurs in the dust and gas nebula associated with this, the Antennae galaxy.

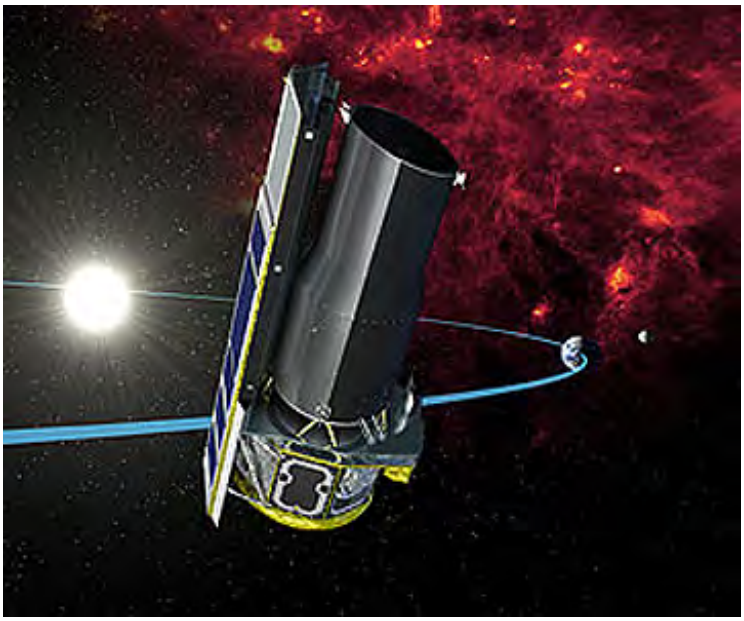


The star GL2591 is surrounded by a dense cloud. Spectra in the Short Wave IR interval sampled by IS0's spectrometer disclose water ice, carbon dioxide ice and silicate particles in the dust grains within the enclosing material.



On August 24, 2003 NASA launched a major new telescope - one of the Big Four of the Great Observatory series in its current astronomy programs - SIRTF (Space Infrared Telescope Facility). Following a contest to rename this powerful new sky-searcher, it is now called the Spitzer Space Telescope (SST), named after the pioneer astronomer Lyman Spitzer. SST is comparable in its capabilities to the HST and Chandra. The orbit is heliocentric and Earth-trailing. Its instruments operate in the infrared between 3 and 180 µm, which includes much of the thermal infrared spectral region. Its primary mission will be to peer through cosmic clouds and dust (usually, transparent in the infrared) to look back in time to see galaxies and stars in their earlier stages of development. A preview of the SST is given at a JPL lecture. Access this at von Karman lectures, entering Webcast into the Format box, and choosing the topic "The Space Infrared Telescope Facility", June 12, 2003, clicking on RealPlayer.

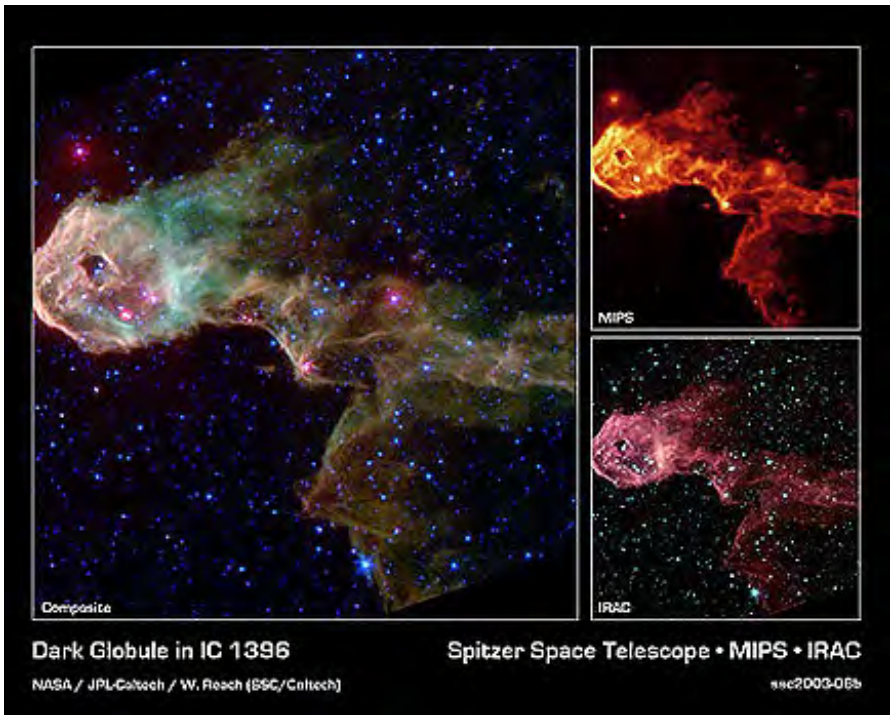
Here is a sketch of the SST in orbit:



First data were released on December 18 of 2003. This panel of four images is typical of first results:

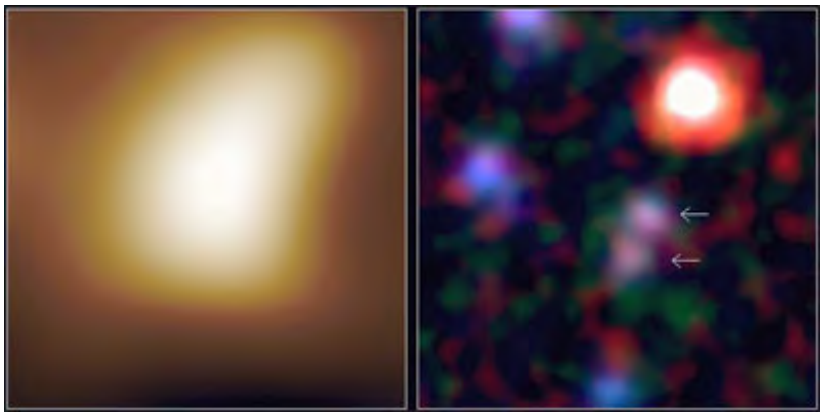
The upper left panel shows the spiral galaxy M81 in a false color thermal infrared composite. Upper right is a Haro-Herbig star seen as a thermal object (in visible light it is masked by clouds). The lower left panel is a view of Comet Schwassmann-Wachman; the remaining panel shows the Dark Globule IC1396 (ordinarily indistinct in visible light).

This illustration enlarges the Dark Globule IC1396 and shows on the right this same image as sensed by two of the instruments on SST.

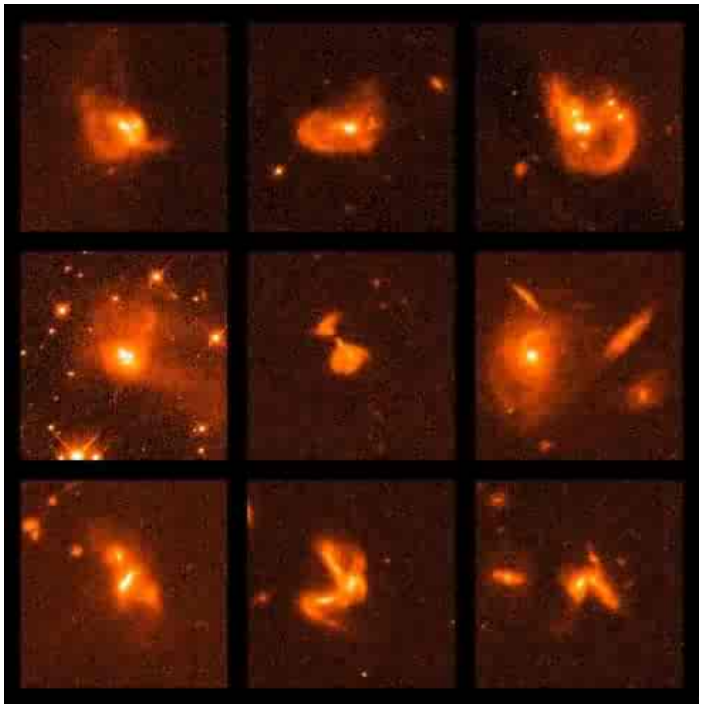


As stated above, one powerful attribute of the SST is its ability to "see" through thick clouds of dust. A region within the Milky Way about 10000 l.y. away contains a great clot of dark dust in which almost no stars are visible. When viewed in the IR by SST, stars and glowing gas were revealed. Among these were some very large stars, demonstrating that stars up to 100 times more massive than the Sun are still forming in our galaxy and, by inference, probably throughout the Universe (thus big stars can be recent in time of formation, even though not long-lived).

This ability to penetrate unresolved dust or close-spaced, very distant galaxies is making Spitzer a powerful tool for studying galaxies formed early in Universe history. The left image below shows what appears to be a uniform nebula imaged in the infrared from the ground by the United Kingdom's SCUBA instrument; nothing beyond is visible. When the SST looked at the same area, at different IR wavelengths, faint galaxies (arrows), very far away and hence seen now in their earlier stages, are now resolved (the bright orange feature is a nearby star in the line of sight.

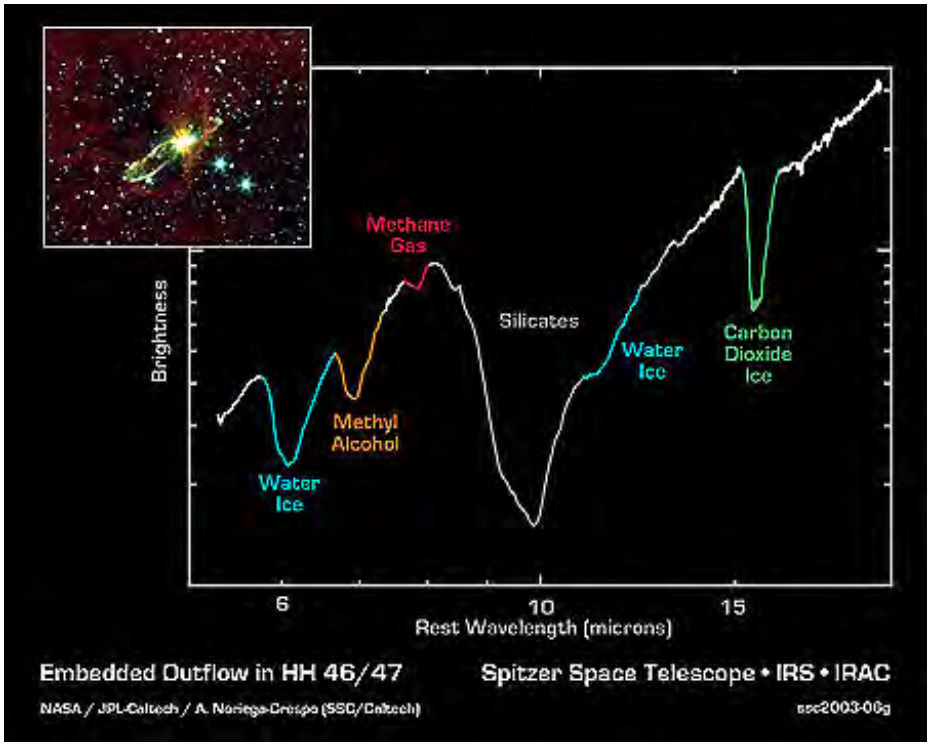


One class of galaxies has proved especially amenable to study using infrared wavelengths. This is the so-called ULIRGs, for Ultraluminous Infrared Galaxies. These galaxies are marked by high production rates of stars. The radiation from the stars heats up the dense clouds of dust and gas, causing them to glow brightly in the IR. Here are images of ULIRGs made by the SST:

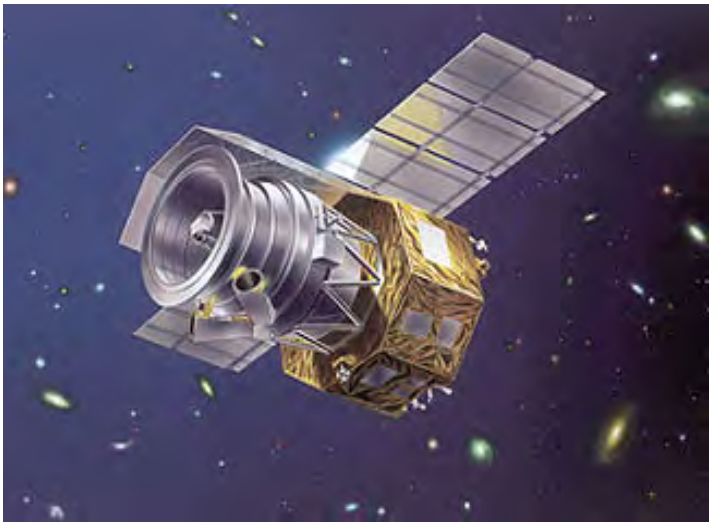


Multispectral images made by different space telescopes covering separated regions of the EM spectrum can lead to some striking images, as seen in this view of Galaxy M82. The Chandra image is rendered blue; the Hubble image of excited Hydrogen is shown in green; and the Spitzer ST image is assigned red:

The SST has a spectrometer that can recognize chemical components of a star and its surrounding cloud. For HH 46/47, shown again in the inset, the spectral curve contains strong absorption bands that indicate the presence of water ice, methyl alcohol, silicate particles, and carbon dioxide.

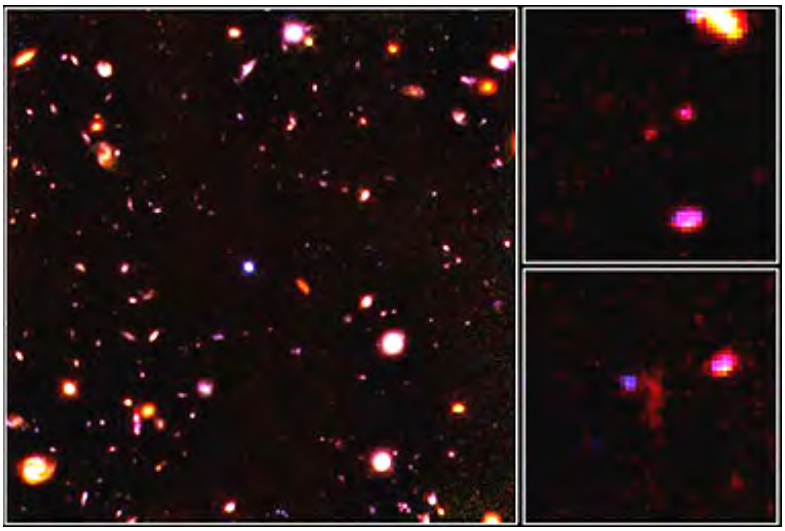


On February 22, 2006 NASDA, the Japanese Space Program, launched ASTRO-F, an Infrared Observatory. Nicknamed "AKARI" ("Night"), the spacecraft looked like this:



Its main sensor has 8 infrared channels. Each produces its characteristic expression of thermal energy distribution, as exemplified:

Infrared imagery has proved especially sensitive to detecting Deep Field galaxies. These have been redshifted by the Universe's expansion so as to radiate in the infrared. This image shows some of the stars at great distances from Earth that glow brightly in the IR:



The Herschel Observatory, shown below, was launched on May 14, 2009. It operates mainly in the far infrared. Herschel has been placed at the Lagrangian point L2 (where the gravity of the Sun just balances that of Earth) about 150,000 km from Earth. Read about it on its ESA website.

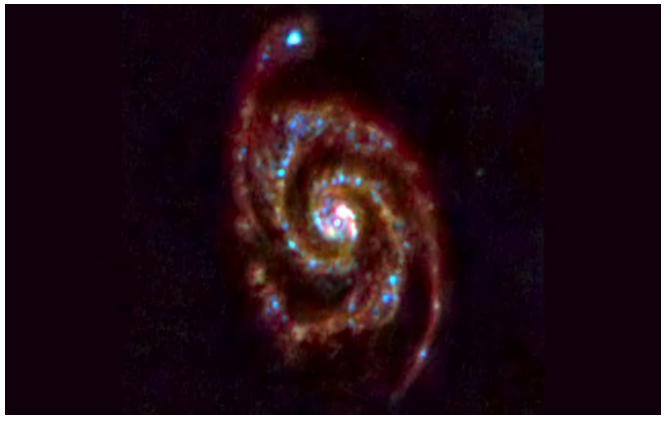
Herschel's science payload consists of three instruments:

Photodetector Array Camera and Spectrometer (PACS), a camera and a low- to medium-resolution spectrometer for wavelengths up to about 205 micrometres. It uses two bolometer detector arrays in the camera and two photo-conductor detector arrays in the spectrometer.

Spectral and Photometric Imaging Receiver (SPIRE), a camera and a low- to medium-resolution spectrometer for wavelengths longer than 200 micrometres. It uses five detector arrays: three to take images of infrared sources in three different infrared 'colours' and two to fully analyse the longer infrared light being released from the source.

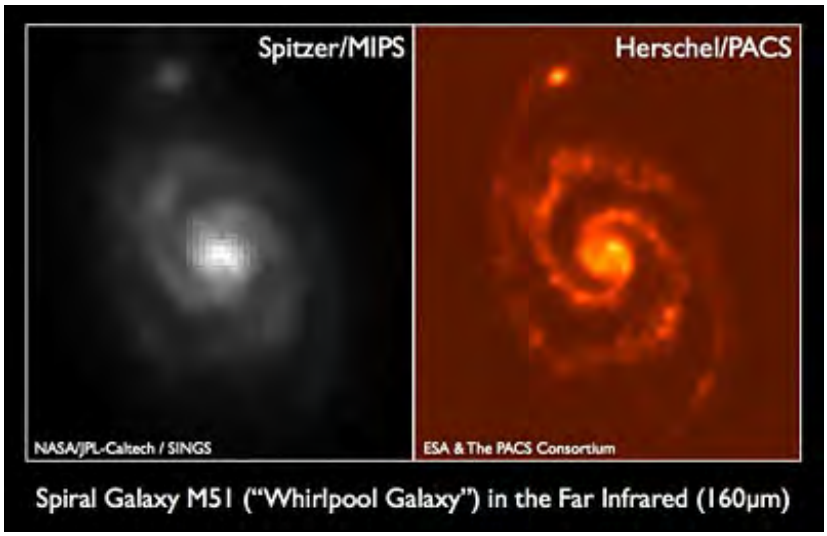
Heterodyne Instrument for the Far Infrared (HIFI), a highly accurate spectrometer that can be used to obtain information about the chemical composition, kinematics, and physical environment of infrared sources.

The first image from Herschel released to the public is shown below. It is the Whirlpool galaxy, near the Milky Way. Contrary to the usual convention, red denotes relatively cooler temperatures while blue is warmer.

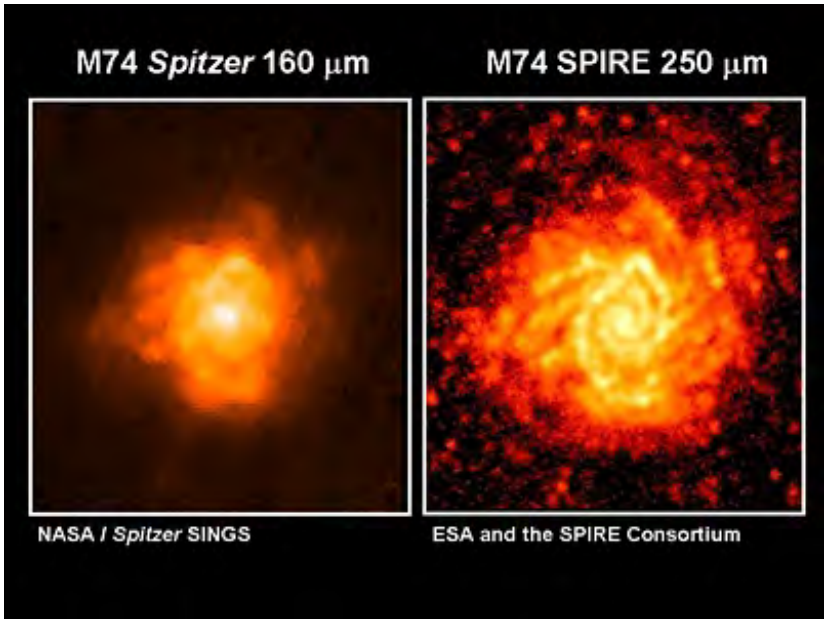


An indication of the improvement in resolution afforded by Herschel is gained from this

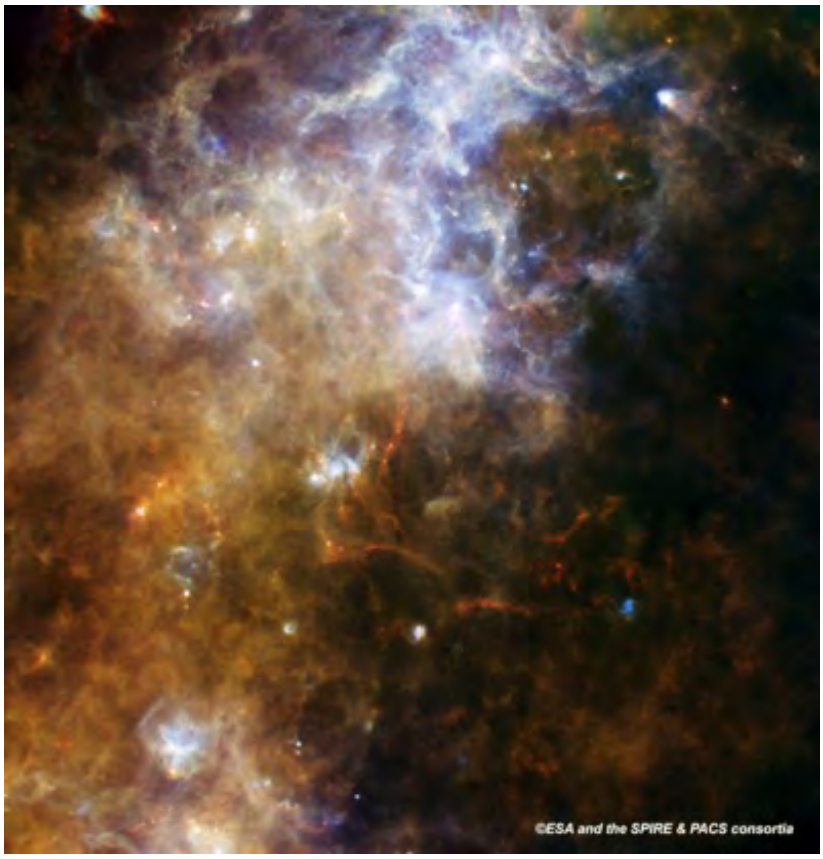
comparison of galaxy M51 as seen by Spitzer (left) and by Herschel's PACS (right):



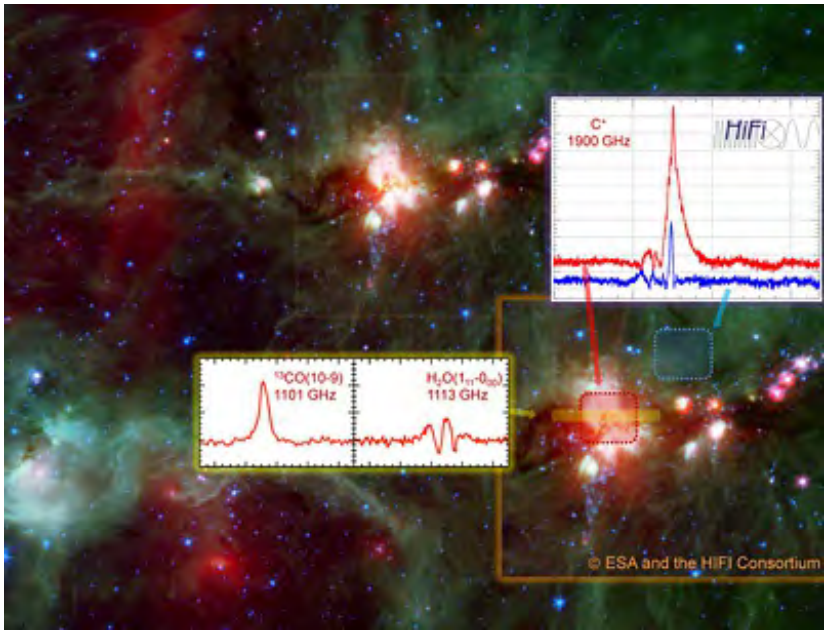
Similar improvements are evident when Herschel's SPIRE instrument is compared with its Spitzer equivalent, using galaxy M74 as the target.



Herschel is spending much of its observing time looking at the Milky Way as it appears in the Infrared. The gases between stars are generally quite cold but the sensitive PACS and SPIRE instruments can detect low temperature outputs, giving a better understanding of how the gas is distributed. Here is a small area within the M.W. as seen individually by these instruments, and then as an image made by combining them:

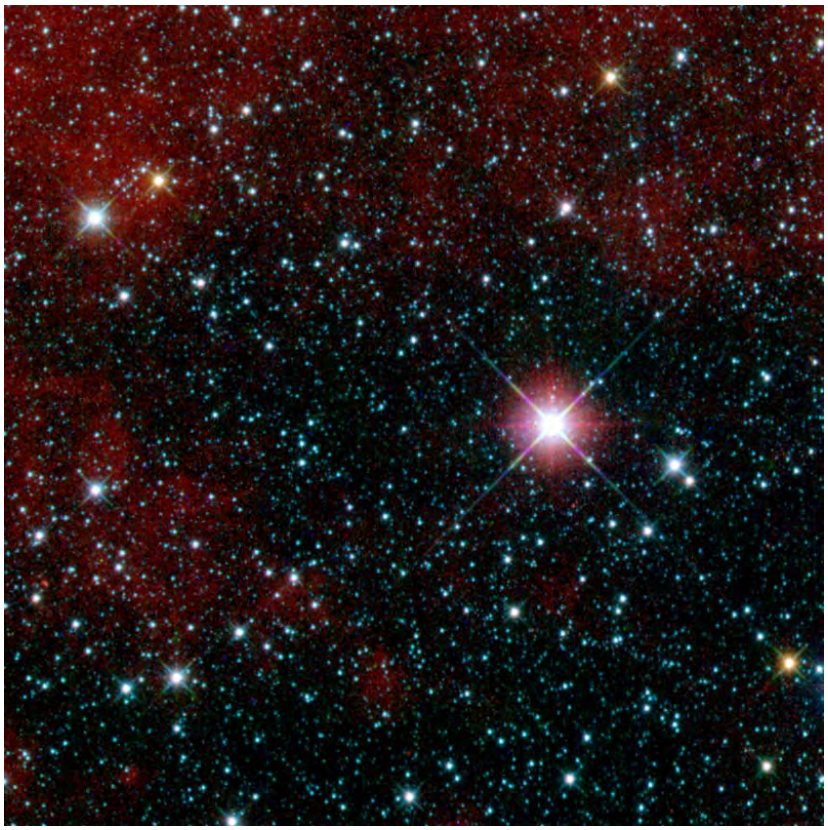


Herschel's HIFI is also working well, as evidenced by this data set that shows detection of carbon in a galactic medium:

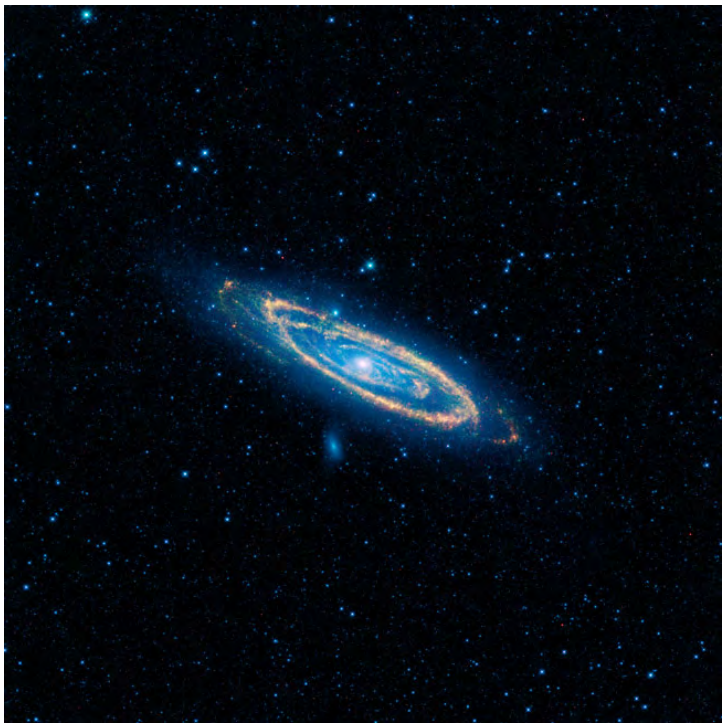


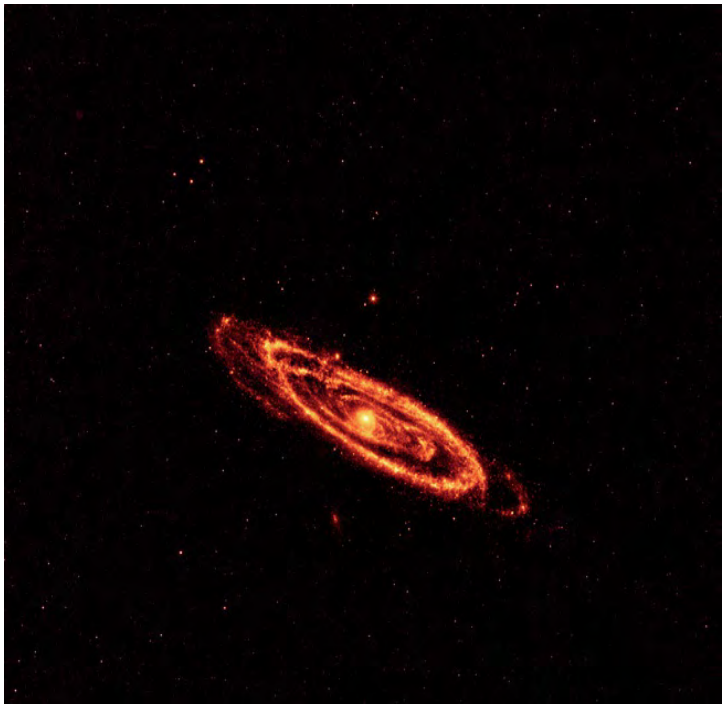
Herschel in mid-summer of 2009 was still in the instrument readiness phase but routine data gathering is expected by November.

Another IR satellite, WISE, the Widefield Infrared Survey Explorer was launched on December 14, 2009. Read about it at this University of California-Berkeley website. It will produce two complete sky maps by October, 2010. These will have more detailed images in the infrared than ever before. Here is the first image:



Typical of the kind of images produced by WISE are these three views, at different wavelengths, of the nearby Andromeda galaxy:





Now, lets move still farther out to longer wavelengths in the EM spectrum. Astronomical objects, in particular galaxies and supernovae, emit the gamut of radiation across the spectrum. Galaxies are usually strong emitters of microwave radiation, in particular in the radio region. Radio waves are generated by excitation of neutral Hydrogen. A good general review of radio astronomy has been prepared by the Haystack group at MIT.

The specialized field of radio astronomy utilizes large "dish" antennas to capture the long wavelength radiation. One of the first radio wave monitors is the famed Arecibo site in Puerto Rico, in which the parabolic receiver is embedded in a limestone sink in the jungle. The dish, 305 meters (just over 1000 ft) wide, is fixed in orientation and must use the rotation of the Earth to examine parts of the astronomical heavens.

The largest movable telescope in the world is the 100 meter radio antenna facility at Effelsberg in Germany. It can both rotate and swing up and down.

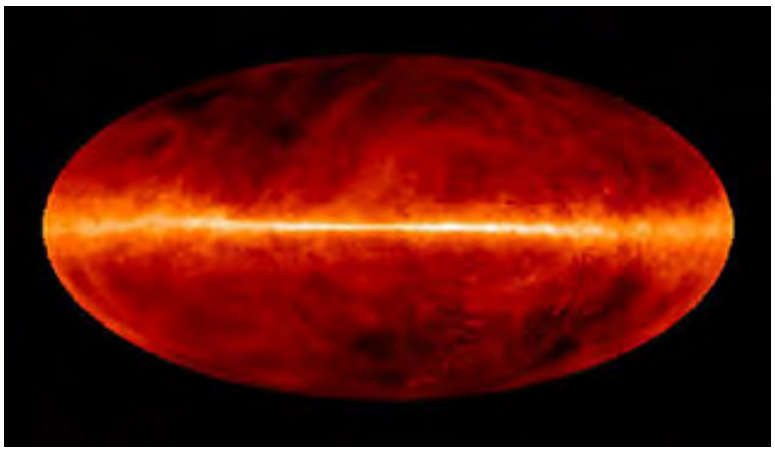
Resolution of celestial targets from which radio waves emanate can be improved by developing a synthesized aperture by means of electronically hooking together individual radio telescopes. A major facility in the National Radio Astronomy Observatory group is the Y shaped array of 27 radio telescopes, each 25 m (81 ft) in diameter, located in the flats 70 miles west of Socorro, New Mexico. This creates an effective resolution of 36 km (22 miles). This Very Large Array (VLA) mode uses principles of Interferometry to process the signals from each telescope as a unit.

In essence, the same signals are received almost simultaneously at different receivers. When added together these may be out of phase and may cancel out or reenforce at specific

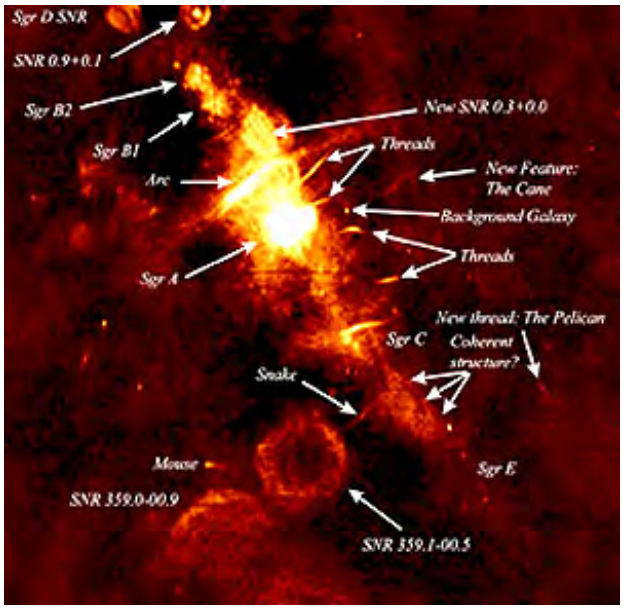
wavelengths; computer processing allows a new interference signal to be produced.

Radio telescopes separated by hundreds and even thousands of kilometers can be tied together by electronic wiring or radio signals to each other to produce an array called VLBI (Very Long Baseline Interferometry). The effect of integrating the telescoope signals is to increase the resolution significantly, so that smaller features in radio objects can be discriminated.

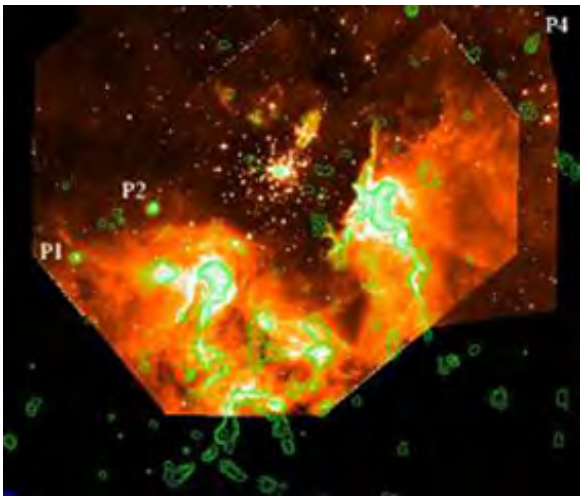
One of the major tasks of radio astronomy was to survey the sky at 21 cm to pick up the distribution of neutral Hydrogen in the Milky Way and the halo around our galaxy. Here is the result:



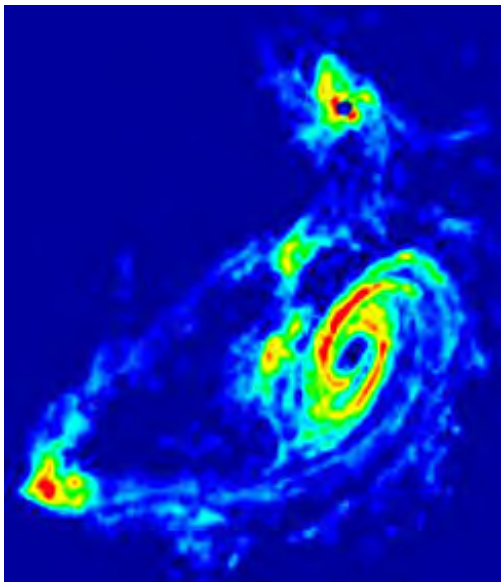
More details about the central region of the Milky Way appear in this radio telescope image made at 90 cm.



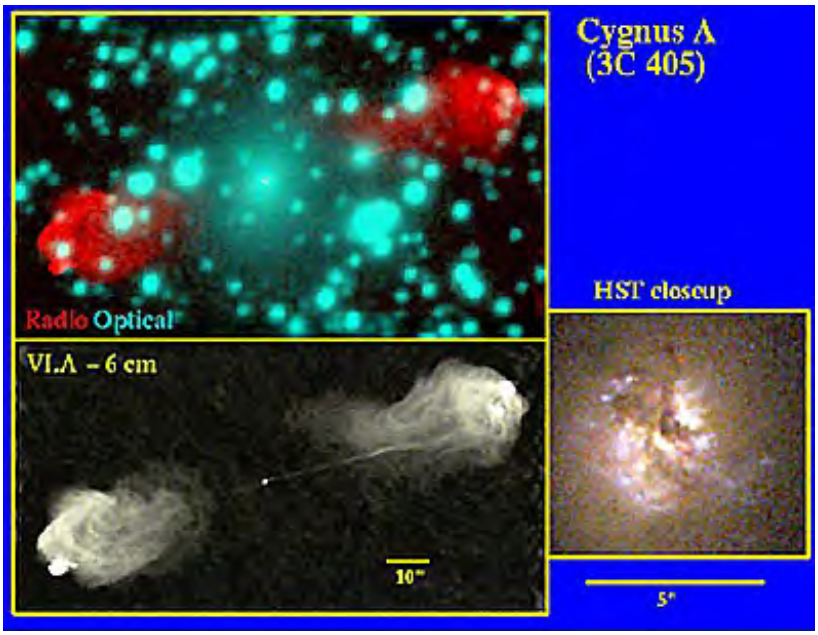
Within the Milky Way, regions of more intense radio wave sources can be identified and mapped. These regions have been called "radio blobs", referring to their somewhat diffuse nature. The blobs consist mostly of ionized Hydrogen gas (plasma). The green areas in nearby galaxy NGC3603 exemplify radio blob distribution.



Most galaxies are so far away that their internal localized sources of radio waves cannot be resolved. Whole galaxies are imaged at the 21 cm H wavelength. Here is M81:

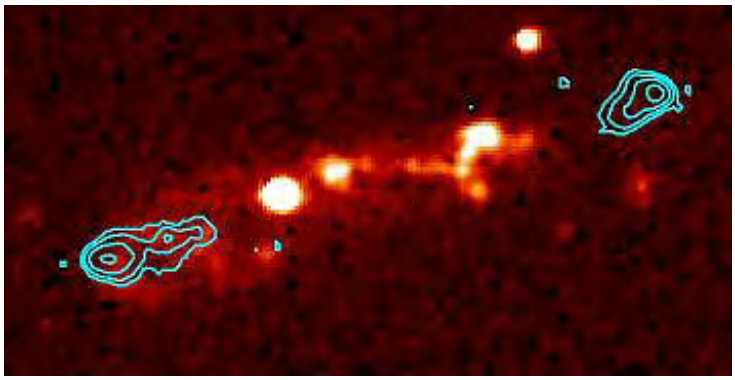


In the early days of radio astronomy, many radio sources in deep space were discovered but when the same region was examined by optical telescopy often no obvious galaxy or other stellar body was found at first. Later observations at non-radio wavelengths have now detected the astronomical feature, usually a galaxy (many galaxies are very strong radio wave emitters). One of the best examples of powerful energy emitters in which visible images do not detect any obvious sources is Cygnus A, from a galactic center about 700 million light years away. Cygnus A is the strongest radio wave emitter in our part of the Universe. Consider these images:

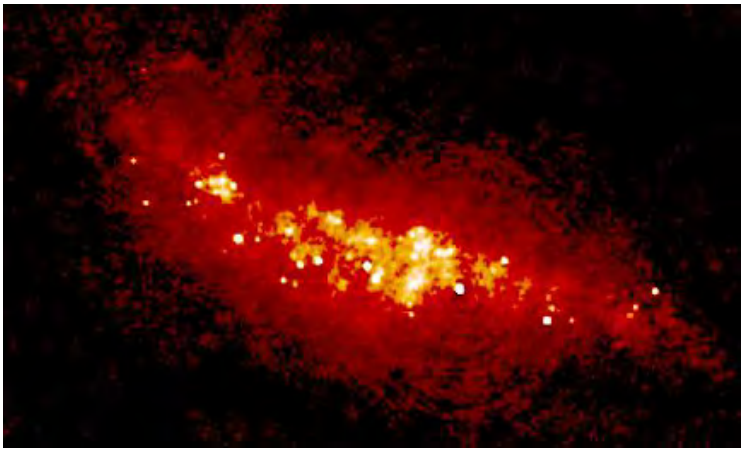


In the above image, the upper left shows a visible light image (star groupings in bright blue) but with no obvious galactic shape. Superimposed, as colorized in red, are two distant lobes representing radio wave signals associated with Cygnus A. The lower left image is another radio wave rendition of signals received at 6 cm. The lower right image, made by HST, reveals some strong radiation coming from the central region of Cygnus A.

Here is a galaxy seen in the Infrared, on which is superimposed the intensity contours associated with two radio sources in the limbs that once seemed isolated from this distinct galaxy.



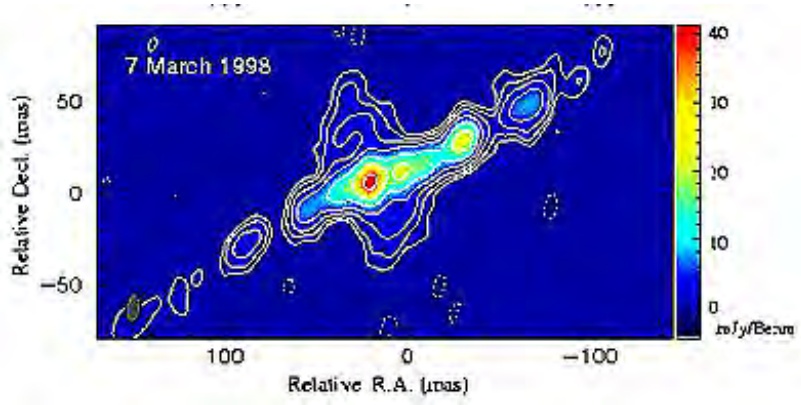
This next image shows an L-Band image of the Starburst Galaxy (M82); this was made at the Jodrell Bank Radio Telescope Observatory near Manchester, England, one of the premier facilities in the field. The signals were obtained from the MERLIN (Multi-Element Radio Linked Interferometer Network) array.



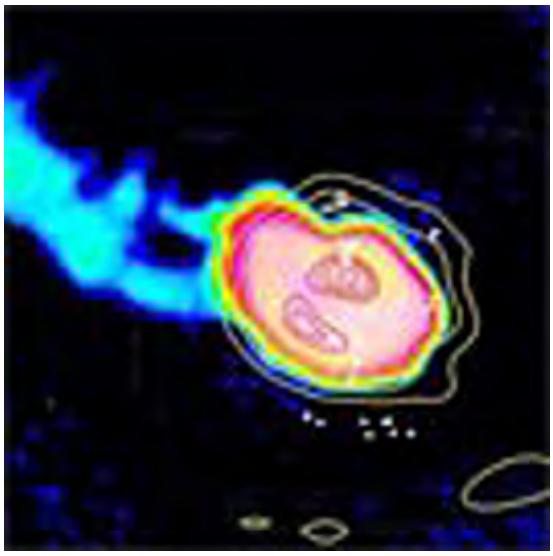
Some of the radio telescope images look remarkably like those made from telescopes imaging at

much shorter wavelengths. A case in point is this image of a supernova, SN2009bb, with a ring of ejected material, as image by a VLA telescope bank:

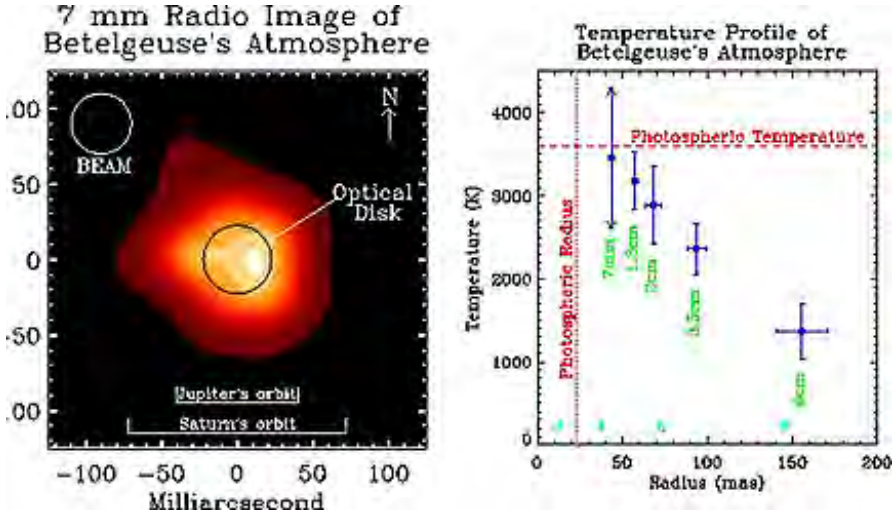
A longer wavelength radio image acquired by the MERLIN VLBI system shows the binary star pair SS433. Contour lines show the extent of radio wave activity outside the central region occupied by the star pair.



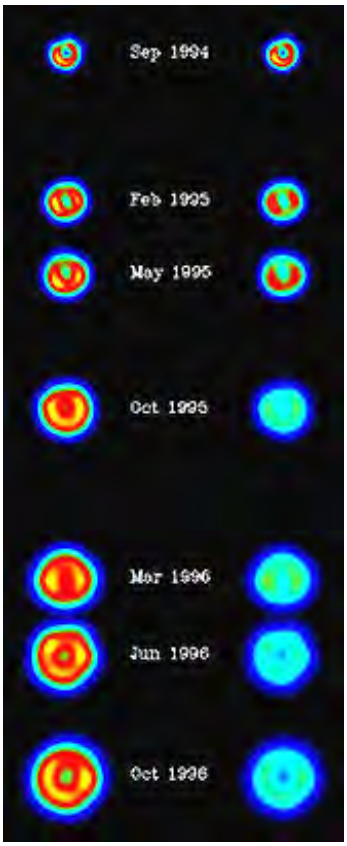
Radio telescopy can identify various organic compounds in space. A spectacular result, using MERLIN, was found in observing a star in the Milky Way from which a gas plume 463 billion kilometers (288 billion miles) long emanates. A large part of that plume is composed of methanol (methyl alcohol; not "ethyl", the drinkable kind - Distillers of the World, Sorry!):



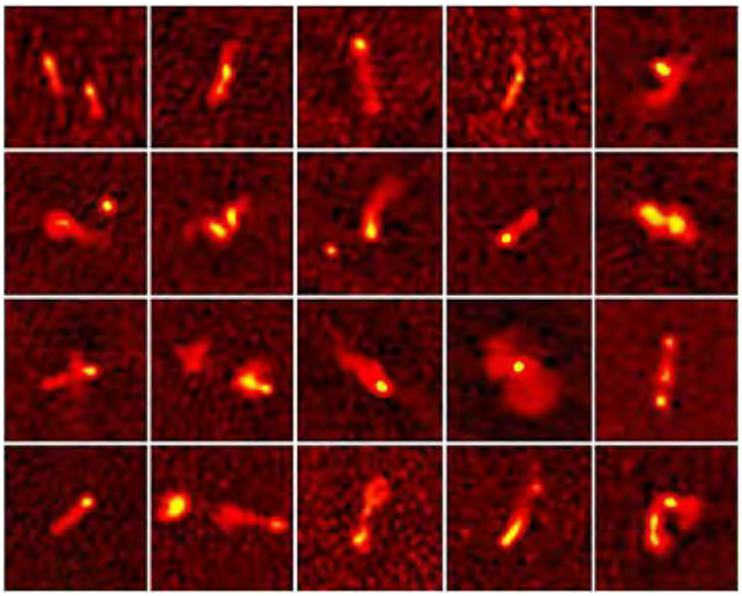
The Red Giant Star Betelguese (see page 20-2) has been imaged within the microwave region (outside the main radio interval) at 7 mm. Under these conditions it was possible to measure a temperature profile (right) in the expanded gas envelope (photosphere) around the star.



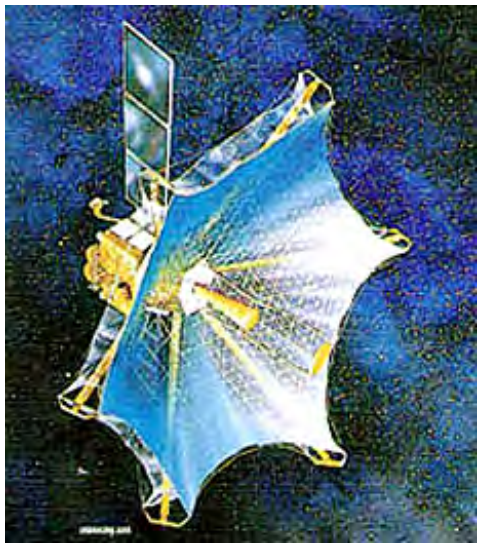
Supernovae (see page 20-6) are strong sources of radio waves. They expand so rapidly that time lapse images taken months apart can monitor their spread and the changes in shape of the radio wave field. Here is such a sequence for Supernova SN1993J in the galaxy M81. The images on the left were taken at 3.6 cm; those on the right at 6.0 cm.



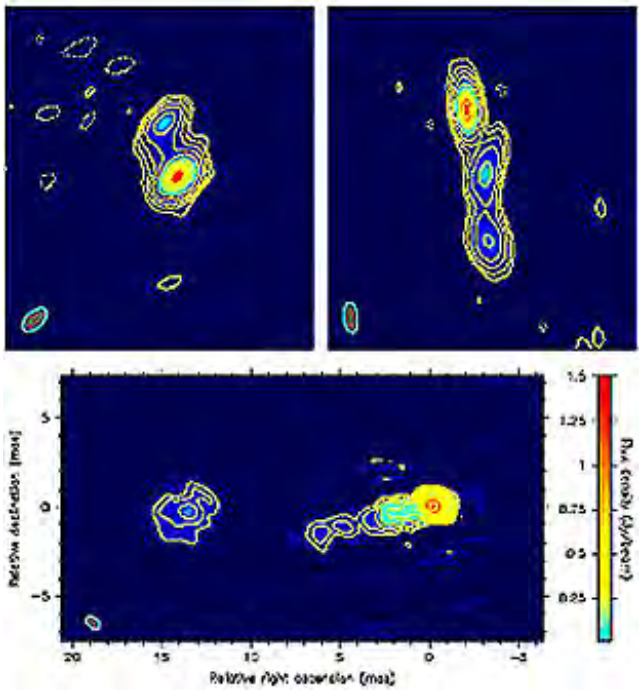
As more radio telescope images of galaxies have accumulated, a distinct pattern has been found with galaxies having a powerful central Black Hole (see page 20-6). One or more powerful jets (material being expelled at speeds approaching that of light) are the hallmark of this type. Below is a series of panels picturing different radio galaxies that show these lobes of ejected materials.



The reader might have had a thought during this review of radio astronomy: Why not put a radio telescope in space? But, wouldn't the antenna have to be much larger than is commonly on satellites? The answer is "No" if the VLBI concept (above) is employed. The Japanese Space Program has developed and launched HALCA (Highly Advanced Laboratory for Communications and Astronomy) in February 1997 as the kingpin in their VSOP (VLBI Space Observatory Project) program. The radio satellite has a 25 m antenna and looks like this:



HALCA's orbit is elliptical, with its perigee (closest approach) at 1000 km and apogee (farthest) at 20000 km. When coupled electronically with one or more radio telescopes on the ground, the effective diameter of the joint system is greater than that of the Earth itself (12755 km). This creates a very high resolution radio wave detector (in some applications, 1000x better than the HST) when used in the Interferometer mode. Although HALCA experienced some trouble in 1999, it did send back considerable data and proved the concept of using multiple integrated radio receivers to achieve exceptional resolution. Here are three images of quasars (see page 20-6) at considerable distances from Earth that illustrate one of the ways in which HALCA data can be displayed:

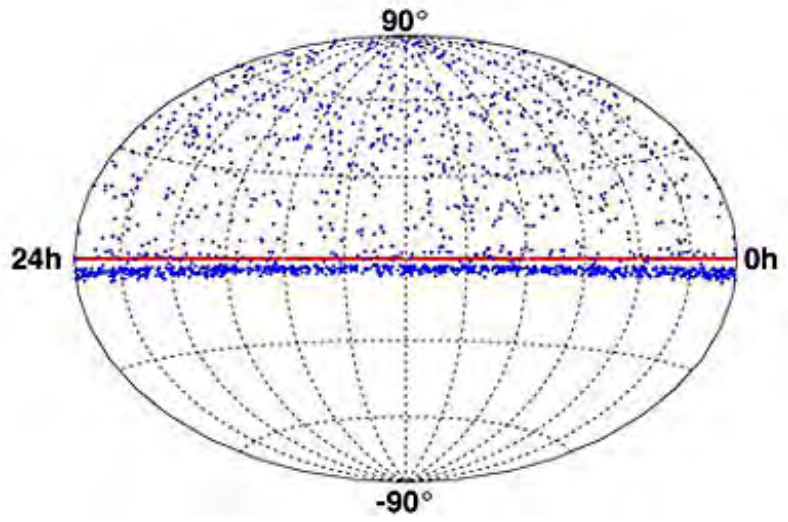


Plans to put other radio telescopes in space are now active. JPL has a brief synopsis of the forthcoming Space Interferometry Mission (SIM) on its SIM web site.

Since we have introduced the specialized technique of interferometry on this page, it is now appropriate to revert back to imaging in the visible spectrum to mention the CHARA (Center for High Resolution Astronomy; operated by Georgia State Univ. astronomers) project which is now commencing operation at the famed Mt. Wilson Observatory (in the mountains north of Los Angeles), shown here:

The large central observatory dome houses the famed 100 inch Hooker telescope that Edwin Hubble used to track down galaxies outside the Milky Way and to measure redshifts, laying the foundation for the Big Bang model. In the above picture are several of the 6 auxiliary optical telescopes tied to the main telescopes. Working in pairs, and later in larger combinations, light from separate components of the array must be combined and synchronized to produce interferometric images in which the waves reenforce rather than cancel. This multiple system produces a baseline (at optimum, 1080 feet) that greatly increases the angular resolution of the central telescope, thus providing images that are expected to exceed the Hubble Space Telescope in sharpness. To get the signals from two or more telescopes into coincidence (the light arrives at any two pairs at slightly different times), one beam is sent through an optical pipe that contain movable mirrors mounted on rails (the "delay line"). The mirror(s) are moved until the extra distance traveled by light to the second telescope (relative to the first) is just compensated enough (equalized) to bring the two signals into phase. This delicate adjustment is made through a computer program that controls pathway adjustments.

One of the most unusual observatory systems now operating productively is AMANDA-II (Antarctic Muon and Neutrino Detector Array), designed to not only detect neutrons but to locate them in the celestial sphere and possibly associate any concentrations of neutrons with discreet sources. The AMANDA is a series of more than 600 glass optical detectors buried in the solid ice 1.5 km below the surface. Here is the first preliminary map of results:



So far, some of these blue dot point sources have been matched with galaxies; others have yet to be correlated with known sources. The neutrinos seem to be generating in the interiors of large galaxies, particularly those with suspected supermassive Black Holes. However, the thick blue band near the equator relates to the Milky Way, which demonstrates other possible stellar

sources.

Astronomy and Cosmology are live and well! Space Observatories are being planned for the next decade or so. Here is a partial list - Click on each to go to its Web Home Page: Herschel; James Webb Space Telescope; Kepler; Laser Interferometer Space Antenna; Nuclear Spectroscopic Telescope Array; Planck Surveyor; Single Aperture Far Infrared Observatory; Space Interferometry Mission; Supernova Acceleration Probe; Wide-Field Infrared Survey Explorer

Ground telescopes are being improved, as new ones are built. Here are several:Atacama Large Millimeter Array; Large Binocular Telescope Interferometer; Large Synoptic Survey Telescope; Pan-STARRS; Square Kilometer Array (radio telescope)

With this examination of observatories that collect data over different parts of the spectrum, we now return to the exposition of aspects of Cosmology by looking next at the some special topics relating to galaxies.

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Some Special Features of Galaxies

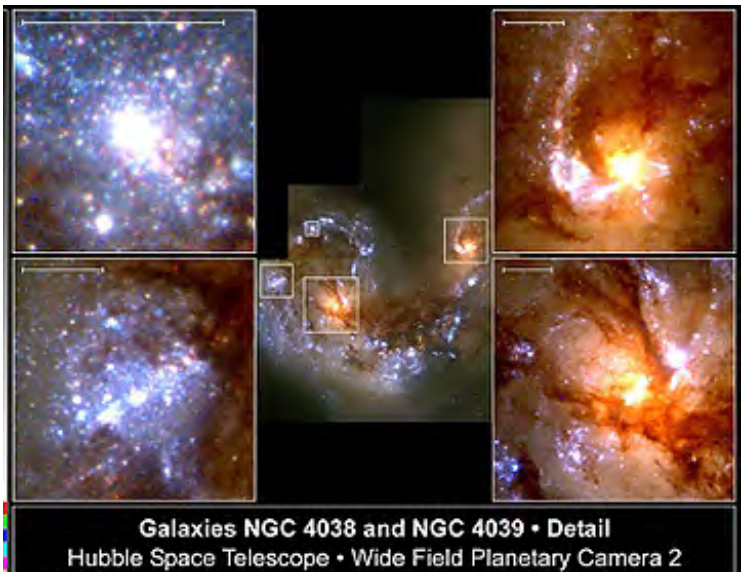
Colliding Galaxies

It is now the general doctrine that many (estimated at more than 50%) galaxies have been involved in one or more "collisions" with other galaxies, or even merged within galaxy clusters. This was more common in early Universe time when galaxies were then much closer. The word "collision" is put in quotes to call attention to the actual nature of the event. In effect, two galaxies that meet 'interact' in the sense that they join into one with only a few stars actually bumping together in a destructive manner, since the distance between stars is much greater than star sizes themselves. The term "merge" is probably more descriptive of the process involved.

It is not necessarily evident from a telescope image that galaxies are colliding. For instance, look at this Hubble view of ARP 274:

This pair was won out in an Internet "contest" run by NASA as the first choice of amateur viewers for the target of 100 hours of observation by the HST to advance our knowledge of the Universe. No rationale for this choice was given: it is extremely unlikely that anything by way of noticeable change will occur in the lifetimes of the contestants. But the viewers seemed to want more detail.

One thing that a collision leads to is the production of new stars as the gases involved are stirred up by the interaction. A release on October 21, 1997 from one of the HST research teams describes important information on star formation within an evolving galaxy system that is undergoing collision. The image below shows a ground-based telescope view of two colliding galaxies which together make up what is called the Antennae galaxy, so-named from the long wisps of luminous gas extending like an insect's antennae. The central panel shows the pair together. Because of its proximity (63 million light years away), it has been a prime candidate for a closer look. The left and right images are much higher resolution HST views (note boxes) of the central galactic mass of merged stars from the two once separated galaxies. More information about the central interior of this developing supergalaxy, and about regions of active star formation appears in this image:



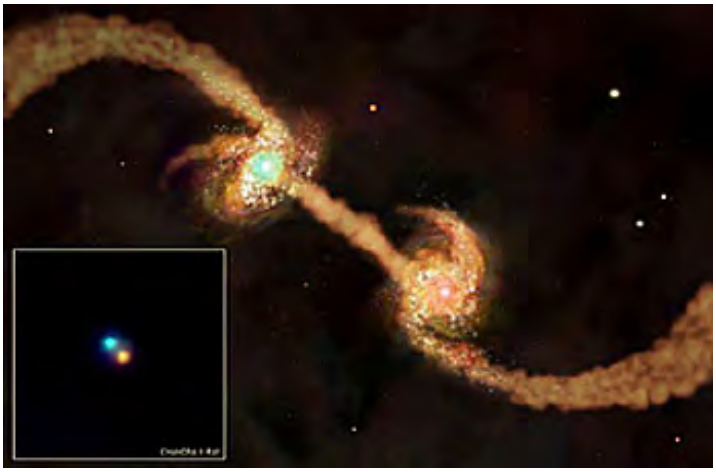
The surprise is the numerous clusters of blue stars. Each appears to be groups of up to a million young (hence bright and hot) individual stars. The clusters likely are still developing, as cold Hydrogen gas in giant molecular clouds (typically 100s of light years across) distributed in pockets through each galaxy are being squeezed during the collision process. They contract and heat up into individual stars as this goes on, often collapsing rapidly enough for many of the stars to explode almost like "firecrackers". Other pre-existent stars are likely to be destroyed as the collision continues. The two orange centers are the older surviving parts of each galaxy.

A closer view of the Antenna galaxies has been made by combining images from several space observatories: Chandra (blue), Hubble (orange and brown), and Spitzer (red), seen next:

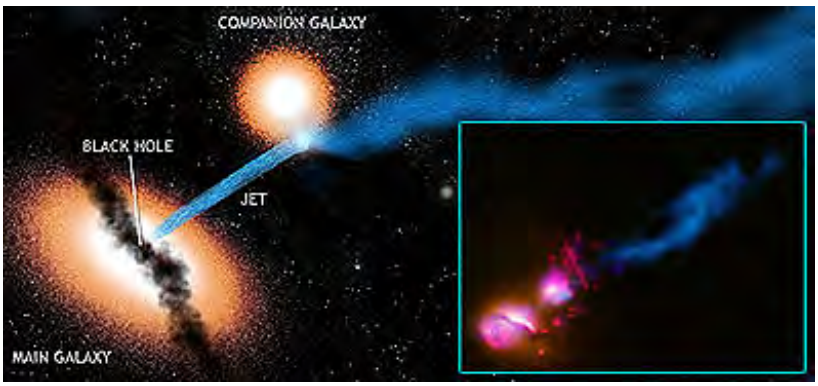
Mayall's Object, ARP 148, is an example of a collision which has drawn out one of the galaxies:

Still another spectacular collision seen by HST is that just beginning between NGC2007 and IGC2163. As the two merge, most individual stars will not collide with another star, since in any such galaxy the distance between stars is actually huge, lessening the changes of direct collisions (although as the pair of galaxies pass through, any given star must always face the possibility of encounter with a star somewhere along their mutual paths):

As galaxies approach, materials (gases; dust) are exchanged between them, new stars are born, and Black Holes grow. This is clearly going on between two (unidentified) galaxies that were imaged by Chandra:

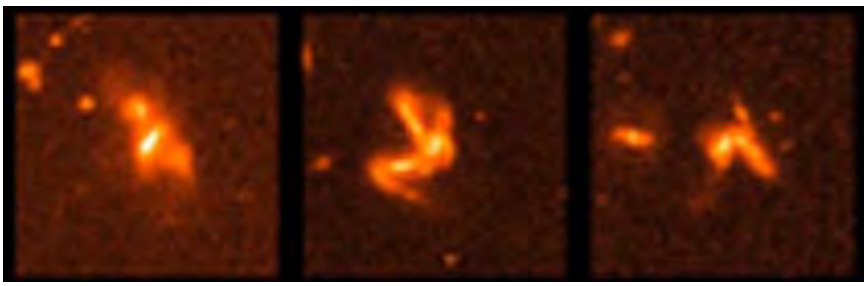


Another example is galaxy 3C321. A jet of gas/dust is being expelled to a smaller nearby galaxy (about 20000 l.y. away), but some of that is also deflected into outer space. The actual Chandra image is in the inset at lower right; an artist's rendering is the main illustration (which portrays the intergalactic jet that is poorly displayed in the Chandra view):



The galaxy pair at ARP 194 shows large blobs of hot material exchanging between the two:

These observations support the growing view that collisions were a more common process in the early Universe (but still happen even now). Perhaps as many as one-third of the ancient galaxies collided during the lengthy period when galaxies were much closer, i.e., Big Bang expansion was less far along. Sometimes more than two galaxies are involved in the colliding process. One solid indication of collisions is the notable irregularity of the galaxy composite, with irregular center(s) and distorted spiral arms. These three HST examples of multiple collisions, imaged in the infrared, illustrate that:



In this recent image, made by combining images obtained by the NICMOS and ACS sensors, 4 galaxies can be resolved individually. As a pair merge, the increased gravitational attraction of the composite can draw in nearby galaxies to foster further enlargement of a galactic grouping.

Another well known grouping of close-spaced galaxies that appear to be headed for some kind of amalgamation is Stephan's Quintet, with three of the five seen in this HST view:

One type of galaxy with a distinct Active Galactic Nucleus is the Seyfert class, mentioned on page 20-3. Seyfert galaxies have strongly ionized Hydrogen, Helium, Oxygen, and Nitrogen in their central region. Interaction between infalling gases and a Black Hole is the probable cause. The Seyfert Sextet is a group of six galaxies, 190 million light years away in the constellation Serpens. They consist of 3 ellipticals and 3 spirals (only five visible in this orientation; the small spiral galaxy seen face on is not in this group, being much more distant). None of the galaxies is more than 35000 l.y. across. This configuration has been interpreted as a congregation of galaxies in the process of colliding and being ripped apart by gravitational interaction. The elongate bright central areas in two regions of the cluster may be the cores of merging pairs of galaxies. Unlike some colliding galaxies, there is no visible evidence of bright new stars being formed in these phases of collision.

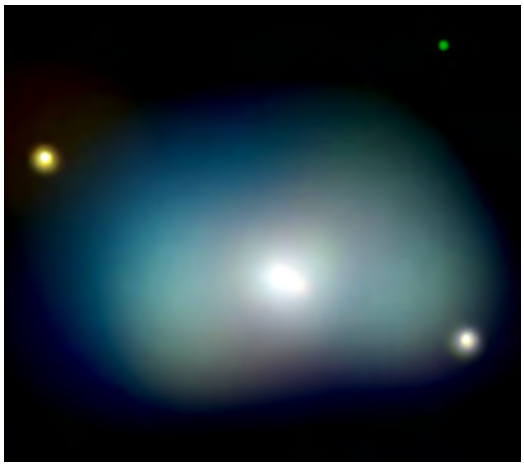
This next pair of spiral galaxies are also starting their collision interactions. Note the stream of gas and dust between them. Stars are forming in this bridge.

The usual end product of the merging of two spiral galaxies is an elliptical galaxy; many elliptical galaxies formed this way. Collisions can also give rise to spiral structure. Some globular clusters also presumably originated from interactive collision.

One of the more visually intriguing results of a collision is the Cartwheel Galaxy (below), another Ringed Galaxy, in which the passage of one galaxy through another generated shock waves traveling at high velocities. As these waves moved outward, they condensed Hydrogen into a huge collection of new stars that lie along the front of the advancing waves. This image released appears to locate one or two galaxies that may be the ones that passed through the Cartwheel galaxy, producing shock waves responsible for several rings in which the new stars are forming. The prominent outer ring contains the largest number of resulting stars:

Here are three more galactic collision images:

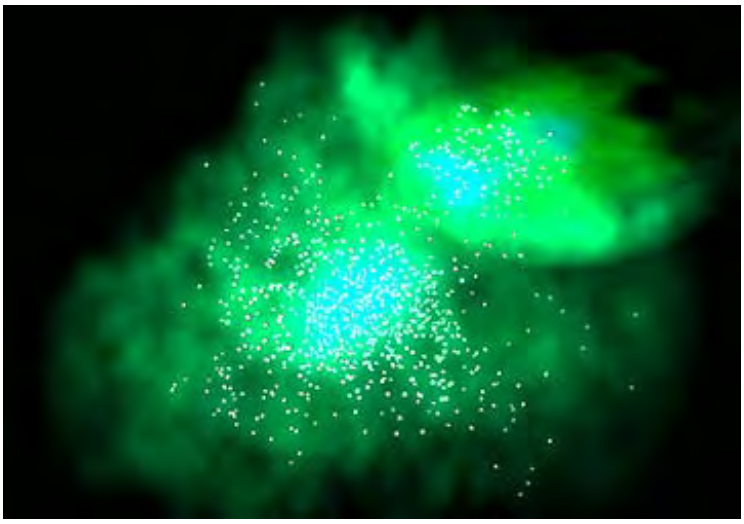
Galaxy collisions can release copious amounts of energy. The Chandra X-Ray telescope has detected a huge release of X-rays stemming from the elliptical galaxy NGC 1700, located 160 million light years from Earth:

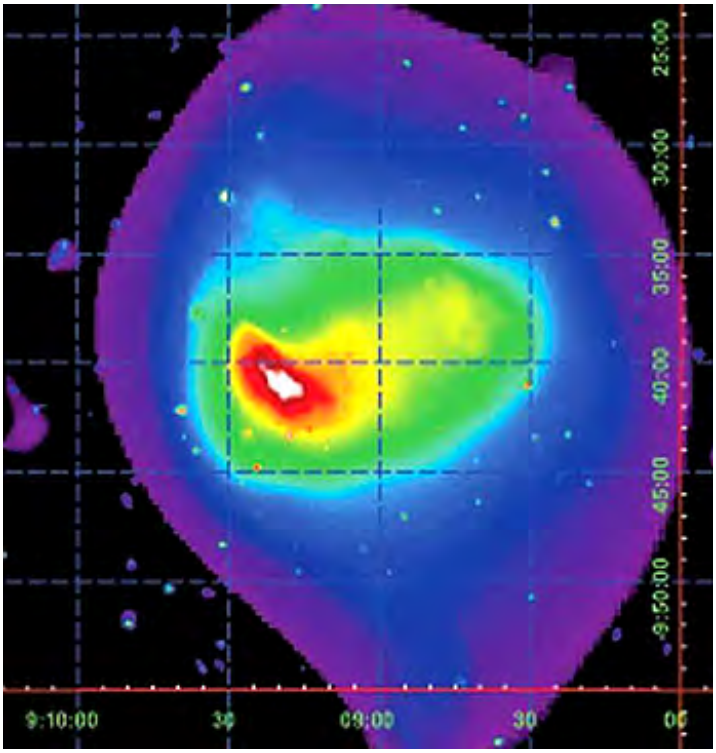


At 90000 light years in diameter, this is the largest X-ray source near Earth yet discovered in the Universe. The emissions come from a vast spinning cloud of Hydrogen gas excited to temperatures in excess of 8 million degrees. Astronomers studying this cloud surmise that the collision was between a spiral and an elliptical galaxy.

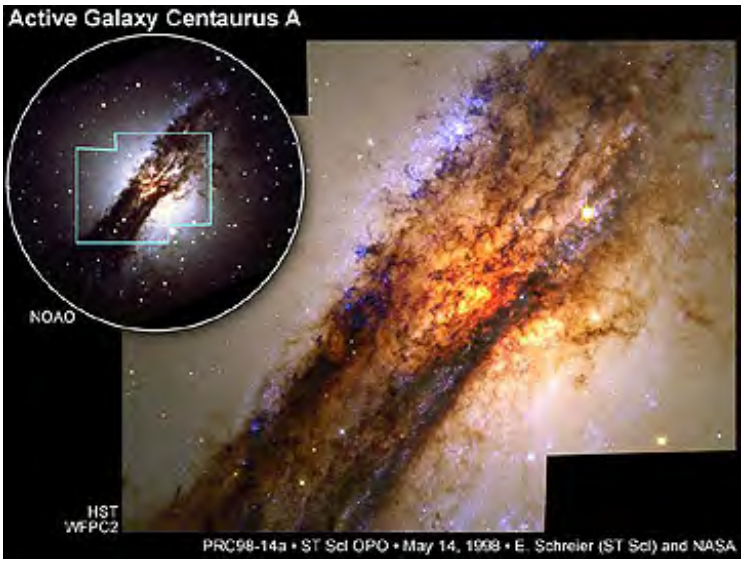
But the biggest collision by far is Abell754, found within the Hydra constellation (but more

distant), in which two small *galaxy clusters* have been colliding over the last half billion years. This yields a tremendous source of X-rays as detected by ESA's XMMM-Newton Observatory. Below is a view produced from data received, but with the image touched up by an artist for emphasis. And beneath that is a plot of energy variations within the colliding clusters that will eventually produce one huge galaxy (supporting the prevailing concept that most galaxies in the Universe have built up by the collision process).



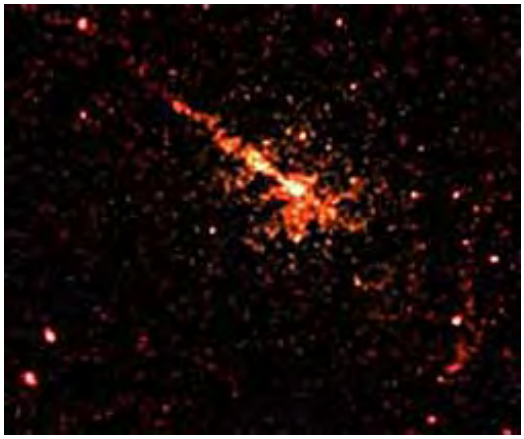


Although now less common as the Universe expands, drawing galaxies apart overall, galaxy collisions are still taking place throughout the Universe, as illustrated by recent HST observations of the Centaurus A (NGC5128), the closest *active* galaxy to our own Milky Way galaxy. Centaurus A, itself much larger than the Milky Way, is a known radiation "hot spot", being the source of intense X-rays and Radio waves. This galaxy is only 10 million light years away; what we see now represents its condition at 10,000,000 years prior to today.



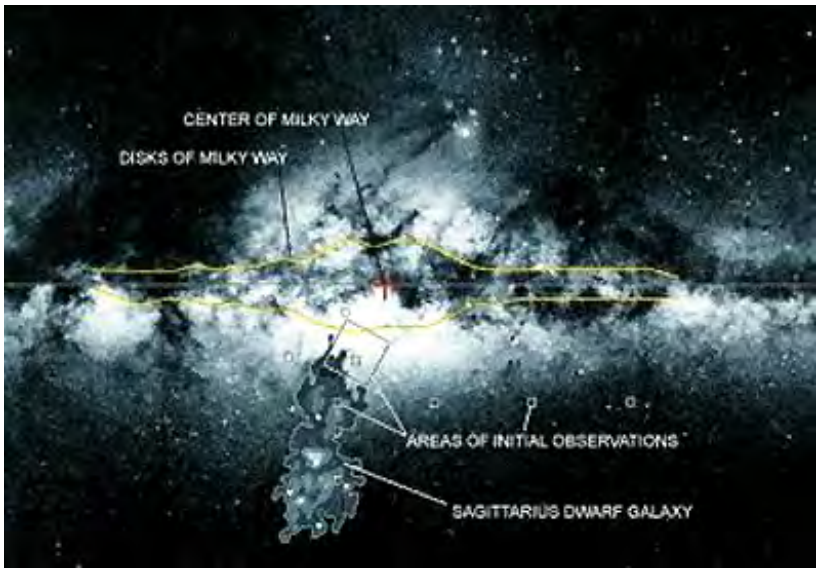
The circular inset shows Centaurus A as seen optically through a ground-based telescope. The detailed view to the right was acquired by HST's Wide Field Camera. An elongate disc, marked by dark dust, is spread across a large white glow that is identified as an elliptical galaxy. This pairing is interpreted to be an intermingling of a spiral galaxy in collision with this elliptical galaxy. The Infrared Camera on HST can penetrate the dust to reveal a hot, turbulent mass of stars, dust and gas from the spiral galaxy falling into the core of the elliptical one, as seen in the larger view.

The Chandra X-ray telescope (page 20-3) captured an unsuspected feature of Centaurus A - namely, a jet of material ejected to a distant of 25000 l.y. from the core. This single jet of intense X-ray energy is roughly at right angles to the plane of the disc.



A Black Hole (see page 20-6) is postulated to occur towards the center of the two interacting Centaurus systems. This B.H. may be as massive as 10 billion solar masses, occupying a volume similar to our Solar System. The Black Hole is "sucking" matter from both galaxies into its growing body. This set of observations is the most detailed yet of the consequences of galactic collisions.

One model of future Universe expansion paths indicates that the nearby Andromeda spiral galaxy could come close to our Milky Way galaxy and might even collide with us, several billion years in the future. There is considerable recent evidence that a small galaxy is presently passing through the Milky Way. Known as the Sagittarius dwarf spheroidal galaxy (or Sgr), its presence has been deduced from motions of certain stars that do not fit the motions of M.W. stars in the spiral arms; also star "tails" stretch out in the galactic halo, suggesting that Sgr is in a broad orbit that has caused it to intersect the Milky Way before. The vast distances between stars keeps interactions to a minimum. Sagittarius is the closest small galaxy to the M.W. Here is an artist's idea of Sagittarius as it is presently encroaching on the M.W.

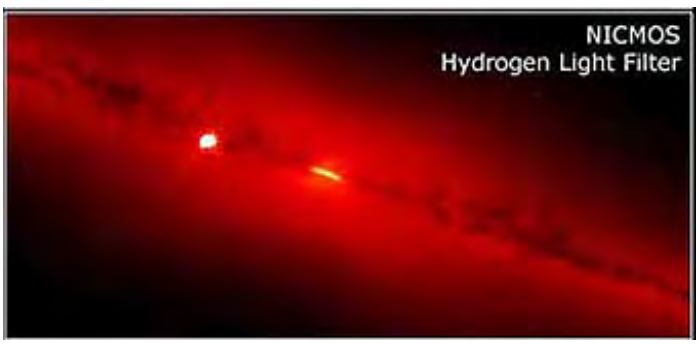


The role of galaxy destruction by merging, leading to a new supergalaxy, has been underappreciated until recently. In the early Universe, proximity of galaxies before expansion had separated them must have led to frequent collisions as the norm. Some astronomers think that most galaxies as seen today were products of earlier collisions.

Galactic Gas

(Note: this subsection was added in January 2002, in response to information summarized in an article for that month in *Scientific American* by Ronald J. Reynolds, entitled "The Gas between the Stars".)

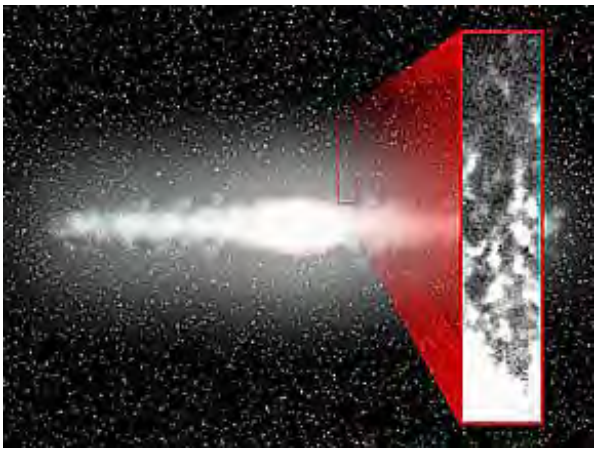
Seemingly empty regions of space - both within and between galaxies -actually contain variably small quantities of matter. The population of identifiable matter (molecules; protons, photons, cosmic rays, etc.) can be very small in space between galaxies and their halos. But, because the totality of intergalactic space is by far the largest volumetrically in the Universe, the small population, expressed as density, is actually the bulk of all known matter. This is particularly true because Dark matter (described on page 20-10), of presently unknown nature, is also present and is predominant. Among the elements, the most abundant species is Hydrogen, occurring as several states; helium is present at about 10% and the higher atomic number element species together constitute only a fraction of 1%. Within galaxies, these same elements make up what has been termed an "atmosphere" to describe the gases and particles not in the associated stars. The dominance of Hydrogen within galaxies is evident from this NICMOS Hubble Space Telescope image of NGC 4013, taken at a infrared wavelength in which Hydrogen appears to glow red:



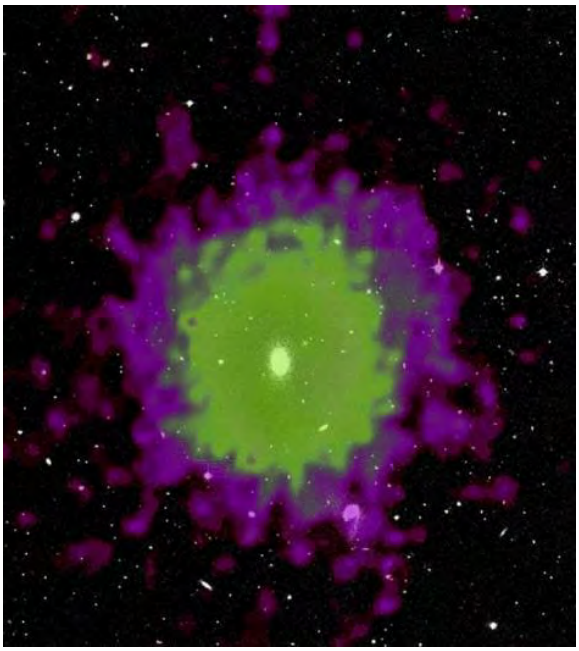
Galactic Hydrogen occurs in the following states within the nebulae present in both the central midplane and to a much smaller extent in the halo : 1) neutral Hydrogen (H_I), found mainly in the central midplane of a galaxy, which has a temperature ~120°K; it is responsible for giving off the 21 cm radiation (1420 MHz) used by radio astronomers to map its distribution; 2) molecular Hydrogen (H_2), with a temperature around 15°K, which, although it comprises only 18% of all

Hydrogen in the galaxy, is concentrated mostly in the gas nebulae in the central plane and is the principal material that organizes into stars; its distribution is mapped both in the 2.2 µm and the far UV regions of the EM spectrum. Outside the gaseous nebulae and predominantly in the halo, the Hydrogen exists in three states, brought about by their higher kinetic temperatures: 1) warm H(I), T = less than 3000°K, about 30% of all the Hydrogen in the galaxy, extending out to 33000 light years (ly); 2) warm H(II), ionized (loss of electron), T =3000-10000+°K, 35% of the Hydrogen, extending out to about 65000 ly; 3) hot H(II), also ionized, T = upwards of 1 million degree K, present at low densities out to about 180000 ly, and making up about 45% of all Hydrogen. Thus, it is now known that nearly all ionized galactic Hydrogen is located beyond the central plane, i.e., in the halo, but make up about 22% of the mass; the stars themselves account for only about 2% by volume but 30% of the mass; the remaining mass is represented by H(I), 35% of which is in the halo and 0.1% in the clouds. To restate this: a spiral galaxy is mostly Hydrogen with some present in stars concentrated in the central plane but also found in smaller numbers in the halo which itself is composed of neutral and ionized Hydrogen whose temperatures are much higher than the Hydrogen gases within the central plane.

Clouds of Hydrogen in the halo region of the Milky Way have been imaged in the infrared. In this next illustration, a view of these clouds is shown in the vertical strip to the right; the side-on image of the Milky Way is an artist's rendition rather than an actual observation.



An idea of just how far a galaxy's gas extends into a halo is evident in this illustration of NGC4325 in which a visible light image of the galaxy itself (center, elliptical, light green) is registered to an X-ray image that displays the tenuous gas (which can reach temperatures here up to 1000000° K) extending to great distances outward diameter to purple edge = 1.9 million l.y.). The purple and green indicate two arbitrarily-divided levels of X-ray emission.

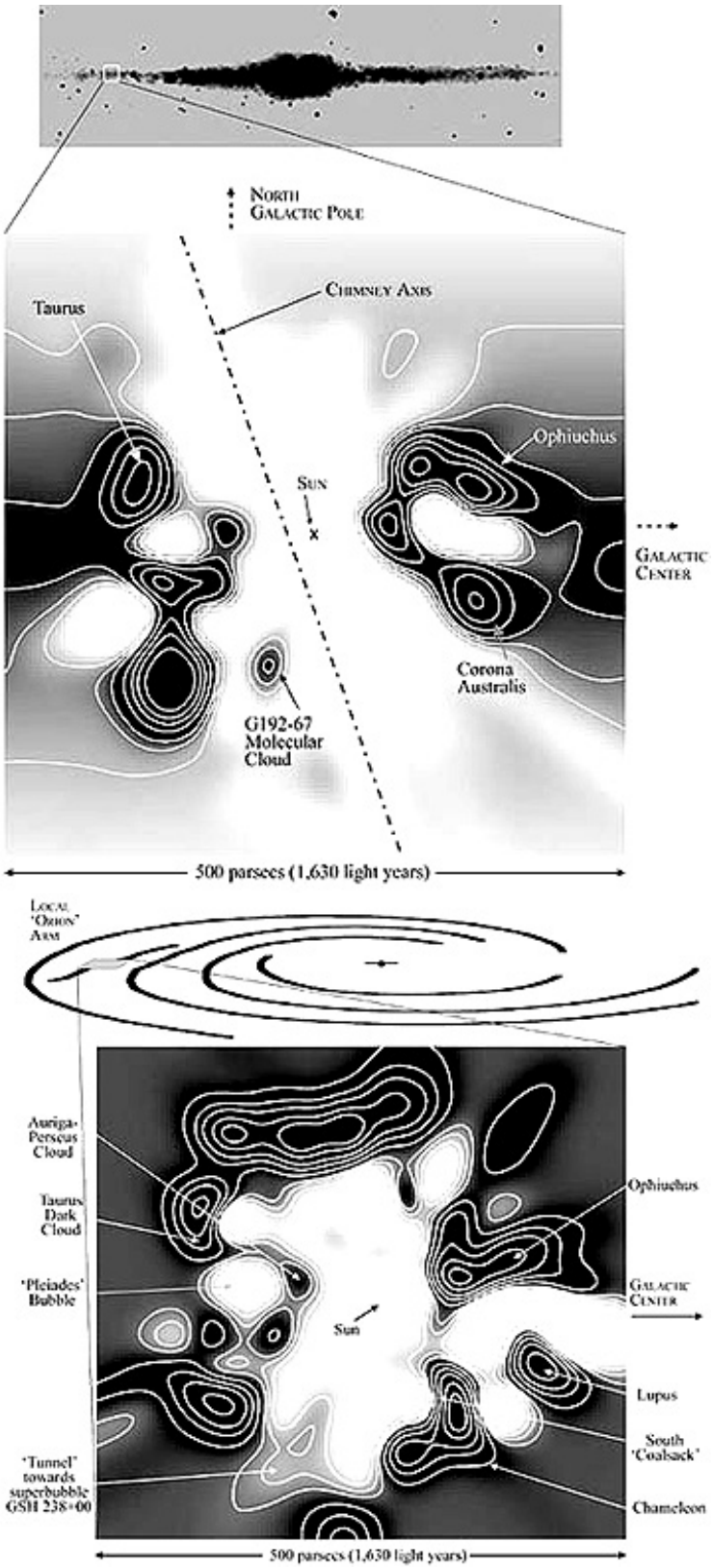


This atmosphere is somewhat analogous to that on Earth (decreasing in density upward) in that its Hydrogen constituents drop in numbers as the outer reaches of the halo are approached. Like the atmosphere of Earth, and similar also to that surrounding the Sun, the galactic atmosphere is dynamically and continuously active. Both within the star-forming region of the central plane and, to a lesser extent in the inner halo, this gas is perturbed by shock waves and expelled particles in stellar winds, generated mainly by supernovae explosions (page 20-6). Many of these explosions are associated with massive, short-lived stars (the O and B types described on page 20-2). The result is "turbulence" driving some matter into denser clots (initiating regions of new star formation but also producing volumes of much lower Hydrogen concentrations called "bubbles"). Conceptually, the galactic atmosphere could be visualized as looking a bit like Swiss cheese, with holes where the bubbles occur.

Indications are that (mostly ionized) Hydrogen is expelled as a plasma from the central plane region after supernova events. This occurs in streamers that are given the descriptive name "columns" and resemble the flares jetted out from the Sun (page 20-3a). They may start as bubbles which rise thermally and draw in Hydrogen gas as they pass into the halo. These move out to distances in excess of 10000 ly, but tend to break up with some material returning to the plane as more diffuse "fountains". Their outward pathways may be controlled by a galactic magnetic field. This predicted phenomenon has not yet been imaged (so its existence is still not proved) but here is an artist's depiction of what it might look like in our Milky Way galaxy (compare this to the illustrations of the Milky Way imaged at various wavelengths on page 20-3):

From Scientific American, January 2002

A survey of gas densities around the Sun in our part of the Milky Way is being carried out by ground telescopes, by FUSE (Far Ultraviolet Spectrometer Experiment), and by CHIPS (Cosmic Hot Interstellar Plasma Spectrometer; also operating in the far UV). In May of 2003, a group at the University of California, Berkeley, and French colleagues, produced maps of the densities out to 1000 light years, as follows:



The top map is a section perpendicular to the plane of the M.W.; the bottom map is a cross-section through the top map in the plane labeled Chimney Axis. The light-toned areas are regions of very hot gases (kinetic temperatures in the 1 million degrees Kelvin range) with lower than average intragalactic densities. The dark-toned areas are higher density, lower temperature regions. The reason for the low density "bubble", actually roughly chimney-shaped with protuberances, is still uncertain. One hypothesis turns to a supernova explosion in the past that may have produced "galactic winds" that blew intragalactic material out of the resulting low density region. At such low densities, atoms of Hydrogen can move greater distances without collisions and thus attain higher kinetic temperatures.

These studies of the processes involving galactic atmospheres are opening up new insights into the factors that lead to star formation during the lifetime of a galaxy. Some examples of

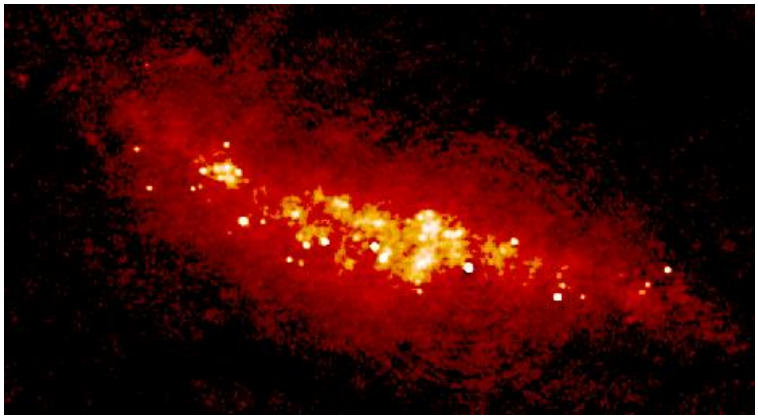
chimneys in nearby galaxies have been imaged by the Hubble Space Telescope and other observatories. Research along the lines described above is likely to improve our understanding of the activities within galaxies that determine their history.

Starbursts and the Active Galactic Nucleus

An article entitled *The Galactic Odd Couple* by Astronomer Kimberly Weaver in the July 2003 issue of *Scientific American* contains so much information that will be synopsized here: Weaver's main thesis is that both Starbursts and Active Galactic Nucleus's (AGNs) are commonplace in galaxies and each seems interdependent in several possible ways on the other and on Black Holes. Furthermore, the two represent the most powerful release of energy (strong in every segment of the EM Spectrum) at galactic scales (hypernovae, quasars, and gamma ray bursts are tops in power output at individual star scales) now known in the Universe. An AGN is thus characterized by strong emissions of energy beyond that of starlight, especially in the radio wavelengths, and by expulsion of matter as radio jets. This diagram shows an artist's model of an AGN, the jets, and starbursts:

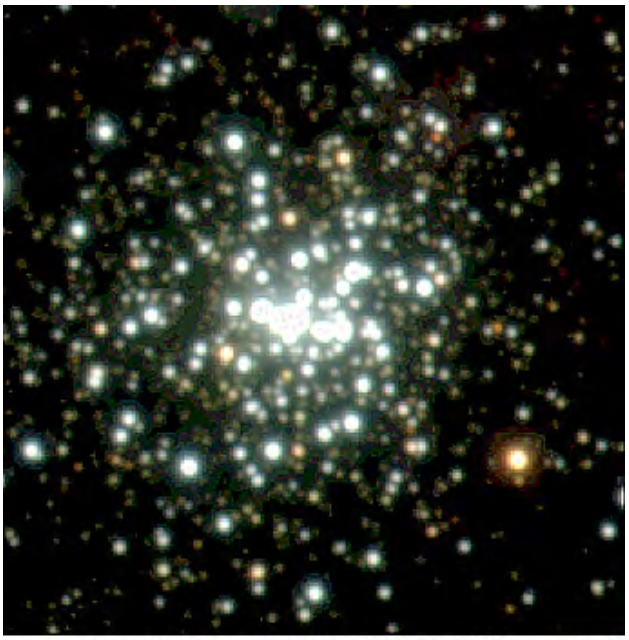
Let's commence by looking at various manifestations of starbursts.

This radio telescope image of M82 shows a number of bright areas which are interpreted to be starburst regions.

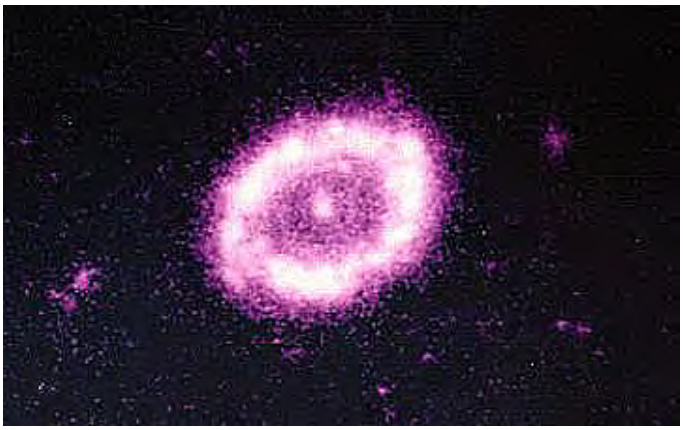


A few more examples of starbursts should affix their appearance and characteristics in your mind.

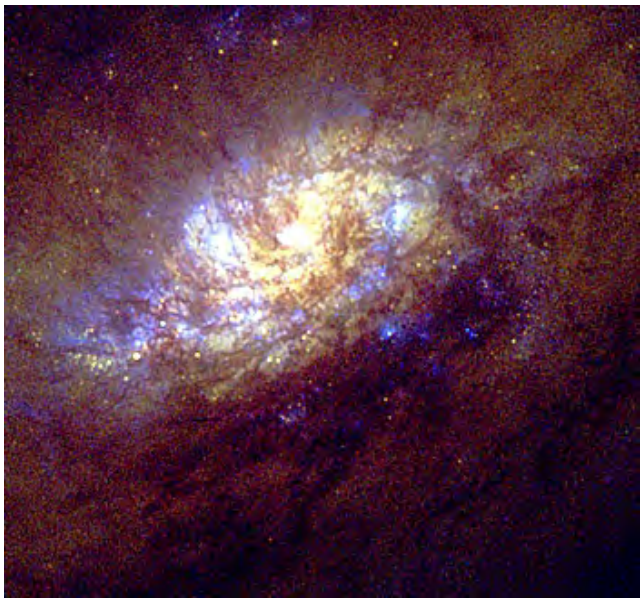
The first image shows a central starburst cluster in galaxy NGC 3603. Below it is a ring of stars in M94, envisioned in the UV. The third image shows a prominent ring of starburst objects just out from the center of galaxy NGC 4314.



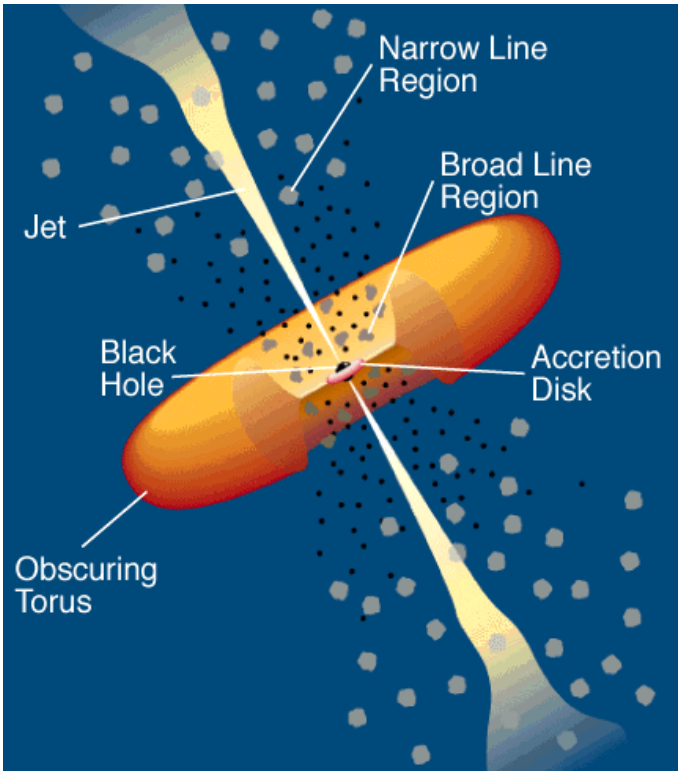
Central Star Cluster in NGC 3603 (VLT ANTU + ISAAC)



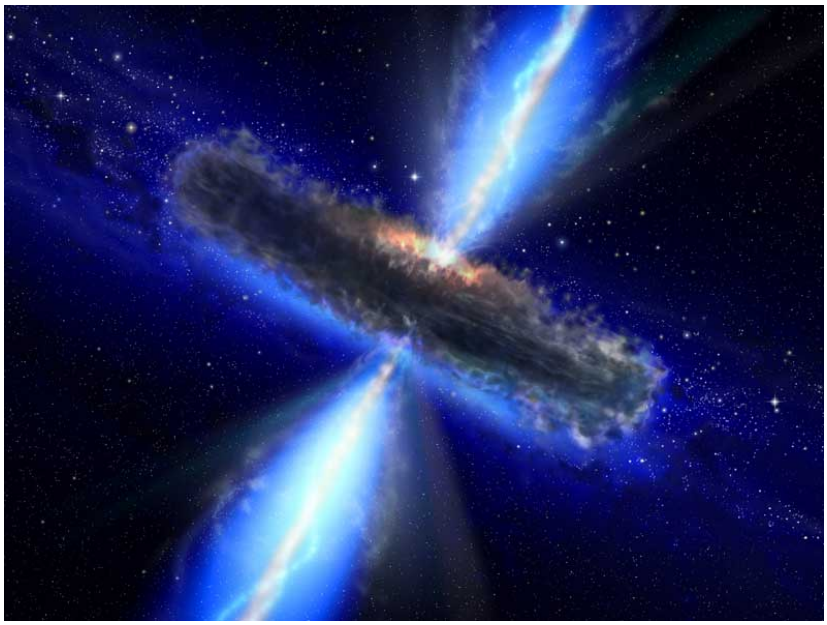
This image of NGC1808 shows a well-defined AGN that may contain unresolved starburst stars as well. Beneath it is NGC9812, showing a mix of AGN excited gas and some starbursts.



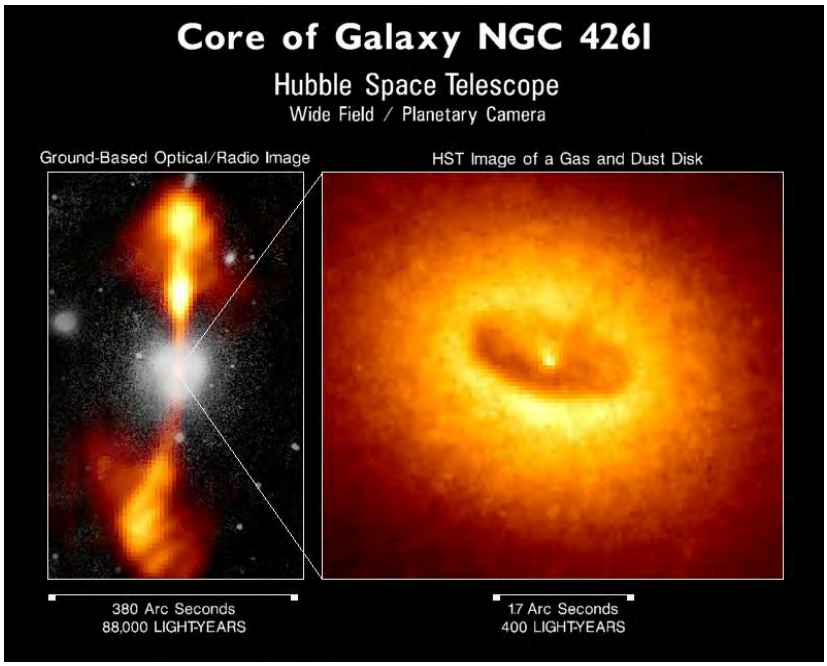
The Active Galactic Nucleus is a region with (usually) a central Black Hole, quasars, a dust-obscuring disc, and often a jet. This is a general mmodel:



An actual example similar to this was imaged by the SWIFT spacecraft.



This illustration shows NGC 4261, with an AGN and a glowing disc.

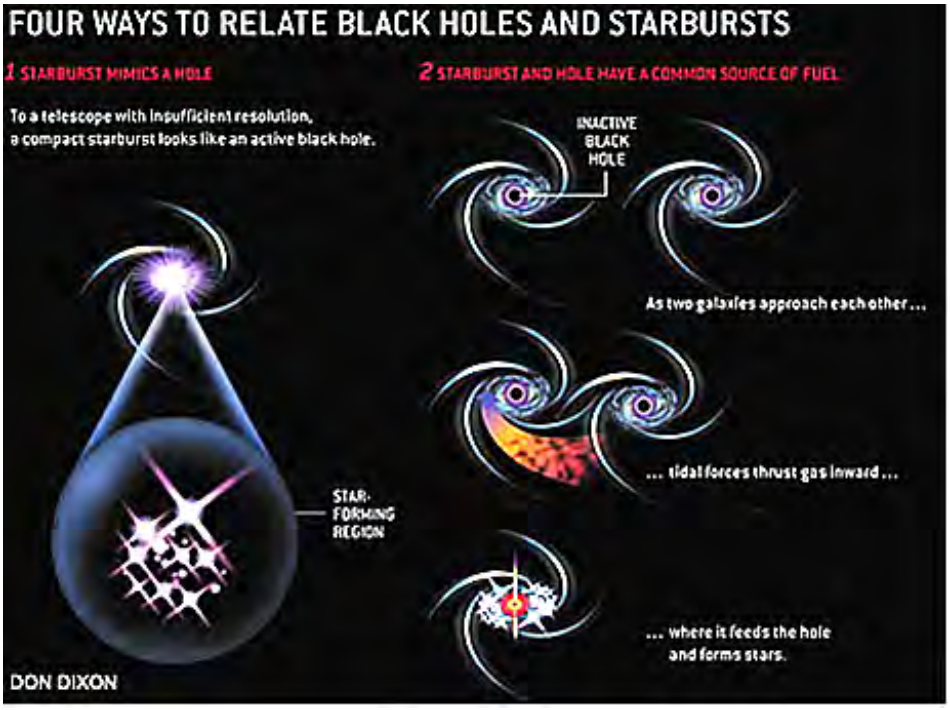


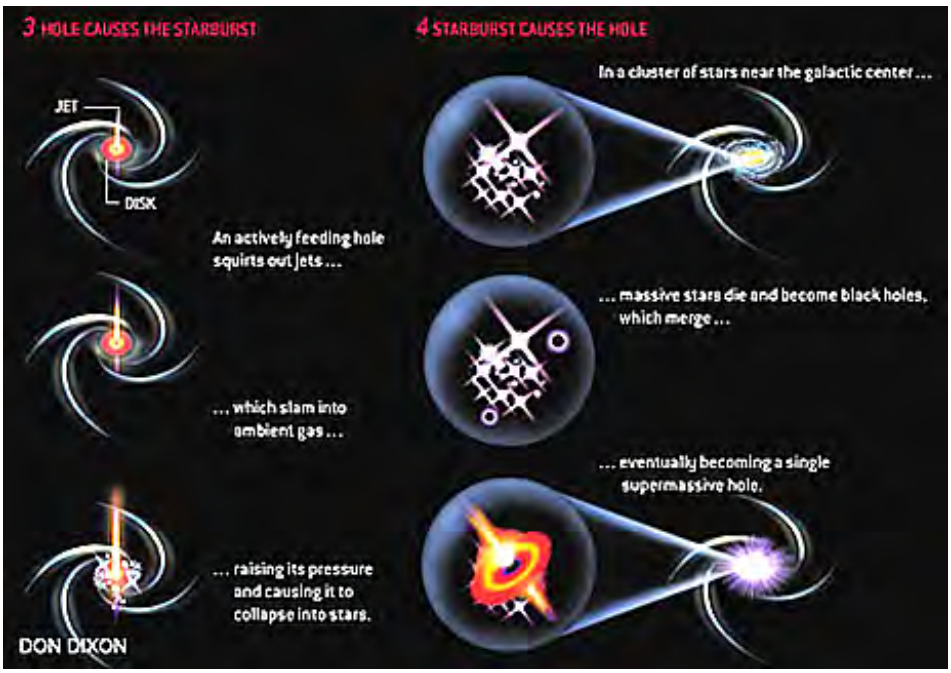
Now to the meat of Dr Weaver's paper.

While Black Holes come in all sizes, the larger ones (supermassive Black Holes; which nevertheless contain about 0.1% of the total galactic mass) are involved in both AGNs and starbursts. Furthermore, B.H.'s may also be responsible for star formation at slower rates (non-bursts). Suupermassive B.H.'s contain a compacted mass up to a trillion times that of the Sun but are only less than 1000 times its diameter. AGNs, as stated earlier, are superbright (although not notably in Visible wavelengths), often outshining the rest of its galaxy. AGNs are superrich in quasars which individually give off huge quantities of radiation. Many starbursts are as brilliant (in terms of energy given off) as the AGNs. A starburst period of activity is usually about 10 million years in length, during which the rate of star formation is up to 1000 times that of normal creative activity in a galaxy. One likely cause of a starburst is intermingling of gases when two galaxies collide (pass through each other), and the density of gases needed to make the stars increases. Both starbursts and AGN formation were much more prevalent in the first 1-2 billion years of Universe history. AGNs are distributed throughout cosmic time (they have been observed in galaxies closer to Earth, hence later in time), in contrast to quasars (next page) which are by far most frequent in the earlier Universe.

The question arises as to whether starbursts and major activity within an AGN occur simultaneously. Observations show that AGNs are often associated with young, bright, massive stars. However, it is often difficult to demonstrate the presence of starbursts within an AGN since the latter is marked by large amounts of obscuring dust (some of which is tied to the starburst process). The AGN can be "dormant" (no conspicuous visibility) until its gases begin to flow abundantly into the supermassive B.H. Starbursts commonly occur at an early stage of this process. Both starbursts and AGNs can be verified as active by examining their spectral - ionized Oxygen and other elements show brighter individual spectral lines.

Four possibilities for the interrelation between starbursts and AGNs - which came first and what causal connections exist between them and between B.H.'s - are cited by Dr. Weaver. We reproduce two schematic diagrams from her Scientific American paper that speak to this:





In the first scenario, AGNs are just manifestations (equivalent to) of starbursts and are possibly not set up by Black Holes. The second case postulates coincidence, i.e., no direct causal relation between starbursts and AGNs, even though both exist as separate phenomena. The third option considers starbursts to be caused by the supermassive B.H. pulling enough gas rapidly towards it to provide a denser environment in which many more than average star formation

events take place. Thus the starbursts result during the period in which gases also react around the Black Hole to form quasar-rich AGNs. The activity controlled by the B.H. that causes the AGN releases shock waves that aid in star burst formation beyond the inner zone. The fourth mechanism argues that starbursts precede the inception and growth of their associated supermassive Black Hole. As each massive star dies, it evolves into a Black Hole or a Neutron star. In time, these many B.H.'s coalesce to form the supermassive B.H. found at or near the center of the galaxy (this subject is treated on pages 20-5 and 20-6). A variant involves collisions of smaller stars that then speed up their lives before exploding into dense cold matter. Generally, this last mode first yields Black Holes of intermediate size (as has been observed in some galaxies) but usually the final outcome is a single large B.H.

But the bottom line in Dr. Weaver's paper is that supermassive Black Holes, AGNs, and Starbursts are a common, perhaps prevalent, set of components in spiral, and perhaps elliptical and also irregular galaxies.

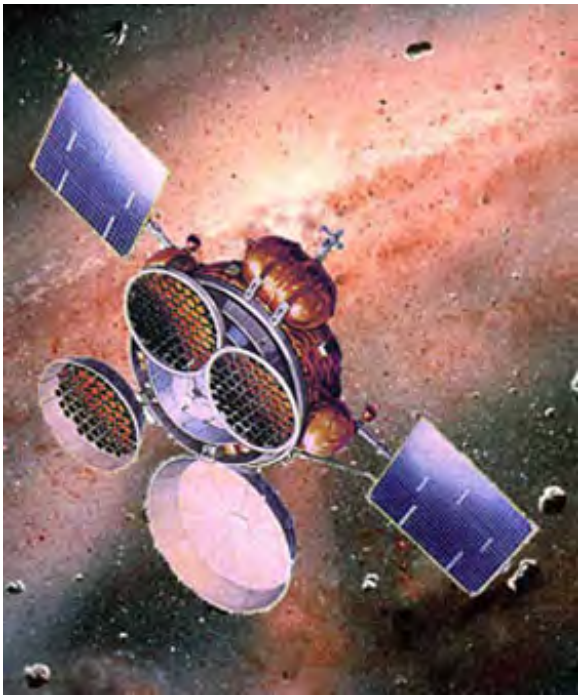
PREVIOUS RST NEXT

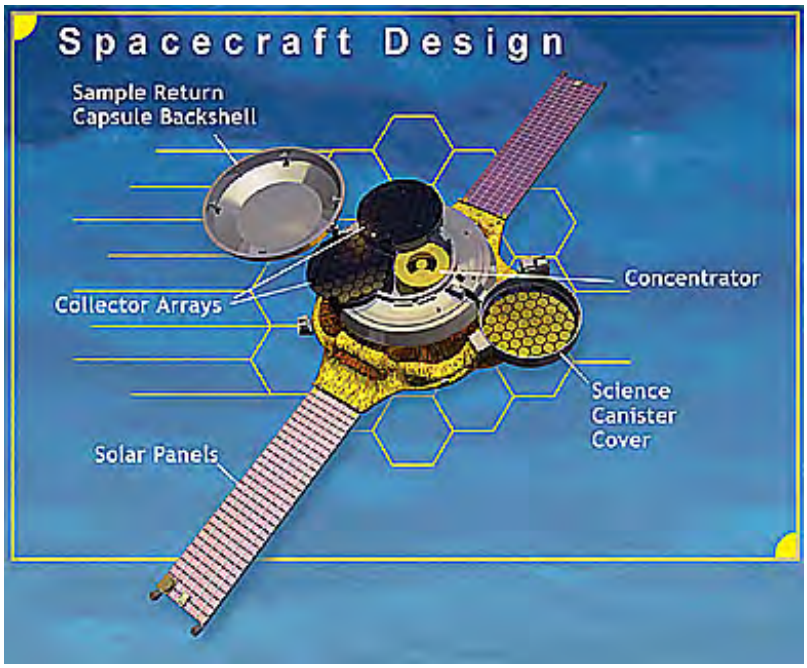
Primary Author: Nicholas M. Short, Sr.

The Genesis Mission

The Sun and its system of planets, asteroids, comets, etc. are closely related, in that the latter formed about contemporaneously with the parent star, both from dust and gas that had previously formed from various cosmic processes including supernova explosions. One prime goal of solar astronomy has always been to learn the details of the processes involved. The various compositions of objects beyond the Sun will depend in large part on the Sun itself. Determining the composition of the outer shells of the Sun can shed light on how the extrasolar bodies themselves formed. Sampling the actual Sun's surface is obviously impossible with current technology. Fortunately, evidence of that composition can be found in the compositional makeup of the solar wind, in particular the isotopic variations of such key elements as Oxygen, nitrogen, and carbon. To gain firsthand knowledge of these isotopes, NASA designed a mission that would collect **solar wind** samples (as atomic particles) and then return these to Earth for analysis. Genesis is an appropriate name for such a mission as it connotes "origin" with respect to the entire Solar System. A good review, by JPL, of the major goals of this mission is found at this JPL science rationale web site. Launch by rocket took place on August 8, 2001. The spacecraft was then "parked" in far outer space at one of the Lagrangian points (L1) in the Earth-Sun gravitational system. Its panels were exposed as shown in the second diagram below and allowed to face the solar wind on a fixed schedule.

The spacecraft as deployed at L1 is pictured in this artist's painting; below it is another rendition in which the major components are labeled:





One of the two collector arrays is pictured below. Individual collector hexagons are composed of various substances, including Gold/Platinum, Silicon, Germanium and Aluminum metals; diamond, and sapphire. Each is selected because of its ability to capture and store certain components of the solar wind. The thickness of a hexagon wafer differs in the two collectors - one way to distinguish any collector that might break loose or into pieces.

The two collector arrays face the solar wind. At the end of the mission the arrays are retracted so that the two nest together and are covered by the Science Canister Cover. The spacecraft is then further covered by the heat-resistant protective Sample Return Capsule Backshell. The spacecraft then is directed back to Earth, a journey of millions of kilometers. As it approaches Earth, the now sealed capsule separates from the solar panels and framework and is directed to enter Earth's environment at an angle that optimizes its chances of passing through the atmosphere intact and slowing down for the final phase of re-entry

The collector array are the first samples of materials from outer space to be returned to Earth since Apollo 17, some 32 years earlier. The final phase leading to retrieval involved a technique used many times for data-holding canisters from military satellites, namely catching the capsule while it is descending in the lower atmosphere using helicopters to snatch it up. The capsule is first slowed by atmospheric drag and as it moves into the troposphere, a drogue parachute is deployed to further decrease its descent speed, followed as it is slowed enough by a main parachute. Assuming the capsule is coming down on or near target, one of two chase helicopters (operated by Hollywood stunt pilots who are masters at this kind of aerial feat) uses a skyhook to snag the parachute and carry the capsule to a very slow, safe ground landing. The

capsule is then supposed to be rapidly retrieved and carried into the nearby "clean room" for removal of the collector arrays that are then flown to Johnson Space Center's facility for storage and examination of extraterrestrial material under ultraclean conditions.

That was the optimistic scenario around 11 AM MST on September 8, 2004 at the Dugway Proving Grounds, a huge federal (Army-managed) complex some 60 km (40 miles) west of Salt Lake City, Utah. The writer watched this event unfold live on a closed-circuit JPL network. The capsule was spotted and cheers went up in the control room. Then, the capsule started tumbling, as seen in this picture.

To the horror of onlookers, no parachute(s) appeared and the capsule plunged helpless at about 300 km/hr (195 mph) to a hard landing in the desert. The explosives slated to open the chute hatches had failed to detonate. The helicopters closed in and took photos like this of the impact point:

Personnel from one of the landed helicopters approached the still hot capsule and took pictures of it. As seen below, the capsule had, in effect, split open. Onsite inspection indicated contamination from desert materials had entered and deposited on the collector array assemblies. Pessimism was high in Mission Control and among scientists that the mission had become a total failure.

The capsule was carefully picked up (it did not fall apart) and brought as is to the nearby clean room at Dugway. This photo shows it to be still largely intact:

Following inspection, fragments of the collector hexagons were removed, as shown below.

Because of the differences in hexagon thickness the fragments can be matched with the proper array holder. There is some preliminary indications that some of the hexagons are still intact. The effect, if any, of the terrestrial contaminations is being assessed. As of this writing (January 4, 2007) there is guarded optimism that some, perhaps most, possibly all of the major measurements as originally intended can be run on individual hexagons. The hexagons have been extracted; they are variably contaminated with desert dust but not with water. A thin film covers most. Thus, the mission now appears not to be a total loss, and may even end up largely successful. But so far, no official report of any results has emerged. This page will be periodically updated with further information or you can check out JPL's Genesis site for the latest status.

PREVIOUS rst NEXT

Primary Author: Nicholas M. Short, Sr. Primary Author: Nicholas M. Short, Sr.

The Hubble Space Telescope (HST)

This Section does not attempt to be a complete textbook on Astronomy. Some topics covered in an Astronomy course are not included here. Before we start the more detailed review of Astronomy and Cosmology in the next 11+ pages, we will examine what most scientists acclaim as the greatest astronomical instrument ever built by Man - the Hubble Space Telescope, or HST. If you may be sceptical of this claim, just feast your eyes on these two images - typical of HST's extraordinary output:

This review of the HST may pique your curiosity about "telescopes" and how they work. The main function of a telescope is to gather photons from a source, concentrate them (focus) so as to improve detectability, and display them as discrete images or numerical data sets. If interest is aroused, try these two websites for a primer on optical telescopes: Wikipedia, and How Stuff Works.

Probably the most famous telescope of all time is at the Mt. Wilson Observatory, in the San Gabriel mountains north of Pasadena, California. The observatory was the brainchild of the astronomer George Hale, who reasoned that building it atop a high peak would avoid problems with the atmosphere and city lights. Work began in 1904 to build the observatory; its first telescope, the 60 inch Hale refractor, began observing in 1908. Here is the present-day view of the buildings including the main observatory at 5715 feet:

Mt. Wilson's most famous telescope - and commensurate with the Hubble Space Telescope in importance - is the 100 inch Hooker (named after its benefactor), begun in 1908 and operational by 1917. It uses a refracting mirror. This is the telescope that Edwin Hubble used to demonstrate the existence of galaxies beyond the Milky Way and to discover that the Universe is expanding. Here it the scope in its current setting:

Prior to 1990 all telescope observations of the heavens were confined to instruments on the ground. These have one obvious advantage: they can be visited by astronomers. But they have a major disadvantage: the atmosphere tends to interfere with the observations, causing distortions and diminution of the radiation signals; building telescopes on top of high mountains - the current practice - reduces that problem. But as populations grew and cities cast more light into the sky, this unwanted background effect caused further reduction in viewing efficiency. Still, ground telescopy remains a mainstay of astronomy. Click on these web sources for lists of Observatories grouped by country and largest ground telescopes. Among the top sites for observatories is Mauna Kea (more than 4200 meters [14000 ft] above sealevel) on the Big Island of Hawaii. Various nations have observatories there. Here is a view of the complex:

Prior to the 1990s, surveying and studying stars and galaxies as visible entities required the use of optical telescopes at ground-based locations. This ground photo shows a typical cluster of observatories, the Kitt Peak complex in Arizona.

Telescope observatories are distributed in mountain ranges across the world. One example is the ESO (European Southern Observatory) Paranal facility in the Atacama desert of Chile. The cluster consists of 18 telescopes, operated by 14 nations, and includes the VLT - four conjoined telescopes each with 8.2 meters mirrors.

In this Section we will see some stunning images acquired by space telesccopes. Images taken through ground telescopes are usually less spectacular but some can rival those acquired by space observatories. Examine below the pair of images (in the visible [left] and infrared [right]) of the Flame Nebula as seen through the newly operational VISTA telescope operated by the

European Space Organization (ES0) on a high mountain in the Chilean Andes:

Before setting out to explore the Universe's development and history since the first moments after the Big Bang, we want to pay homage to what this writer (NMS; and many, many others) consider the greatest scientific instrument yet devised by mankind - the first of the Great Observatories: the **Hubble Space Telescope** (which we will often refer to as **HST**). No other instrument has advanced our knowledge of astronomy and the Universe as much as this splendid observatory in outer space. Perhaps no other astronomical observatory has captured the public's imagination, with its numerous sensational pictures, as has the Hubble. HST has provided many extraordinary views of stars, galaxies, dust clouds, exploding stars, and interstellar and intergalactic space, extending our view to the outermost reaches of the Universe. HST has brought about a *revolution* in our understanding of Astronomy and Cosmology. One good reason for placing this HST review on this second page is simply that many of the subsequent illustrations of the Cosmos used in this Section were made by this telescope.

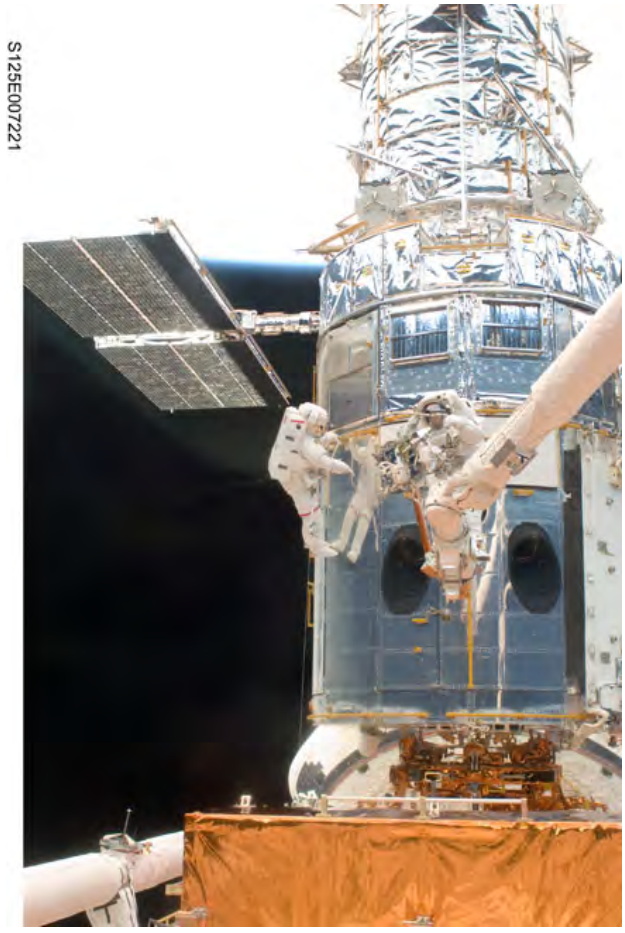
This, the most versatile optical telescope up to the present and perhaps the penultimate remote sensing system, receives its name to honor Edwin Hubble, the man who confirmed much about the existence, distribution, and movement of galaxies, leading to the realization of an expanding Universe which in turn brought about the Big Bang theory. Here he is at work in the 1920s on the 100-inch Mt. Wilson telescope:

As the space programs developed, astronomers dreamed of placing the telescopes in space orbit where viewing conditions are optimized. HST is the outgrowth of a concept first suggested in 1946 by Lyman Spitzer who argued that any telescope placed above Earth's atmosphere would produce significantly better imagery from outer space. (Spitzer has been honored for this idea by having the last of the Great Observatories named after him; see page 20-4.) The HST was launched from the Space Shuttle on April 24 of 1990 after 20 years of dedicated efforts by more than 10000 scientists and engineers to get this project funded and the spacecraft built. Here is the HTS in the Bay of the Shuttle:

HST Docked in Shuttle Discovery • STS-82 Servicing Mission Day 3 • NASA Photo

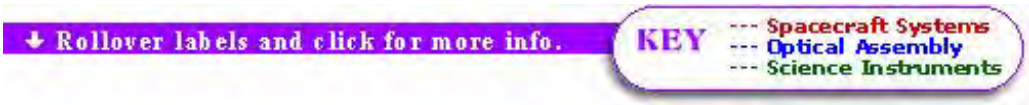
And this is a photo of the HST in orbit, as seen from the Shuttle:

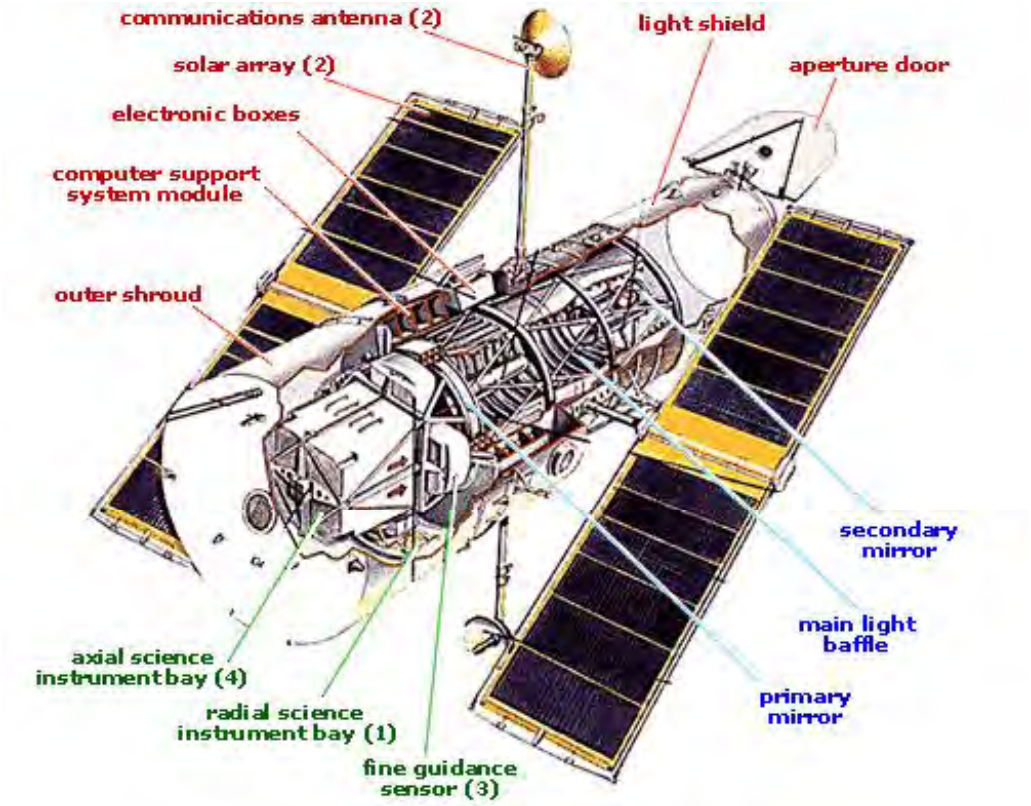
The HST is big - commonly described as the size of a school bus. An idea of the "bigness" - and why it just barely fitted into the Space Shuttle's cargo bay - is evident in this photo taken during the fourth repair mission, in which the astronauts provide a convenient scale.



A general description of the Hubble Space Telescope and its mission is given in this review by the Space Telescope Science Institute.

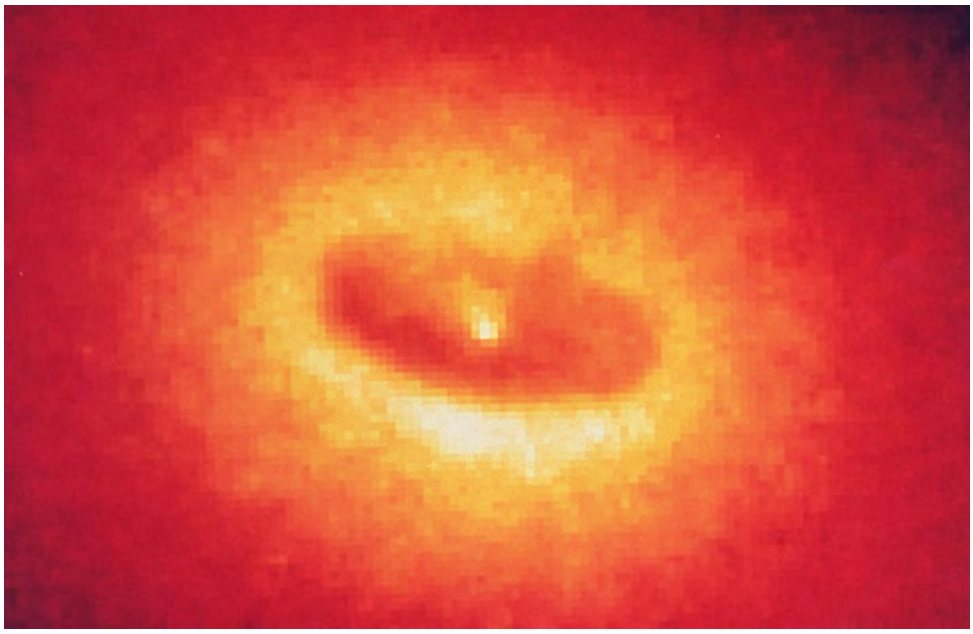
This cutaway diagram shows the major features and components of the HST:





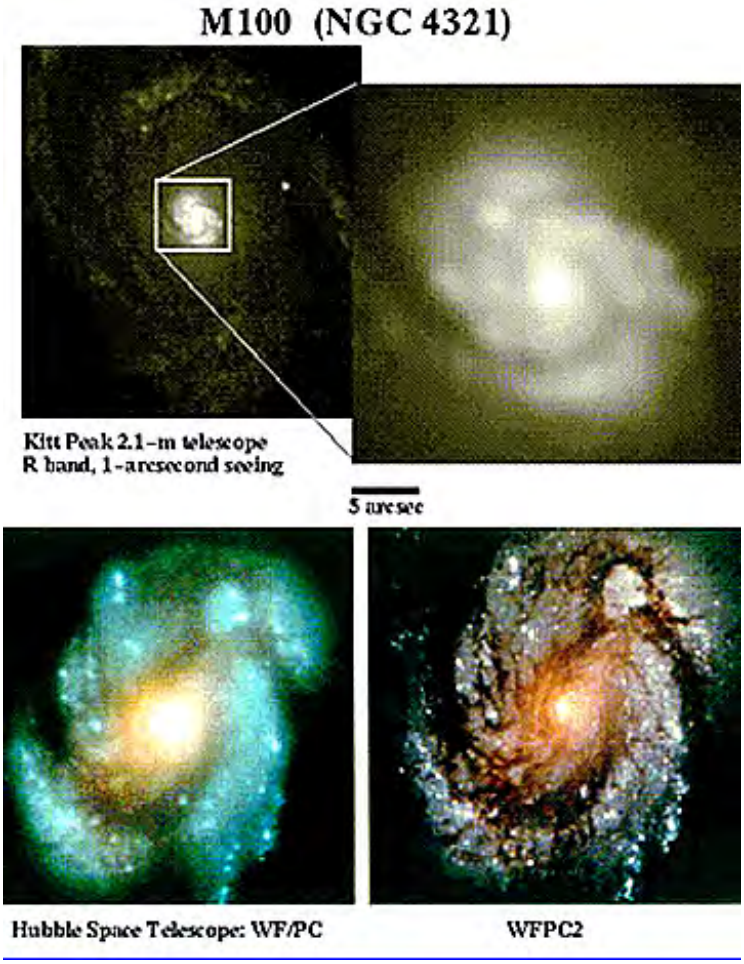
But, as scientists examined the first images they were dismayed to learn that these were both out of focus and lacked expected resolution. HST proved unable to deliver quite the sharp pictures expected because of a fundamental mistake in grinding the shape of its primary (2.4 m) mirror. The curvature was off by less than 100th of a millimeter but this error prevented focusing of light to yield sharp images. The distortion that resulted is evident in this early HST image of a star:

This dismaying result was a major blow to astronomers. NASA was urged by the scientific community to find a way to salvage HST. This took three and a half years to come up with and apply the solution. But during that time, HST did some limited but useful work. Consider this somewhat blurred image of NGC 4261. The small central bright spot was interpreted to be caused by the action of a black hole

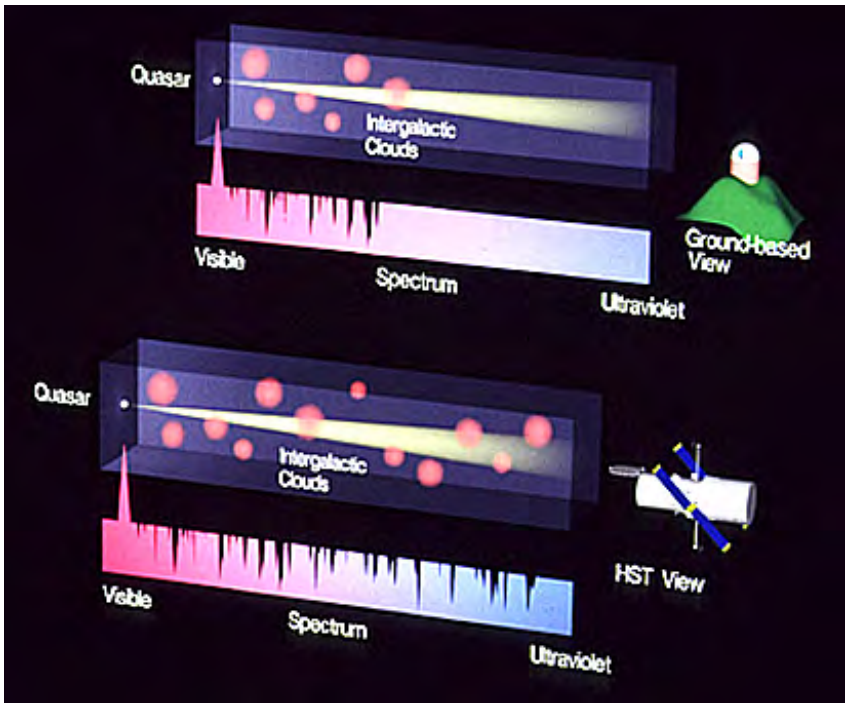


Astronomers and engineers put their heads together to solve this egregious problem and designed optical hardware that could restore a sharp focus. In December of 1993 the Hubble telescope was revisited by the Space Shuttle. (This mission to salvage the HST is a definitive answer to the critics of manned space missions - only human intervention could remedy an otherwise lost cause.) At that time 5 astronaut spacewalks succeeded in installing corrective mirrors and servicing other sensors. The package was known as COSTAR (Corrective Optics Space Telescope Axial Replacement).

After the first servicing mission, the striking improvement in optical and electronic response is evident in the set of images below made by the telescope, which show the famed M100 (M denotes the Messier Catalog number) spiral galaxy viewed by the Wide Field Planetary Camera before (bottom left) and after (bottom right) the correction. For an indication of how much HST improves astronomers' views of distant astronomical bodies, one of the best earth-based telescope images, from Kitt Peak, is shown at the top:



Another way to judge the improvement that HST provides by being above the atmosphere is to compare absorption spectra for Hydrogen in the Visible and Ultraviolet coming from a quasar source as recorded by a ground based telescope and HST.

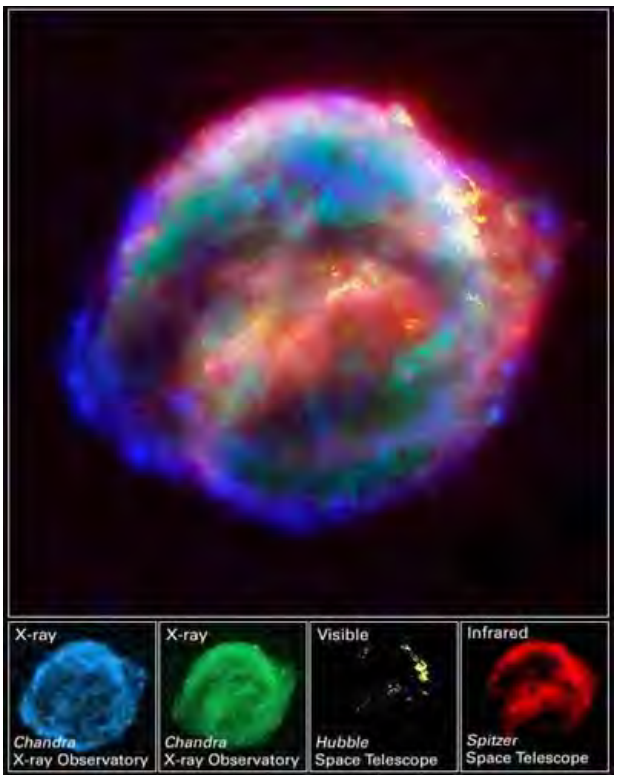


The increased sensitivity of the HST instrumentation, unimpeded by atmospheric absorption, provides more detected Hydrogen lines in both the UV and Visible regions of the EM spectrum.

In some respects, the HST shares remote sensing features found on Landsat. HST has filters that narrow the wavelengths sensed. The filters range through part of the UV, the Visible, and the Near-IR. This permits individual chemical elements to be detected at their diagnostic

wavelengths. The resulting narrow band images can then be combined through filters to produce the multicolored imagery that has made many Hubble scenes into almost an "abstract art" form - one of the reasons that the general public has taken so positively to this great instrument. As an example, here is an HST multifilter image of the Crab Nebula in which the blue is assigned to radiation from neutral hydrogen, the green relates to singly ionized oxygen, and the red doubly ionized oxygen.

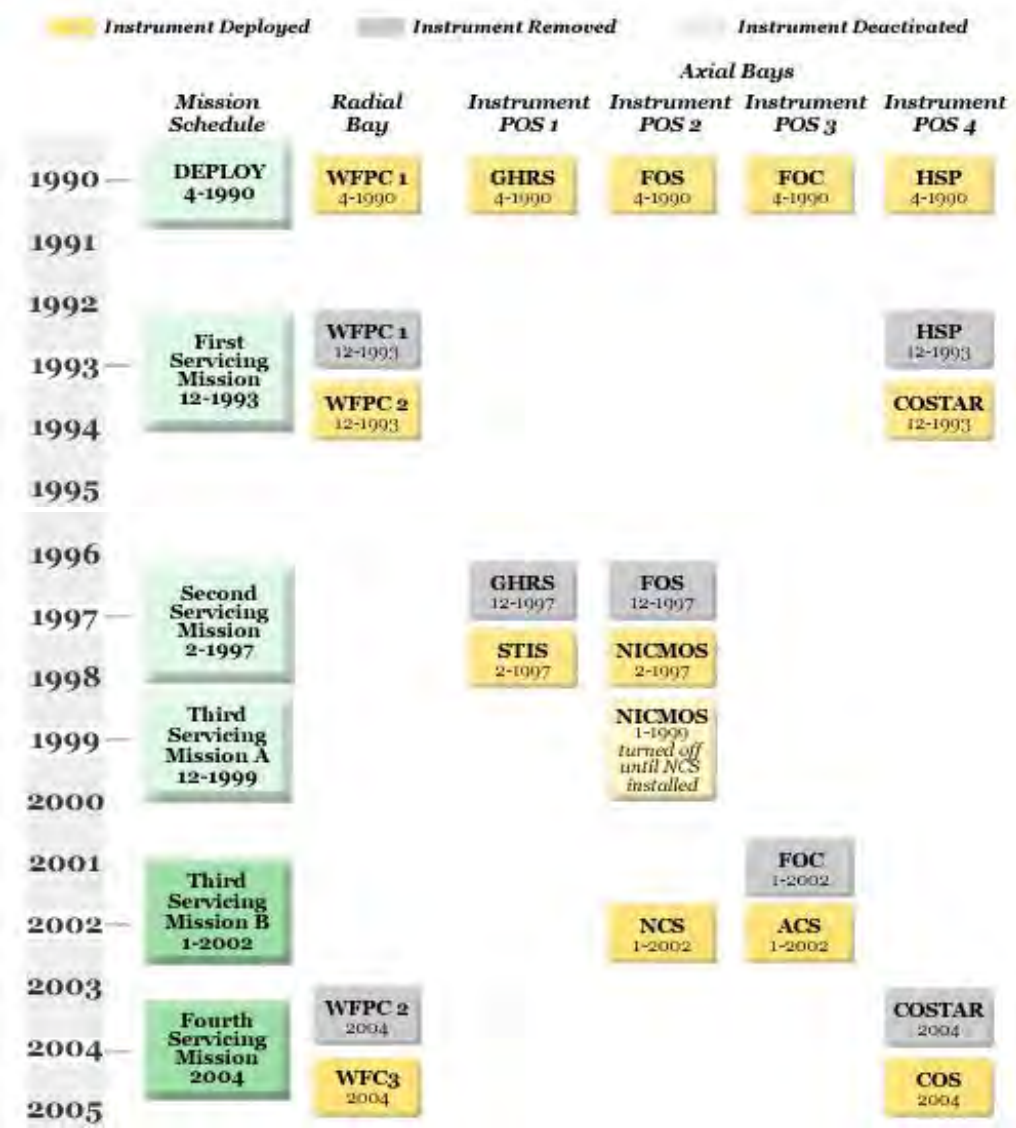
HST images can be combined with those made from other space observatories that sense at wavelengths outside the visible. This provides information on chemical composition as well as temperatures and the types of radiation involved. Consider this example:



These multiwavelength images give rise to one technique for picking out galaxies that are located at various great distances from Earth - the so-called Deep Field galaxies (page 20-3a) that formed early in the Universe's history. These galaxies are moving away at ever greater velocities. The redshift method (see page 20-10) of determining distance relies on the Doppler

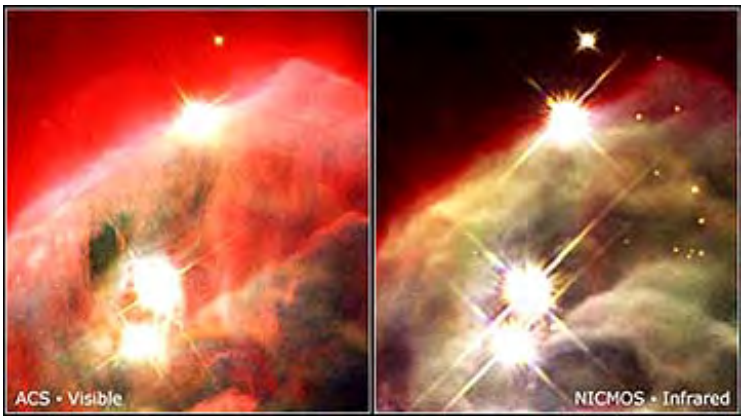
effect in which motion relative to the observer reduces the frequency (lengthens the wavelength towards/to/past red) of light radiation as the galaxies move away from Earth as a result of expansion. Those ever farther away, moving at progressively greater velocities, experience increasing redshifts. A galaxy emitting light at some maximum frequency can be imaged through, say, a narrow bandpass blue filter. This frequency translates to a specific redshift and hence a particular distance. A galaxy farther away has its redshift toward/to the green and will appear brightest through a green filter. Filters passing longer wavelengths will favor detection of greater redshifts - thus galaxies still more distant from Earth. Younger/closer galaxies may not even shine bright enough at shorter wavelengths to be detectable in filters whose bandpass cutoffs exclude those wavelengths.

Information on both original Hubble instruments and those added later appears in this site prepared by the Space Telescope Science Institute. The history of HST in terms of instrument placements and servicing missions, from the early days to the present and a look to the future is given in this chart prepared by the Space Telescope Institute:

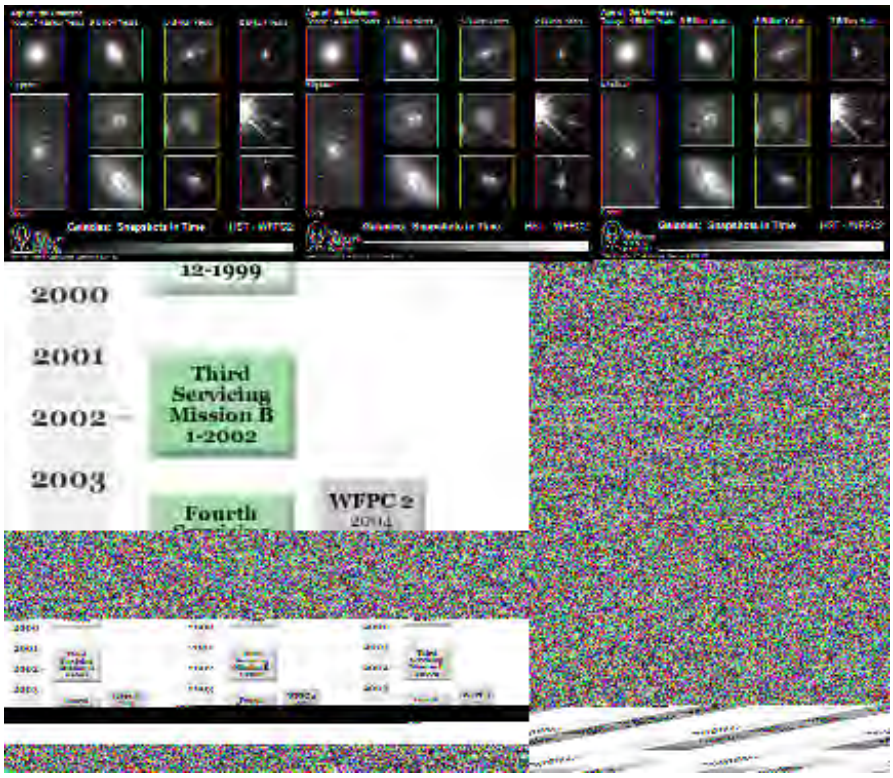


The original 5 instruments onboard HST were: the FOC (Faint Object Camera); FOS (Faint Object Spectrograph) GHRS (Goddard High Resolution Spectrograph); HSP (High Speed Photometer) and WFPC1 (Wide Field Planetary Camera); added since (by subsequent visits using the Space Shuttle) are NICMOS (Near Infrared Camera and MultiObject Spectrometer); STIS (Space Telescope Imaging Spectrograph); ACS (Advanced Camera Surveyor); FGS (Fine Guidance Sensor); and WFPC2; future additions (by Shuttle flights) may be the COS (Cosmic Origins Spectrograph) and WFPC3.

Thus, HST has been further improved even beyond its initial ten year life expectancy - now extended well into the second decade of the 21st century. A third Shuttle servicing mission was successfully completed in two stages: December 1999 and March of 2002. In addition to replacing or "repairing" existing systems on the satellite bus, a new instrument, the ACS (Advanced Camera for Surveys) was added; it represents a tenfold improvement in resolution and clarity. During this repair mission the NICMOS (Near Infrared Camera and Multi-Object Spectroscope) sensor, out of working order for nearly three years, was repaired and upgraded. This pair of images, ACS on the left and NICMOS on the right, shows the improved quality of imaging of part of the Cone Nebula, bringing out more details of the dust that dominates this feature:



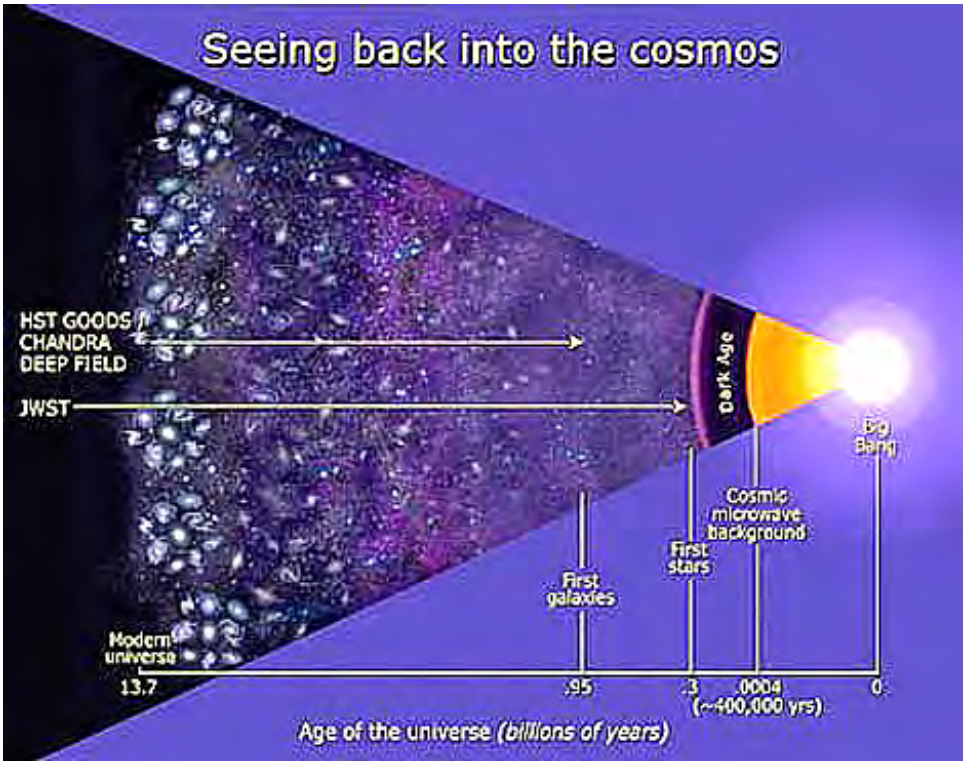
Many of the most informative HST images can be viewed on the Space Telescope Science Institute's (Baltimore, MD) Home Page . HST has imaged numerous galaxies at different distances - almost to the edge of space - from Earth that are therefore also at different time stages in the general evolution of the Universe. The following illustration shows both spiral and elliptical galaxies (but not the same individuals) at 2, 5, 9, and 14 billion years after the Big Bang in a sequence that represents different stages in this development.



The Hubble Space Telescope has had a remarkable impact on the study of the Universe. In its honor, the Astronomy Picture of the Day (APOD) web site, in celebration of its 10th anniversary, has compiled a collage of a variety of the more spectacular images acquired by HST, supplemented with a few images made by other instruments. This is reproduced here; be on the lookout for many of the individual embedded images in this montage elsewhere in this Section.

However, technology and design are allowing ground-based telescopes to "catch up" with the HST, at least for those galaxies that are relatively close to Earth. **The resolution and clarity of some recently activated ground telescopes are on a par with their Hubble counterparts**, at least within the depth range (lookback time) of a few billion light years. This results from better detectors, improved optics, the ability of a ground telescope to dwell on the target for much longer time spans (allowing buildup of the incoming radiation to generate a bright image), and, for some location on high mountain tops, above most of the atmosphere. This is illustrated with this pair of images which show a Highton Compact Group galaxy (HCG87) imaged by ESO's southern hemisphere telescope (left) and by the Hubble ST (right):

This diagram summarizes the current and anticipated status of space telescopes' ability to see back in time towards the earliest events following the Big Bang:

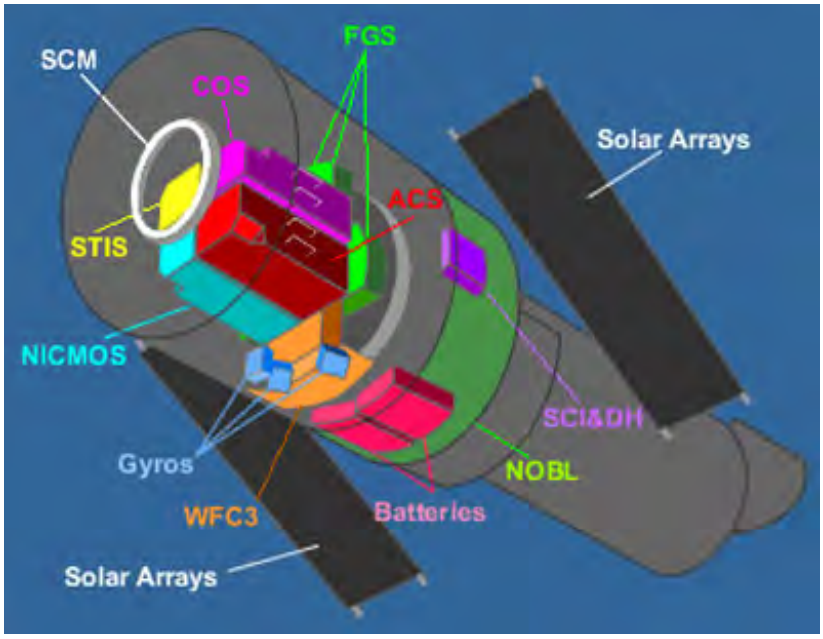


However, the Hubble Space Telescope remains the premier astronomical instrument - in many opinions, the finest instrument of any kind yet made - in the stable of space observatories. But, being complicated in its electronics and optics, like any fine instrument it has a finite lifetime. Being out in space, it is not easy to repair the HST whenever a major failure occurs.

Hubble has now been visited five times (1993; 1997; 1999; 2002; 2009) already for repair and upgrade. However, its components are now well beyond their planned lifetime and will likely fail in the next few years. Following the Columbia disaster, the perils of space travel for humans caused NASA to decide against another servicing mission that could be too dangerous at the higher altitude in which HST orbits. This raised a storm of protest and expressions of dismay from both the scientific community and an involved public. Sensitive to this outcry, the then NASA Administrator, Michael Griffin, ordered a "rethink" of that decision and on October 31, 2006 he announced that, with the resumption of the Shuttle program, the HST has been scheduled to be visited by astronaut-repairmen in the Spring of 2009 to rescue it from eventual failure.

The principal tasks (see below) were carried out in 5 (dangerous and challenging) EVAs. The ACS, which has had some periodic problems, apparently failed totally in January, 2007, owing to an electronic short-circuit; it is too large to be replaced. Other instruments and components have failed, or will soon, so that if not fixed or replaced, the HST would cease to function in the

foreseeable future. This diagram indicates the major modifications and replacements that were executed during the 2009 mission:



This paragraph provides more details about the changes, all of which were successfully made:

* Installation of three new rate sensing units, or RSUs, containing two gyroscopes each to restore full redundancy in the telescope's pointing control system

* Installation of six new nickel-hydrogen batteries to replace the power packs launched with Hubble in 1990

* Installation of the Wide Field Camera 3 (in place of the current Wide Field Planetary Camera 2), providing high-resolution optical coverage from the near-infrared region of the spectrum to the ultraviolet

* Installation of the Cosmic Origins Spectrograph, sensitive to ultraviolet wavelengths. COS will take the place of a no-longer-used instrument known as COSTAR that once was used to correct for the spherical aberration of Hubble's primary mirror. All current Hubble instruments are equipped with their own corrective optics

* Repair of the Advanced Camera for Surveys

* Repair of the Space Telescope Imaging Spectrograph

* Installation of a refurbished fine guidance sensor, one of three used to lock onto and track astronomical targets (two of Hubble's three sensors suffer degraded performance). The refurbished FGS, removed from Hubble during a 1999 servicing mission, will replace FGS-2R, which has a problem with an LED sensor in a star selector subsystem

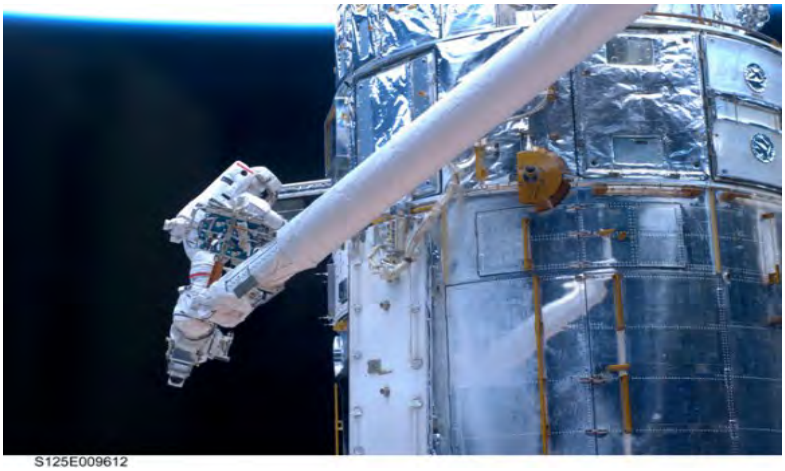
* Installation of the replacement science instrument command and data handling system computer

* Attachment of new outer blanket layer - NOBL - insulation to replace degrading panels

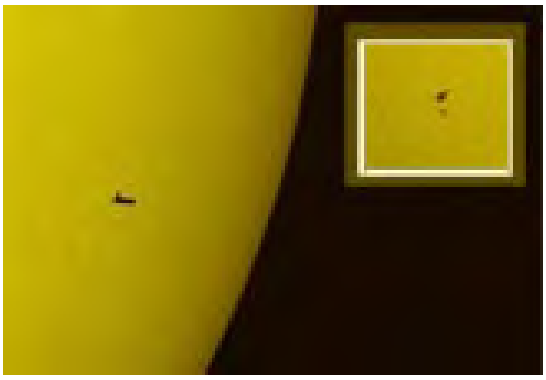
* Attachment of the soft capture mechanism to permit future attachment to a deorbit rocket motor or NASA's planned Orion capsule

The launch of the Space Shuttle Atlantis to begin the repair mission took place on May 12, 2009. The first space walk on May 14, shown below, successfully installed the new Wide Field Camera:

This photo, taken by one of the astronauts during the third spacewalk, shows astronaut Grunsfeld next to the Hubble.



On May 13th an amateur photographer captured the picture shown below, which made all the evening news broadcasts. It shows the Shuttle and the Hubble during the brief moment they passed in front of the Sun.



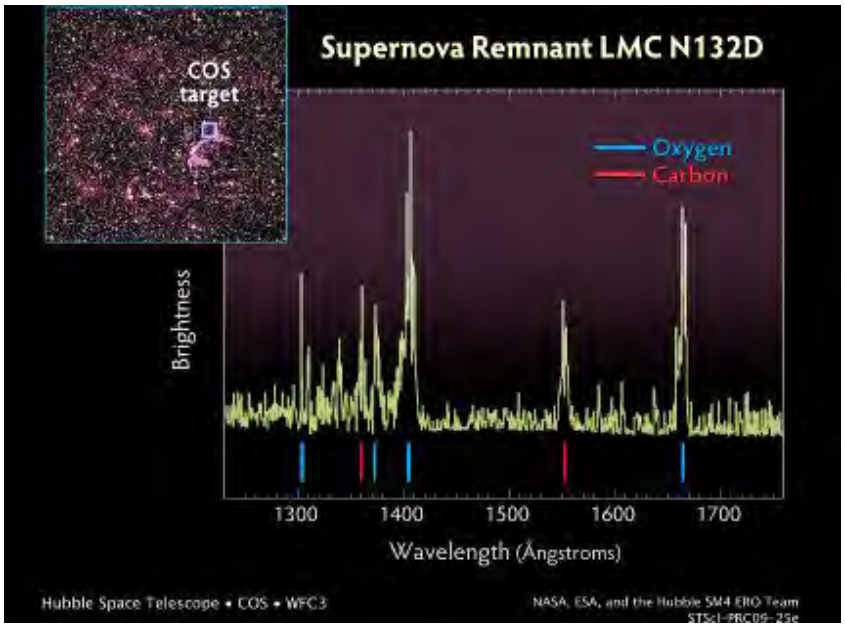
Baring the unforeseen, the HST should continue to operate in its improved state for years (with luck, into the 2020s).

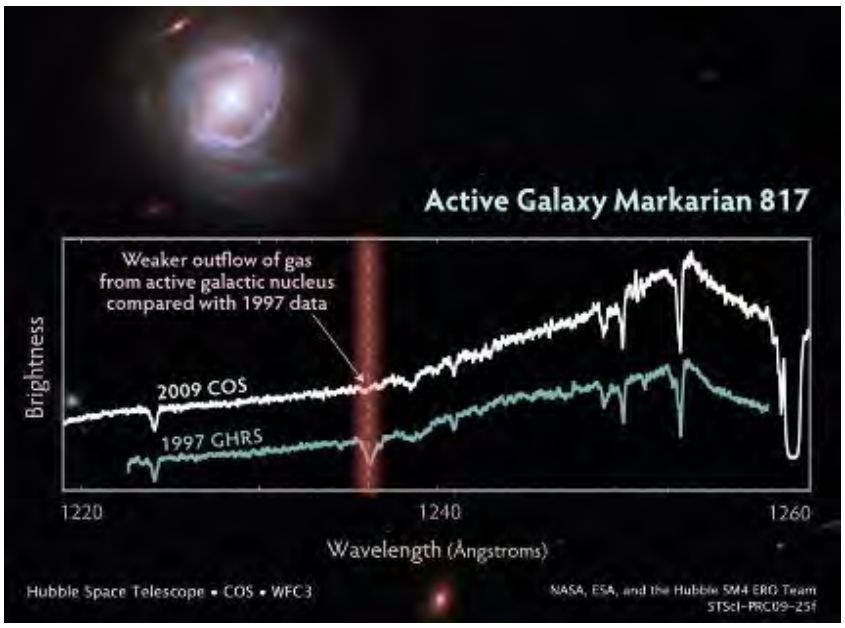
It took more than 3 months for the new equipment to "gas out" and otherwise be checked for functionality before routine observations could begin. On September 9, 2009 NASA released the first new images and spectral data from the refurbished HST. Here are some representative samples, some with commentary only in their captions:

The first pair of images show Stephan's Quintet, a group of colliding galaxies. The top image was taken with the old Wide Field Camera, the bottom with the new WFC3:

The nearby (3200 light years) planetary nebula NGC6302, called the Butterfly Nebula (there is another by that name, M2-9) is shown here first with part of it from the new WFC3 imagery on the right and a previous image made by the WFC2 on the left. Then, the full WFC3 image of this beautiful nebula is shown; compare that with the best ground image made by the European Southern Observatory telescope, shown beneath it:

HST is not just an imaging system. It also has instruments that can make other kinds of measurements. Here are two such products:





As we finish up this introduction to the HST, the writer would like to place an enlarged image of his favorite among all Hubble images seen to date: It is of the Carinae nebula, and shows one of its hydrogen gas and dust pillars from which stars are born. A few stars have already formed. More stars will eventually develop, destroying their pillars of creation over the next 100,000 years, and resulting in a new open cluster of stars. The pink dots around the image are newly formed stars. The technical name for the stellar jets are Herbig-Haro objects. How a star creates Herbig-Haro jets is an ongoing topic of research, but it likely involves an accretion disk swirling around a central star. A second impressive Herbig-Haro jet occurs diagonally near the image center.

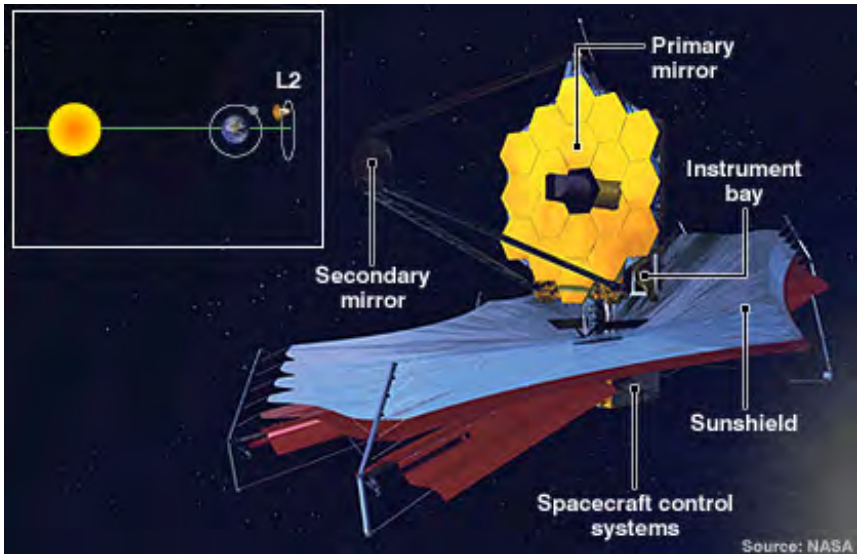
The inquisitive reader of this Section may well ask: Is this the real appearance of the nebula; does it have the blues that give it the visual flare one sees? The answer is that the assigned colors are somewhat arbitrary or selective, done to enhance the appearance. The ways in which colors are chosen is the subject of this Hubble website: The Meaning of Color

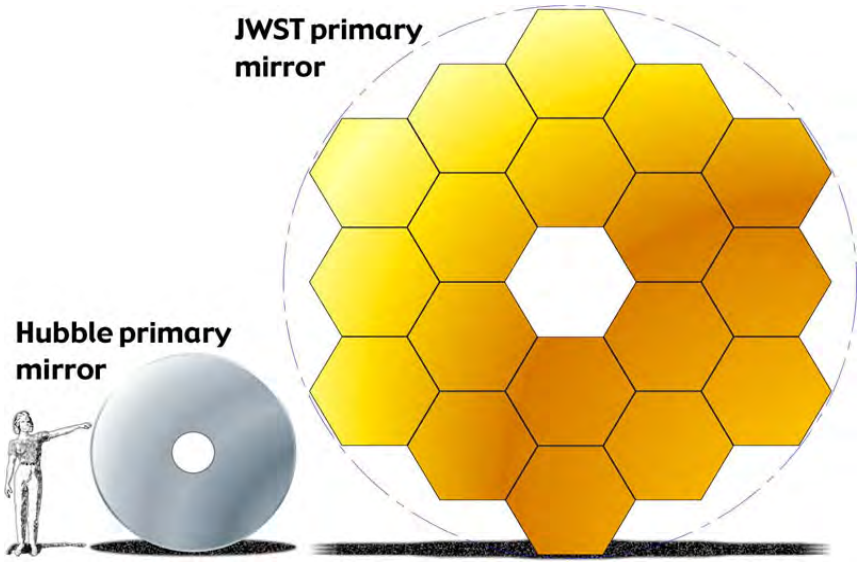
This image inspires this statement: May the best happen to this most supreme of scientific instruments!

A significant number of other space telescopes have been placed in orbit. Most have instruments that cover other parts of the EM spectrum beyond the visible. Among these we mention here: SWIFT (gamma-ray bursts) the Chandra Telescope (X-ray region), XMM-Newton (X-ray region), FUSE and Galex (UV), and the Spitzer Space Telescope (Infrared). Most of those telescopes operate in various parts of the spectrum as described on page 20-4. For a comprehensive listing of nearly all space telescopes launched or planned consult this Space Observatories website.

As scientists learn ever more about the Cosmos from existing ground and space telescopes, they are constantly advocating the need to have a more powerful and sophisticated telescope in space as the eventual HST replacement. NASA and the astronomical community always seem to have new telescopes on the drawing boards. The big follow-up being planned by The Space Telescope Institute and Goddard Space Flight Center is NGST which stands for the *Next Generation Space Telescope*. In 2002, this telescope was formally renamed the James Webb Space Telescope (JWST), to honor the second NASA Administrator for his many accomplishments in galvanizing the space program, including his role in the Moon program. A launch date has been set for no sooner than 2014. It will move far from Earth to "park" at a

Lagrangian point (about 1000000 miles away, where the Earth's and the Sun's gravitational forces balance out). A separate launch will place a big heat shield to block out the Sun's rays to keep the sensors at about 20° K above absolute zero. The JWST is an unusually shaped spacecraft, with a very large primary mirror, as evident from these three pictures:

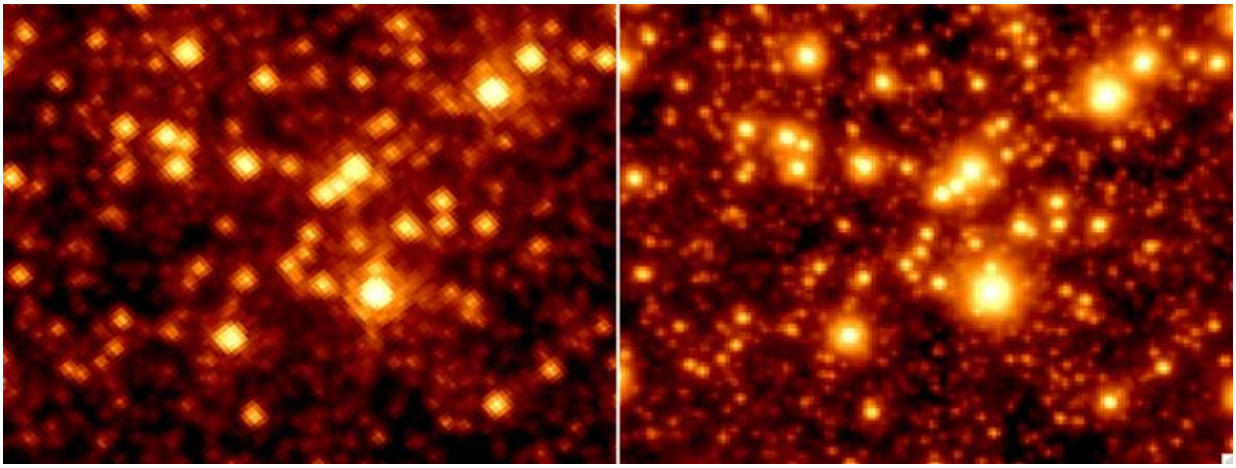




The HST has the detection capability and resolving power to look back to about the half billion years whereas the JWST will be able to detect and image events taking place about 300,000

years after the BB. Earlier than that will be difficult to examine by visual means because of the opacity of the Universe prior to that time. The principal scientific goal of JWST is to obtain improved information about the Universe's history between about 1 million and 2 billion years. The telescope main sensor will concentrate on the infrared region of the spectrum, with a range between 0.6 and 28 μm. Because of the spectral wavelength redshift that results from the expansion of space (see page 20-9), the visible light from these early moments in the Universe's history will have now, as received, extended into the infrared. (For further information, check out Goddard's JWST site.)

To reiterate, space telescopes have the advantage over most ground-based telescopes because they are above the distorting atmosphere. But those on the ground are constantly being improved, and more are being built on high mountains above most of the atmosphere. Several recent ones rival the HST. The Large Binocular Telescope in Arizona, which began operating in 2010, has captured images that are slightly sharper than HST can produce. This image pair shows the same region of outer space as imaged by HST on the left and LBT on the right:



These images of both ground and space-based views into the Cosmos are solid proof that Astronomy has entered a "Golden Age" in the last quarter century.

The Birth, Life and Death of Stars

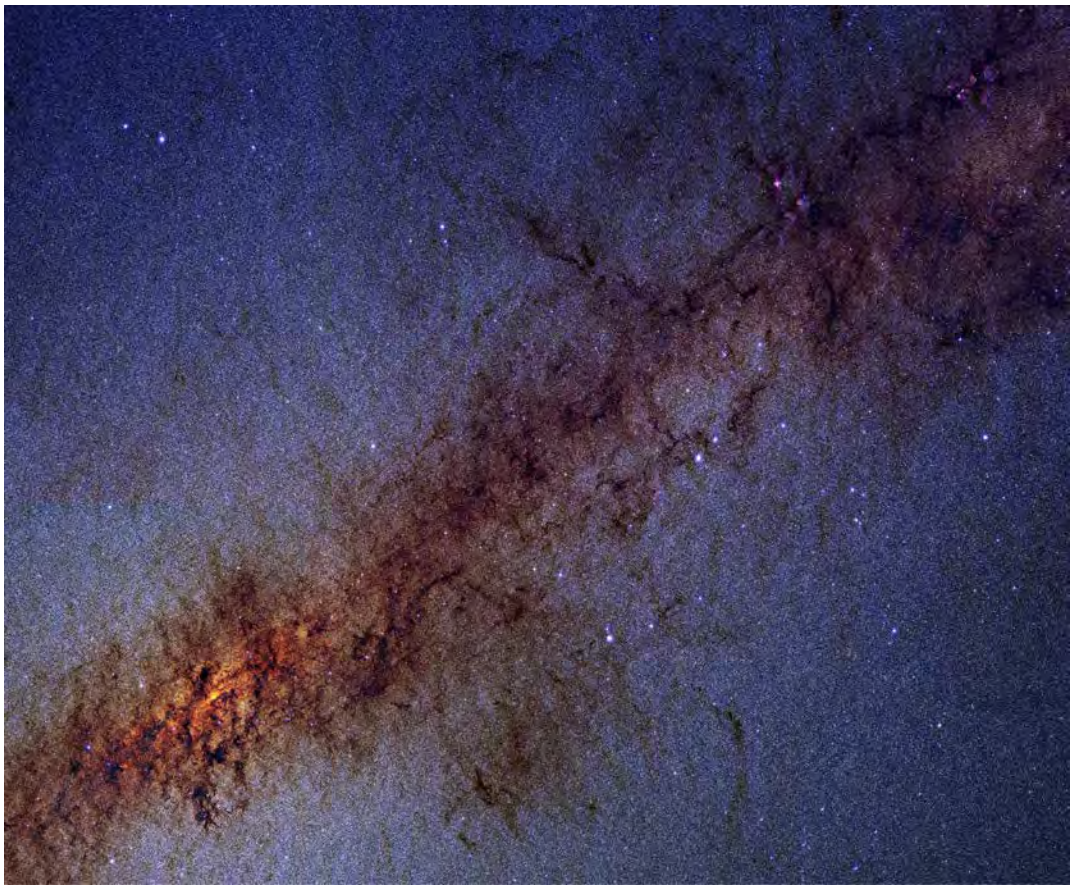
On the first page of this Section, we stated that stars are a *fundamental unit* making up the visible Universe. We repeat the definition of a star given in the Overview on that page: **A star is defined as a massive, spherical astronomical body that is undergoing (or has undergone) burning of nuclear fuels (initially Hydrogen; as it evolves elements of greater atomic number are both produced and consumed as well) so as to release energy in large amounts of both luminous and non-luminous radiation (over a wide range of the EM spectrum)**. The vast majority of stars are found within or close to collections of stars in vast assemblages called galaxies. Almost all stars used as illustrations on this and the next page lie within our own galaxy, the Milky Way, since those in other galaxies are too far away to be resolved (the majority of other images of features smaller than galaxies shown in this Section are also "local" in the sense that they are found within the Milky Way as well).

We concentrate on this page on the inception, evolution, and demise of individual stars. A helpful Web Site that supplements the content of this page has been put online by Prof. Nick Strobel of Bakersfield College (Note: his original online Astronomy Notes have disappeared off the Internet; this link is to a mirror site in Denmark) . Also recommended is the University of Oregon site cited in the Preface (especially relevant are Lectures 15-18, and 20). Another source of information is at the Wikepedia Star web site. Before reading the next two pages, it may be profitable for you to get an overview of Star Formation by reading relevant pages from the just-cited Oregon lectures.

Stars have fascinated humans since way back in prehistoric times. Stars are a favorite subject of artists. Here is a painting by Vincent Van Gogh, which is firmly embedded in the writer's (NMS) memory from childhood since my uncle had made a copy (art was his hobby) that hung in my grandmother's home.

Most stars, seen through telescopes, are rather bland and non-descript, appearing as uniform bright round objects. Some, when photographed, take on an added feature that make them appear more dramatic. These spokes or rays, seen in this image, are not real or part of the actual star; they are artifacts related to the mounting in reflection type telescopes produced by a diffraction process that is preserved in the photographic product:

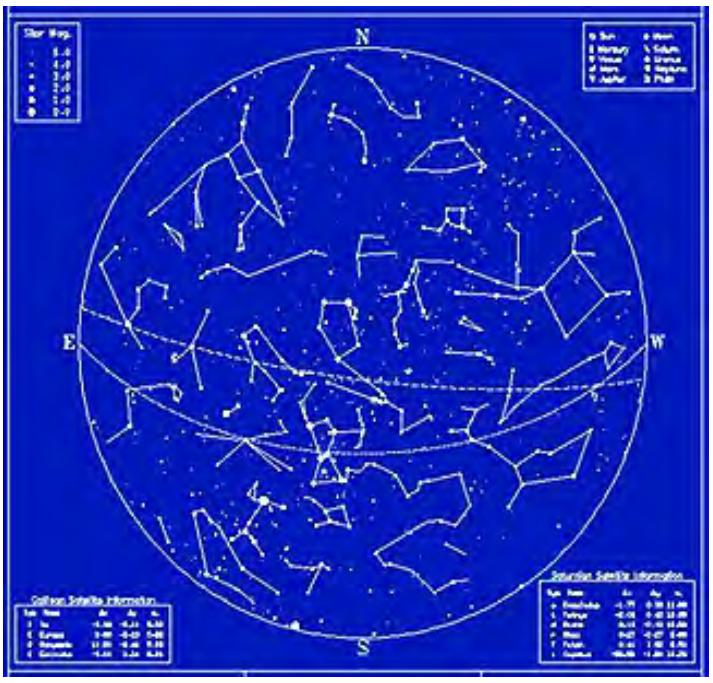
The number of stars in the Universe must be incredibly huge - a good guess is 100 billion galaxies each containing on average 100-200 billion stars or (10^{11} times 10^{11}, which calculates as 10^{22}). More recently, an independent estimate using other means states the value as 7 x 10^{22} stars; that survey made by Simon Driver of the Australian National University is said by him to still be a conservative underestimate. To make this point, here are two views of stars near us that indicates the densities just within a small part of the Milky Way - our own galaxy, the host of our parent star, the Sun.



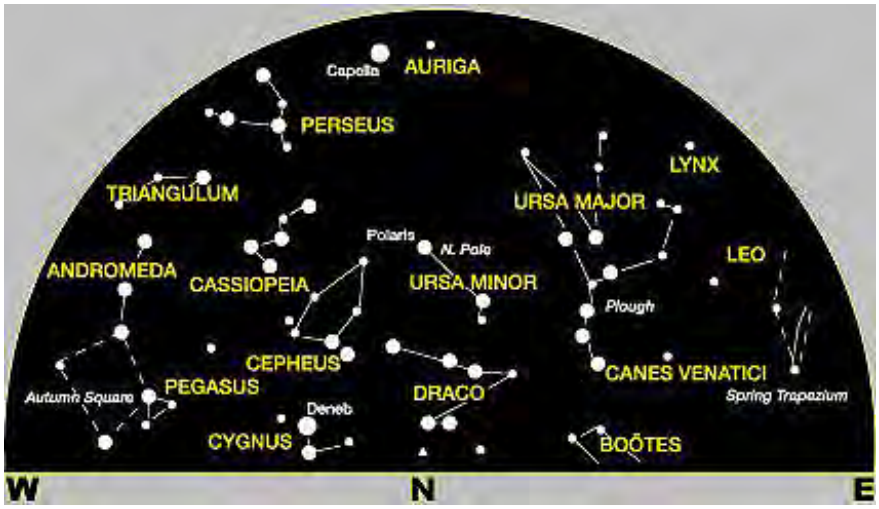
Yet on a very clear, moonless night in dry air (say, in the western Great Plains of the United States), one sees at any one location, without using a telescope or binoculars, only about 2500 "star" points within the Milky Way in the northern hemisphere and a similar number in the southern hemisphere (some of the points are nearby galaxies). Catalogs exist that list 9100 'naked eye visible' stars that have been located and inventoried as to position. Under normal (unaided eye) viewing conditions, outside the Milky Way band only a few individual stars can be seen by eye alone (less than 200 in a typical suburban [which cuts down detection because of city lights] setting with humid air; this number increases to more than a thousand in a rural area with an arid climate). (The best viewing experienced by the writer [NMS] was during a July night

on the plains of South Dakota.) Most of the observed stars are close to Earth (first few hundred light years) in the halo of the Milky Way, but several are planets, and a small number are nearby galaxies (the unaided eye can discern a galaxy only as an apparent single light source). The brightest stars are generally those closest to the Sun (around 1 to 100 light years away). Only when powerful telescopes are used does the astronomer realize by estimate or extrapolation that billions of galaxies exist; by inference we deduce that these probably contain stars in numbers similar to those that can be roughly counted in the Milky Way (in many tens of billions). The picture below recreates a typical view of part of the sky on a clear night:

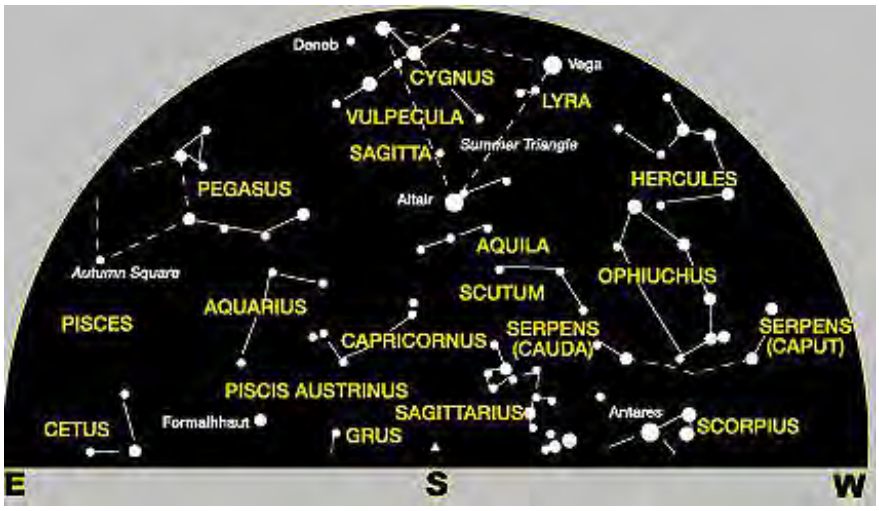
Wherever you happen to be on the Earth's globe, as you look up into the sky you will see a myriad of hundreds to a few thousand stars (depending on atmospheric conditions) that extend from directly above to the horizon in any, and all, directions. Some groupings of stars form distinct patterns that the ancients imagined were signs from the gods telling humans how to live their lives. These were given names and were called constellations (discussed later on this page). Here is a "star map" with some common constellations (left unnamed):



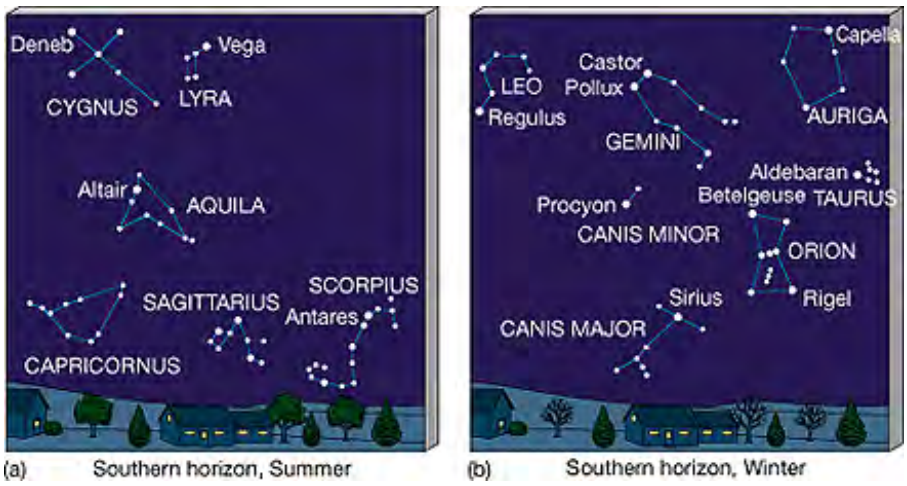
Let us talk a bit more about constellations. That in itself is not truly a scientific topic but which remains useful to astronomers as a convenient way to locate stars by reference to "Sky Maps". Such maps contain the *Constellations* - patterns of certain visible stars (a few were actually galaxies but this was not known at the time) that the ancients imaginatively discerned in looking at individual stars within narrow patches of the celestial hemisphere which seemed to be distinctive and readily recognized. These arrangements were given fanciful names, of gods, animals, and other descriptors from their everyday experiences. This began with the Babylonians in Mesopotamia, and the system was expanded to 88 named constellations by the Greeks and Romans some 2000 years ago. (Astrologers, Psychics and Fortune Tellers have used constellations as "signs" and for horoscopes for several millenia.) This next pair of illustrations shows some of the major constellations in the Northern hemisphere, plotted in two half circle fields, one for summertime, the other for winter; the look direction is towards the north:



The constellations visible from the Southern hemisphere are different:



In each Hemisphere the stars and constellations shift positions with the seasons. To illustrate this, we repeat here an illustration first included on page 19-2:



To the ancients, the stars were all equidistant on the Celestial Sphere and varied only in brightness. Modern astronomers now know that individual stars/galaxies in the light points that make up the constellation pattern are actually located at various distances from Earth, which together with diifferences in size account for their different brightnesses. Most of the defining stars in a constellation are located in our galaxy, the Milky Way.

As we shall see next, the groups of constellations also shift with the seasons and with the place on Earth where the stargazer is positioned (also, different constellations are seen by those in the Southern Hemisphere of the Earth than those in the Northern Hemisphere [lookers at the Equator will see some of the constellations visible in each Hemisphere]). More about the constellations can be found at the Star Map Internet site.

How do astronomers - or anyone - locate a particular star among the plethora that can be seen? You can imagine that the pinpoints of light, which can be stars, galaxies, planets, asteroids, comets, and large spacecraft, are located over a hemisphere of infinite depth that extends from your local horizon to the zenith point directly vertical to your location (latitude). Although you cannot see it, there is a complimentary hemisphere either to your south (if you are in the northern terrestrial hemisphere) or to your north (for southern hemisphere dwellers). These two hemispheres constitute the **Celestial Sphere**. On any given night, the various heavenly bodies will seem to travel across the sky in arcuate paths, owing to the rotation of the Earth. The Moon follows a definite path on any given night and the Sun goes from horizon to horizon on a path a few degrees displaced from the Moon; these too shift with seasons and location. Some insight into this sphere as it relates to a person on Earth is given by this Celestial Sphere plastic globe:

How does one locate any point of light on the Celestial Sphere? By consensus, the Earth's North and South Pole (through which pass its rotational axis) are made to coincide with the North and South Poles of the Celestial Sphere. In the northern hemisphere a star called Polaris - the North Star - coincides almost exactly with the extension of the Earth's rotational pole onto the Sphere.) Here is one way to determine the location of any light point (e.g., star) on this sphere:

From Unsold & Baschek, The New Cosmos, 4th Ed.

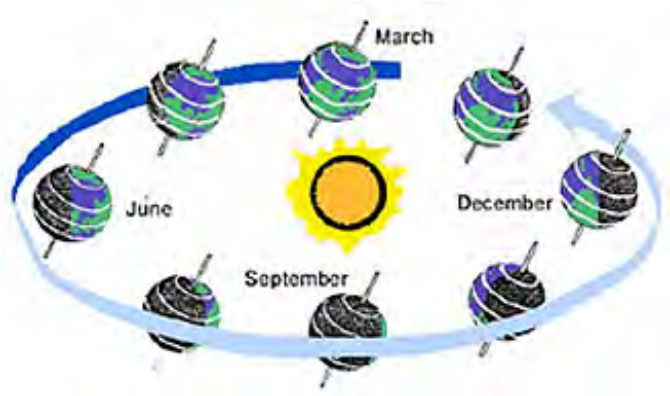
In this figure, the Celestial Sphere is held fixed (does not rotate). Here, the frame of coordinate reference is a great circle plane (equatorial) passing through the horizon as defined by true N(orth), S, E, and W. A second great circle contains the N & S directions and two points directly above the observer at any location on Earth - the zenith - and below - the nadir. At some given moment, the star's position on the Sphere can be specified by fixing its azimuth in degrees with respect to North or South and its altitude in degrees from 0 to 90°. (These are equivalent to longitude and latitude in the geographic coordinate system.)

If one were looking straight up overhead while standing on the North Pole, the relative locations of the constellation at that moment would constitute a reference Celestial (Hemi-)Sphere (same for South Pole but with different star patterns). Locating a given star would be done using the above azimuth-altitude spherical geometry. But globally the location is complicated by the three major types of motion involved in the real world: 1) on any given night the stars will move in arcs across the sky (centered on the Celestial Pole) in arcuate patterns; thus each star shifts position

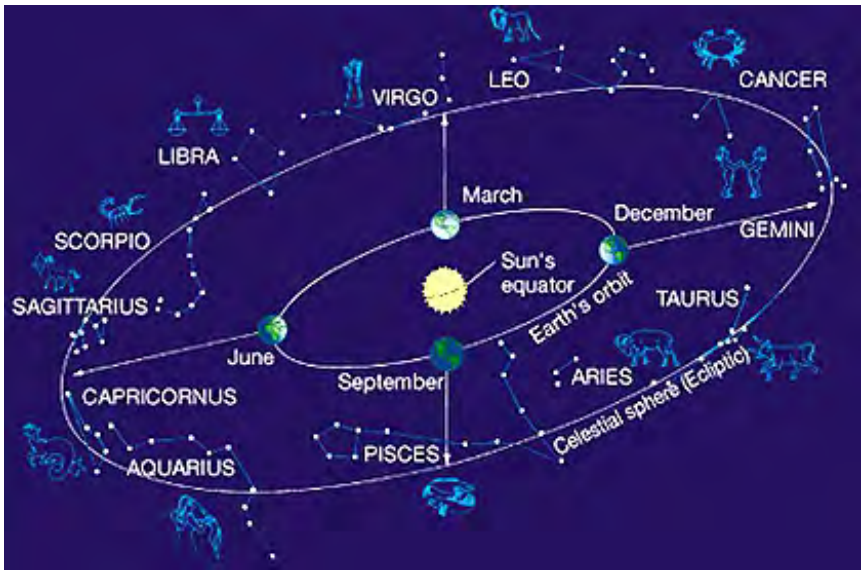
such that at different times it will lie at different points along its arc; 2) as the Earth revolves around the Sun (giving the different seasons), the positions of the stars will shift; 3) at any specific point on the Earth's sphere away from a Pole, the particular portion of the Celestial Sphere that is visible will differ (different hemispherical segments will be visible at different localities);

The effect of the Earth's rotation on the stars can be illustrated by a simple experiment: point a wide-angle lens camera up towards the hemisphere and keep its shutter open for hours. This would be the result:

Various constellations seemingly occupy different positions in the sky as the months and seasons pass during an earth year. Actually, the Earth's rotational axis retains its direction in space relative to the fixed stars (around the celestial poles) as it makes its annual revolution around the stars. There is an apparent shift of the constellations owing to the position of the Earth at each season:

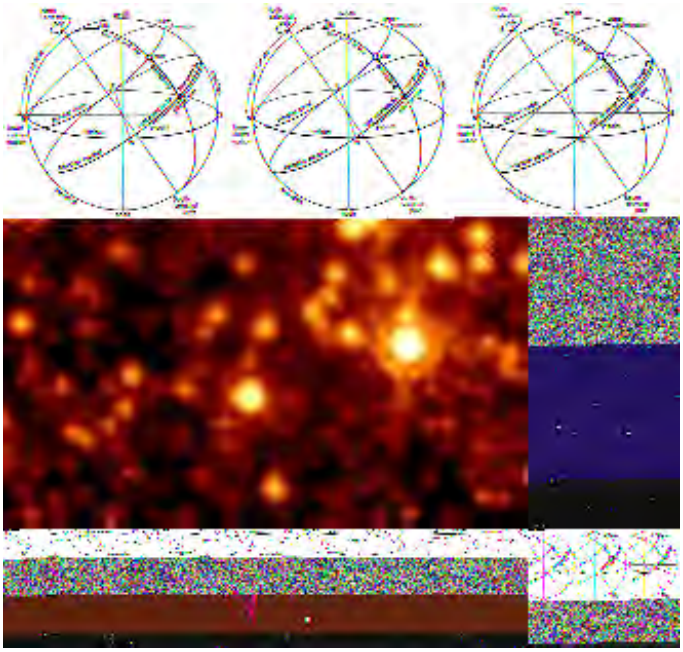


The ancients picked certain constellations to relate to the months they had named. These form the **Zodaic** - a word that has 'Zoos' as a root, to denote the animals perceived in most of the constellations chosen. For example, Cancer is actually a Latin word for 'crab'. Here are the 12 zodaical signs:



As the seasons progress, the Sun has different locations within the constellations on each seaons' starting date at the time of its daily zenith when crossing the ecliptic, as follows: Vernal Equinox(Spring) = Aries; Summer Solstice = Cancer; Autumnal Equinox = Libra; Winter Solstice = Capricorn.

Because the altitude and azimuth of a star are constantly changing in response to Earth's motion and to one' position on the terrestrial sphere, it is not always useful to rely on the above horizontal coordinate system to catalog the positions of stars. A more convenient coordinate system for cataloging purposes is one based on the celestial equator and the celestial poles and defined in a similar manner to latitude and longitude on the surface of the Earth. This location system takes the hours of observation into account. This means that one must consider both the specific time (usually relative to Greenwich Mean Time) when the observation is to be made (say, through a telescope in an Observatory) and the geographic location of the Observatory. Examine this diagram, in which the Celestial Pole is tilted 23° to accommodate the tilt of the Earth's rotational pole:



From Unsold & Baschek, The New Cosmos, 4th Ed.

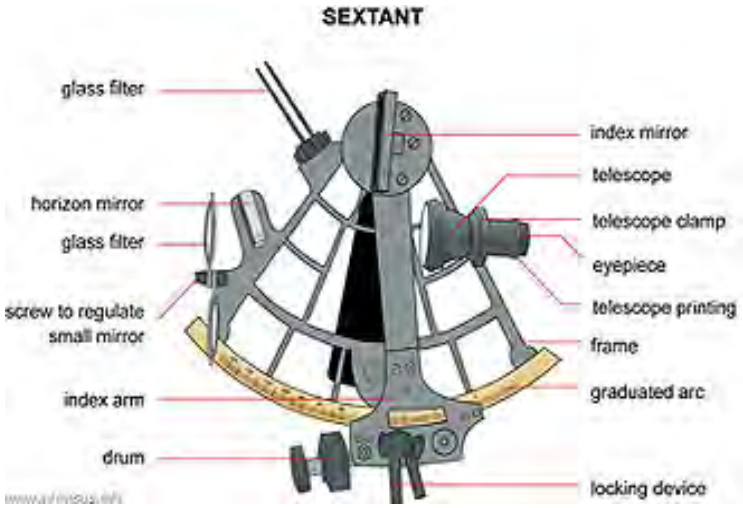
In this system, known as the Equatorial Coordinate system, the analog of latitude is the declination, δ. The declination of a star is its angular distance in degrees measured from the celestial equator along the meridian through the star. It is measured north (as +) and south (-) of the celestial equator and ranges from 0° at the celestial equator to 90° at the celestial poles,

being taken to be + when north of the celestial equator and - when south. The zero point chosen on the celestial sphere is the first point of the constellation Aries, γ, and the angle between it and the intersection of the meridian through a celestial object such as a star and the celestial equator is called the Right Ascension (RA) of the star. RA is sometimes denoted by the Greek letter α and is measured from 0h to 24h along the celestial equator eastwards from the first point of Aries, i.e., in the opposite direction to that in which hour angle is measured.

Because of the rotation of the Earth, a reference called the hour angle (HA) increases uniformly with time, going from 0° to 360° in 24 hours. Defining the observer's meridian as the arc of the great circle which passes from the north celestial pole through the zenith to the south celestial pole, the hour angle of a star - which changes with time of day - is measured from the observer's meridian westwards (for both northern and southern hemisphere observers) to the meridian through the star (from 0° to 360°). The hour angle of a particular object is therefore a measure of the time since it crossed the observer's meridian - hence the name. For this reason it is often measured in hours, minutes and seconds of time rather than in angular measure (just like longitude). The hour angle is referenced to the hour angle of the Constellation Aries at Vernal Equinox.

In practice, if one wishes to locate a particular star, galaxy, or planet at some moment, the main steps are to look up declination δ and RA in a star catalog or Ephemeris, read the time using a sidereal clock, and adjust the telescope settings to the hour angle τ and the δ value, taking into account location and time. (This time is usually *sidereal time*, exactly 24 hrs earth time. The so-called 24-hr solar day is not really precisely 24 hours because the Earth has moved along its orbital path a specific distance during a full rotation, so that it is 3 minutes and 56 seconds longer that the sidereal day,)

The relationship between the stars and Earth's time and space location proved a means of marine navigation even for the ancients. But precise determination of latitude and longitude requires two important capabilities: 1) fixing the location as declination of a star, and 2) establishing accurately the time relative to some place of reference (Greenwich). The development of the sextant (shown below) and of a Chronometer made this possible.



The above is only a part of the story about the celestial sphere and ways to find individual planets, stars, and galaxies. Additional treatment of how to navigate the sky at night is found on the aforementioned University of Oregon website. Perusal of this review is highly recommended.

Interesting, but back to Astronomical Science. The standard model for a star is, of course, our Sun. The Sun is typical of most stars; as we shall note shortly, these stellar bodies vary from about 0.1 to 100 times the mass contained in the Sun. Without a telescope, under exceptional viewing conditions (using binoculars), about 9000 individual stars can be seen in the wide celestial band that is the central disc of the Milky Way (M.W.) galaxy. Others elsewhere in the celestial hemisphere make up about 2000 points of stellar light that can be seen (in clear air,

away from urban light contamination) by the naked eye. Some are nearby within our galaxy and are not particularly large, while others are mostly stars of the Giant/Supergiant types in the halo (see below) around the Milky Way. Still others (a minority) are galaxies that lie in intergalactic space beyond the Milky Way but mostly within a billion light years from Earth. The solar planets are interspersed with these cosmic bodies. Telescopes can resolve countless more stars in the M.W., can recognize millions of galaxies, and can pick out some individual stars in nearby galaxies.

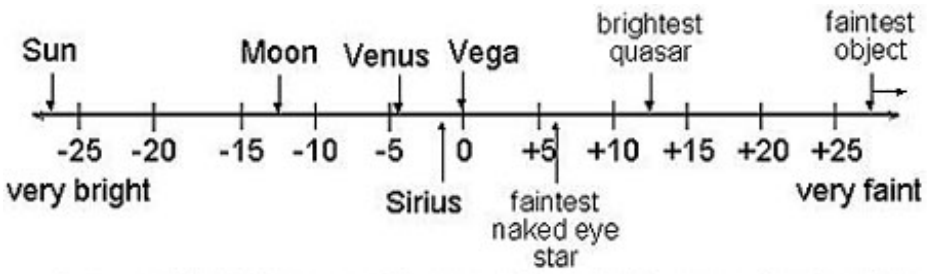
A degree of luminosity of an object in the sky (galaxy; star; glowing clouds; planet) can be represented by its *apparent magnitude* - a measure of how *bright* it actually appears as seen by the telescope or other measuring device. This magnitude is a function of 1) the intrinsic brightness which varies as a function of size, mass, and spectral type (related to star's surface temperature) and 2) its distance from Earth. (Magnitude as applied to a galaxy, which seldom shows many individual stars unless they are close [generally less than a billion years away], is an integrated value for the unresolved composite of glowing stars and gases within it.) The brightness of a star can be measured photometrically (at some arbitrary wavelength range) and assigned a luminosity L (radiant flux). For two stars (a and b) whose luminosities have been determined, this relationship holds:

$$L_a/L_b = (2.512)^{m_b - m_a}$$

from which can be derived:

$$m_b - m_a = 2.5 \log(L_a/L_b)$$

To establish a numerical scale, some reference star(s) must be assigned an arbitrary value. Initially, the star chosen, Polaris, was rated at +2.0 but when it was later found to be a variable star, others were selected to be the 0 reference value for m. The magnitude scale ranges from -m (very bright) to +m (increasingly faint) values. The very brightest objects have larger negative numbers. The more positive the number, the fainter is the object (planet; star; galaxy); very distant galaxies, even though these may be extremely luminous, could have large positive apparent magnitudes because of the $1/r^2$ decrease in brightness with increasing distance. The Sun has the value - 26.5; the full Moon is -12.5; Venus is -4.4; the naked eye can see stars brighter than + 7; Pluto has a magnitude of +15; Earth-based telescopes can pick out stars visually with magnitudes down to ~+ 20 (faintest) and with CCD integrators to about +28, and the HST to about +30. Thus, the trend in these values is from decreasing negativity to increasing positivity as the objects get ever less luminous as observed through a telescope. Each change in magnitude by 1 unit represents an increase/decrease in apparent brightness of 2.512; a jump of 3 units towards decreasing luminosity, say from magnitude +4 to +7, results in a $(2.512)^3 = 15.87$ decrease in brightness (the formula for this is derivable from the above equations, such that the ratio of luminosities is given by this expression: $10^{(0.4)(m_b - m_a)}$. Below is a simple linear graph that shows various astronomical objects plotted on the apparent magnitude scale:



Apparent brightnesses of some objects in the magnitude system.
From *Nick Strobel's Astronomy Notes.*

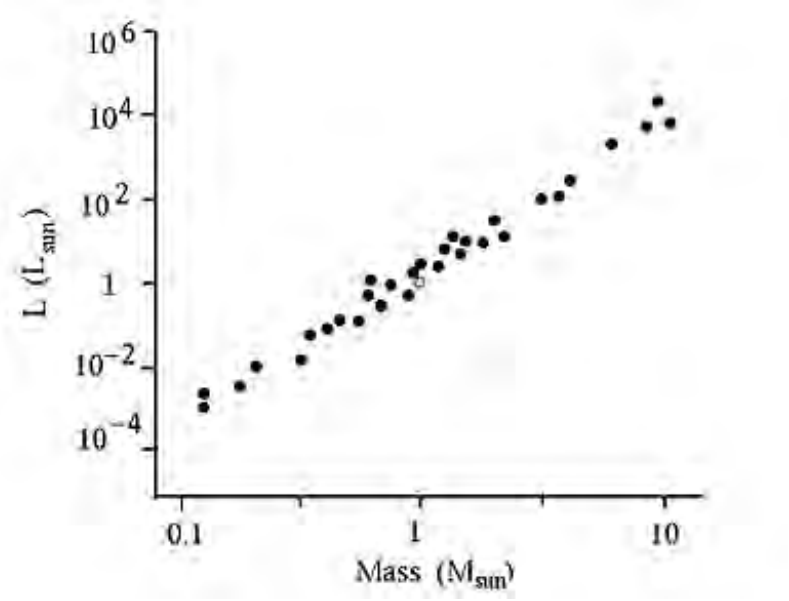
Absolute magnitude (M) is the apparent magnitude (m) a star would have if it were relocated to a standard distance from Earth. Apparent magnitude can be converted to absolute magnitude

by calculating what the star's or galaxy's luminosity would appear to be if it were conceived as being moved to a reference distance of 10 parsecs (10 x 3.26 light years) from Earth. The formula for this is:

$$M = m + 5 - 5\log_{10}r,$$

where r is the actual distance (in parsecs; 1 pc = 3.26 light years = 206,265 A.U. = 3.086 x 10^{16} meters) of the star from Earth. Both positive and negative values for M are possible. The procedure envisions all stars of varying intrinsic brightnesses and at varying distances from Earth throughout the Cosmos as having been arbitrarily relocated at a single common distance away from Earth.

Both luminosity and magnitude are related to a star's mass (which is best determined by applying Newton's Laws of motion to binary stars [a pair; see below for a discussion of binaries]). The graph below, made from astrometric data in which mass is determined by gravitational effects, expresses this relationship; in the plot both mass and luminosity are referenced to the Sun (note that the numbers are plotted in logarithmic units on both axes):



There is a relationship between absolute magnitude (here given by L for luminosity) and mass (given by the conventional letter M; which accounts for replacing the absolute magnitude M with L). Here is one expression:

$$\frac{L}{L_{Sun}} = \left(\frac{M}{M_{Sun}}\right)^4 \quad \text{for } M > 0.43M_{Sun}$$

$$\frac{L}{L_{Sun}} = \left(\frac{M}{M_{Sun}}\right)^{2.3} \quad \text{for } M < 0.43M_{Sun}$$

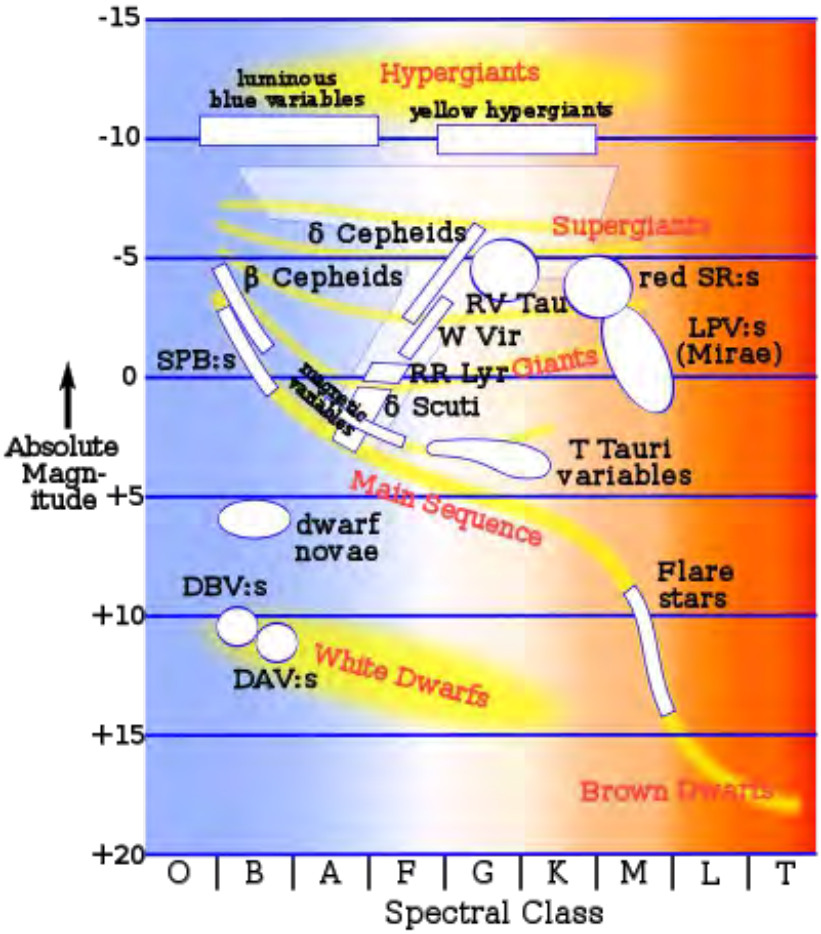
In the above, both L and M for a given star are ratioed to the values determined for the Sun. Note the two different power exponents. It seems that some stars obey a fourth power, others a 3, and a few are just the square of the mass. The most general expression in use is given as $L = M^{3.5}$. There are relatively few stars with mass greater 50 times the Sun. Very rarely, we can find a star approaching 100s solar mass, but these are so short-lived that nearly all created before the last million years have exploded, with their mass being highly dispersed, and thus ceasing to send detectable radiation.

If the Sun were envisioned as displaced outward to a distance of 32.6 l.y., its apparent magnitude as seen from Earth would be -26.5; its absolute magnitude would be changed to +4.85. A quasar, which is commonly brighter than a galaxy, has an absolute brightness of - 27

(note that in the absolute scale increasingly negative values denote increasing intrinsic brightness).

Measurements of location in the celestial sphere, distance from Earth, and magnitudes of stars can be done through ground-based telescopes or from space observatories. One space telescope, Hipparcos, was dedicated exclusively to obtain precise measurements of these parameters. Launched in August, 1989 by ESA, it operated through March of 1993. High precision data were obtained for more than 118,000 stars and less precise results for another 1,000,000+.

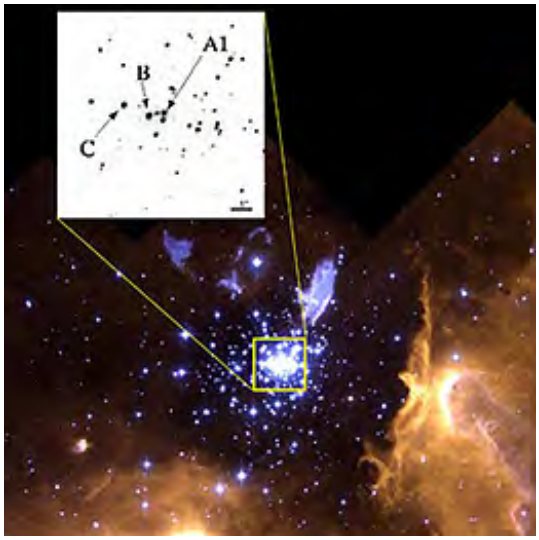
One classification of stars is that of setting up categories of star types in a series of decreasing sizes and luminosities (see also the discussion below of the Hertzsprung-Russell [H-R] diagram). These are the Luminosity/Type Classes: Ia, Ib: **Extreme Supergiants (Hypergiants)**; II: **Supergiants** (Betelgeuse); III: **Giants** (Antares); IV: **Subgiants**; V: **Dwarfs** (Sun): VI: **Subdwarfs** (metal poor); VII: **White Dwarfs** (burned out stars); VIII: **Brown Dwarfs**. Thus in this list luminosity decreases from left to right (the mass and size of a star, which determines its luminosity, also decreases left to right). The oddity in this classification is the omission of a category of "Normal"; a star is either a Giant or a Dwarf. The diagram below relates this hierarchy to star brightnesses (magnitudes) which decrease from top to bottom; the spectral types (indicated by letters) range from high luminosity/mass (left) to low luminosity/mass (right) but these letters do not relate to the Classes as plotted (the diagram is idiosyncratic in that the top Class - Hypergiant - is an O spectral type and the bottom Class - Brown Dwarf - is a T spectral type:



Another classification is based on mass. Starting with the least dense and progressing to the most dense, this is the sequence: **Supergiant**; **Red Giant**; **Main Sequence** ; **Red Dwarf**; **Brown Dwarf**; **White Dwarf**; **Neutron Star**; and **Black Hole**. This is also ranked according to (decreasing) size (diameter). The sequence just shown covers a mass range of about 1000. Few stars are larger than 100 times the mass of the Sun; few likewise are smaller than 1/10th of a solar mass. Very large stars burn up their atomic fuel rapidly and are not around more than a

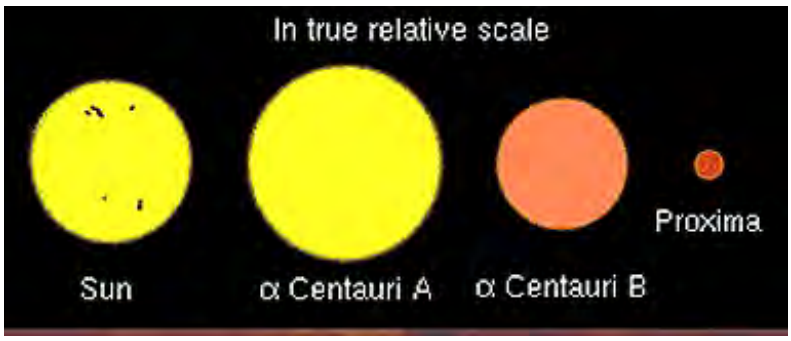
million years or so. Very small masses do not get hot enough to begin the Hydrogen fusion process that initiates a star's life.

Supergiants are dimensionally wide, and also luminous. NGC 3603 is the largest star in the Milky Way, being 116 times the mass of the Sun:

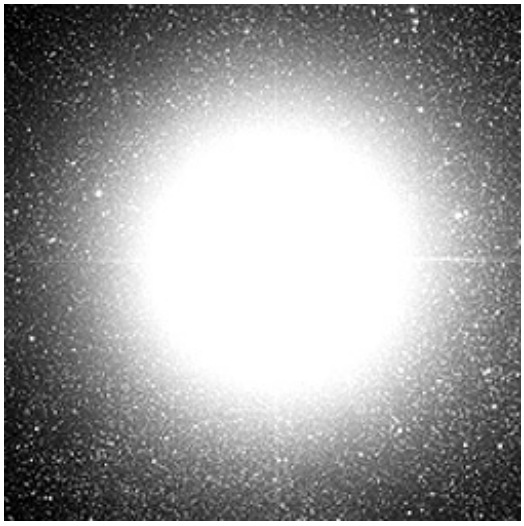


The brightest star in the northern hemisphere of the sky is Sirius, an A type star (see the H-R plots below and accompanying paragraphs which explain the letter designation of stars) of apparent magnitude -1.47 that lies 8.7 light years away. Here is how it appears through a telescope:

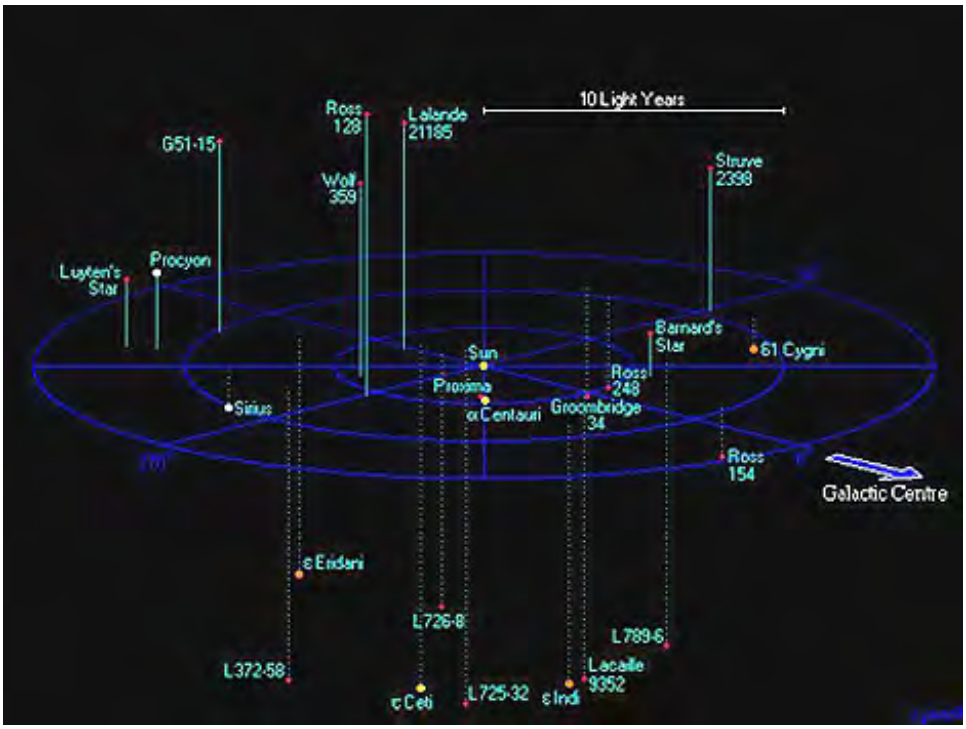
Closest to the Sun is the red dwarf Proxima Centauri, being 4.2 light years away. Just slightly farther away (4.4 l.y.) are Alpha Centauri A and B (visible in southern hemisphere), stars similar to the Sun that are among the brightest in the heavens.



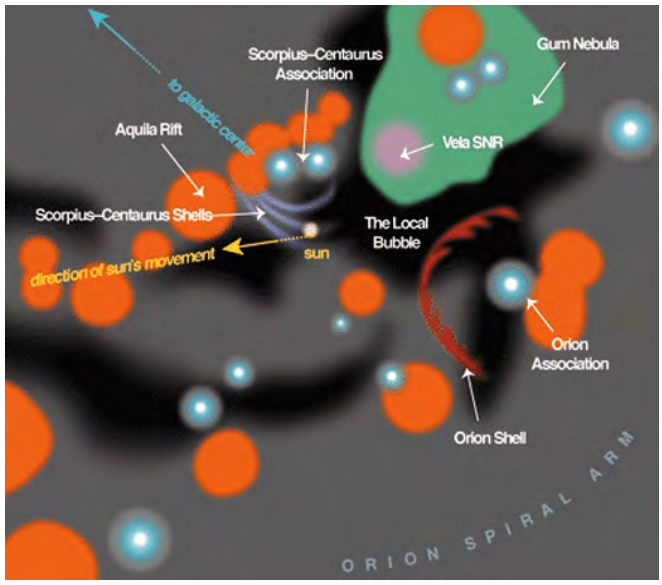
Here is a telescope view of Alpha Centauri A:



The map below is a plot of the distances from Earth (outer circle is 13.1 light years in radius) of the 25 nearest individual or binary stars or local clusters in our region of the Milky Way Galaxy. Many of these stars are red dwarfs (see next page):



(Information Bonus: Just beyond this map's edge is the star Vega (27 light years away). It has two claims to fame: 1) it alternates with Polaris as the North Star used in navigation; the Earth's precession brings Vega into this position every 11000 years, and 2) It was the nearby star used as the host for an extraterrestrial civiliation in Carl Sagan's extraordinary science fiction novel "Contact" (later made into the movie of the same name "starring" Jodie Foster; in the film contact was made with a planet near Vega as a signal picked up by the Socorro, NM radio telescope array - as initially interpreted that signal consisted of a string of prime numbers [those divisible only by themselves and 1]).



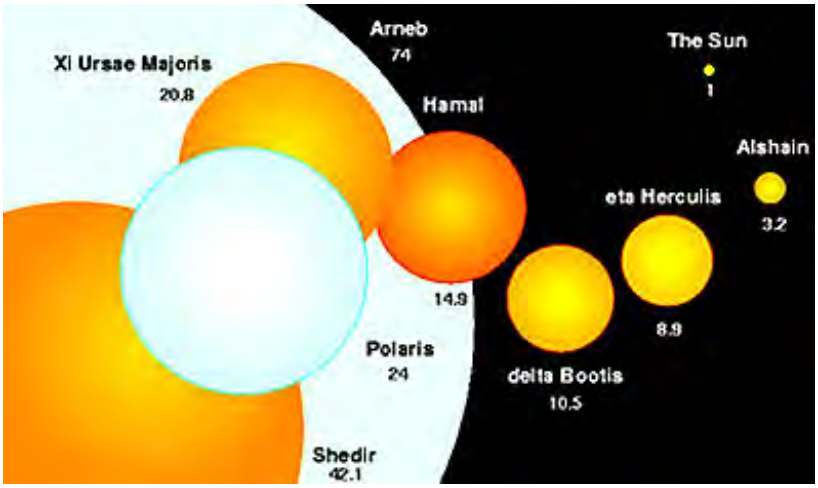
Lets look farther out into our galactic neighborhood. Referring to the above diagram, the following is extracted *verbatim* from the caption accompanying this image that was displayed on the Astronomy Picture of the Day Website for February 17, 2002: "What surrounds the Sun in this neck of the Milky Way Galaxy? Our current best guess is depicted in the above map of the surrounding 1500 light years constructed from various observations and deductions. Currently, the Sun is passing through a Local Interstellar Cloud (LIC), shown in violet, which is flowing away from the Scorpius-Centaurus Association of young stars. The LIC resides in a low-density hole in the interstellar medium (ISM) called the Local Bubble, shown in black. Nearby, high-density molecular clouds including the Aquila Rift surround star forming regions, each shown in orange. The Gum Nebula, shown in green, is a region of hot ionized Hydrogen gas. Inside the Gum Nebula is the Vela Supernova Remnant, shown in pink, which is expanding to create fragmented shells of material like the LIC. Future observations should help astronomers discern more about the local Galactic Neighborhood and how it might have affected Earth's past climate."

The largest star so far measured in the Milky Way is Mu Cephi (in the galactic cloud IC1396), seen as the orange disc (also called Herschel's Garnet star) near top center of this HST image. Located about 1800 light years from Earth, it is almost 2500 times the diameter of the Sun.

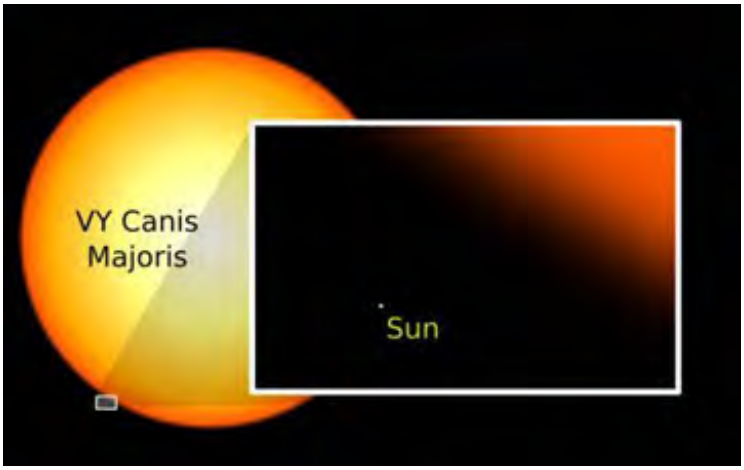
This is an example of a rare type of star known as a Hypergiant (see next page). Another even bigger star (2800 times the solar diameter; 2.4 billion miles) is Epsilon Aurigae (in the constellation Auriga, the Charioteer), residing in the Milky Way about 3300 light years from Earth. This star, also known as Al Maaz (Arabic for he-goat) and visible to the naked eye) is considered by many astronomers to be the "strangest" star in the firmament. Every 27 years this star (magnitude 3.2) undergoes a diminishing of brightness (about 60000 times greater than the Sun) lasting about 2 years. The last such event was in 1983; the next in 2010. It is thus one of a class called "eclipsing stars". The cause of this regular pattern of luminosity change is still uncertain;

some astronomers think it is caused by the passage of a second massive star across Epsilon Aurigae's face but that binary is so far undetected, leading to the hypothesis that the drop in luminosity occurs when a cloud of dark material (dust) orbiting the star as a clump obscures Epsilon Aurigae each time it moves through the line of sight to the Earth.

Most stars bigger than the Sun are not as huge as Mu Cephi or Epsilon Aurigae. The majority are no larger than about 100x the diameter of the Sun. This diagram illustrates the relative size of some common stars (setting the Sun's diameter as 1), which establishes our star as rather ordinary in the size scheme within the Milky Way:



One of the largest stars whose size can be accurately determined is VY Canis Majoris, in the Milky Way about 5000 l.y. from Earth. Best estimate of its diameter is about 2100 times that of the Sun.



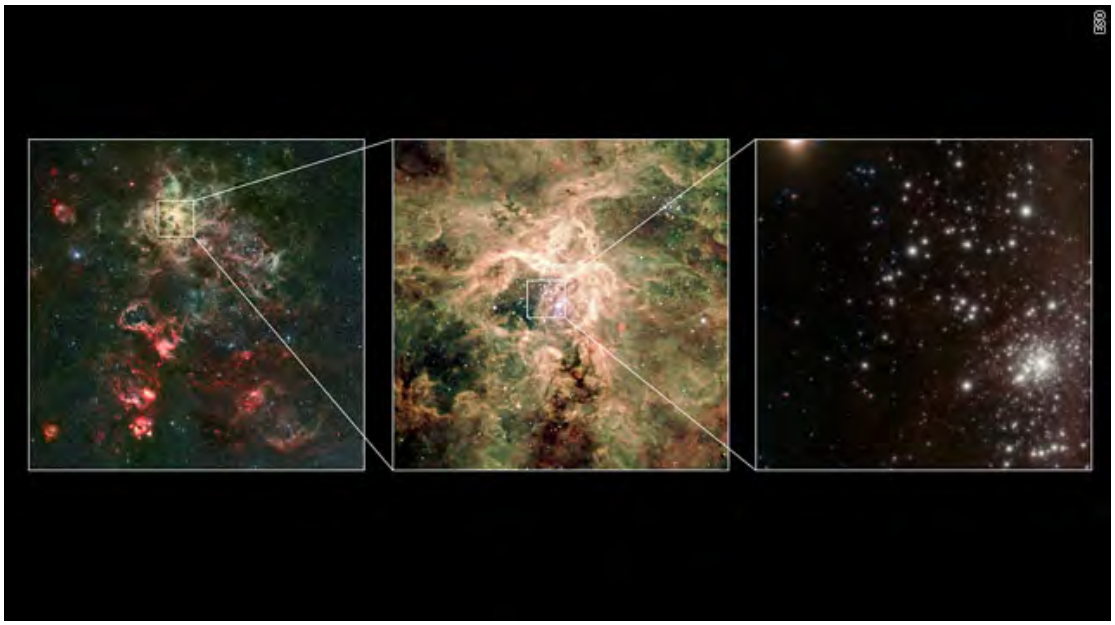
VY Canis Majoris is a Hypergiant star. Here is the best telescope view of this massive stellar body.

This star is a superb example of how one view within one segment of the spectrum gives a specific impression of an apparently simple visualization but is misleading in that a different spectral region discloses a much different appearance. VY Canis Majoris is actually surrounding by a much larger cloud of gas of varying composition, as evident in this pair of HST images.

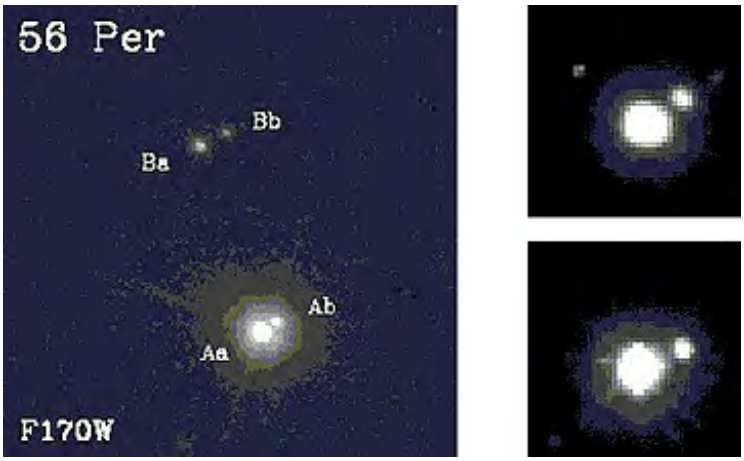
What is going on is that the star, because of its huge size, is destined to be short-lived and is already unstable, throwing off much of its mass. It is likely to be destroyed in less than another 100000 years as a supernova.

Another appropriate way to distinguish stars by size is to rank them according to mass. In the Milky Way, the Arches cluster contains the most massive single star (about 130 solar masses) found yet in the M.W.; at the time theorists thought this is a reasonable upper limit throughout the Universe:

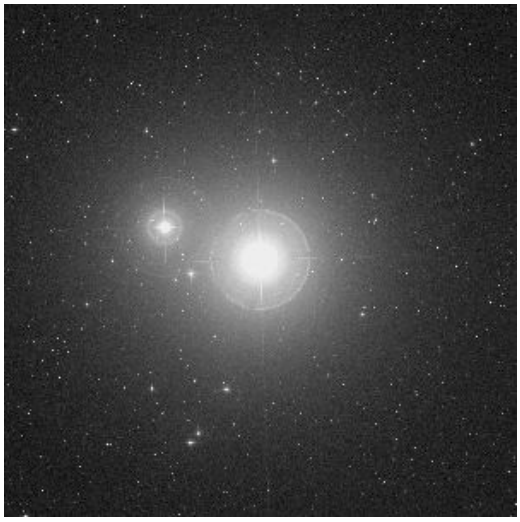
But, a new record for "most massive" star was claimed in July 2010 by astronomers using the ESO telescope in South America. Star R136a1 lies within a young cluster of stars in the nearby Large Magellanic Cloud. It has 320 times the mass of the Sun and 10 million times the solar luminosity. This young (about 1 million years) star will be short-lived. It is shown here among other large stars:



More than half of the stars in a typical galaxy are also tied locally to a second star as a companion (the mutual interrelation of two stars is referred to by the term 'binary'), such that each of the pair orbits around a common center in space determined by their mass-dependent mutual gravitational attraction. This arrangement is exemplified by the image made by the HST Faint Object Camera (FOC) of the Persei 56 group:



Mizar, one of the stars making up the Big Dipper, was first observed by Galileo in 1650. He believed he saw two stars that seemed to revolve about each other. This is the first record of a true binary. Mizar A and Mizar B are shown here in a modern telescope view (each one is itself mated to a small second star, so that this is actually a paired set of binaries).

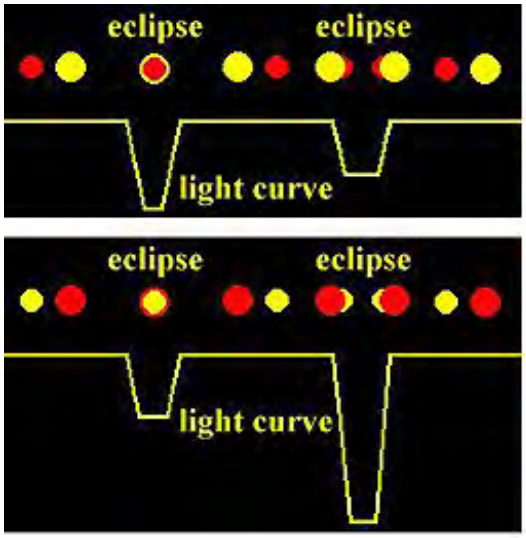


Some stars are grouped into more than one companion; ternary groupings (three stars orbiting about a common center of gravity) are fairly common. Here is an image of four stars orbiting as a unit about a gravity center in the galaxy M73.

Binary star systems are recognized by three means: 1) visual, through a telescope (as in the above two images); 2) by periodic drops in brightness caused by passage of one star across another (eclipse; an uncommon observation condition); and 3) by measuring spectral characteristics in which both a Doppler shift towards the red and the blue occur as one star moves away and the other towards Earth (and the reverse) along pathways of their mutual

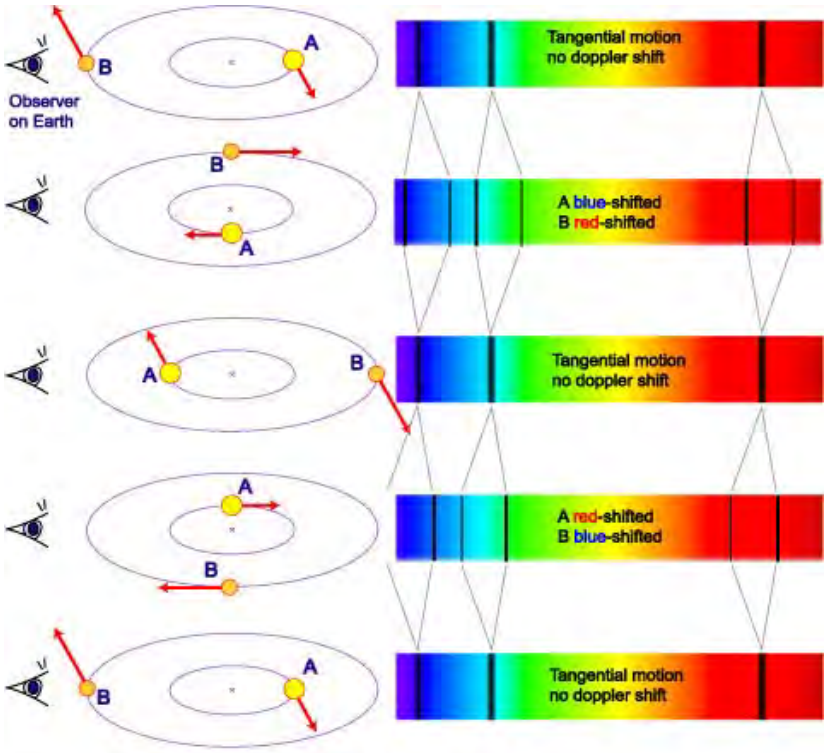
orbits.

To demonstrate the second means, examine this diagram which shows the brightness levels (and magnitude variations) for a binary star in which one is larger and brighter than the other (thus there are different decreases in brightness when the brighter star passes in front of the less luminous star, and vice versa). Incidentally, this method is also used to hunt for and verify planets associated with stars.



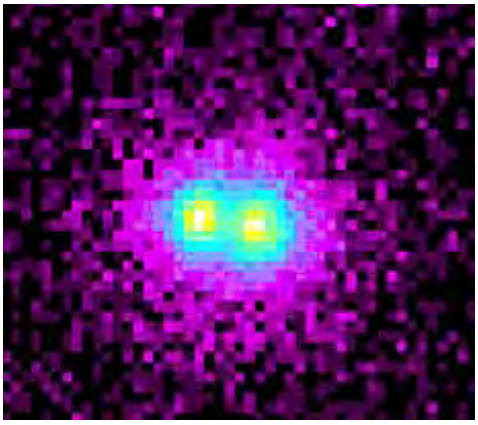
Spectral line shifts are used to study the motions of binary stars. We will treat stellar spectroscopy in detail on page 20-7 As a preview, the spectral method can be illustrated by looking at a pair of spectral strips for two similar stars that are mutually orbiting:

Bright lines for Hydrogen appear in the top and bottom (dark background) strips. This fixes a reference location for excited Hydrogen in the rest state. The two center spectral strips include the same Hydrogen lines, the first strip acquired from one and the second the other star. Note that the lines in one have moved to the left and the other to the right of the reference lines position. The spectrum on the bottom center has been blueshifted (see page 20-9) towards shorter wavelengths; the spectrum at the top center has been redshifted towards longer wavelengths. This is explained thusly: The bottom star is in motion towards the observing system on Earth whereas the top star is moving away from the telescope. This would occur when the two stars are aligned sideways to the line of sight and are moving in opposite directions around a common center of gravity. This diagram amplifies this explanation:



A Spectroscopic Binary System
High-mass star A and lower-mass B orbit around a common centre of mass. The observed combined spectrum shows periodic splitting and shifting of spectral lines. The amount of shift is a function of the alignment of the system relative to us and the orbital speed of the stars.

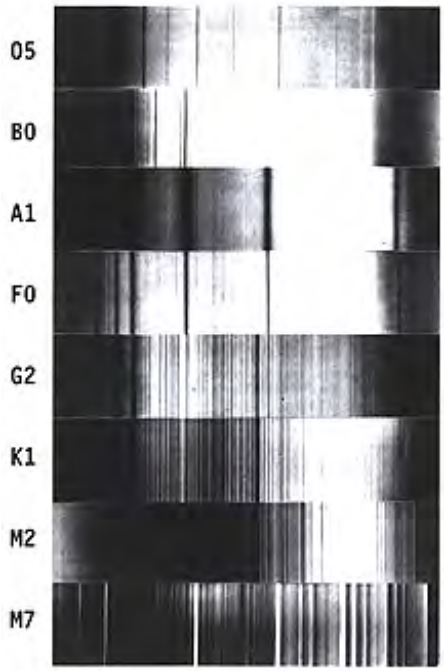
The Chandra X-ray Observatory has imaged a close binary pair in the M15 Galaxy. Prior to obtaining this image, the object was thought to be a single star, but at X-ray wavelengths, it is now resolved into a faint blue star and a nearby companion believed to be a neutron star giving off high energy radiation. Thus:



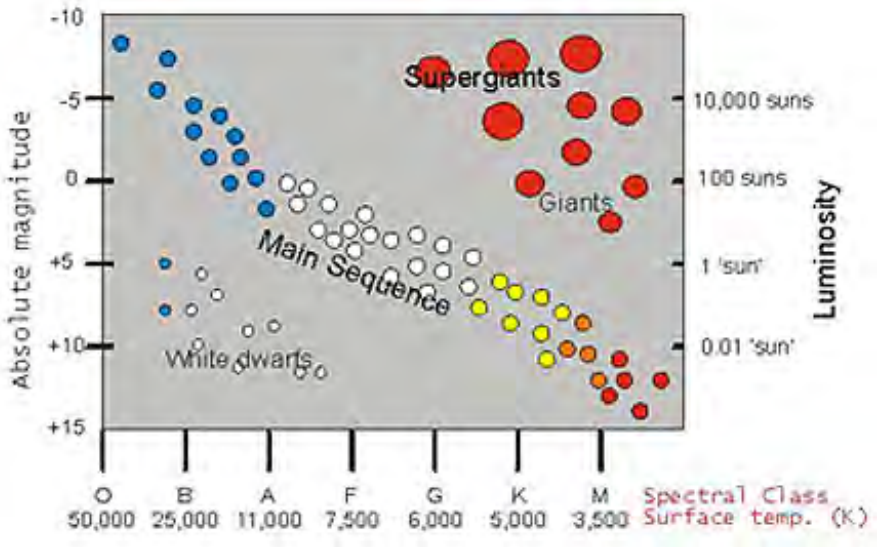
Most binary stars exist as two separated entities. But in 2008 an observation of two binaries that actually are enjoined (like Siamese twins) by shared solar matter were reported from a galaxy 13 million l.y. away. This is an artist's enhancement of one of these peanut-shaped pairs:

Turning now to stellar evolution, to preview what will be examined in some detail later on this page and the next page, the pattern of a star's history follows a pathway that, depending on its total mass, eventually splits into one of two branches (> or \>), as it leaves what is known as the Main Sequence. This is: 1) Development of a large cloud of denser gas made up of predominantly molecular Hydrogen (H_2) + dust --> Protostar --> T-Tauri Phase --> Main Sequence (**if mass less than 8 solar masses**)--> Red Giant --> Planetary Nebula --> White Dwarf; or 2) (**if mass greater than 8 solar masses**) Main Sequence --\> Supernova --\> Neutron Star or Black Hole (if **mass [size] is greater than 50-100 solar masses**).

The star types are categorized into Spectral Classes which are defined on the basis of certain chemical elements that become excited at different temperatures and give off characteristic radiation at specific wavelengths. The classes are designated by the letters (O, B,...etc.) assigned to each group. Here are spectra for some of the different classes (this is treated in some detail on page 20-7).



Both star classification and evolution can be summarized in a graphlike chart that consists of a plot of luminosity (vertical axis) or, alternatively the related magnitude parameter, versus star surface temperature which is expressed also by (correlated with) the star's visual color (note also the Spectral Type designations at the top). This is known as the Hertzsprung-Russell (H-R) Diagram. (The masses of the stars in the diagram increase to the left on the abscissa; Red Giants are big but have low mass densities [many less than the Sun, since mass was lost in evolving to that state].) Most known stars lie along the Main Sequence; they describe a stage in which a protostar reaches some fixed size and mass and commences burning of most of its Hydrogen before changing to some other star type off the sequence. Here is a H-R diagram:

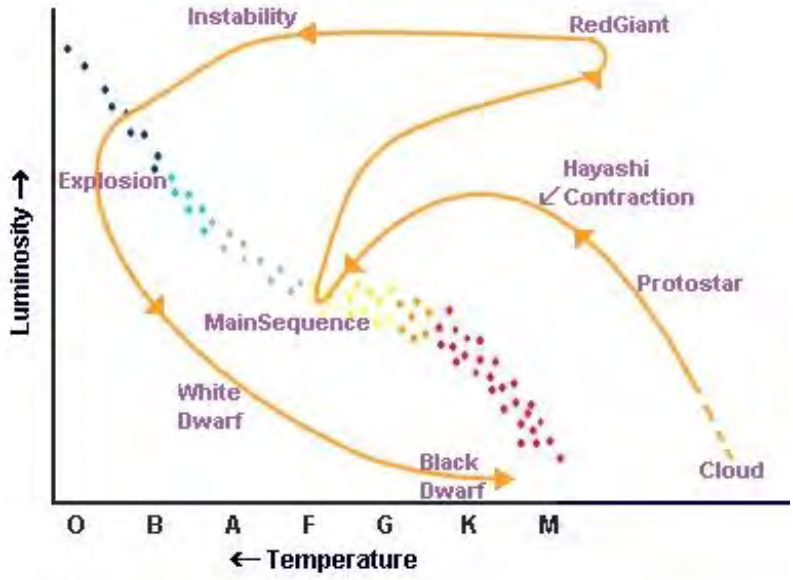


The Sun (also called Sol) is a G type star on the Main Sequence. Very hot stars on the M.S. include the Blue White stars. Red Dwarfs are M stars. Large luminous but low temperature stars form several Giant classes. Small, still luminous and very hot (at surface) stars make up the White Dwarfs.

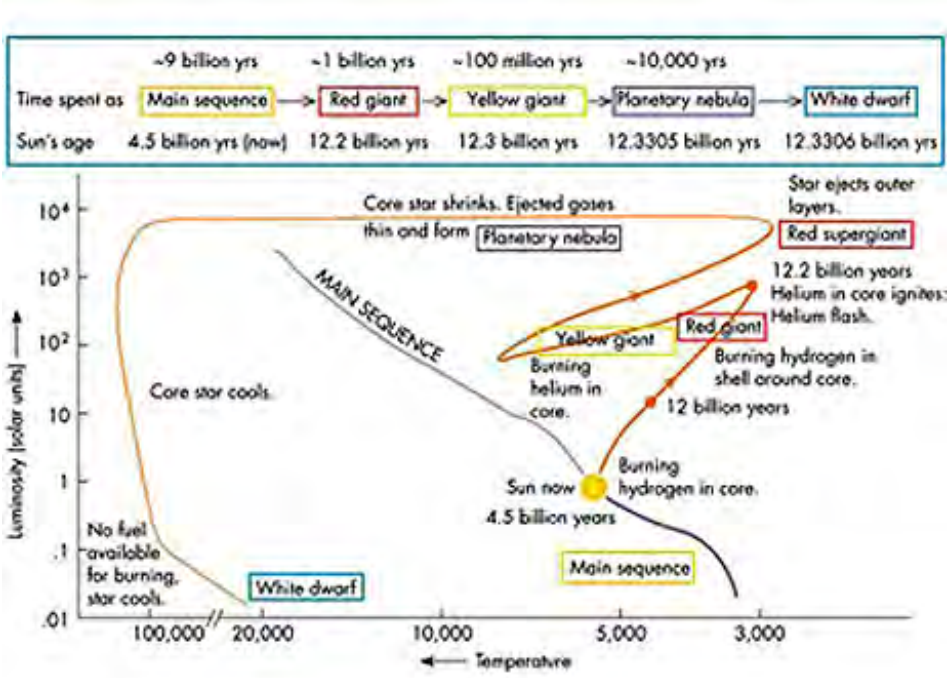
The above H-R plot also shows along the right ordinate the relative sizes of each star compared with the Sun. As far as we now know, stars do not completely vanish, but survive as dwarfs or Black Holes (but the latter in principle can disappear by evaporation as Hawking radiation).

Among the off-Main Sequence evolved star groups are four types of Giants (Sub; Red; Bright; Super), T Tauri. These are discussed again on this page or elsewhere in this Section. Not shown among the Dwarfs is the recent designation of LT for Brown Dwarfs. Note that the letters at the bottom include some like B0 and B5 or K0-K5; this denotes subdivision of each class into temperature subclasses (0 being hottest and 5 coolest in a class). Temperature ranges (in °K) are: O class = greater than 30000; B = 11000 - 30000; A = 7500 - 11000; F = 6000 - 7500; G = 5000 - 6000; K = 3500 - 5000; M = less than 2500. Colorwise, the first three are all "blue-white" stars, F is bluish to white; G is white to yellow; K is yellow orange; and M is red. Although not directly shown in an H-R diagram, there is a systematic increase in mass of a star going from the right to the left end of the plot.

A star on the Main Sequence will follow some pathway during its subsequent history. To illustrate this progression, look first at this evolution diagram for a star the mass of the Sun. The first diagram extends the history of a F star by showing the sequence of star stages from its very inception as a nebular mass that grows into a protostar, then to the M.S., next, as it burns most of its , off the M.S. as a Red Giant, followed by an explosion, and subsequent evolution into a final dwarf state.:



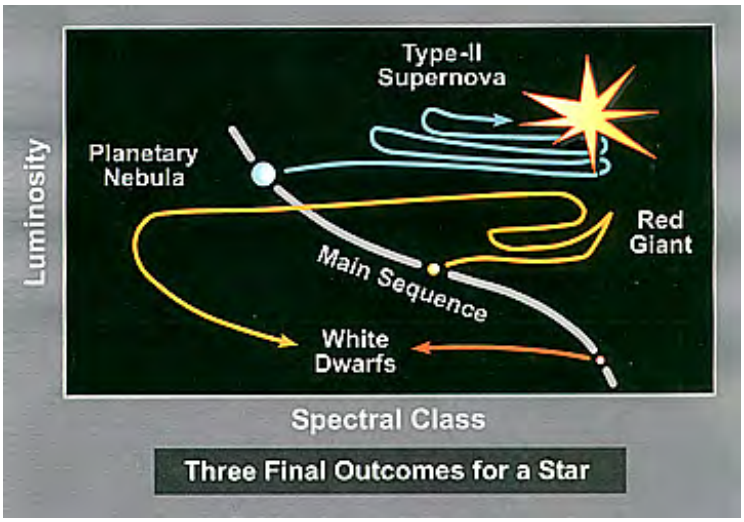
This second diagram follows the history of a G star (which is the path to be followed by the Sun in about 5 billion years) after it leaves the Main Sequence:



The key steps in the progression are 1) exhaustion of the main nuclear fuel; 2) change to a Red Giant, with shedding of some mass; 3) explosion to the Planetary Nebula phase, dispersing much of the star's mass into interstellar space; 4) survival of a central core as a White Dwarf star.

The pathways of protostars to the Main Sequence depend on their mass (in multiples of a solar mass) at the stage when they commence proceeding to the M.S and initiate Hydrogen fusion. The times involved in this transition will vary systematically with mass; thus, a 15 solar mass protostar takes only about 10000 years to reach the M.S. whereas a 2 solar mass star may require up to 10,000,000 years for the process to begin fusion:

This next diagram shows the evolutionary history of three stars of differing mass at the upper, central, and lower ends of the Main Sequence after they leave the M.S.:



These pathways are somewhat generalized: **A Sun-sized star (Class G) eventually becomes a Red Giant and then a White Dwarf. A smaller star (Class M) can evolve directly into a White Dwarf. A much larger star (Class B or O) will destroy itself as a Supernova that yields a planetary nebula (the gaseous remnants from the explosion; page 20-2a) but may retain some of its mass as a Neutron Star.**

The largest number of individual stars in galaxies fall in a narrow range between just under 1 solar mass to about 10 solar masses. During their evolution to Red Giants, they follow this internal history of burning (fusion) of the initial Hydrogen:

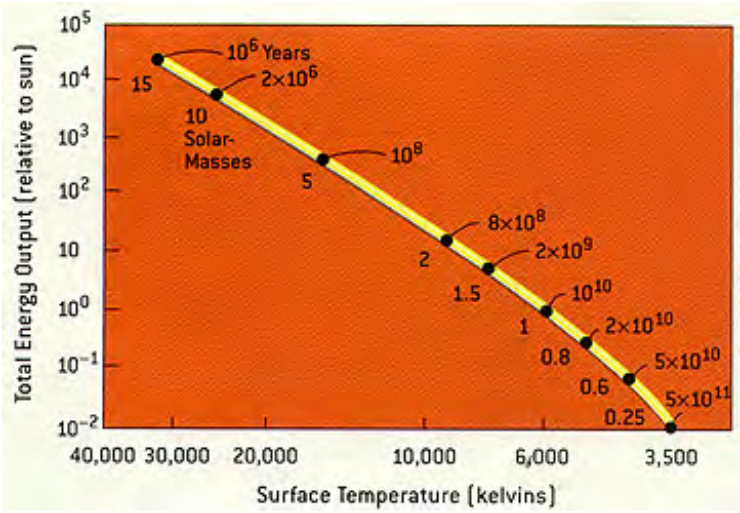
As these burn their Hydrogen fuel into helium, they start to contract and begin to burn that helium and further brighten, cast off some of the outer Hydrogen, and become luminous (for stars under a solar mass of 2.3, there is a short-lived large increase in luminosity known as the helium flash phase). Then, as the helium burns to carbon (which organizes into a core of degenerate carbon and some O(oxygen); see page 20-7), such stars follow what is known as the asymptotic giant branch (AGB) pathway which begins with a second Red Giant state.

A star's precise position along the Main Sequence depends on its total mass of H fuel that collects during the formative phase into the gas ball. Some stars (e.g., Type M) have masses as low as 1/20th of the Sun (1 solar mass is the standard of reference as is the luminosity of the Sun, also set at 1), whereas others fall within a range of greater masses that may exceed 50 solar masses (Type O). The high mass stars on the Main Sequence are brighter and bluer whereas those at the lower end of the M.S. tend to be yellow to orange. The initial quantity of mass in a star is the prime determinant of its life expectancy, which also depends on its evolutionary history and final fate. As a general rule, small stars may take more than 50 billion years to burn out completely, stars in the size range of the Sun live on the order of 5 to 15 billion years, and much bigger stars carry their cycle to completion in a billion or less years. Stars whose masses are similar to the Sun's actually will burn about 90% of their Hydrogen during their stay on the Main Sequence. Stars with greater than 50 solar masses may complete their M.S. burning in just 20-30 million years.

The lifetime spent on the Main Sequence is approximately proportional to the inverse cube of

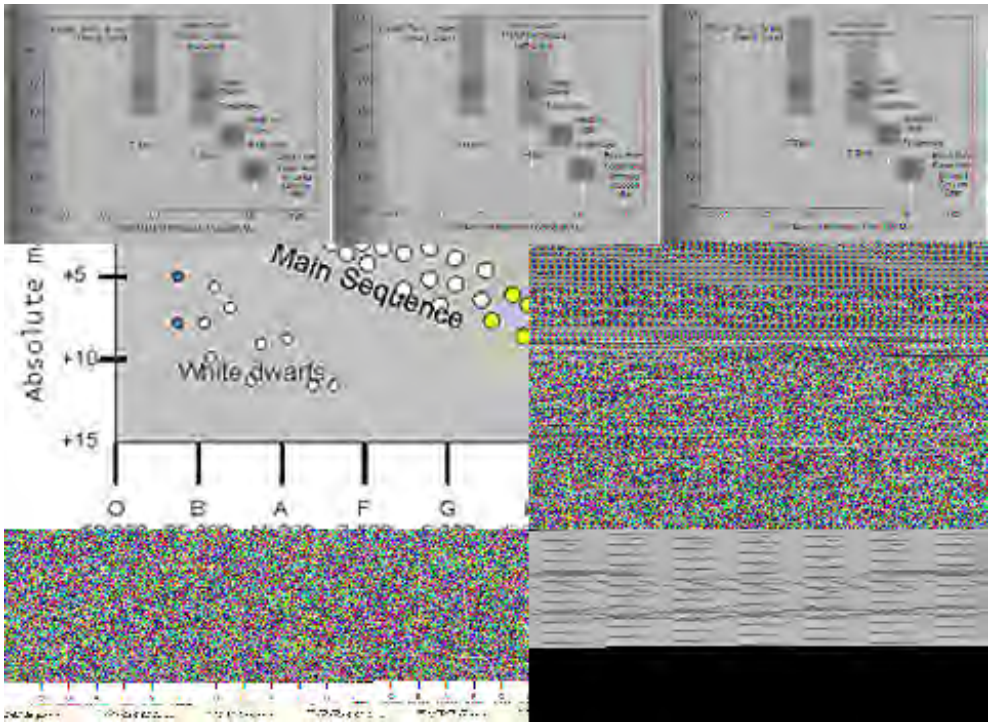
the star's mass (this is true for most stars, especially massive ones; stars less than a solar mass have lifetimes closer to the inverse 4th power). O and B blue-white stars may last only a few million years. Red Dwarfs can potentially last a trillion years or more.

The relation between size (mass) and age is shown in this next diagram (check the values on the curve itself, not the abscissa/ordinate values); the most massive stars have the shortest lifetimes.



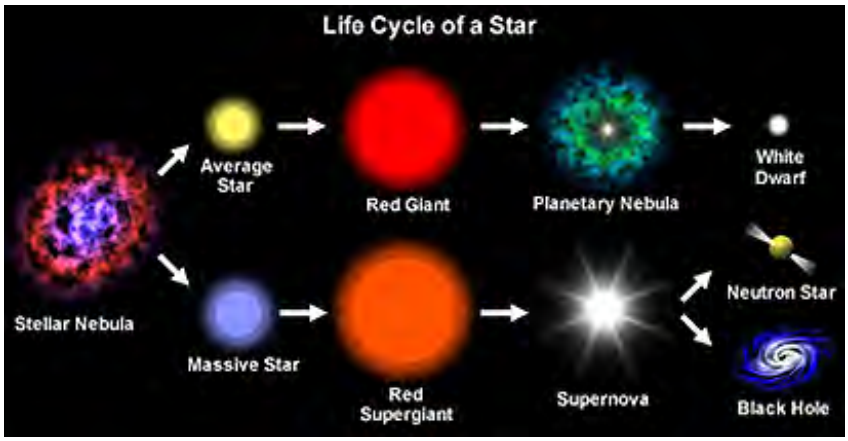
From B.C. Chaboyer, p. 53, Scientific American, May 2001

The history (from onset in the nebular phase) and fate of stars (at the end of their history) of all sizes (and different masses) can be conveniently summarized in this Evolution diagram; the various pathways depend on starting mass:

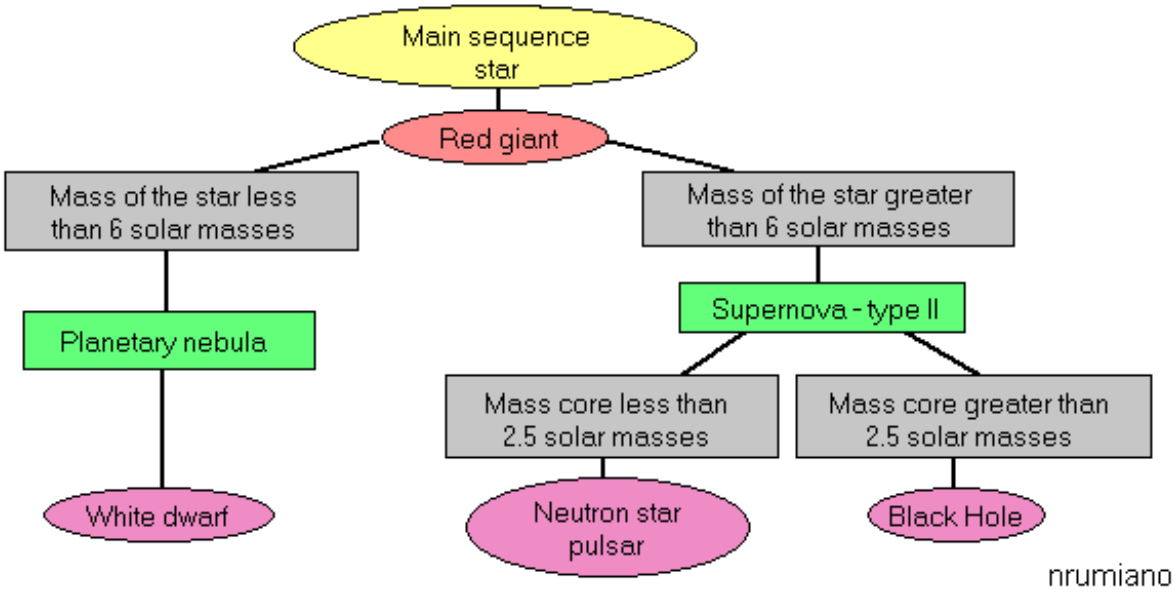


From J. Silk, The Big Bang, 2nd Ed., © 1989. Reproduced by permission of W.H. Freeman Co., New York

A generalized and simpler version of this shows the pathways followed by stars about the size of the Sun and stars that are much more massive:



A similar diagram that shows additional information is this:



Of special interest are the end products of each evolutionary path. After burnout or explosion, small stars end up as White Dwarfs; intermediate stars as Neutron Stars; and the largest stars as Black Holes. These end products will be discussed later.

Now to a more detailed discussion of the history of stars as expressed in the above diagrams. We shall begin by zeroing in on several of the common modes by which stars are born.

Stars develop within galaxies in clumps of gas and dust called Stellar Nebulae (also called *Giant Molecular Clouds* [GMC]) These nebulae are composed mostly of H_2)that are subjected to progressive sub-fragmentation, aggregation and contraction of the gas and dust into centers of higher density that become the sites of star birth. (More on GMCs is found on page 20-3; also do not confuse this type of nebula with "planetary nebula", the dispersed matter that makes up the residue from explosive destruction of a star, as described on the next page [20-2a].)

A typical GMC is a few hundred to a few thousand light years across. Its temperatures range from about 40 to 300 °: K. The total mass is on the order of 10^4 to 10^6 solar masses. Densities vary but 1000 atoms of H is a characteristic value. However, pockets of much higher density develop, commonly leading to eventual star formation. Galaxies contain hundreds to thousands of GMCs.

While the main ingredient of GMCs is molecular hydrogen (e.g. pairs of Hydrogen atoms bonded to make up molecules), the bulk of its mass may be due to a combination of the following molecules: formic acid, carbon monoxide, ammonia, acetylene, methane, methyl formate, ethyl alcohol (the "drinkable" kind) and hydroxyl radicals; particulate dust is also common.

These nebulae represent localized concentrations of gases brought about by several processes such as the driving force of shock waves from supernova explosions and intergalactic magnetic fields. The clouds turn very slowly but this helps to develop "seed" locations - internal denser regions that bring the gases toward them because of greater gravitational attraction. The H-He atoms in these denser local regions assemble into gas balls (the stars) and dust clouds by collisions and gravitational forces at initially low temperatures (100's of ºK) in a turbulent process of condensation, generating heat (in large part dissipated as thermal radiation). Thus, molecular Hydrogen clouds are the regions of gas where most new stars are born.

These clouds (GMC) are usually "photogenic" and hence many breathtaking images have been shown to the public. Let's start with an example of a mature galaxy in which star formation is continuing. A case in point is NGC 604, about 1500 l.y. wide, at the edge of galaxy M33; thus, this is the most active region in an already formed, but still primitive, galaxy in which Hydrogen gas has concentrated and is collapsing into new stars. We will take four looks at different scales. Here is the galaxy, which is 2.7 million light years away, with the reddish (from Hydrogen excitation) NGC 604 in its upper right:

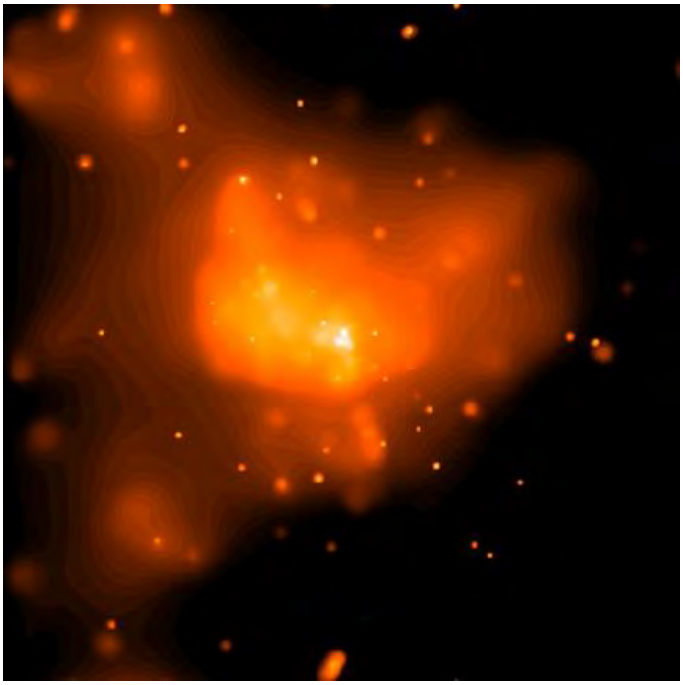
Seen through a telescope at the Kitt Peak (Arizona) Observatory, this GMC appears to consist of excited gases and stars seemingly associated with it (but some may actually be at different distances in the foreground):

As viewed by the Hubble Space Telescope, NGC 604 now shows some of the details of gases being moved about in a very irregular pattern, with stars forming as bright dots.

A later HST image of the central part of NGC 604 shows a characteristic feature, the development of a large number of small starbursts within the central part of the circulating gas medium.

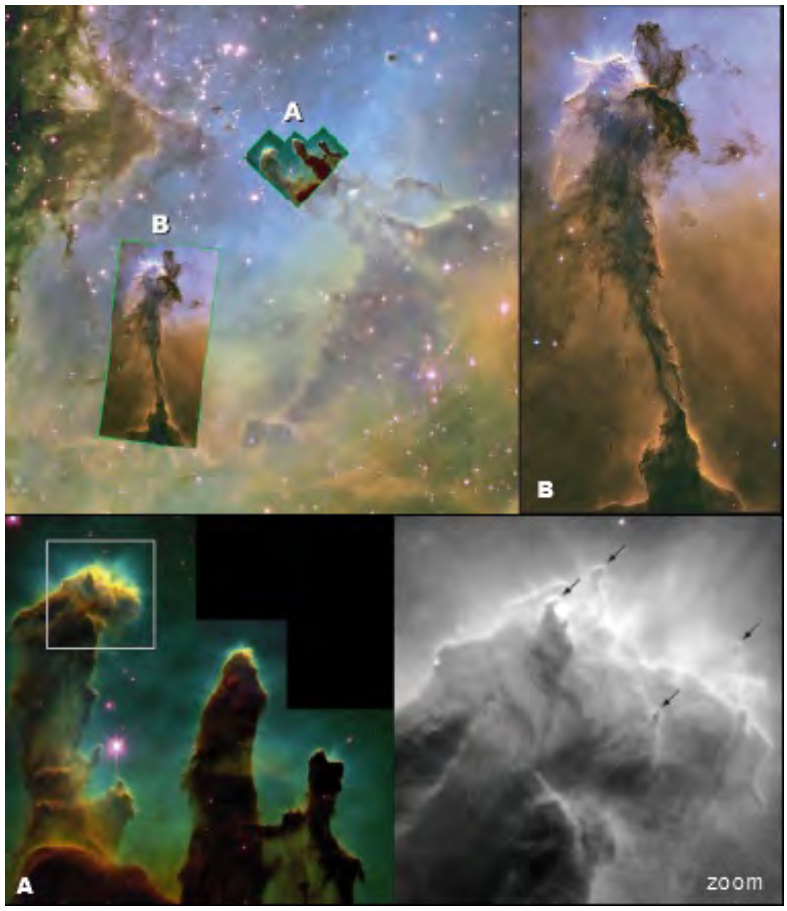
One of the most active regions of star formation is the central cloud in the galaxy NGC 1569, some 11 million light years away. The gas cloud shown in this next image is about 5000 light years in maximum dimension:

In the Milky Way the best known GMC is Sagittarius B2, near the center. Here is a view in which radio telescope data have been assigned orange colors to make it visible:



Most stars originate from within the GMCs. The gist of the **process of star formation**, which will be amplified later on this page and the next, goes like this: GMCs and similar nebulae consisting of mostly Hydrogen, some Helium, and varying amounts of other elements (produced from within stars in earlier generations), along with dust-sized collections of solidified Hydrogen and hydrocarbons, both build up and cool within the early Universe (and to a lesser extent continued to accrue as the Universe evolved). Temperatures dipped to a few hundred degrees Kelvin or less. With cooling myriads of patches within the GMCs increased in density. At various times individual pockets of gas contracted further with more cooling, developing local conditions in which gravitational attraction caused them to form into gas balls that both grew in size and heated up from the compression. At some stage the process accelerated such that the gas ball collapses on itself. When temperatures reached about 10,000,000 ° K, the Hydrogen nuclei (protons whose electrons had been stripped off by then by the thermal energy) - which tend to repel one another because they have like + charges - become energetic enough that they can then fuse. This thermonuclear fusion initiates the stage where the hot gas ball is now a true star. The high temperatures are maintained until the Hydrogen fuel supply is exhaused (various successive fusion processes can form some of the heavier elements [up to Iron], as described on page 20-7). An equilibrium sets in between thermally-driven outward expansion involving radiative processes and gravitation-driven inward contraction (see below).

To refine this idea: A general model of star formation is exemplified by studies of gas and dust clouds called pillars in the Eagle Nebula (discussed again in the first page dealing with galaxies). The next italicized paragraphs are taken from the Internet (one source: Softpedia).



The above image of the Eagle pillars is one of the most famous pictures taken by the Hubble Space Telescope. Within the nebula, stars begin in the especially dense clouds of molecular hydrogen gas (two atoms of hydrogen in each molecule) and dust that have survived longer than their surroundings in the face of a flood of ultraviolet light from other hot, massive newborn stars (off the top edge of the picture). This process is called "photoevaporation. This ultraviolet light is also responsible for illuminating the convoluted surfaces of the columns and the ghostly streamers of gas boiling away from their surfaces, producing the dramatic visual effects that highlight the three-dimensional nature of the clouds. The tallest pillar (left) is about a light-year long from base to tip.

As the pillars themselves are slowly eroded away by the ultraviolet light, small globules of even denser gas buried within the pillars are uncovered. These globules have been dubbed "EGGs." EGGs is an acronym for "Evaporating Gaseous Globules," but it is also a word that describes what these objects are. Forming inside at least some of the EGGs are embryonic stars - stars that abruptly stop growing when the EGGs are uncovered and they are separated from the larger reservoir of gas from which they were drawing mass. Eventually, the stars themselves emerge from the EGGs as the EGGs themselves succumb to photoevaporation.

In the first detail view, region A, we can see three columns of dust and gas (mostly hydrogen) illuminated by the young, hot, massive stars on top of the pillars. In the zoomed region these stars are indicated by arrows.

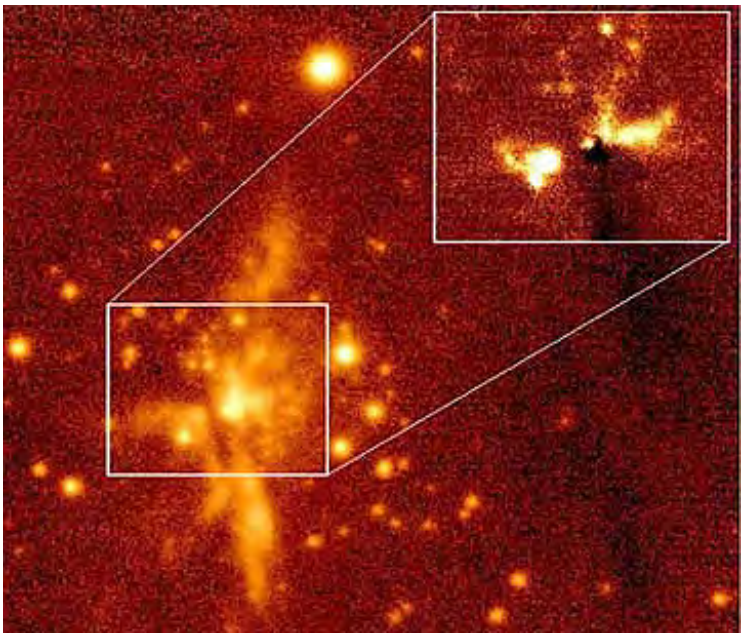
The stars' intense radiation heats the surrounding gas, making it glow. Moreover, this radiation is responsible for "sculpting" the columns through a process called photoevaporation - the light pushes away the feeblest particles and leaves behind only the larger conglomerates, denser gas globules called EGGs (Evaporating Gaseous Globules).Photoevaporation in the Eagle Nebula has cut newly forming stars off from the cloud feeding them. While some of the EGGs are large enough to eventually become stars, others may never make it.

In the second detailed view, B, we can see another stellar nursery (see larger image). The odd-looking tower is 9.5 light-years (or about 92 trillion kilometers) high, more than twice the distance

from our Sun to Alpha Centauris (the nearest star).

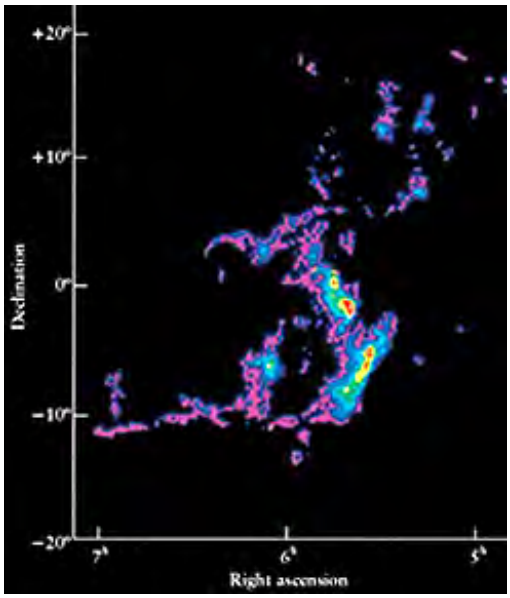
This giant cosmic sculpture is created by the ultraviolet light coming from the newborn stars. The stars at the top of the tower heat the gas creating the wing-like features. They also create a shock front that pushes against the darker cold gas and that will move down across the tower. Insofar it has just started, lighting up the "wings". This intense pressure compresses the gas, making it easier for stars to form. The threads of material in the center of the tower are also stellar birthing areas. They are roughly the size of our solar system. In this portion of the tower the formation of stars prevents other stars from forming because the light of the newborn stars is pushing away the gaseous material, dispersing it.

Now, consider this example of localized individual star formation. Stellar object 07427-2400 is a young forming massive star about 100000 years old located 20000 light years from Earth. It has a huge protostellar disc (GMC) of accreting molecular Hydrogen that is spiraling into its massive central star (now about 100 times the luminosity of the Sun). In the process, shock waves are produced that move against the disk, making it luminous also by exciting the Hydrogen and ionized iron. The IRAS Observatory has produced this image

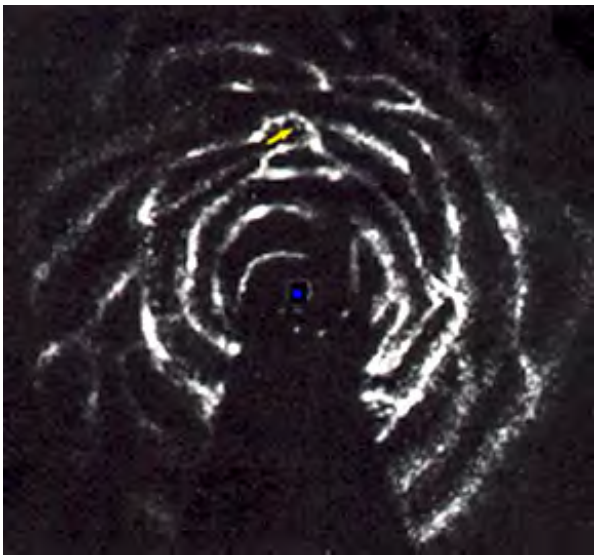


This image of the Trapezium nebula shows, in the inset, the birth of four individual stars from the gas and dust in the nebula:

One way to study GMCs is to plot the distribution of excited carbon monoxide (CO) dispersed within the molecular Hydrogen. In this state CO produces two prominent emission lines at 1.3 and 2.6 mm in the near radio wave segment of the EM spectrum. (H_2 does not emit strong signals in the radio region.) Here is the CO pattern that occurs in the Orion Nebula, a GMC which also contains regions of strong H_{II}, i.e., ionized H (see below).

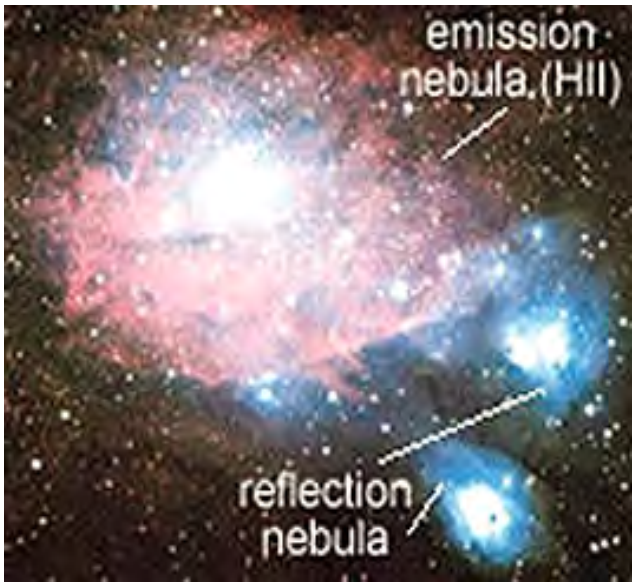


Outside the clouds, H and He also are dispersed, at much lower densities, as the principal elements distributed in interstellar space. The density of free H (mostly neutral) in that space is estimated to be between 3 and 8 atoms per cubic meter. This atomic Hydrogen when excited but not ionized is detectable by its signature at a 21 cm wavelength as determined through radio telescopy, representing photon radiation given off when excited Hydrogen reverts to its lowest energy state. But, in spiral galaxies most atomic Hydrogen gas has been rearranged in long streamers between arms of existing stars, as seen in this 21-cm radio telescope image of the Milky Way.



When GMCs heat up to temperatures above about 5000° K, the Hydrogen can be further ionized (see Page 20-7 for a discussion of the different ionized states of Hydrogen and their characteristic spectral lines). This gives rise to strongly emitting clouds that are referred to as *H_{II} Regions* (Atomic Hydrogen is denoted by H_I; singly ionized [loss of one electron] by H_{II}). One prominent line used to image and study H_{II} regions is Hα, whose line lies at 0.656 μm - the N3 --> N2 transition in Balmer series. These clouds are photogenic and deserve several examples here. First, an emission nebula as imaged by a telescope used in the 2Mass project (inventory of stellar objects in the Visible-Near IR):

Note this image which contains an emission cloud (pink) and two smaller reflection clouds (molecular Hydrogen) (blue):



The cloud contains a multiplicity of stars. In this next case, a cloud similar to the Eagle Nebula example contains just one star. This is typical, but is hard to observe in galaxies beyond the Milky Way. As will be discussed on page A-11, as a star begins to burn and send out strong illuminating radiation in the visible, the remaining gas and dust will become lit up as a distinct enshrouding cloud. In the image below a single star in the nearby Large Magellanic Cloud (a cluster of stars within the Milky Way's influence) is responsible for illuminating the irregular gas/dust cloud that has not yet (if ever) organized into a disk or ring but is likely to dissipate in part by further infall into its parent star.

Before organizing into an galaxy or after a galaxy has formed, the initial nebulae can have irregular shapes. Some nebulae appear dominated by dark dust, mixed with Hydrogen. These may have elongated shapes, some of which are described as "pillars". Part of the Eagle nebula contains such dark dust concentrations, as seen here:

A close view of one of these pillars (said by many as the most fascinating image yet obtained by the HST) is shown on page 20-11. Another type of dark dust-rich clot, with sharp boundaries, of star-forming material is called a "Bok Globule" (see several examples on Page 20-4), which commonly produces a large number of massive O-type stars, the brightest on the Main Sequence, that have short life times. Here is a typical grouping of dark patches that belong to the Bok Globule category:

A pair of Bok Globules in IC 2944 appear to be merging in this HST close-up:

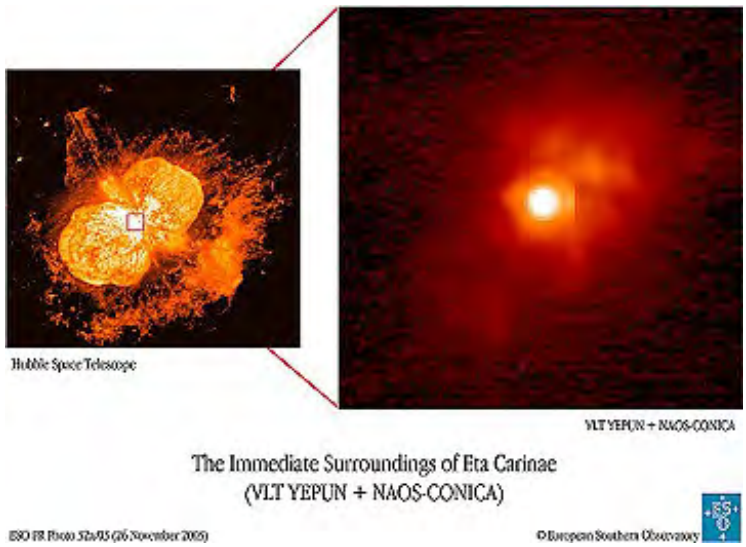
Dust is a major constituent of most galaxies. The dust often obscures the presence of many/most stars in a galaxy, at least as viewed in visible light. As we shall see on page 20-4, using sensors in telescopes (ground and in space) that image in wavelengths both shorter and longer than the visible can detect features and characteristics not evident in the narrow visible light spectral range. Thus, Bok Globule BHR 71 contains many luminous stars but these are masked by the dust, as seen in this ground telescope view (Observatory in Chile):

One of the largest nebulae is the Carina Nebula, seen only from Earth's southern hemisphere. It lies within the Milky Way at a distance from 6500 to 10000 light years from Earth. Here is a famed Hubble Telescope view, shown first without annotation and then with named components:

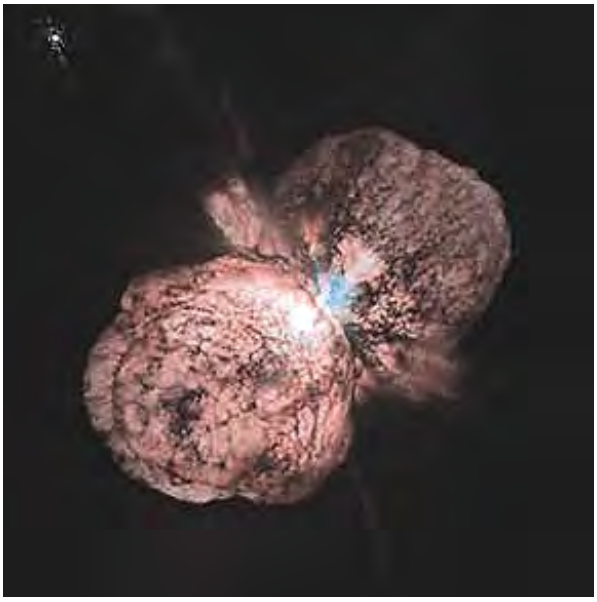
A subdivision of the Carina Nebula is the Keyhole nebula, some 9000 light years from our own galaxy. Its size is about 200 l.y. in diameter. It is classed as a dark nebula, but in this rendition computer processing brings out its rich colors. (Note: the term *nebula*, derived from the Latin for "cloud", has multiple meanings. In the early 20th century, the word was applied to bright objects in the sky that Hubble and others showed to be galaxies. Now, the term is restricted to any collection of Hydrogen gas and dust that may occur outside of a galaxy, as intragalactic material, or as remnants of exploding stars. A good review of the types of nebulae is found at The Web Nebula web page.

Here is a bright nebula about 7500 l.y. away that lies just before the Keyhole nebula. When imaged in the Infrared by the Spitzer Space Telescope (next page), the dust clouds and pillars of this nebula are revealed to contain newly forming stars, probably caused by shock wave compression of Hydrogen gas related to star flaring or bursting events. This is one mode of star formation that has been confirmed in other nebulae.

Within the Carina nebula is the very bright star Eta Carina, which displays two tear-drop plumes of gases and dust being ejected in opposition.



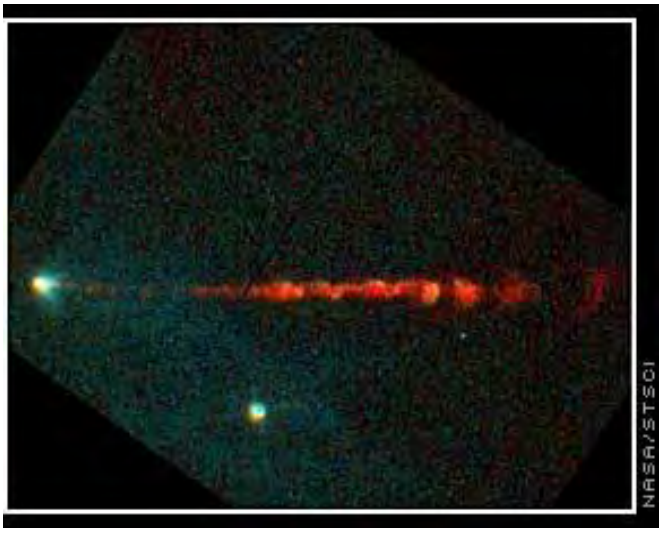
The star, first discovered by Herschel in 1677, began to flare up to brighter magnitudes in the early 1700s, faded, flared to a lesser extent in the 1800s, and became less bright by 1900. This HST view shows the gaseous material ejected in two directions; the star however is still present, thus this is not a supernova (page 20-6) but may be a nova.



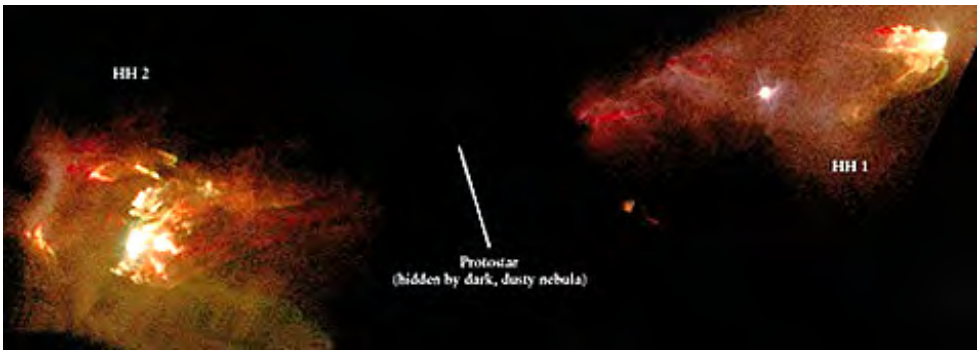
The HST Wide Field Camera has recently imaged a small cluster of stars in an early stage of their organization. This is in the Small Magellanic Cloud, about 200,000 light years away in the Milky Way galaxy halo. It and the Large Magellanic Cloud are considered small galaxies. This "cloud" (almost 10 light years wide) consists of glowing Hydrogen gas within which numerous stars are embedded. At least 50 of those that can be resolved appear to be young, massive stars. As time continues, these stars will enlarge as gravity pulls in the surrounding nebular material. Because of their large sizes, their destiny is to rapidly burn up their Hydrogen fuel, and eventually explode as supernovae (see below), many ending as neutron stars.

As a large number of stars develop from a nebula, and become luminous as Hydrogen-burning ensues, processes including radiation pressure from starlight will allow the stars to be seen through the diminishing dust and gas. The nebula may continue to produce more new stars if it draws more Hydrogen from beyond its boundaries, but generally nebulae tend to use up available H_2 and may deactivate. Stars may then form elsewhere as new clouds develop and reach conditions favoring stellar generation.

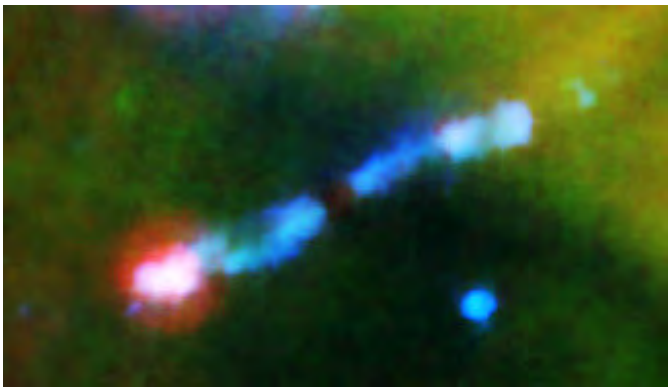
Individual stars develop along fairly well known blueprints. A central clot of mainly gas organizes and is surrounded by an envelope usually enriched in dust. As the protostar heats up, some of its material is ejected by magnetic forces as jets, such as in this example:



The expulsion of these high speed gases and charged particles can cause parts of the surrounding nebular masses to be excited and glow in luminescent patches. This phenomenon is known as Herbig-Haro (HH) Objects. Here is one example:

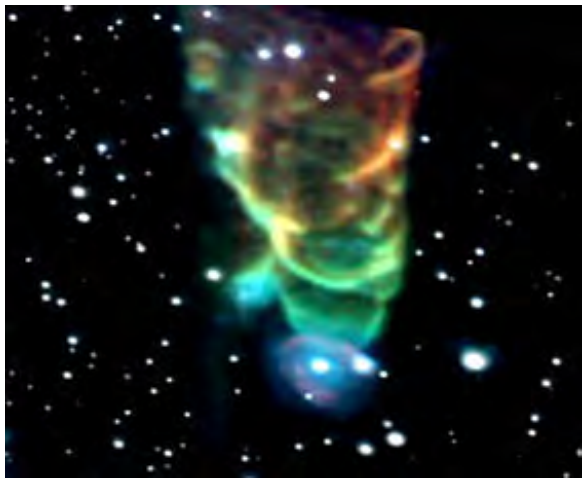


Gas jets are often developed during the HH phase. The jet from the nascent star HH211-480 contains discernible water (in its spectra). Beyond it (to the right) the gases have collected in a luminous "cloud".

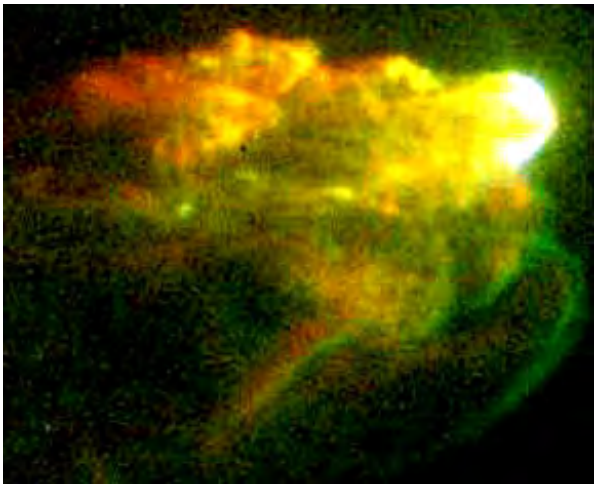


Instead of expelled jets, some stars have a "tail" analogous to that of a comet, produced by ablation as a star moves through space. Beta-Mira-Head-C is an example:

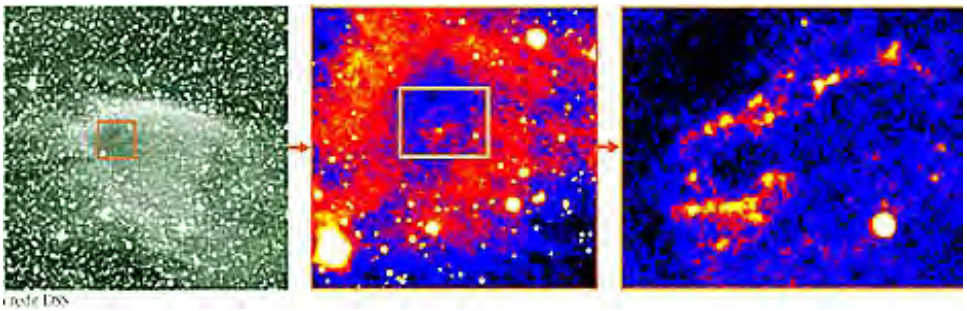
Sometimes the formation process during the HH stage produces an effect known as a "space tornado". The gases and dust involved seem to organize in a swirl like the winds of a terrestrial tornado, but on a huge scale. Magnetic lines of force, and electrical currents, may be involved. Here is an example:



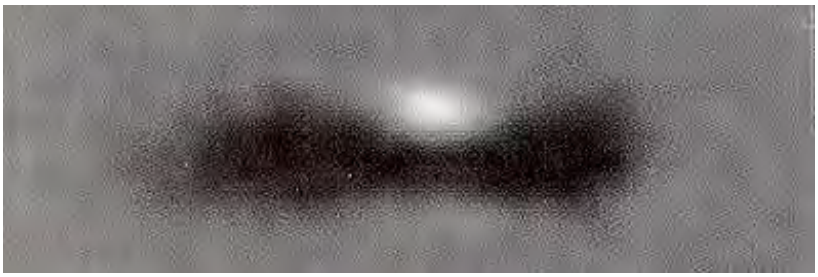
The emergence of these objects at two opposing sides (bipolar) of the protostar is typical. The next HST view shows this HH effect in a glowing "cloud" which is located near the end of a jet (bright hemisphere, to the right) passing through it.



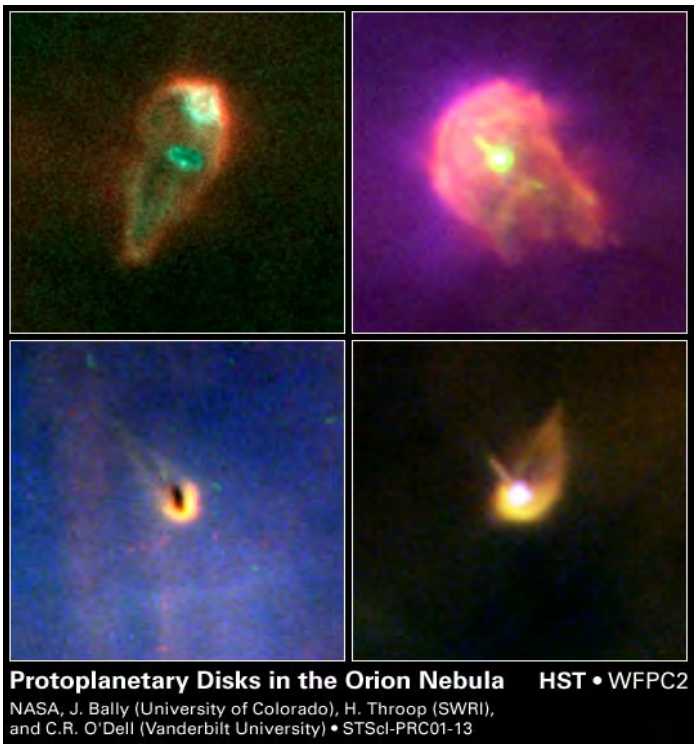
In the nascent star phase, the dust and gases form a very large volume of organizing material called a "globule" (at least some of these are Bok Globules; see above). A globule in the inner Milky Way, designated DC303.8-14.2 shown below, was first detected by ESA's IRAS satellite. In this trio of images, the left image of the globule, obtained during the Digital Sky Survey observations, shows the extent of the nebular mass seen in visible red light. The center image made by Kimmo Lehntinen's team using VLT ANTU telescope at ESO's Observatory in Chile, is the inset of the left image shown here in color from several infrared band images on this telescope; it shows a distinct ring of gases and dust that emits strongly in the infrared. The right image (of the inset of the center image) indicates several jets of the Herbig-Haro type involved in the early stages of formation of the eventual star.



From the above discussion we conclude that the dominant behavior during the pre-Main Sequence history of a protostar is marked by light gases continuing to inflow and build up the star's mass and size. Much of the dust remains as a thick disk outside the star, such as this example:



Developing stars with dust-rich accretionary disks are common, perhaps the norm, in galaxies. As will be further explained on page 20-11, disks like this are the potential conditions that lead to planet formation. Such stars are called protoplanetary disk types; this is contracted to give the term *proplyd*. The Orion nebula is usually cited as the type location for proplyds (which are, in effect, planetary nurseries). Here are two images that show typical examples; note that a discrete spherical star has not yet formed in most:



Now, lets discuss what happens as the star approaches pressure-temperature levels capable of

initiating Hydrogen fusion. When more matter accrues within a growing nebula, its internal gravity continues to increase and draw in still more gases. Gravitationally-driven collapse into forming stars induces compression and further heat rise. The protostar phase is reached as temperatures rise to 2000 - 3000° K. At ~10,000° K, the H begins to ionize (electrons stripped away) and, in the process, loses some heat energy by radiation which tends to slow or counter the compression. Over time, the cloud eventually reaches a density that requires it to then undergo local clumping of gases into clots that grow into still denser concentrations to become stars (these smaller clots can exist for much of the galaxy's life but are the sites of further star formation). Here is a Hubble Space Telescope false color view of the central Orion nebula, which appears to be in an early stage of organization into stars (hence, a younger nebula). (See page page 20-3 for other Orion images, treated on that page from a galaxy formation viewpoint.)

This next view of part of the Orion Nebula is inserted here to make a special point. While Hubble, Chandra and other space telescopes, and some of the large ones on Earth usually seem to provide the most spectacular images, small telescopes operated by "amateur astronomers", if used effectively, can yield their own superb views. The image below was made through a 14 inch "backyard telescope" by Russell Croman using filters and exposing on a tracked target for 7 hours. (Check out his web site for many other astronomical photos taken by him.) The color output rivals some HST images. Red in the image highlights sulphur-rich parts of the nebula; green show Hydrogen enrichment, and blue singles out Oxygen.

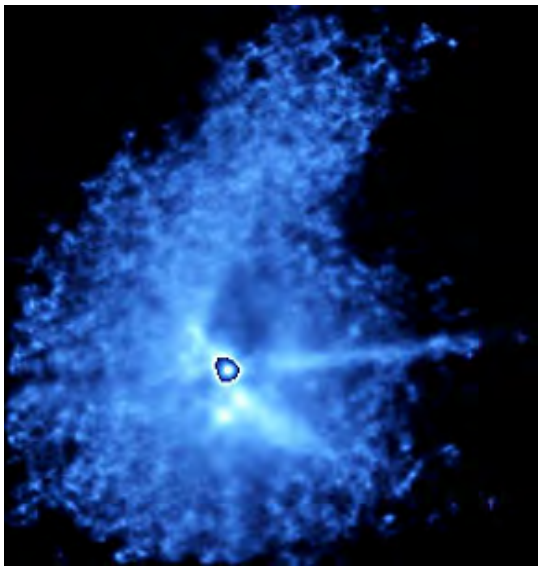
The message given by this example is that anyone - including those who are not professional scientists - can participate in the exploration of the Cosmos.

That message is re-enforced by this next image, also made by astronomer hobbyists di Cicco and Walker. It shows a part of the night sky above a house outside of Boston. The sky was photographed over a cumulative forty hours. There are broad wisps that are actually individual nebulae in the Orion-Eridanun supergroup. These nebulae are in the Milky Way and are relatively close to Earth - hence show up as large. Each nebula hosts many (unresolved) stars.

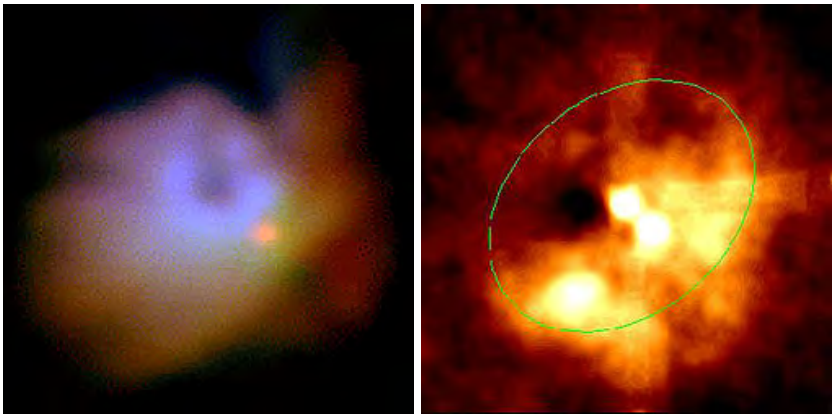
In July, 2003 a report was released stating that the Orion nebula contains the hottest stars yet discovered in the Universe. Temperatures were obtained using Chandra X-ray data. The 3 hottest were supermassive stars shown in the right panel below:

The single hottest of these stars reaches a surface temperature of 60 million degrees Centigrade (108,000,000° F), more than double the value of the previous record holder.

As the early stages of star formation proceeds, the cloud tends to gather around the star in a more isolated manner, removed from neighboring gas and dust nebula. It may then enter the T Tauri phase at which the growing star starts to generate strong stellar winds. The cloud disk still can exceed 150 A.U. in dimension. This telescope image shows the glowing cloud (rendered here in blue, but actually of a different color) around the incipient, still poorly organized central star (a binary pair).



Here are two more T Tauri stars, the one on the left showing the nebular shield that masks the bright growing star and the one on the right showing another T Tauri star as seen in the infrared:

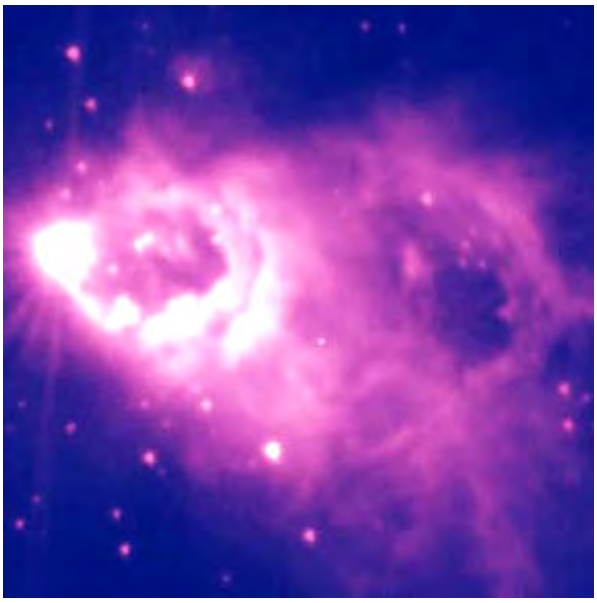


The star now rapidly contracts as it passes through the Hayashi phase. This relies on the proton-proton nuclear reaction which releases radiation energy that causes a notable increase in luminosity. However, hydrostatic equilibrium (see below) is not yet reached as the growing star continues to experience disruptive convection.

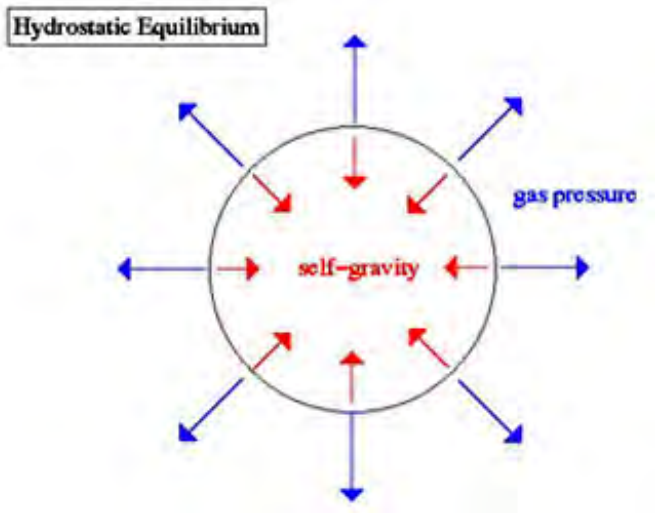
When a star has finally organized into its Hydrogen-burning sphere, it may eject and dissipate its remaining nebular material as shown in this image of what is now known as McNeil's nebula (named after its initial discover, an amateur astronomer):

For stars of masses near that of the Sun, it takes about 10 million years to work through the protostar phase and another 20 million years to join the Main Sequence. More massive stars reach the Main Sequence more rapidly. Below is a view taken through the Japanese Suburu Telescope of S106, which has a mass density twenty times that of the Sun, that began to burn only about 100,000 years ago. This star, 2000 l.y. from Earth, still is showing dust and gas flowing into the central body.

An early stage of another massive star, AFGL2591, 10 times the size of the Sun, has been viewed in infrared light by the newly operational Gemini North Telescope on Mauna Kea, Hawaii. Some 3000 l.y away in the Milky Way (located against the backdrop of the Constellation Cygnus), the central region of the forming star is still disorganized. Infalling material continues its growth but also sets off a return outflow of gas and dust.



After a star has moved onto the Main Sequence, the history of its life cycle there will be a continuous (somewhat oscillating) "contest" between contractive heating during stages of gravitational collapse and expansive cooling by thermal radiation outbursts whenever rising temperatures increase Hydrogen ionization. Generally, an evolving star tends to seek out a balance [hydrostatic equilibrium] between inward gravitational forces and outward radiation pressure developed from the burning of the star's nuclear fuel. This is illustrated in this simplistic diagram:



In its early life, the contraction phase ultimately dominates, so that a star's deep interior temperature eventually will be raised above 10^7 K (varies with star size), at which stage a fundamental nuclear reaction within the Hydrogen gas commences. This involves thermonuclear fusion: p + p => H^2 + e^+ + neutrinos (H^2 or deuterium is a single proton and a neutron and e^+ is a positron [emitted]). That change of state results in thermal energy release which contributes to continual rises in temperature. Deep within the star, an alternate but dominant fusion process involves melding of 4 single protons into a single helium nucleus consisting of two protons and two neutrons. As temperatures increase further, some protons, neutrons, deuterium (and minute amounts of tritium [H^3]) combine (in a three step process) into helium (He^4 nuclei [2p, 2n]) which migrates into the star's interior towards its core. In these reactions, some of the mass is converted to energy ($E = mc^2$) which radiates outward as the source of the star's luminosity and which produces the outward pressure that counteracts inward forces owing to gravitational contraction. Luminosity varies as the fourth power of a star's mass (thus a star with twice the mass of the Sun shines 16 times brighter).

Helium remains stable until temperatures approach 100 million° K, at which state it reacts with

more protons and neutrons to transmute into other elements of higher mass numbers (see below). More massive Main Sequence stars can generate Carbon; some of this element may be in the star initially if it is formed from previous gases and particles that contain carbon produced in earlier star generations. This carbon-enriched star, as its temperature rises and interior pressure increases, can go through another fuel-burning process known as the CNO. Through a series of steps as reactions of Carbon with Hydrogen protons take place, first C^{12} is converted to isotopes of Nitrogen or O^{15} but reaction with He^4 will lead to C^{12} again plus energy released as positrons and neutrinos.

When the H => He process reaches a steady state, gravitational contraction no longer dominates (attains a balance called *hydrostatic equilibrium)*), the star's total radiant (EM) energy output per second (defined as its luminosity; also referred to as brightness) becomes constant, and the star reaches a stable state on the Main Sequence (M.S.), populated by stars that are primarily in the Hydrogen-burning stages. This equilibrium - in which inward directed gravity forces are more or less countered by outward radiation pressure - is maintained during most of the star's life on the Main Sequence. These stars spend up to 90% of their total lives on the Main Sequence.

To re-enforce these statements about a star's early history, review this chart:

The page on Stars is continued on page 20-2a, accessed by the Next button at the bottom here or at the top of this page.

PREVIOUS RST NEXT

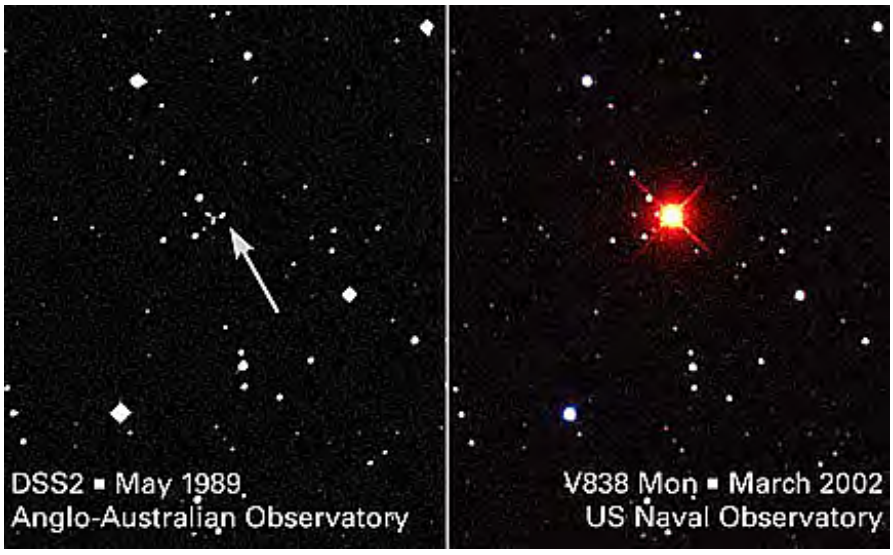
Primary Author: Nicholas M. Short, Sr.

Novae, Supernovae; Neutron Stars and Pulsars; Quasars and Black Holes; Gamma Ray Bursts; and Star Collisions.

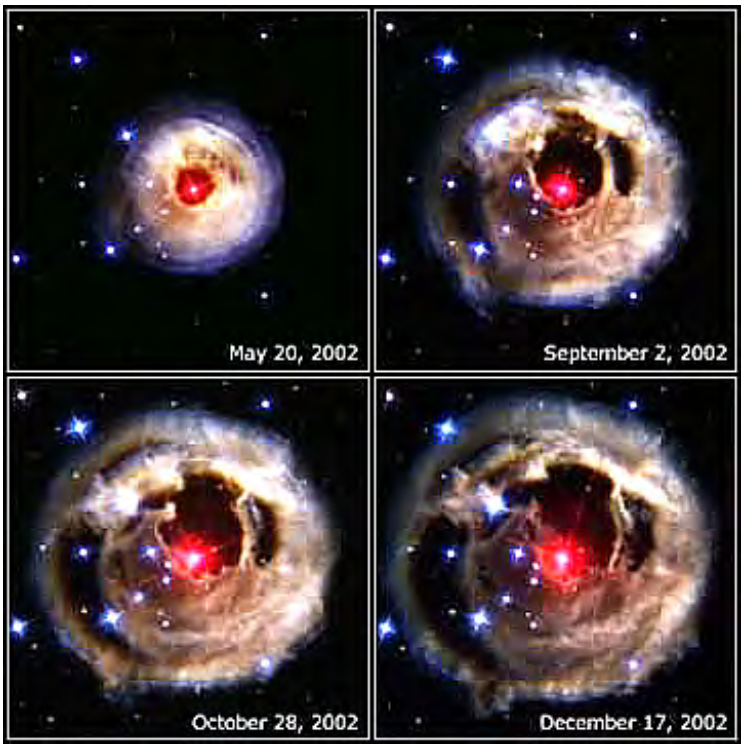
We return now to special events that are associated with astronomical bodies on the scale of individual stars.

Novae and Supernovae

Periodically, a bright object appears in a galaxy and remains that way for days to months. It is referred to, erroneously, as a *new* star or *Nova* (Latin for "new"; plural, Novae). It is, in fact, a star that for more than one reason experiences a major flare-up that later dies down leaving the star intact but with loss of material. Star V838 in the Monoceras constellation underwent such a flare-up of a blue star, releasing considerable material as it reached a luminosity of ~600000 times that of the Sun.

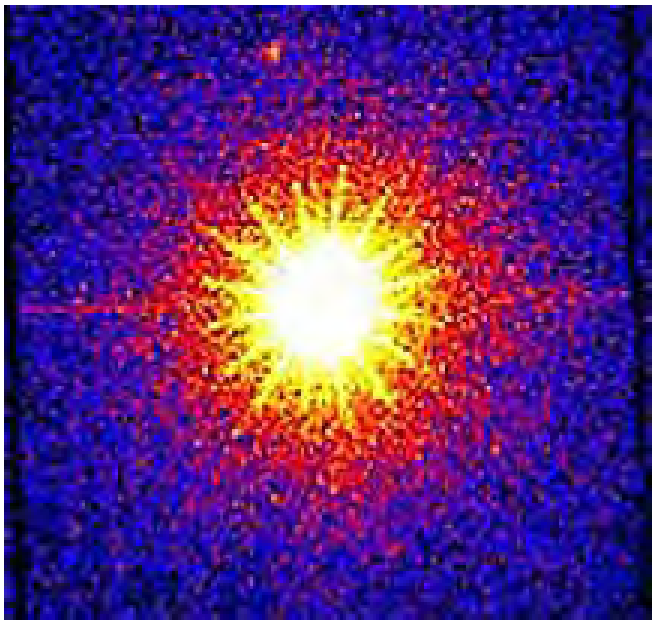


Soon after its discovery, this Nova has been examined in close detail by the HST, yielding this dramatic sequence of images:



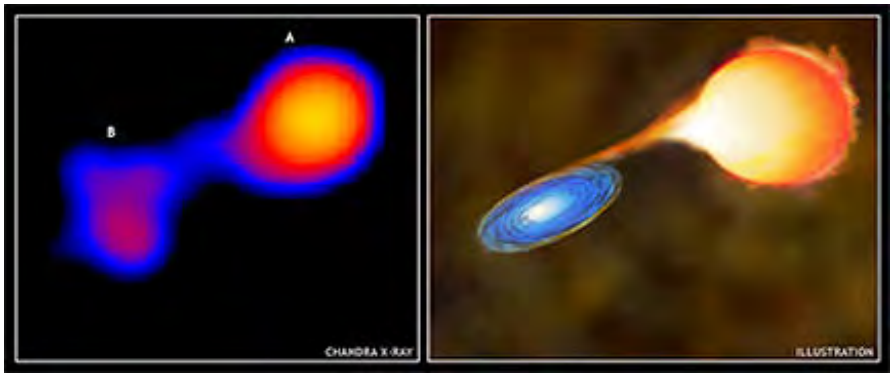
Recently, a new hypothesis about its effects on possible planets orbiting it contends that the flare-up consumed these planets. There is, of course, no direct evidence that there were any planets around this star. But the spectra obtained for the Nova phase of V838 show a strong enrichment in Li, Al, Mg and other elements that could have been concentrated in planetary bodies that were caught up in and destroyed by the expanding shell of gases in the flare-up. This is the predicted fate for most planetary systems as the parent star expands its gaseous envelope. Our Solar System will likely be destroyed by such a process in about 5 billion years hence (see page 20-11).

Novae events involve release of huge amounts of Gamma ray and X-ray radiation. The XMM-Newton spacecraft sensors on September 22, 2006 captured this X-ray image of a Nova near the center of the Milky Way:

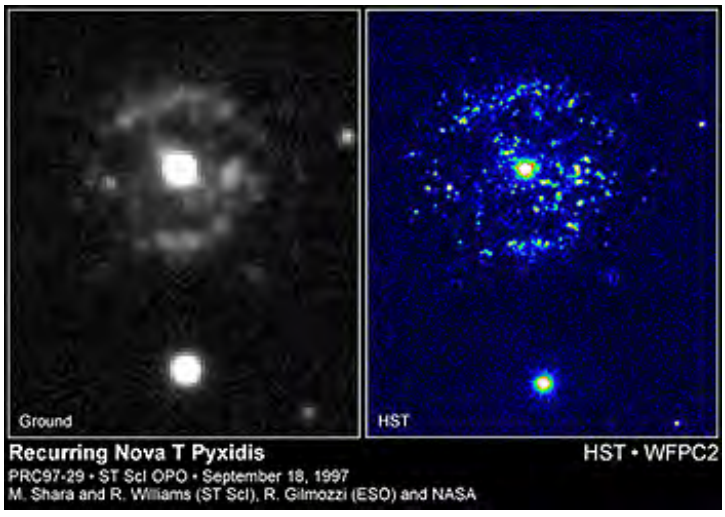


Novae are common events in individual galaxies. One way in which these occur is as follows: What in fact is being observed is a binary star system, one member of which is a White Dwarf and the other a Red Dwarf or even more massive Main Sequence or Red Giant stars. The process involves stripping off of Hydrogen from the larger companion star which streams toward and is added as an accretion disk to the White Dwarf, whose gravity controls the activity. This

artist's concept in the right illustration below shows what happens (in reality, the material being removed is often not be luminous which is why the actual process is not observed around such star pairs). On the left is an actual case, in which Chandra has produced an X-ray image of star Mira B, a White Dwarf, receiving material pulled from its companion, Mira A, a Red Giant.:



This accretion process causes a buildup of Hydrogen gas around the White Dwarf until compression under the strong gravity raises the temperature to 10^7 °K, at which condition nuclear fusion occurs. This causes a sudden brightness of the White Dwarf and a rapid consumption of the accreted Hydrogen some of which also may be expelled. The process repeats through a number of cycles, at time scales of 1000s to 10000s of years per flare-up. Novae are therefore recurring events, without star destruction at each occurrence, in distinction to the Supernovae described below. The illustration below shows a Nova near its peak brightness; the specks around it are ejected Hydrogen (the star near the bottom may [?] be the source of the accreted material):



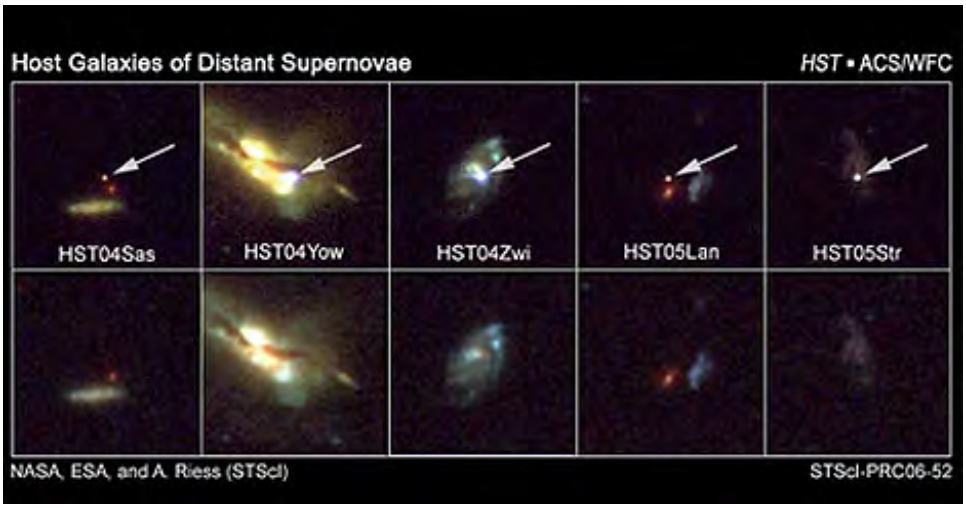
This star, normally a small ordinary type in the Milky Way about 20000 light years from Earth, is an *eruptive variable* in which the Hydrogen-burning undergoes a significant flare-up, enlarging the star somewhat but not directly shredding off or expelling significant mass. At its peak, the star had about 600,000 times the energy output as the Sun but in time settled back to its prior state.

As hinted at above, the hallmark of a Nova event is that it RECURS (happens more than once; repeats), because while explosions shed some stellar material each time the star brightens, the process of Hydrogen-burning continues until all outer material is removed. This has been the case for star RS Ophiuchi, some 5000 light years away, which has become a Nova 5 times in the last 108 years. Now a white dwarf, this star erupted in February, 2006 and became bright enough to be found by the naked eye. It is also a strong source of X-rays. It has shedded its outer atmosphere in a near-perfect ring, as shown here:

More massive stars, originally with 8 to 50+ solar masses, burn their gaseous fuel (in the plasma state [atoms are ionized]) much more rapidly until nuclear processes force the gases away at high velocities from the core in an explosion whose early stage may be seen from Earth for a few years as a hugely luminous event called a Supernova. Red Giants are the most common star type that is destroyed in this way. One such very bright event was imaged by HST on April 28, 1998 in the spiral galaxy NGC3982; the Supernova is the large blue-white object in an arm off the galactic center. Supernovae occur, on average, about once every 30-50 years in a galaxy.

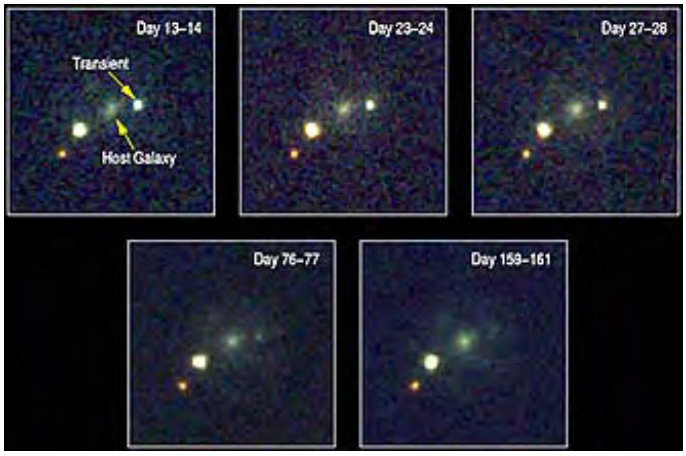
Credit: H. Dahle

The hallmark of a Supernova event might be verbalized by the magician's dictum: "Now you see it - now you don't". This refers to the short span of time in which a Supernova shows its distinct characteristics - from a few years to a thousand or more years. Consider these Hubble views of five Supernovae:



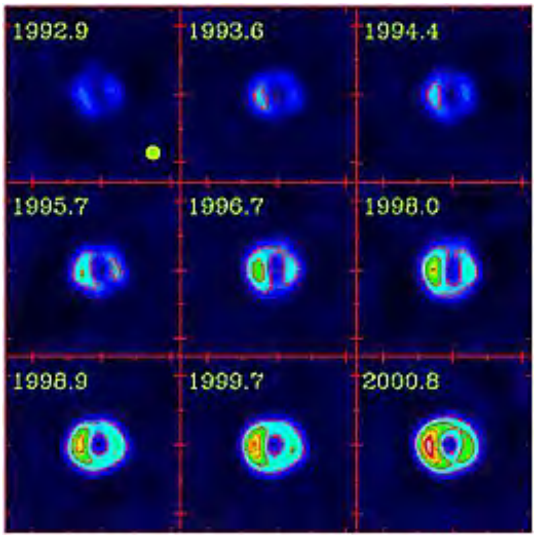
In each case, the Supernova - a very bright spot - appears for a short time (weeks to months), then disappears.

The rapid rise and decrease of luminosity during a Supernova event (labelled "Transect") was captured visually through a telescope looking at GRB 011121 (a Gamma-ray-burst source; see below):



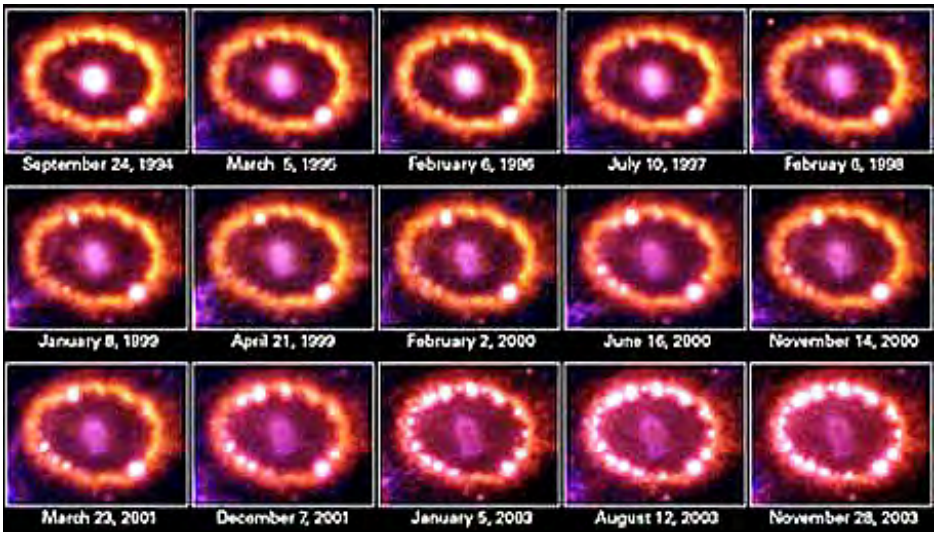
In February, 1987 the brightest Supernova in nearly 500 years, SN1987A (located in the Large Magellanic Cloud), was first discovered in the southern hemisphere skies by an Observatory in Chile. Here is a before-after image made by a telescope at the Anglo-Australian University:

Since 1987, it is being continuously monitored both from Earth and from the HST, providing a "stellar" example of the self-destruction of a star by catastrophic explosion. It appears to still be in a declining luminosity phase going into the 21st Century.



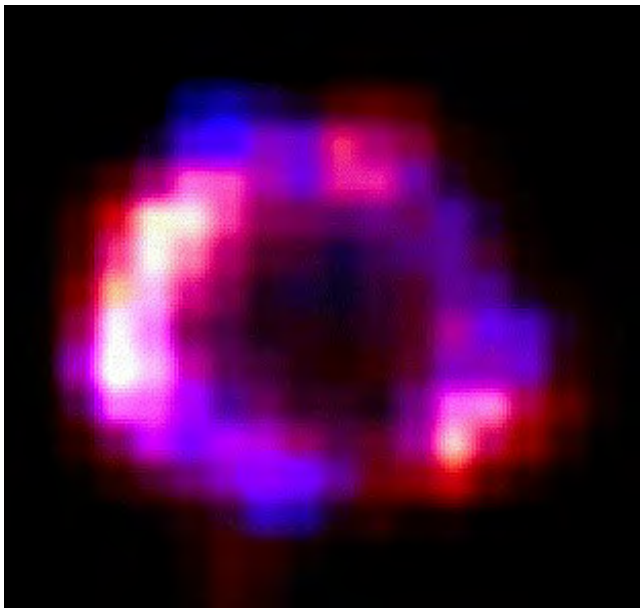
This next image is one of the most spectacular views of 1987A yet acquired by the HST. The single large bright light is a star beyond the Supernova environs. Around the central Supernova is a single ring but associated with the expansion of expelled gases are also a pair of rings further away that stand out when imaged at a wavelength that screens out much of this bright light.

Visual changes around the star remnant and its surroundings have since been observed over the last 15+ years, as shown in this sequence made with the Wide Field Camera on the HST.

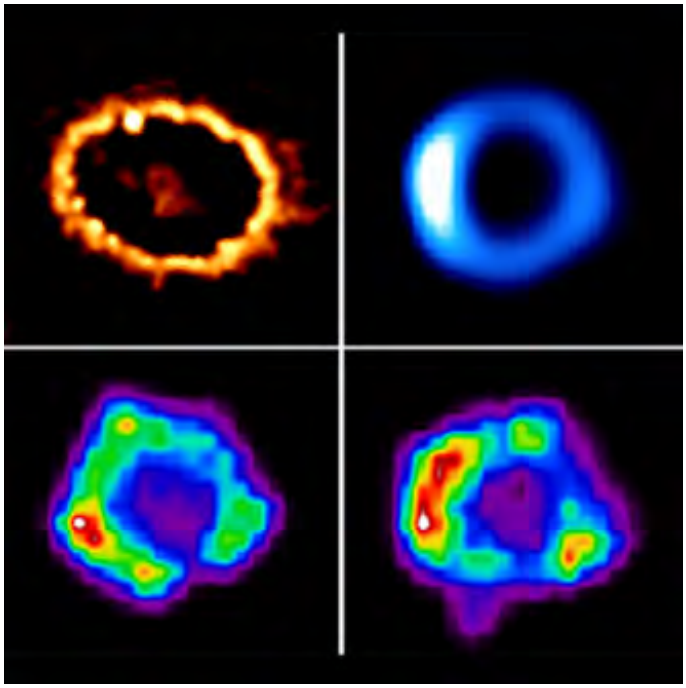


This Supernova is also expressive as a concentrated source of X-ray, UV radiation, and Radio

waves. Here is the first Chandra X-ray image of SN1987A:

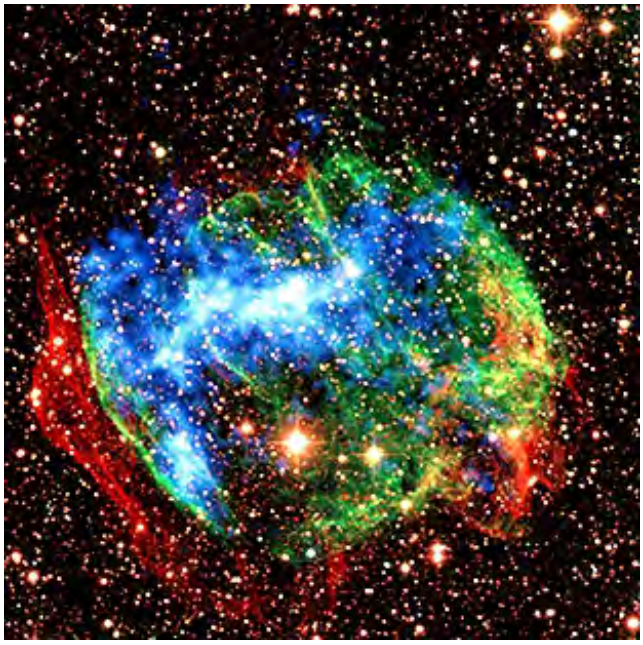


The next figure shows SN1987A seen in the optical range by HST (upper left), by an Australian Radio Telescope (upper right), and by Chandra on two dates (lower left: Oct 1999; lower right: Jan 2000):

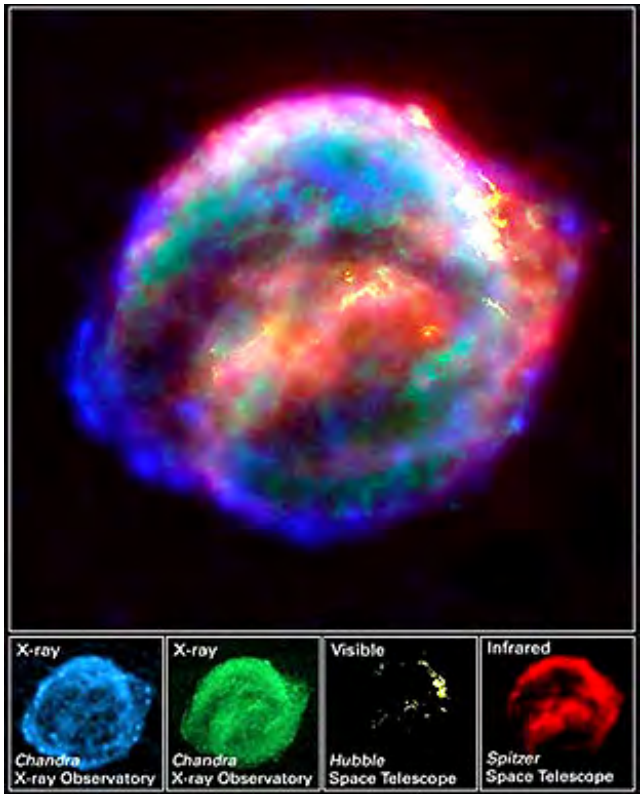


The SN event began about 167,000 years ago, based on distance measurements but its light burst is only now arriving at Earth. The star 20000 years earlier first cast off an envelope of gases as it expanded to a Red Giant. As its core collapsed, it finally exploded violently in seconds, pushing away exterior gases driven by shock waves, and releasing a huge burst of neutrinos as the core protons and electrons were squeezed into neutrons. It is heating of the gases in the ring by these shock waves that has now been producing first a few, then more, of the bright light spots in the ring. In time, it is predicted that the spots will merge and the whole ring will become bright.

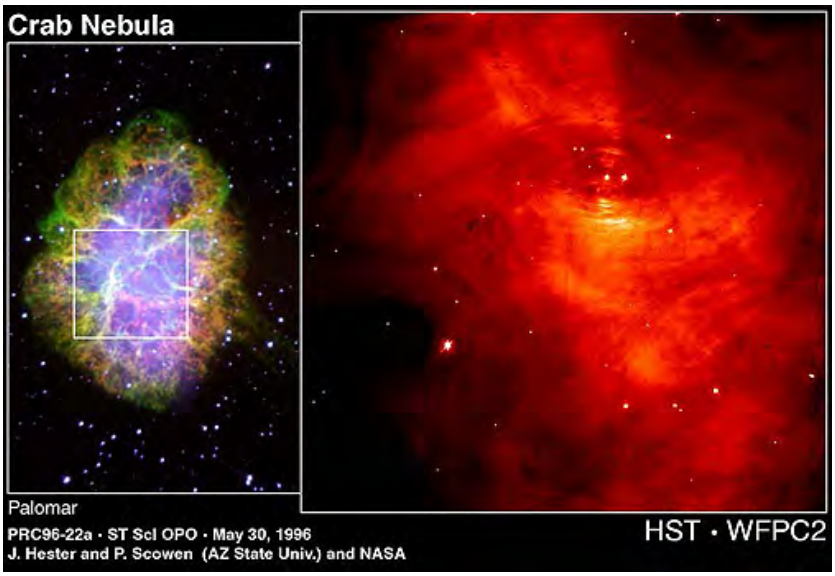
As will be demonstrated by subsequent Supernovae images shown on this page, a SN is one of the most photogenic phenomena observed in the Cosmos. This is supported by this image - one of the most beautiful ever acquired - of SN W49B made from a Chandra X-ray image rendered blue and two Palomar 100 inch ground telescope images registered in the green (Visible) and red (Near IR):



Supernovae in our galaxy and others nearby can appear as very bright light sources sometimes visible to the naked eye. One of the most famed in history is known familiarly as Supernova Kepler, named after the great astronomer who first observed it on August 8, 1604. In celebration of the 400th anniversary of its sudden appearance as a "new star" brighter than any of the planets, astronomers have produced this composite of Chandra, Hubble, and Spitzer Space Telescope images (the one on the right is recording mainly [in red] the dust in the Supernova):



Another prime example of a bright, long-lasting Supernova is this Palomar telescope view of the Crab Nebula (left), with an HST Wide Field Camera view of the volume within the square shown on the right).



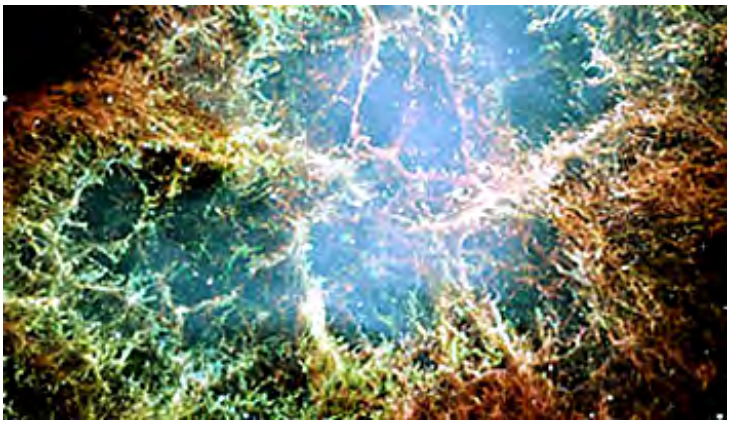
Lets take a closer look at this nebula - perhaps the most studied by astronomers to date - by first showing a ground telescope view made by the CFHT and then an HST view. Both show remarkable detail, once again proving that for closer astronomical objects ground telescopes can compete with the excellence achieved by the Hubble Space Telescope.

The Crab Nebula is famous in history. It was first observed on July 4, 1054 A.D. by Chinese astronomers as a suddenly appearing bright light, seemingly within the Taurus constellation,

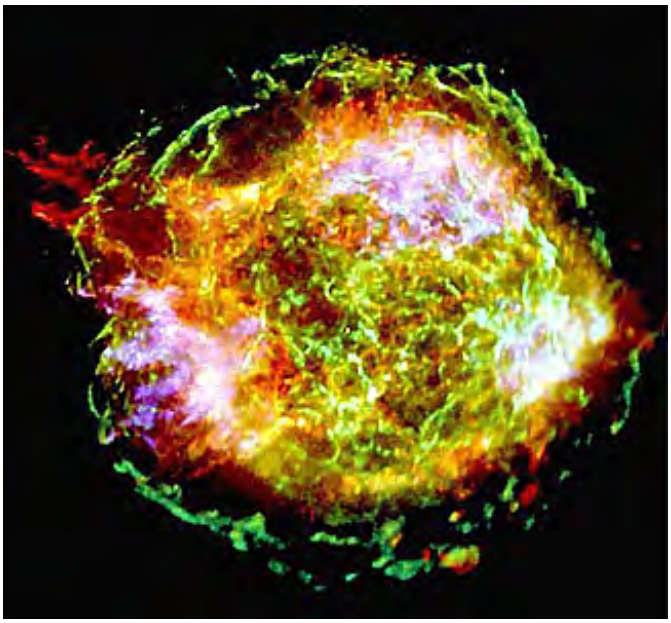
which remained intense enough so that for a few years it could be seen even during the day. Modern telescope views show that filaments are streaming from the explosion center at speeds up to half that of light. This Supernova is, like others in general, an extremely energetic event, radiating from short wavelengths (Gamma rays) through the visible and into the long wave Radio region. A Pulsar-Neutron star (see below), rotating 30 times a second, has been detected in its central region. (Recall that the Crab nebula was imaged in four spectral regions, as displayed on page I-3 in the Introduction).

The Crab Nebula has a notably different shape when imaged with X-ray radiation by the Chandra Telescope. We show this X-ray image combined with a Visible light image made by the HST. A ring structure emerges and a jetlike protuberance extends roughly perpendicular to the ring.

Recently, the HST returned images of the Crab nebula that show the details of the excited gaseous filaments now extending far out into space from the neutron star core. The principal element in many of these filaments is identified by its (process-determined) color: Hydrogen = orange; Nitrogen = red; Sulfur = pink; Oxygen = greenish.

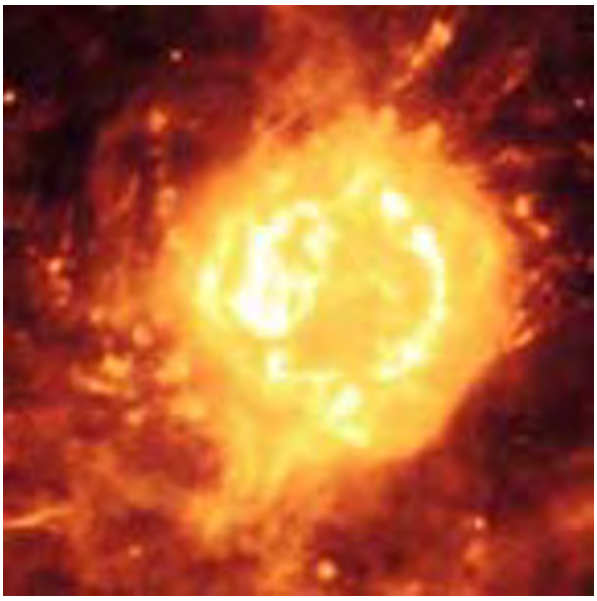


This next image demonstrates how long exposure times can bring out much more detail in a distant astronomical object. Chandra has looked at Cassiopeia A, with an exposure time of 11.5 days. X-ray wavelengths bring out the distribution of Fe and Si, along with other elements. Of special note are the red jets emanating from the still expanding gaseous matter. The surviving Neutron star is not evident in the image.



An HST image shows the filamentous structure of the Cassiopeia A Supernova. The star that blew up in this constellation was about 10000 light years away. Thus, the event took place at that star around 10000 years ago. Historical records note a bright star made a first appearance in the late 1600s in the sky location of Cassiopeia A; what we see today is the dispersal of material after about 400 years. In this rendition, Oxygen-rich clouds of gas/particles are blue; Sulphur is red.

The Spitzer Space Telescope has examined Cassiopeia A in the near IR. This image of the gaseous matter around the burst star looks much like a fireball as we would see an aerial bomb burst on Earth:



However, a surprise greeted investigators when this burst was imaged using a combination of IR bands. Observe this image.

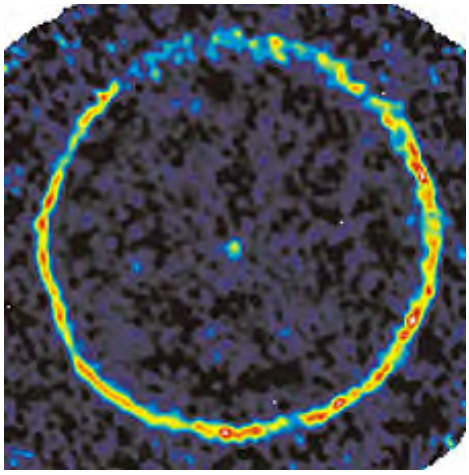
One of the bands used to make the image is centered on 24 μm. The excited gaseous material making up the glowing filaments was determined to be moving near the speed of light, so that pictures taken a year apart show distinct positional differences. This does not fit a simple growth of the explosion nebula that began 324 years ago. The tentative interpretation: the Neutron star that remains after the Supernova can produce "echoes" by repeating blasts that re-energize the gaseous material. Calculations indicate a Neutron star event about 50 years ago that sent a blast wave through the outwardly progressing gas. A mechanism to cause this Neutron star activity is still speculative.

From the preceding images, it should be obvious that Supernovae are the "spectacular fireworks show" that delights both astronomers and the public alike when the resulting images are widely displayed. In recent years, astronomers have become quite adept at spotting a Supernova soon after it explodes and then training a variety of sensors - both ground- and spaceborne - to preserve the high moments of the event's expansion. Here is still another "sensation", SN49, in the Large Magellanic Cloud:

Also in the Large Magellanic Cloud is Supernova N312D. It is 163000 light years away. The image below, made as a composite of HST and Chandra images, shows the extent to which it has expanded and dispersed after 3000 years. By this time the excited gases are diffuse so that the visual stage of the Supernova is no longer dominant. :

Another Supernova example is Eta Carinae, in the 19th Century the second-brightest star in the sky (southern hemisphere) but today too faint to be seen with the naked eye. When processed using a combination of red and UV filter images from HST, the central part appears as an apparent "cloud" of matter which is actually mainly a light burst from this Supernova, now some 10 billion miles across, that resulted from the explosion of a star 150x more massive than our Sun.

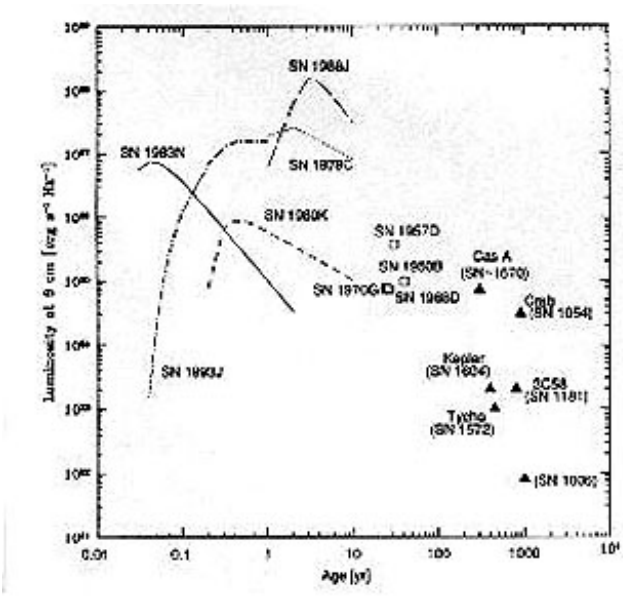
The Red Giant, TTCygni (in the constellation Cygnus), is a carbon-rich star which as it explodes expels carbon monoxide (CO) in a discrete ring that has now advanced to about 0.25 light years from the central Giant.



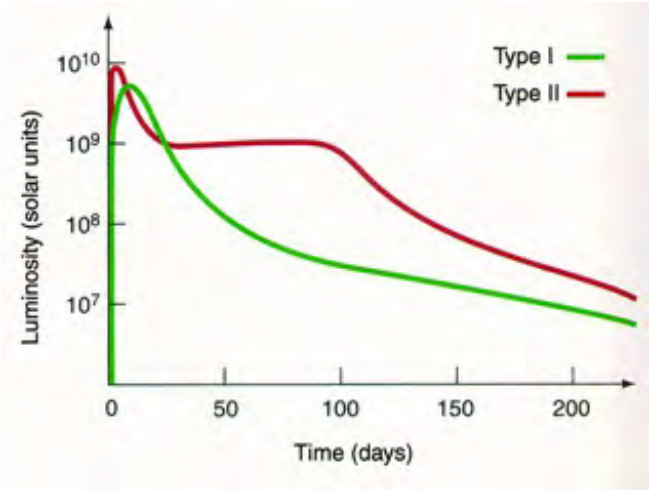
A variant of the gas distribution around a Supernova is sometimes referred to as "a stellar geode", (the term "geode" is an analogy to rocks which contain cavities, usually lined with crystals). N44G, in the Large Magellanic Cloud (160000 l.y. away), is a star which acts like a Supernova to drive surrounding gases into a "bubble" using stellar wind and UV radiation. This HST view uses a red filter to detect Hydrogen and a blue filter to respond to sulphur excitation. The cavity is presently about 35 l.y. in diameter. More than one explosion is suspected (perhaps several Supernovae). This cavitation process is relatively rare.

Once a Supernova is spotted, its rather short history can be monitored in terms of changes in

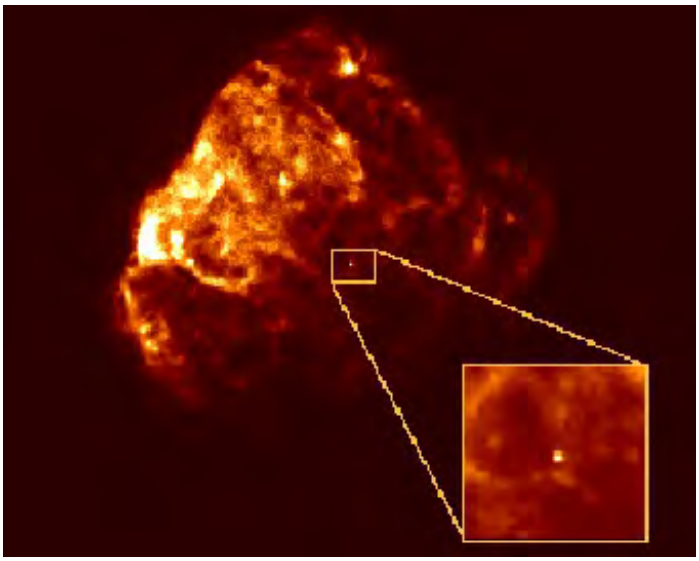
luminosity over time. The graphs below plot brightness variations for several Supernovae of recent vintage and for older Supernova whose remnants are still visible.



Astronomers have distinguished between two general types of Supernovae, separated by the intensity of the luminosity and by the pattern of decreasing light output over time. These are simply labeled: Type I and Type II Supernova. The basis for each type is 1) a Type I Supernova has no Hydrogen in its spectra, and 2) Type II shows Hydrogen in the spectra. Type I is further subdivided into Ia, which results from a thermonuclear explosion of a White Dwarf star, and Ib and Ic, which are caused by collapse of layered massive stars (with iron cores) which then blow up as shock waves (powered in part by neutrinos) expel the layers in huge explosions, leaving Neutron stars if the initial mass was 8 or above or Black Holes if mass was much higher. Type II stars are responsible for dispersion of heavier elements (made in the layers by fusion of initial Hydrogen, then higher atomic number elements, with increasing T and P with depth) into intergalactic space (page 20-7). Type II Supernovae are characterized by asymmetric Type II has proved particularly useful as another "standard candle" - any class of stellar or galactic objects whose (known) intrinsic luminosity (total power output) remains fairly constant at a specific time in their evolutionary history - in the quest to determine distances to far away stars/galaxies and to relate these to rates of expansion. The two types are shown here in this generalized plot:



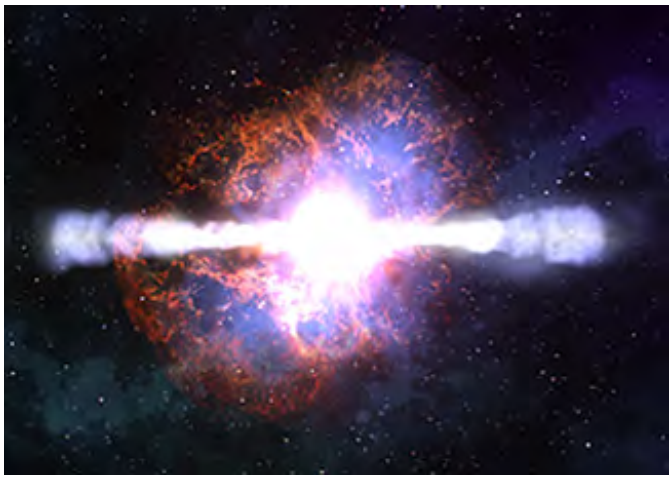
A classic example of a Type Ia thermonuclear explosion is Tycho's Supernova, first observed by the astronomer Tycho Brahe in 1572. Seen below, its gaseous and particulate constituents consist mainly of Silicon, Iron, Nickel, and other heavy constituents. Two prime examples of a Type II core-collapse Supernova are 1987A and the Crab Nebula, shown above. Here is another Type II Supernova, Puppis A, which shows the remnant neutron star:



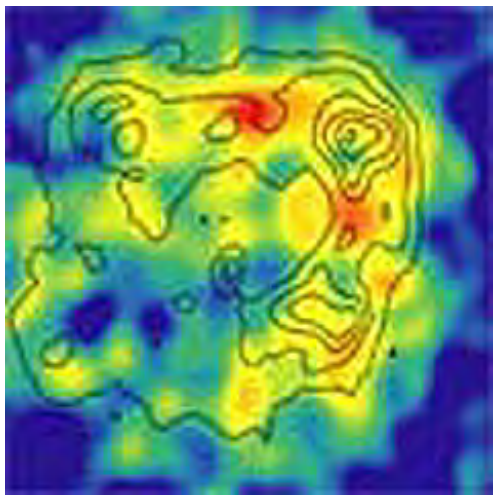
The Ia type Supernova has come center stage in the recent recognition that the Universe is now accelerating rather than slowing down (see page 20-10 where the behavior of this type is considered in detail). Type 1a results when a White Dwarf has grabbed so much matter from a neighboring star (with which it is paired; see top of this page) that it undergoes an implosion followed by a sudden explosion. This event is accompanied by a characteristic spectrum. Type 1a's are less common than the Types I and II; a 1a occurs on average about once every three years in a galaxy.

A star close to the Sun that explodes as a Supernova (or hyperNova; see below) can send shock waves and high-speed particles to distances that could envelop the Earth. This is very unlikely at any given time, such as NOW. But, statistically it is finitely possible, and could be one cause of mass extinctions of life on our planet. A group of astronomers have pointed out that a large number of O and B stars occur in a nearby cluster positioned in the sky near the meeting of the Scorpio and Centaurus constellations. Some ones in this cluster may have passed through Supernovae stages. That Earth may have been affected is implicated by evidence of a deficiency of interstellar matter (including gas) in the so-called "Local Bubble" within which the Sun lies. A consequence of this is that there is less material in our neighborhood that absorbs or impedes light from more distant parts of the Universe; this improves viewing conditions of those cosmic sources. There may be geologic evidence for Supernovae material having reached the Earth: marine deposits dated at 2 and 5 million years are enriched in an iron isotope that would be expelled during a Supernova explosion.

In December, 1997, astronomers observed a localized event in deep space which released more Gamma ray energy at that point than has been calculated to emanate from the entire Universe under a normal state. Because of their similarity to the short-lived, bright Supernovae, such events have been termed Hypernovae, which produce at least several orders of magnitude more energy (10^{53} -10^{54} ergs) than associated with a Supernova (~10^{51} ergs), but they seemingly form by a different mechanism. The initial flare-up may take only a few seconds to actuate but the effects can last for weeks to months. Hypernovae are tied to the destruction of stars with masses at least 20 times that of the Sun. Some HyperNovae seem related to Gamma Ray bursts, described below. HyperNovae were most common in the early Universe when very massive stars underwent rapid burning of their Hydrogen fuel to heavier elements and finally exploded with fusionable fuel was expended. This is an example of a Hypernova - visually it looks like a Supernova but the measured energy release is much larger:

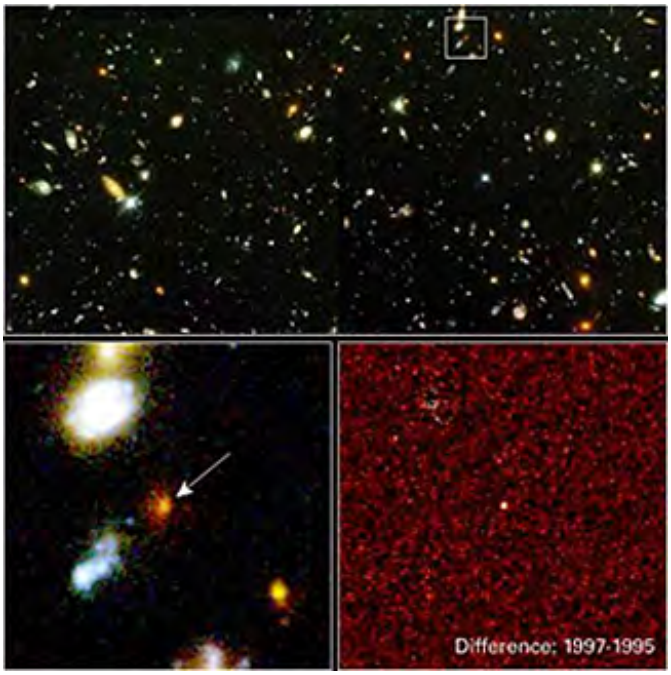


In November of 2004, a group of British scientists announced the results of a sophisticated study of a Supernova that exploded about 1000 years ago which is shedding new insight into the ubiquitous cosmic radiation that permeates space. This event is still growing in the region near our Sun so that its effects have been now carefully documented. The detection system is known as H.E.S.S., for High Energy Stereoscopic System. As presently configured, four Chernkov telescopes located in the mountains of Namibia are tied together in an array. Together, these provide high definition of a form of blue-colored radiation (the Cherenkov effect) caused by the cosmic rays interacting with atoms in the upper atmosphere. This Supernova has proved to be a major source of very energetic cosmic rays (short wave Gamma rays). This strongly suggests the one predominant mechanism for production of such high speed particles is part of the Supernova explosion process. Here is a plot of the large area (about twice the diameter of the Moon, but, of course, invisible to the eye) of the expanded radiation field as picked up by the H.E.S.S. observation system:



A star is in the Aries constellation, about 440 million light years away, experienced massive shedding of material as first seen on February 18, 2006. Its Gamma ray burst phase was especially long - more than 2000 seconds - and powerful - about 25x greater than a typical GRB. Many astronomers believe this event is a precursor to a huge Supernova explosion; telescopes will follow its history for weeks thereafter. Here is a preburst image of the star and then a few days later as the SWIFT satellite detected a ring of radiation flung off the now brighter star:

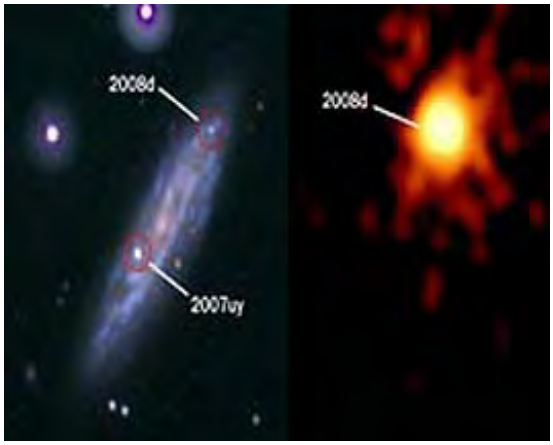
The Supernova process has certainly gone on constantly throughout cosmic time. Logically, one would expect this phenomenon to have acted regularly as far back as the first stars. One of the oldest observed Supernovae occurred at least 10 billion years ago:



In May of 2008, announcement was made of a supernova within the Milky Way that occurred about 140 years ago, in the 1870s. It was not seen optically because it took place within a thick, shrouing cloud of dust near the galactic center. It was discovered using X-rays from Chandra. Here it is:

At almost the same time, another announcement was made of the discovery, quite by chance since the astronomer was looking at another earlier discovered supernova in the same galaxy 90 million light years away, of the first moments of a supernova. Using the Swift space observatory, the beginning consisted of a burst of X-rays. This image shows the galaxy, NGC2770, and the

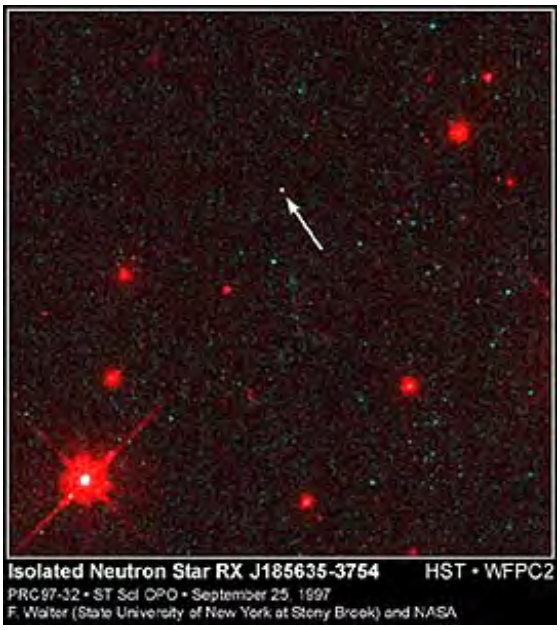
location of supernova 2008d and the earlier SN2007, and the X-ray burst of SN2008 that was detected on January 9, 2008:



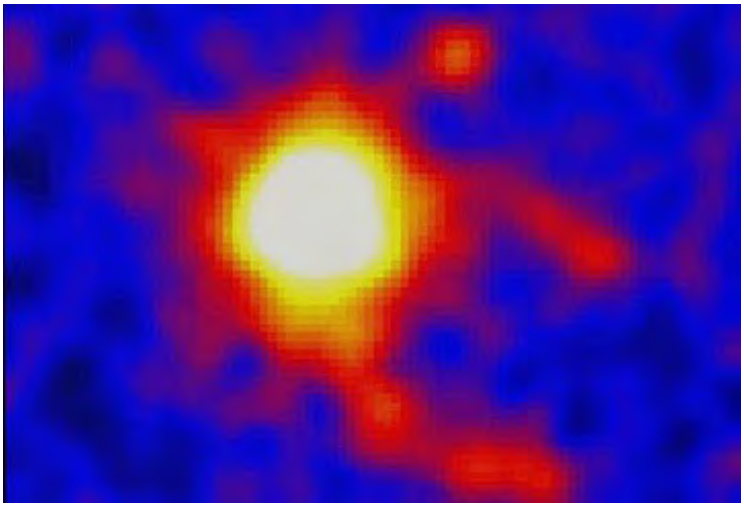
Neutron Stars and Pulsars

The end product of a Supernova event associated with stars greater than about 8-10 solar masses is a *Neutron star* (see also page 20-2a). Such stars develop from strong internal pressures that create neutrons by intense squeezing together of protons and electrons (remember: p + e ---> n); these neutrons are also *degenerate*. (Degenerate matter describes a condition in which the pressures exerted by the mass [as in a gaseous state] no longer depend on temperature but only on the [high] density reached at this stage; the matter is said to no longer obey the classical laws of physics). During the formation of a Neutron star, the prior state star (which may have a core as heavy as iron) develops a degeneration pressure that rises until it is capable of instigating a gravity-driven collapse down to a remarkably tiny size.

This class of stars winds up as small objects only a few kilometers wide but containing matter equivalent to 4-5 solar masses. Their densities can exceed 10^{14} gm/cc (or 10^{7} denser than White Dwarfs). (A feel for this extreme density is gained from this comparison: A volume equivalent to a lump of sugar would contain 100 million metric tons [as measured on Earth] of neutron star matter.) These stars can be detected by telescopes that gather Gamma-ray, X-ray, and Radio radiation. Obviously, being of such small size Neutron stars are very hard to find by optical telescope, even though they can glow with intense radiance, unless they are very near to Earth within the M.W. galaxy. The HST has now provided the first-ever image in Visible light, shown below, coming a Neutron star. It is shining just in front of a nebular dust mass whose distance is just 400 light years away. (The light is produced by processes involving photon escape from a surface whose temperature exceeds 10000°K; the surface area is quite small in keeping with the miniscule size of the star.) The size of this object has been estimated to be only 28 km (16.8 miles) making it the second smallest intrinsically radiating object beyond our Solar System discovered to date by visual means.

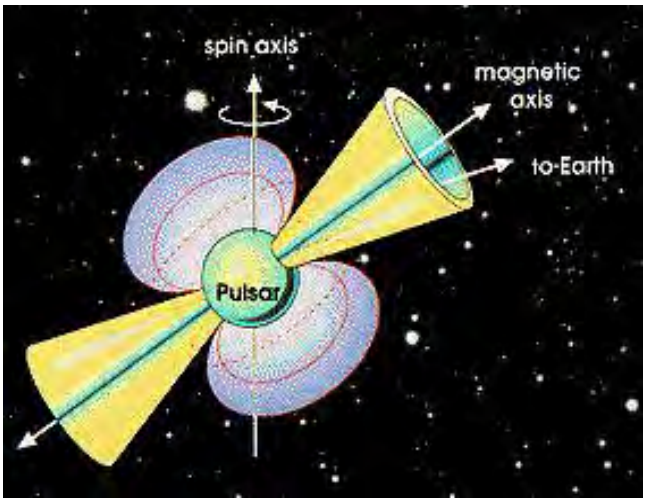


The XMM-Newton satellite has produced a strong Gamma ray image of a Neutron star just 500 light years beyond the Solar System, as shown here:



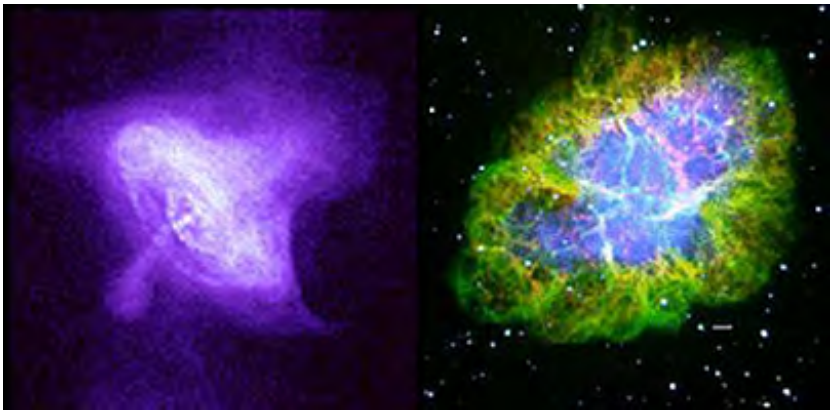
This smallest imaged star, Geminga, is about 20 km (12 miles) in diameter and has a mass 1.5 times that of the Sun. It rotates at a speed of 4 revolutions per second. Its hot surface exudes strong X-rays and Gamma rays, some extending out as filaments (tails), being driven by its huge magnetic field. Electrons and positrons are also a part of the filament, the result of the electric field built by the rotation of the star within its magnetic field. The electrons accelerate outward but some evidence shows that the positrons are coaxed back to the star to settle into hot spots.

Some Neutron stars, called *Pulsars*, are known to have intense magnetic fields and to emit directional beams of strong pulses, best observed by Radio astronomy but also very evident in the X-ray region, in extremely regular intervals (with periods from about 1/1000th of a second to several seconds) whose cyclical nature is related to their (often rapid) rotation; the Earth must lie within the beam's solid angle in order to detect this Pulsar action (the pulses therefore are bursts of radiation from a constant beam detected intermittently from Earth. That is much like a searchlight's beam which, while sweeping continuously, appears to the viewer only when aligned momentarily as it passes through its cycle). This is illustrated by this diagram:

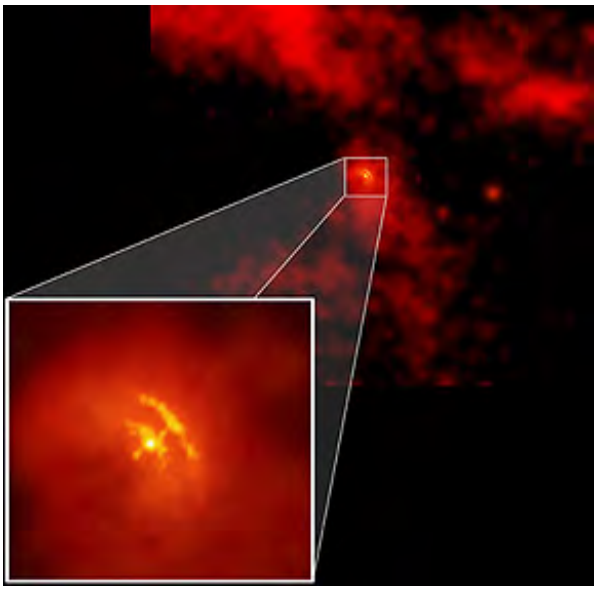


Pulsars are formed by the Neutron star's immense gravity pulling gas from Supernova debris (most Pulsars seem associated with Type II Supernovae), such that this gas is accelerated to a half or more of the speed of light (thus approaching relativistic speeds [those near light speed]); this gas "detonates" when it strikes the Neutron star surface. The magnetic field tends to funnel the fast-moving gas and particles onto narrow parts of the Neutron star's surface which become "hot spots. This releases great quantities of energy extending in the spectrum from Radio to X-ray regions. There are thousands of bursts of energy that rise from the surface many times each second giving rise to the periodicity detected by Radio telescopes and by X-ray observatories such as Chandra.

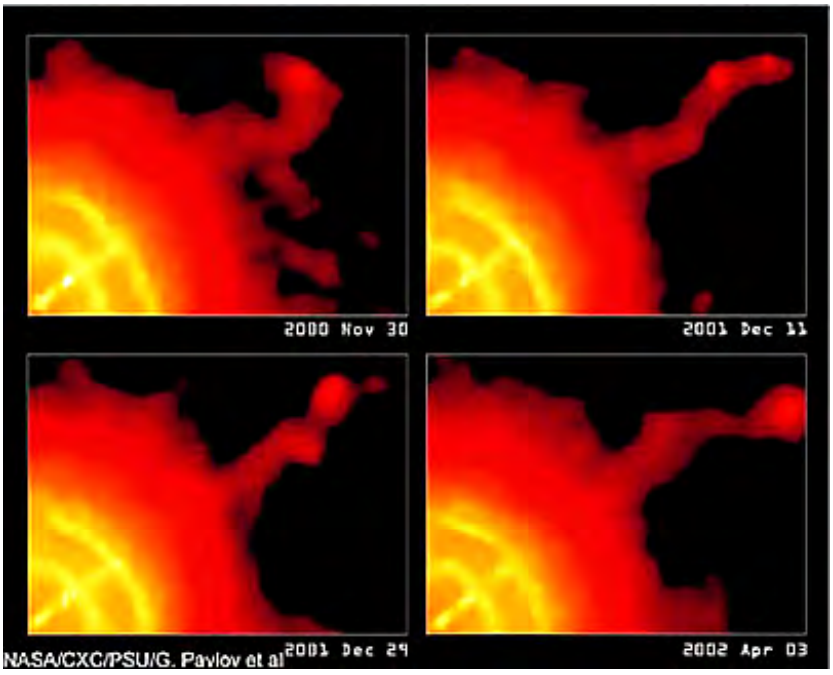
One of the Pulsars that has been extensively studied lies in the heart of the Crab Nebula which we have examined earlier. It shows the development of a pair of short jets of very hot gases that radiate strongly in the X-ray region of the spectrum.



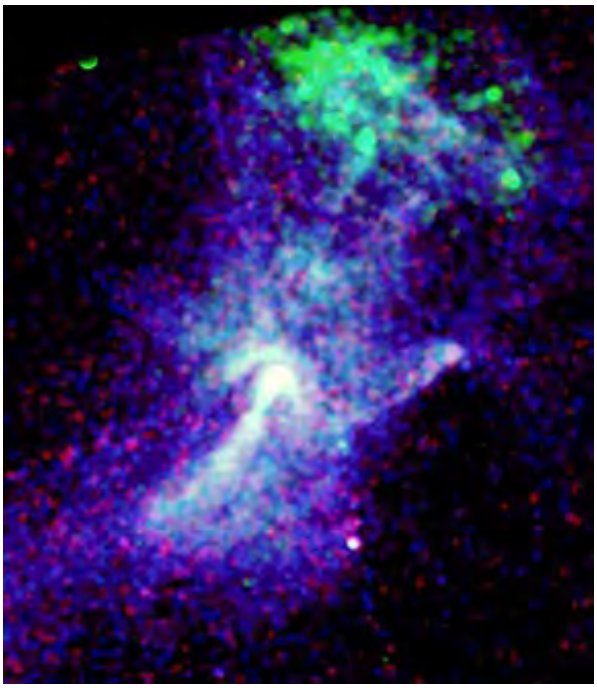
There is a very strong Pulsar in the Vega constellation. Here is a Chandra view of this feature, expanded in an inset:



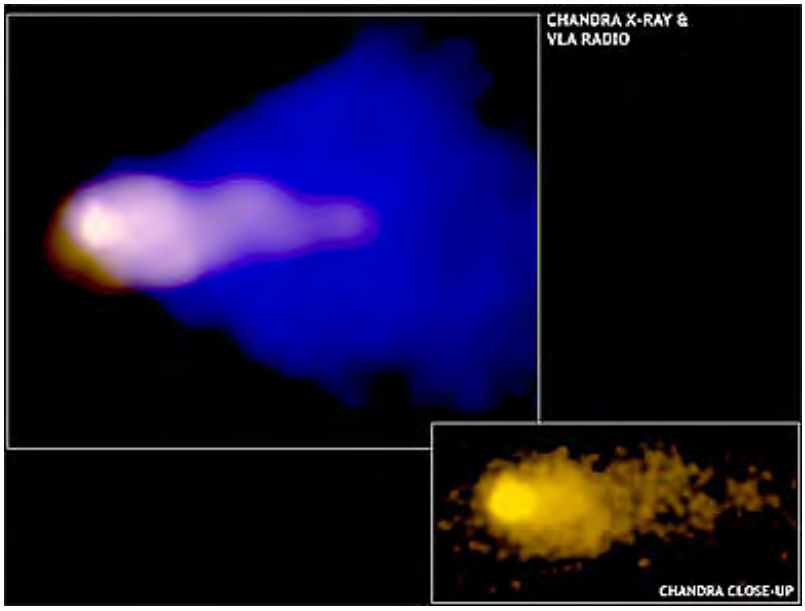
The jet (about 0.5 light years in length) around the Vega Pulsar continually shifts its position and shape, as monitored over many months by Chandra. This bespeaks of a significant variation in the configuration of its driving magnetic field.



Pulsars can be irregular in shape. This asymmetry is mainly the result of unequal distribution of X-ray excited gas around the central Neutron star. This is evident in PSR B1509-38.

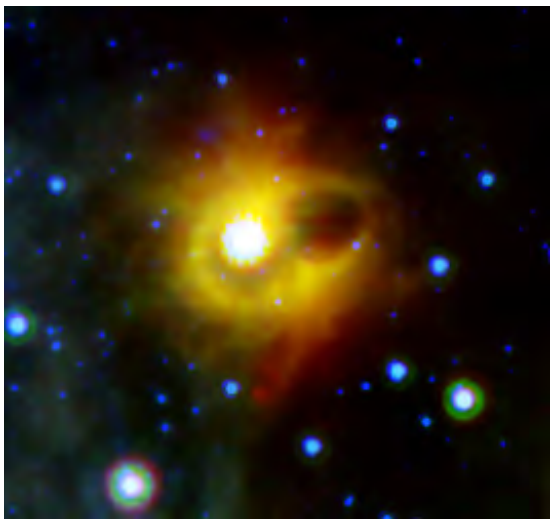


Some Pulsars leave a distinctive trail of excited particles behind them, resembling patterns seen in wind tunnel experiments. These are, in fact, dubbed "Wind Pulsars". The best documented example so far is the Mouse Pulsar, moving at a speed of 2 million km/hr (1.25 million mph) through space. This image combines data from Chandra and a Radio telescope:/p>



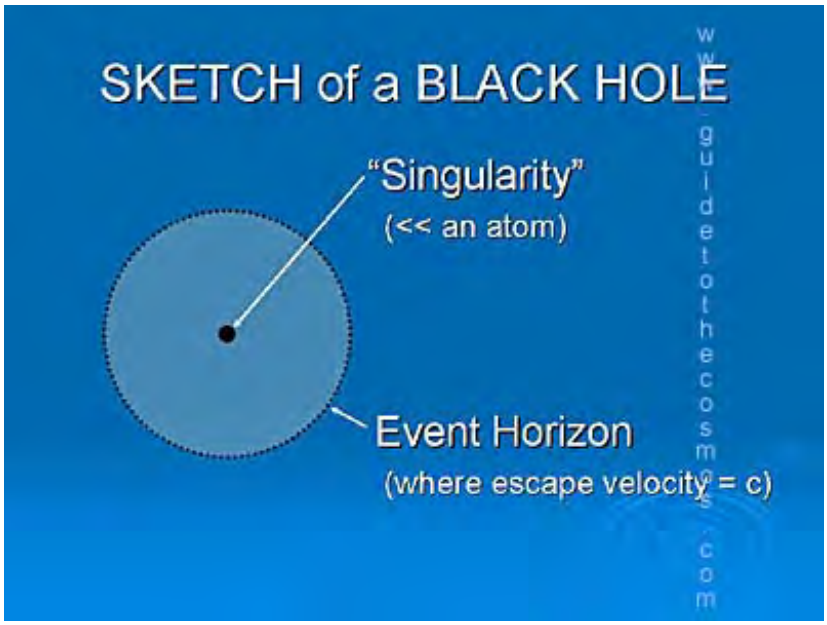
Most Neutron stars have very strong magnetic fields up to 10^{12} gauss (a normal star's field has a strength of around 100 gauss). A rare subclass of Neutron stars is called a *Magnetar*, or more formally, an AXP (Anomalous X-ray Pulsar). Only 15 have been found so far but they are probably much more common (it is estimated that about 1 in 10 pulsars are also magnetars). An AXP has a magnetic field measuring around 10^{14} Gauss (the current record holder, at 10^{15} Gauss, is SGR 1806-20, about 1000 times greater than a typical Neutron star and a million billion times that of the Sun's 5 Gauss). A Magnetar is similar to an SGR (Soft Gamma-ray Repeaters), another Neutron star variant that undergoes periodic variations in energy output. Both AXPs and SGRs are detected by their pronounced X-ray signals. The Rossi Explorer satellite is well-suited to study Neutron stars. The Magnetar N 39 has been examined by the HST; while not directly seeable, its presence is evident in the Visible as a collection of filamentous strands formed from shock waves released when a giant star exploded some thousands of years ago leaving a Magnetar behind.

One magnetar, SGR 1900+14, has been found with a ring 7 light years across. This is how it appears in a Spitzer IR image (the ring appears to be heated dust):



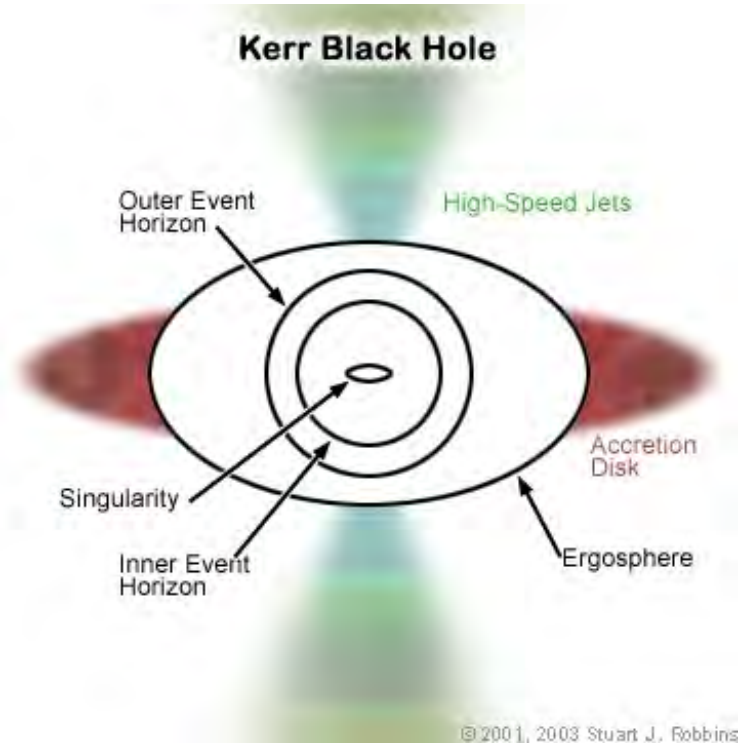
Black Holes and Quasars

As seen in the star evolution diagram on page 20-2, when gases and other matter in stars having solar masses much larger than the Sun gravitationally contract into small, compact bodies, the result is a *Black Hole* (B.H.), so called because the gravity associated with its extremely dense mass (if the mass in the Sun were to collapse into a Black Hole, this would yield a density of about 10^{22} grams per cubic meter; larger stars would produce densities several orders of magnitude greater). The center of a typical small B.H. is much smaller than a single atom. Consider this diagram:



The supermass characterizing a Black Hole prevents all detectable internal radiation originating from within from escaping beyond its *event horizon* (sphere of influence inside of which extreme gravitational forces preclude any mass or photon radiation from leaving). The distance from a B.H.'s center to the horizon is known as the Schwartzchild radius . Since the B.H. is itself invisible (black), its existence must usually be inferred from its gravitational effect on surrounding stars and interstellar matter. Before their observational discovery, Black Holes were predicated to exist from General Relativity considerations. Black Holes indeed have such strong gravitational influence that they notably warp the fabric of Einstein's spacetime dimensionality.

A terminology has evolved to describe Black Holes:



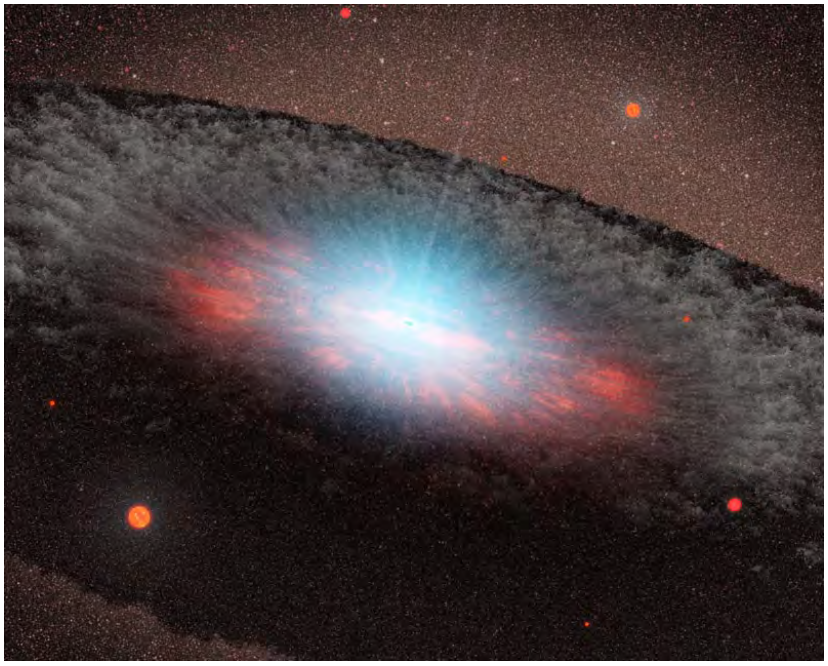
The hole itself is known as a singularity. This is the very center of the black hole, and is where the mass of the original star (and all acquired matter) lies. In a Kerr black hole (a black hole that assumes the star's core was spinning and had a magnetic field when it collapsed), the singularity is theorized to be ring-shaped. In a black hole that does not spin, the singularity is a dimensionless point of infinite density. Moving out from the center, the next part is the inner event horizon. Between the inner event horizon and the singularity, space is believed to be relatively normal - except for the fact that all objects are drawn towards the singularity and

cannot escape. Next out is the outer event horizon. This marks the boundary at which the escape velocity is greater than the speed of light, and all known objects are drawn into the hole. This also marks the "outer edge" of the black hole; we cannot see into it, for no form of known radiation can escape the gravitational pull from this point inward. The ergosphere is a region of space where all particles are drawn in a circular path that match the hole's rotation. However, within the ergosphere, matter and energy can still escape the hole's grasp. The outer edge of the ergosphere is called the static limit. This is the distance that matter must maintain in order to keep a stable orbit and not be trapped by the hole's rotation. Other parts of a black hole are present only in "active" black holes. The accretion disk is matter that has been trapped in orbit around the black hole. It will gradually be pulled into the hole. As it gets closer, its speed increases, and it also gains energy and begins to emit light. The only physical part of a black hole is the singularity. The other terms refer to mathematical boundaries. There is no physical barrier called an event horizon, but it marks the boundaries between types of space under the influences of the singularity.

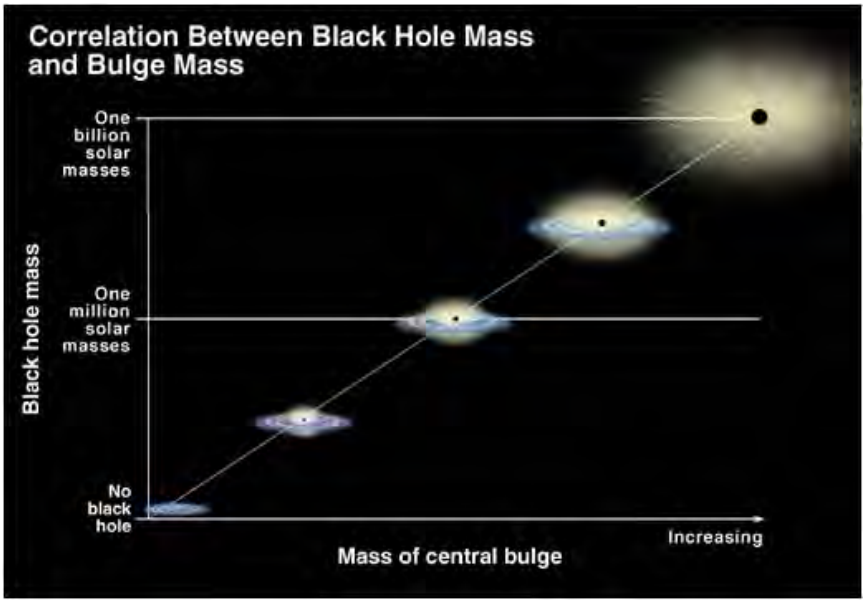
Black Holes cannot be directly imaged, since their supermass prevents light radiation from escaping. But they are usually surrounded by luminous matter excited as it spirals into the B.H. Dust also is concentrated in the region beyond the B.H. The galaxy M64 has such central luminous matter.

When entire galaxies are imaged, the central Supermassive B.H. usually appears as a large bright object (likened to a super star) surrounded by a glowing cloud (which is both gas and closely packed stars) as shown in this infrared image:

The Internet abounds with "artist's conceptions" of Black Holes. Here is one example, showing the bright central location of excited material in a galaxy:

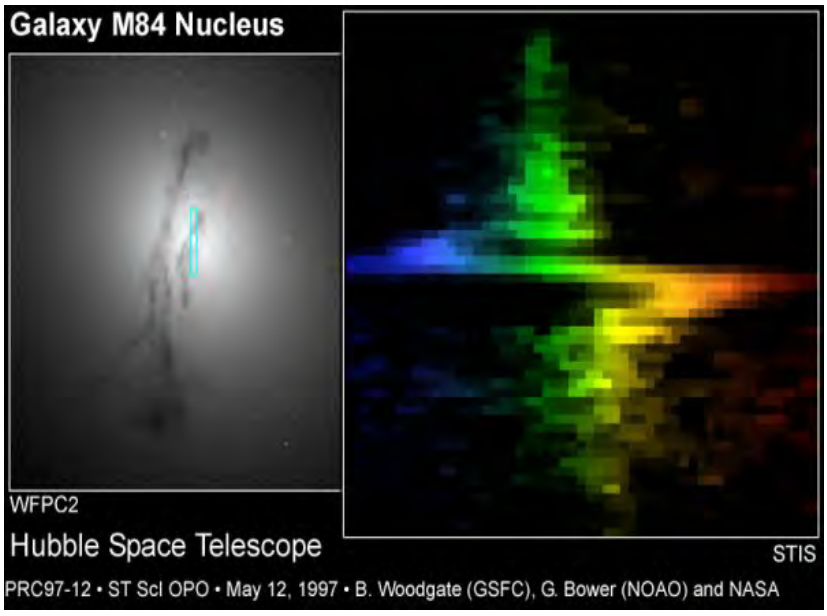


Black Holes can vary in dimensions, the smallest in the general class being much less than a kilometer in diameter but packing dense mass equivalent to about 3 solar masses. (Theory indicates that mini-Black Holes can be as small as a few centimeters or even microscopic in size; so far, none this small have been detected; one school of thought holds that there are countless numbers of tiny B.H's distributed within galaxies and in intergalactic space that contribute to holding the Universe intact.) Black Holes whose masses are similar to that of most stars such as the Sun, and are the products of the final stages of a typical star's life, are known as Stellar Black Holes. The range of mass (with the Sun's mass [solar mass] being the unit of measure) of Black Holes is considerable, as shown in this diagram:



Humongous B.H.s can contain masses derived from billions of infalling stars and galactic matter, attaining sizes exceeding that of our Solar System. These are known as **Supermassive Black Holes**. Supermassive Black Holes are now thought to be the customary objects at the *center* of spiral and other galaxy types, having built up from millions of stars and other matter converging inward as though moving to a drain. Many cosmologists believe that at least one supermassive B.H. is present near the center of any galaxy. Its role is to keep the galaxy from flying apart (thus its gravitational forces are the stabilizer of the galactic structure). The presence of a Supermassive Black Hole, or any size Black Hole for that matter, cannot be confirmed directly, since Black Holes themselves are invisible insofar as being sources of detectable electromagnetic radiation. They are recognized mainly from their effect on stars and gases, causing these to be excited in distinctive ways (such as the Quasars soon to be discussed).

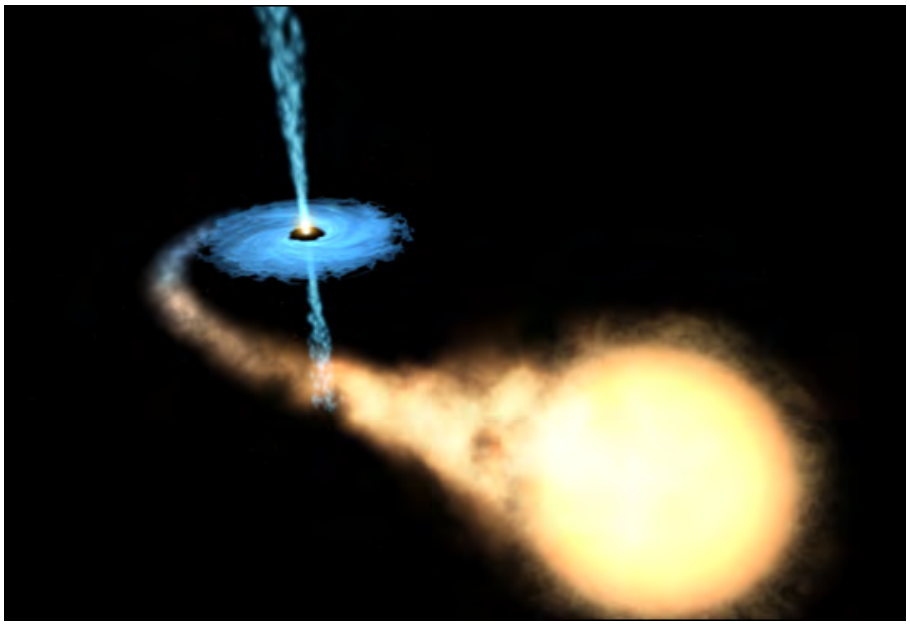
An example of a distinctive detection mode is afforded by the galaxy M84, at whose center a Supermassive Black Hole is postulated. EM radiation coming from its central region shows two opposing spectral wavelength shifts, one a redshift (towards longer wavelengths) and the other a blueshift (shorter wavelengths). This is attributed to the rapid velocities (400+ kilometers/sec) of whirling gases on either side of the central Supermassive Black Hole, as these rotate around the B.H. Thus infalling material is set to spiraling around the B.H. that captures it by intense gravitational attraction. This illustration shows the observed effects:



There can be other visible secondary indications. The HST view of NGC7742, a Seyfert type 2 active galaxy, shows a large glowing central region (an AGN), within which a supermassive Black

Hole is postulated. Its bright center probably represents a quiescent Quasar state, there are periodic flare-ups resulting from energy release when stars spiral past the B.H. horizon into its interior; note the ring of bright hot, largely younger stars beyond and the faint spiral arms further out.

Many Black Holes are the end product of Supernovae explosions of Red Giant stars, as these burn up their fuel and reach the stage where Fe becomes the dominant element (which does not further ignite by fusion; see page 20-7). There are various models of how Black Holes form. These range from explanations of the mode of origin in a binary star system to the origin of central Black Holes in a galaxy to formation of Black Holes in intergalactic space. Illustrative of the origin of a B.H. in a binary system is this diagram; the B.H. results from the explosion of a Red Giant, followed by buildup of matter (mostly gas and dust) in a surrounding accretion disk, and continued supply of new material drawn from the companion star.



Smaller stars end up as Neutron stars which in principle can coalesce into B.H.'s. The larger B.H's have masses from millions to billions greater than the Sun. At the other extreme small B.H's may have only a fraction of a solar mass, perhaps up to a billion tons occupying a tiny volume such that the density of just a teaspoon-full of this compact matter is still enormous. This extraordinary density is possible because under the great pressure that formed the B.H.

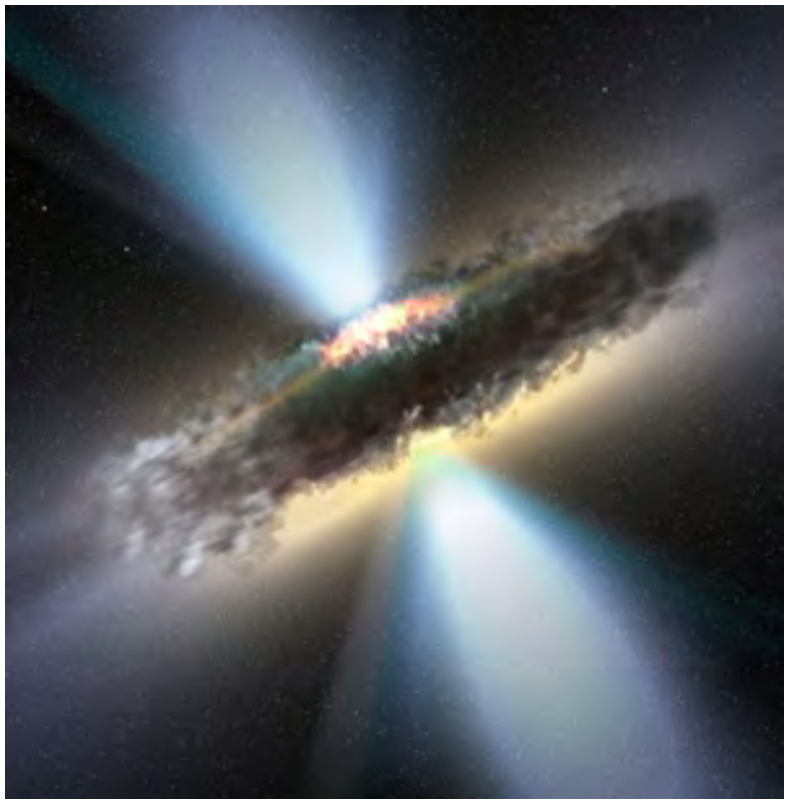
electrons and the atoms themselves become very closely compacted. Smaller B.H's may be ubiquitous - millions of remnants from earlier explosions within the Milky Way and galaxies in general; they may even exist within the Solar System but are too small to affect its spatial fabric and perturb planetary orbits.

A Black Hole generally is so small - yet so massive - that its spacetime expression produces a curvature so pronounced that all internal energy and radiation is seemingly trapped beneath the B.H. (within its horizon). An exception may be *Hawking radiation* (named after Stephen Hawking who devised the theory) consisting of particles created by quantum processes and driven by the gravitational energy within and around the Hole. The mechanism by which this process takes place is an excellent example of "quantum weirdness". In the 'empty' space just outside the B.H.'s event horizon, virtual particles and antiparticles are constantly created (as happens in general in this environment throughout the Universe). Under the strong gravity field around the B.H., one of these particles, the one with positive energy, is likely to be propelled away while the other is captured and dragged into the B.H. Antiparticles have negative energy and those brought into the B.H. react with B.H. particles to reduce the mass of the Hole and thereby lower its gravitational field. This B.H. gravitational field, in turn, loses the energy it provided to make the virtual pair. The escaping particles constitute the Hawking radiation, which is too "faint" to be detected from Earth but nevertheless causes the B.H to slowly "evaporate".

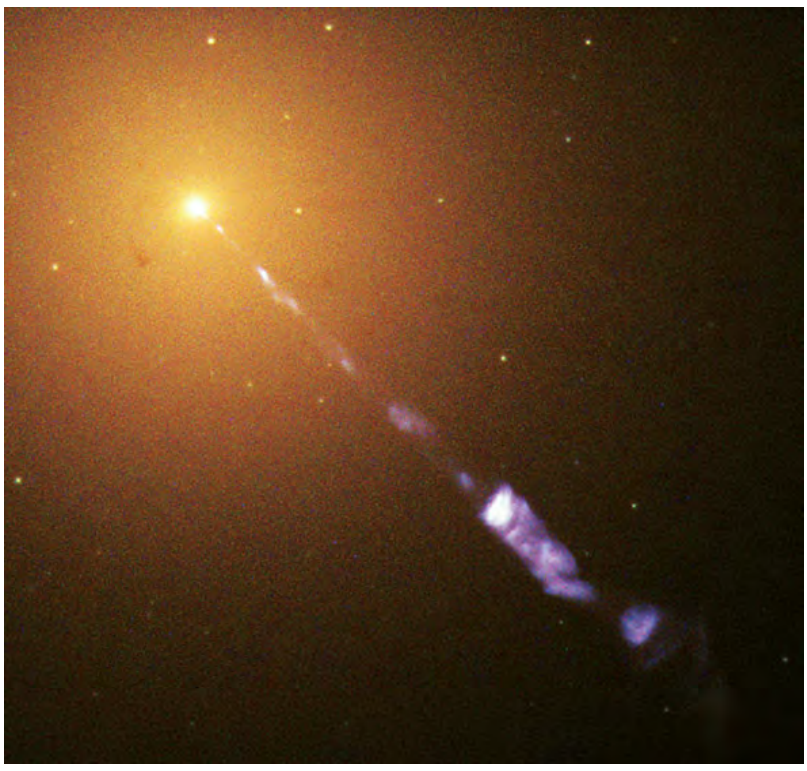
This escaping (emitted) radiation is most effective for tiny Black Holes and provides a means by which they can dissipate over extremely long times through this evaporation. While based on sound theoretical reasoning, Hawking radiation has not yet been directly detected. But if it is proved to exist, it provides a mechanism by which countless numbers of small primordial B.H.'s that formed at the outset of the Universe, because gravity was so intense then, have since vanished. At the present time, astrophysicists are learning more about B.H.'s by computer modeling and simulating their behavior.

Black Holes are also capable of ejecting matter in jets or streams of particles moving in beams almost at the speed of light. (Jets also occur during star formation and during late stages of star death). This next image, made by the Swift telescope, shows the star, around which a Black Hole is orbiting, and two distinct jets.

On a galactic scale, perhaps the best known image to date that shows distinct jets of luminous material emanating from a central Black Hole in two opposing directions perpendicular to the galactic plane is this artist's sketch based on a Chandra image of NGC1365; the galaxy itself appears here as a dark dust/gas cloud obscuring its many stars.:

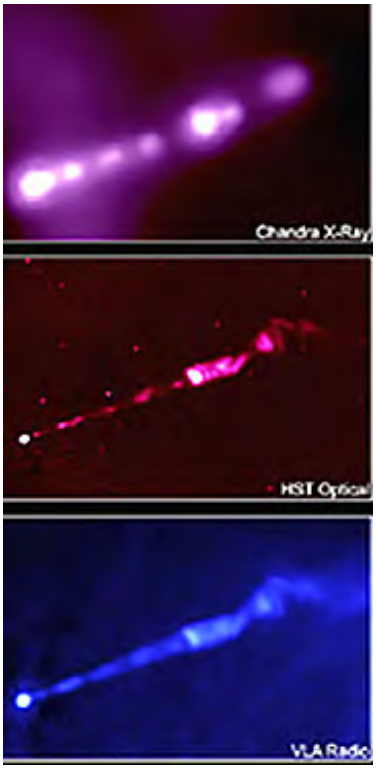


A single jet of high speed luminous particles from a central Black Hole in the M87 galaxy. Here is a Fermi gamma ray view of the well-known galaxy, M87, in which its billions of stars are not resolved so as to appear as a yellowish-red glow. The central "star" is actually light emitted from the exterior around a B.H., probably as a Quasar.



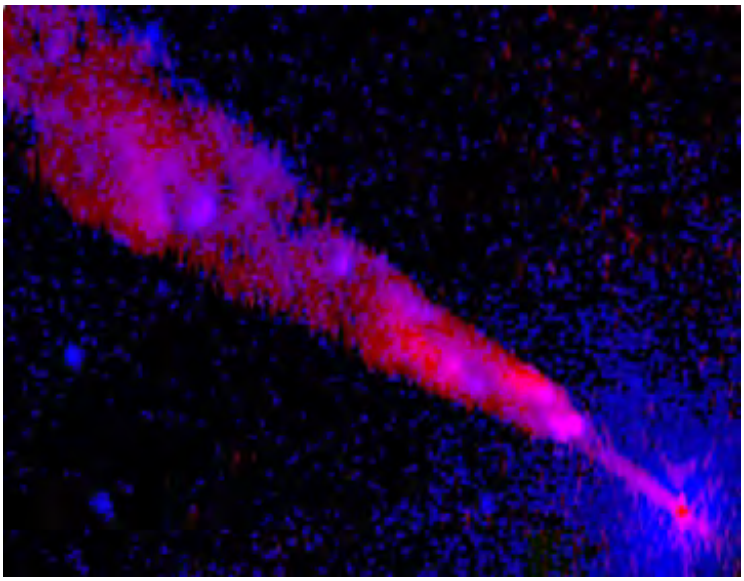
Below are three more views of the jet streamer from M87; the top is imaged by Chandra in the X-ray region; the center is visible light; and the bottom from Radio waves. The origin of such

streamers, found also associated with other galaxies, is still imperfectly known. But, the Black Hole(s) causing this ejection of gas and particles are the source of strong, directional electromagnetic fields. The gases may be excited by synchrotron radiation, causing photons whose energy levels extend over most of the Electromagnetic Spectrum.

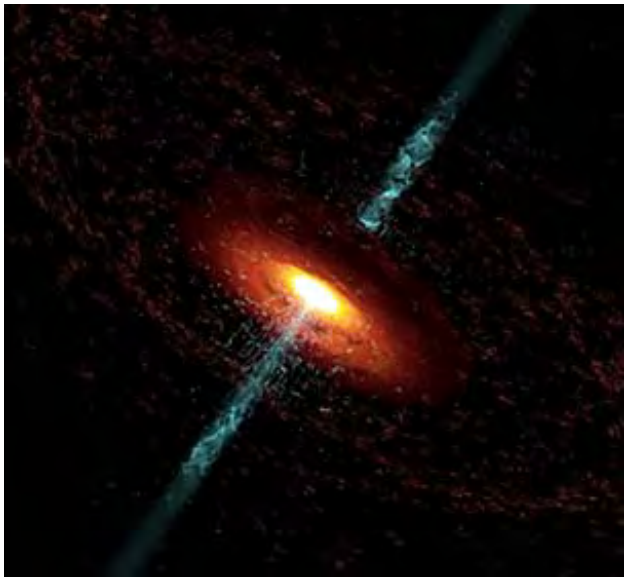


Still another example of a jet associated with the presumed central Black Hole in a galaxy is Centaurus A (NGC5128) located some 11 million l.y. from Earth. On one side the jet is obvious but it has a faint companion on the other side. This jet pair lines up with the axis of rotation of the galaxy. The image, made by Chandra, is converted to a visible view using data sensed in the X-ray region of the spectrum:

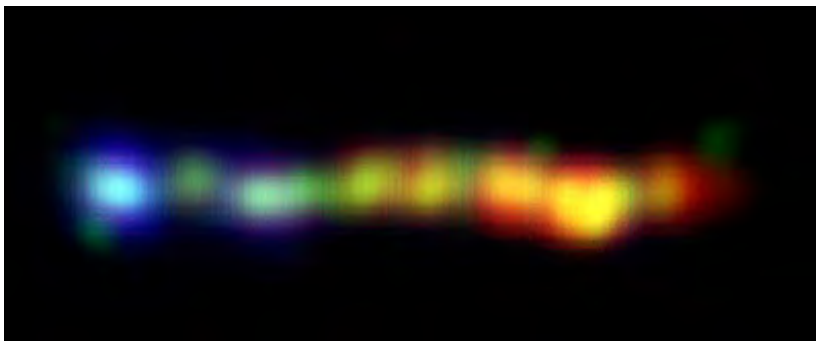
This jet, which follows magnetic lines, is even more splendidly displayed in an HST image that is combined with a Chandra image, as shown below, with the strongest X-ray signals shown in blue:



Probably the best image to date of a paired jet associated with a supermassive Black Hole is that recently captured by HST as it trained on Quasar 3C120; the jet, composed of X-rays and electrons, follows strong magnetic lines:



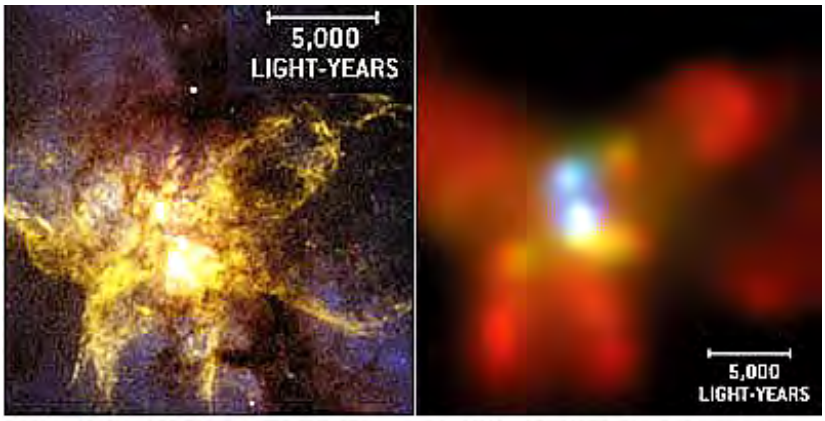
A jet emanating from Quasar 3C273, which is powered by a Super Black Hole, extends out 100000 light years, as seen in this color composite made from a Chandra X-ray image (blue), a HST Visible image (green), and a Spitzer IR image (red):



These jets may be the same phenomena commonly detected by Radio astronomy, as illustrated near the bottom of page 20-3.

Galaxies are thought to have multiple Black Holes, most of which are relatively small. Evidence is growing that most spiral galaxies, at least, have one or more large Black Holes in their central regions. Such B.H.'s eem to serve as a stabilizing influence on the maintenance and evolution of a galaxy, causing stars and stellar gas and dust to migrate inward and be dragged into the Hole. Matter is constantly being attracted into the B.H. such that over time all of the galaxy will converge into the B.H. and thus be wiped out.

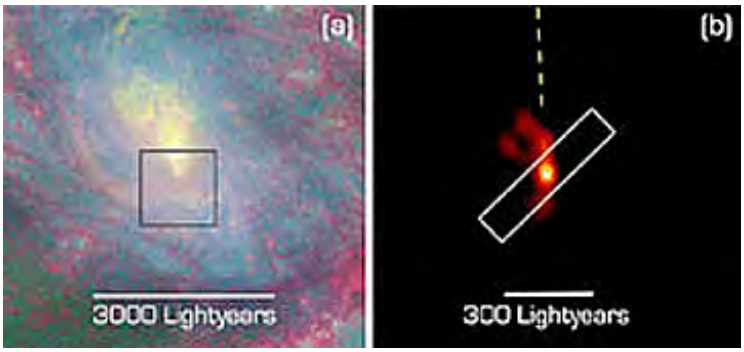
The first case in which two supermassive B.H.'s occur in the central core of a galaxy has been found in NGC6240. This irregular galaxy is shown in visible light in the left HST image below; the Chandra image on the right indicates a pair of Black Holes, which create a strong X-ray signal (in blue; weaker X-rays in red and yellow) as infalling material is heated to very high temperatures. Astronomers predict that these B.H.'s will eventually merge by collision.



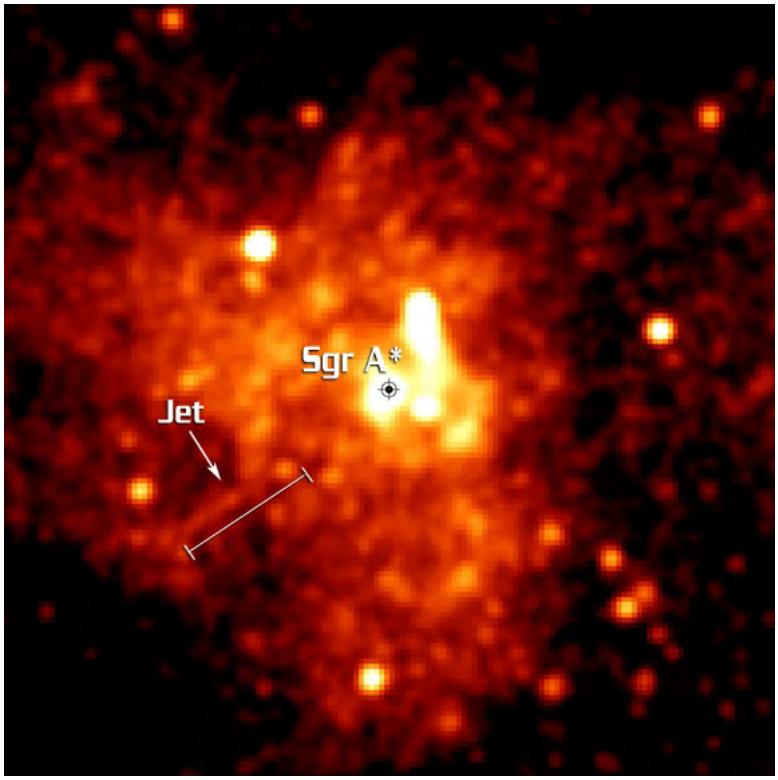
NASA/ROELAND P. VAN DER MAREL and JORIS GERSSEN *Space Telescope Science Institute*

NASA/CXC/STEFANIE KOMASSA *Max Planck Institute for Extraterrestrial Physics*

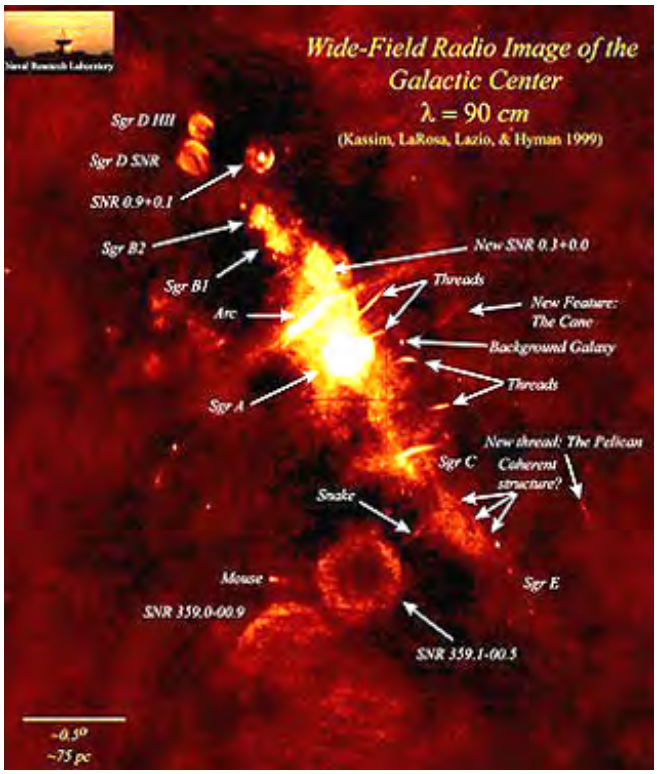
The amount of gas and dust surrounding a central Black Hole can be much greater than farther out in a galaxy. In NGC 1068 (also known as Messier 77), the HST image on the left of this next figure shows reflected light from the dust in blue, ionized Oxygen gas in yellow, and ionized Hydrogen gas in red; on the right is a Chandra image of its center in which the orange-red corresponds to highly energized material around the immediate B.H. that is emitting X-rays.



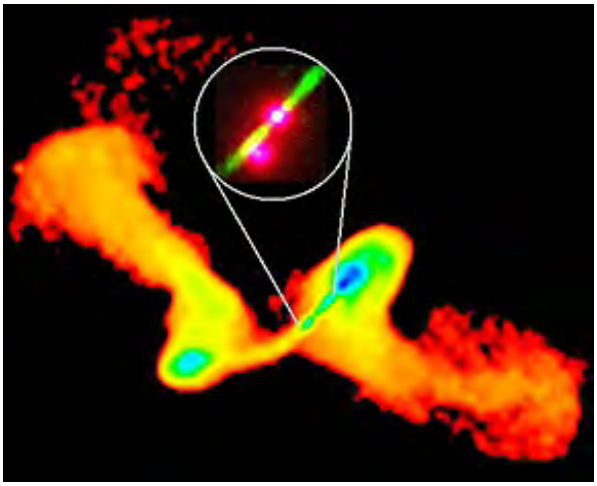
Chandra and Radio telescopy have now established that the central region of the Milky Way galaxy has a supermassive Black Hole (located in the celestial hemisphere at a point close to Sagittarius A); perhaps there is more than one in this inner part. Proof of the presence of a large B.H. in the M.W. was hard to come by, because the central region is shrouded by dust. As Infrared images of this region accrued, the stars within the dust region were imaged. This allowed determination of their orbits as time lapse views permitted calculation of their movements. Many stars showed just the pathways expected from theory that would occur if a B.H. were sucking in these bodies. Later surveys of the central region using Gamma ray and X-ray radiation to image the behavior of the B.H.-seeking stars confirmed the presence of a very large B.H. at the centerpoint of the Milky Way. However, as this Chandra image discloses, the strongly radiating dust and gas at X-ray wavelengths does not single out Sagittarius A or any other manifestation of the B.H.



The size of the M.W.'s central Black Hole has been hard to determine because of this masking matter. It mass has been estimated to be about 2.6 to 4 million solar masses. Early estimates placed the diameter of a sphere to its event horizon at about 1.5 to 23 million kilometers (1 to 14 million miles). Recently, studies done by penetrating Radio waves, using Radio telescopes, has shed light on its dimensions, so that the upper value is considered close. This Radio wave image shows Sagittarius A, a bright object that may be the glow of excited radiation around the B.H.:



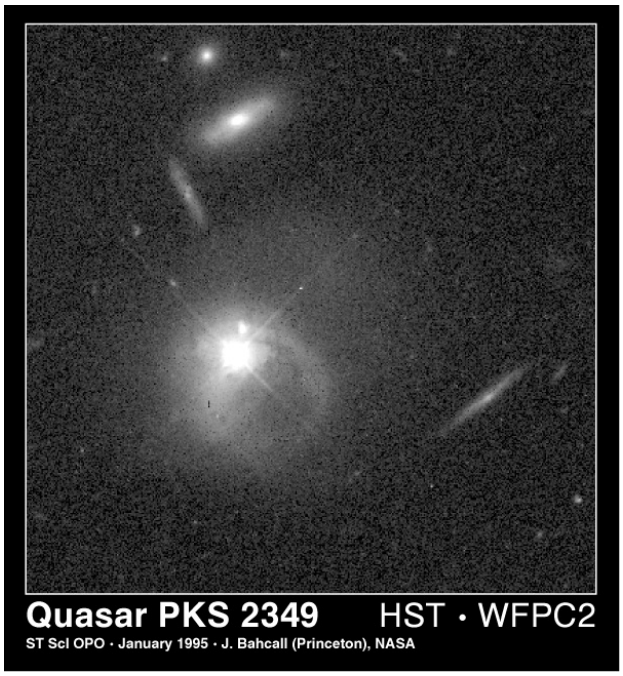
In principle, Black Holes should sometimes collide (but the consequences are not yet defined explicitly), especially when two galaxies collide with the B.H's at their centers then interacting. Evidence for this is sparse. However, such an event is postulated for the observations by the Wide Field Camera on HST of NGC326. In the main view below is the pattern of jet lobes from that galaxy seen a few years ago. In the offset second image is a more recent observation in which the orientation of the principal jets has now shifted more than 90°. The favored explanation is that two Black Holes have now interacted causing the spin axis of one to shift notably.



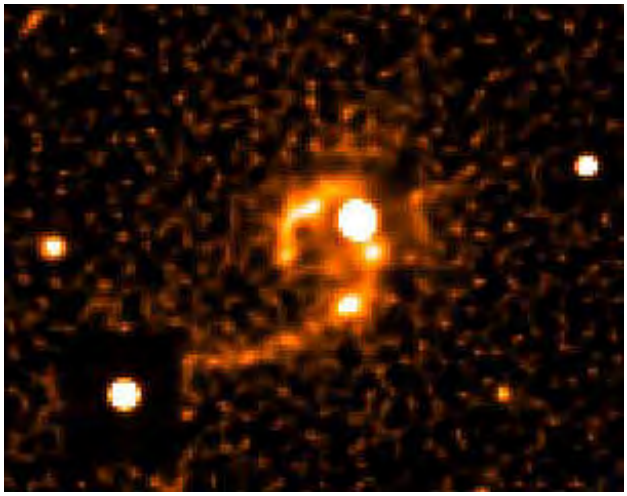
As implied above, Black Holes play a large role in the life of a galaxy. Recent UV observations by the Galaxy Evolution Explorer (Galex) finds a broad relationship between B.H. size and galaxy size (the larger the first, the bigger the second; the study was confined to elliptical galaxies but probably holds true for spiral ones as well). It was also found that for large central B.H. galaxies, the production of new/young stars in the inner regions became distinctly sparse. This has been attributed to B.H.-controlled heating of Hydrogen gas to temperatures too high to form stars and/or to expulsion of the gas from the inner galaxy.

A Black Hole's incredible gravity pulls in particles from outside the event horizon until their velocities are accelerated to nearly the speed of light. Matter is literally torn apart upon entering the Black Hole. As these particles close in, monstrous energy releases produce continuous bursts of energy outside the horizon, a process believed responsible for most *Quasars* (a

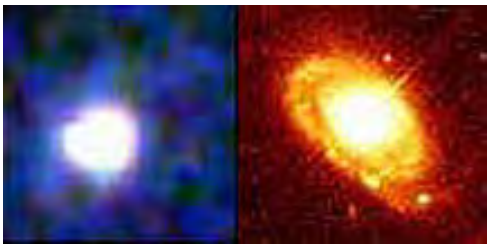
contractive term for "quasi-stellar" to describe a star-like appearance even though the observed feature is not a single star). Quasars are extremely bright objects (very high luminosity, comparable or even exceeding that of an entire typical galaxy), being considered by most astronomers to be the glow of radiation bursts ("hot spots" of Gamma radiation, X-rays and Visible light) from both stellar and interstellar matter continuously infalling into the central regions of active galaxies, whose cores are probably supermassive Black Holes. While the majority of Quasars are located at or near a galaxy center, some occur in the spiral galaxy arms or in the regions beyond an elliptical galaxy's core. They were initially discovered as intense Radio wave sources detected by Radio telescopes. Now it is known that most Quasars are not accompanied by Radio waves (less than 2% are dominantly Radio sources, in which that wavelength region marks energy developed by synchrotron radiation) but are instead sources of more intense, shorter wavelength radiation. Here is an HST optical image of one (and possibly several) Quasar(s):



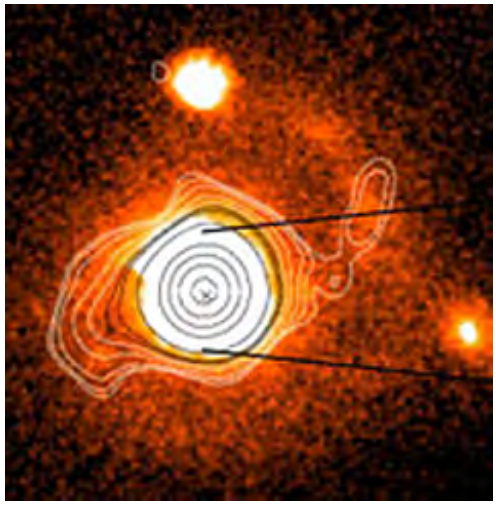
Quasar HE 1013-2136 at a distance of 10 billion l.y., imaged by an ESO telescope on a Chilean mountaintop, seems to be drawing gases from a galaxy to the left:



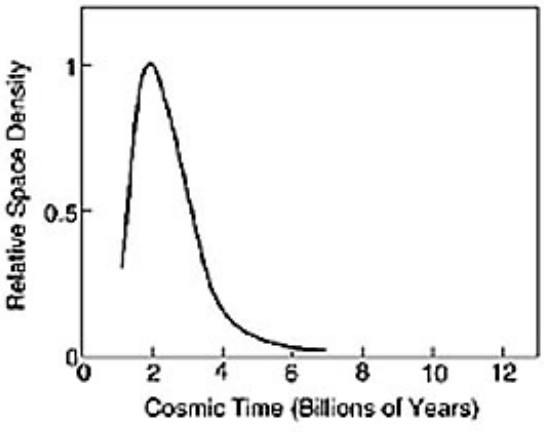
This pair of images shows a Quasar in Visible (bright in the blue) and Infrared light.



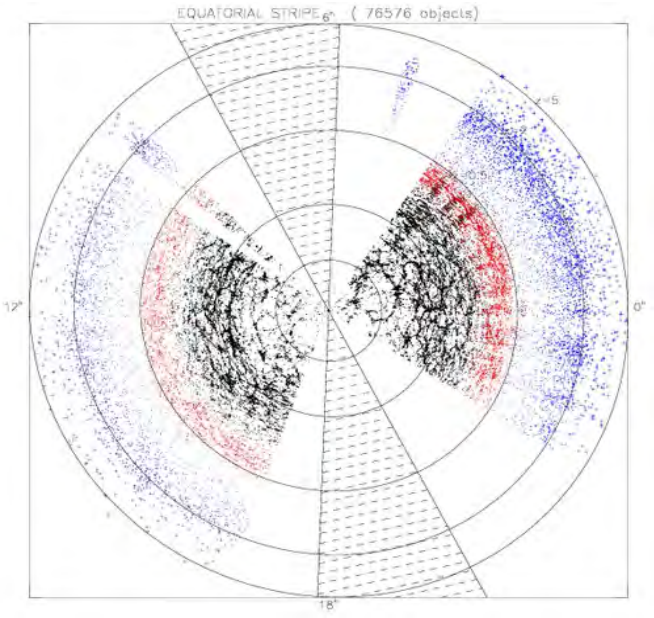
The powerful Quasar qso 1 Zw 1, as seen in the Infrared, is also a strong Radio source (contours superimposed).



Most Quasars are so far away (but some more recent ones are nearby) that light arriving at Earth left the Quasar source when the young Universe was only about 1/4 to 1/6 its present size. Thus, most (estimates in excess of 75%) Quasars formed early in Universe history and many, particularly the larger ones, have since become either greatly diminished ("dormant", with occasional flare-ups) or are now extinguished in today's time frame. This generalized (smooth) plot of Quasar history, both in terms of time since the Big Bang and when the numbers of galaxies relative to the expansion size of the Universe are normalized to 1 (maximum), illustrates these points:

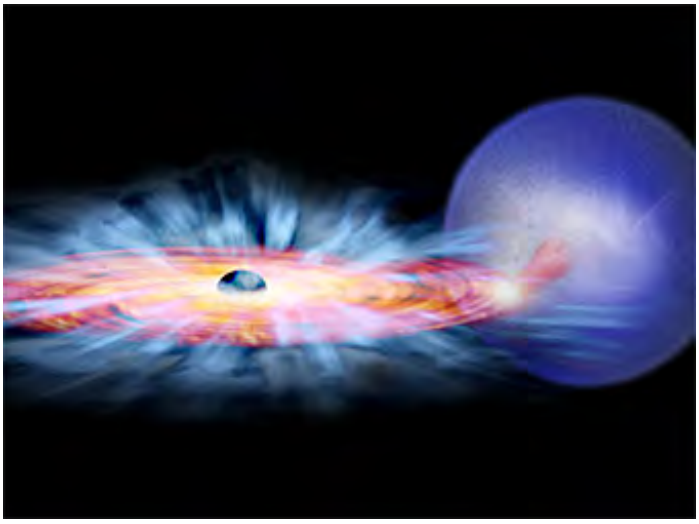


This distribution of quasars in time is also evident in this Sloan Sky map portrayal, in which the outer blue dots are Quasars (at that distance they are young in terms of Universe age), the red dots are Luminous Red Galaxies (clusters of elliptical galaxies), and the black dots are closer to the Milky Way and are relatively old:



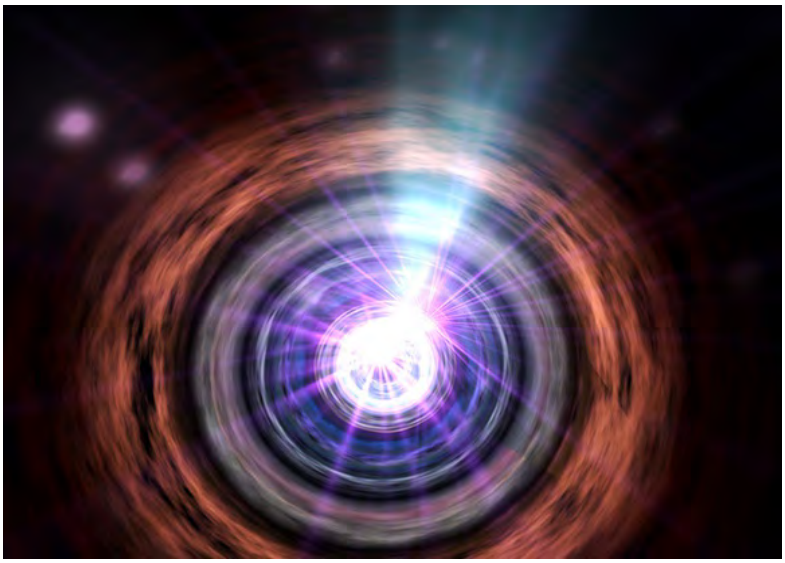
Since Black Holes can still form in young cosmological time, i.e., recently, throughout the Universe, conceivably they are giving rise (usually after only millions of years) to new Quasars. Quasars are made visible because of emission of light resulting from energy conversion as stars and interstellar gases are gravitationally sucked into supermassive Black Holes.

Perhaps as much as 50% of the EM radiation in the Universe is related to Quasars around Black Holes. The Quasars result from material being pulled off nearby star(s), transferred as stellar winds along magnetic lines from the stars, and accumulating in a disk around the Black Hole. A study of Chandra data for J1655 leads to this pictorial interpretation:

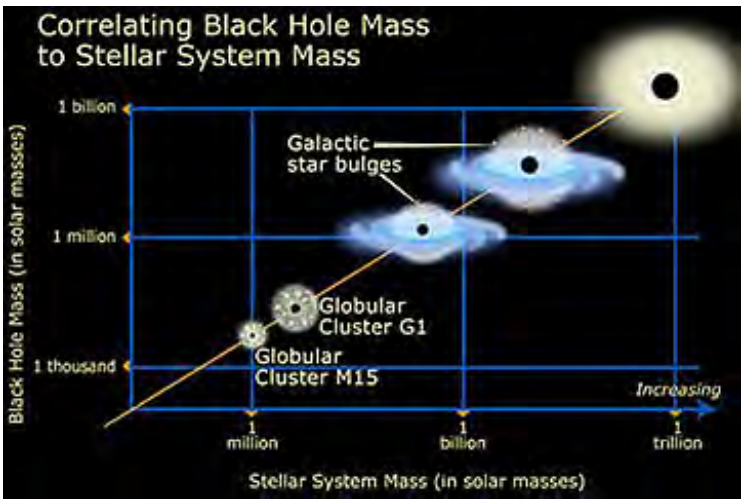


This may be the most common mechanism for Quasar production.

Very energetic quasars emit their radiation primarily in the Gamma Ray segment of the EM spectrum. Such quasars are called *blazars*. Here is an artist's conception of a blazar, based on telescope observations, in which there is a jet moving out of each end; in this view, the direction of look is straight down parallel with the central jet.



Black Holes that occur outside galaxies, or in a star-sparse region within a galaxy, do not attract enough material to become readily visible by virtue of the excitation of incoming matter. But their presence is often suspected where an X-ray or gammma-ray source is observed without a corresponding visible body. Recently, Black Holes have been detected in Globular Clusters by analyzing the patterns of movement and velocities of stars that can be resolved in the assemblages making up the clusters. These B.H.'s have estimated masses intermediate between the small isolated ones mentioned above and the Supermassive ones described in the previous paragraph. Although the numbers of points in the following plot relating B.H. mass to stellar assemblage mass are still few, a general trend that fits size to a straight line is evident:



In the early years after first postulated and then discovered, Black Holes were treated almost as a curiosity, without any special importance in the initial phases of the Universe's history. But, with the discovery that most (if not all) galaxies have B.H's in their core, there is a growing belief among astronomers that they are the necessary starting point in the formation of a galaxy, serving as the nucleus or core that attracts the matter that eventually organizes into a galaxy. Recent reports of both observational and theoretical studies now offer two important ideas: 1) both Black Holes and Neutron stars are more abundant in the inner or central part of a galaxy - a fact related to the idea that massive stars tend to form more readily in the core region; and 2) in early cosmological time Black Holes had a definite symbiotic relation to the processes that form and develop galaxies, i.e., massive B.H.'s can serve either as a nucleus for a growing galaxy or at the least aid in gathering matter into organized gas clumps that evolve into primitive galaxies.

In some respects, the smallest Black Holes are an approximation to the supersingularity postulated as the starting point of the Big Bang except that they have finite dimensions of meters to several kilometers and even much larger for those in galactic centers depending on their amounts of mass (can be equivalent to the cumulate mass of hundreds of millions to

billions of Suns). One theoretical class of Black Holes consists of concentrations of extreme densities collected in "points" as small as 10^{-15} meters.

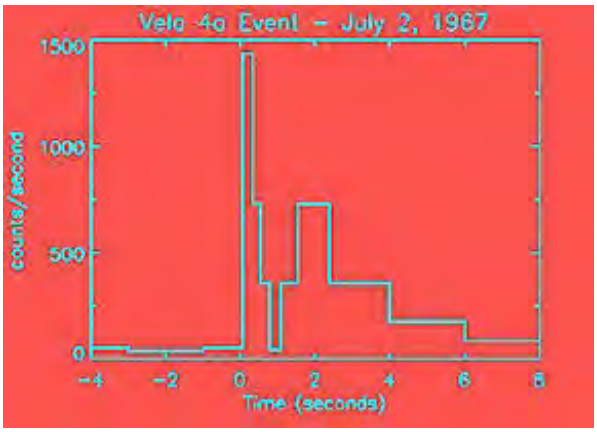
Some Black Holes are thought to be the sole surviving remnants of galaxies that have been completely swept into them. Other Black Holes may have formed during the first seconds of the Big Bang. There are increasing indications that supermassive Black Holes were in existence within the first billion years of the Universe. Many of these are either relics of the B.B. or remnants of early Supernovae.

Speculatively, one future outcome for the Universe (depending on the ultimate mode of expansion [see page 20-8]), after 50 b.y. or so, could be a collection of billions of Black Holes that eventually converge upon themselves to coalesce into a single ultra-dense Black Hole that ultimately would become the singularity for the next Universe (in this model, any number of successive Universes, exploding and contracting cyclically, is feasible). Such a concept of repeating Universes (treated in more detail on page 20-10) is referred to as the "Big Crunch", or even more colloquially, as the "Bounce" in reference to the repetition of an explosion after total collapse to the B.H. singularity.

Gamma Ray Bursts

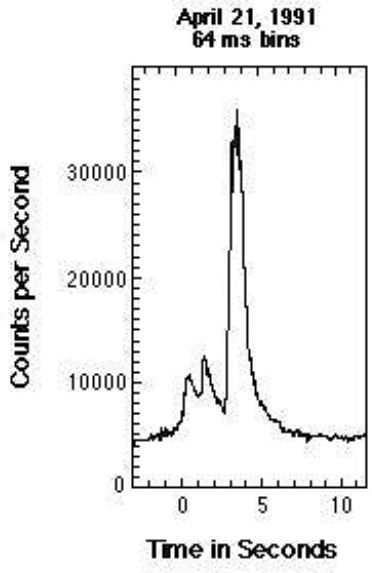
Black Holes almost certainly play a role in what are called Gamma Ray Bursts (GRB). These are the most intense and copious releases of energy observed in the Universe - less than that of the Big Bang itself but much more than given out by Supernovae or Quasars. GRBs can at their outset release enough energy to give them a luminosity calculated to be 10^{19} greater than that of the Sun. They are characterized by extreme outputs over very brief periods, measured in seconds to minutes at their peak. At least one GRB is observed each day somewhere in the Universe, so they are rather common events, albeit less frequent than Supernovae.

Despite being the largest rapid release high energy events in the Cosmos, GRBs were unknown (sometimes mistaken for ordinary Supernovae) before 1967. The manner in which they were discovered is an interesting example of serendipity: Nuclear explosions on Earth release large quantities of Gamma ray energy. In the 1960s, the U.S. was seeking ways to detect Soviet nuclear tests, so it built and orbited Gamma ray, X-ray, and neutron detectors on military satellites. In the U.S. Air Force Vela program, the Vela-4 satellite detected many Gamma ray events, all at times that failed to correlate with any known nuclear blasts on Earth. These Gamma ray events were all proved (eventually) to emanate from well beyond Earth. Here is a plot of one of the first records:

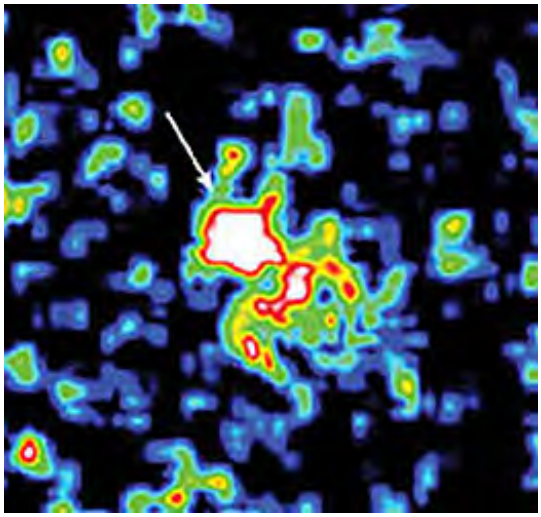


GRBs give off tremendous amounts of energy extending through all wavelengths of the EM spectrum. The diagnostic signature of the GRB that separates it from Supernovae is the predominance of high energy Gamma rays over very short time periods. GBRs can be subdivided into two types: short burst (around 2 seconds) and long burst (more 2 seconds; initial emissions on the order of 20-30 seconds, with a few extending up to an hour). This time spike has been observed in GRBs detected by more sophisticated sensors that monitored such events. Thus,

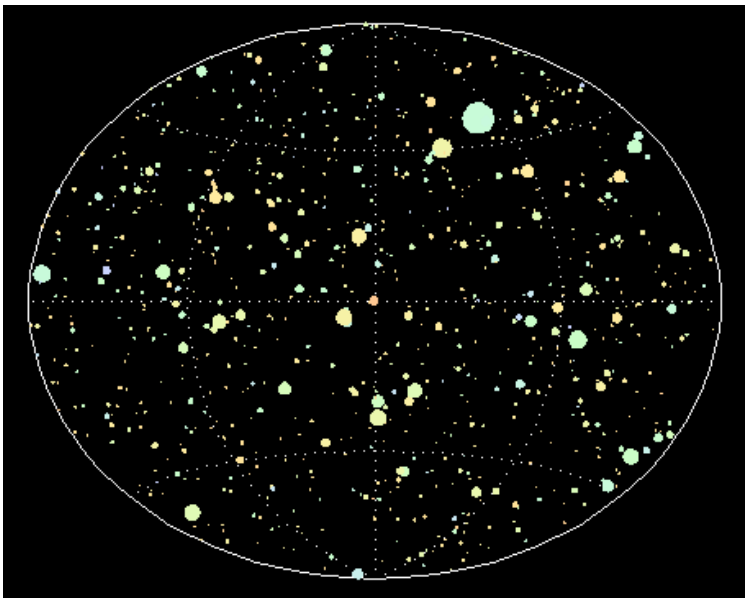
this example:



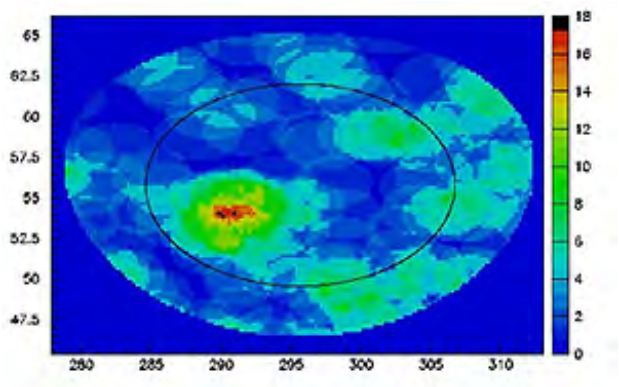
These GRBs puzzled astrophysicists. They were first thought to be in the Milky Way. And in fact some were actually located in our galaxy, where they occur on average about once in 10000 years. Afterglow radiation from one such event was observed on February 28, 1997 in the M.W. itself by an Italian X-ray satellite called BeppoSAX:



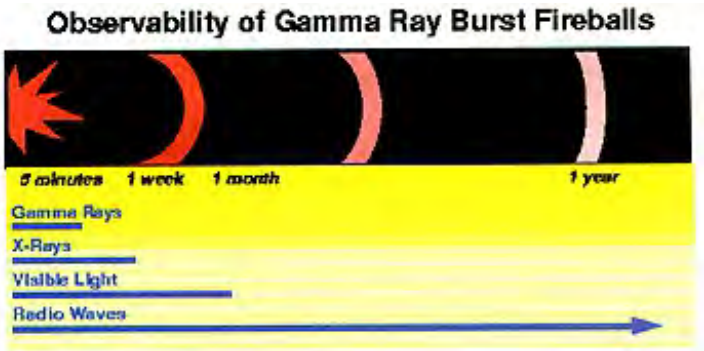
But, the frequency of occurrence, which as more observations were confirmed indicates at least one GRB every day, suggested that the vast majority of GRBs were located in galaxies well beyond the Milky Way. As more records of these events accumulated, it became evident that GRBs are not concentrated in specific regions of the sky but are distributed at random (isotropic) over the entire sky. GRB's are also randomly distributed in time - occurring anywhere in the Universe (thus over the full extent of time since the first galaxies). A large number seem to be distant, near the outer part of the observed Universe, and hence were most common in the early history of the Universe. Here is a map of the sky showing many of the larger GRBs, as detected by the Compton Gamma Ray Observer and BeppoSAX.



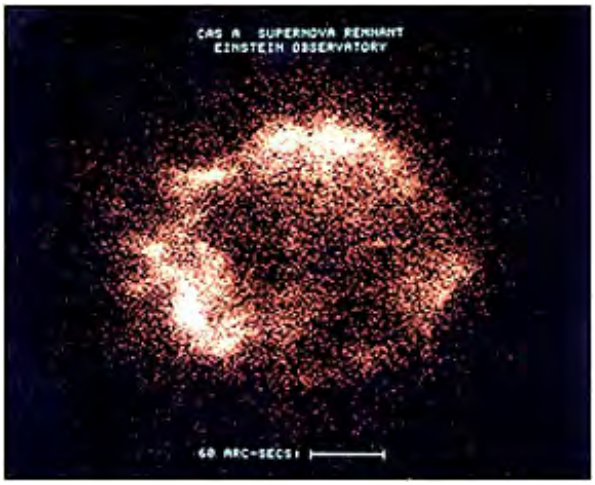
The BATSE (Burst and Transient Source Experiment) instrument on the Compton Gamma Ray Observatory (CGRO; see page 20-4) was particularly suited to detecting GRBs. Here is one image of an event that occurred several billion light years away:



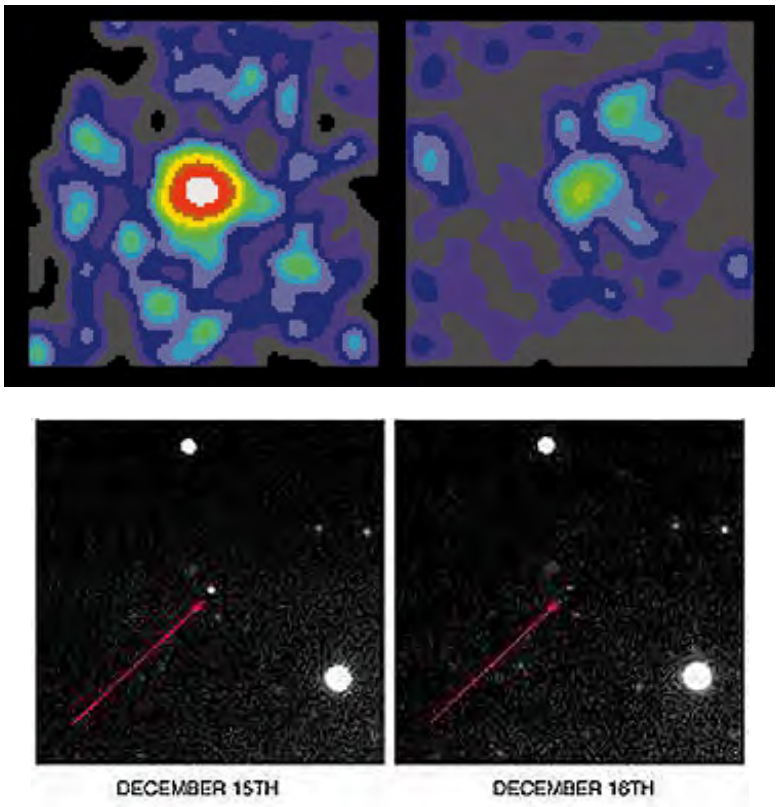
These GRB events should generate radiation at wavelengths longer than those of Gamma rays. As studies of them expanded, traces of individual events were sought by other satellites that monitor at different wavelengths. The problem is that evidence of a GRB diminishes rapidly at shorter wavelengths. However, in time such events were picked up at various wavelengths when alerts were given and the sky locations established. Now, with experience this is the time frame for durations of GRBs over a range of wavelengths:



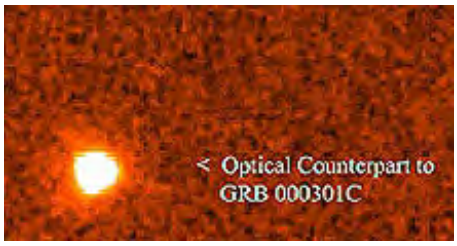
These signs of lower levels of energy at longer wavelengths persisting around a GRB are grouped under the general term "afterglow". X-rays proved useful as GRB signatures provided the searching satellite(s) could check out the source region within a few days. The X-ray emissions persist over periods of hours to days. This is one X-ray image of a presumptive GRB that was located in a galaxy nearby (some has classified this as a hypernova):



Images acquired by BeppoSAX were especially helpful in the sky survey for GRBs. The top illustration consists of two intensity contoured images typical of X-ray renditions; note the reduction in intensity in just four days between February 28 and March 3 (right). Below it is a pair of BeppoSAX images taken first on December 15, showing the GRB as a bright dot and then on December 16 as the afterglow had faded away.

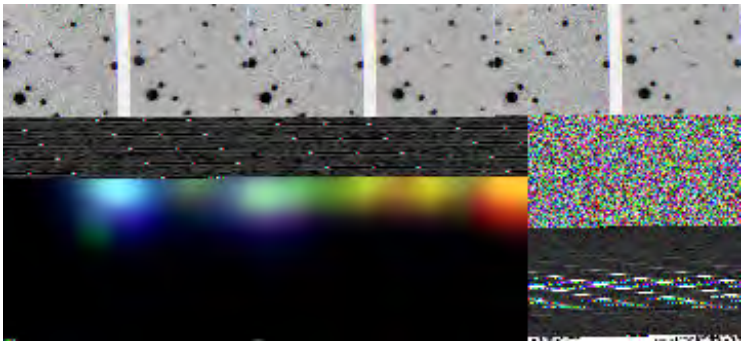


Special attention was given to finding GRBs at visible (optical) wavelengths, since these are especially capable of measuring red shifts by which approximate distance to the source can be estimated. About half the GRBs give off light in the Visible for durations of a week or more. The HST and the Keck Observatory in Hawaii were pointed at targets reported by other observing satellites. Here is the HST image of event GRB 000301c.

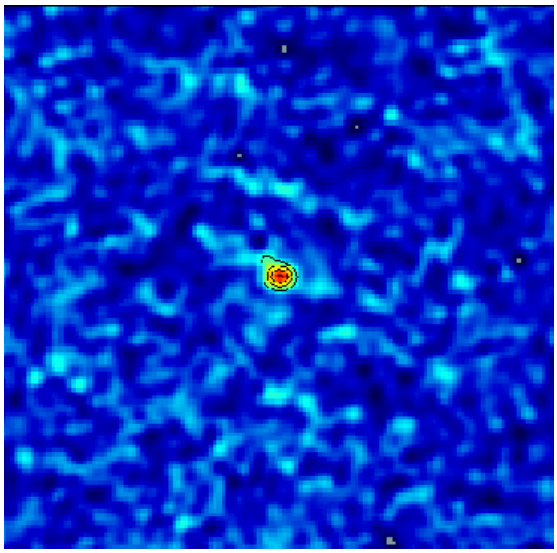


A ground telescope imaging of another GRB shows the burst as seen in visible light (here the

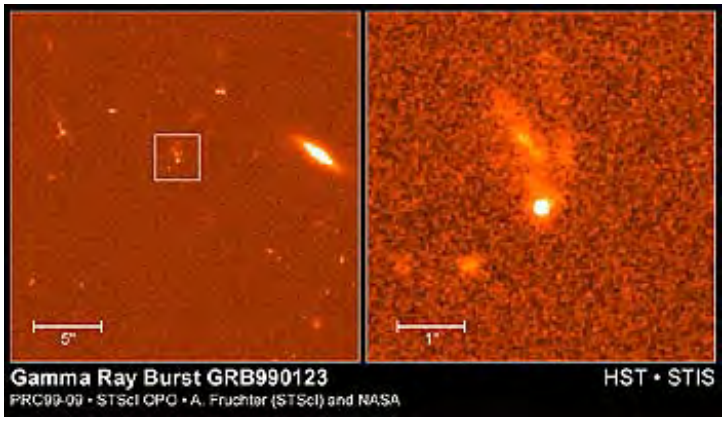
print is a negative) at 21 hours (left) and 8 days (right) after first detection. The rapid fading of the galaxy-sized feature is evident (note arrows)



Although not used a lot for this purpose, Radio telescopes have detected and imaged GRBs. Here is one made by the VLA group:



One very important GRB event led to some intriguing information that indicates that this phenomenon occurred much more often early in cosmic time (but continues til the present) and helps to confirm the huge amounts of energy involved. Its magnitude is equivalent to 100 million billion solar radiances. On December 14, 1997 the CGRO registered this event. Word was sent to BeppoSAX operators and to the HST and Keck telescopes to look for it as rapidly as possible. All succeeded. This is how the event was imaged by the HST:

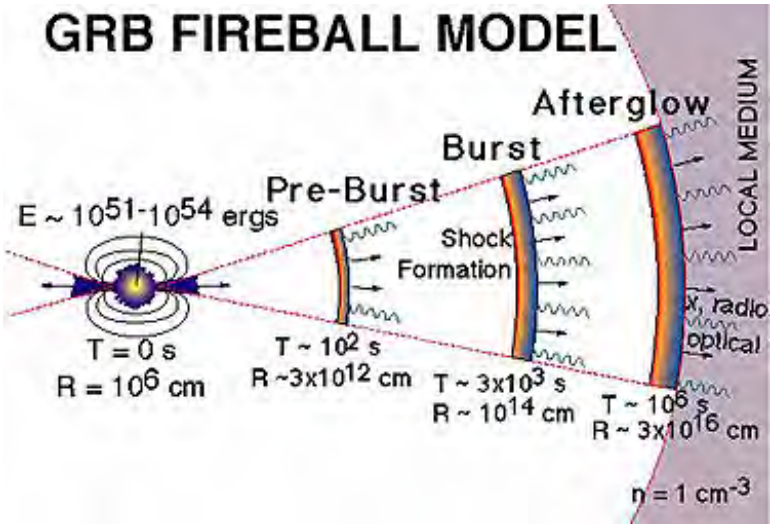


The image on the right was taken on January 23, 1999 during its maximum. When a redshift distance measurement was made on the GRB, it was found to be some 12 billion light years from Earth, proving the surmise that GRBs have probably been part of the Universe's history since soon after the Big Bang. It was also the brightest object yet found at that far distance from Earth.

Thus, the pattern found for most GRB events is rapid emission of Gamma rays followed, as they fade, by the dominant radiation passing through X-ray, Visible, and Radio wavelengths, with the whole sequence being over in less than a few months.

GRBs are of such high interest that another dedicated satellite has been placed in orbit to look for these and similar events. This is HETE-2, the High Energy Transient Explorer, launched October 9, 2000 (the first HETE failed to separate from its third stage rocket). It is described at this MIT site.

The cause(s) of GRBs continues to be uncertain and tantalizing. As an aid to the following discussion, use this Fireball Model to provide a framework for the starting energy, the expansion of the GRB, and the time involved in reaching the afterglow phase:

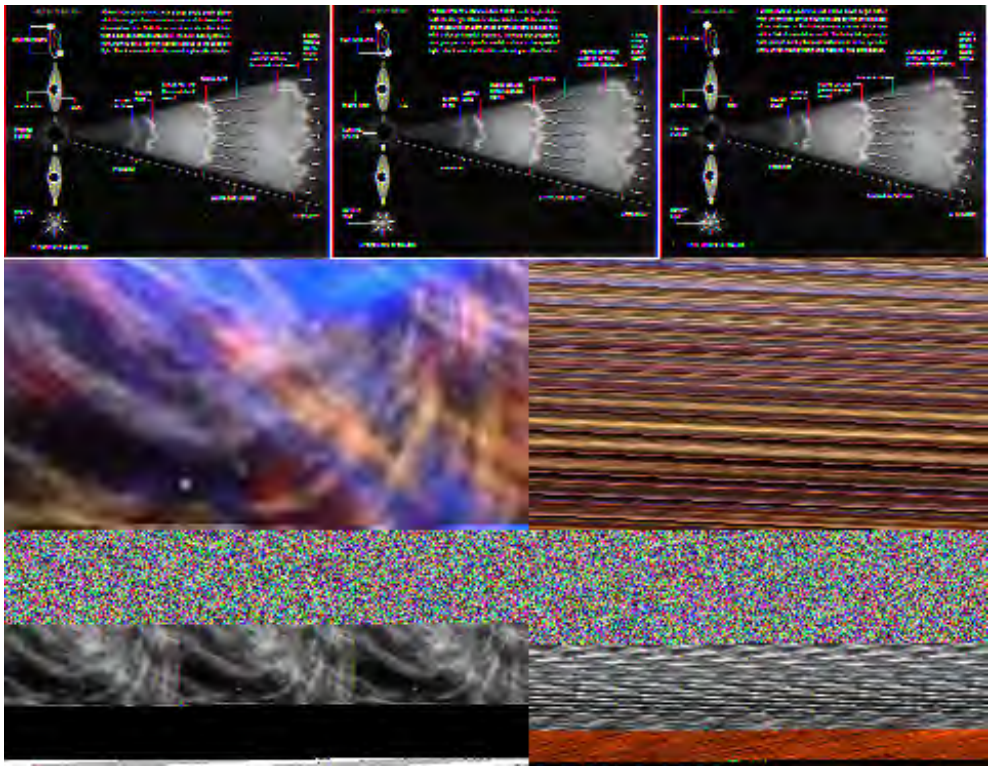


The early idea of the explosion of material sucked in and around a neutron star (see top illustration on this page for a similar example) has been challenged. But, a variant postulates a role for a binary pair of Neutron stars which, if they should collide, should produce a huge release of energy. Still others attribute the GRBs to some involvement with Quasars. One school holds them to be the outcome of giant Supernovae (Hypernovae) which generate very powerful short-time energy release levels. A recent hypothesis takes still a new tack - the GRBs are associated with large clusters of galaxies which together have such a strong gravitational pull that they accelerate matter both within and around the galaxies to high speeds that, upon colliding with intergalactic matter, release energy at the Gamma-ray level.

Another hypothesis, known as the Paczynski Model (also known as a Collapsar event) and now the most favored explanation, starts with a supermassive (type O) rotating star that collapses to form a Black Hole that continues to draw more material around it until a critical state is reached that requires an intense Supernova-like explosion producing the GRB fireball. Essentially, all the mass involved is suddenly converted to energy in obeyance to the Einstein $E = mc^2$ relation. There are indications that this energy release may be directed, something like the beam associated with a Pulsar. This is an artist's conception of the beam that seems to be formed for at least some GRBs:

Until the release of information in June 2004 about a GRB only 35000 l.y. away - either in or just outside the M.W. - no nearby events had ever been confirmed. The image below shows W49B as a color composite made from Chandra X-ray data (blue), and Palomar telescope images taken in the IR (green and red). The estimated initial release of energy over a 1 minute time span is 10^{13} greater than that of the Sun in that timeframe. The image represent the GRB status soon after the burst; the colors indicate enrichment in Iron. The postulated beam associated with collapsar events was not oriented straight at the Earth and hence is not visible here.

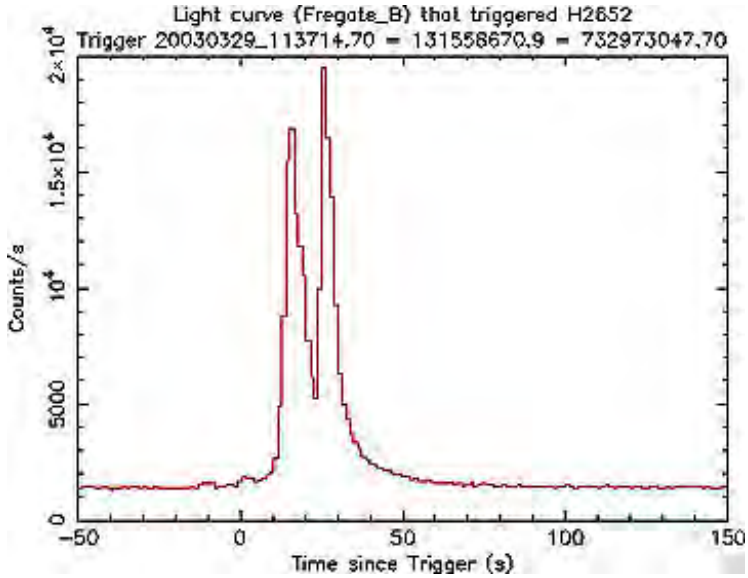
Some of the above information has been extracted from an article in the December 2002 edition of Scientific American, entitled "The Brightest Explosions in the Universe", by N. Gehrels, L. Piro, and P. Leonard. The article contains this illustration that summarizes the authors' ideas on the formation of GRBs:



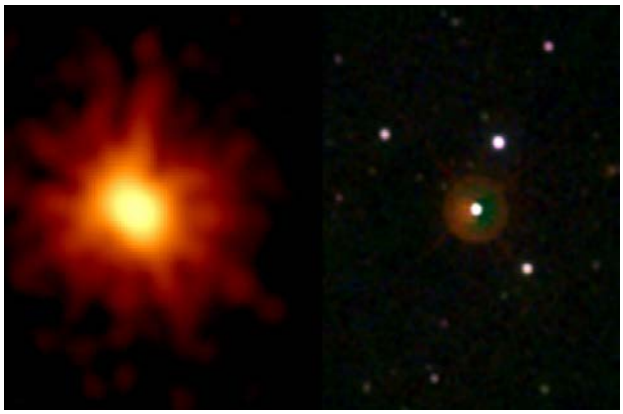
From Scientific American, December 2002

In their model, similar to some others proposed, GRBs are definitely associated explosive processes that will end up forming Black Holes. In one common mechanism, a massive star collapses and explodes as a Hypernova, leading to a disk of matter/energy surrounding a Black Hole; this is a fast process in the sense that at a critical time, the Hypernova ensues without anything discernible obviously leading up to it. Alternatively, over a long span of time (millions of years, the same end result could occur as two neutron stars mutually orbiting finally crash into each other. The wedge to the right of the 'Central Engine' conforms to a jet that carries the photons released in the GRB outward at near light-speed. This material moves outward as "blobs" that catch-up and coalesce forming internal shock waves that generate the Gamma bursts. With expansion over time, the high energy photons are replaced by those of progressively lower energies represented by X-rays, Visible light, and Radio waves as the emissions encounter the galactic/intergalactic medium. The final result is an afterglow that fades over time.

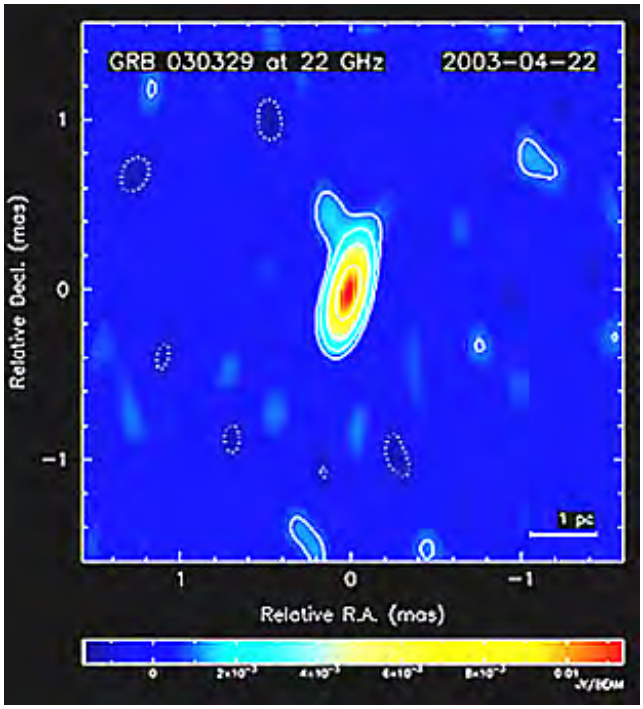
On March 29, 2003, HETE-2 captured a GRB (HETE Burst H2652 is also listed as GRB 030329 and SN2003dh) in a galaxy 2.6 billion light years distant and sent the occurrence of this event back to Earth so quickly that many observatories were alerted quick enough to train their telescopes on it within minutes. Thus, for the first time the earliest stages of a GRB could be monitored. This event proved one of the brightest ever observed. This plot of HETE data shows how brief was the main phase of the event.



The SWIFT satellite obtained this image of GRB 030329:



A Radio telescope image of this GRB taken on April 22 shows a progressive distribution of decreasing energy moving outward. This seems to confirm the "fireball" model for ejection of matter (an alternate explanation, that material is ejected in huge blobs [the "cannonball" model], is apparently not valid for this observation). The expelled material moves at nearly the speed of light.

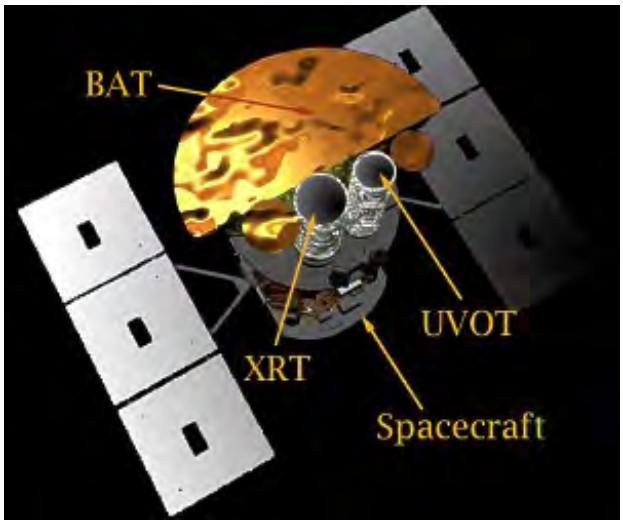


Spectrographic data showed that the initial burst of H2652 was rich in excited Silicon and Iron.

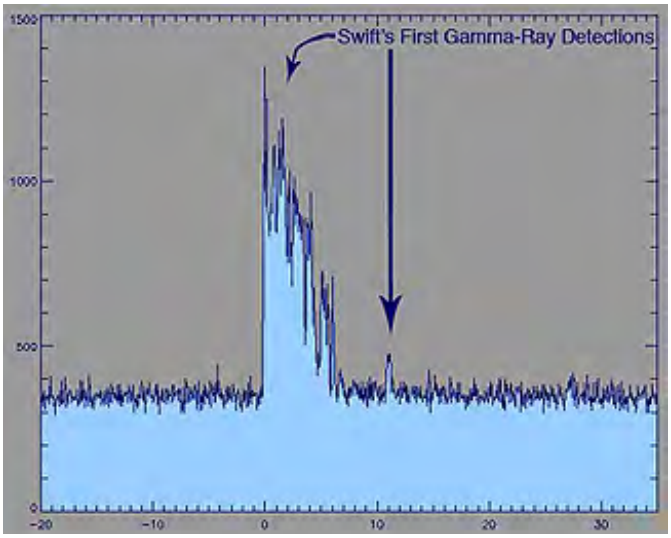
These elements would be produced in a star whose mass is at least 30x that of the Sun, which would give rise to temperatures and pressures that generate nuclear reactions that fuse nuclei into Si and Fe. These are the conditions that favor a "super-Supernova", another way of referring to hyperNovae. Astronomers believe this is convincing evidence for that mode of generation of many (perhaps most) GRBs.

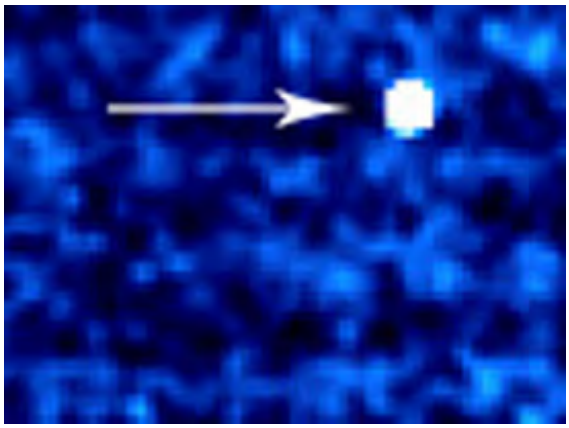
As knowledge of the roles of the predominant dark matter and dark energy (see pages 20-9 and 20-10) is increasing, another explanation (Louis Clavelli) - somewhat conjectural - has emerged. This holds that dark energy under the right conditions acts upon ordinary matter to convert it to dark matter. From theory, this should give off intense bursts of energy as the conversion proceeds, witnessed by us as GRBs.

Another space telescope dedicated to GRB detection is NASA's SWIFT, launched on November 20, 2004. This spacecraft has three detection systems that can be activated within minutes of the one always on that picks up the Gamma-radiation. SWIFT is capable of finding and monitoring as many as 2 GRBs per week, far more than previous instruments, and will follow the changes to the stage when Black Holes as the end product is reached. SWIFT also can "see" back to the early days of the Universe out to almost 14 billion light years away. The spacecraft and its three telescopes are shown here:

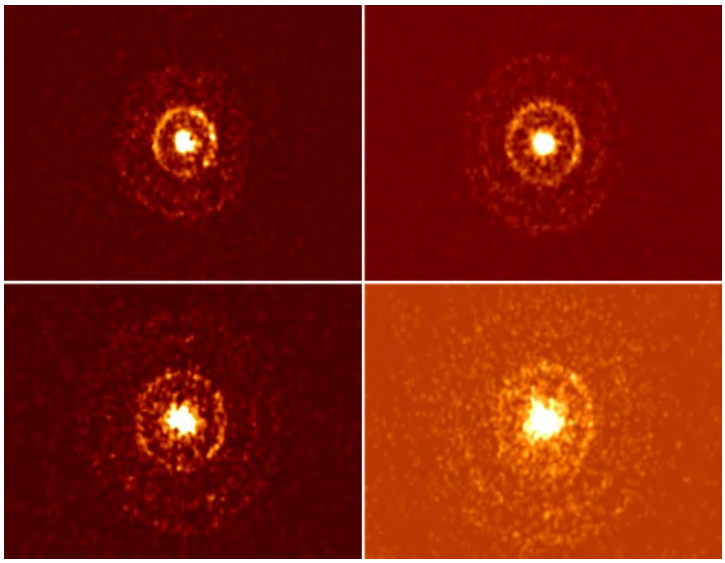


In January, 2005 9 GRB events were detected by SWIFT. The first detection was in December, 2004, as shown below as an energy plot and the actual image derived from the data:

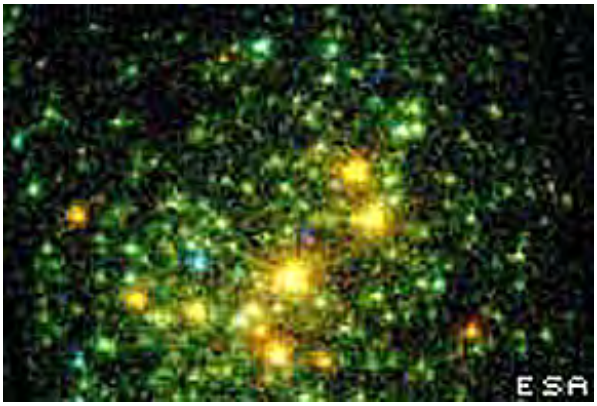




The four panels in the illustration below show sequential steps in a gamma ray burst:



As mentioned above, GRBs are all short-lived, even in human terms, lasting from hundreds of seconds to a few days. One very short duration type, which releases much less energy, is known as *Gamma rays flashes*; this lasts for milliseconds to a few seconds. These have been observed by SWIFT and by earth-based telescopes. Their cause(s) may be Neutron star pair interactions but other mechanisms have not been ruled out. Here are several that occurred simultaneously:



Problems with explaining GRBs are compounded by observations (using EGRET, NASA's orbiting Energetic Gamma Ray Experiment Telescope, part of the Compton X-ray Observatory) of about 170 sources of continuously emitting high energy Gamma rays. Thus, these do not display short-lived bursts. This class was first discovered by the Compton Gamma Ray Observatory (page 20-3). They may be associated with clumps of supersymmetric particles (page 20-1) including a type called the *neutralino*.

Gamma Ray Bursts are of great significance to astrophysicists from a theoretical standpoint. But, they also should be of interest to the general public - to all of humanity, since they pose a

potential threat to all life on Earth. If one should occur in this part of the Milky Way, and be pointed in the right direction, the gamma rays could be fatal both from their direct effect on living tissue and from the likelihood that they would blow away the ozone layer that protects Earth from solar irradiation. This may actually have happened at the end of the Silurian Period about 400 million years ago when a major extinction occurred, which has not yet been linked with any other cause. However, there is a very small probability that the any star near the Sun will be of the right size to produce a hypernova; but, the neutron star collision cannot be ruled out right now since such binary pairs are not easily detected.

In sum, there seems to be two entirely different mechanisms that release the huge amounts of energy associated with GRBs. Longer duration GRBs result from Hypernova explosions. Shorter GRBs are probably the final instantaneous phase of collision between two Neutron stars. Needless to say, GRBs continue to fascinate cosmologists since they represent the largest and fastest explosive events beyond that of the Big Bang itself. As they are better understood, they may reveal the action of physical processes only now being speculated upon, and suggested by particle physics experiments. The next big step in studying GRBs and the continuous types will be the launch of GLAST (Gamma-ray Large Area Space Telescope), perhaps as early as 2007. For more information about this phenomenon, consult this Wikipedia web site.

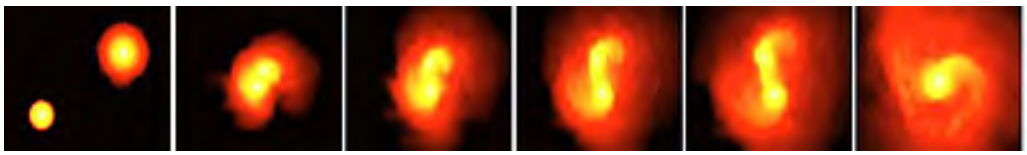
Colliding Stars

At least some of the GRBs and X-ray bursts may stem from collisions or mergers of (usually two) stars (check back at the bottom of page 20-5 for the earlier review of galaxy collisions). As recently as the 1970s astronomers considered collisions to be rare stellar events. Although an actual collision has as yet not be observed by HST or other astronomical satellites and ground telescopes, phenomena associated with certain stellar configurations have now been postulated (attributed) to either head-on or glancing encounters between stars.

As we have seen, stars in the arms of spiral galaxies or the fringes of elliptical galaxies are very widely spaced and hence the probability of collision is low. But star distributions in central cores of these two types show much closer spacing (denser).

To appreciate the significant increase in density, if one counts the stars that are about 25 light years from our Sun, the number would be about 100 but if the same 25 l.y. volume is set around the center of a globular cluster, that star number rises to an order of about 1,000,000. This crowding means that those stars are very closely packed and hence capable of numerous collisions. Three processes make collisions much more likely: 1) a process called "evaporation", in which stars approach others and some are then flinged out of the grouping which contracts to such densities as to make collisions inevitable; 2) gravitational focusing, in which approaching stars have their pathways deflected so that two stars now follow a collision course; 3) tidal capture, in which neutron stars or Black Holes latch onto nearby stars and in time draw these into the high gravity

Theoreticians have developed computer models to simulate pictorially different modes of collision. Shown here is the sequence of change as two Sunlike stars are merged:



The end result of a collision depends on several factors: 1) whether there is a direct hit or a glancing encounter; 2) the relative size (mass) difference between colliding bodies; 3) the terminal speed of each body. The process can be as brief as an hour; or as long as days to years (this rapid time for completion is one reason while such events have yet to be observed in "real time"). Any two of the 7 density types shown near the top of page 20-5 can experience a collision. In some combinations, such as a White Dwarf striking a Red Giant, the end result is two

White Dwarfs (one being the incoming member; the other [Red Giant] dispersing and losing so much of its gas by the interaction that only its core remain which quickly evolves into the new White Dwarf, Or, one star remains relatively intact as the second star is incorporated within it.

In this last case, the result is that the now coalesced star pair has gained considerable mass. This means that it now appears to be a bigger star, and since the total mass determines the rate of Hydrogen fuel consumption, the new, brighter star would appear as though it will burn its Hydrogen mass much faster and thus appears to have a shorter lifetime - hence seems younger. A specific case: if two stars, each with a mass of the Sun (5 billion years old) that has a total burn out time of 10 billion years, collide and form a single star with twice the mass, the now more luminous composite star would have a life expectancy of 800 million years. This seems to be the best explanation of "Blue Straggler" stars - much brighter than the majority of stars in a globular cluster. This is evident in this HST image of NGC 6397:

An excellent summary of stellar collision processes and consequences is offered in the November 2002 edition of Scientific American under the title "When Stars Collide", by Michael Sharma.

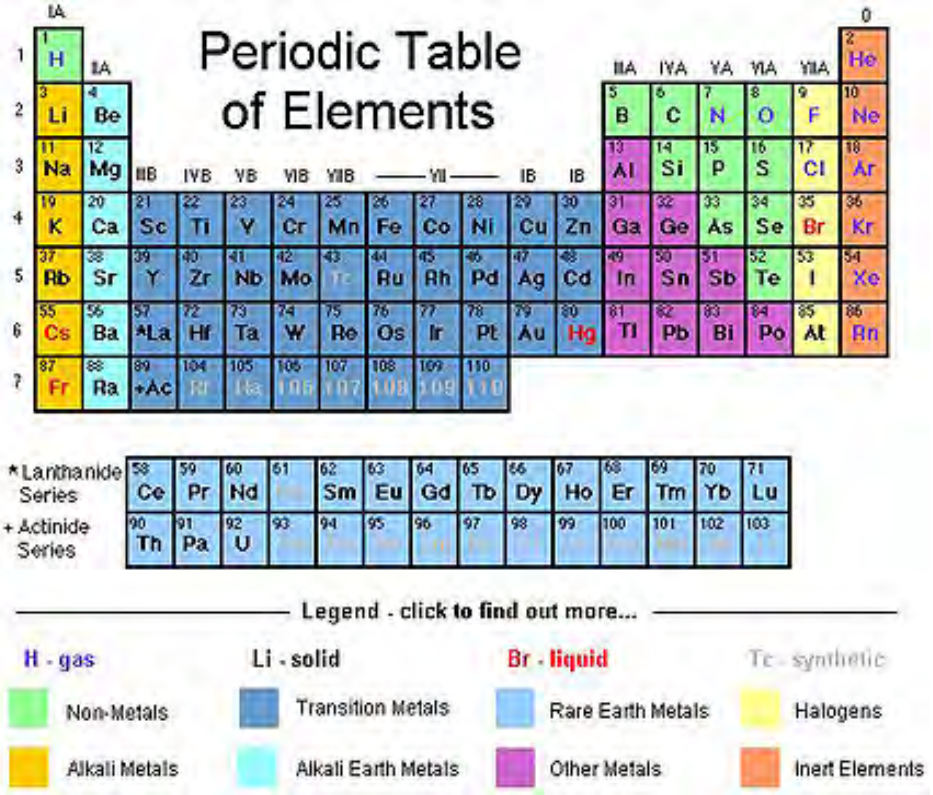
PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Spectral Analysis of Star Composition;

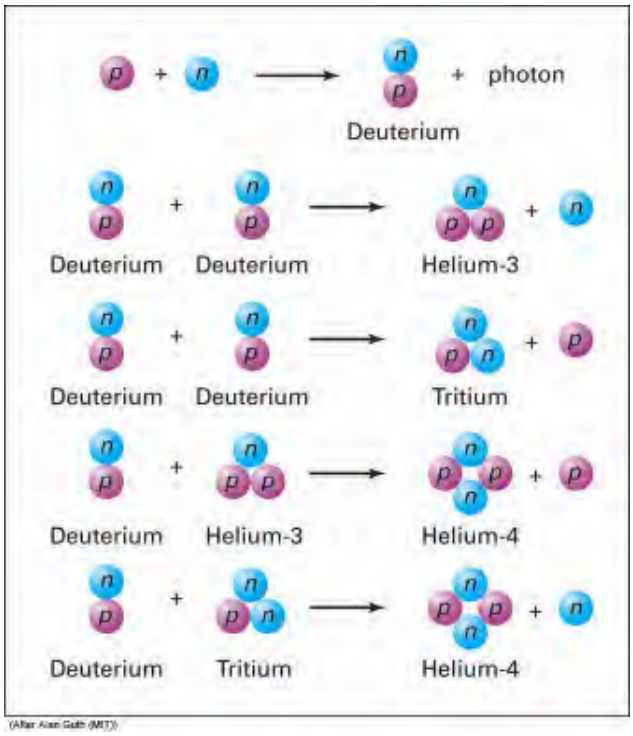
Element Synthesis in Stars

On this page we look into how the elements that are found in Nature were created in the past. As a reminder, here is the well-known Periodic Table of the Elements (we will not discuss its development per se, but if curious consult any Chemistry text):

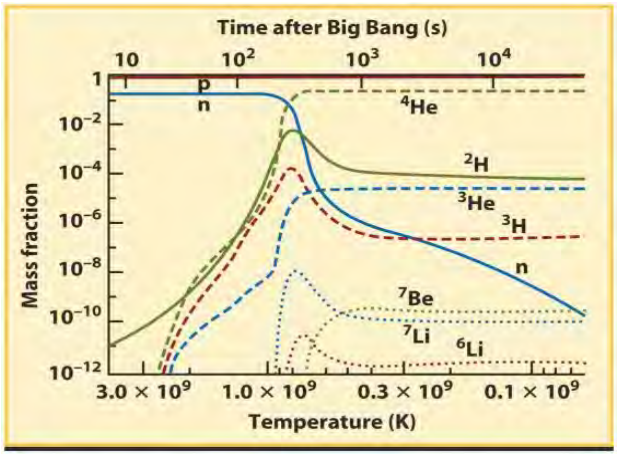


The number in the upper left of each box is that elements Atomic Number N = the number of protons in its nucleus. The Atomic Weight A (not shown) = the number of protons and neutrons in the fundamental isotope of that element (e.g., Fe = 56, which consists of 26 protons and 30 neutrons). (Note: On this page we will discuss element formation by a variety of nucleosynthesis processes; a useful review of some of the fundamentals is found on the Internet at nucleosynthesis. Some of this information appears also in paragraphs below.)

As a general statement: Elements up to N = 5 were created by primary synthesis during the Big Bang (but some of elements N = 2 to 4 were formed later in stars); elements from N = 5 to N = 26 were formed inside the first and subsequent stars by fusion; elements from N = 27 to N = 94 (93 = Neptunium and 94 = Plutonium can occur naturally but are rare) were generated by neutron capture (or later Beta decay) in supernova explosions. The Big Bang produced just two elements in abundance: Hydrogen at ~75% and Helium at ~25%, by atomic weight, or in the ratio of 7 to 1 by numbers of atoms. During the first minute of the Universe temperatures fell rapidly through an initial trillions of degrees Kelvin to a few thousand degrees. These high temperatures were necessary to induce fusion. This is a basic scheme for Hydrogen-Helium isotope production within this time:



This shows that a fraction of the Hydrogen atoms initially produced became the isotopes deuterium and tritium, which contain 1 and 2 neutrons respectively). The only other elements produced in the beginning were Helium, and a smattering of Lithium and Beryllium. Hydrogen is the fundamental constituent of the Universe, from whence all others of atomic numbers 2-4 (in addition to the primordial He, Li, and Be) and higher have been created after the Universe's opening moments. The path of this initial elemental isotope production can be shown in this diagram:



Most of the 88 elements of atomic number greater than 4 that occur naturally on Earth (21 more elements have been created solely in the laboratory by particle accelerators, etc.) or have been detected in stars have been, and are being, continuously created *not in the first few minutes of the Big Bang* but throughout subsequent Universe time *within stars* and are constantly being redistributed through destruction of stars (mainly by supernovae events) and reorganization of the debris into new stars, dust clouds, and under favorable circumstances into planets. Ever newer (younger) stars, as well as the interstellar medium, are becoming progressively richer in elements of atomic numbers greater than H. Since no new matter appears to be created since the Big Bang, the amount of primary Hydrogen must be decreasing. Because so many of the stars in the early Universe were massive, short-lived, and subject to explosions, the heavier elements were more rapidly produced and released in the first few billion years than, say, the present.

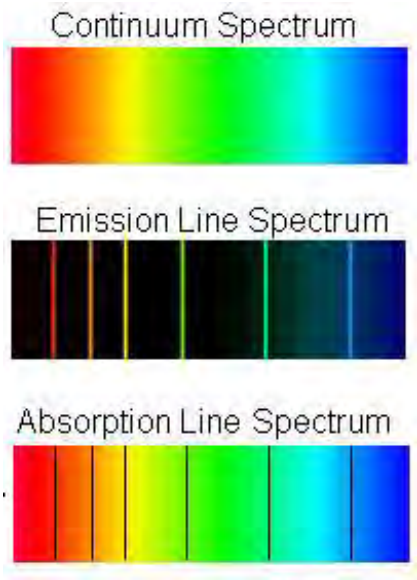
The production of heavier elements almost certainly began as soon as the first stars formed in the first half billion years. Most such stars were massive and hence heavier elements were

developed in their cores. These stars, being short-lived, exploded as supernovae and spread the interior elements into the growing Universe. As direct proof of the appearance of various element species in this timeframe, Carbon monoxide has been detected spectrally in the light given off by one of the oldest quasars yet examined.

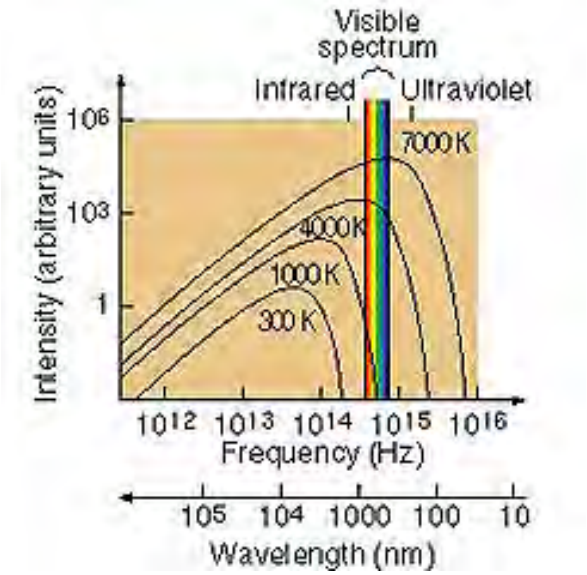
A good review of element production in stars are found at a site maintained by the Wright Center for Science Education; Tufts University.

Before we examine the element-forming processes involved, it is instructive to review how the composition of stars is determined. That determination is done primarily by spectroscopic analysis of the radiation emanating from the outer shells of a star, including (by analogy to the Sun) its chromosphere and photosphere. The principles involved in spectroscopy, as it is applied to obtaining spectral data of the Earth primarily by sampling reflected or emitted radiation of solar-irradiated surface and atmospheric features, were covered on page 13-6 (a review of that one page may prove helpful in working through the present page).

As applied to star analysis, four kinds of spectral data are relevant: 1) Continuous spectra; 2) Emission Line spectra; 3) Absorption Line spectra; and 4) Blackbody radiation spectra. The first three are illustrated here:

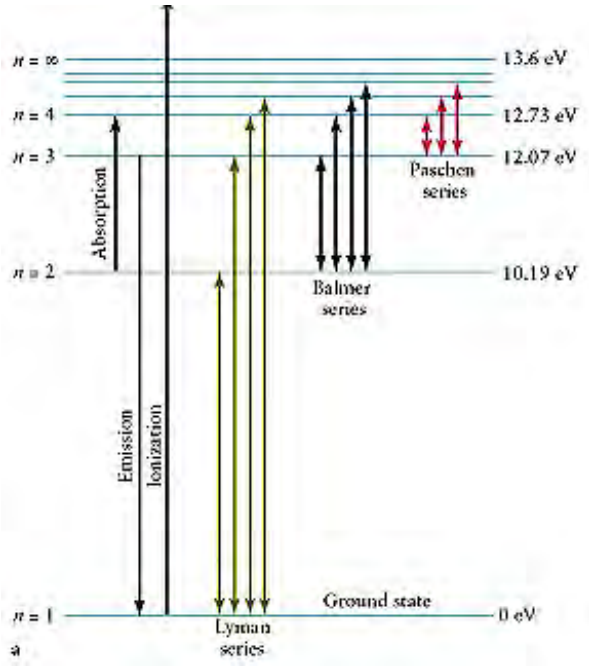


A *Continuum Spectrum*, particularly as it applies to the UV, Visible, and Near IR segments of the Electromagnetic Spectrum, is that produced by white light, i.e., all wavelengths (all colors in the Visible) in this region are present in essentially equal proportions. An *Emission Line Spectrum* results when photons of narrow, specific wavelengths are emitted during excitation of the elements present in an emitting body; each (colored) line is diagnostic (identifies) some particular element. An *Absorption Line Spectrum* occurs when the emitted radiation from a hot body passes through cooler gases containing the same elements as those responsible for the emitted photons which interact (by absorption processes) and are removed from the continuum (as shown by black lines). A Blackbody Spectrum, discussed in detail on page 9-1 and 9-2, is a continuum spectrum that is associated with the thermal state of an emitting body considered to be a *perfect radiator*, especially one heated to incandescence, in which its spectral curve peaks at some wavelength which varies systematically with temperature. This figure should remind you of the information presented on page 9-2 dealing with Wien's Displacement Law.

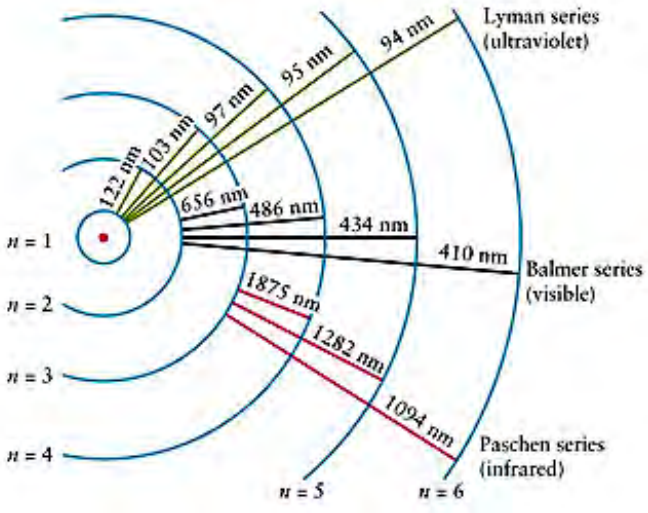


Now, in more detail: Each element has a series of spectral lines that are diagnostic, being found in fixed locations in a spread of the spectrum as determined by the wavelengths of emitted radiation resulting from excitation of electrons into higher energy levels (recall the formula: $\Delta E = h\nu$). Emission lines relate to light (including UV and IR) radiation passing unimpeded from the source. But, starlight normally must pass through the star's atmosphere; if the outer gases contain the same elements as those from its surface, the emitted radiation will be absorbed at the characteristic wavelengths, giving rised to absorption spectra. The image below is the spectrum for our Sun, with the dark absorption (Frauenhofer) lines correlating mostly with Hydrogen and Helium:

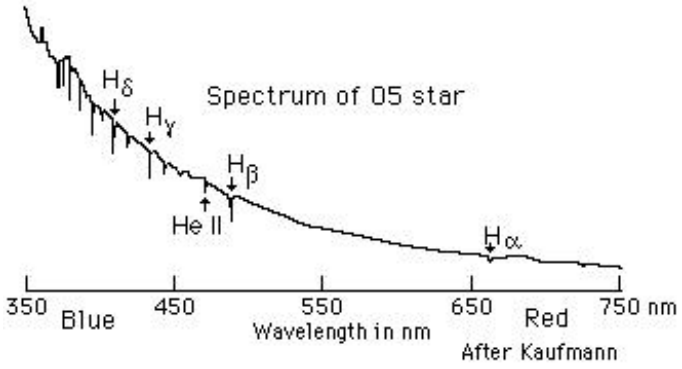
Since Hydrogen is by far the most common element in the Universe, comments on its spectra are in order; the principles involved in the generation of Hydrogen's spectral lines apply to all other elements. Radiation from excited Hydrogen is detectable over most of the EM spectral range, but important and diagnostic radiation at specific wavelengths used by astronomers extend from the Ultraviolet through much of the Infrared range. Emitted radiation results when the single electron in the neutral Hydrogen atom is excited by various forms of energy (e.g., heat, electrical current, particle bombardment) such that the electron is displaced from its ground state to one or more of the various energy levels associated with the possible orbital levels surrounding the nucleus. These are energy levels that are discrete (specific values) in terms of the quantum states possible when excitation has occurred. These levels are, by convention, represented by the letter "N" and are expressed as integers from 1 through 2, 3, 4, 5,6, ... infinity. In the ground state, the electron resides in level 1 (or shell, as is often depicted in the Bohr atom model). When excitation energy is provided, the electron can "jump" to higher (quantized) levels, as, for example from N = 1 to N = 3. That energy is calculated by the familiar Planck equation: $\Delta E = h\nu$, where the ΔE is the energy required to move to a specific level, say from 1 to 3, shown as $E_3 - E_1$, h is the Planck constant, and ν is the frequency (its reciprocal is the wavelength λ. In the higher energy states (multiples of N greater than 1), the electron may remain for a time in a metastable mode but for most of the transitions the electron almost instantly returns to a lower energy state (either to the ground state N = 1 or to one of the lower levels of N than the level first reached by the electron. When the return occurs, the excitation energy is given off as photons whose specific frequency (or wavelength equivalent) is determined by ΔE. Examine this diagram:



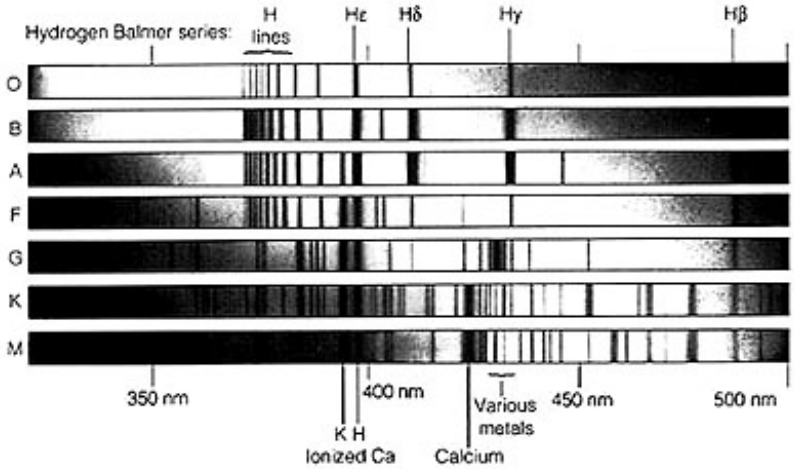
For the Lyman series (of transitions expressed as spectral lines of very precise wavelengths), the electrons will move to different N levels and then revert to the N = 1 state. For the Balmer series, the reverted level is N = 2; the Paschen series, N = 3. To illustrate with specific values, consider the Balmer series, in which the four principal lines, designated as H_α, H_β, H_γ, and H_δ, require (in the same sequence) energies (hν)of 3.02×10^{-19} , 4.07×10^{-19}, 4.57×10^{-19}, and 4.84×10^{-19} Joules (J), and give off photons whose wavelengths (state here in nanometers [µm x 1000) are 656.3, 486.1, 434.0, and 410.2 nm respectively. The Balmer wavelengths are all in the Visible region of the spectrum. The Lyman series occurs in the Ultraviolet and the Paschen series in the near Infrared segments of the EM spectrum. There are other series (not named) elsewhere in the EM spectrum. Now, look at this next diagram - a variant of the one above but with added information:



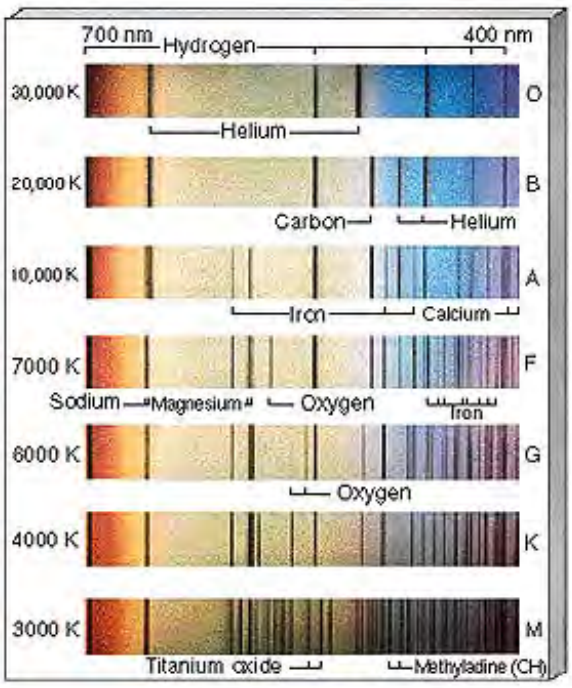
All of these lines are found in solar spectra. A spectral curve (the spectrum as plotted on a strip chart recorder) from an O-type (very hot) star produces absorption spikes for the Balmer series in the Visible; it looks like this:



As was first treated on page 20-2, letters in the sequence O-B-A-F-G-K-M refer to spectral classes of stars; the sequence is also an observed temperature indicator with each letter denoting a range of temperatures, with O hottest (greater than 10000°K) and M coolest (less than 3000°K), Typical spectra for the different classes of stars on the Main Sequence will include lines for Hydrogen, Helium, and other elements, shown as follows:

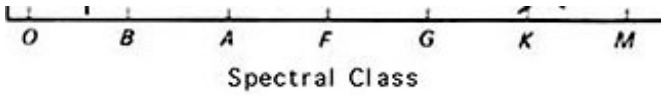


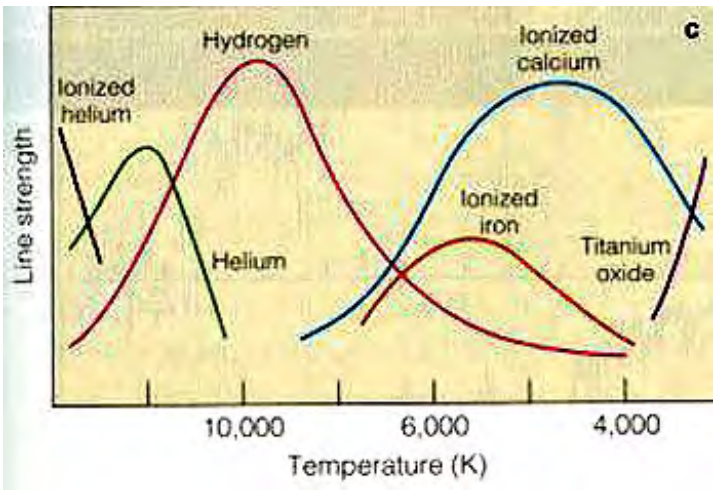
The following are principal spectral lines within the Visible spectrum representing the different stellar classes, with surface temperatures plotted on the ordinate:



This next diagram helps to categorize the star spectral classes O through M, in which for each class a range of spectral lines of certain individual or several elements are diagnostic and may predominate. Thus an A star shows strong Hydrogen lines with some neutral Helium and ionized

metals contributing their lines whereas a K star spectrum is predominantly that of Calcium and excited neutral metals.



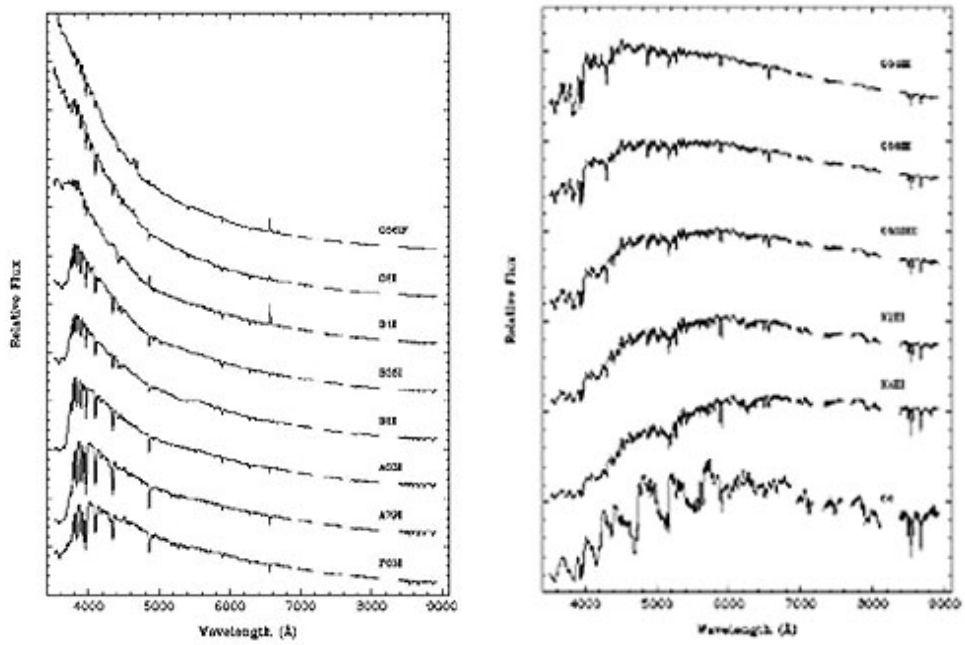


This can be restated in the following chart that names the star class, its intrinsic surface color, a characteristic surface temperature, and the principal diagnostic spectral lines.

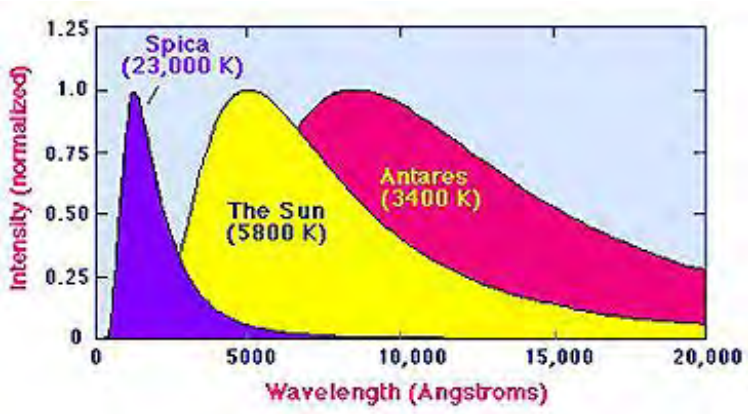
Spectral Class Characteristics

Spectral Class	Intrinsic Color	Surface Temperature (K)	Prominent Absorption Lines
O	Blue	41,000	He^+, O^{++}, N^{++}, Si^{++}, He, H
B	Blue	31,000	He, H, O^+, C^+, N^+, Si^+
A	Blue-white	9,500	H(strongest), Ca^+, Mg^+, Fe^+
F	White	7,240	H(weaker), Ca^+, ionized metals
G	Yellow-white	5,920	H(weaker), Ca+, ionized & neutral metal
K	Orange	5,300	Ca^+(strongest), neutral metals strong, H(weak)
M	Red	3,850	Strong neutral atoms, TiO

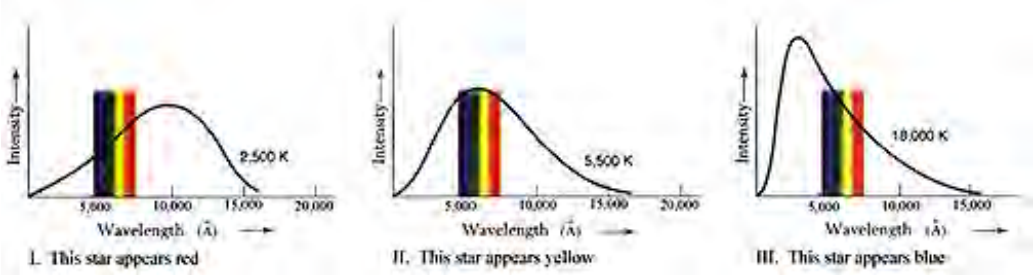
The stars off the Main Sequence will, of course, show different spectral patterns depending on their compositions. Below are two sets of spectral curves, with individual lines noted as downward spikes in part of a Blackbody spectrum (see page 9-2) in the spectral range from 4000-9000 nm (0.4-0.9 µm) range. The left (or top) set covers spectra from Blue Giants; the right from Red Giants. The shift in peaks is a function of temperature. The left group is dominated by Hydrogen lines; in the right group some lines include calcium.



These plots suggest that the shape of the overall Blackbody spectrum will vary as a function of temperature. This is apparent in these generalized Blackbody spectral curves for a very hot star (Spica), the Sun, and the cool star Antares:



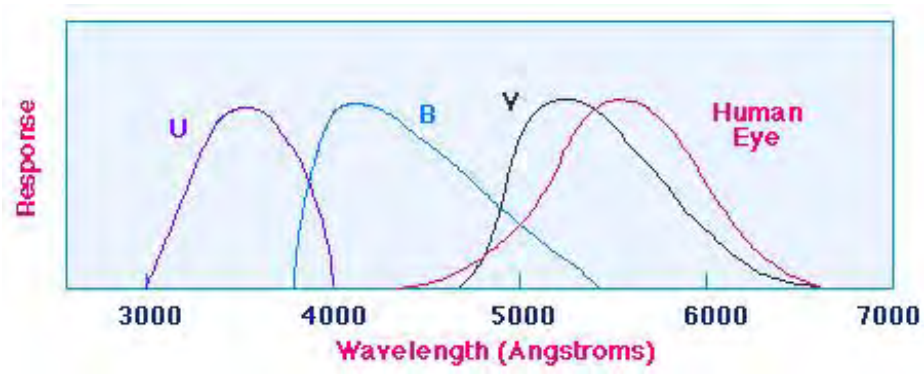
The general Blackbody curve as it shifts with temperature also aids in showing how individual stars display the colors astronomers assign to them. Consider this illustration:



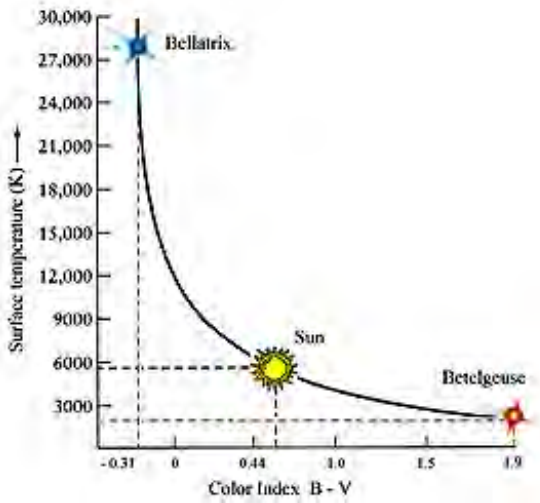
The left curve, for a cool star, shows that the part of the highest part on the curve intersected by the color spread in the Visible spectrum is associated with red, hence such stars are defined colorwise by Red. In the middle curve, the high point on the curve is straddled by yellow; a Sunlike star then is Yellow (or Orange). The right curve, for a hot star, has the visible blue at a higher intensity than green or red and hence defines a Blue star (actually, as it appears, such a star is a bright bluish-white).

This suggests that color can be used in the Letter classification. Astronomers have developed a Color Index system of relating stars to their surface temperatures. A given star is observed through a telescope at three different wavelength ranges, one (U) centering in part of the

Ultraviolet, a second (B) in the Blue, and a third (V) in the longer wavelength part of the Visible. The starlight passes through three filters, as shown:



The intensity of light received through each filter can either be expressed in flux terminology or, more commonly, converted into an apparent magnitude value "m" appropriate to the spectral range (e.g., m_B). In turn, this magnitude must be converted to an absolute magnitude M and then corrected for atmospheric effects to produce what is termed a bolometric magnitude M_{bol}. This is necessary so that all stars are compared in brightness at the same fixed distance. A Color Index value in the UBV system is then calculated as B - V (and/or U - B), by mathematically subtracting the bolometric magnitudes, as for instance, m_B - M_V. This is a B - V vs temperature plot for three classes of stars (O. G, M):



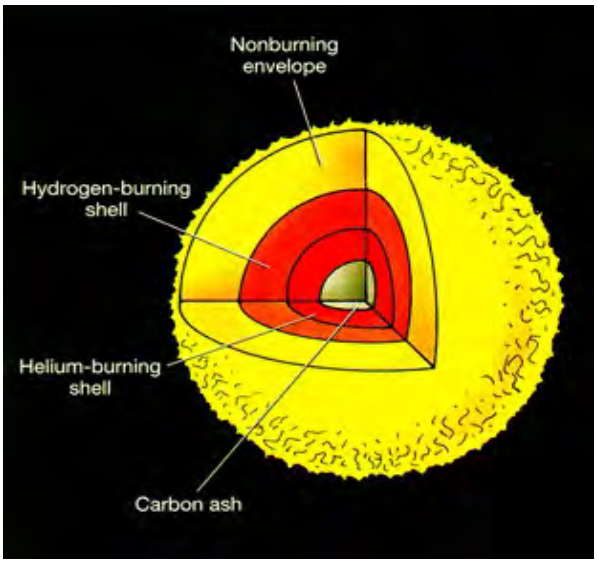
The Index can have positive or negative values. Hotter stars have C.I.'s that are negative or slightly positive; as magnitudes decrease with lower temperatures the Index becomes more positive. The Sun's B - V Color Index is +0.62. In the third illustration above the star Spica has a B - V of -0.22 and Betelgeuse a value of +1.85. A hotter star than the Sun would have a smaller +C.I. or, with increasing temperature, values that become negative.

Now, with this background, let us turn our attention to how elements of atomic number above 4 have been produced by stellar processes. These processes are described by the general term "nucleosynthesis". A good overview of this subject is found at this Wikipedia website; near its bottom are links to specific types of processes that will only briefly be mentioned on this page. As a generalization, the processes fall into two broad categories: 1) nuclear fusion, that is responsible for element synthesis up to Iron (atomic number 26) and 2) a variety of processes for elements heavier than Iron that includes neutron and proton capture and nuclear fission (as exemplified by radioactive elements).

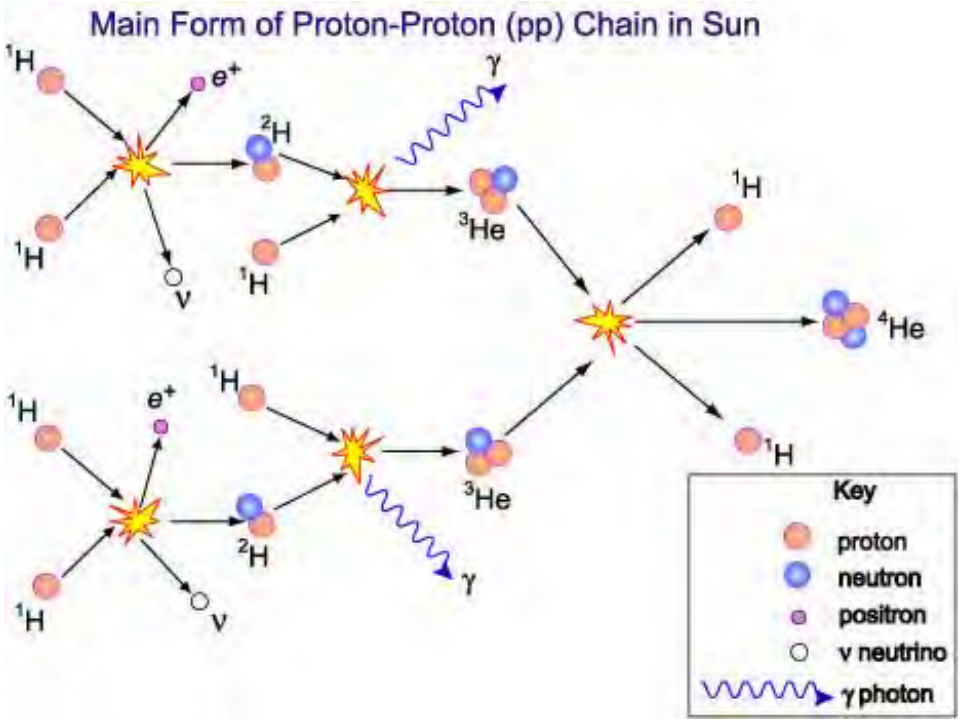
In the first half billion or so years after the Big Bang, the elemental chemistry of the Universe was quite simple. Hydrogen and Helium dominated, with very small amounts of several slightly heavier elements produced during the early days. As galaxies began to organize from clots of slightly denser Hydrogen, the first stars formed. At that time many (most) were very massive O and B

types. These have very short lifetimes, sometimes burning their fuel in a few million years. Their fate is to explode as supernovae, as described on page 20-6. Even smaller stars that work through the Red Giant stage have, or will, eventually cast off a considerable amount of their elemental constituents enroute to becoming White Dwarfs.

Stars are the furnaces in which the elements beyond H, He, and some Li are created (*stellar nucleosynthesis*) by successive steps in nuclear fusion in which more and more protons and neutrons are joined into stable to unstable nuclei. These steps occur during the fusion process as a given star heats up to higher temperatures. The development of shells of elements with mass numbers greater than 2 is shown here for a star of overall mass and size similar to the Sun.

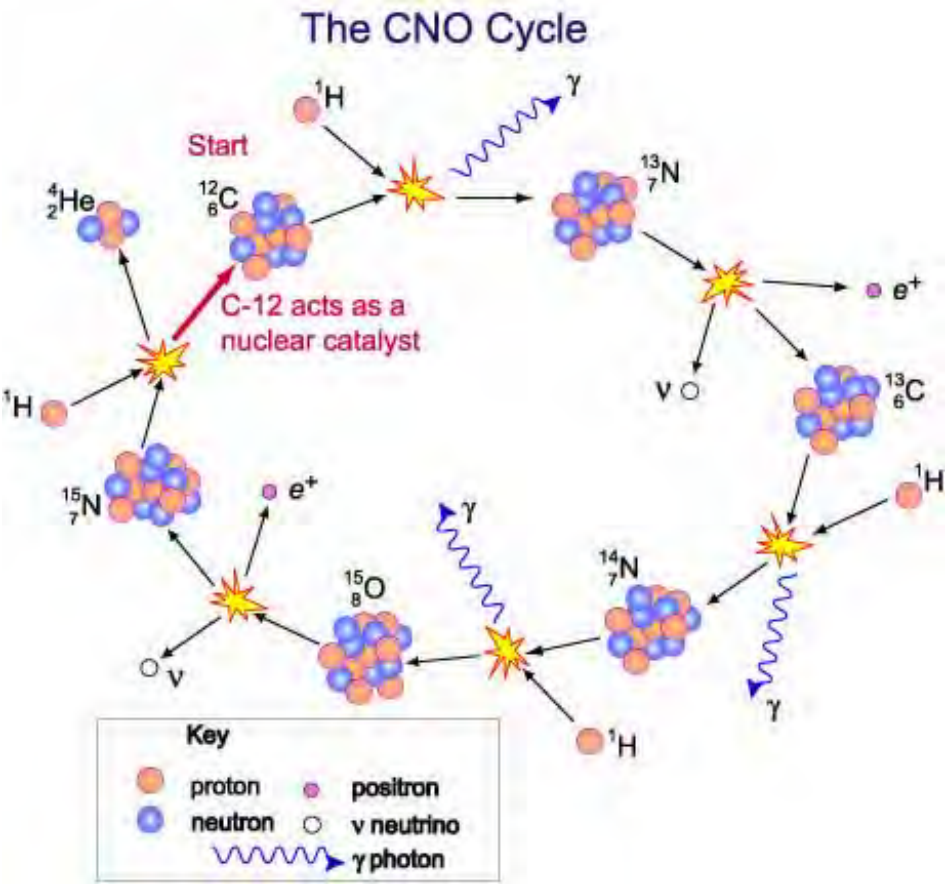


The most basic nuclear reaction is the proton-proton process, which takes place in the Sun, shown here:



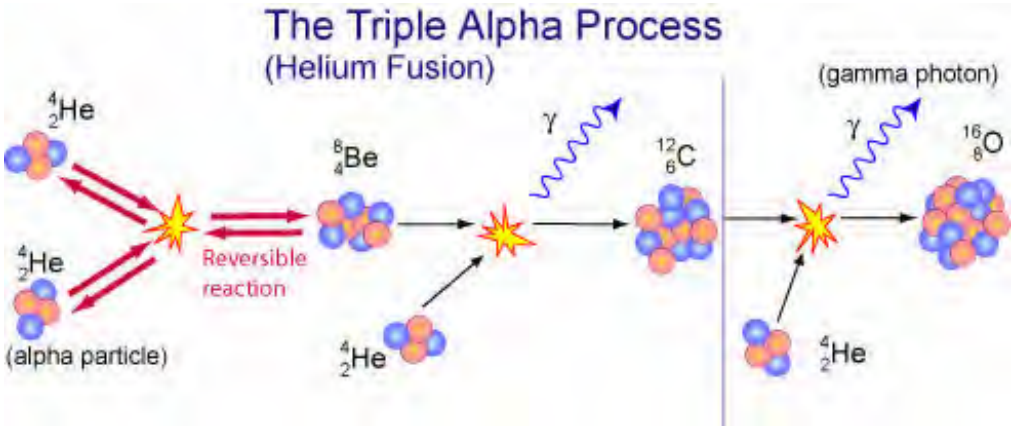
Stars with solar masses between 1 and 10 (those that follow the asymptotic giant branch [AGB] described on page 20-5) tend to burn their Helium into Carbon and some Oxygen and Nitrogen, following the so-called CNO cycle, but do not form elements of higher atomic numbers. As a massive Hydrogen-rich star contracts and experiences greater pressures, Helium is the first nuclear product within its core region. The energy released from fusion, along with continuing densification, yields higher temperatures (1-2 x 10^8 K) that transmute this innermost Helium into

Carbon (by fusion of three Helium nuclei) while producing new Helium at the next outer shell, but with Hydrogen still dominant. This CNO (Carbon-Nitrogen-Oxygen) burning cycle is illustrated here:



Once Carbon is formed in abundance, more Helium is generated as an end by-product of the CNO cycle. In this, some C^{12} reacts with protons to generate, in successive steps, N^{13}, N^{14}, N^{15} and then O^{15}. After this last step, that unstable Oxygen isotope can fuse with a proton and then decay by fission, thereby releasing an alpha particle (He^4, stripped of its electrons) causing the reversion to C^{12}.

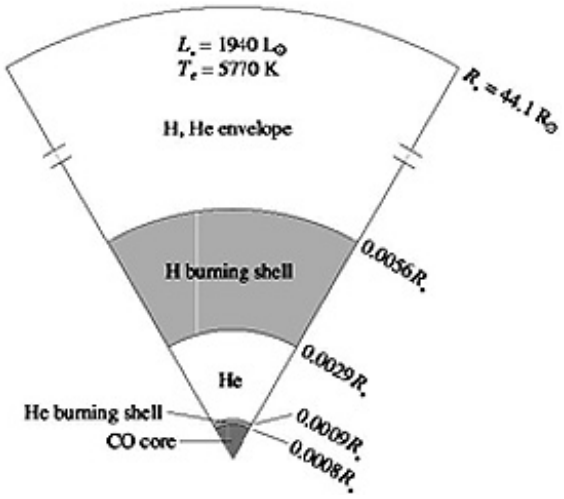
Older stars can produce Carbon by the interaction of three alpha (Helium nuclei) particles.



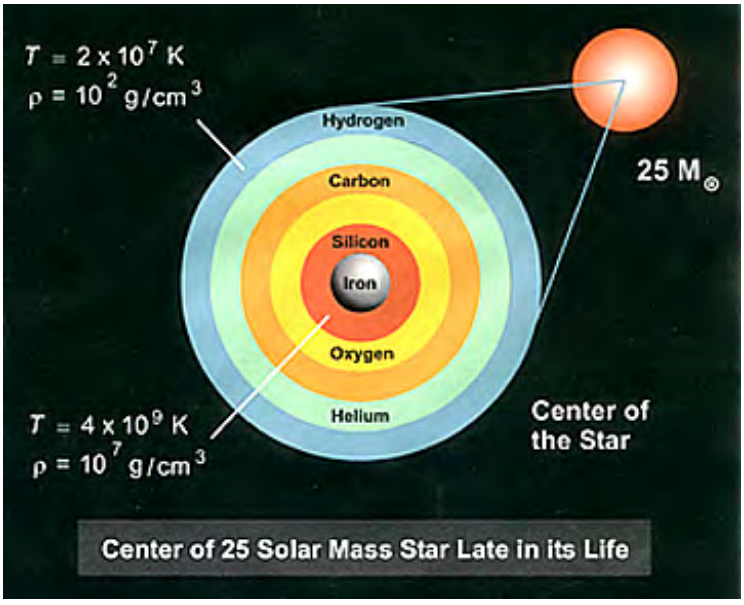
Production of Carbon (6 protons) is of great interest to humans since it is the key element in the organization of life. In principle, the fusion of three Helium atoms will produce this element but the likelihood of all three coming together at once is very low. The common pathway is the fusion of one Helium atom (2 protons) with a Beryllium atom (4 protons) that has formed as an intermediate element by fusion of two Helium atoms. But the resulting Beryllium isotope has a very short half-life and should not be available to react with Helium. However, a property called resonance obviates this problem, allowing the reaction to occur in stars that are massive enough

to reach the Helium-burning, Carbon-generating stage.

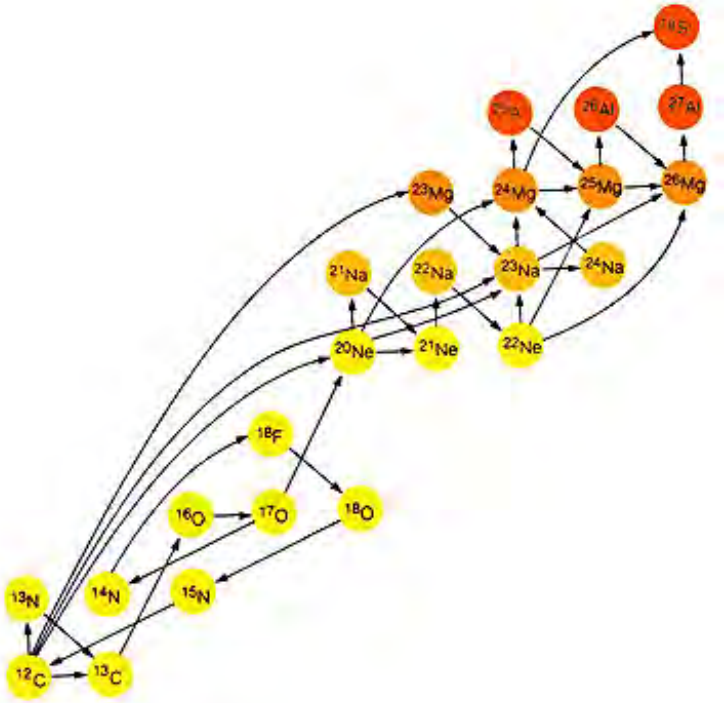
Carbon is an end product in stars at or just above solar mass that end up in the Red Giant phase. This diagram of a Red Giant shows the distribution of elements after fusion has produced these element shells;



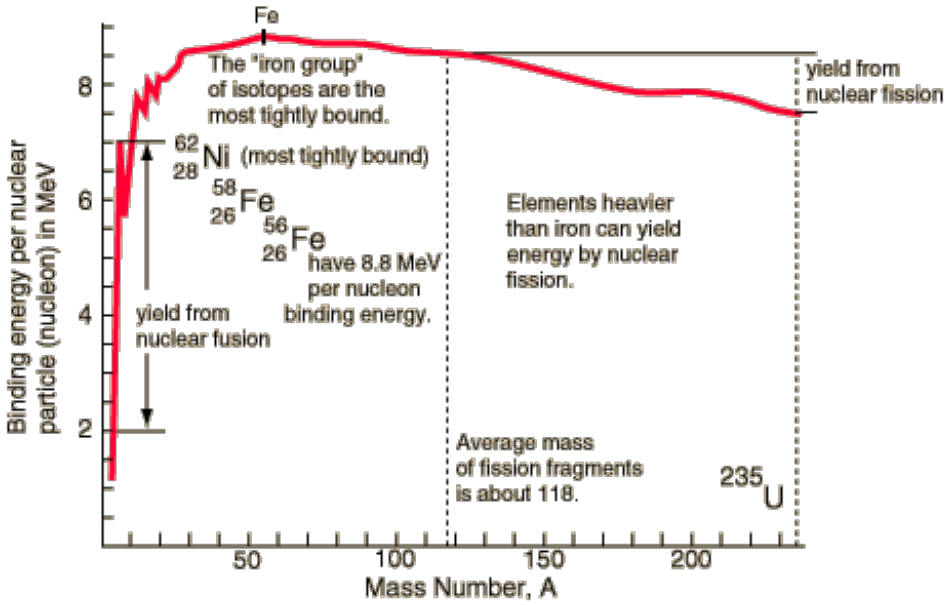
A typical, but somewhat generalized sequence of nucleosynthesis of elements of atomic number higher than Oxygen (>8) is depicted in the figure below for a star composed initially of 25 solar masses (M_O) of Hydrogen, but now is approaching (there is some Hydrogen left) its final stage of evolution (before exploding as a supernova) in which the star consists of a sequence of elements formed progressively with depth as it heated up and contracted. Stars with greater than 10 solar masses can proceed to the Iron core stage; a Sun-sized star reaches only the Carbon core stage. Here is a generalized diagram showing the element synthesis by fusion within the core of a Red Supergiant; note the variations in temperature and density proceeding to the core's center:



Almost anywhere in the fusion pathways for elements from Carbon to Iron a variety of possible reactions of nuclei are possible. This can be complicated, and is beyond the scope assigned to this page. But, to illustrate the complexity here is a simplified diagram showing element production by fusion for Carbon to Silicon.



Ever greater contraction, with concomitant temperatures reaching > 5 x 10^8 K for elements like Sodium and Magnesium, 1 x 10^9 K for Oxygen, and approaching 3 x 10^9 K for Nickel, Cobalt, and Iron, can progressively generate the elements listed in the figure up to Iron (plus others of lesser atomic numbers) in amounts proportional to the comparative solar masses indicated. Thus, a star massive enough to ultimately achieve an Iron core also contains elements of lower atomic numbers in its outer shells, broadly distributed in the relative positions shown in the figure, reflecting response to the fuel to outwardly decreasing densities and temperatures. Iron (atomic number Z [no. of protons = 26]; mass number A [number of protons + neutrons = 56]) is the heaviest element producible directly by stellar fusion. In fusion, nuclear binding energies for the new nuclides increase gradually up to Iron (A = 56) but the mass of a fused nuclide is less than the sum of the fusing constituents. The missing mass is converted to energetic particles (E = mc^2), given off as Gamma rays, neutrinos, positrons, and others; thus the fusion process is always an energy-releasing one. The binding energy after A = 56 gradually diminishes; those of higher mass numbers cannot form by fusion - they are generated either by neutron or proton bombardment or by beta decay.

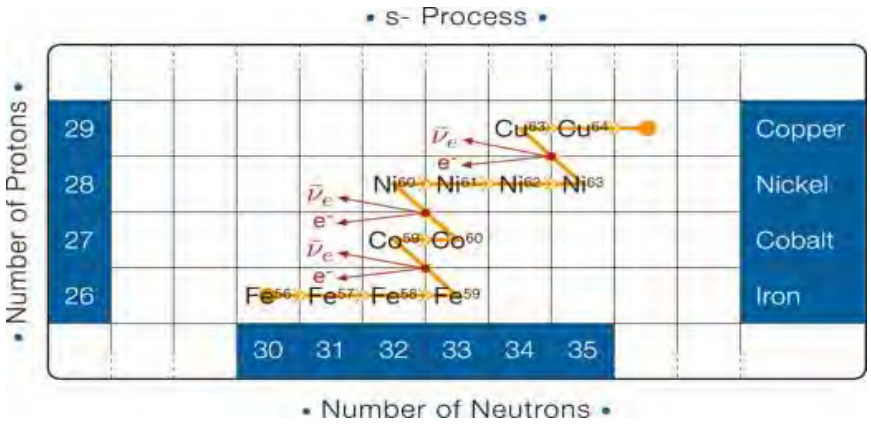


After stars which have become enriched in the elements between C and Fe made through

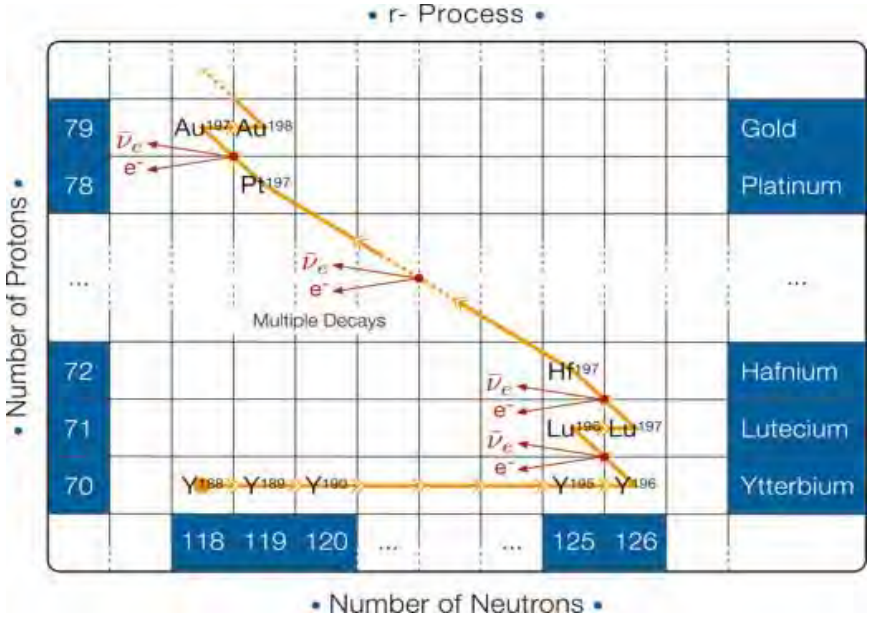
fusion undergo destruction that shrinks them to White Dwarfs, these dwarfs will be composed largely of the highest atomic number element reached as the star enters the Giant phase. Many of the White Dwarfs around 4-6 times as massive as the Sun will consist primarily of Carbon. Neutron stars, the end product of more massive star explosions, are not composed of any specific element since protons and electrons have been forced together to make neutrons, thus destroying the elemental identity reached by these stars prior to this extreme transition.

As evidenced in the above plot, elements with A greater than Fe have decreasing binding energy and to form require energy input from non-fusion processes (principally neutron capture). Because those stars capable of synthesizing elements up to Fe have masses greater than 10 solar masses, such stars at their end stage of fusion will rapidly (over spans of hundreds of years) collapse and explode (fly apart) as supernovae when their dense cores become unstable. This - the Beta process - gives rise to intense neutron fluxes that manufacture various elements including those with A > 56. There are two variants of this process: 1) the Beta S process - slow, involving neutron capture, then proton decay, and another neutron capture, which produces elements in the Iron group just above atomic number 26 and 2) the Beta R process - rapid; capture of neutron(s) before any neutron to proton decay occurs, producing heavier elements.

This is the scheme involving the S process, which takes place mainly in very luminous Red Giants



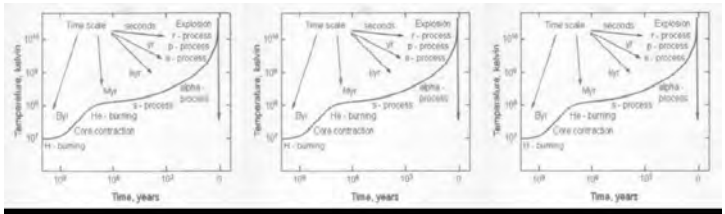
Here is a general plot that illustrates the R process, which takes place in a supernova:



Check the Wikipedia site at the link provided earlier on this page for information about the p-process, the alpha-process, and others not treated here.

As suggested in the preceding paragraphs, element synthesis is time dependent. Some

elements are formed almost instantaneously, such as those made by the R process. Others are produced continuously over long periods in stars of suitable size and temperature. This chart summarizes the time-temperature relations among the processes:



After a supernova event, the synthesized elements become rapidly dispersed into interstellar space. These heavier elements, along with H, He and the A < 57 elements (which include O, S, C, N, Fe, Mg, Ca, Al, Na, and K - the dominant constituents making up the planets), can thereafter collect into new nebulae (clouds) that may reorganize into additional stars, setting up further nucleosynthesis. The elements were mostly created in the first few billion years when rates of star formation, burning, and explosive destruction were higher than present, but the process of element production still goes on. Elemental materials not reincorporated in stars are available to organize into compounds that make up the dust, gases, and particles from which planetary bodies are assembled.

Some elements including Uranium and some isotopes of lower A number elements are radioactive and can decay by alpha or beta processes, or fission, to form new elements (with lower A numbers, i.e, numbers of protons). Over time these have built up to various extents within the Universe. Alpha decay involves ejection of a Helium nucleus (A decreases by 2; Z or mass number decreases by 4). Beta decay results when an element captures a neutron that then decays into a proton + electron + neutrino; the electron is ejected as a Beta ray (A increases by 1). Another process, nuclear fission, results in an atomic nucleus breaking apart into two new elements, each with N's significantly different from the parent (example: Uranium decay into barium plus Krypton or Lanthanum plus Bromine). The next two diagrams depict the decay history of Uranium-238 into new elements; slant decay is by an Alpha process; horizontal decay is by a Beta process; a half-life is the time for one half of all atoms of the element to have decayed (the first cycle leaves one-half; the second cycle one-fourth; the third cycle one-eigth, and so forth :

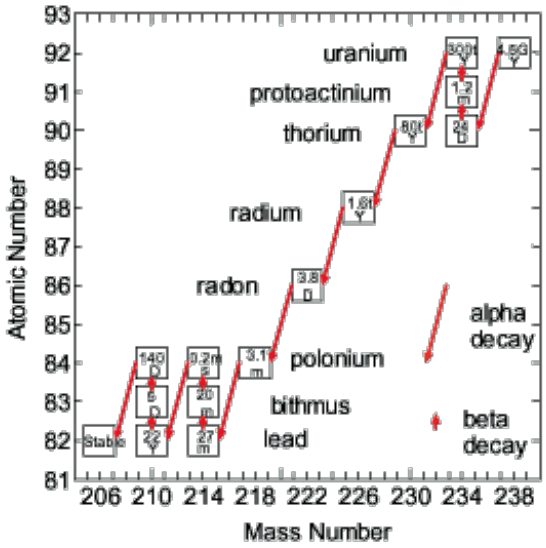
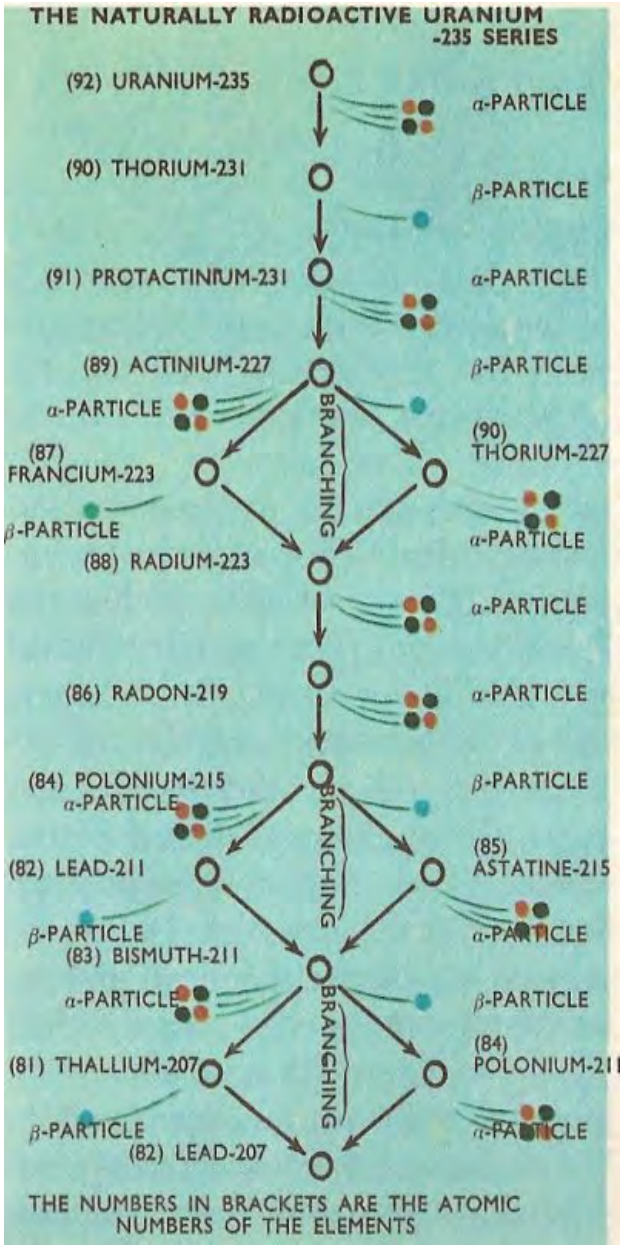


Figure 7 Decay Scheme of Uranium Series

URANIUM 238 (U238)
RADIOACTIVE DECAY

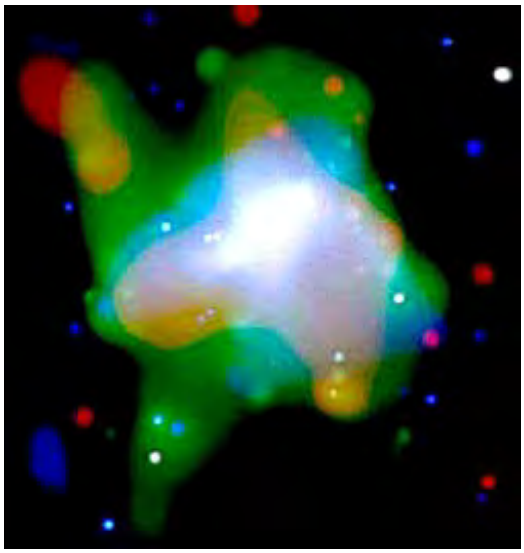
type of radiation	nuclide	half-life
	uranium—238	4.5×10^{9} years
α	thorium—234	24.5 days
β	protactinium—234	1.14 minutes
β	uranium—234	2.33×10^{5} years
α	thorium—230	8.3×10^{4} years
α	radium—226	1590 years
α	radon—222	3.825 days
α	polonium—218	3.05 minutes
α	lead—214	26.8 minutes
β	bismuth—214	19.7 minutes
β	polonium—214	1.5×10^{-4} seconds
α	lead—210	22 years
β	bismuth—210	5 days
β	polonium—210	140 days
α	lead—206	stable

Uranium-235 (which is just 0.75% of all Uranium, but is enriched to make fissionable weapons-grade material for atomic bombs) has a different decay sequence:

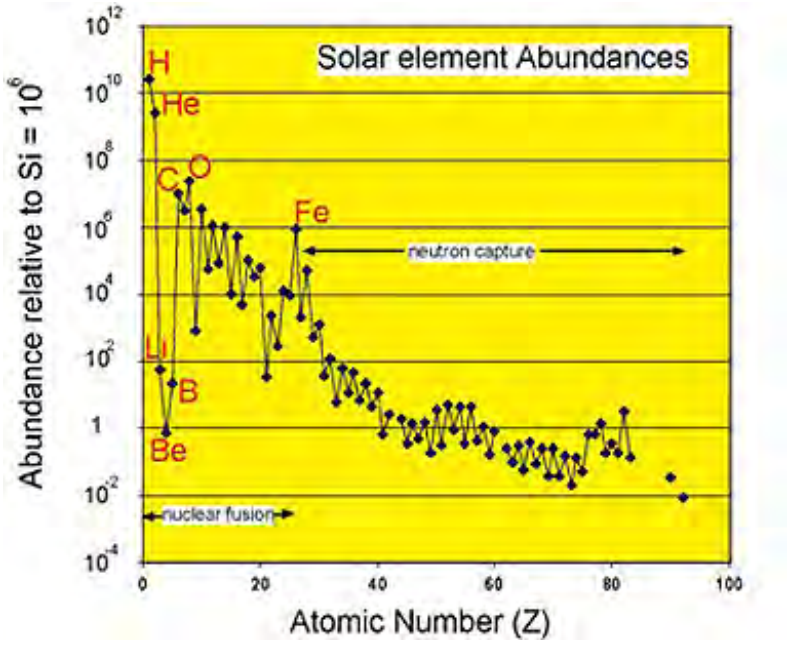


As explained earlier, stars capable of synthesizing the heavier elements up to Fe are also larger and thus fated to be destroyed explosively. In so doing, they expel and disperse the heavier elements in mixes of dust particles and gases. These re-collect over time in nebular masses that become the new "nurseries" for later (younger) stars. Many of those in turn will give off the heavier elements in surface expulsions as Red Giants strip down and if large enough as supernovae. Thus the interstellar space is continually gaining a new chemical mix of elements, tending towards loss of Hydrogen/Helium and proportionately higher percentages of the elements of the remainder of the Periodic Table. As more stars form, not only do they contain some fraction of these elements but the associated dust/gas clouds may by then have enough of those elements we associate with planets and organic matter.

There is growing evidence that a significant fraction of the heavier elements were and are being produced in the myriads of Dwarf galaxies (most still undetected because of their reduced luminosities) that pervade the Universe. Many of these undergo extended periods of star formation and correlative supernovae bursts, releasing these elements to intergalactic space. In this Chandra image of NGC 020724, a Dwarf galaxy about 7 million light years away, giant bubbles of hot (10 million degrees) supernova gases undergoing rapid expansion have been shown spectroscopically to contain enrichments of Oxygen, Neon, Magnesium and Silicon.

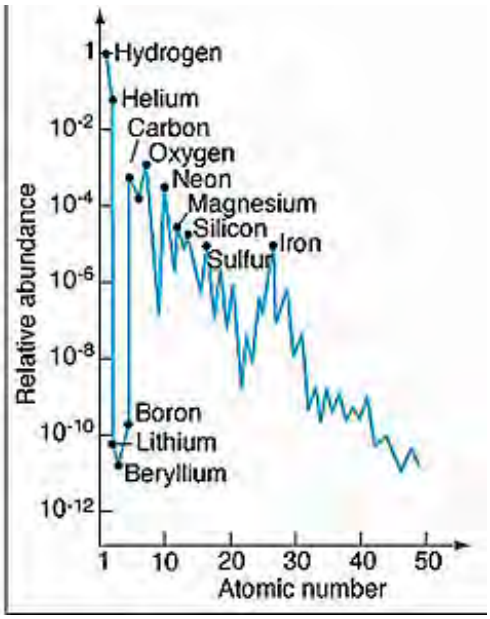


Having now reviewed how the elements are produced from Big Bang and subsequent stellar processes, we should mention something about the relative abundances of the various elements throughout the Universe. This turns out to be a difficult task for one obvious reason. Spectroscopic measurements of elements from the distant stars are strongly biased towards only those elements in excited states at or near the stellar surface. Those elements residing principally in the interior do not contribute to surface radiation in the same proportions as actually exist in a star. Only estimates based on stellar interior models can be made. The situation is better for our own star, the Sun. When element distributions are stated as *Cosmic Abundances*, they actually are rough estimates made from *Solar Abundances*.



(Note that all elements up to Iron are produced within stars by fusion; those beyond Iron (greater Atomic Numbers) are produced by Neutron capture)

This plot is similar to one purported to represent the entire Universe, i.e., the element abundances are primarily for the stars (arbitrarily, the ordinate values are logarithmic fractions of 1, which refers to Hydrogen):



Several comments about this (these) abundance plots are in order: *First*, the general trend is towards ever decreasing abundances as the atomic number increases. *Second*, there is a distinct zig-zag (up-down) pattern to the whole curve. For example, between Carbon and Oxygen there is a decrease (the element is Nitrogen); between neon and magnesium the decrease element is sodium; the largest drop is between Oxygen and neon, the element that thus decreases notably is fluorine. The reason for this fluctuating pattern is just this: elements with odd numbers of nucleons (protons and neutrons) are less stable, resulting in one unpaired (odd) proton or neutron - those that pair these particles result in offsetting spins in opposite directions that enhance stability (all this is part of the quantum theory of nuclear arrangements). *Third*, there is a huge drop in abundance for the Lithium-Beryllium-Boron (Li-Be-B) triplet. This results from two factors: 1) At the Big Bang, nuclear processes that could fuse the proper H or He isotopes into Li and/or the other two were statistically very rare and hence inefficient, and 2) Some of the Li-Be-B that formed and survived may be destroyed in processes with stars.

The cosmic abundance of these elements in terms of numbers of nuclear particles is further reported in this table:

TABLE 21.1 Cosmic Abundances of the Elements

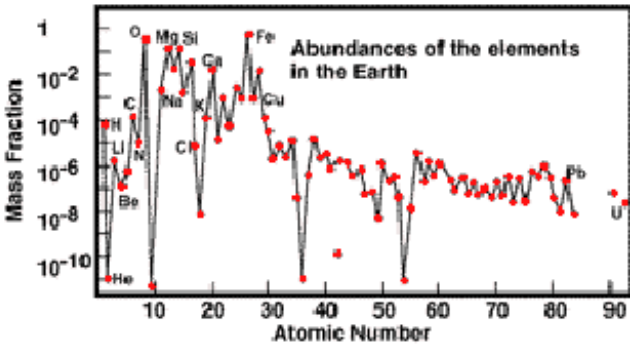
Elemental Group of Particles	Percent Abundance by Number*
Hydrogen (1 nuclear particle)	90
Helium (4 nuclear particles)	9
Lithium group (7–11 nuclear particles)	0.000001
Carbon group (12–20 nuclear particles)	0.2
Silicon group (23–48 nuclear particles)	0.01
Iron group (50–62 nuclear particles)	0.01
Middle-weight group (63–100 nuclear particles)	0.00000001
Heaviest-weight group (over 100 nuclear particles)	0.000000001

**The total does not equal 100 percent, because of uncertainties in the abundance of helium. All isotopes of all elements are included.*

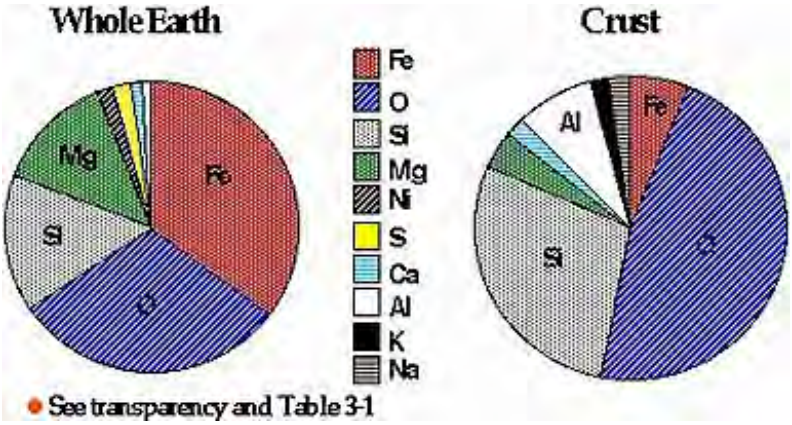
Copyright © 2005 Pearson Prentice Hall, Inc.

The best estimate for the dominant elements in the stars is almost the same as that of the Sun: 91.2% H, 8.7% He by numbers of atoms and 71% H and 27% He by mass.

The abundances of the elements for the entire Earth (probably similar in the inner planets) is quite different from the Sun and stars. Here is a plot using the same type of graph as shown above:



If we consider only those terrestrial elements that make up the crust/mantle silicates and the core (together nearly all the bulk mass of the Earth), this plot results:



A somewhat easier task is to compare stars and galaxies in terms of their *metallicities* - a ratio of all amounts of all elements with atomic numbers greater than 2 to the amount of Hydrogen + Heliumpresent. Astronomers use the word "metal" differently from chemists. A metal for a chemist includes only those elements in the Periodic Table labeled IB through VIIIB. Astronomers simply include all elements (including those with non-metallic properties) beyond He as "metals".

The "metal" composition of the Sun is fairly well known. Actually, the measurements are made on the chromosphere, the dominantly Hydrogen gas which constitutes the solar atmosphere. The source of spectral radiation, however, comes mainly from the photosphere. The relative numbers of elements within the compressed body of the Sun is different, but good estimates can be made based on element distribution models. One element that gives many strong spectral lines is Iron (Fe). This element is chosen as an indicator of the Sun's metallicity; it proxies for all metals whose amounts tend to vary systematically with the Iron concentration. Both the amount of Iron and of Hydrogen present at the surface can be calculated from the strengths of selected Hydrogen and Iron spectra derived from analysis of their absorption lines as their quantized radiation passes through the chromosphere.

From these compositional data a quantity determined as the ratio of amount of Iron to amount of Hydrogen (Fe/H) can be calculated for the Sun. It is arbitrarily set = 1. Corresponding ratios are determined for either individual stars or for galaxies (in which the Fe/H depends on the gross or composite average of these two elements resulting from radiation emitted by all stars, intragalactic gas, and halos within a given galaxy). By convention, the Fe/H ratio values are expressed as $\log_{10}$ numbers. This is a commonly used formula for comparing Fe/H ratios of stars to that of the Sun:

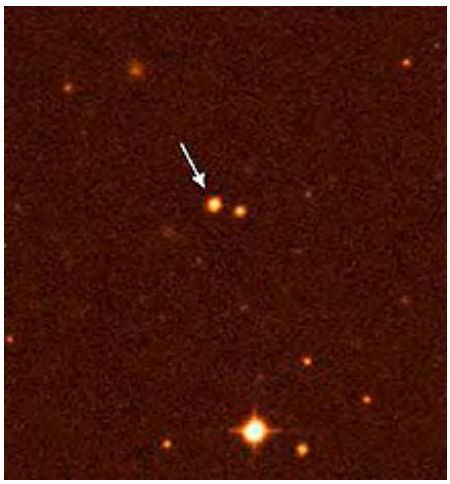
$$[\mathrm{Fe/H}] = \log \frac{(\mathrm{Fe/H})_*}{(\mathrm{Fe/H})_\odot} = \log (\mathrm{Fe/H})_* - \log (\mathrm{Fe/H})_\odot$$

Thus the Sun's Fe/H is the log of 1, which is the number zero (0). A star with a ratio of 1 to 100 yields a log value of -2; this also means that the metals abundances are 1% of that established for the Sun. A star whose log is +1 contains ten times as much metals as the Sun. Measurements for thousands of stars have established that the range of log values is from -4 (very metal-poor) to +1 (very metal-rich).

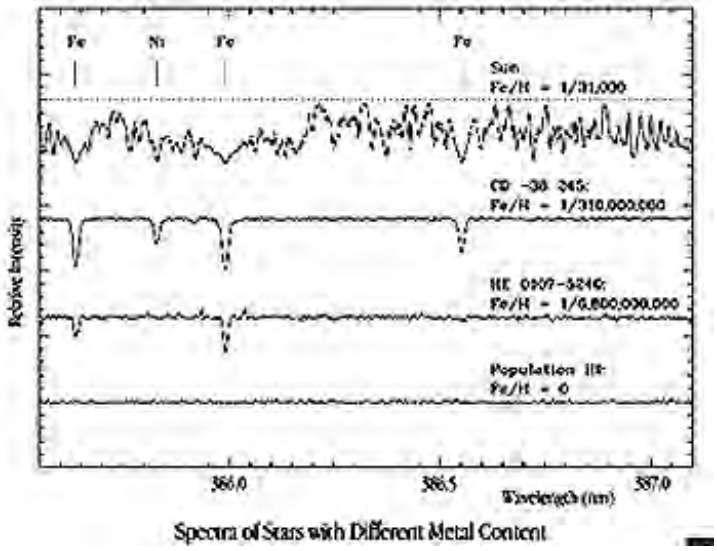
Some general observations about the characteristics of stars as indicated by their metallicities: 1) the disk portion of a galaxy has a range of metallicities, with Population I stars having values > -1, i.e., towards smaller negative numbers to positive numbers less than +1, whereas Population II stars have negative values beyond - 1; 2. globular clusters and halo stars are metal-poor (values more negative than -1); 3) metal-rich stars are in the red segment of the Color Index and metal-poor stars are blue; 4) although there can be complexities, in general metal-poor stars (initially, Population III; the first stars had a near-zero metallicity) are young in appearance (either near the outer limits of the Universe which show stars that formed in the first few billion years after the Big Bang or stars formed more recently from gas clouds that have had little contribution of heavier elements from supernovae) and short-lived; 5) metal-rich stars from F, G, K, and M positions on the Main Sequence are redder than stars of similar sizes (masses); and 6) dust around a star will make it redder.

Overall, the rule of thumb is just that a star will show a metallicity that depends on prior processes that have changed the composition of the interstellar gas in the neighborhood in which it forms. This is a function mainly of the number of supernovae that have occurred previous to the formation of the star and the amounts of metals each ejected that then became mixed into the cloud that supplies the star (and other stars growing from this cloud). Since, over time the gas composition in the interstellar medium should progressively enrich in metals, then those stars that are metal-rich tend to have organized in later stages of a galaxy's history.

From the above it follows that stars that are extremely metal-poor are likely to be first-generation and thus primitive. HE0107-5240, a small star in our Milky Way some 36000 light years from Earth, has an age estimated to be at least 12 billion years, making it one of the oldest stars examined to date. Here it is:



This star is extremely metal-poor, as indicated by this series of spectra:



Spectra of Stars with Different Metal Content

The ratio of Fe atoms to H atoms for the Sun is 1/31000. The Fe/H ratio for HE0107-5240 is dramitically lower, 1/6,800,000,000. Its composition then indicates that it had to form early in Universe history when enrichment of the heavier elements had not yet built up significantly in spite of the spate of early supernovae.

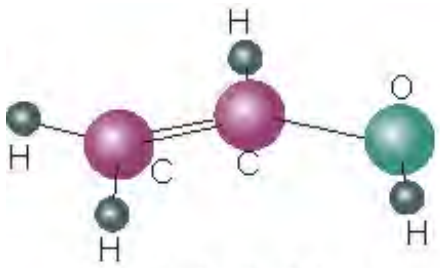
Metallicity has a practical significance to this Earth's inhabitants. Life does not apparently form around all stars from O to M types. It tends to develop around those stars that will produce planets of the right composition. Thus, the degree of element abundance and enrichment become a vital factor. The dust around a star has a composition that is related to the extent of metallicity found in that star. Only a small fraction of all the stars is likely to have a suitable metallicity that extends to its surrounding dust and gases, such that planets with Earthlike conditions are produced. We were lucky.

Now, putting the above information on element production in the context of the existence of thinking organisms: Life on Earth (and probably elsewhere) is based on Carbon and a few other elements (principally H, N, and O, but smaller amounts of P, S, and traces of Fe and other heavier elements). Where does the Carbon come from? As we have already alluded to, scientists generally agree that all matter was originally present in the form of Hydrogen, Helium and miniscule heavier elements, and that the bulk of elements up to Fe were constructed by fusing together Hydrogen nuclei (protons) with neutrons and electrons. This process involves the conversion of a small amount of mass to energy, according to Einstein's $E = mc^2$. In the 1930s, Hans Bethe showed that the energy radiated from the Sun and stars could be produced by either or both of two sequences of nuclear reactions: (1) the "proton cycle" in which protons are fused to form Helium nuclei (each having 2 protons and two neutrons); (2) the "Carbon cycle" in which a Carbon nucleus (6 protons and 6 neutrons) absorbs a Helium nucleus to form an Oxygen nucleus. But there did not seem to be any way to get from Helium to Carbon. The most obvious path would be to combine two Helium nuclei to produce an isotope of Beryllium with 4 protons and 4 neutrons, but that isotope is not stable (it takes one more neutron to produce a stable isotope) and calculations showed that it would not last long enough to pick up another Helium nucleus to yield Carbon.

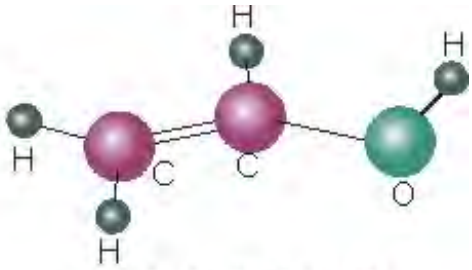
However, in 1953 Fred Hoyle predicted that the Carbon nucleus has an excited state with just the right energy to match the energy of Beryllium plus Helium, producing a resonance which allows the reaction to produce Carbon in the interior of a star. His prediction was confirmed by experiments at CalTech, and the Helium --> Beryllium --> Carbon reaction is now considered a crucial step in a more general scheme of nuclear reactions that produce all the heavier elements. Hoyle later noted that his prediction was a successful application of the "anthropic principle": the Universe must have the properties needed to allow the evolution of life, otherwise humans wouldn't be here to study it (and this page doesn't exist!). If some of the physical constants had slightly different values, the Carbon nucleus would not have an excited state with the right energy to make this reaction go, and Carbon-based life (especially **us**) could not exist.

This brings us to a subject worth exploring in some detail: organic molecules in interstellar space. An attempt was made to find a good single review article on the Internet. The result was chaos: many sites were useless, some were meagre, a few touched upon the topic without really saying much. So, nothing profound or extensive will be considered here, other than a few comments in the next paragraph. There is a synopsis of organic chemistry, astrobiology, and the nature and evolution of life on Earth given on page 20-12 of this Tutorial. That will answer some of the questions about organic compounds that have various extraterrestrial distributions.

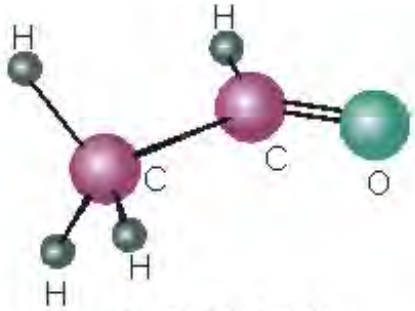
A variety of organic molecules have been found in meteorites (the Murchison carbonaceous chondrite has about 500 different carbon-based compounds). Analysis of the spectra for comets and asteroids shows them to contain various organics. Titan also has organic molecules, as discussed in Section 19. Interstellar dust clouds and giant molecular clouds enroute to becoming stars have yielded about 120 organic compounds when studied through telescopes. Carbon monoxide, methane, formaldehyde, and simple alchohols are among these. Others include PAHs or Polycyclic Aromatic Hydrocarbons. Examples are shown here:



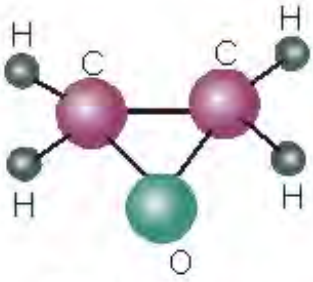
Vinyl alcohol (syn)



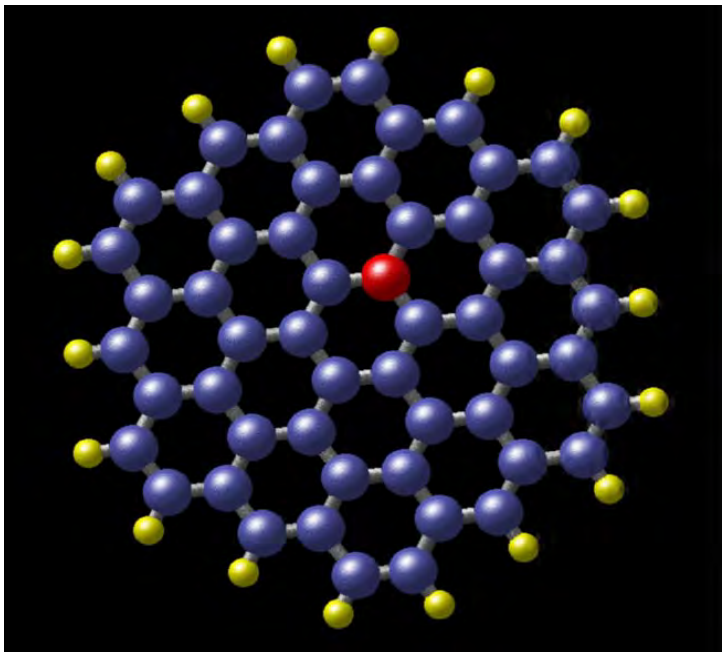
Vinyl alcohol (anti)

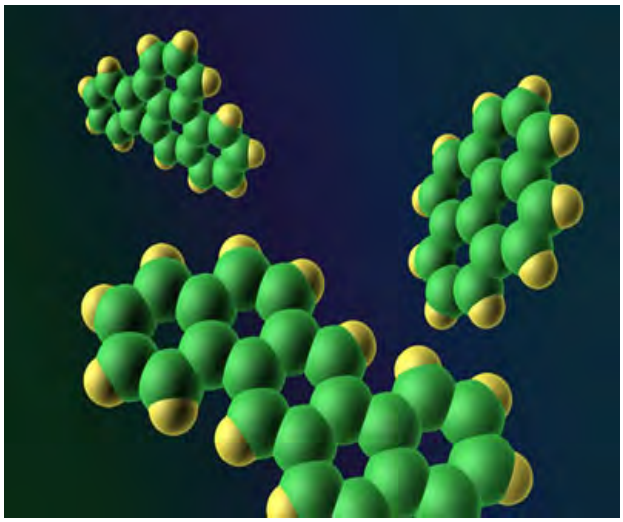


Acetaldehyde

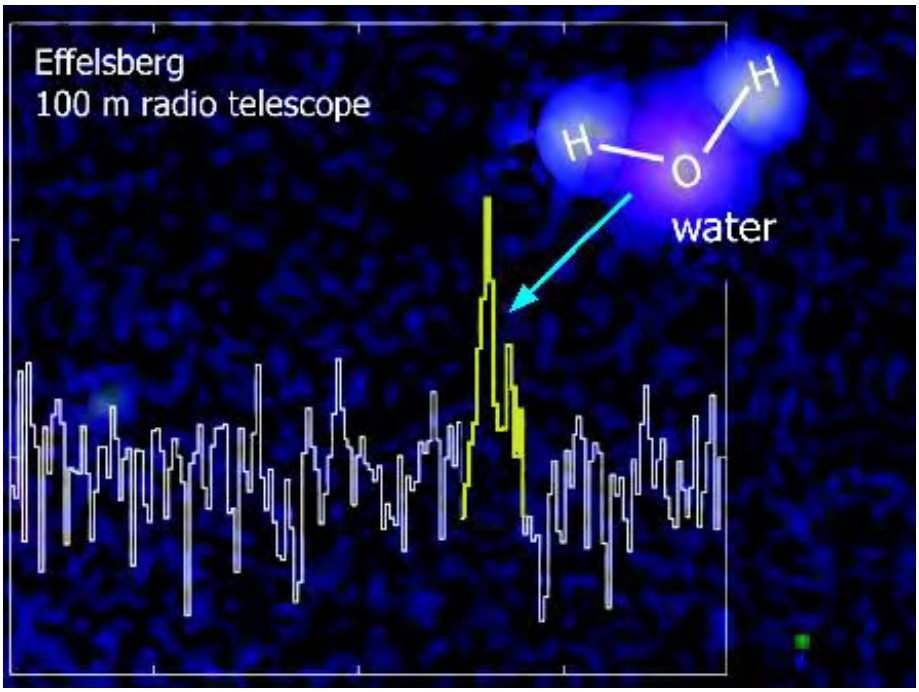


Ethylene oxide





For life to develop wherever carbon organics are present, water seems to be an essential ingredient. Water is found throughout the Solar System and in the interstellar clouds within the Milky Way galaxy. It is also present in some galaxies. The farthest from Earth is a galaxy 11.1 billion light years from Earth; here



As a bottom line aside, remember what we commented on in page 20-1: that the atoms and molecules making up everything on Earth including our own bodies have had a continuous existence from varying time spans in Universe history. Many elements have originated in supernovae that fed material from our galaxy into the cloud leading to the organization of the Sun and its planets. Hydrogen in the organisms of Earth, and in you particularly, is traceable to the very beginning of cosmic time since many (perhaps all) of the individual Hydrogen atoms and all protons in today's living creatures were formed during the Big Bang itself. Thus, parts of **you and me**, mainly as the Hydrogen in our body water and most organic molecules, are in actuality *13.7 billion years old*, having participated in the processes during the first few minutes to become end products during the Big Bang key stages. Others of our atoms came to be during stellar evolution. Most of the elements higher in atomic number than Helium have, at one time or various times, also resided in interstellar gas and dust. Most startling of all, one form of immortality guaranteed to each of us, is that our body atoms will live on for all of the future, since the fundamental precept that "matter [made during the first minute] is neither created nor destroyed" since the Big Bang at least (although atomic species beyond Hydrogen were produced later) has so far not been disproven. We truly are **Star People** who have had our moment of existence in the Grand Panorama of Cosmology. If there is no heaven or afterlife of

the spirit, then after death at best our now dispersed body constituents are assured of another kind of eternity (mostly being incorporated into inanimate objects; some atoms could find their way into living forms in future generations [this, in addition to our potential for creating life by the usual means of reproduction]).

PREVIOUS RST NEXT

Primary Author: Nicholas M. Short, Sr.

Space-Time and Expansion of the Universe

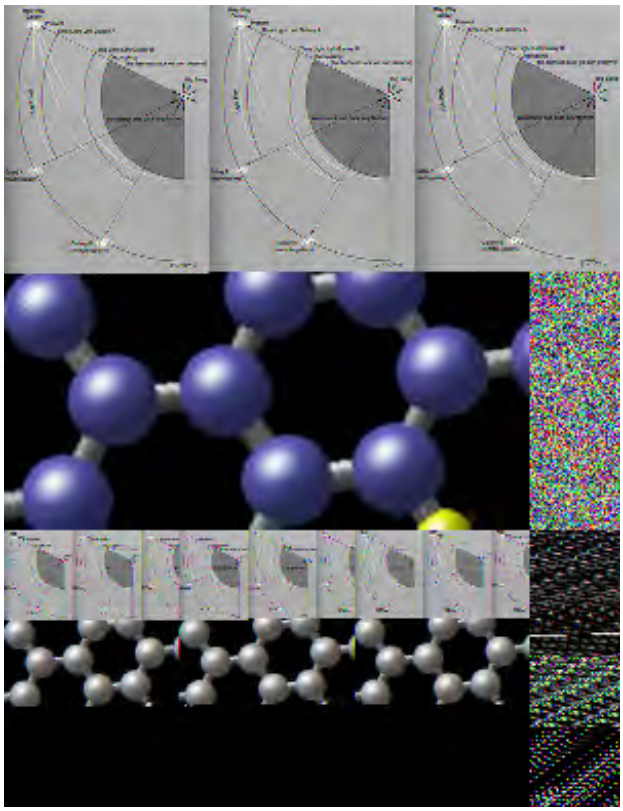
So far, this Section has concentrated more on Astronomy as such but with some consideration of topics that are primary to Cosmology. Now we turn to a more-in-depth treatment of cosmological ideas.

Shifting our attention to the creation of space aspect of the Big Bang model: The Bang effect, starting from a singularity event, requires that all matter created (photons, atoms; subatomic particles) moves apart systematically in a continuously *expanding* Universe. That the best model for the Universe's growth over time is EXPANSION has now been accepted by almost all astronomers. As a result of this expansion, galaxies, each holding a fraction of this matter (as gas and dust), are receding from one another over a range of velocities (as demonstrated by the spectral red shift phenomenon [see next page]) that depend on their relative proximities and on their apparent ages as a function of distance. The *apparent* age of a given galaxy (remember, most galaxies started to form early in cosmic time [first 1 - 2 b.y.], which is their real age) when the astronomer sees it is just the time elapsed for light to have traveled (at ~300,000 km/sec [186,000 miles/sec]) the total distance it is now from Earth.

Thus, if a galaxy is shown (see page 20-9) to be 9 billion l.y. away today (a distance of ~8 x 10^{19} kilometers), light has taken that long to get here and the galaxy must therefore have been in existence at least 9 b.y ago (and likely longer, perhaps to the first billion years of the Universe). We see this galaxy as it was then, in an earlier or younger stage of development, not as it has since become up to our time. i.e., is now. That galaxy has changed during the last 9 billion years, as stars within it have burned out and new ones formed. If somehow we could image the galaxy in its present (current) state, it would appear different from what we can actually see that represents its state 9 billion years ago. And, we cannot say anything directly about the galaxy prior to (older than) 9 billion light years, since light from any more ancient time has long since moved past the present-day Earth.

Of course, that galaxy and the Earth (within the Milky Way) were much closer to each other 9 b.y. ago and have continued receding from one another since. Today the galaxy is much further away from Earth and light leaving it now will take much longer than 9 billion years to reach us. In the early moments of the Big Bang, the two galaxies were indeed very close, as were all galaxies including those now at the maximum observable distance from Earth; since then the two have separated a distance specified by that covered by their mutual travel and increasing over the 9 billion light years of expansion. This seeming paradox is addressed by applying Einsteinian concepts of gravitationally moderated space-time (his theory of General Relativity). The age assigned to a galaxy at any distance from Earth depends on its distance (in light years) and the age chosen for the beginning of the Big Bang. To calculate this: Age we now see = Stage of Development = Age of Universe - Time elapsed during the travel time of light [l.y.] from galaxy to us. Assuming the Universe's age to be 14 Ga (recall that a "Ga" refers to Giga-annum or a billion years), the galaxy 9 Ga away shows us today what it looked like at 14 - 9 = 5 Ga after the Big Bang. Again, this is repetitious: If the galaxy formed 1 billion years after the Big Bang, we will have missed all evidence of its growth and evolution for the 4 Ga time interval prior to the 9 Ga moment in which we observe it now. To gain an idea of what galaxies older than 9 Ga looked like, we must peer back farther into spacetime, i.e., search for galaxies that can be "dated" at, say, 11 or 13 Ga, etc. in terms of measured light-years away. In the other time direction, if humans are around one billion years from now, and decide to observe this same galaxy, they will see it as it has changed (evolved) to an older stage of its existence one billion years later.

Perhaps a diagram will aid in comprehending these ideas of the effects of expansion on the perceived ages of galaxies at different distances:



From J. Silk, The Big Bang, 2nd Ed., © 1989. Reproduced by permission of W.H. Freeman Co., New York

In the diagram, the Big Bang itself is the zero point in time and space on the right end of the arcuate wedge diagram. Both straight lines from this point are non-linear time lines (the one on the right ends at 20 billion light years - an upper limit for the age of the Universe chosen arbitrarily by Professor Silk as a maximum estimate at the time he fashioned the figure. Now an age based 13.7 billion l.y. is believed to be more factual, so this illustrative diagram is out-of-date in that respect; the line on the top shows in words major events during the Universe's history). The long circular arc can be thought of both as a distance measure and another expression of the time involved in light speed travel of photons from any point on the curve to observers on Earth in the Milky Way - it is the timeline for the present day. The shaded area lies within the 300,000 year span from the Big Bang moment until the end of the Radiation Era - any event in the then smaller Universe is forever hidden from detecting/monitoring by the opacity of early Universe prior to Decoupling (see page 20-1. The first luminous matter to form after this Era (300,000 to 500,000 years after the Big Bang) has taken just under 20 (think 14 hereafter) billion years to get to Earth, covering a distance of 14 (minus 0.5) billion light years as traced by the long white curved line. Galaxy B as seen appears several billion years younger (that is, is in a stage of development at, say, 3 billion years after the BB, so that the distance traveled by its light path (white) is less. Galaxy A seems much younger still in terms of *when its light left it* for Earth; for that case, the light path is shortest.

Nevertheless, both Galaxy A and B can be nearly the same actual age, since the bulk of the galaxies are believed to have organized over a time interval of a 1-2 billion years at most after the Big Bang. (However, keep in mind that many astronomers believe that some [possibly many] galaxies have formed at various times subsequent to the era of main galaxy formation - about a billion years after the Big Bang - and would thus appear younger [in earlier stages of evolution] at various distances from Earth within the 4 dimensional mix of galaxies distributed throughout the expanding Universe.) But, in fact the appearance of Galaxy B shows it to be in an earlier stage of development than Galaxy A. This seeming paradox is simply the relativistic effect of time of travel: B took longer to get here because it is further away and at the time of departure was

developing at an earlier time in Universe history. Light from A left later, when that galaxy was in a stage of evolution that appears further along, and has traveled a shorter distance, being located also closer to the Milky Way (B is further away because it is located among those galaxies that have been receding from Earth as an observation point at faster rates; see below).

Now, consider this thought experiment: Suppose you peer through a telescope one night and happen to see the moment of detonation - the earliest stage - of 3 supernovae. They all will disappear as bright objects within hundreds of years. You might conclude that they all were happening right now or at least recently. Measuring the recessional velocity redshift of each, from which distances to observer can be calculated, you find that Supernova A is 2 billion l.y. away, B is 7 b.l.y from Earth, and C is 12 b.l.y distant. This last supernova actually occurred 12 billion years ago, the others 7 and 2 billion years ago respectively. They are indeed *not* simultaneous events, and all three in today's time have no trace of their existence (some of their materials have been incorporated in other stars and the rest are dispersed in intra- and inter-galactic space). Yet, as you witnessed them suddenly appear, you might conclude they were happening NOW - not so. The one distinction among them is that the luminosity of the three decreases in the time-distance sequence from brightest A, then B, and C with the lowest luminosity; this is simply the effect of increasing distance making objects of similar absolute magnitudes progressively decrease their apparent magnitudes.

Intuitively, one might imagine while looking outward and back in time that the farther out we see, the older are the galaxies. But as shown above, this is in fact just the opposite of distribution of developmental ages. That is one of the insights that Relativity has led us to. Also, intuitively we might (subconsciously) harken back to the belief of the ancients up almost to the 20th Century that the Milky Way in fact lies at the center of the observable Universe. Since we don't actually know where the outer boundaries of the Universe are (if the geometry described on the next page is selected to be that of a finite Universe; other models lead instead to a boundary-free infinite Universe), we cannot say where we *actually* lie within the horizons reached by telescopes. We could in fact be at (near) the center but this is highly improbable (a quirk of pure chance). More likely, the Milky Way is not equidistant from all the outermost galaxies seen so far, and is not near any center (expansion models do not allow for a discoverable center) - a positional state that may not even be valid. Oone can however speculate about the spatial distribution of galaxies relative to the singularity which doesn't have a location since the space associated with this Universe did not exist until after the Big Bang.

We on Earth are likely subject to the *Copernican Cosmological Principle* which says that there is no preferred location in the space within the Universe we dwell in. Thus, an observer at any galaxy regardless of location will see the same general picture of a Universe with galaxies seemingly moving away as the Universe expands. An astronomer on a planet in a galaxy 7 billion light years from Earth should therefore see the same patterns of movement and age relationships in a Universe that likewise appears homogeneous and isotropic (General Cosmological Principle) that we have deduced from our vantage point. He/she would, however, see a different set of sky map features - a different zodiac and different constellations (these are largely comprised of stars resident in the Milky Way galaxy).

Because the above ideas may seem bizarre and perhaps contradictory, we will reiterate for emphasis their essence in this (seemingly repitious) statement: Galaxies detected by the HST as very far away, say 10-13 billion light years, will look to us like they are among the youngest in the Universe. What we see of these distant objects is their status or appearance at that time, when they were young; we can make no claim as to their condition today, except that it is reasonable and probable that they would, if we could somehow transport ourselves to their neighborhood, resemble those we do see close by today. As the distance between Earth and the galaxies we choose to observe becomes progressively less, in terms of light years, we see them as ever further along in their evolutionary history (assuming that the vast majority of galaxies formed during the first several billion years since the Big Bang). Thus, a galaxy 2 billion light years away appears to an Earth astronomer as it was just two billion years ago - it is more evolved than a galaxy 7 billion light years away, which displays a stage of development when the

Universe was just about half its present age.

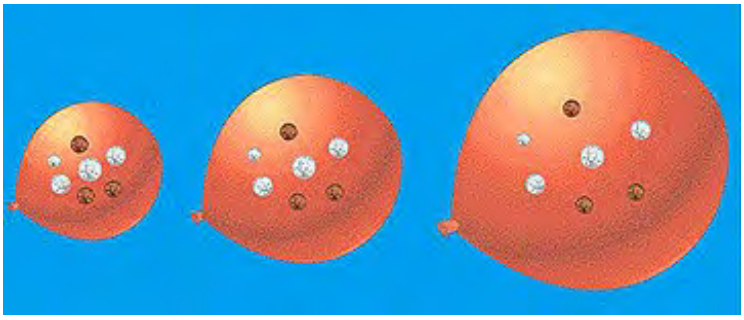
From the grand concepts of Relativity (perhaps the greatest "thought" of the 20th Century), it is concluded that space itself is the entity that is doing the expanding rather than just the matter within it simply flying apart. As one looks ever farther towards the expansion front, evermore new space appears and the matter it will contain can be said to be "created" and organized from the events ordained by the Big Bang. The expansion of space is four-dimensional (3 spatial and one time dimension); this is often referred to as *spacetime* (the first illustration in this Section, on page 20-1, is a spacetime diagram). Since the Big Bang, as time increases after the singularity-activated, ever-larger regions of the Universe become progressively visible.

It is important to note that, while space is the entity doing the expansion, there is no corresponding expansion within galaxies (their stars are not drawing apart from each other, and in fact over time they tend to contract as the stars move inward towards the usual Black Holes at their centers). A major reason why galaxies (and our own Solar System) do not seem to participate in the general expansion, which applies primarily to the large-scale Universe, is the effect of "local" gravity (from the stars within a galaxy, plus some contribution from nearby galaxies in clusters) that roughly counterbalances the enlargement of space even as it serves to maintain equilibrium with galactic angular momentum.

For the Universe, any point on the sphere is equivalent to any other point but the idea of a center on the sphere itself has no meaning. Only when one point (such as Earth) is chosen as an observing reference does any meaningful framework of spacetime geometry have any value in discussing "location". Earth, in Babylonian and Ptolemaic models, and in those adopted by Christian theology, was once held to be the center of the Universe. Later, the Sun was accepted as center. Now it has become obvious that our Solar System occupies a non-important position part way out from the central region of our Milky Way galaxy. This galaxy is as typical ("normal") as most others in the Universe. Its location within the Universe cannot be specified with respect to an ultimate center, since none can be found; it would be extraordinarily fortuitous if the Sun and the Milky Way galaxy were actually at or near the "center" of a finite (and spherical) Universe. (Thus, some civilization on a planet around a star in a distant galaxy could likewise elect to place themselves as *apparently* at the center of their perceived Universe). But, again, no real center can be fixed anywhere since all points on the expanding geometry (see next page for the spatial alternatives, which include non-spherical and contracting models) have equal weight and significance. In a sense, the fourth dimension has a "center", namely the singular point in time when expansion began.

It is difficult to draw upon ordinary visual experience to envision this expansion but the analogies described in the next several paragraphs can be useful.

To help picture in one's mind the more abstract 4-dimensional (time included) expansion of the Universe, the example is often cited of the spreading apart of polka dots on a balloon's surface that is progressively enlarged as the balloon is inflated.



This analogy is a good working description of one mode of the enlargement of space (spherical) if one simply ignores the everyday experience of a balloon having an interior - instead think only in terms of its surface; the analogy is not meant to imply that the Universe coincides with an expanding surface of a sphere within which there is nothing (but air or gas in a real world

balloon). A spherical surface has the special property of being dimensionally finite but has no edge as such; it curves continuously on and through itself. In the balloon example, its surface is in effect a 2-D curved surface. The balloon's surface can be thought of as representing the surface of an expanding sphere on which the Universe's components - galaxies, intergalactic matter, "empty" space between stars and between galaxies, radiation - are located. The universal space being generated associates with the equivalent balloon surface (however, the galaxies, unlike the dots, do have dimensional depth). The advantage of the balloon example is that it helps us, as 3-D dwellers who have problems in picturing an expanding space with Einsteinian curved dimensionality, to picture the separation of points on a finite, curved, closed space geometry, by working on the 2-D analog that illustrates some of the properties of cosmic expansion. The disadvantages: This is only a surface; the interior is empty and ignored; also the spots would stretch out on the balloon surface (ignored in the illustration) A word of caution: other expansion geometries - hyperbolic and flat - are possible and lead to different end results, as discussed on the next page.

Another commonly cited "mental picture" is the *raisin loaf* analog. Consider bread dough(with a high dosage of yeast) in which are embedded randomly but uniformly spread out numerous raisins.

The dough (envision this either in the traditional shape or as a sphere) is then baked in an oven. As it cooks, the entire loaf expands outward indefinitely (in reality this process will eventually stop). With growth, the raisins become progressively farther apart but the individual raisins largely retain their original size. During expansion the raisins farthest from any one chosen as a frame of reference will move the most and hence at a faster clip (velocity) than those close to each other. While helpful in the visualization, this model too has problems. The raisins near or at the edges do not see raisins in all directions, only inward, which violates the Cosmological Principle. And, as one stands back from the baking loaf, the bread "universe" has a conceptual center. The most serious shortcoming: The analogy does not give a sense that the outer raisins appear as though younger, i.e., represent earlier stages of development; there is no indication that large separations involve long transit times for information to move at the speed of light (almost nothing in our Earth experience can duplicate this, although the very brief delay when someone on television is communicating with another half way around the globe, which results from the large distances that the signal [moving at light speed] must travel out and back is an actual example of a relativity effect). So then, this analog is imperfect and aids only in envisioning a part of the expansion picture.

There is a third analogy that may prove helpful.This is quoted directly from Joseph Silk's *The Big Bang*, 1989 (comments in parentheses are the writer's [NMS]):

" *We can try to visualize the initial expansion (of the Universe) by imagining an immense swarm of bees (matter; later, the galaxies) crammed into a tiny hive. Suddenly the beekeeper removes the hive, and the bees rush off in all directions (at once). Any given bee will observe its neighbors to be moving away from each other. Suppose all the bees fly in straight lines but in random directions. The swarm of bees will steadily spread out, covering an ever-increasing volume, and the fastest bees will be farthest away. A simple relation connects the velocity of any bee with the distance traveled."*. To this I add: there always remain some bees close-in to the hive site; and some bees don't get too adventurous and start flying at greater speeds past other bees further out, i.e., they elect to stay in their relative positions after release. And, like the bees, within galaxy clusters, the net motion of individual galaxies tend to be in all directions (i.e., random) including

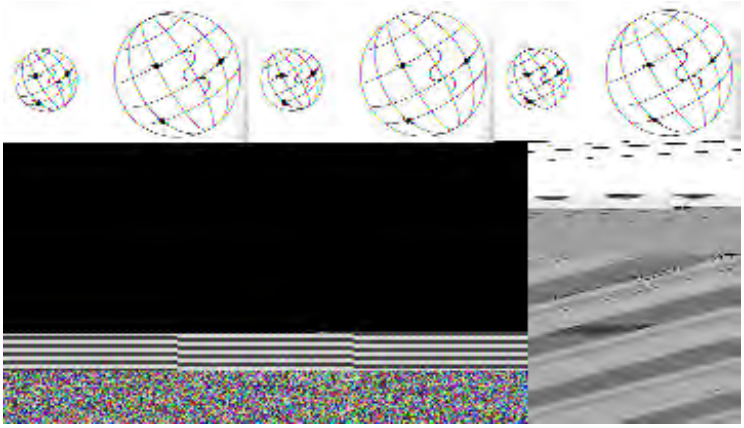
some coming towards Earth even as the cluster groups themselves are moving both outward and apart.

The bee model still doesn't quite give the proper picture. We will try here to supplement the above idea with this thought experiment: Imagine you are in a spherical room completely filled with marbles. Then, conceive of the boundary of the sphere as now moving outward and at the same time, all the marbles move away from each other, as though propelled by some gas. This continues indefinitely with the marbles progressively moving further apart. No gaps or holes develop within the expanding marble complex, so that the density of marble distribution, while ever decreasing, is uniform. As one looks around while within this array of marbles, each moving away from its neighbors as the whole group draws apart, it is impossible to determine just where the original center of the complex was - in effect, there seems to be no center.

Another failure in the balloon and raisin bread models is that it is hard to perceive that the objects farthest away must move fastest during the expansion. A simple way to visualize that idea is to carry out this rubber band experiment: Take a wide rubber band and penmark five marks on it - one [make it an X] at the center as a reference, then one dot on either side an inch apart and one more dot on each side separated by another inch from each of the inner paired dots; holding the band around your two forefingers, gradually stretch the rubber band outwards about three inches and observe the shifts of the dots against a backdrop of a ruler, placing and holding the center X at some number; the two inner dots move some distance and the two outer dots also move but over a greater distance; also, the increase in separation between inner and outer dot on either side is larger than between the X and inner dot; since velocity (v) = distance (d) divided by time (t) and the time of the stretch was the same for all dots, the outer dots are said to move apart at a greater velocity than the inner pair; this is a relative effect since any set of points along the rubber band do not actually move at different speeds as it pulls apart. This manner of motion is inherent to cosmological expansion; otherwise, if all points expanded at the same velocity along radii of a sphere from a common starting point, all would simply migrate outward as an expanding shell creating an increasing void bounded by their inner edge (here, the balloon analogy may help to picture this).

With the sophistication of the Internet it is now possible to create "movies" that help one to picture the progressive expansion of space. One has been put together by Ned Wright, whose Cosmology Website was cited in the Preface. Click on his name and watch the Universe grow.

Perhaps these analogies and movies have assisted you in picturing some aspects of cosmic expansion, with its relativistic time delays. We need now to consider expansion in a more quantitative way. Let's delve into another pictorial way to visualize this notion of expansion with the help of this diagram (a variant of the balloon case) which presents the process as a 2-D portrayal using circles (the concept depicted works just as well for the 3-D [balloon] version):



From J. Hawley and K. Holcomb, Foundations of Modern Cosmology, © 1998. Reprinted by permission of Oxford University Press, Oxford, Great Britain.

The circle on the left depicts a sphere with a radius r_1 on which a coordinate system (essentially, latitude and longitude lines) has been traced. That describes expansion from an initial point (radius r_0). As expansion continues, the circle on the right now has a radius r_2. The coordinate system has correspondingly expanded so that the coordinates of any point, such as locate any of the three "Saturn" discs (note that they remain constant in size even as they separate; this is analogous to the above-mentioned statement that galaxies themselves do not expand in proportion to space expansion), have changed only in scale. From this one can define a basic function called the Cosmic Scale Factor, given by the symbol R(t) which describes the changes in spatial separation (lengths or distances in a three-dimensional Universe) in an expanding system as a function of time. Scale factor is not the same as "scale" in the conventional sense. We can refer to a model of a sailboat as at a scale of 1/40th the size of the same boat as it really is used; scale in this usage does not have a time connotation.

As a hypothetical example, let's say that the Scale Factor is 3 - in some unit of time, two galaxies separated by a beginning distance of d will after t time become separated by 3d. The Scale Factor considers this question: How much does the distance between two points in an expanding Universe increase between t_1 and t_2? The answer is given by the scale factor which in turn relates to the rate of expansion.

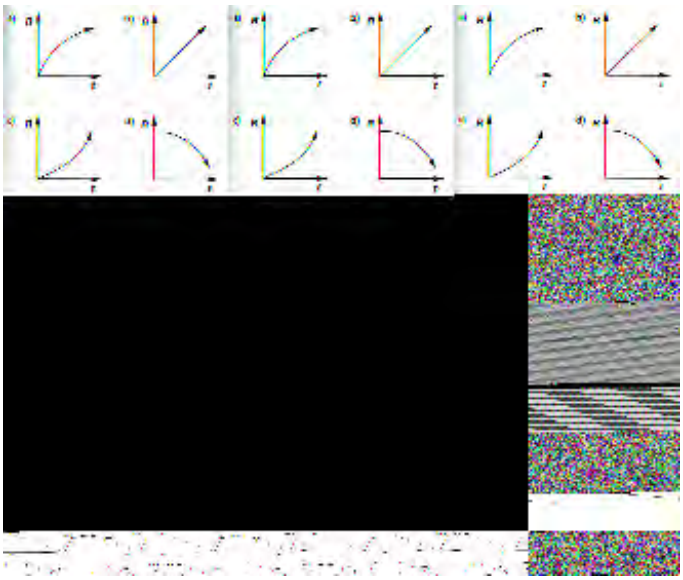
In one mode of application to the expanding Universe, the current scale factor is set at 1 and all preceding factors are decimal fractions of 1. So, a scale factor of 0.001 at, say, 400000 years after the Big Bang would indicate the Universe to be 1/1000th its present size, or a Hubble horizon of 13.7 million light years. Used this way, the scale factor is a direct measure of the extent of expansion at the time specified.

For the general case, this simple equation applies to cosmic expansion: $\mathbf{R(t) = r_n/r_0}$, where r_n is the radius at some specific time and r_0 is the initial radius (for the Universe, the singularity point). Thus, the amount or rate of expansion (or contraction) can be adjusted by a given Scale Factor; if not defining a linear function one value will yield a faster (slower) rate than another that is numerically less (greater). For a given span of time, separations (length spreads) will be greater for higher R's than lower ones. The coordinates are said to be co-moving, that is, they enlarge during expansion but all x, y, and z points referenced to them scale proportionately with R while maintaining their same relative positions. The Scale Factor is a fundamental geometric property that is relevant to a description of an expanding Universe.

The scale factor R(t) participates in expansion in what is known as the Minkowski Expansion equation:

$$\Delta s^2 = (c\Delta t)^2 - R^2(t)(\Delta x^2 + \Delta y^2 + \Delta z^2$$

Four general modes of change of R with time are depicted in the next diagram. Note that in three of the four cases shown R varies in magnitude with time.



From J. Hawley and K. Holcomb, Foundations of Modern Cosmology, © 1998. Reprinted by permission of Oxford University Press, Oxford, Great Britain.

Graph **a** shows a decreasing rate of expansion, **b** a uniform or constant (linear) rate, **c**, an increasing rate, and **d** a negative rate of expansion (i.e., a contraction).

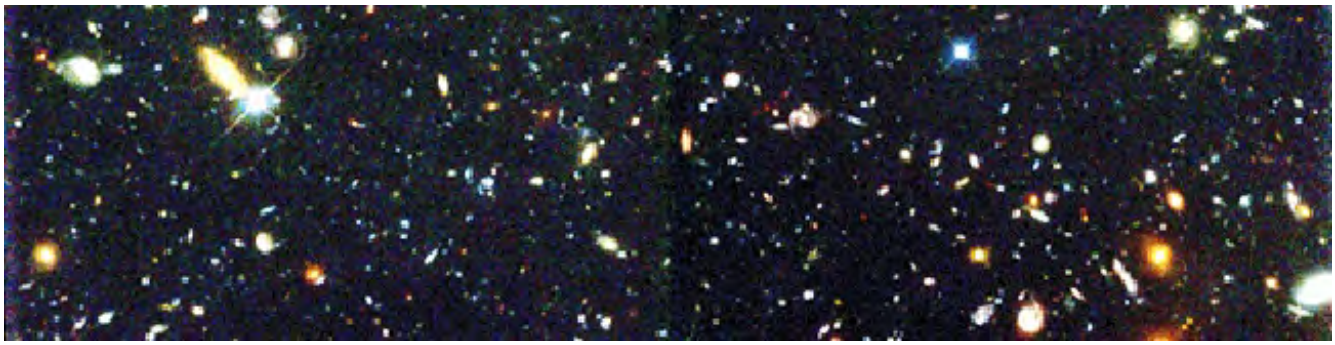
Note that for a given increase in expansion over some time from t_1 to t_2, points that are farther apart at t_1 expand at progressively greater velocities than those nearer each other; thus, they cover greater distances in a unit of time (we shall see on the next page that this ever increasing velocity outward is associated with progressive increases in wavelengths of light as shown by the redshift phenomenon).

This can be further elucidated with this diagram:

From J. Hawley and K. Holcomb, Foundations of Modern Cosmology, © 1998. Reprinted by permission of Oxford University Press, Oxford, Great Britain.

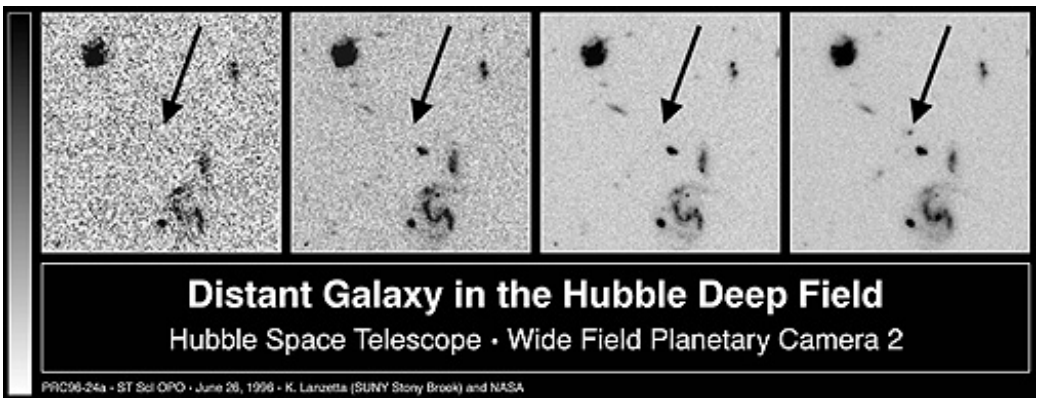
Let the upper row represent the position of three galaxies at t_1 and the lower the later expanded location at t_2. The elapsed time is $(t_2 - t_1) = \Delta t$ (Delta t = a finite interval of time). Initially, each galaxy is separated by a distance d. Following expansion, A is now separated from B by 2d and from C by 4d. Therefore B has moved with respect to A (the observation position) at a recessional velocity (d/t) of $(2d - d)/\Delta t = d/\Delta t$ and C from A at $(4d - 2d)/\Delta t = 2d/\Delta t$. Thus, the velocity of recession of C with respect to A is twice that of B to A. (Returning to the balloon analogy, one can see that farther dots recede faster than closer ones relative to some dot chosen as the point of observation.) The relative velocities will depend on the Scale Factor. As determined from red shift studies (next page), in this dynamic Universe any two galaxies are moving relative to one another at different recessional velocities which depends on their distance apart; the velocity between one of these galaxies and still a third that is twice as far away will be double (twice) that of the first pair considered.

Let us turn our attention to what has been learned in the last decade about the outer reaches of the observable Universe. This next illustration is a panorama of galaxies imaged by HST in which some of the light points are located in what has been called the Deep Field region, i.e., in that region of space with galaxies that are the farthest from us, in terms of light-year travel distances. This was discussed earlier on page 20-3a.



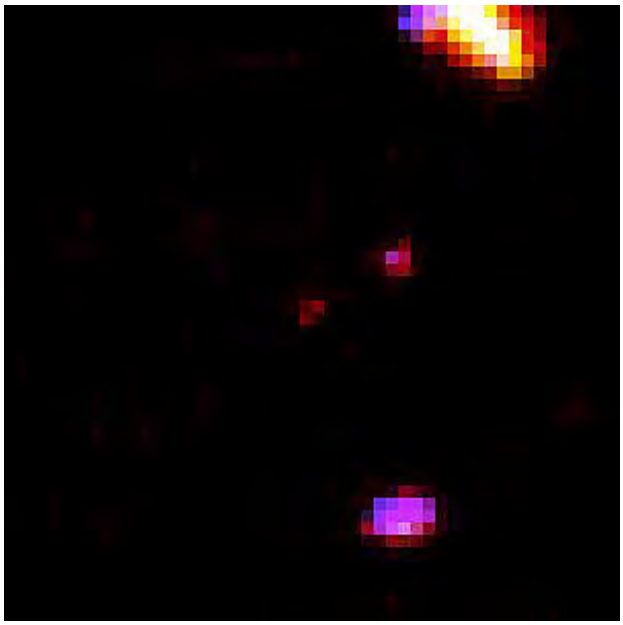
The apparent 2-D aspect (i.e., flat; lack of depth) of this image is misleading in that many of the galactic bodies (stars cannot be resolved) - generally, the ones appearing to be larger, with discernible structure - can be shown to be relatively close. Size, then, is mostly a function of distance although galaxies do vary in dimensions to some extent. Most of the smaller ones may be 9 to 12 billion light years or more away. Those objects farthest way - near the edge of the (so-far) observed Universe - appear as tiny blue specks (about 1,000,000,000 times fainter than can be detected by the human eye); their red shifts (see next page) of greater than 3-4 indicate them to be the fastest moving in the expanding Universe as predicted from the Big Bang model.

A few of these extremely distant objects may well be galaxies that formed in the first billion years. An example below (which also illustrates how the Hubble scientists go about the process of making their observations and drawing their deductions) shows a small region of the Deep Field (and several closer galaxies) as imaged by HST in four spectral bands from the UV to the Near-IR.

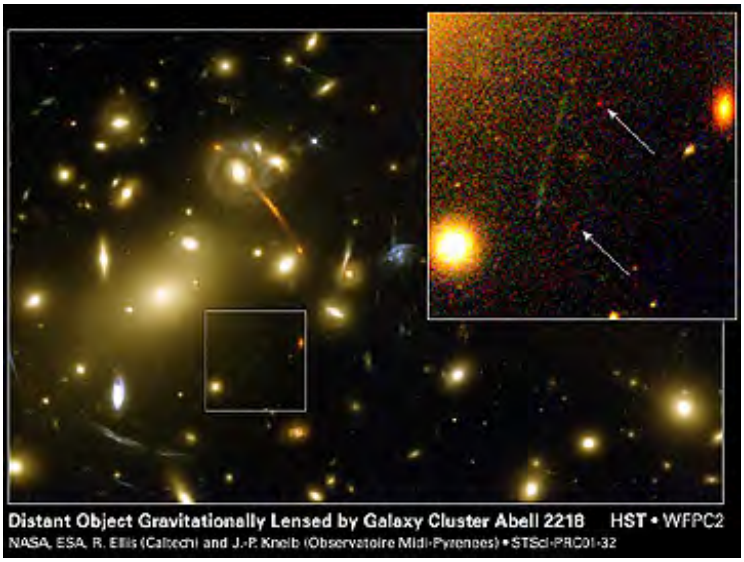


The arrows point to a location within this region in which no object appears in the left three panels (UV and two Visible bands) but a dark spot under the arrow in the right panel (Near-IR) shows an object (presumably a galaxy) in that position. Its red shift (see next page), now out of the visible red into the near infrared, is great enough (> 5) to fit an object near the time-space edge of the Universe corresponding to the earliest period of galaxy formation. Thus, detection of very distant galactic objects is very much dependent on wavelengths sampled.

The more distant galactic objects detected so far (by the Hubble Space Telescope [HST]) are at least 7 billion light years, and some perhaps 10-13 or more billion l.y., away from us. HST has now found galaxies whose redshift distance calculates out as 12 billion l.y.(and one at 13 b.l.y.); these are so faint that the view seen here represents an exposure time of 36 hours as the NICMOS (Near Infrared and Multi-Object Spectrometer) instrument gathered radiation from a very narrow field of view pointed near the top of the Big Dipper.

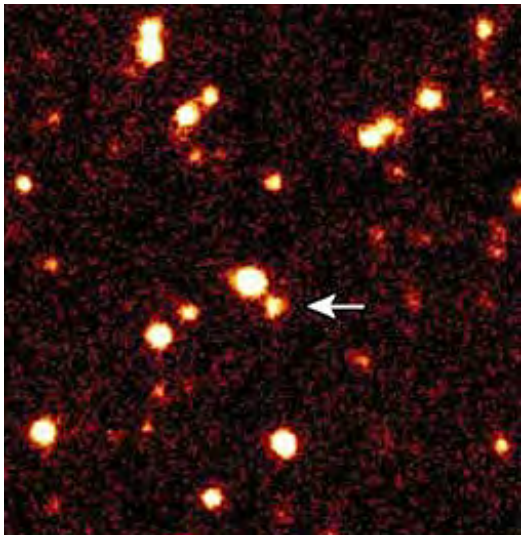


In October, 2001, a sighting of what then is the farthest galaxy to be detected was reported. The larger field shows the galactic cluster Abell 2218 which lies a distance of about 2 billion light years from Earth. Look at this image, made using the HST in conjunction with the Keck Telescope:



The small square in this image is enlarged and inset in the upper right. Within it, arrows point to (apparently) two red spots. This double image is made visible by the gravitational lensing effect of the massive Abell 2248 cluster. The pair is actually a single small galaxy at a distance from Earth calculated from a redshift of z = 7.1 to be 13.4 billion l.y. (in this case, the Hubble constant that gives a Universe age of 14 billion years was used) . Thus, the galaxy has presumably formed within the first billion years of Universe time since the Big Bang and is visible only because of the emergence of radiation after the "Dark Ages" of the early Universe (see page 20-1). The reason for the double red dots is another case of the effect of image splitting owing to the Einstein's General Relativity mechanism of optical bending of light as a result of gravitational influence of mass, concentrated in galaxies and stars (see Preface to this Section). This Einstein mechanism is thus responsible for the gravitational lens effect exerted largely by the Abell cluster, which magnifies this small (500 light years across), still growing (in its infant stage) galaxy that gives us an insight into the appearance of the first galaxies.

That distance record may have since been supplanted by a new value of 14 billion light years. Japanese astronomers, using the powerful Subaru telescope, made multiple exposures of the objects in a small segment of the celestial sphere. They added (superimposed) the luminosity measurements from these repeat observations so as to get a brighter image. Thus:

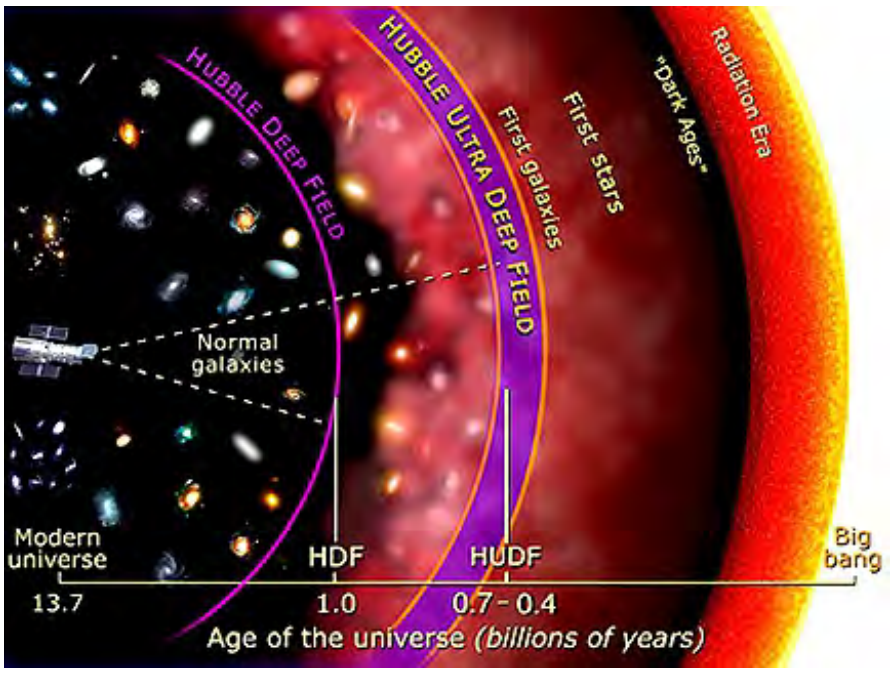


The notable brightness and irregular shape of this galaxy are attributed to an hypothesized large number of young giant stars undergoing explosions (most likely, supernovae), with its luminosity possibly added to by rapidly forming quasars. It is conceivable that this galaxy (and presumably others) is actually undergoing destructive disruption and may have since disappeared.

As the galaxies began to form and afterwards evolve - producing visible entities whose relative motions with respect to one another can then be monitored - they all continued the earlier expansion motions imposed since the Big Bang. This, in effect, means that they are all moving generally in the direction of expansion while at the same time are each moving apart from its neighbors (thus, all galaxies tend to spread apart from one another, although they occasionally can cross paths and collide). In any (and all) region(s) of the Universe, such as the Local Group which contains our Milky Way and many nearby galaxies, the movements of these galaxies, while expansive overall, tend to be in various directions including towards one another (thus ensuring some collisions), but with the majority outwards. A few, such as Andromeda, are currently moving towards us. And whenever galaxies that form a group or cluster are examined, some proportion of the total will show shifts towards the blue (not necessarily into the blue). This multi-directional motion pattern results from gravitational interactions that mutually perturb their relative movements (see the bee analogy above). This is observable only nearby; at great distances the larger red shifts swamp any variation caused by more immediate conditions around the observed galaxy(ies).

To sum up, General Relativity demands some degree of curvature for expanding space. This is controlled by the balance of gravity against the forces imposed by the initial Big Bang. The distribution of matter determines the geometry of space. Space in some way will curve back on itself, maintaining the requirement that it has no (edge) boundaries. If space completely curves back on itself, the Universe is closed; if curved in other ways, it is open. We can only see that part of the (visible) Universe in which light has had time to get to our observing station. In any mode obeying General Relativity that recognizes perpetual expansion, the Universe will be infinitely large; this means that it will grow continuously forever. As time from the present moves onward, we will see further beyond the present horizon; thus in another billion years, the Universe will be at least that much larger by the amount of expansion that has taken place in this interval and the most distant galaxies will then be a billion light years further "out" than in relation to their positions today. But, one must remember this effect: Any two galaxies moving generally in the same outward direction during that billion years are both in forward motion but since the motions are relative, what has changed is their new separation a billion years hence. And, keep in mind that there is no actual center in the expanding distribution of objects in space; from our observation location, however, we seem to be at a "center" in that all galaxies appear to be receding from us (the exception: a few galaxies - which are indeed moving along with the expansion - will appear as though moving in our direction (like a bee that temporarily turns inward toward the hive point).

Review this cosmological history diagram to affix these ideas in your mind:



Having read this page, you are now urged to read this online Scientific American article on Misconceptions about the Big Bang.

PREVIOUS rst NEXT

Primary Author: Nicholas M. Short, Sr.

Evidence for the Big Bang and the Expansion of the Universe:

The Redshift; Stellar/Galactic Distances; Age of the Universe; Cosmic Background Radiation; Expansion Models; Dark Matter/Energy

What evidence leads to and supports the Big Bang model? A good review of the resulting expansion (and calculated rates) and ages derived from these observations can be found in a Scientific American article (October, 1998; pp. 92-96) prepared by Dr. Wendy L. Freedman.

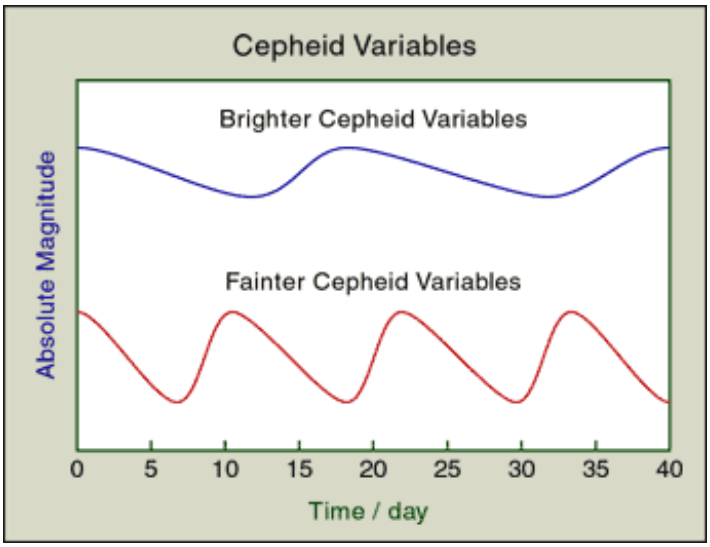
Two accepted lines of proof for the Big Bang, already described, are restated:

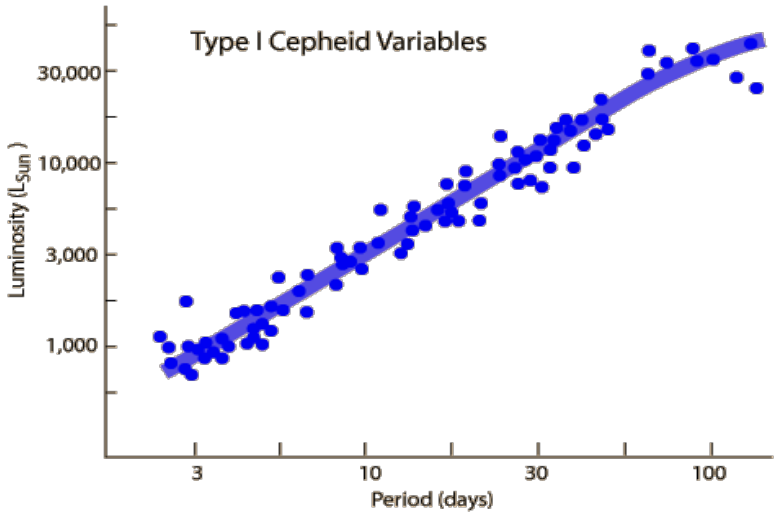
1) The details of the creation physics and progressive emergence of various elementary particles during the first minute of the Big Bang (the Standard Model and its variants; see page page 20-1) are consistent with a model based on Big Bang precepts; these particles are the outcome of a history that can be predicted and explained by Quantum and High Energy Physics, that is, the theoretical production and sequence of particles seems verified by the observed amounts of H, He, and Li atoms in the Universe; and 2) The observations, particularly from HST, of the farthest galaxies as being more primitive in appearance and development - are precisely what is expected from the expansion model in which those parts of space (in which the galaxies are embedded) that have moved the fastest are now the most distant. Thus, we see them in earlier stages of evolution when they were younger as we look back in time outwards from our frame of reference,.

But, even more convincing are two other physical observations that are best explained by a Big Bang origin for the Universe, especially in terms of its expansion behavior: redshifts of light (towards longer wavelengths) from the stars as a composite source in galaxies and cosmic background radiation.

The Hubble Law and the Cosmological Redshift

In the 1920s, Edwin Hubble made revealing observations through optical telescopes. Using Cepheid stars as distance indicators he found that some of these were much farther away than any of those in the Milky Way. He knew the following about Cepheids in the M.W.: 1) they brightened and dimmed in a regular pulsating way, and 2) the higher the frequency of pulsation, the brighter (more luminous) the stars were in their radiation output (since distances to other stars adjacent to the Cepheids were fairly well known [see below], the luminosity of the Cepheids could be deduced). These two plots establish these ideas:

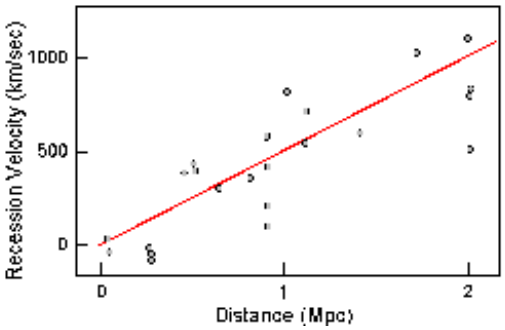




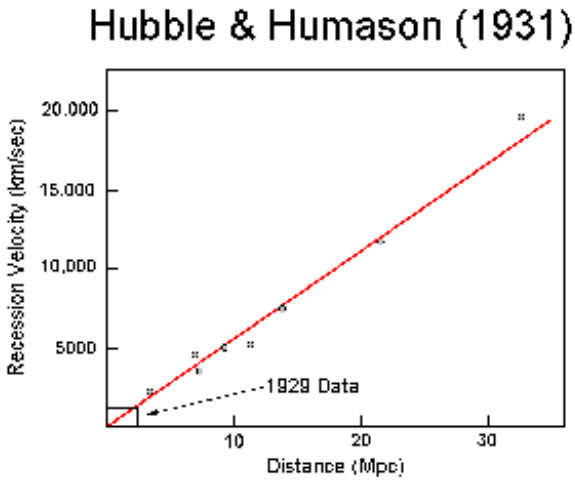
As early as 1924, Hubble had found Cepheids in a great cluster of stars we know as the Andromeda galaxy. When he applied the period-luminosity relationship to these Cepheids, they were notably dimmer (less luminous) than he expected. The most logical conclusion: the host Andromeda galaxy was much farther away (hence, the relative dimming) than any stars in the Milky Way, whose size had been established as about 100000 light years. Thus, the Andromeda collection of stars had to be a galaxy beyond the Milky Way. This deduction is one of the great achievements in Astronomy and all of Science for that matter.

Relying on more Cepheid measurements, by 1929 Hubble had observed various galaxys that were less than a billion light years from Earth but outside the M.W. Here are the data he published then:

Hubble's Data (1929)



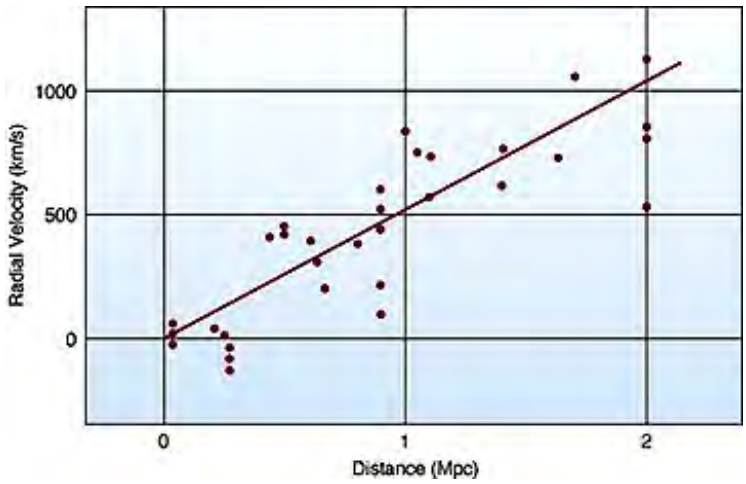
Working with others, he extended his observations to even farther distances, publishing this plot in 1931:



Comment: This discovery was worthy of a Nobel Prize. But in those days, Nobels were awarded for Physics but not Astronomy. Hubble championed the cause to change that and after his death, astronomers became eligible for this prize.

Hubble's other great discovery is the now strongly confirmed fact that the velocity of expansion (of the Universe) is directly proportional to its distance from the point of observation (for us, on Earth or in nearby space).

Hubble noted that, as recessional velocities V_r were measured for stellar sources over the range of astronomical distances D that could be measured accurately at that time, the plot of V_r/D disclosed a straight line relation whose slope has a value H, known as the Hubble Constant, named after him. This, the Hubble Law, is the fundamental statement of the Big Bang model. We repeat here his first published plot of velocity versus distance.



From Astronomica.org

The resulting straight line plot is easily described mathematically, in the basic Hubble equation:

$$\mathbf{V_r}\ \text{(velocity of recession)} = \mathbf{H_0 \times D}\ \text{(distance)}$$

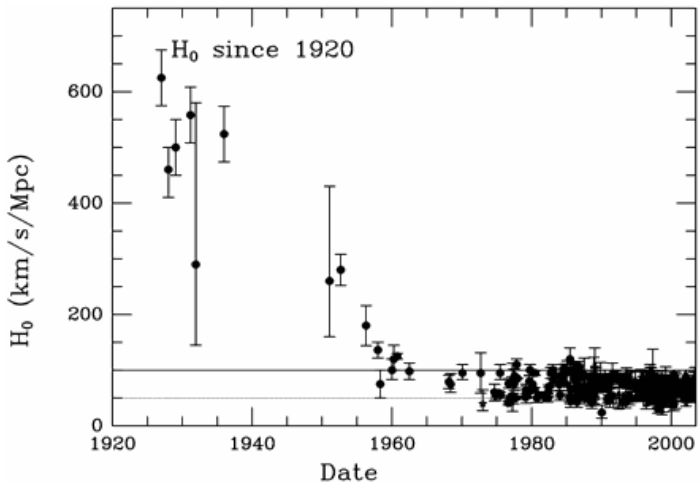
The scale constant is designated by the letter H, and is called the Hubble Constant (stated as H_0, which refers to the present day value; H has been different in the past). It is normally given the units of Km/sec/Megaparsec (an alternate form is Km/sec/million light years). The current value for H_0 is 71.4 Km/sec/Megaparsec or 21.5 Km/sec/million light years). The uncertainty in the value is about +/- 10%.

The prime information derived from this equation is that **objects (such as galaxies) appear to travel at ever increasing velocities as their distance from the observer (Earth) becomes ever greater**. The upper limit to expansion rate is the speed of light (although some

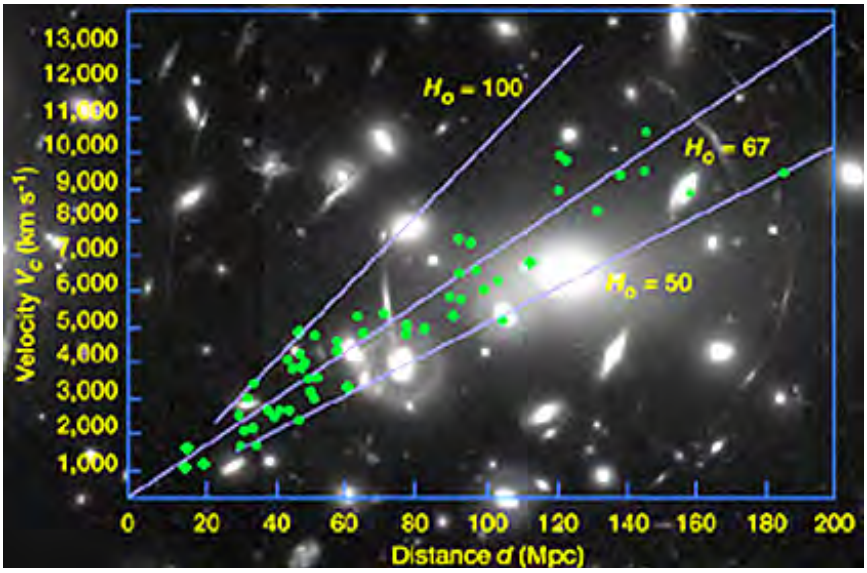
interpretations of inflation suggest that this huge leap in dimensional enlargement occurred at greater than light speed). The current rate of expansion is specified as one light year per Earth year (think about this and its logic should be revealed).

Note that the units for H are in velocity per specified distance. Thus, the velocity at one million light years is 21.5 km/sec; this means that a galaxy at that distance from Earth is moving away from our planet at that expansion speed. At two million light years it is 2 x 21.5 or 43 km/sec, and so on. If the Universe is indeed 13.7 billion years old, the velocity of the farthest observable galaxies, formed about 13 billion years ago would seem to be 1300 x 21.5 = 27950 km/sec. However, such a calculation is not straightforward since H has varied with time.

Not only has H *really* changed to other values in the past, its value as determined by astronomers has changed in the last century. In the last 20 years, with the advent of space observatories, the range of values have converged on the presently accepted number. This graph shows this history:



One problem troubling Hubble in the early years after his discovery is that when he used the first value for H he derived to calculate the age of the Universe, it came out around 2+ billion years, a number in stark conflict with the then accepted age of the Earth at about 4 billion years. The contradiction resulted from very imperfect - and too small - estimates of distance to the nearby galaxies he used. As more trustworthy values were obtained, and elliptical galaxies further out were better fixed as to distance, an improved curve resulted (but still applicable to redshift z values [see below] of less than 1). The diagram below is a recent plot of galaxy velocity (in km/sec; converted to kph by multiplying by 3600) versus distance (in megaparsecs) of each galaxy from Earth; the green dots denote specific galaxies for which "reasonably good" measurements have been made (other galaxies have also been so measured but their values are not on this diagram).



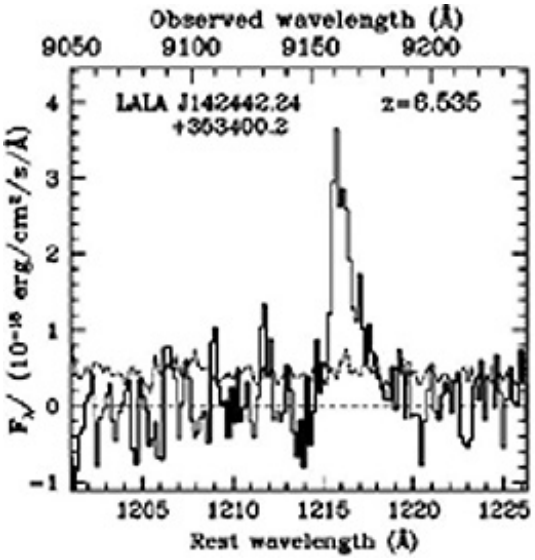
Most of these values come from galaxies 5 billion or less light years away. H_0 is the present-day Hubble Constant whose precise value is still a major goal in cosmological research; its spread of estimates is related to the uncertainties both in determining the redshift and in fixing the distance of a galaxy at the time light now received left it.

The Hubble Law works best (gives a straight line) from plots of V versus D involving galaxies a few billion or less light years away; uncertainties as to the correctness of distances further out cause an increasing scatter of points in the plot that suggest (or mask) some degree of non-linearity related to the cumulative effects of the curvature of space.

Although called a "constant", H has in fact varied in value over time. In this, it behaves much like the three non-linear plots of R (Scale Factor) versus time shown on the previous page. R describes how distances (as a measured parameter) change over time; H relates distance traveled in a unit time span (usually either Megaparsecs or million years) at each distance moving outward from the point of observation. The two are related. H refers to the relative rate of change of R. The reason that H has different values going back through the past is that it is unlikely that the expansion rate of the Universe has itself been constant since the Big Bang. One model of expansion was strongly influenced by deceleration due to gravitational forces pulling back on the enlarging universe, which means the rate of expansion has been continually decreasing, giving rise to a systematically changing H over the past (its value would increase as we move back in time towards the outer Universe). But now, new evidence for a gradual acceleration about midtime in the Universe's history (see next page) would also affect the variability of H. At best, we can now only determine with reasonable accuracy the value of H_0, which proxies for the current value that takes into account the variations in earlier eons of the Universe. We can also say that H was at its maximum value relative to the present right after the extremely large (anomalous) expansion rate of Inflation; we cannot measure this value since we are unable to determine any redshifts until the Universe became transparent.

Hubble was led to the Velocity-Distance Law by interpreting the observed redshifts of stars and galaxies. To appreciate this, we need to explore in more detail what redshifts are.

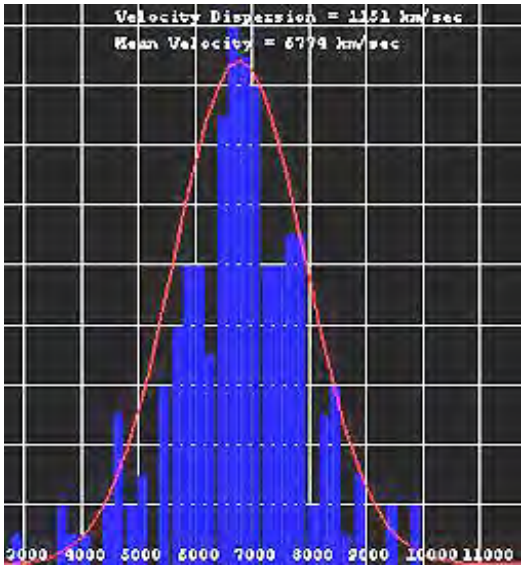
The concept of "redshift" has been mentioned several times on earlier pages in this Section. Redshift measurements have become the best indicators of cosmic distances - how far away are the stars and galaxies observed through telescopes (see next subsection for a discussion of how these distances are determined). These distances are inputs in calculating the rate of cosmic expansion. There is a systematic relationship between relative velocities of sources of radiation (such as galaxies) as these recede ever faster at increasing distances from Earth and the measured wavelengths of spectral lines from these sources. This is manifested as the ***redshift*** - so-called because light from a given wavelength such as in the UV and blue experiences a systematic decrease in frequency (ever longer wavelengths) towards (and beyond) the visible red as expansion velocities increase (this concept is discussed below on this page). As an example, the Lyman Alpha (Hydrogen) spectral line (see page 20-7) normally measured at 1216 Angstroms (in the far UV) when "at rest" on Earth is shifted to 9160 Angstroms (in the near IR) for a galaxy whose redshift is 6.54.



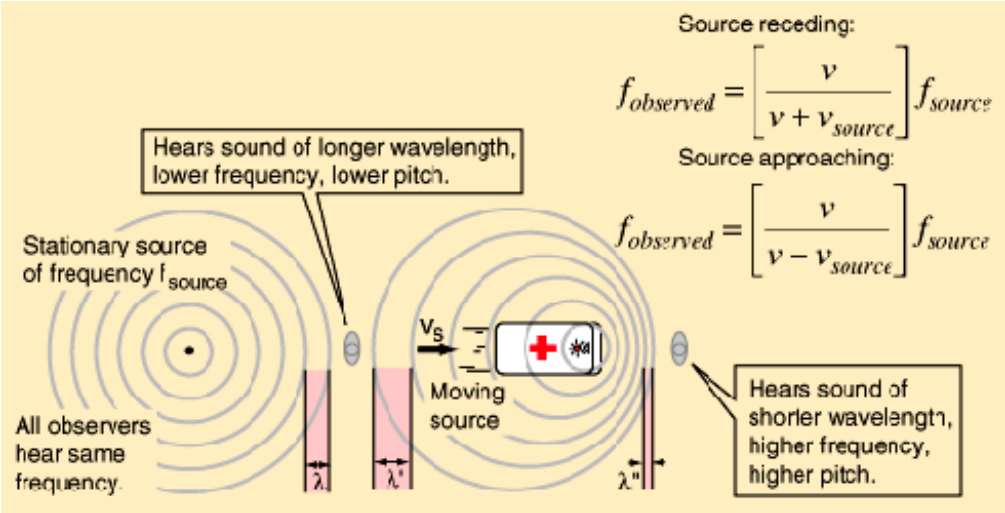
The redshift phenomenon was formalized by V.M. Slipher in 1912 but, in fact, H. Robertson noticed a bit earlier that the farther nearby galaxies were from our telescopes, the greater was the redshift. However, Edwin Hubble in 1924 has received credit for promulgating this redshift-velocity-distance relationship because he included many more galaxies as data points. He thus is recognized as the key individual behind the Expanding Universe model, from whence later came the Big Bang conception of its origin. (Note: Hubble himself never completely accepted the implications of his observations and had doubts about the Big Bang and most of the Universe models described below; for many years after drawing attention to this phenomenon he continued to prefer a Steady State rather than an Expanding Universe, although his position on the latter "mellowed" near the end of his life.) The Hubble Law can be formulated in terms of the redshift value z (where z is the recessional velocity divided by the speed of light; see below) for any celestial object (almost always a galaxy), thusly:

$$d = \frac{v}{H_0} \approx \frac{cz}{H_0}$$

Some of Hubble's observed redshifts led to estimates of galaxy velocities centered around 24 million kph, about 2% the speed of light. Galaxy velocities vary, as indicated in this histogram. Some galaxies can go as fast at 0.1 light speed; a few even faster.



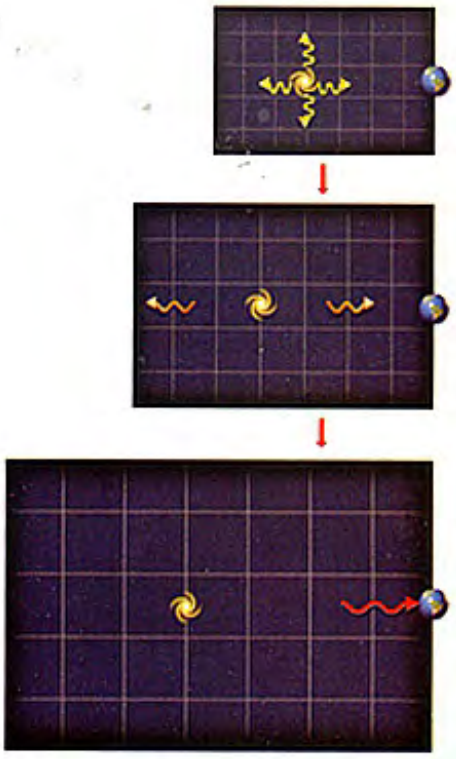
Let us now look into the details of the concept of "redshift". Redshift is analogous to, but not the same as, the familiar Doppler shift as applied to sound. Increases in recessional velocities are associated with changes in the wavelength of light being received, such that as the velocity becomes greater the wavelength becomes longer, i.e, moves to higher values (say, from 0.4 to 0.6 μm in the visible; wavelengths in other regions of the EM spectrum also are shifted towards greater values). This change is very much like the Doppler effect studied in Introductory Physics: this shows the influence of motion towards or away from the observer of a signal of some given wavelength, resulting in a systematic wavelength shift. One manifestation of a wavelength shift's effect, which can be experienced in everyday life on Earth, is exemplified by an audible phenomenon - recall the sound of a whistle or horn on a fast-moving train as it approaches and then moves past where you are stopped at a crossing. Or, perhaps more familiar is the change in pitch of a steady ambulance siren as it approaches you and then falls systematically as the ambulance recedes after passing (lower frequencies). This wavelength shortening (higher pitch) on approach and lengthening (lower pitch) with recession is called the *Doppler effect*, which results from velocity and/or position changes (relative motions) between moving source and stationary receiver.



In a sense, the lengthening of wavelength as light sources (mostly galaxies) recede from Earth at progressively increasing velocities and distances is seemingly analogous to the above Doppler effect. But, strictly speaking, this familiar effect as observed by us on Earth is not the same as applies to cosmic distances (although it is a good approximation for nearby galaxies in relative motion away from our observing location).

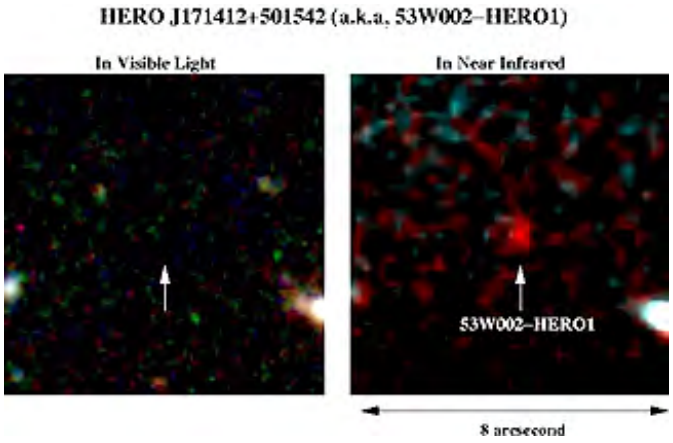
As applied to more distant objects seemingly moving away from us during Universe expansion, the wavelength shift actually results from a different mechanism known as the *Cosmological Redshift*. From a relativistic standpoint, while Dopplerlike in its consequences, the cosmological redshift is analogous to the "stretching" of light caused by the progressive increases in distance resulting from the *continuous expansion of (curving) space*. This in turn results in proportional increases in *recessional velocities* (thus in the formula for velocity v = d/t, it is the d that changes with respect to steady time progression) with increasing distance from Earth (recall the rubber band analogy on page 20-8).

The causative influence of expansion resulting in a stretching or elongation of wavelength is evident in this diagram:



From: Misconceptions about the Big Bang, by C.H. Lineweaver and T.M. Davis, *Scientific American*, March 2005
ILLUSTRATIONS COURTESY OF ALFRED T. KAMAJIAN

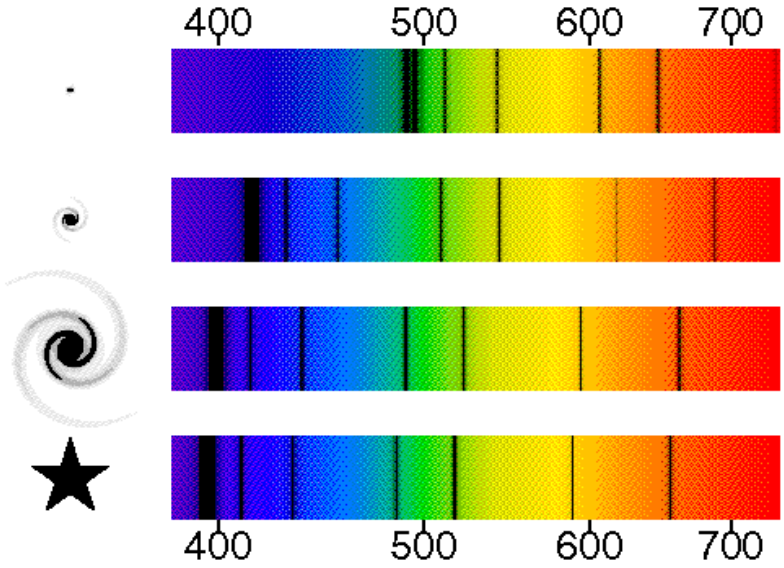
A recently reported observation of a type of galactic body called a HERO (Hyper Extremely Red Object) may be the result of this cosmological redshift. Check these two images:



On the left, the object is not detected in visible light; but it appears as a red blotch in the near Infrared. The object, at least 10 billion light years from Earth, has been found to be speeding away from us at nearly the speed of light. One interpretation considers this object to be red (from a large proportion of older stars) at the time its light left the source 10 b.y. ago . But another considers this object to be composed in large part of bright, bluish stars, perhaps even farther away (13 billion l.y.) but owing to the cosmological redshift the light as received has been stretched to near Infrared wavelengths (but assigned red in this false color rendition).

Redshift phenomena are effectively studied from the spectral states of starlight (see page 20-7). As a star or galaxy emitting radiation recedes from an observing (measuring) spectrometer (somewhere on or near the Earth), the wavelength associated with a particular line will be shifted towards the red (longer wavelength-lower energy end of the visible spectrum) and even into the near Infrared. What is measured is the displacement ($\delta\lambda/\lambda$ = the incremental wavelength shift ratioed to its initial wavelength λ) of this line to a new apparent wavelength relative to its [rest state] wavelength in a spectrum obtained by exciting the element on Earth in an emission

or absorption spectrometer. Before modern electronic measuring methods, the spectra were commonly recorded on a photographic plate showing multiple lines that result from the spectral spread of wavelengths characteristic of all detected elements) representing an element in its ground or some excited state in the visible. The diagram below shows four sets of spectra, for hydrogen, with the bottom representing a star in the Milky Way, and each successive spectrum upwards representing galaxies at increasing distances from Earth. The leftward shifts are towards longer wavelengths.



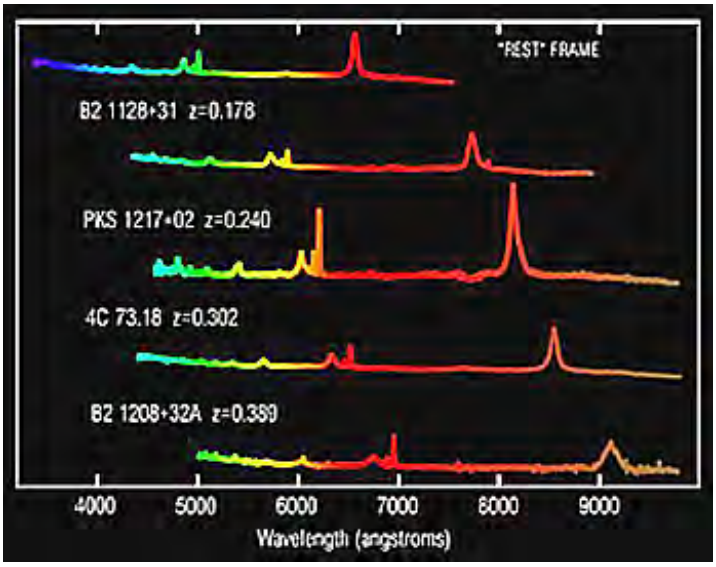
This next illustration shows telescope images and spectra from five galaxies at increasing distances from Earth.

From J. Silk, The Big Bang, 2nd Ed.

This oft-cited diagram, traceable to Hubble's work, can be misleading without some interpretation. To pick out and thus intrepret these spectra, start with the Virgo galaxy example (top right). The top and bottom lines are the same emission spectra for this spectral interval

(unspecified as to Angstrom units; the lines are white instead of black because the photographic plate is printed as a negative) obtained by spectroscopic analysis on Earth of a sample containing elements observed in stars . The two leftmost lines are the H and K spectra for the excited Ca^{++} state. The spectrum from the galaxy appears as a long lenticular white smear in the middle between the two reference spectra. The vertical arrow points to the now shifted H and K line pair, which here appear black because they are absorption rather than emission lines. In the spectral image second from the top, the horizontal arrow leads to the position of the line pair (which does not reproduce well on this page) for a galaxy in Ursa Major, now shifted notably to the right. In the three succeeding spectal images, the horizontal arrow carries to the position of the two (hard to see) dark H and K lines after each greater redshift. From these observed shifts, the recessional velocities listed under each spectral image have been calculated. These could be plotted on the distance-recessional velocity diagram above, and would fall within the general distribution shown thereon.

Today, the spectra are more commonly recorded as continuous tracings on a strip chart. The next figure shows a spectrogram recorded by a Kitt Peak National Observatory telescope in which the top spectrum (obtained at rest in the laboratory) has peaks for three Hydrogen lines at 4340 A (in the blue); 4860 A (green) and 6552 A (red). The next four are spectra from distant quasars at progressively greater distances.



Source: M. Corbin

The displacement of a spectral line owing to redshift can be used to calculate the Cosmological Redshift value **z** associated with a source simply from the *rest* wavelength of a given line and the *obs*erved wavelength of the same line displaced by the source's motion.

The Cosmological Redshift **z** is given as:

$$z = (\lambda_{rec} - \lambda_{em})/\lambda_{em} = V_r/c, = H_0 r/c$$

where λ_{em} is the wavelength of EM radiation given out in the past (then) at the emitting galaxy or star, λ_{rec} is the shifted wavelength received today (now) at the detector (on Earth), V_r is the recessional velocity for the particular redshift, c is the speed of light, r is the distance to the star whose redshift is measured, and H_O is the Hubble Constant. The above equation applies to low to moderate z's but for large z's, which are attained as the velocities near that of light speed (and are characteristics of the early moments of the Universe) a modified expression must be used:

$$z + 1 = (1 + v/c)/(1 + v^2/c^2)^{1/2}$$

Using the **z** value, the velocity **v** of receding motion of the source is given by:

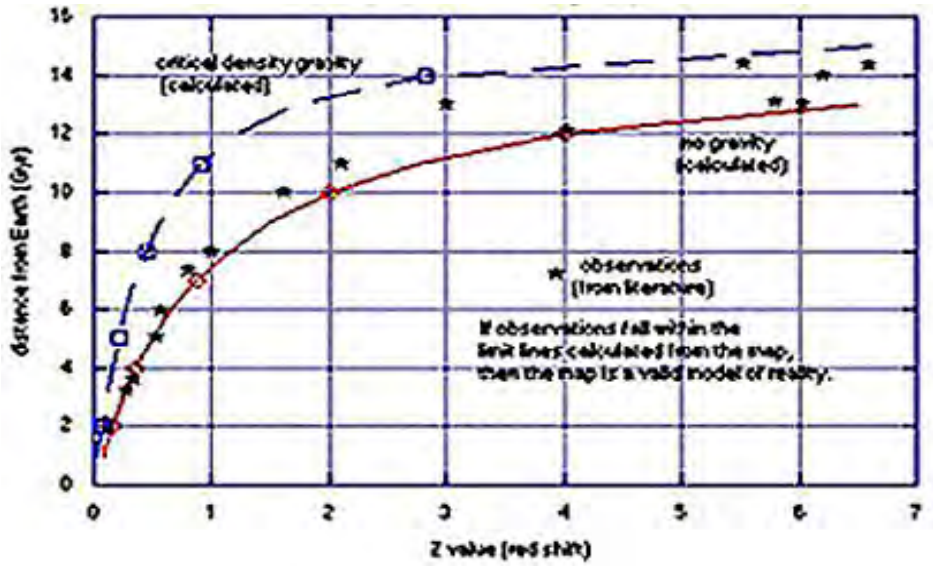
v = cz (1 + 0.5z)/(1 + z)

Since the redshift is velocity dependent, its magnitude is a direct indication of the rate of recession, i.e., the larger the shift, the greater the velocity. The redshift z is a number that represents the fraction by which spectral lines from a luminous source shift towards longer wavelengths. Values of z range from less than one for closer sources and have risen for the most distant sources (early time galaxies) to numbers around 6.

If instead the source advances towards the observer, the shift will be towards the blue (shorter wavelengths). Since it is postulated in the Big Bang model that all sources are apparently moving *away* from one another, a blueshift would seem anomalous. However, this occurs, for example, when spectra are acquired from a rotating spiral galaxy in which arms on one side (from the center) may indeed be moving away but the other side must be approaching from opposite directions. Likewise, some galaxies in a local group may appear to be moving towards Earth towards Earth, but the entire group is still receding relative to our galaxy.

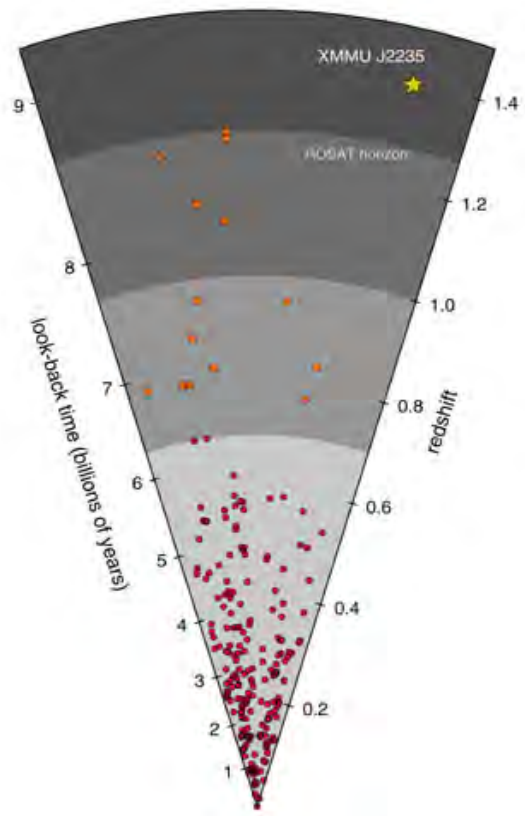
Another mechanism can cause redshifts, namely, the effects of gravity on radiation. This *gravitational redshift* is a consequence of General Relativity. When light leaves a massive gravitational source, such as a White Dwarf, gravity causes a shift towards a longer wavelength (conversely, light passing into a huge gravitational field will undergo a blueshift). The massive body thus slows down photons representing a range of energies as these escape from it, causing a loss in their energies that results in reducing their frequencies and increasing their wavelengths. This effect has been observed for light grazing supermassive bodies, including Black Holes. Overall, the effect is *localized* or confined to individual bodies, and normally the shift is very small, so that even the cumulative effects of light reach Earth from the outermost reaches of Space are quite small compared with the motion-induced Cosmological Redshifts related to expansion. Nevertheless this local redshift must be accounted for when individual receding galaxies are used in determining the cosmological-scale redshifts.

There is another, more general effect of gravity, shown in the plot below, which shows the redshift curve for a Universe with maximum gravity influence versus no gravity at all. The ordinate is distance in billions of light years. This range of possibilities is pertinent to the accelerating Universe model discussed on the next page.

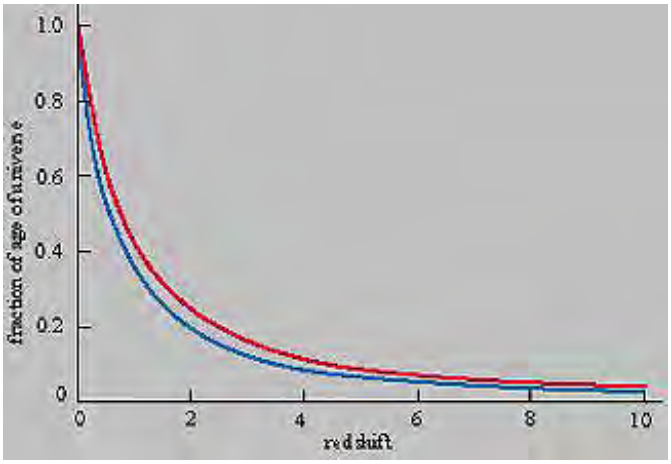


Most redshifts measured so far include the lower values of z obtained by examining a range of "normal" galaxies at distances from Earth under about 7 billion light years.

As stated above, most redshifts are observed to fall within the range of a fraction of 1 to about 3. These are all associated with galaxies that are less than 10 billion years (in lookback time) from Earth. This is evident in this plot:



As galaxies lying beyond 10 billion light years are observed, their redshifts begin to rise at more rapid rates. Thus:



The ordinate denotes relative age: The present time is given by "1", with nearby galaxies that appear most fully evolved (to us in the present time) having very low redshifts. The exponental drop in the curves (the red curve applies to a Universe with 70% Dark Matter; the blue curve described a Universe without Dark Energy [Cosmological Constant = 0]) shows that the maximum rate of increase in the value of 'z' occurred when the Universe was less than a relative 0.2.

Higher redshifts have been found for galaxies that are strong radio sources and even larger values (around z = 5 to 6.5) from very distant quasars (mainly those which display their effects in the first two billion years of Universe history). Values of 'z' increase rapidly towards infinity for Universe events older than the first stars. For instance, at the time of Recombination (page 20-1) z = 1000. This is the general relationship as tied to major cosmological entities:

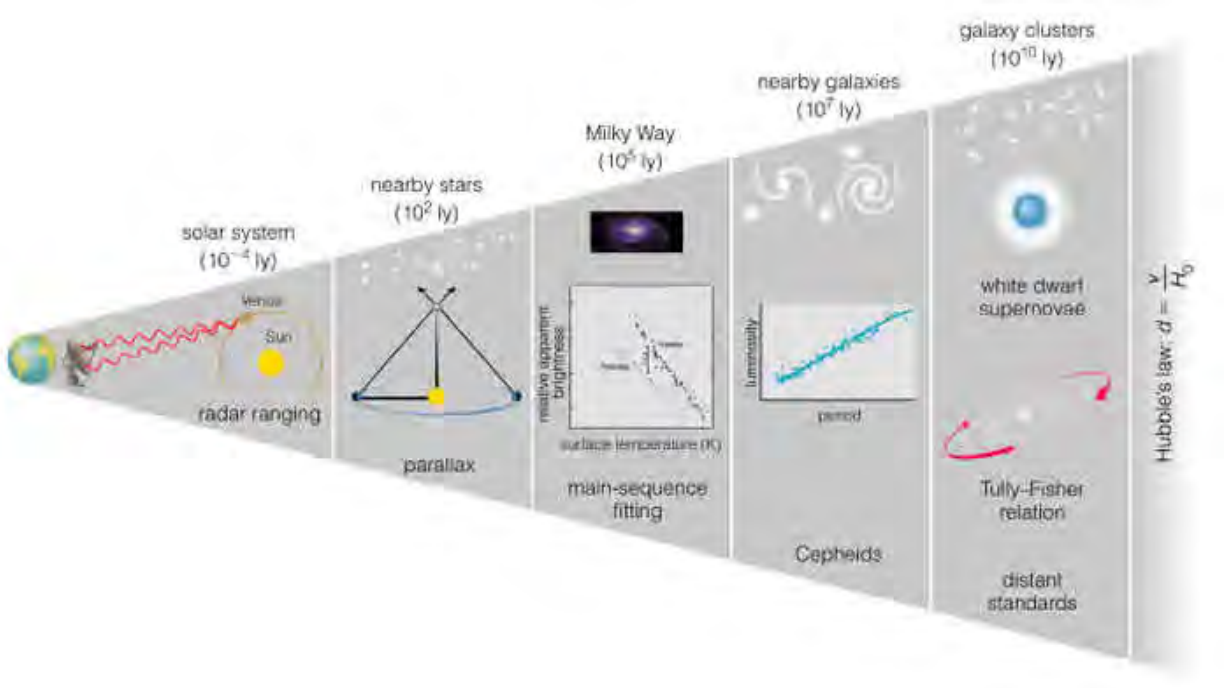
The most distant galaxies are hard to observe but their redshifts can be greater than 10. Theory permits calculations of redshifts from the Recombination Era back through the first minute of the Big Bang. These values have been taken from Joseph Silk's The Big Bang (3rd Ed, page 68): At Universe age of 2 billion years, z = 5; Decoupling at 380000 yrs, z = 10^3; Radiation Era, 1st minute after the BB: z = 10^9; Hadronic Era, at t = 10^{-10}sec, z = 10^{10}sec; Planck time, at 10^{-43}sec, z = 10^{32} ;

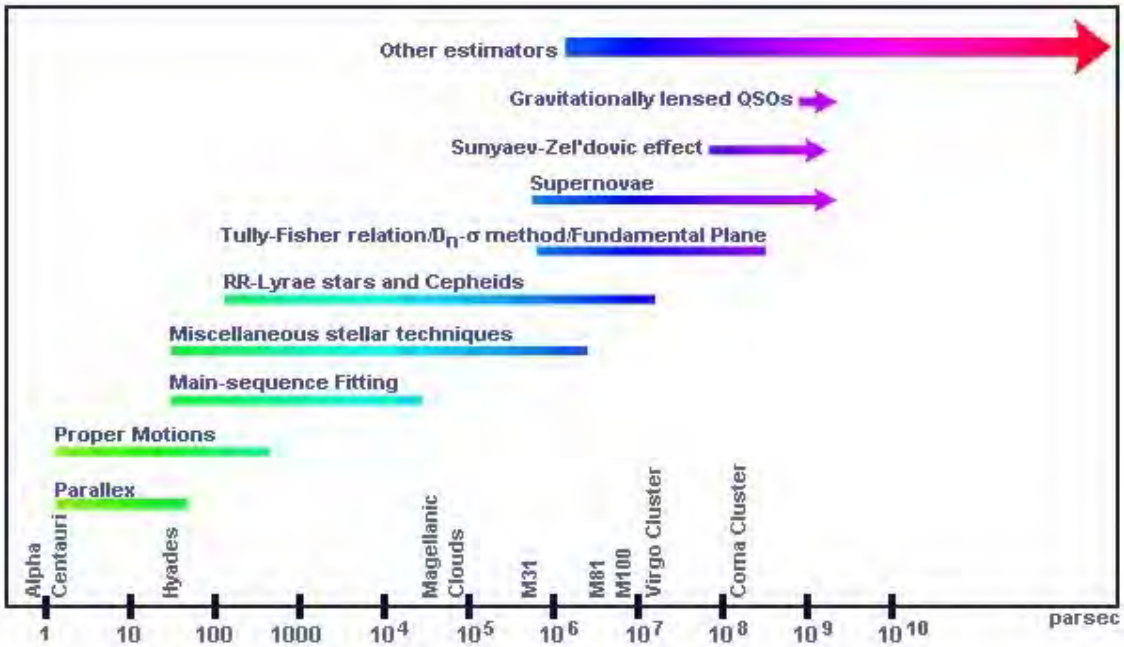
Now with this overview of redshifts, the reader is encouraged to work through a review on this Wikipedia website.

Stellar and Galactic Distances (from Earth)

To calibrate and then apply the redshift to estimate R (Scale Factor; previous page), and to calculate the Hubble contant H, the distances to the stellar bodies each with a specific redshift must be determined. Over the last century various methods of estimating distances have been developed. A good, in-depth review of the principal methods used in distance determination is found at Ned Wright's Cosmology site.

Some of these methods will be discussed in this subsection. Use of multiple methods applicable at different distances is called the Cosmic Distance Ladder. We will start with an overview given in these two illustrations and a summation:





This is a synopsis of most of the distance measuring methods:

1) .For the Solar System: Radar is the most accurate measuring method.

2) For nearby stars: The parallax method, using either Earth's orbit to occupy different positions or an orbiting satellite, to determine how far away are stars in the Milky Way Galaxy; works well out to about 3000 light years.

3) For some stars farther out in the Milky Way: The standard candle method uses stars whose true luminosity L can be determined.

4) For stars out to the edge of the Milky Way: The Cepheid Variable star method is employed.

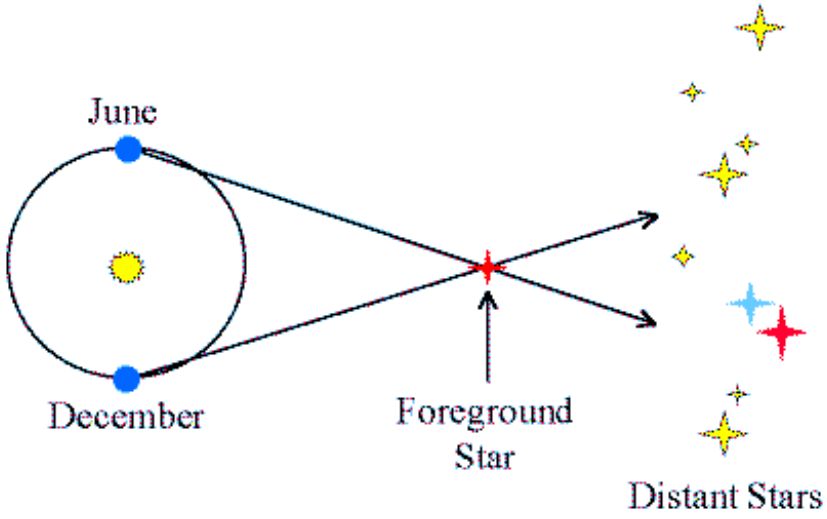
5) For many more distant galaxies out to several hundred million light years: The Tully-Fisher method utilizes measured brightness and galaxy rotation (using the Doppler effect).

6) For galaxies out to about 10 billion light years: This depends on observations of White Dwarf supernovae, whose intrinsic brightness is well known.

Some of these methods have accuracies of +/- 10% or better.

Now, to more detail: Distance measurements obtained for nearby bodies, e.g., in our own Milky Way galaxy, can be made on visible stars whose magnitudes can be directly ascertained. One technique is that of parallax observations. While not fully explained here, the gist of this technique can be sensed by this simple experiment: Hold your index finger first about 6 inches in front of your nose and rapidly alternately close your left eye and then right one repeatedly. Your finger will appear to shift back and forth relative to a fixed background, perhaps seeming to displace several inches. Now, put your finger full out (about 24 inches) and do the same thing. Note that the displacement is now less. This is the parallax effect (first discussed in Section 11). The amount of shift decreases with increasing distance and that distance can be determined by simple trigonometry.

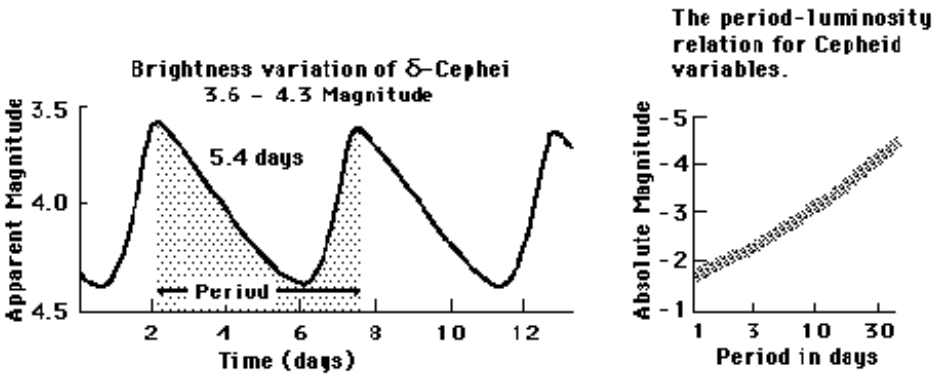
As used to measure stars within about 100 parsecs (326 light years), the left and right eye positions are proxied by the positions at opposite points in the Earth's elliptical orbit six months apart. A star's apparent shift relative to distant background stars, even though proportionately much smaller than that of the finger experiment, is sufficient to provide an accurate distance measure for stellar bodies close to Earth.

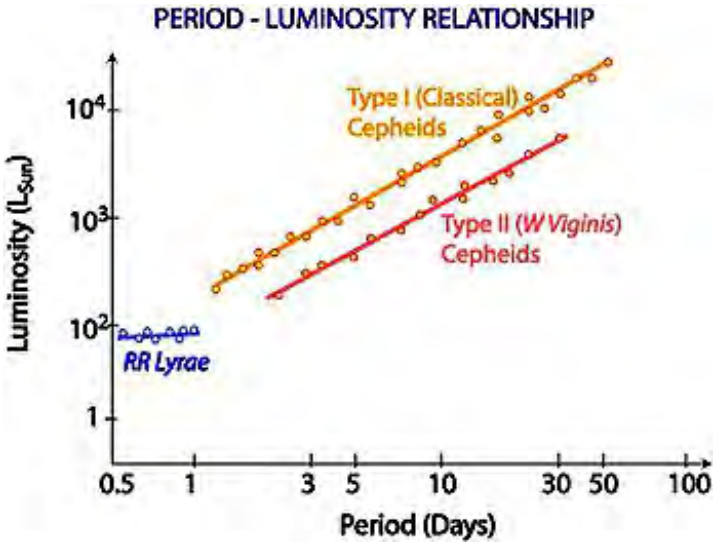


The standard candle method requires use of objects (stars; galaxies) whose intrinsic luminosity L is known. At some distance d, the L (in watts) is spread over a sphere of area $A = 4\pi d^2$. What is actually measured is brightness b; b = L/A. The distance d is then calculated as d = (L/4πb.

Approximate distances to closer host galaxies containing separable stars rely on determining the intrinsic luminosity of certain types of individual stars. One class is the so-called pulsating stars, i.e., those whose luminosities vary systematically over periods of days to several months. These include stars that have used up nearly all of their Hydrogen fuel and are enroute off the Main Sequence towards then becoming Red Supergiants. During this phase of their history, their atmospheres expand rapidly with a rise in luminosity, only to revert back to their previous state during a cycle whose time is that of a regular period. What happens is this: the star in its more compact state has a specific internal pressure; at some point the nuclear processes cause the star to expand, increasing its diameter by a factor around 2. The pressure gradient decreases until a condition is reached in which gravity now reverses the process causing contraction. The expansion-contraction repeats at its characteristic, nearly constant time period (in Earth days) for a long time before a particular pulsating star evolves into a more stable Red (Super)Giant. Most stars showing this phenomenon have initial masses from 5 to 20 times that of the Sun. More massive stars have longer periods of expansion-contraction and are also more luminous to start with.

One class of periodically pulsing star groups are the RR-Lyrae stars whose periods are in hours to a single day. More important are the *Cepheid* Supergiant stars. Cepheids were first discovered by astronomer Henrietta Leavitt in 1912 in the nearby Magellanic Clouds; she then showed them to have regular, pulsating variations in luminosity proportional to their pulse periods (in so doing, determined that the brighter the star, the longer its period P). Cepheids flare up to peak brightnesses, then dim down, over periods of days to weeks. Using the parallax method, the distances to some of these were independently fixed and their absolute magnitudes M were calculated. Since these distances varied (within the Milky Way and in the Magellanic Clouds), the various M values could be associated with their corresponding periods in the cycle, thus establishing the M-P relationship. Of course, Cepheids having the same values of M but located at widely varying distances from Earth will experience an apparent decrease in brightness m depending on distance (and subject to the $1/d^2$ relation that defines the falloff in brightness with distance). These ideas are illustrated for one type of Cepheids (δ-Cephei) and for the more general case:.





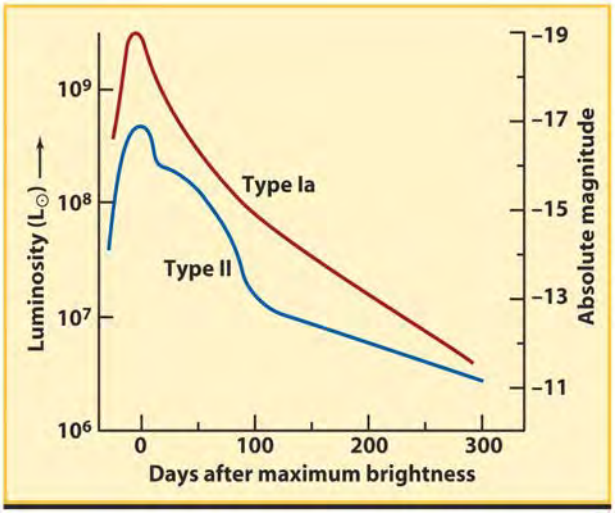
Once absolute luminosity for a given Cepheid is calibrated from this relation, the drop in apparent (observed) brightness m from that value owing to its specific distance d can then be included in the following equation to determine that distance to this star:

m - M = 5(log d/10)

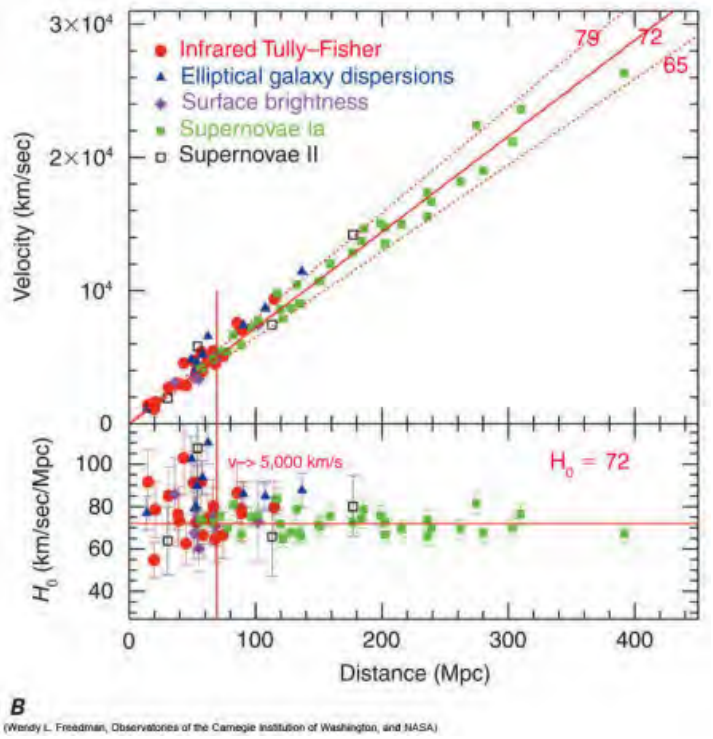
In the 1920s, Edwin Hubble firmly established the relation that the longer the period, the greater is the increase in the intrinsic (absolute) brightness in a Cepheid. He applied this pulse cycle approach to these stars in different galaxies and over a range of distances. It was Hubble's use of primarily Cepheid-derived distances that led to his first major hypothesis of an expanding Universe, after also introducing the redshift relation. Some of the values he used were not highly accurate (but were later corrected) so that his initial postulated rates of expansion were considerably off-the-mark.

The Cepheid variable star method works well out to a distance of 50 million light years (roughly, out to Virgo). For galaxies farther away, other methods of measuring distances to them (such as the rich cluster-brightest galaxy indicator which gives usable approximations out to 10 billion l.y.) have been worked out and applied (these have varying degrees of accuracy.

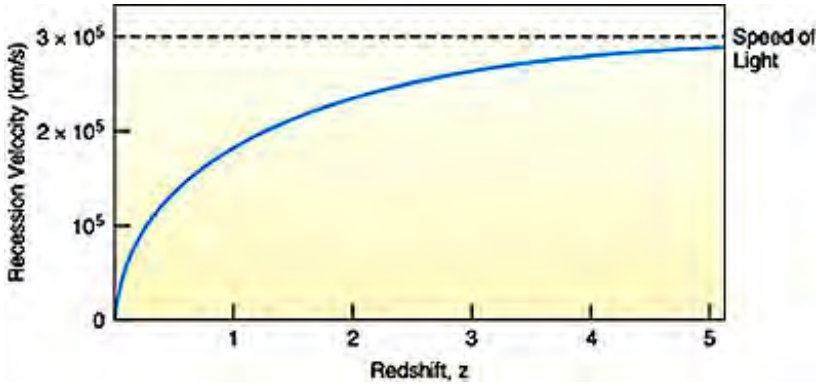
The best among these is the Type Ia supernova explosion (discussed in more detail on the next page). This occurs when a white dwarf star has a larger binary companion. Material stripped from the larger star accumulates around the dwarf until a critical mass (~1.4 solar masses) is attained, at which point it explodes releasing luminant energy of a narrow range of L. (Details about White Dwarfs are presented on this Wikipedia website). Thus it qualifies as a standard candle star. This explosion is a Type Ia supernova that is so bright that it is readily detected in very distant galaxies. This plot show the relation of luminosity to time history of Type 1a and Type II supernovae:



This next diagram was shown first on page 20-1. It is repeated as a summary of some of the Cosmic Ladder methods:



As earlier stated, cosmic distances are used to relate redshift values to recessional velocities. When redshifts begin to exceed about 1, the speeds of the objects concerned begin to approach relativistic values, i.e., they are ever larger fractions of the speed of light. Thus, although the actual speeds continue to increase, the incremental rate of velocity increase itself decreases (slope asymptote approaches 0). This gives rise to a redshift vs recessional speed curve that is like this:



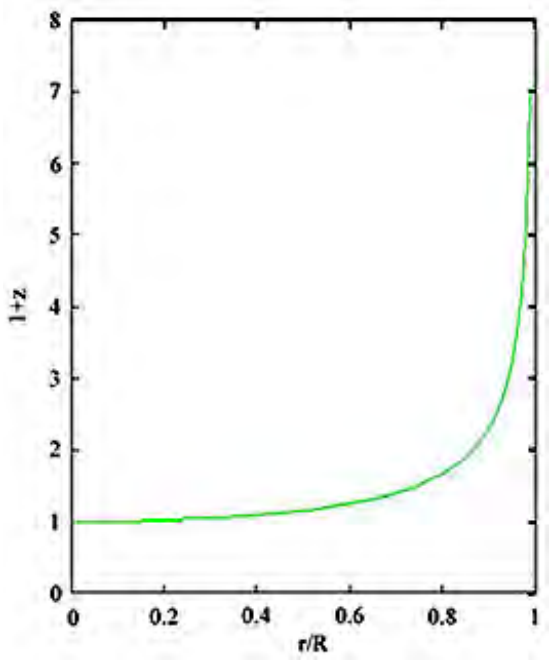
From Astronomica.com

Another relationship: $z = 1/R(t_{em}) - 1$ describes the redshift in terms of the Scale Factor R pertinent to t_{em} which refers to the particular time when the light was emitted . This relationship can also be cast in the following way:

$$\mathbf{D_{now}/D_{then} = R_{now}/R_{then} = z + 1 = \lambda_{rec}/\lambda_{em}},$$

in which D_{now} is the distance to the emitter when the light is received and D_{then} refers to the distance in the past when light left the emitter.

A plot of z + 1 versus r/R (the distance out to any galaxy ratioed to the distance out to a Universe's edge, set at the Scale Factor R) shows the exponential character of this curve (generalized here):



We see a redshift (towards longer wavelengths) because the Universe had a different Scale Factor when the light left the emitter. The redshift is due to the relative expansion of space (increasing "D's" [for distance]) rather than actual speeding up of more distant galaxies. Look at the two circle drawing shown earlier on page 20-8. Note the S-like curl that represents part of a wavelength train. It has a shorter wavelength in the left circle; as the circle expands with its enlarged coordinates, note that the wavelength on the right is now longer.

Before new data from the HST and other observing systems were acquired, the prior estimates of the value of H_0) had fallen between 50 to 100 km/sec/Megaparsecs (a parsec is 3.26 l.y). (In some expressions of H, megaparsecs are replaced by 1 million (10^6) light years; thus 75 km/sec/Mpc = 23 km/sec/10^6 l.y.) One goal of the Hubble Telescope is to better zero in on the most accurate value of H_0 - essential to an accurate estimate of the Universe's age. From most recent best estimates, a range of H_0 (value at the present time) between 65 and 79 km/sec/Mpc is considered the most likely to contain the eventual most accurate value (still being sought).

A good discussion of some of the above information on the Hubble Law is found at these two Wikipedia-1 and Wikipedia-2 sites.

Cosmic Ages: The Age of the Universe

The Hubble constant H_0 provides a direct way of estimating the time from the Big Bang (Universe's "birth") to the present. Currently, 13.69 (rounded to 13.7) billion years (using

H_0 = 71.4 km/sec/Mpc, based on WMAP data [see below on this page]), with an uncertainty of +/- 10%, is the most widely accepted value.

The general relation for the Universe's age (since the Big Bang) is given by the expression:

$$t_0 = 1/H_0$$

.

This formula is deceptively simple. Just putting in a value for H_0 yields a number that is not years as such. The proper units must be included. Here we will run through the calculation that leads to the end-result age for a value of H_0 = 71 km/s/Mpc (s = sec; Mpc must be converted into mega-light years). In the actual calculations, when H units (in the Mpc mode) are adjusted to give an answer in billions of years, the formula becomes:

$$\text{Age } t_0 \text{ (years)} = 979.3 \times 10^9/H_0$$

Inserting 71 km/s/Mpc into the formula yields an age of 13.7 billion years.

Most textbooks on Cosmology and relevant Internet sites do not show the details of the calculations in regard to units involved. This is done here:

Age of Universe (10^9 years) = $1/H_0$ (**M**pc x **s**ec/**k**m) x (3.09×10^{19} **k**m/1 **M**pc) x (1 yr/3.155×10^7 **s**ec)

The bold letters refer to units that cancel out.

The lower the value of H_0, the larger is t_0 and thus the Universe becomes older; smaller H_0 yield younger ages.

The first reported (before 1995) HST-derived ages fell between 8-12 Ga, anomalously low compared with pre-HST reported ranges of 12 - 18 Ga. This was especially confusing in that separate evidence and theoretical calculations indicate some distant galaxies might well be 14 Ga and possibly older. This Age Paradox - stars seemingly older than the Big Bang's start time - proved particularly troubling to cosmological theorists for several years. The problem was minimized with further studies of nearby globular clusters which contain very old stars. These clusters formed along with the organization of the oldest galaxies around which the clusters are tied by gravity within the galactic halos. Data from the Hipparcos astronomical satellite led to a redetermination of globular cluster luminosities, and correlative rates of fuel consumption. From this new information the average ages of clusters was reduced by 14% so that their oldest stars (Red Giants) could not be older than the 13 Ga cited above. This, together with the more refined 13-14 billion year Hubble age (see below), obviates the discrepancy posed by the Paradox. One consequence of this most recent age estimate is that the farthest galaxies whose distances from Earth is said to be 13.4 billion l.y. must lie near the observable edge of the Universe. One galaxy has now been found at 13.23 b.l.y, and in time more will be detected that are even farther away (older).

Over the past 7 years, observational data analyzed by HST Teams whose prime task is to try to pin down the Universe's age using a better determined Hubble constant suggested in May of 1999 a best estimate for the Hubble constant of 70 km/sec/Mpc. (That number also coincides with the local expansion rate based on redshift-distance measurements for galaxies near the Milky Way.) For the H_0 range they arrived at, an age of *13.97 billion years* would result - a value reasonably close to the more recent WMAP results. The age uncertainty represents an accuracy variation to within +/- 10% for the value of this constant. Their value depends on analysis of redshifts in 18 galaxies within 67 million l.y. from Earth; in these they have found up to 800 Cepheid variable stars which are considered reliable indicators of large distances. From the

combined determinations for the 18 galaxies, this best estimate of expansion rate gives an increase in velocity of 256,000 km/hr (160,000 mph) for every 3.3 million l.y. farther out the stellar entity (galaxy or individual stars) is from Earth.

A group of astronomers associated with Marshall Space Flight Center colleagues have used 38 galaxy clusters spread over distances from 1.4 to 9.3 million light years observed by the Chandra (X-ray) Observatory space telescope combined with radio telescope data to obtain intrinsic luminosity data from which distances can be calculated. Using these, they derived a value for H_0 of 77 (+/- 15%) which translates into an age of 12.7 billion years for the Universe.

Many astronomers disputed the above age specifications based on the galaxy distance model, citing older ages according to their calculations and their interpretation of H values using different inputs. In the late 1990s most cosmologists (e.g., Alan Sandage and associates) had accepted a small range of values of H_0 that yield ages centered on 14 Ga; those ages are now close to the preferred "best estimate" of 13.7 Ga (see above). However, a vital note of caution: As more galaxies at great distances from Earth are detected and measured astrometrically, so that their intrinsic brightnesses, distances, and redshifts are known with notable accuracy, the value of H_0 could be recalculated to a lower number. This would mean an older Universe (greater than 14 Ga) and would mean that the oldest galaxies now detected lie inside the limits of the knowable Cosmos. Said another way: there may be considerably more space beyond our present observable Universe, which is where our time horizon now extends, and this additional outer volume would likely contain galaxies. This can be assessed when/if we can see the outermost, already detected galaxies in such detail that we can specify how primitive or early they are in their evolution. If they appear to be in the first stages of formation, if we know enough about their rates of growth, and if galaxies indeed to form within the first billion years after the Big Bang, then these galaxies are probably near the edge of the expanding Universe, with little or no space beyond. This does not rule out an infinite Universe, if it is destined to continue expanding into an infinite future.

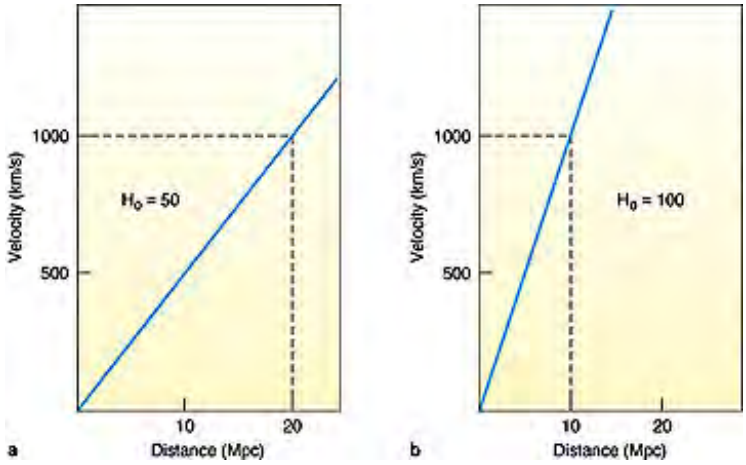
However, the Hubble Age also can be modified depending on whether the Universe is open, closed, or flat, and may be influenced by the type of space involved (see below). In the absence of gravity the value of t_H is $1/H_0$. The Hubble age for a Universe with flat expansion varies as the relation $t_H = 0.67/H_0$ (this applies to the Einstein-DeSitter Universe [see below]). For an open Universe, t_H falls between 1 and 0.67 but 1 is usually chosen; an open Universe seems the best model at this time. For a closed Universe, t_H can be less than 0.67. These several cases for ages that are less than 1/H seemingly point to Universes that began less than ~14 billion years ago. But, if the ages of the most distant galaxies, now only estimated from distance-brightness relations, prove to be around that value, then the resulting paradox - parts are older than the whole - will need to be explained away. To some extent, resolving this paradox can help to specify the type of Universe that actually exists, since age-incompatible situations would seem to argue against the types that don't fit.

Just when it seemed that astronomers have finally firmly fixed the value of the Hubble constant so that the Universe's age and size can be considered as accurate, a new set of data - if confirmed by additional observations - have cast doubt on the best number for H_0 and hence for estimates of Universe age and size. A group of astronomers led by Dr. Kris Stanek of Ohio State Universe have spent years developing a new technique for determining distances to galaxies and stellar objects that would avoid several steps in the process - hence reducing errors. Their method involves determining the mass of mutually orbiting binary stars, which allows calculation of intrinisic brightness that is compared with apparent brightness, from which distance is then established. They used two large bright stars in M33, the Triangulum galaxy. Their method came up with a distance from Earth of 3.0×10^6 l.y. Previously that galaxy had been rated as 2.6 million light years away.

If, after repeated testing and detailed assessment by fellow astronomers, their method holds up

and becomes accepted as the best distance estimater, this could imply that the 15% greater distance the OSU team found for M33 is valid for other galaxies at various distances. The value for H_0 would need to be adjusted downward by 15% and thus cosmic age and size of the Universe would need to be increased by 15%. At the moment, the "jury is still out" on this new result.

Nevertheless, from the above, variations in the chosen value for H_0 have a major, definitive influence on two fundamental cosmological parameters that scientists seek to know "exactly" - the size of the observable Universe and the age of the Universe. This notion is brought home by considering the consequences of changing the H_0 value, as is done in this figure:



The question to ask in interpreting these H curves is which one leads to a younger Universe; which Universe is smaller? Check the conclusion by clicking on this asterisk *.

The essential factors determining the Universe's age are its overall density (mass and energy) and the value of the Deceleration Parameter (related to the Hubble Scale Factor), as discussed elsewhere on this page. These specify the rate of expansion which in turn reveals how long it takes for galaxies to get to the farthest reaches of *observable* space Observable space is defined as the limits or horizon defined as the farthest bodies that have emitted radiation which has had time since the beginning of the Universe to travel to Earth's observing stations. This will be marked by the first vestiges of materials capable of emitting detectable radiation during the early moments of the Big Bang. So far, detectors covering optical and other spectral regions have not yet picked out these oldest sources, so the currently observable Universe presently is smaller than the total observable Universe.

The Hubble equation specifies that the fastest receding objects must be farthest away; conversely, those near the Milky Way are the slowest moving. Thus, in an expanding Universe, with all galaxies ultimately drawing apart from each other, those progressively farther away must travel at proportionately greater speeds, but at the same rates in all directions, to preserve an overall uniformity of spatial relations during these expansive movements. As a general rule, the greater the *lookback* time, the smaller was the size of the Universe at such times, and the hotter and denser is the early expansion status of matter and energy. (Lookback time connotes the idea that the farther out in space one looks, the further back in time [earlier] is the event or stage of development associated with objects [e.g., galaxies] when light left them; a large Lookback time means a younger age]).

Because most galactic measurements made on distant galaxies show red rather than blue shifts (the latter are seen for mostly nearby galaxies moving towards us [Andromeda is approaching Earth at ~360,000 kph] or can be noted in individual spiral galaxies as one arm moves towards Earth), this evidence for overall (net) recession is the principal proof for the Big Bang expansion model. The redshift is related to *recessional velocities* (ratioed with respect to the speed of light) by an exponential curve in which the velocities rise rapidly towards infinity as that speed in approached. Most measurements of z from less distant galaxies afford numbers between 0 and

1 (for example, z = 0.1 represents a distance of about 1 billion light years). Farther out galaxies showing redshifts of 1.2 correspond to ages in light years of about 8 billion years; HST has now observed many galaxies with z's up to 2+. Distant quasars, some about 10-11 billion l.y away, have shifts of 3 - 4 or higher (at an observed age much earlier in Big Bang time). Several galaxies have measured z values of 5-6 and one now has a value of z = ~10 (reaching to about 90% of the speed of light); these are thus formed during the first billion years of the Universe.

Here is an image obtained during the Sloan Digital Sky Survey (SDSS) showing a galaxy with a redshift of 5.82 that is unusually bright (a quasar is inferred as the cause).

Recessional velocities as a function of distance of cluster galaxies from Earth as the observational frame of reference can be calculated from the Hubble equation and z values. Choosing a Hubble constant that gives 14 Ga as the age of the Universe, a galaxy recedes an additional 25 km/sec for each million l.y. further out one looks through space. For a cluster in the Virgo Constellation, at a distance of 78 million light years, the recessional velocity is ~ 1200 km/sec. For the Bootes cluster, at 2.5 billion l.y., the velocity has increased to 22000 km/sec. Galaxies whose distance is about 5 billion l.y., attain velocities approximately one-third the speed of light (100000 km/sec). The most distant observed sources (mainly quasars) reach recessional velocities approaching light speed. The same type of velocity distribution would be ascertained at any other observational point (such as set up by the distant galaxy "civilization" referred to earlier) in the Universe.

As HST observations accumulate, it is becoming evident that, with its resolving power, structure in galaxies can still be recognized out to about 4 billion light years. Present evidence is that beyond a z value of 2.75 no well-formed spiral galaxies can be confirmed to exist (but at least some are likely). Those that lie farther out seem to be ellipitical or commonly "dismorphous" (no regular form). Since these are older, this implies that spiral galaxies may not develop until later in galactic evolution. Some of the earlier-formed spirals have one or more extra arms compared with younger ones (the Milky Way has 3 major ones).

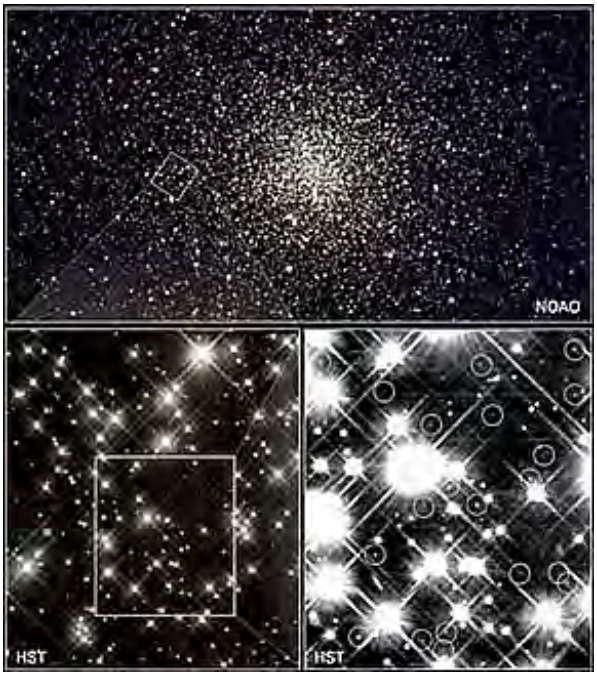
The discussion in the above paragraphs is confined to redshift measurements that can be made from observable astronomical phenomena such as galaxies and quasars. There is another aspect which is more theoretical, namely, the redshifts in the earlier history of the Big Bang prior to the onset of the Decoupling Era (before which no direct observations is possible). At the initial Planck Time of 10^{-43}, the redshift z is calculated to be 10^{32}. After one minute - the beginning of the Radiation Era, z drops to 10^{9}. In the first 1-2 billion years after the B.B., the redshift decreases from about 30 to 6. The latter is near the maximum value determined so far by direct measurements - the galaxies with that value are about 13 billion l.y. away.

This systematic increase in redshift going back in time accompanies the expansion of the Universe. The process of enlarging space leads to a lengthening of the wavelength of light - hence the progressive rise in the redshift value of z. Photons that have to travel greater distances, from further out in the expanding Universe, appear as though they have decreased

energy - hence longer wavelengths (Planck's Radiation Law), i.e., shifts from blue to red. Since redshift depends on the velocity of a receding object, it follows that the maximum velocities of galaxies are found in the outer reaches of the observable Universe. This is logical: if all matter/energy was concentrated at a singularity at the time of the Big Bang and then dispersed thereafter, those manifestations of matter such as the galaxies that are farthest from the observation point (for us, Earth) must have been traveling at the fastest speeds - in other words, if all matter started from the same point, matter now farthest away had to travel at the highest velocities.

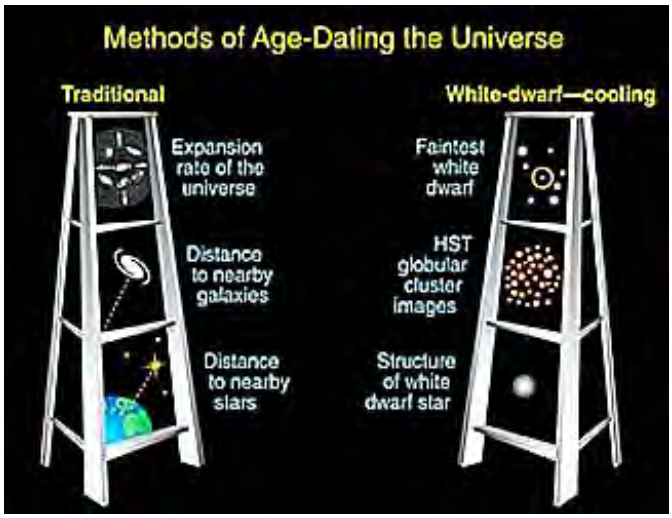
There is also another theory which can, in principle, modify the implications of the observed redshifts, namely, that the velocity of light is *not* constant but has changed over time by gradually slowing down: this is the "tired light" concept which, while intriguing, has so far not been supported by data or observational proofs, although it does seem to have a relation to the expansion aspect of stretching light to longer wavelengths (paragraph above). It has its supporters; some cosmologists and quantum physicists have postulated that the current values of certain fundamental parameters have changed with time, having different values (especially in the early moments of the Big Bang) that evolve into their present numbers as the Universe grew. Even though evidence for this is presently lacking, this is not trivial or frivolous speculation but falls into the time-honored scientific methodology of proposing seemingly outlandish theorems or propositions capable of explaining some phenomena and then conducting experiments to confirm or deny the idea.

The age of the Universe is a fundamental value which cosmologists seek with great care and effort to establish accurately. What will help in settling on a "best value" would be an independent measurement using a technique other than the recessional velocity extrapolation. In April of 2002, a seemingly reliable second method has been reported. It is based on knowledge of the time involved in White Dwarf stars burning out their remaining fuel to reach a "glowing ember" state. Theory sets a fairly precise time span for this to occur. In the earliest stages of galaxy formation, Globular clusters will contain rapidly produced white dwarfs as large stars burn their Hydrogen over a brief time and then enter the Dwarf stage. The "embers" that are very old are hard to detect by telescopes. But, the Hubble ST has been used on a globular cluster near the Milky Way to search for these embers; by taking a long exposure image (cumulative 8 days, spread over 67 days) these faint White Dwarfs were detected, as shown in this set of images of stars within cluster M4:



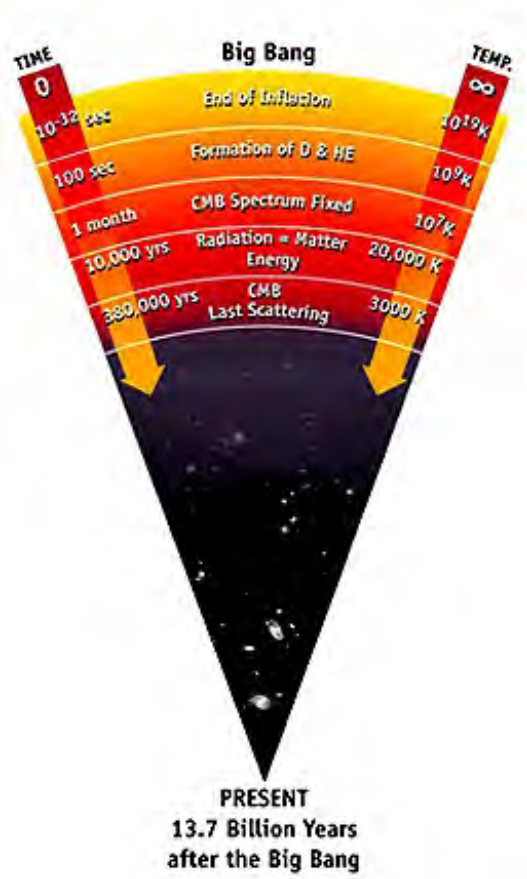
As reported by Dr.Harvey Richer and his colleagues, calculations place the age of these White Dwarf "cinders" at between 13 and 14 billion years. By adding ~1 b.y. (typical time for the first Globular clusters to develop) to these values, this independent age assessment falls right within

the same range now generally accepted from recession measurements. The two methods of determining Universe age, using a "ladder" approach to arrive at the final values, are shown schematically in this diagram:



Unless fatal flaws are discovered in either or both methods, it seems for now that an upper limit of 14 billion years will stand as the actual age of our Universe.

We can summarize the age information that includes key events in the Universe's early history by posting this diagram which includes some of these events (most of those were described on page 20-1):



Note that we can only look directly back at stars and galaxies, such as those that formed in the early Universe, to a time approximately 380,000 years after the Big Bang.

Cosmic Background Radiation

Another solid proof for the Big Bang was the discovery that **Cosmic Background Radiation** (CBR; also referred to as Cosmic Microwave Background [CMB] radiation in the diagram just

above) peaks near the wavelength of 1 mm (1000 µm [micrometers]) which lies at the far IR/microwave boundary region of the EM spectrum. This is the peak wavelength expected from a radiant blackbody source whose temperature is now 2.73° K. George Gamow and his colleagues Ralph Alpher and R. Hermann had first predicted such radiation (their estimate of its peak was at 5° K) in 1948. The term "afterglow" can be applied to this radiation. The radiation consists of photons (whose electromagnetic wavelengths vary with the blackbody temperature distribution at any given time after the BB) created during the first BB minute as matter and anti-matter annihilated almost completely. We first discussed CBR on page 20-1 and suggested the reader consult this Wikipedia website for more information.

The CBR now evident as pervasive throughout space (both intragalactic and intergalactic) can be traced to an equilibrium state between nucleons, electrons and photons that was arrived at when the Universe had cooled to about 10 million °K approximately 6 months after the Big Bang. Evidence of what it was doing during the Radiation Era, up to Decoupling, had been lacking because of the opacity brought about by scattering and internal entrapment of photons (see page 20-1) within the early Universe during the next 350,000 years. At that time, as the temperature dropped to about 3000 °K, almost all electrons (the principal scatterers) and protons were able to combine as Hydrogen atoms that no longer scattered the photons so that light and other radiation emerged from the radiation "fog" which was fully lifted by 1,000,000 years after the BB. With the resultant transparent Universe, CBR first became detectable, displaying the higher temperatures it then possessed in the still early Universe. (If anyone had existed at that time, a near-infrared radiation detector would have been needed since that was the spectral range for a temperature of 3000° K; the CBR had cooled through the temperature range that would be in the Visible while the opaque Dark Age still dominated.) From Decoupling to the present time, the CBR has experienced a redshift of ~1200. This relativistic redshift resulting from expansion is involved in causing the temperature of 2.73° K that is recorded today.

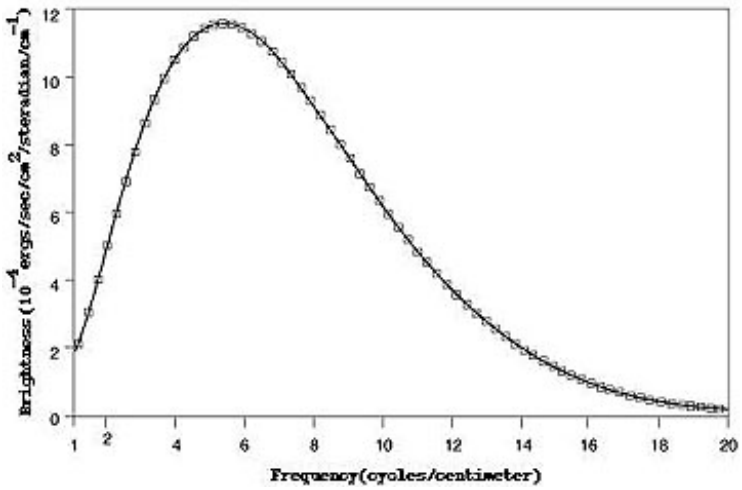
The photon radiation throughout the Universe now being measured is a manifestation of the present-day Cosmic Microwave Background (CMB), inherited from the original radiation (much hotter and therefore then of much shorter wavelengths in the Infrared) released at the Big Bang. Astronomers commonly refer to the CMB as the general residue of photons that were produced and released during particle interactions in the first minute of the Universe - colloquially, the CMB is the remnant of the "burst" of radiation that marked the "explosion" of the Universe (but which really didn't explode in the sense of detonation of a nuclear device in which there is an initial "flash" of light). It is also referred to as the "afterglow" of the BB. This radiation seems to be very uniform and isotropic throughout the Universe. The vast majority of all photons found in the present Universe are tied up in the background radiation. However, despite their huge numbers, it is estimated that they comprise only about 1/50000th of the mass contained in all the galaxies. The present ~3° K value is consistent with a predictive model that requires very energetic high temperature radiation (mainly gamma rays, with much shorter wavelengths) that constituted the early CMB released soon after the Big Bang to cool drastically by *adiabatic* (no energy added or removed) *thermodynamic expansion* (a good Earth analog: expansion of an air mass is accompanied by release of heat with resultant cooling) within a Universe having at the least the presently observed spatial limits. Mechanistically, as space is stretched the original shorter wavelength photons experience a corresponding lengthening of their wavelengths into the microwave region and so lose energy ($E = hc/\lambda$) which in turn is expressed as a much lower temperature.

The extraction of a weak radio telescope signal (after receiver noise was subtracted) in the microwave region at 7.3 cm (4.1 GHz) was made in 1965 by R. Wilson and A. Penzias (for which they received the Nobel Prize in Physics; their discovery was somewhat accidental at first since they were trying to track down what they perceived as "noise" in their radio telescope). (Actually, a similar signal was first detected in 1961 by E. Ohm, then verified by B.Burke, but not connected to the CBR prediction.), with its correlation to cosmic background radiation. Correlation of Wilson and Penzias' discovery and its implication for the predicted background radiation was then confirmed by R. Dicke and his group at Princeton.

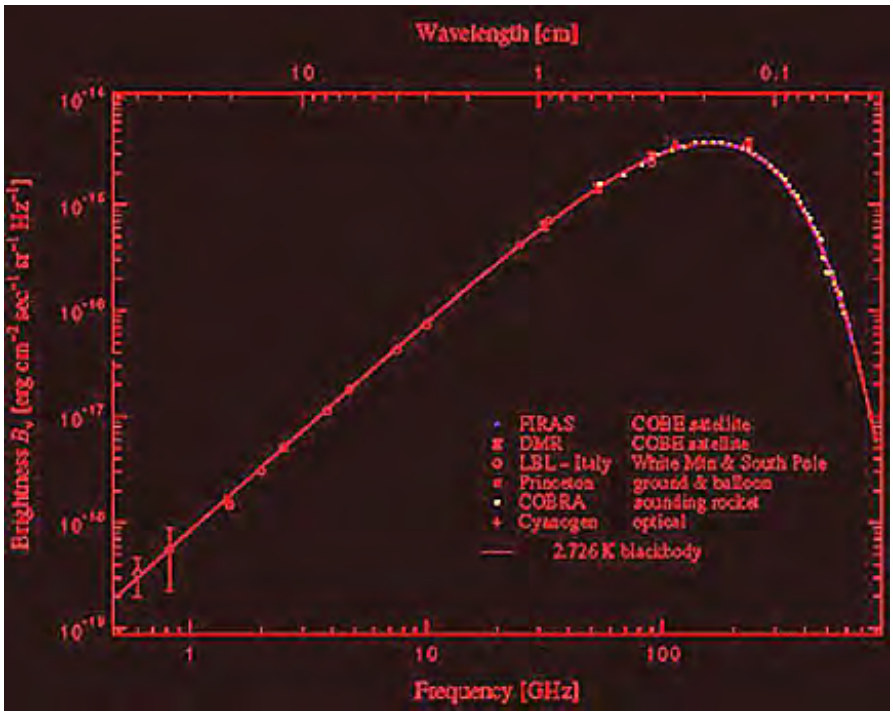
This accomplishment, along with the work by Hubble, the theory of General Relativity by Einstein, the pioneering concepts of a primordial singularity by Lemaitre, the Inflationary Model by Guth, and supporting contributions by numerous cosmologists, astronomers, physicists, and mathematicians, taken together, make up the foundation concepts that support and explain the Big Bang in its present form. Further discoveries will likely lead to refinements but the fundamental premises and the proper numbers predicted from the general model now seem to be solidly substantiated.

The value of satellites in understanding CBR is well illustrated by COBE (Cosmic Background Explorer), launched in 1987 (check out its current COBE website. Earlier attempts by Smoot and others to map the apparent non-variant (uniform) background radiation over the entire sky using balloons and aircraft, to make measurements above the atmosphere which blocks out (absorbs) radiation in the .001 to 0.1 m wavelength region of the spectrum, gave strong hints of radiation uniformity but were subject to imprecision. With COBE, the mapping process was greatly improved so that a detailed chart covering the full sky was assembled in just a year. COBE verified the high degree of uniformity of the present background in all directions and also confirmed that the general expansion is extremely uniform in all directions.

COBE took extremely accurate readings over much of the wavelengths involved in constructing a blackbody radiation curve. These measurements were then combined with those covering other wavelengths and obtained by different means to produce this classic blackbody radiation curve (see page 9-2 for a review of blackbody radiation) in which the COBE values were so accurate that error bars could be omitted. (When the COBE curve was first displayed to participants at an Astronomy conference, the audience was moved to give a standing ovation; such an extraordinary curve with all points precisely on the best fit version is the dream of all experimental scientists.) When compared with curves determined experimentally for blackbodies of different temperatures, the best fit was to a 2.726° K body; demonstrating that the CBR radiation fits that curve at better than 99% accuracy (an astounding achievement seldom attained in most scientific measurements).

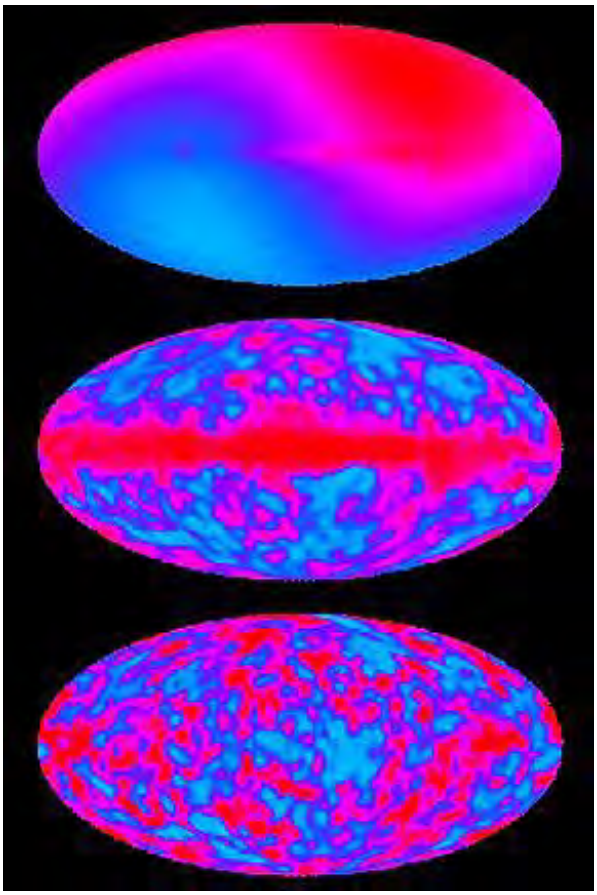


A variant of this includes measurements made by other CMR measuring experiments (different systems).

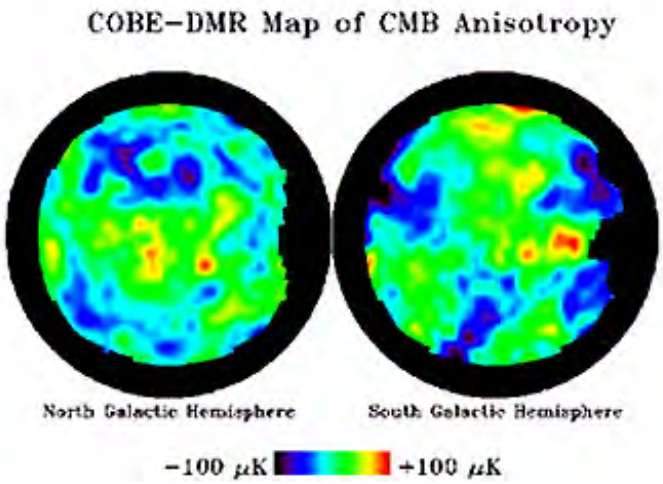


The two plots differ because of different brightness and frequency units and log values are used in the second diagram.

COBE allowed the mapping of radiation in the early stages of the Universe, (specifically, at the close of the Radiation Era some 300,000 [perhaps to 500,000] years after the Big Bang, when the plasma in the expanding Universe had cooled sufficiently to become transparent to photons) to an accuracy such that it showed variations in temperature and density as slight as 1 part in 10000 during the first billion years after time zero. Said another way, COBE proved the residual radiation after the Big Bang was **smooth** to within a fluctuation of 0.01 percent. (Had it been notably rougher, such irregularities would have forced the Universe either to collapse on itself or develop mostly black holes instead of stars.) It also established a range of +/- 30 microKelvins as the range of differences around the average CMR temperature; these irregularities are of the order of 1 part in 100000. The maps below show the broad distribution of these minute temperature differences (ripples) across the early Universe as detected by COBE's DMR (Differential Microwave Radiometer) using data collected at 53 and 90 GHz. The blues represent slightly cooler and reds slightly warmer temperatures - thus also define regions of greater and lesser densities.



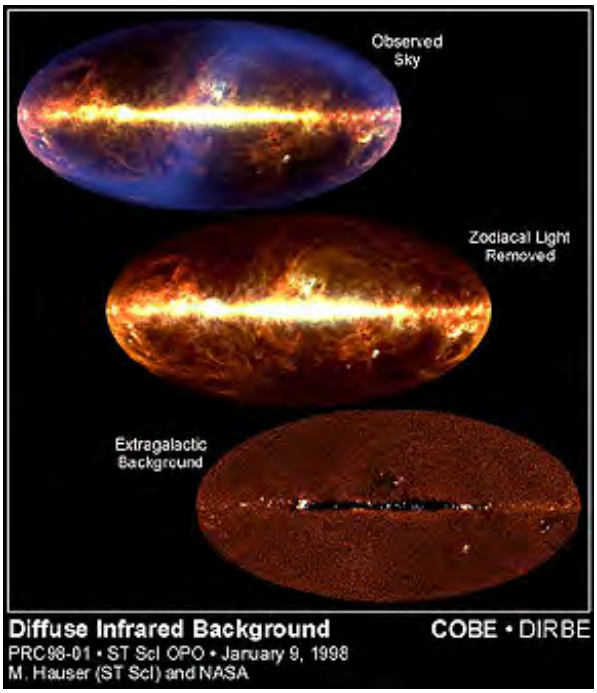
The top map is the "raw" data plot in which the dipole effect caused by the Doppler motion of the Milky Way galaxy has not been removed. The middle map results when the dipole effect is eliminated, but the radiation from the Milky Way (central band) has not been compensated for. The bottom map is the final product with both dipole and galaxy effects removed - this is the one usually cited as the model for CMB distribution. Another such plot, using different colors, recasts the distribution in terms of the northern and southern hemispheres of the celestial sphere:



These small differences were, however, vital in allowing matter to break from the initial extreme uniformity into regions of slightly cooler, denser conditions where the protogalaxies could begin to form. Eventually, in the early Universe these seed fluctuations promoted localized clotting of particles that became gravitational centers whose growing attraction of more matter led ultimately to development of the billions of galaxies that populate the Cosmos as we now know it.

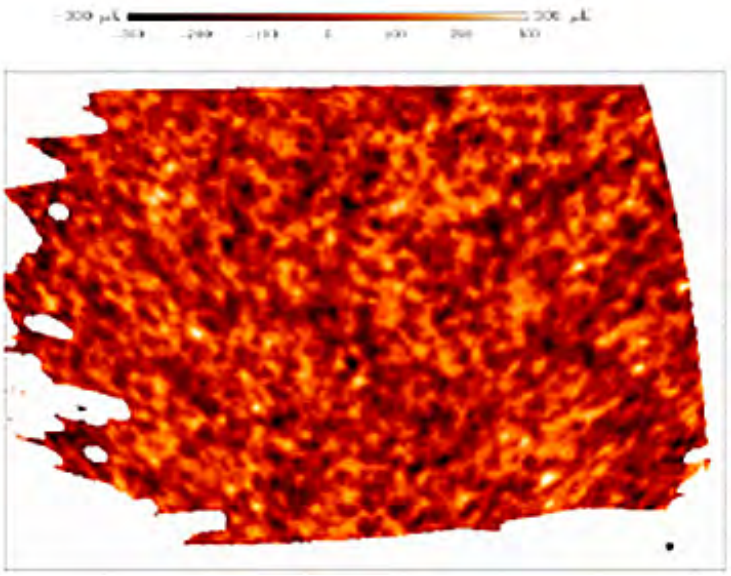
COBE has allowed an estimate of the total energy in the Universe by sampling yet another part

of the spectrum. This results from painstaking analysis of radiation in the far Infrared using the Diffuse Infrared Background Experiment instrument onboard. This measures heating of the dust distributed throughout the Universe, using windows at 140 and 240 µm. However, the overall background is "contaminated" by dust and other sources within and around the Milky Way, the Earth's atmosphere, and other sources, which require correction. The procedure is indicated in this figure:

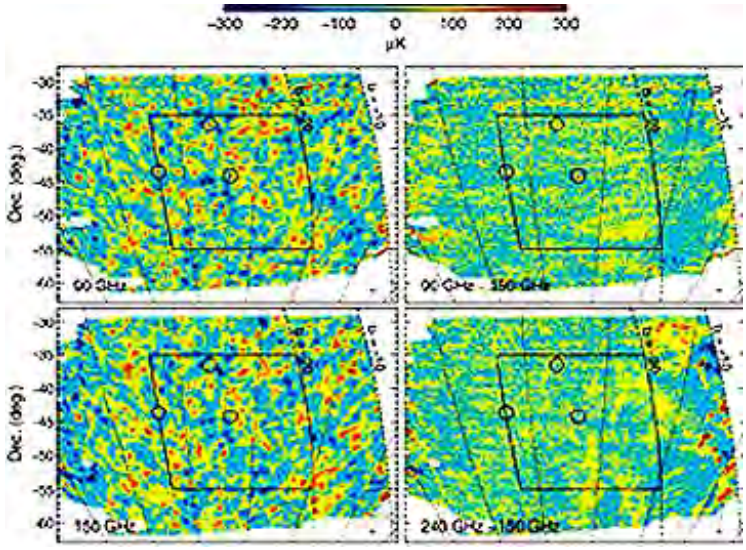


The upper panel shows a sky map of the Infrared radiation for the whole Universe with a bright central band representing the Milky Way contribution. The central projection is the change after Zodiacal light is removed. The bottom panel is the residual Infrared radiation for the Universe after the Milky Way Galaxy's influence has been removed. The net effect is that there is much more starlight in the Universe as "fossil radiation" than heretofore suspected owing to the masking by dust (ranging from near-Earth to intergalactic) whose influence is now accounted for with this corrective DIRBE inventory.

In April, 2000 a group of scientists presented the results of project BOOMERANG (acronym for Balloon Observations of Multimetric Extragalactic Radiation and Geophysics) One output was a more detailed map of 3% of the sky which shows variations (with a 35x improvement in resolution) in CBR at the end of the Radiation Era - which also signals the beginning of the Decoupling Era marked by the recombination of protons and electrons to form Hydrogen atoms. This map was constructed by measurements obtained with a passive microwave telescope suspended on a balloon for 11 days at approximately 36400 meters (120,000 ft) above the Earth's atmosphere over the Antarctic. The variations depicted are in units of microKelvins.



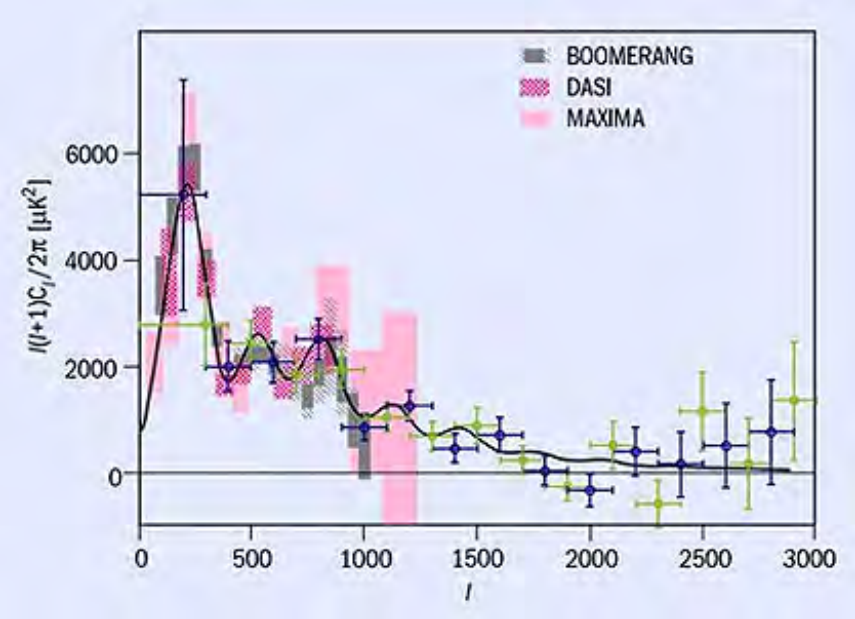
Here are several more maps from this experiment using radiation detected at different wavelengths. The upper and lower left maps are at 90 and 150 MHz respectively; the two right maps are differences between 90 - 150 (top) and 150 - 240 (bottom) MHz.



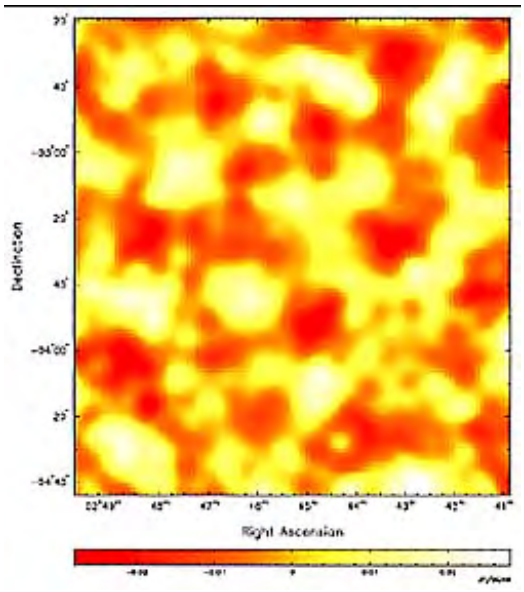
The Boomerang scientists envisioned the early Universe to be full of 'sound waves' compressing and rarefying matter and light, producing 'acoustical peaks'. They used this model to calculate the distribution of Dark Matter (30%) and Dark Energy (65%), with the remainder about 4.5% Ordinary Matter, within the components of the physical Universe.

COBE and Boomerang results are confirmed, with more detail, by the CBI (Cosmic Background Interferometry) experiment run jointly by CalTech and the NSF. The CBI is located in dry air in Chile's Atacama desert, at an altitude 0f 5080 m (16.700 ft). It started data collection in 1999. Here is an onsite photo of this sensitive instrument:

Thirteen 1 meter diameter dish antennae are synchronized in an array with a broad frequency baseline from 26 to 36 GHz. Each dish receives a different wavelength signal, and interferometry is used to integrate the data from which a power spectrum is produced, as shown by the solid line curve, with values from other CMR instruments also shown:



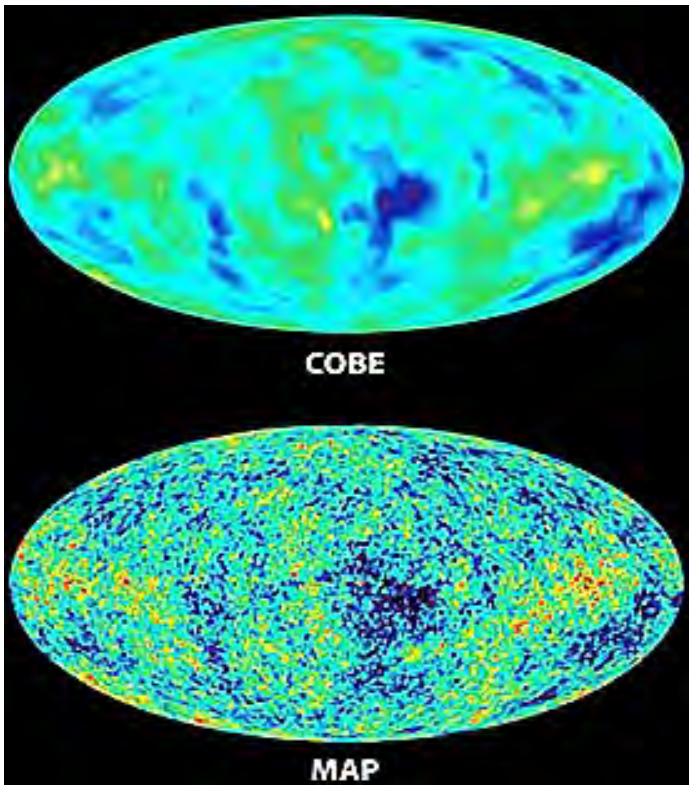
This next figure is a map of the background radiation over an area equivalent to about 2.2 degrees in declination units (2 widths of a full Moon). The differences being measured are temperature values in microKelvins (μK) that vary around the mean sky temperature of 2.73.. °K.



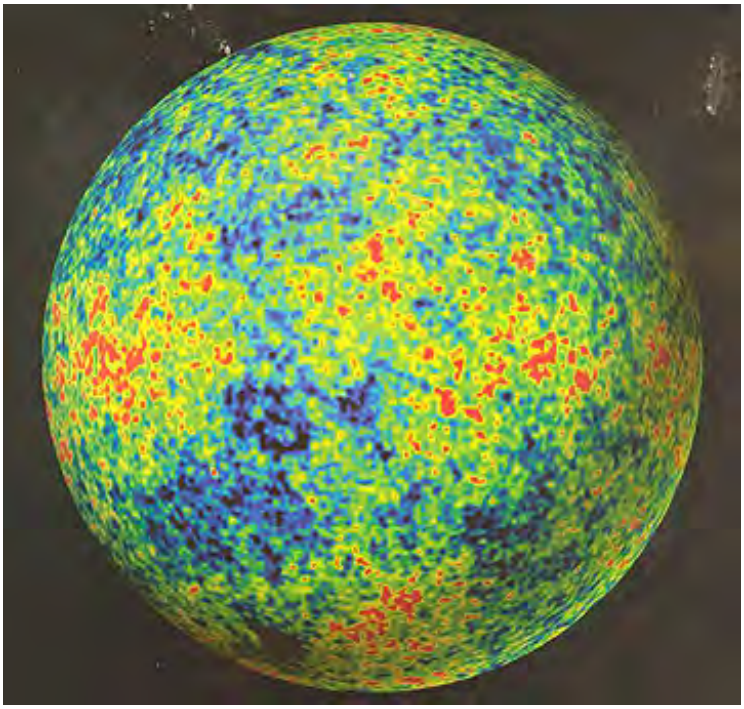
What is being sensed are small temperature differences (range of ~100 µK) when the CBR was around 3000° K. The yellows indicate slightly hotter regions compared with cooler reds and blacks. Associated with these differences are variations in material density - the hotter regions have higher densities, indicating matter has already begun collecting and interacting to generate heat (possible indication that early stars had formed). This observation supports the idea that matter in the Universe at this early time was unevenly distributed, thus pointing to the first stages of (increasing) density/gravity variations required to initiate the process by which galaxic clusters form.

The results from COBE proved of such import to understanding the early Universe, especially the small but vital fluctuations it detected, that a more sophisticated satellite, WMAP (Wilkinson Microwave Anisotropy Probe), was launched in July of 2001. Background information on this important new astronomical observatory can be found at NASA Goddard's WMAP site.

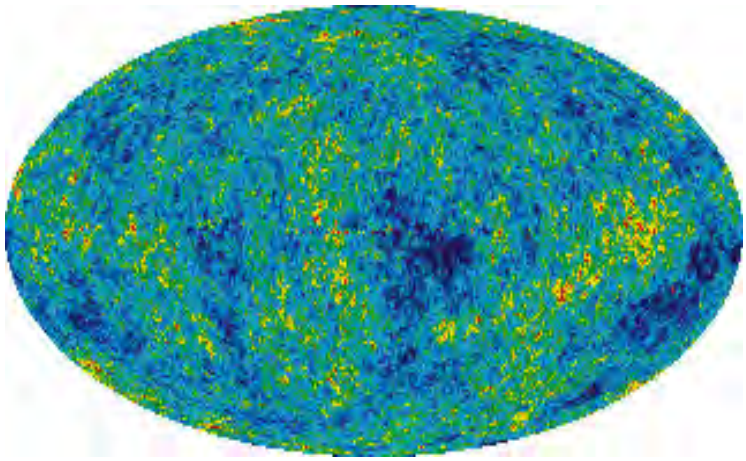
The long-awaited preliminary results from MAP were announced at a press conference on February 11, 2003. By then MAP was renamed WMAP, honoring the late David Wilkenson, a leader in the field from Princeton. Cosmologists at this conference stated that the WMAP results were among the most important in the last half century in deciphering the history of the early Universe. The higher resolution of WMAP, in terms of ability to measure even smaller temperature variations, is evident by comparing the new all-skies thermal map from WMAP with the equivalent coverage by COBE:



This pair of plots clearly demonstrates the great leap in resolution provided by WMAP, leading to much more detail about the very slight but signficant variations in CBR temperatures. A better view of the distribution of the very small but important blackbody background temperatures is afforded in this projection of WMAP measurements that describe conditions less than a half million years after the Big Bang; three additional spherical maps are needed to cover the entire sky:



Slight differences from the first WMAP shown above appear in the version below (Spring, 2009) which represents five years of data gathering.



The variations in color in this, and other maps above, give the impression of notable differences in value. Actually, the temperatures vary by very small amounts. The temperature differences give rise to slight variations in density. But these variations were key to the gradual buildup in differential densities that eventually brought about differences in gravitational attraction leading to local clumping into the first stars and larger scale clumping into the first galaxies. The variations represented by the spread of colors as seen in the WMAP map are an indication of the spatial distribution of the clumps in the early Universe; even though the CBR is now much cooler than in these early times this distribution pattern has persisted.

Some very far-reaching conclusions about the Universe have been drawn from interpretations of the WMAP data. One is a new (but still not necessarily the most accurate, although an accuracy of +/- 200 million years is claimed) age for the Universe of **13.7 billion years**. The value has superceded the earlier 14.7 billion years that came out of COBE and other studies. This is based mainly on what is believed to be a better estimate of the Hubble Constant: 65 km/s/Mpc. Another WMAP conclusion is strong confirmation of the **reality of Inflation** during the first fraction of a second after the Big Bang.

WMAP leads also to a better estimate of the amount of detectable Ordinary Matter in the Universe and values for the invisible matter/energy that so far has eluded direct recognition and measurement. These have been reset at 4% for Ordinary Matter, whereas Dark Matter is 23% and Dark Energy 73% (but the results offer no clear indication of the nature of these dark states). (Only a fraction of Ordinary Matter is luminous [gives off detectable visible light], so that the vast bulk of the Universe's constituents is in fact non-luminous and thus hard to detect.) The time when the Universe first became **transparent** is now given as **~380000 years after the BB**. Indirect evidence from WMAP data suggest that **massive stars** had begun to organize even earlier, perhaps over an interval of about **200,000-300,000 years post-Big Bang.** .

COBE and WMAP data have been used to refine many of the fundamental physics and cosmological parameters as shown in this table:

However, the list below focuses on what cosmologists consider the 10 most important parameters whose values have been better calculated using the CMR data:

The English cosmologist, Dr. Martin Rees has cited his own list of the irreducible number of fundamental parameters that determine the development of the Universe we can measure in his book *Just Six Numbers: The Deep Forces that Shape the Universe*, 2000, Basic Books. These are: 1) **N**; = ratio of the electric force holding atoms together to the (much weaker) force of gravity, 10^{36}; if this number were larger then only a miniature and short-lived Universe would have formed; 2) **ε** = a measure of the strong nuclear force, determined from the energy released in the fusion of Hydrogen to helium (differential of 0.007); if much different than this value, a different mix of chemical elements results, with carbon very scarce; 3) **Ω** = amount of matter/energy in the Universe (see below); it is the ratio of actual density to the critical density (see below); if too high, the Universe collapses and if too low, expansion would be so fast that there would be insufficient time for stars and galaxies to form; 4) **λ** = antigravity force (Cosmological Constant); if too large, the Universe would have expanded so rapidly as not to develop as it has; 5) **Q** = force needed to dissemble a gravitationally stabilized cosmic structure (star; galaxy), measured as the ratio of the energy needed to overcome the gravity force to the energy bound in the rest mass of the cosmic body, given as 10^{-5}; if Q deviates from this value, a smaller number would have prevented the ripples that gave rise to galaxies, leaving only a dispersed gas but if much larger, only black holes would exist today; 6) **D** = number of spatial dimensions (3) needed to sustain life in our planet and similar bodies; 2 or 4 would have doomed our existence.

In addition to cosmological parameters, there are also many basic physical constants that are involved in both Cosmology and Physics as practiced on Earth. A discussion of these is given at this web site. More discussion of many of the parameters, as they bear not only on the origin and development of the Universe but on the appearance of organic life and ultimately Man, is given

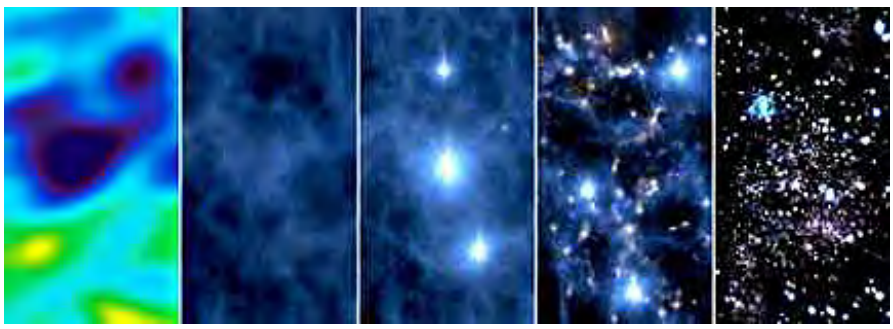
in the philosophical discussion subsection near the bottom of page 20-10.

The data displayed in the WMAP, CBI and other CMR maps also bear on the model that predicts the Universe had undergone a dramatic Inflation in its initial moments, and in effect provide a positive test of that concept. They likewise point to the notion of a flat Universe that will expand forever (see below).

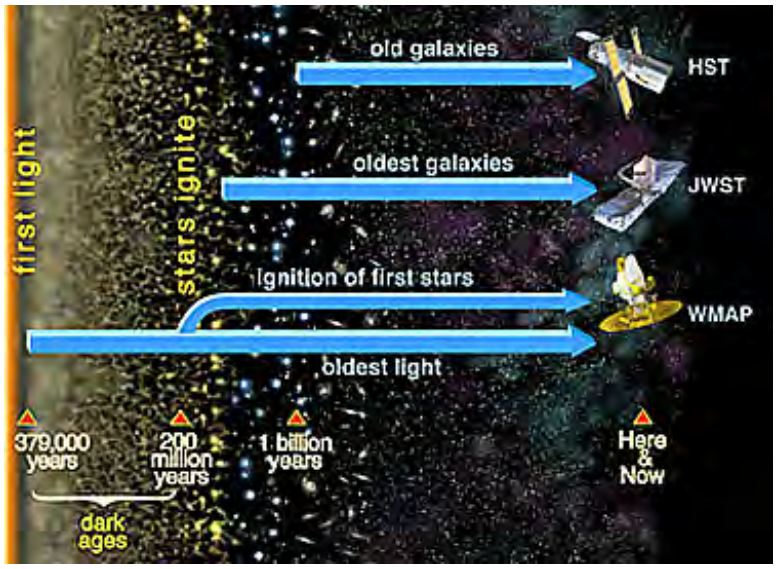
A recent announcement from Hubble scientists carries this cosmic background concept into the visible radiation realm. Based on estimates of quasar populations at the farthest reaches of observable space (the Deep Field region), extrapolations of visible light sources to the entire Universe can be made. Results suggest that most of these sources are now accounted for and that the total amount of visible light which persists throughout the Universe is approximately of the order to be expected (by calculation) from the same model that predicts the amount of Cosmic Background Radiation. In other words, as different parts of the EM spectrum are analyzed for total energy involved, the numbers remain consistent with expectations and thus support the energy distribution predicted from the Big Bang model. The overall notion of an *expansion* appears on firm ground based on the ever accumulating scientific evidence.

(A CAUTION: A report issued in November, 2003 presents some support for a very KEY topic of controversy in Astrophysics and Cosmology: Have some or even all of the fundamental constants been *constant* throughout the Universe's history? The speed of light is a leading candidate for dispute and ingenious arguments indicating possible variation. In the report, evidence is cited that the strength of the attraction between nuclear protons and orbiting electrons may have been much greater in the early days of the Universe. The role of quintessence (top of next page) is cited as the factor responsible for this. Other constants are being challenged but until incontrovertible proof is accepted, the "rule of thumb" is to stay with the values (subject to possible minor modifications) cited above.)

Some of the recent ideas on the start times for the first stars and galaxies received support and specificity from the WMAP results. The first stars began to form as Supergiants about 200,000,000 million years after the Big Bang. The first galaxies began to organize some three hundred million (300,000,000) years later (possibly earlier). This next diagram depicts these stages (from left to right): 1) initial stages of CBR variations; 2) clots of matter prior to organization as stars; 3) the first supergiants; 4) developing galaxies; 5) galaxies after the first billion years.



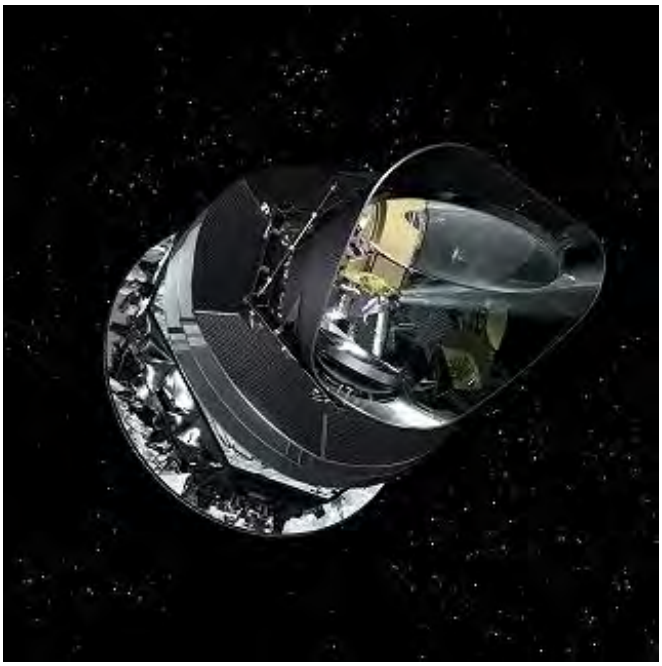
The time lines for the first stars and galaxies as measured by different space telescopes (JWST is the James Webb Space Telescope planned for 2010; its mission will focus on the early eons of the galaxies, so that the starting time shown above is a "best estimate" for now) are shown in this diagram. Of special import is the new estimate of when the first stars started to form - about 200 million years after the Big Bang.



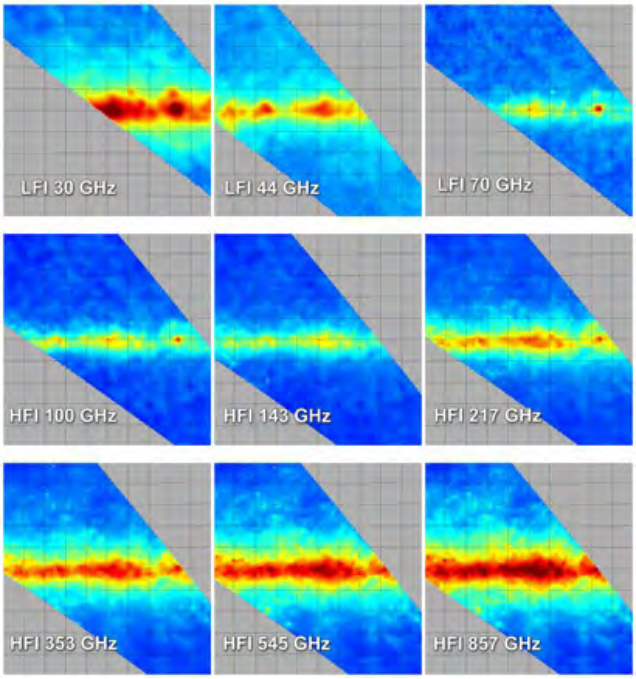
A major future objective of WMAP still to be addressed is to measure extremely small temperature fluctuations that should support/confirm the existence of gravitational waves. These were first postulated by Einstein as a consequence of his General Theory of Relativity. Gravitational waves represent moving disturbances within gravitational fields that are generated by various interactions of matter and/or energy, such as collisions of Black Holes or Neutron stars. With their force particles, the gravitons, they are analogous to electromagnetic waves, with their photons, except that gravitational waves can move unimpeded through matter that itself interacts with photons by absorption. Like the graviton, gravitational waves have yet to be detected but their behavior and influence within the Universe can be simulated with computer-based models. As gravitational waves move through space, they cause the geometry of space to oscillate (stretching and squeezing it). The wavelength of a gravitational wave depends on the mechanism of its generation.

Theory holds that gravitons and gravitational waves must have first been created during the Inflation period between 10^{-38} and 10^{-35} seconds at the outset of the Big Bang. These waves participated in the extreme expansion during those moments and as a result their wavelengths were greatly elongated. The inflationary gravitational waves played a key role in bringing about the slight variations in the distribution of matter and energy during the Radiation Era which ended in the Decoupling Era at which time photons were no longer scattered - this latter period is the earliest in which Cosmic Background Radiation could then be detected. WMAP is seeking to determine more exactly the temperature fluctuations in the CBR field which correspond to the pertubations imposed by the gravitational waves. In theory, these waves are detectable by analysis of the CBR coming from the Cosmic Microwave Background; gravitational waves will cause the radiation to be right or left polarized whereas density variations in the CMB will induce radial polarization (the two modes of polarization must be separated and distinguished by Fourier analysis.

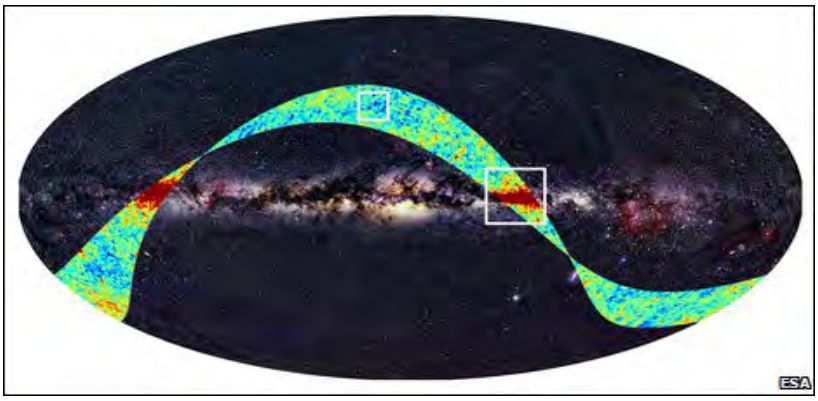
The CBR phenomena are proving to be especially fruitful in the study of the Cosmos. It is not surprising then to learn that more satellites are to be put into space to observe the CBR. Next up is Planck, which was launched with its sister satellite Herschel on May 14 of 2009. It will eventually reach its final location at L2, the second Lagrangian point between Earth and the Sun. That location is approximately 1,500,000 km (930,000 miles) from Earth. Planck will have instruments that can measure very small temperature differences using data from 9 different bands. Here is an artist's rendition of Planck. Check out its homepage at this ESA website.

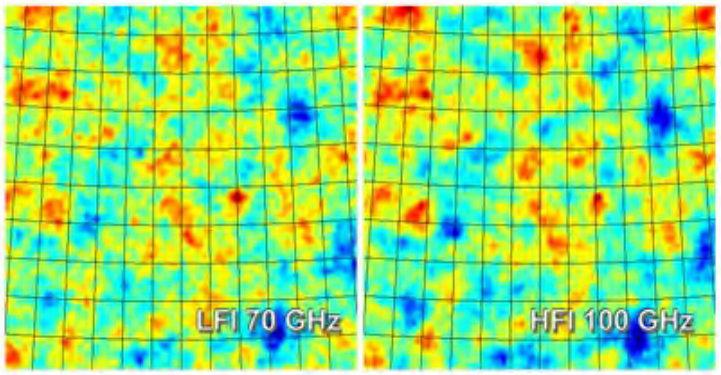


In August, 2009 Planck began returning data from which images are made. This example shows the Milky Way (central band) and adjacent areas of the sky (covering a 20° by 20° area) at 9 different microwave wavelengths (LFI refers to Low Frequency Instrument and HFI is the High Frequency Instrument):

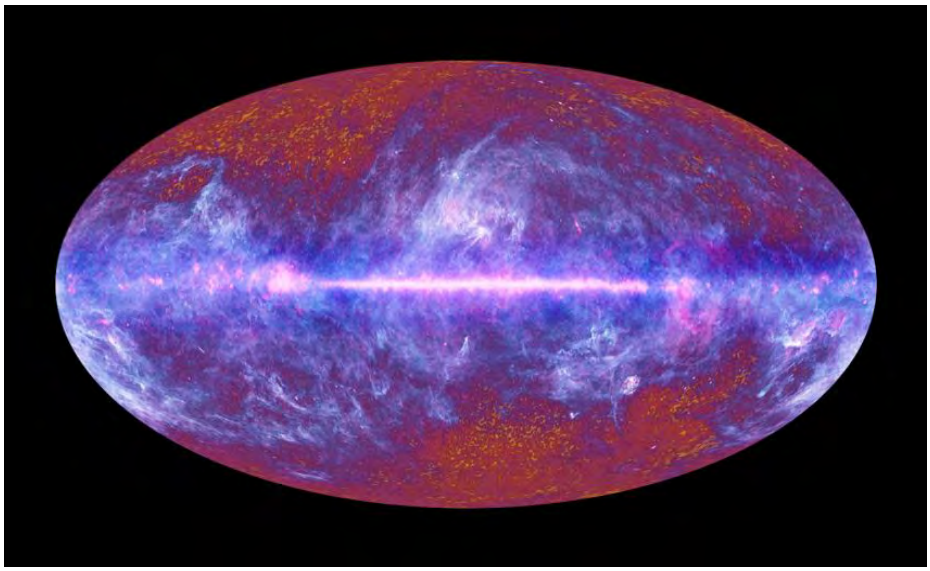


The first full sky Planck microwave map shows a strip that actually covers 360° but is curved as shown on this projection. The map has been superimposed on a visible light image that highlights the Milky Way (which shows up as the dark red brown patches in the map). The regions off the Milky Way represent temperature differences (as with WMAP, red is warmer than blue) associated with the time of First Light about 300,000 years after the Big Bang. This rendition is not impressive as such, but the image below it shows some details in a 10° by 10° segment of the sky at high latitude off the Milky Way).

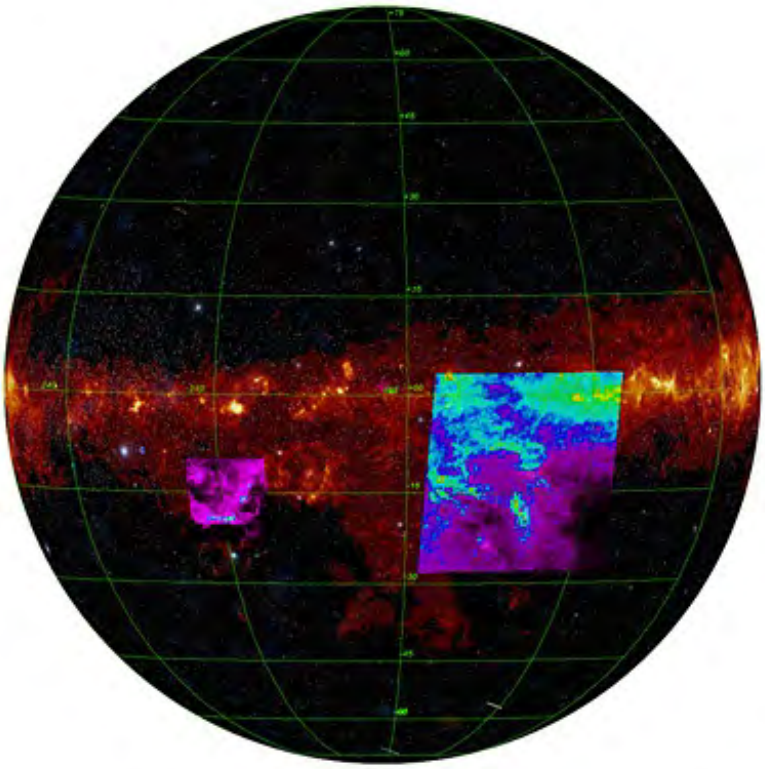


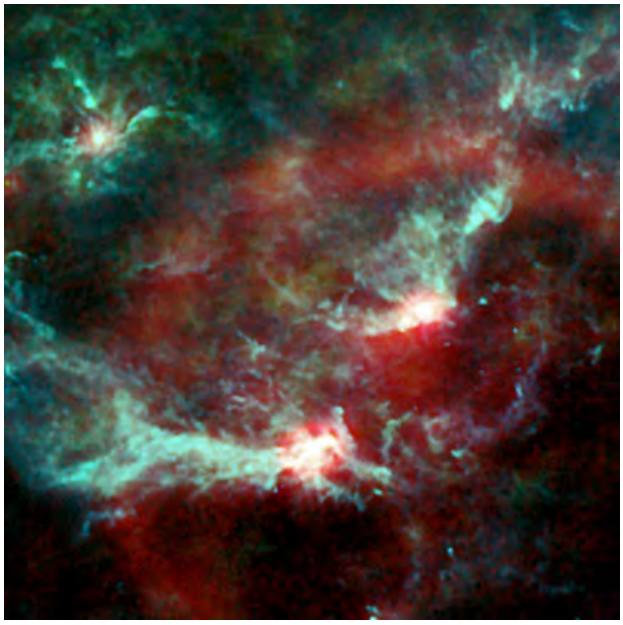


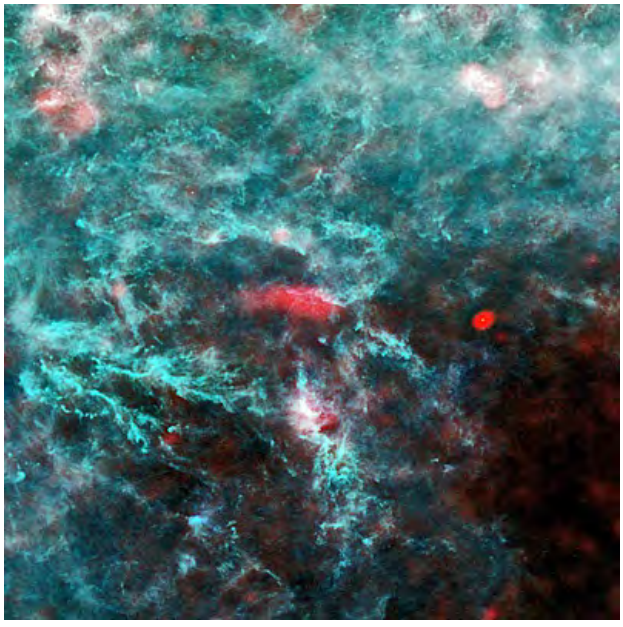
In July of 2010, the Planck team released the first full sky map made by the Planck telescope. Most of the blue in this map is associated with the Milky Way galaxy. Much of the CBR is rendered in red. A few anomalies are evident but it will take up to 2 years for the team to produce details of variations within the radiation.



This is a low resolution map. Planck can produce false color composites of selected regions using any 3 of the 9 bands in its sensor. These show much detail. Two examples are the Orion Nebula and star formation in the Perseus cluster. Their locations are indicated in the circular map shown first:







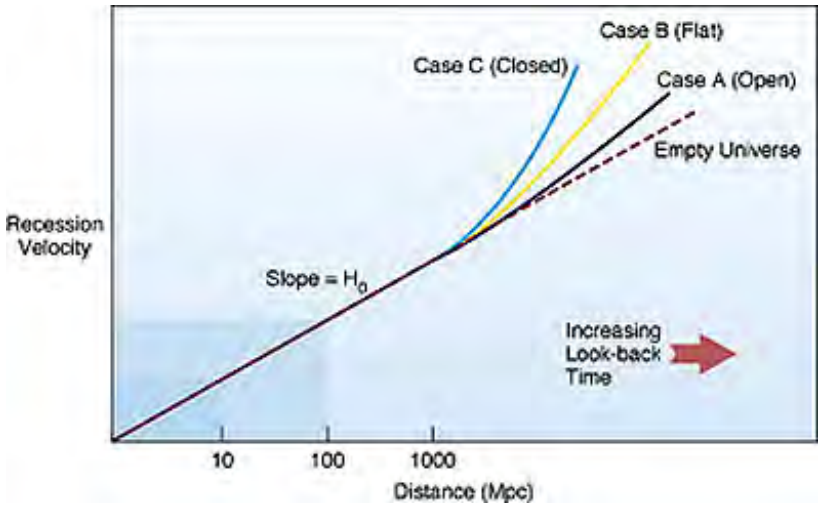
Planck will produce a series of image strips covering the entire sky that will become available in (or soon after) 2010. The resolution will be superior to that of WMAP. This should aid in improving our understanding of the state of the early Universe.

Evidence for the Expanding Universe

The modes of behavior of the Universe over time can be classified in several ways: 1) it follows either Newtonian or Relativistic physics; 2) it commenced with or without a Big Bang (i.e, expanding vs steady state); and 3) for the Big Bang case, the growth has been controlled either by Standard Model physics or has been influenced by the Cosmological Constant (or both?).

As a fundamental conclusion drawn from the general acceptance of the Big Bang model for the Universe's origin and development, the initial small space that developed in the first minute has been continuously enlarging - a process analogous to expanding in the manner described on the previous page. However, the precise nature of this expansion, still not fully known, depends on the specific expansion model, as we shall see below. This is related to the amount of mass/energy available to control or influence the expansion. As we will see in the following paragraphs, proposed geometries of the expanding Universe range from spherical to hyperbolic to flat. The duration of expansion ranges from finite to infinite. The terms "open, closed, flat" refer to certain constraints on the curvature of space and on its expansion history.

The type of Universe "shape" model - open, closed, flat, spherical - is a factor in the change in the Hubble constant (and the corresponding redshift) with time. A generalized relationship depending on expansion models is shown in this next plot:



Before reviewing the various models that were proposed in the 20th Century, we pause to briefly describe a useful and (deceptively) simple view of the Universe embodied in the term Hubble Sphere. This is almost a synonym for "observable Universe" but with one unique property. The sphere moves - it is just that which can be observed at any arbitrary point in the Universe. Earth has its own Hubble Sphere. But a planet in a galaxy 5 billion light years is another arbitrary point and has its own Hubble Sphere. The sphere at each such point has its Hubble Length, which is just the distance outward from the observing point, as an *arbitrary center* (remember, the Universe actually has no meaningful center), that light has traveled in 1 Hubble time ($t_H = 1/H$). In this framework, that distance is represented as the farthest out that a particular observer at a point can look with the best telescopes to see the first evidence of the Big Bang (which is not really possible owing to the opacity soon after the BB); it is closely related to Lookback time (time for light from an emitter to reach Earth or any other point of reference). Consider the Hubble distance to be the radius r for a sphere that encloses all of the Universe that can presently be seen. That outer limit boundary is, of course, a *time horizon* and not an actual physical surface encompassing the sphere. As we progress into the future and our instruments "see" still farther, the apparent surface of the sphere moves outward with the increase in r_H.

There are galaxies beyond the Hubble Sphere; they just haven't been seen yet but will come into view later. Beyond the outermost galaxies, assuming they occur at light year distances equivalent to that of a precisely known Hubble Age, we cannot as of now specify "What's there".

Let us now say a few things about the **size** of the known Universe. It would seem to be determined by the Hubble Distance (D_H), which relates to the Hubble Age, around 14 billion years. This is the distance out to the *event horizon*, the farthest out in spacetime that we can see discrete particles or objects in the Universe. To quantify the distance in Earth kilometers [or miles], just multiply the distance that light travels in 14 billion years by the speed of light. Thus: 14,000,000,000 b.y. x 300 x 10^4 km/sec x 3600 sec/hour x 24/hrs/day x 365.4 days/year. For this case, the result, which I will call D_H, is 1.3245 x 10^{24} km, or about 1.3245 septillion kilometers.) From the Hubble Sphere model, one might assume that the sphere has a diameter of 2 x D_H, particularly when one is aware that the event horizon is essentially the same looking outward, say from the North Pole at the northern celestial sphere and from the South Pole at the southern celestial sphere.

But, this is not so. In relativistic space expansion, the distances outward in opposite directions from the Earth framework are not additive. This is due to the fact that all points in the singularity that are now galaxies were next to each other at the beginning and have simply drawn apart with the expansion of space. With no meaningful center, we can only state for now that space has expanded some finite amount in 13.7 billion years. Euclidian size is not a valid way to look at the Universe, whatever "shape" it may have, as implied from the paragraphs later on this page. In trying to think about "size" there is a further complication. The expansion during the Inflation period (see page 20-1) may have proceeded at rates faster than the speed of light. If so, the Universe may really be much bigger than what we deduce from event horizon distances. We get our idea of distances only from measurements of z and H as determined from we see now in the Universe after the galaxies formed. Prior to those times, inflation expansion, yielding much greater z and H values, could have pushed the outer edge of the Universe to distances well beyond what can be detected as *apparent* event horizons.

One interesting corollary to this reasoning is that in principle cosmologists can detect galaxies that appear to be receding as speeds that when calculated seem to be exceeding the speed of light. This is because the Hubble constant isn't constant, it is increasing. This apparent contradiction to Einstein's relativity demand that light has one fixed finite value is explained on pages 40 and 42 of the previously cited Misconceptions about the Big Bang, by C.H. Lineweaver and T.M. Davis, *Scientific American*, March 2005.

So, what can we say about our understanding of the size of the (our) Universe. Its minimum size must be at least as far out in spacetime as we can see galaxies, quasars, and supernovae - 14+ billion light years to the currently known event horizon. We cannot [yet] see timewise to anything before the Radiation Era; Cosmic Background Radiation, which traces to about 300,000 years, is pervasive and thus not location-specific. The maximum conceivable size is infinity, with "outer limits" reachable only in infinite time. If the Universe is indeed infinite, its present outer limits are not fixed in any way, as they will enlarge forever in their expansion towards infinity. If the Universe is proved to be finite (possibly contrary to the most likely scenario - see below), then its boundary is almost certainly beyond the event horizon we now see - there are more galaxies farther away which will become visible as time progresses and D_H lengthens. The safest conclusion now reached under currently postulated scenarios is just that the Universe must be larger than the presently determined horizon distance.

But thoughts on size are changing. We can set a lower limit of the **observable** Universe at the currently most favored age of 13.7 billion years, that is, we can now see out to that part of the outer Universe that contains all the stars within a "sphere" in which light has traveled no longer than 13.7 billion years at its present speed. And we see that value in any direction we look. Perhaps this means that we are near the center of a finite Universe. If so, its diameter would about 29.4 billion years. But when the fact that the Universe is now accelerating (as discussed

on the next two pages), this must be taken into account.

Dr. Neil Comish and colleagues at Montana State University have done a preliminary calculation on how much the Universe has expanded since the first radiation around 13+ billion years was released and is now just being received on Earth. But because the Universe has expanded tremendously since the Big Bang, the distance that light has been traveling continues to increase during the 13+ billion year timeframe. The source of the primordial light leaving the very early Universe from any point (say, the first star), has thus itself moved much farther than the time-travel distance. Their model leads to an astounding number - 78 billion light years to the edge of **today's** Universe and a possible diameter of a spherical Universe of 156 billion l.y. (so large both because of the effects of spatial expansion and acceleration rates since about 6 billion years ago. One might conclude that this requires different values for the speed of light in the past, which might seem to violate the General and Special Laws of Relativity, but the astronomers point out that this speed remains constant while the distance a photon released at the beginning of the Universe must travel is what is increasing. This estimated diameter and other figures for size are obviously still controversial but the "true" size should be increasingly refined in coming years.

A summation: Lets say that we detect a galaxy that is 13.5 billion light years away. It has taken that long for light near the dawn of cosmic time to reach Earth. Is that the size of the Universe? An emphatic NO! It must be larger. At the time of light departure, the actual Universe THEN was much smaller - about a 1000 times less than at present. Over the 13.5 b.y. travel time, the Universe has expanded (not at a uniform rate but with the decreasing acceleration in its first half of existence and now replaced by an increasing acceleration, as described on the next page). The important point is that growth of the Universe during the 13.5 b.y. time of travel has stretched out the actual size even as the light had to continue to make its way to Earth observers. The fact that this burst of light 13.5 b.y. ago ultimately reaches us happens because the total growth rate is notably less than the speed of light. Relativity is built into this description (consider the analogy of one walking from one end of a train to the other even as the train moves along track; the total distance traveled is some combination of person walk plus train progress). Hopefully, this paragraph may clear up any uncertainties you retain after working through other relevant parts of this Section. But, if you would like more information on the size and shape of the Universe, consult this Wikipedia website that discusses the Observable Universe.

Major Models for the Spacetime Universe:

Relativity has played a vital part in the models of the Universe that remain the most plausible. The expansion of the Universe (in terms of rate of change of the Scale Factor $R = r/r_0$) from a relativistic framework can be summarized as the Friedmann equation. For the distribution of matter in the Hubble sphere, the equation considers the contributions of both the gravitational potential energy and the kinetic energy of expansion. We give it here in two forms, the first as a differential equation:

$$(dR/dt)^2 = (8 \Pi G)/3 \rho R^2 - kc^2$$

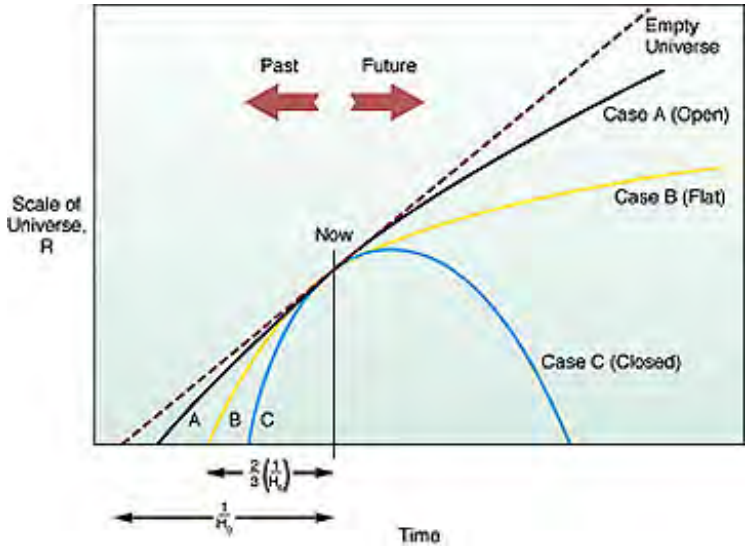
And the second (introducing H):

$$H^2 - (8 \Pi)/3 G \rho = - kc^2/R^2$$

In these equations, Π (pi) is the familiar constant (ratio of a circle's circumference to its diameter = 3.14159...), G is the Universal Gravitational constant (6.6726×10^{-11} $m^3/kg/sec^2$), ρ is a Greek letter denoting the average density of the Universe, k is a curvature constant in which values of 0, +1, -1 represent flat, spherical, and hyperbolic geometries respectively, R is the Scale Factor for the observable Universe, H is the Hubble Constant, c is the speed of light, and t is time. A solution to the Friedmann equation depends on which Universe model is being tested, as the

group described next has different values for key parameters.

Several cosmological scenarios, named after the scientist(s) who first proposed each (several scientists came up with more than one model), for various modes of expansion lead to different end results (shown graphically below for four general models).



In common, they all obey the Cosmological Principle, which states that the Universe is both *homogeneous*(defined as being the same at any point within space) and *isotropic* (as one looks out, appearing the same in any direction. This boils down to having essentially the same average distribution of matter/energy in all directions) on the largest scales (500,000 light years and larger. This is not violated at the scale of galaxy clustering since at the universal scale these tend to be "smoothed out" by having much the same patterns anywhere one looks. Open models also must be consistent with the restriction placed by the Second Law of Thermodynamics which from a cosmological standpoint states that over time the *entropy* (a measure of disorder of a system) must ultimately increase to (or towards) a maximum (total disorder). Interpreted at a universal scale this would lead to complete dispersal of galaxies and their stars (perhaps rearranged as randomly distributed Black Holes) and blackbody temperatures approaching zero. A corollary holds the initial singularity to have minimum entropy which then rapidly increases during the first moments of the Big Bang.

Note that when the above curves are extrapolated back in time, they strike the horizontal axis at different positions (times). This means that the age of the Universe will vary relative to the particular model being considered. Thus, although the current Hubble time ($1/H_0$), which depends on the accurate determination of the rate of expansion, leads to an age or duration of the Universe, that value can be modified when (and if) a particular expansion model is shown to be the best or valid one.

The following table (modified from Hawley and Holcomb, 1998) summarizes the principal Cosmological Models that have been developed and tested by calculations. They fall into two groups: Non-Big Bang and Big Bang. Another distinction category: Models in which the Cosmological Constant L (see below) is a factor (upper five rows of table) and the Standard Friedmann (or Friedmann-LeMaitre) models in which L is not involved (i.e., is O; bottom three rows); the three standard models also have Deceleration Parameters q (defined below) that include the value 1/2 in some way.

MODEL	GEOMETRY (k)	L	q	FATE
de Sitter	Flat (0)	>0	-1	No BB; exponential expansion; empty
Steady				

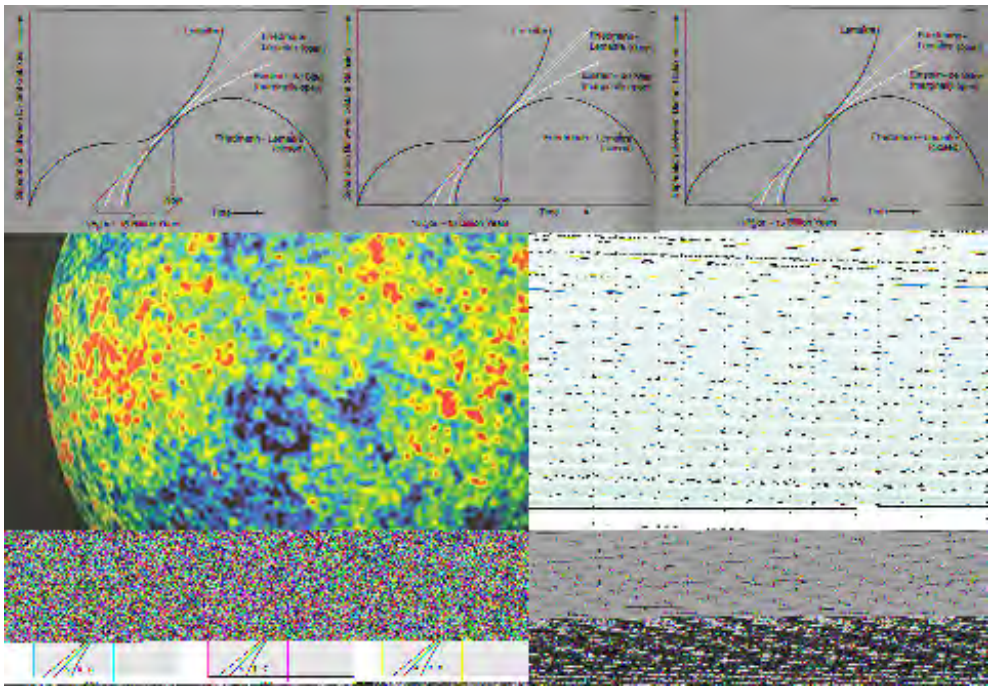
Steady State	Flat (0)	>0	-1	No BB; uniform expansion
Einstein	Spherical (+1)	L_c	0	Static; H = 0; now, gravity balanced by repulsive force; may be unstable
Lemaitre	Spherical (+1)	$>L_c$	<0	Expand; hover; expand
Negative L	Any	<0	>0	Big Crunch
Closed	Spherical (+1)	0	>½	Big Crunch
Einstein-de Sitter	Flat (0)	0	½	Expands forever; density at critical value
Open	Hyperbolic (-1)	0	0<q<½	Expands forever

q = The Deceleration Parameter: denotes the rate of change with time of the Hubble Constant and R; a positive value indicates acceleration; negative = deceleration.

L = The Cosmological Constant, introduced by Einstein to his field equations for General Relativity in order to provide some constraint to gravity (a counter-effect) to avoid an inevitable collapse of the Universe; if + (repulsive) L counteracts gravity; if - (attractive) L may be equivalent to the vacuum energy density associated with particles at the quantum level. (L in texts is also given by a capital Greek letter Λ). The current value for energy density within the observable Universe is between 1 and 5 x 10^{-26} kilograms per cubic meter.

The Steady State, de Sitter, and Einstein Universes, all non-standard, are currently not supported by observational evidence.

The more general diagram above showing four alternative expansion models can now be redisplayed in terms of the names associated with some of the specific models described in the above table:



From J. Silk, The Big Bang, 2nd Ed., © 1989. Reproduced by permission of W.H. Freeman Co., New York

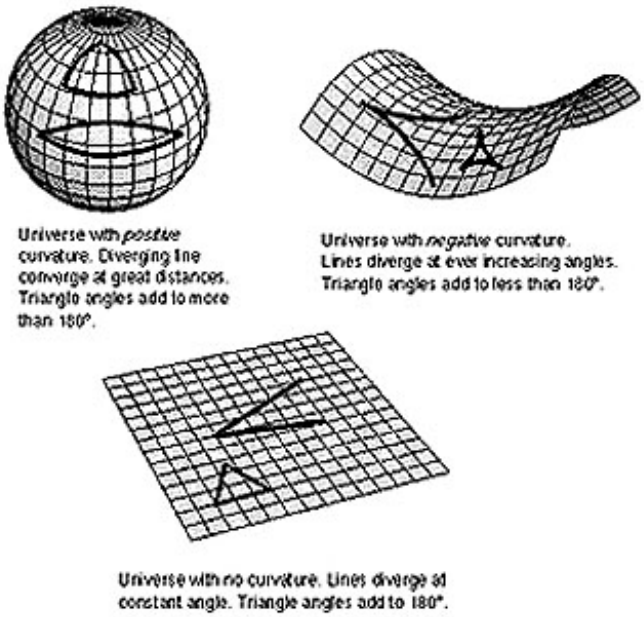
The nature and shape of the Universe depends on its mass density (including energy forms that relate to mass according to the $E = mc^2$ equivalency). The key parameter is the **Critical Density**, symbolized as ρ_{crit}, and is calculated as $\rho_{crit} = 3H^2/8\pi G$. ρ_{crit} is derived from the Friedmann equation; conceptually, it is the amount of matter/energy needed to lead to flat spacetime, in which the Universe is neither open or closed. Using the current value of H as 22 km/sec/Mly, ρ_{crit} is 9.7×10^{-27} kg/m^3. Since Hydrogen is by far the most abundant element in the Universe, ρ_{crit} can be presented as an average of 5.8 atoms of H per cubic meter. ρ_{crit} is postulated to be just that total mass/energy that causes the Universe neither to expand forever nor to collapse on itself, i.e., it is flat and will just stop expansion after infinite cosmic time has elapsed. Thus, a flat Universe is one that expands at the "balanced rate" that permits it to just avoid an eventual collapse.

Ω is a fundamental parameter that is the *ratio of the actual density to the critical density*. In terms of the Friedmann equation, $\Omega = \rho/\rho_{crit} = 1 + kc^2/H^2R^2$; k is a measure of curvature and R is the radius of a Universe considered to be spherical . Ω values have been determined for all the prime constituents of the Universe and have these (approximate values). Ω_{de} (dark energy) = 0. 73; Ω_{dm} (dark matter) = 0.23; Ω_{at} (atomic matter) = 0.044; Ω_{ph} (photons) = 5×10^{-5}; Ω_{nu} (neutrinos) = 3.4×10^{-5}. These will sum to Ω_{tot} (total) = 1. In the more general case, Ω is often expressed as Ω_Λ for all Dark Energy and Ω_M for all matter, both Dark and Ordinary.

In the more general case, Ω is often expressed as Ω_Λ for all Dark Energy and Ω_M for all matter, both Dark and Ordinary. The history of Ω_Λ and Ω_M over cosmic time has been one of progressive increase of the first at the expense of the second. Thus, Ω_M was nearly one (1) at the beginning of the Universe and now has diminished to 0.27 (0.23 for Dark Matter and 0.04 for Ordinary Matter) will continue to decrease in the future as Ω_Λ increases beyond the present 0.73, causing ever increasing expansion.

As a practical measure it is estimated that, if all atomic matter - both galactic and intergalactic - is redistributed to spread uniformly through space, its mass density will average just under 6 atoms per cubic meter. This is notably lower in the parts of space far from galaxies. This near vacuum space will contain mostly Hydrogen atoms. The least populated parts of intergalactic space are "almost totally empty", not quite a true void having about 1 atom per cubic meter, but consists of Hydrogen atoms, traces of other atomic species, ions, infinitesimal amounts of ice and dust (including possible organics), virtual particles and Dark Matter/Energy - all in very low amounts. The intergalactic density distribution also likely varies, becoming higher as galaxies are approached. Individual stars and nebular clouds are found in parts of intergalactic space as local regions of concentrated matter.

There are three general density-controlled shapes available as options for the configuration and expansion of the Universe. Their geometric characteristics are depicted in this next figure. Note that two properties help to define the nature and behavior of each shape: 1) What happens to so-called parallel lines in traversing the shape, and 2) What is the sum of angles in any triangle drawn on the shape?



This figure was copied from Nick Strobel's Astronomy Notes. Go to his site at www.astronomynotes.com for any updated versions.

The spherical shape is said to have no boundary in that one would always remain on its surface and if "walking" along a great circle would always return to the starting point. It has positive curvature. Hyperbolic space is one that has negative curvature: although difficult to visualize, and best described mathematically, descriptively it has been likened to a horse's saddle; this geometry has the peculiar spatial attribute that movement away from the lowest point on it can go either "downhill" or "uphill", depending on direction. Flat space has minimal (zero) curvature and obeys the precepts of Euclidean geometry. Spherical, flat, and hyperbolic space have this defining property in which the sum of the angles in a triangle are greater than, equal to, or less than 180° respectively. In flat space, parallel lines remain parallel in this geometric sense; this provides a means to test the type of Universe geometry that corresponds to reality. This implies that light beams from a distant source do not converge or diverge. So far, evidence is that lines of radiation travelling in space remain parallel unless disturbed by gravity from massive bodies. Both flat and hyperbolic space can extend indefinitely (to infinity) in contrast to spherical space (but, in principle, if it expands continuously forever that too could lead to a kind of infinity).

These three general types of shape can also be depicted in spacetime cone figures, such as this one, showing from left to right the steady, decelerating, and accelerating expansion models:

From the above, one theoretical way to distinguish which shape best describes that of the Universe: send two light beams oriented parallel to each other but separated by some distance. In the flat Universe, these beams will always remain parallel. In the hyperbolic Universe, the beams diverge; in the spherical Universe they will eventually converge and cross each other. (Theoretically, the test could be compromised by the beam being affected by strong gravitational forces, as has been demonstrated for predictions made by Einstein's General

Relativity.)

If the the total density (choosing M as the sum of all matter and all energy; remember that energy can be stated in terms of mass by Einstein's famed equation, restated: $m = E/c^2$) distributed throughout a finite Universe of some size or volume V is less than the Critical Density, space is *hyperbolic* and open; if greater than critical, *spherical* and closed; and if equal to critical, space is *flat* (at least at the scales we observe it). This can also be expressed as the *Density Parameter* or **Ω** which is the ratio of the actual densities ρ of matter and energy present in the Universe to ρ_c, the density that would apply to a Flat Universe expanding to infinity. The Open, Flat, and Closed Universes are associated with $\Omega > 1$, $= 1$, and < 1 respectively.

Each of those models fits at least *some* of the general observations of the Universe but has failed on other accounts. Considering all of these models: If the Universe is open or flat, the Universe will expand infinitely but at different rates depending on the parameters associated with each model. The closed and negative L models, in contrast, predict finite expansion followed by eventual contraction and thus at some time the Universe returns to a singularity state. For each of the models, the expansion geometry and the behavior from the onset (the Steady State model has no "beginning") to its eventual fate (Crunch; Expansion) depends ultimately on the matter density that characterizes it.

As stated earlier on this page, the presently favored model for the Universe's geometry is "flat" (although an Ω_{tot} of 1.02 [see table above] suggests it may be slightly open). This seemingly contradicts the notion of expansion of a sphere. But when the size of the sphere is huge (radius estimates vary but can be greater than 42 billion light years), if one is located at any point "on" the sphere, the curvature is so slight that the "surface" appears flat. The relevance of "flat" is that it describes a geometry that is Euclidean.

The first five models are all non-standard and were devised when Einstein's Cosmological Constant seemed to have some essential validity.

This is an appropriate point at which to elucidate this famed Cosmological Constant. Einstein himself spent many years in calculating properties of various Universes. Einstein favored a static Universe (neither expanding nor contracting). But gravity should cause the Universe to eventually collapse. To offset that effect, Einstein postulated a form of vacuum energy density that would be positive in the sense of countering gravity (in a sense, causing offsetting expansion that just balanced to contraction from gravity). This he named the Cosmological Constant. It is usually represented by the Greek letter Λ and is expressed in this equation, in which ρ is the density term, G is the Universal Gravitational Constant, and c is the speed of light:

$$\Lambda = (8\pi G/3c^2)\rho_\Lambda$$

He introduced Λ to his field equation that described the Universe's status in terms of the Friedman model (see above), yielding this equation:

$$H^2 = \frac{8\pi G\rho}{3c^2} - \frac{k}{R^2} + \frac{\Lambda c^2}{3} \quad \text{where } \Lambda = \text{cosmological constant}$$

When Hubble and others presented ever more convincing evidence that the Universe was actually expanding, Einstein refuted his Cosmological Constant, calling it the "greatest blunder of my life".

In recent years variations of the Cosmological Constant are again becoming fashionable to explain some of the phenomena essential to a changing Universe, as we shall see on the next page which dwells upon an Accerating Universe. Its possible equivalence to the concept of u>vacuum energy density that may be the nature of Dark Energy (next page) has rehabilitated

the idea. It may also have been a key factor in the Inflationary Stage of the early Universe; a rapid increase in L (the Lambda-force) could be the driver behind the tremendous expansion then but that increase had to be short-lived and L must revert towards zero or the Universe would have long since "blown" away.

Values for Λ are hard to come by even in modern textbooks. The current values are associated with vacuum energy density. The ones quoted in the Wikipedia web site on the Constant are "on the order of 10^{-35} s^{-2}, or 10^{-47} GeV, or 10^{-29} g/cm3,[5] or about 10^{-120} in reduced Planck units."

The Einstein Universe is a static one, with spherical geometry. It was put forth by this great scientist as an attempt to apply General Relativity to Cosmology. The idea of the Big Bang had not yet captured the attention of cosmo-scientists. In order to keep the Universe "going" instead of collapsing under its own gravity, Einstein invented his Cosmological Constant L to balance the attractive forces. While now considered notably incorrect, this type of Steady State model stimulated others to propose variants that incorporate expansion. The de Sitter Universe is a strange one, being empty and never undergoing a Big Bang. Its value of q being negative (-q) denotes an accelerating Universe. But in working back towards time zero, its representative R(t) value never attains zero, which means that it has no beginning, i.e., has an infinite past. While theoretically interesting, the model defies most observational parameters, and the very notion of the Big Bang concept. The Steady State Universe was formulated by Hoyle and others as an "antidote" to the Big Bang model. It accepts expansion and implies that the Universe has no beginning or end. In order to preserve the matter density distribution determined for the Universe, the Steady State model requires a "creation field" in which new matter (mass) must be continuously created through time to balance the rate of expansion. Another model (not in above Table) also does not start with a Big Bang; this, the Eddington-Lemaitre model, is closed and finite and is static initially but thereafter starts expanding when the galaxies begin to form by Hydrogen gas condensation.

The Lemaitre model, derived from the Big Bang concepts, begins with a rapid increase in R during the early Universe but then experiences an extended period when R(t) remains nearly constant (owing to the effect of L being greater than L_c) so that expansion is minimal ("hovers") until much later resuming at an accelerated rate (read the modern version of this resumption of acceleration on the next page). The Abbe George Lemaitre (a Catholic priest from Belgium who also was a physicist) was the first to consider the starting state to be one of extremely high (approaching infinity) density (he called the singularity a "primeval atom"; later, George Gamow applied the Greek word "ylem" [primitive matter] to everything contained in such a singularity).

Among the three standard hot (high temperature) Big Bang models, the *Open Universe* model (also known as the Friedmann-Lemaitre model) predicts that expansion continues forever at an essentially constant rate through an infinite and unbounded space based on hyperbolic geometry (in which light can follow both positive and negative curvature simultaneously). Evidence so far suggests that a (nearly) *flat* Universe model, whose density is at the critical density (in which Ω = 1, the condition that there is just enough matter distributed throughout the Universe to cause it to expand forever even as it endlessly slows down) accounts for many of its observed properties, so that the *Einstein-DeSitter Universe* is currently the model most widely held to approximate reality. This model is in accord with current estimates for the age of the Universe.

The mental picture one gets from the word "Flat" as we have been applying it to cosmic expansion may be somewhat illusory. One meaning - just as the surface of a large balloon may appear flat to an ant at some point on it, so the Universe may in fact be spherical but acts as though flat within the region open to our direct observation (we experience this on Earth as our local surroundings appear flat out to the horizon but would show its real curvature if we were orbiting astronauts). However, flat on a Universe scale may mean just that - flat in our

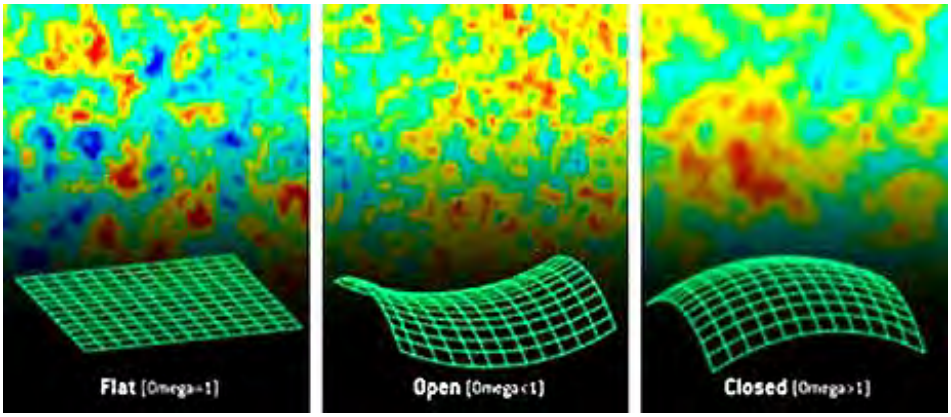
experiential Euclidian sense - imagine a table top that keeps expanding forever from within itself (not just by growth at the edges) in two primary directions; points at different parts of this infinitely growing top would all separate from each other. Table tops do have a third dimension (thickness), as presumably does a flat Universe, but expansion in that dimension may be finite.

Closed Universes follow spherical geometries. The prime model shows greater rates of expansion in early cosmic time with decreasing rates of augmentation thereafter. (This is not the same as the incredible but brief expansion almost at the very beginning of the Universe if inflation indeed is a real phenomenon.) Thus, the components of the Universe move outward powered in part by the inertia imparted by the energy release at the Big Bang. However, the mass/energy level is high enough for gravity to effectively pull on galaxies, stars, and other matter so as to gradually slow the expansion to a zero rate. Thereafter, the condition becomes one of increasing deceleration. The rate of separation between galaxies diminishes with time until, at some future time, expansion ceases and galaxies then draw closer at ever faster rates until all matter and radiation converge to a singularity (perhaps 50 b.y. in the future), undergoing what has been called the Big Crunch.

The Crunch concept remains intriguing. A group at Penn State University has carried out computer calculations that strive to combine General Relativity and Quantum Physics to derive a Quantum Gravitational model of a contracting Universe. This results in matter and energy in that previous Universe coming together in the singularity we know as the Big Bang. As contraction proceeds to its last moments, at its terminal stage quantum forces become repulsive, causing a quantum ***Big Bounce*** that seems similar in its properties to a Big Bang. Their model retains a consistency that does not rule out this possibility but the difficulty of trying to find confirmative evidence for this earlier Universe is only addressed in terms of the model being plausible without any direct proof.

This raises the possibility of *Repeated Universes*, as singularities explode, expand, ultimately contract to the next singularity, and then repeat the cycle indefinitely, or, even infinitely. However, this scenario seemingly would violate the entropy restriction in that the singularity should have a minimum rather than maximum state of disorder that is the outcome for every model. Multiple Universes (next page) that evolve simultaneously, or at different "times", are a possible consequence of the *Chaotic Inflationary model* which in recent years has gained favor as a variant of the inflationary version of the Big Bang. These Universes, however, have no likelihood of contact with one another, so that their existences may be unprovable.

To sum up this topic - the shape of the Universe. Evidence is building that the Flat condition is the most likely to describe this configuration. However, this "flatness" may be illusory: the Universe may be so large that what is viewed and perceived as "flat" is actually part of curved space - as viewed the curvature is too slight to be determined (imagine being on a small part of a very large balloon). Results from COBE and WMAP seem to offer solid support for the idea of real or apparent "flatness". Consider this diagram:



The closest fit of size variations of cosmic background radiation fluctuations, as determined by calculations, to the observed COBE and WMAP data is that of a Flat Universe.

Nevertheless, whether the present Universe is open, closed or flat is still being debated, despite the ever-stronger support for a Flat Universe. The key factor is the mass density of the Universe. If this amount lies below a precise value, then gravitational forces will be insufficient to finally halt the expansion that would eventually result in all matter throughout space being pulled back into closure; in the open case, expansion is forever in (apparently) the one and only Universe. As of now, inventories of mass within the Universe have come up way short of the amount actually identified as to their nature needed to maintain a closed Universe but with further observation and experimentation the gap is narrowing. If the interpretation of WMAP data is essentially "on target", then the condition that the total Universe mass is close to the Critical Mass allows for a fairly precise estimate of how much mass must be found and accounted for.

The bulk of the missing mass (and energy particles that have mass) is believed to exist as **Dark** (non-luminous, i.e., does not give off [detectable] electromagnetic radiation) Matter of still uncertain types and/or as neutrinos, and thus remains at present "invisible" to astronomers' detectors. Theory presently holds that most proposed kinds of Dark Matter do not interact with baryonic (Ordinary) Matter, although some Dark Matter may be baryonic. Dark Matter is currently difficult to detect and thus its amount is hard to quantify. Dark Energy, a closely related topic, is mentioned briefly in this subsection but will be examined in some detail on the next page. Suffice for now to state that the main role of Dark Energy seems to be that it has some property the counteracts the effect of gravity in slowing down or decelerating the Universe's expansion. Dark Energy thus produces an accelerating force which is the best current explanation for the recent observation that the Universe is now expanding or speeding uup(and appears to have resumed a net expansion in the last few billion years).

Some of the Dark Matter seems to associate with the galaxies. Theory says that Dark Matter plays a key role in galactic development and stability. The galaxies themselves contain insufficient mass to provide the gravitational strength needed to hold them together. Since they are intact, it is a logical inference to assume the existence of Dark Matter which supplies the gravitational forces that maintain them. A computer simulation reported in late 2003 indicates **the bulk of Dark Matter occurs in large "clumps" distributed throughout intergalactic space**; see further discussion below. Black holes may serve as collectors or maintainers of this Dark Matter.

A recent estimate states outright that Dark Matter and Dark Energy together comprise ~95% of the Universe's mass (remember Einstein's mass-energy equivalency), with varieties of normal or baryonic matter accounting for the remainder are thus less than 5%. This means that most of the matter in the Universe presently is <u>invisible</u> to detection from Earth (the billions of luminous galaxies, each with billions of stars therefore make up only a tiny fraction of the Universe's total mass).

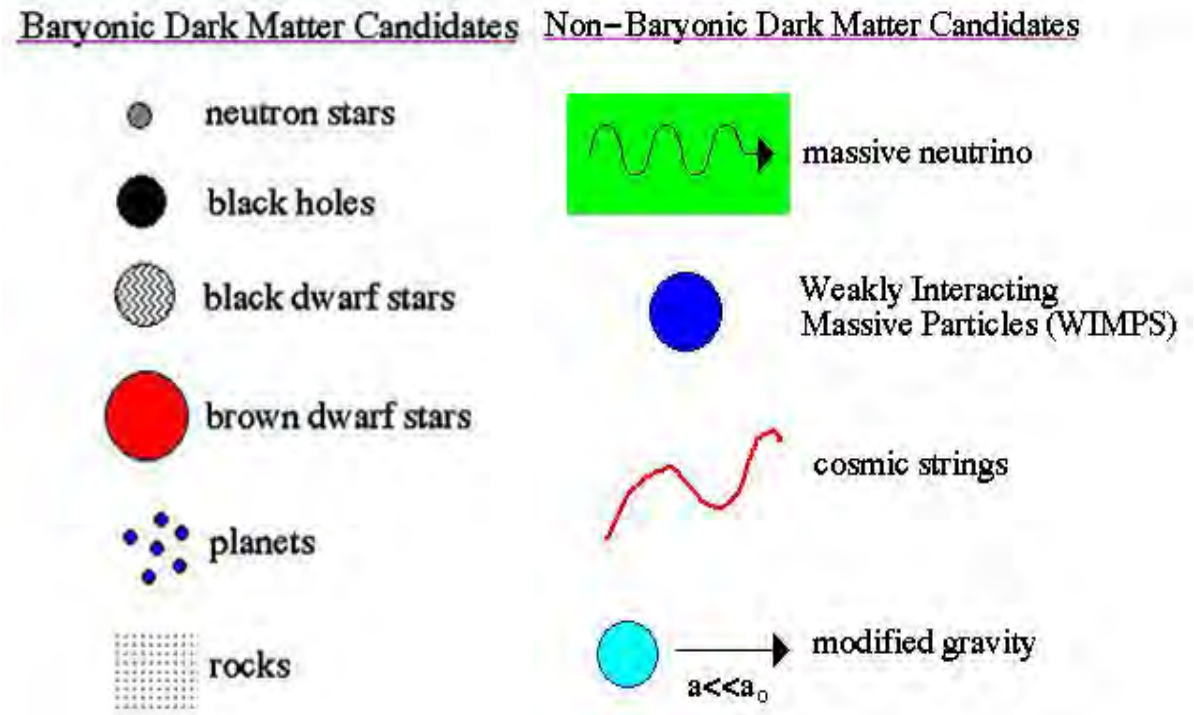
This humongous amount of Dark Matter is postulated by inferences drawn from observed effects; indirect evidence comes from the behavior of galaxies and galactic clusters which seem to need this superabundance of mass to account for their stabilities and motions. For example, the velocities of stars in outer galactic spiral arms is much greater than predicted from the Newtonian $1/r^2$ force law, implying excess external mass. Thus a sheath of invisible Dark Matter/Energy acts to prevent galaxies of spinning apart by holding the fast-moving outer stars in a gravitational bind.

Another line of evidence comes from the gravitational lens effect (see text and illustrations in the Preface page of this Section) - much more mass than observable is needed to account for the degree of bending of space as predicted by General Relativity. This is revealed by a greater curving of light from more distant light sources (therefore displaying larger displacements than expected) than would be caused only by the mass of the specific galactic cluster whose gravitational influence is being tested. Another sign of concentrations of hidden mass relates to

directional movement of galaxies near enough to observe and measure this motion. Close to home, our Local Group (including the Milky Way) of galaxies is moving through space in the direction of the Constellation Centaurus at greater than expected velocities, under the influence of an invisible mass concentration dubbed "The Great Attractor."(by itself, the Milky Way is moving at about 2.1 million km/hr towards the region around the Constellation Leo). Also, huge masses of glowing gas whose molecules are rapidly moving and hence indicate very high temperatures have been detected; being very hot, they should fly apart but clearly are holding together, indicating the attractive action of great quantities of invisible mass.

There is evidence that older, more primitive stars (that have a paucity of those heavier elements that were produced in stars and dispersed by supernova explosions) contain around them higher concentrations of Dark Matter. But, much (most?) of the missing mass may be tied up in Black Holes; billions probably exist throughout the Universe, and many, if not most, of the galaxies have Black Holes in their central cores. A Black Hole with immense mass has been verified at the center of the Milky Way, around which the inner stars revolve about the Hole at speeds up to 3 million km/hr as they spiral inward to eventually be sucked in (by comparison, the Sun orbits around the galactic center at ~790,000 km/hr and the Earth around the Sun at ~108,000 km/hr).

The composition of Dark Matter is still speculative. Candidates are shown in these diagrams, modified from that which appears in the page dealing with the nature of the Dark Universe found on the University of Oregon Astronomy site we have referenced several times in this Section:



Some of the baryonic Dark Matter occurs in what is called MACHOs (for MAssively Compact Halo Objects), consisting of baryons (protons, neutrons) and other matter (probably some fraction of the neutrinos pervading space) in the non-radiating dark halos now known to distribute around galaxies and in intergalactic space (see below). In the halos they constitute most of the faster moving Hot Dark Matter (HDM). MACHOs contain enough extra mass to provide the gravitational boost that holds galaxies together (motion in the spiral disk would otherwise cause a galaxy to fly apart). Dwarf galaxies and Wwhite and Black Dwarf stellar remnants - too small for ready detection in more distant space - may also abound in this material. Black Holes, Neutron stars, and Dwarfs, undetectable by visual means but identified by their gravitational effects on nearby visible stars, also make major contributions. In fact, the same (as yet undetected) material as occurs in MACHO's may also make up planet-sized Black Holes, whose numbers in the Universe can be huge; this is not yet verifiable if a B.H. is "standing alone", i.e., does not have a companion star(s) feeding it material that becomes excited and luminous.

One MACHO type is the neutrino whose existence has now been proved. Once thought to be massless, a tiny but real neutrino mass has only recently been verified - a determination that could account for much (most?) of the "missing" Universe mass. Neutrinos (electrically neutral) are found not only in haloes but appear to pervade the "empty space" of the Universe. An announcement in June 1998 by a Japanese-American research team may dramatically change the role of neutrinos in the mass balance sheet. Using a deeply buried detector and a huge array of detectors, they have been able to capture light signals set off by neutrinos that disclose these extremely small particles to have an infinitesimally minute mass, about one ten-millionth that of the electron. But, because of the extremely large population of neutrinos throughout the Universe (like the Cosmic Background Radiation, they too are a residue of the Big Bang, being especially produced in abundance during the first few minutes when protons were being fused into Deuterium and Helium nucleii releasing energy as neutrinos), the cumulative mass of these particles may make a sizeable contribution to the inventory of missing mass. But other theoretic MACHO(s) that remain to be discovered (either from future particle accelerator experiments or from still-to-be-built, more sensitive space detectors) will likely constitute the bulk of the MACHO population.

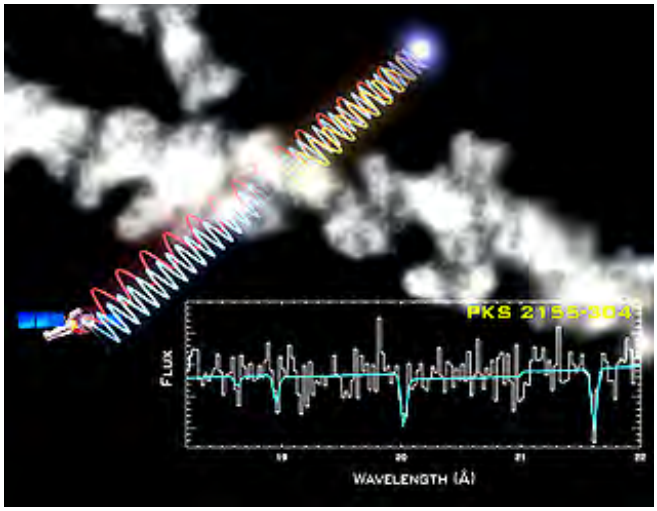
Concerning neutrinos as a component of Dark Matter, until recently there was still considerable uncertainty as to whether neutrinos (there are several varieties) have any mass at all. They rarely interact with matter and are thus very hard to detect. At this instant, billions of neutrinos are passing through your body, and perhaps one or two at most will meet with an atom. Experiments in several countries are attempting to find out more aboutthe elusive neutrino. Below is one example of an experimental detection setup. Several thousand feet below the surface, in a large alcove in a nickel mine near Sudbury, Ontario (the overlying rock screens out most other high energy particles that might produce false signals), is a 12 meter diameter sphere containing heavy water (deuterium instead of Hydrogen). Around it are a bevy of event detectors. A few events have been ascertained, but the number of solar neutrinos appears lower than theory had predicted.

If the neutrinos are finally assigned a specific (but tiny) mass, they must be a major component of Dark Matter. The estimated ratio of Dark Matter neutrinos to Ordinary Matter neutrinos is 100,000,000 to 1.

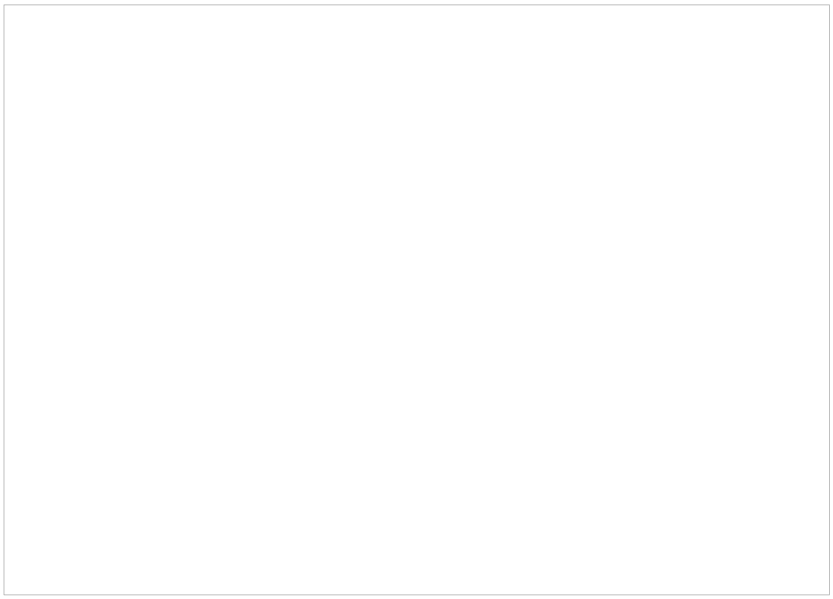
A significant fraction (perhaps the bulk) of the missing matter, however, seems to occur in Cold Dark Matter (CDM), probably in the form of slow-moving WIMPs (Weakly Interacting Massive Particles) - elusive matter/energy that does not interact with electromagnetic (EM) or strong nuclear forces. These appear to be heavy nonbaryonic particles (50 to 100 times the mass of a proton) of still unknown nature. They may also include two specific particles that have been

postulated but not yet found experimentally: the *neutralinos* and the *axions* - both predicted to exist according to SuperSymmetry theory. Several high energy accelerators are either now operational or are being built (the best and biggest is CERN's Large Hadron Collider [LHC] in Switzerland) to find direct evidence of the existence of WIMPs or some similar form of matter.

Recent reports from several investigator teams claim that much of the Dark Matter is accounted for by the very hot gases that are believed to pervade the intergalactic space that appears "empty" compared with galaxies and nebular clusters of visible gases. A clue to the presence of these gases (undetected at Visible and longer wavelengths) was found in UV spectrum data but the signal was weak. However, when Chandra data for X-ray radiation was collected from intergalactic regions, indications were for very hot gases in signficant concentrations in the so-called void. An estimate of gas densities was obtained by calculating the weakening of certain wavelengths as radiation from a gas-rich region passes through a nebular mass. This is shown diagramatically:



The relative amounts of Dark Matter and Dark Energy have been calculated based on data collected in the past few years. This is conveniently displayed in this pie-chart diagram, made in the mid-1990s, that includes all physical entities:



In this diagram, Ordinary (detectable) "Matter", mainly in the star-studded galaxies, amounts to only 0.5%. Photon radiation moving about the Universe contributes just 0.005%. The MACHOS make up the Ordinary (including nonluminous) Matter (4%) and WIMPS comprise the Exotic Dark Matter (25%). The bulk of the ingredients in the Universe is what is known as "Dark Energy", and accounts for ~70% of the estimated total mass/energy; conjectures as to its nature are described on the next page in the paragraphs that deal with the repulsive energy that could

relate to Einstein's Cosmological Constant.

The WMAP data have resulted in a further refinement in the percentages of Dark Matter and Dark Energy. This diagram shows a shift to a 3% larger amount of Dark Energy with a 2% decrease in Dark Matter:

In February 2003, F. Nicastro and his associates at the Center for Astrophysics of Harvard's Smithsonian Astronomical Observatory reported a more precise calculation of the numbers for Ordinary Matter. Of the 4% that comprises the total amount, about 0.4% is luminous (in the visible range) as the stars making up galaxies. The remainder consists of H, He, and other Baryons that occupy both galactic halos and beyond in intergalactic space. These exist in a high temperature (10^5 to 10^7 degrees Kelvin) "fog" that they have detected using both FUSE (Far Ultraviolet Spectrometer Explorer) and Chandra (X-ray) data. This fog is left over from the galactic formation process and serves to provide the extra mass needed to gravitationally bind together galactic groups (they studied the Local Group around the Milky Way) in clusters.

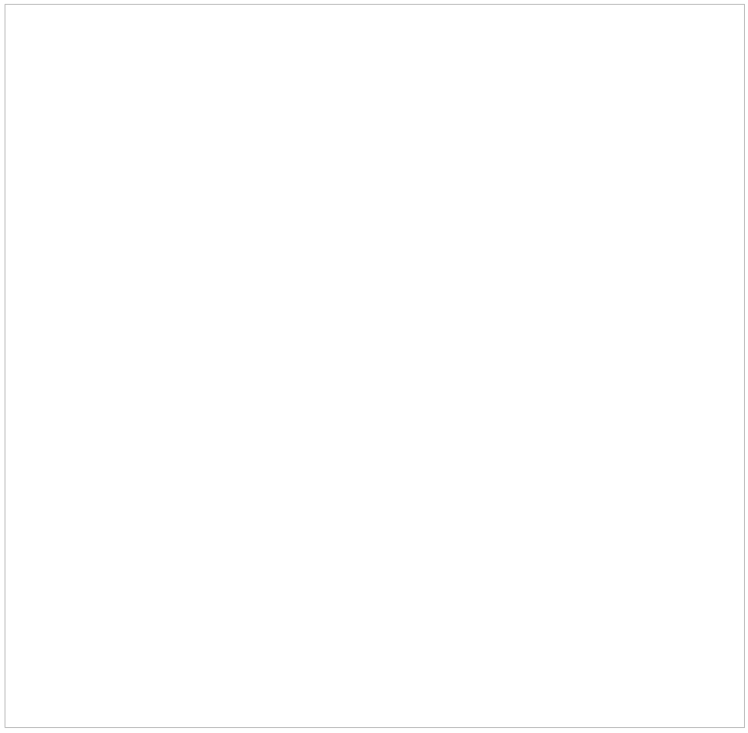
The roles of Dark Matter and Dark Energy in the observed Universe are still being worked out. Both, or perhaps Dark Matter in particular, seem to be needed to keep the galaxies intact and to prevent them from collapsing into each other. It appears that the gravitational forces from the matter within a galaxy are insufficient to keep them from flying apart and dissipating into intergalactic space. Dark Matter should provide the mass needed to furnish the gravitational stability that maintains the integrity of the galaxies once formed. In fact, several models for its existence and nature consider **the bulk of Dark Matter to occupy large invisible halos surrounding the galaxies, serving as the major gravitational "glue" that holds them together.** And, as will be discussed on the next page, **Dark Energy, as inferred by theorists, is a positive force that is responsible for the increasing expansion of the Universe (thus, it counteracts gravity)**

An imaginative explanation for why Dark Matter (in terms of mass) has so far eluded scientists as to proving its existence and nature has been proffered by Dr. Jonathan Feng and his group at the Univesity of California-Irvine. They postulate this matter to be hidden particles that reside in an extra dimension beyond the three spatial ones that we sense. This is plausible if the multi-dimensional space described at the bottom of page 20-1 is a reality. This model may be provable by this means: the extra-dimensional mass particles would tend to collect around bodies that have strong gravitational pulls. Under this influence, the particles would tend to collide with each other more often, producing neutrinos with especially high energy. Experiments to test this hypothesis have been proposed.

One test for the existence of Dark Matter (and probably also Dark Energy) is that of the (Einstein) **gravitational lensing** effect (bending of light as it passes by a massive galaxy). If only the luminous matter in the galaxy, supplemented by assumed mass in contained black holes, is operative, the amount of deflection should be less than observed. In fact, that deflection

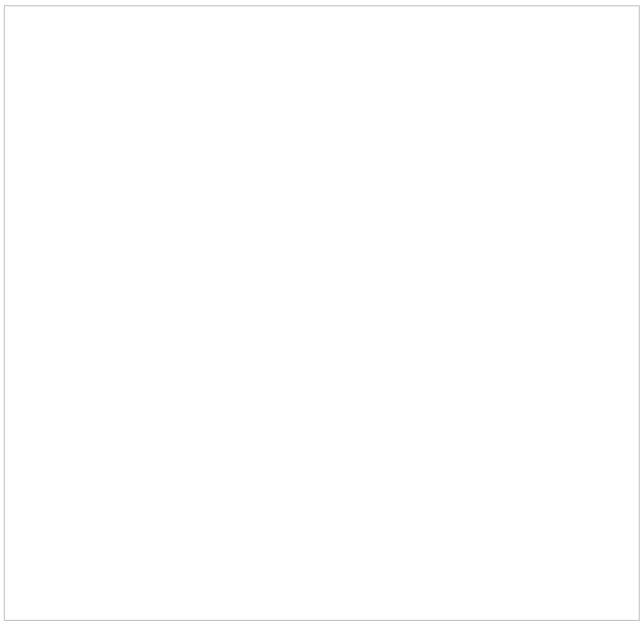
is greater than this simpler case, and thus has been attributed to at least MACHOS and/or some amount of WIMPS that are emplaced in and around the galaxy. Using gravitational lensing techniques, the size of the Dark Matter distribution concentrated around galaxies has been estimated to be in a roughly spherical arrangement (the 'halo') up to about 5 times the radius of each galaxy examined so far.

Gravitational lensing has been used to create this image of a group of galaxies, in which the pinkish halo around each is the estimate of the size and position of Dark Matter holding each intact:



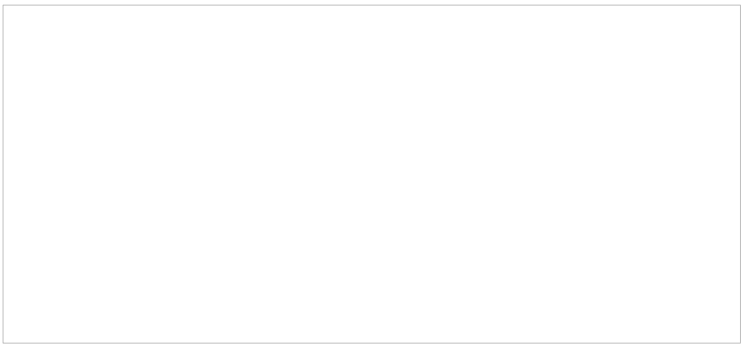
Incontrovertible proof that Dark Matter and Dark Energy really exist, and insights into their nature, rank near the top of priorities that astronomers and cosmologists intend to address in the first decade of the 21st Century. A number of approaches besides gravitational lensing have been proposed. One is to compare the size of a galaxy as seen in visible light to the size of its associated hot matter (the "fog") that gives off strong X-ray radiation. Look at this pair of images of a galaxy in the Virgo galactic cluster:

Another, oft-cited, example of the X-ray envelope is the NGC 2300 galaxy group, consisting of a number of individual galaxies as seen in Visble light that do not show up as individuals when X-radiation (measured by ROSAT) is assigned colors to indicate its extent.

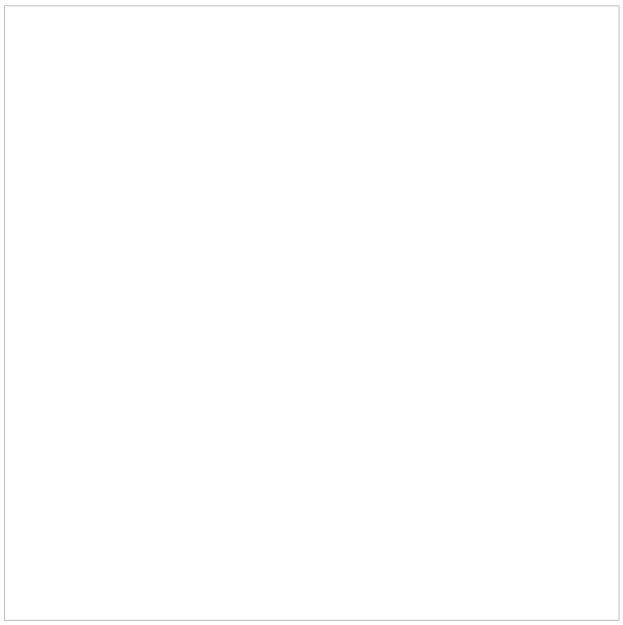


How do these X-ray images indicate the presence (existence) of Dark Matter? The argument is that the galaxies would have flown apart by now but are held together by the mass associated with the Dark Matter that encloses them. Consult the University of Oregon page cited above for illustrations that define this process.

A variant to the size argument is to study the shapes of galaxies that have characteristics explainable by the presence of Dark Matter. MGC 720, some 80 million light years away, shows differences between Visible and X-ray images that can be postulated as imposed by Dark Matter in the halo surrounding the galaxy. As seen below, the galaxy has a flattened elliptical shape when viewed in Visible light but as imaged by the Chandra X-ray Telescope the core appears round and the surrounding excited matter is distributed spherically. A calculated model for this difference works well when a concentration of Dark Matter is introduced. There is a hypothesis that galaxies will attract greater amounts of Dark Matter surrounding them than would be found in intergalactic space.



In June, 2003 another report about the action of Dark Matter gave a strong indication of just how much mass is involved. Below is a Chandra view of the galaxy cluster Abell 2029, located about 1 billion light years from Earth. This X-ray image shows a central radiating mass (an elliptical supergalaxy that resulted from merger of multiple galaxies) and a huge cloud of glowing hot gas that is interpreted as under direct control by this Dark Matter, which is estimated to be equivalent to a hundred trillion times the mass of the Sun.

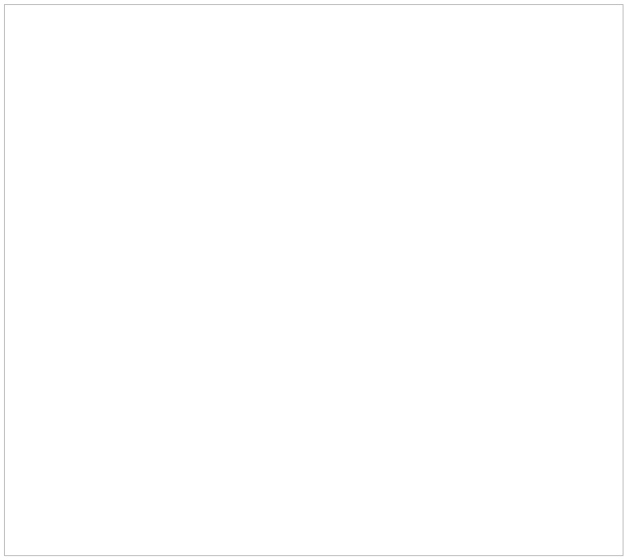


In July of 2003 another image was released which may actually show the Dark Matter as a faint glow (arbitrarily rendered by a blue tint) pervading a cluster of galaxies:

To obtain this image, the HST was trained for a total of 120 hour of light-collection time on the cluster CL0024 which presently is some 4.5 billion light years away. Spread through the galactic group is the uniform glow which in itself has not been identified as to nature. But its distribution is much as expected for Dark Matter based on theory. If this glow indeed is really a property of Dark Matter, then it produces some luminosity, i.e., emits EM radiation that is normally so weak as to be undetected in shorter exposures of visible wavelength observations from both space and Earth telescopes.

Another similar image of two galaxy clusters in HH47 (in the Vega constellation) was released in December 2005 by a Johns Hopkins group. These astronomers used an inventory of stars that bend light to determine that there must be a surplus of Dark Matter (shown in purple) that is needed to allow and maintain the clustering involved. Their interpretion of this matter is that it in the "strange" state wherein individual particles do not collide as they surround the galaxies. The implication is that Dark Matter itself has an uneven distribution in the Universe, tending to be more dense around galaxies and galaxy groups. Here is a defining image published in their report:

Another line of evidence for the presence and role of Dark Matter is associated with NGC1404, the Formax galaxy cluster, shown below. The Chandra X-ray image shows galactic gases in blue, heated to as high as 10 million degrees Celsius. The gas clouds are moving at high speeds toward an (invisible) center. The presumption of a large, unseen mass of Dark Matter around the cluster gives the best (theoretically) explanation of the behavior of what has been detected. This cluster may lie at the intersection of two filaments of stretched out Dark Matter; that speculation is in line with the ideas of a network of filaments in the early Universe described on page 20-2.



Recently, studies of Dark Matter in the Ultraviolet suggest that one of its constituents appears to be ionized Hydrogen which does not produce a detectable signal from the myriads of Hydrogen clouds now known to populate galactic haloes and intergalactic space. But, chemical processes in these clouds produce ionized Oxygen, uncombined with the invisible ionized Hydrogen, which can be detected. This form of Oxygen is widespread in much of space beyond the galaxies and, together, with the ionized Hydrogen may account for a significant fraction of the daark matter.

Another potential approach to studying Dark Matter, including an attempt to prove its existence as commonplace within the Universe (rather than just a fleeting existence during collider experiments on Earth), is now underway with the installation of a 2.8 meter telescope, called the South Pole Telescope, located close to that point in the Southern Hemisphere. It will be capable of measuring small changes in conditions (primarily temperatures) around galaxy clusters; these changes would verify expected variations caused by Dark Matter/Energy as predicted by models now being developed.

From the above discussion, one might ask "Where is most of the Dark Matter?". Since it has yet

to be found directly, the answer may lie in these possibilities: 1) in Black Holes; 2) in non-luminous galaxies (evidence for a "Dark Galaxy" has recently been reported); 3) present but "invisible" within galaxies; 4) in roughly spherical "halos" around each galaxy; 5) dispersed throughout intergalactic space.

An interesting alternative to Dark Matter as the stabilizing force is that of a concept called MOND, for Modified Newtonian Dynamics. This is reviewed by its originator, Dr. Mordecai Milgrom of the Weizmann Institute in Israel, in his article "Does Dark Matter Really Exist?" in the August 2002 issue of Scientific American or the October 2003 issue of Discover Magazine. Check this to be familiarized with its essentials. His ideas are derived from observations that such spacecraft as Pioneer 10, now 9 billion miles from Earth, are slowing down (the expectation would be for increasing velocities as the Sun's gravity weakens with distance). This would seem to imply that solar gravity is actually increasing as the distance lengthens.

In a nutshell, Milgrom postulates a constant of acceleration, called a_0, that determines the gravitational behavior of matter. Acceleration values larger than a_0 are consistent with the Newtonian Second Law in which F (force) = m (mass) times $\underline{a}$ (acceleration such as we observe experimentally for gravitational processes or for imposed accelerations [e.g., a rocket launch]). (Recall in the Preface that Newton's Second Law works well on Earth and in the Solar System but is not accurate when galactic scales are involved.) But in his proposal values smaller than a_0 modify the behavior of galaxies, including the rates at which they rotate, in a way that obviates the need for Dark Matter's contribution to galactic gravity. Using this non-Newtonian approach, Milgrom's mathematical analysis of galaxy dynamics based on values of $\underline{a}$'s that are less than a_0 reproduces most of the observed effects in the orbital patterns and rates of motion of stars within galaxies. He has tentatively determined that a_0 decreases the Newtonian proportionality by one 10-billionth of a meter per second per second. This, rather than Dark Matter, causes stars in a galaxy to move faster than expected. Milgrom's model, initially disregarded by most cosmologists, is now attracting serious attention.

An excellent, clearly written article by Martin Rees and Priyamvada Nataranjan, entitled *A Field Guide to the Invisible Universe* appears in the December, 2003 issue of Discover Magazine

As this page has been intermittently added to with new information, we must close our discussion of Dark Matter with two more ideas. One simply contends that "some" of the Dark Matter is Ordinary Matter lying beyond the limits of the observed Universe. This matter then has its own gravitational influence on that Universe as though it were actually within the observed part. Interesting but not readily testable.

The second new idea relates to the concepts of Higgs fields and particles (Higgs Bosons) mentioned mainly in footnote 4 on page 20-1. Drawing upon interrelations between particles as treated by the SuperSymmetry model and Higgs fields, theory states that there may exist something called the LSP (for lightest superpartner). The LSP is the decay end product of a series of superpartners (those that can be postulated as complementary to ordinary Standard Model particles) created in the early stages of the Universe. The LSP contains Higgs-induced mass. If real, LSPs can account for much of the Dark Matter.

In August of 2006 a claim was made that direct evidence for action by Dark Matter had been found. Astronomers from the University of Arizona and elsewhere, examining the Bullet galaxy cluster using Chandra and other telescopes, were able to determine the mass within the cluster. They observed two kinds of intergalactic material of gaseous/particulate nature that had separated within the cluster. Thus:

Their interpretation assumed that within the cluster a smaller galaxy had passed through a larger galaxy (the stars are normally so far apart that few collisions occurred). The larger galaxy should have stripped off matter from the smaller, leading to some segregation. Most of that matter ought to reside in the interstellar gas that provides the bulk of the gravitational forces within the cluster. But they argue that this degree of gravity would be insufficiently strong to keep the galaxies intact during the process. They hypothesized that Dark Matter was needed to preserve the separated gases. If indeed Dark Matter was present in the amount they calculated, then it would bend starlight to a specific extent in the manner which allowed Einstein to prove Relativity. This was observed in about that level. The argument then concluded that Dark Matter rather than the gas was the factor that led to this set of observations. While the proposal is tantallizing, as usual other astronomers have come up with objections and alternate explanations. One such alternative considers gravity to vary in strength over cosmic distances.

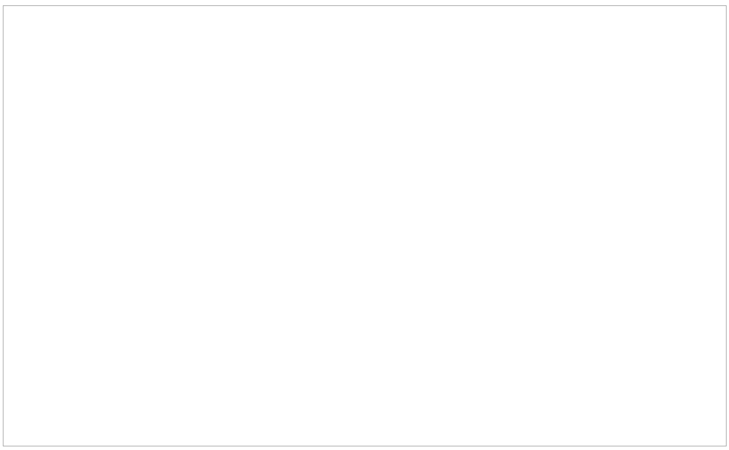
An even better example of direct imaging of dark matter is this view of two colliding galaxies (official identity: MAGSJ0025). The blue denotes dark matter, as imaged by HST; the red is a Chandra X-ray image of gases:

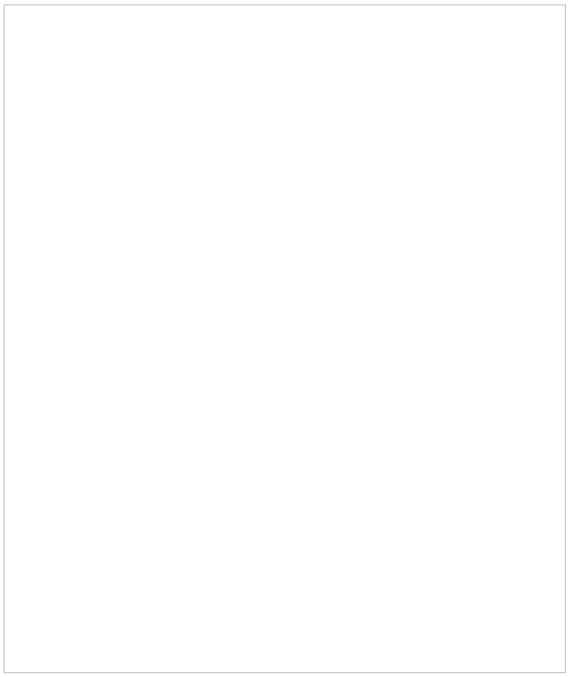
Dark matter (blue) in MAGSJ0025.

The importance of the Dark Matter concept is that it appears to provide just enough total matter to provide the gravity that holds the Universe together. Those who favor the "Goldilocks Hypothesis" for the Universe's formation - not too hot, not too cold, just right" - argue that too much matter (and/or too little velocity of expansion) would have led to gravitational pull strong enough to counteract the expansion of the Universe, causing it to collapse on itself; too little matter (too much velocity) would have prevented matter from organizing into stars and galaxies. Such results would have been a Universe that did exist but never attained the right conditions for the appearance of biological beings and ultimately humans (and conceivably intelligence elsewhere; next page).

Dark Matter seems essential to the formation of stars and galaxies as well. There is insufficient Ordinary Matter to organize and maintain galaxies. Pockets of Dark Matter - whose distribution appears to be related to the ripples or inhomogeneities found by COBE and WMAP - were needed to assist in bringing clots of matter together to start galaxies. Today's galaxies seem held together by higher densities of Dark Matter that comprises the halo around both spiral and elliptical types; this matter also influences the rotational velocities within a spiral galaxy, such that its outer parts move at about the same rates as the inner parts.

A painstaking study by an international group of astronomers (NASA, ESA, CalTech) has produced the first 3-D depiction of Dark Matter in a selected region of the Universe. Using Hubble Space Telescope data, they detected this matter indirectly, through its effect on passing light. This yields information on concentrations of Dark Matter at various distances from Earth, using the standard candle approach to determining just how far away is each clump of Dark Matter. In this way they could arrive at how the matter was distributed at different times of Universe history. These two diagrams show aspects of their study:





Now, on to Dark Energy *per se*, which comprises the bulk of the physical entities that make up the Universe.

---

*The steeper sloped H case represents a Universe that is more rapidly expanding, has covered less distance, and is younger (10 billion years); the slope for H = 50 has expanded more slowly over a longer time (20 billion years) but has covered a greater distance (bigger Universe).

Primary Author: Nicholas M. Short, Sr.

1-1: The two landmarks that best tie the TM and MSS subscenes together are the river, with its distinctive bends, and an arrowhead-shaped mountain (Catawissa Mountain) south of the river. Approximately 20% of the MSS subscene (and less than 10% of the full scene) coincides with the TM subscene. What TM shows alone is some area north of the river. **BACK**

1-2: Farming! Farms occupy most of the land in the valleys. The chief crops are corn, soybeans, some wheat, oats, hay (alfalfa and timothy), pumpkins and squash. In the June Landsat TM subscene, most of the barren fields (light buff) are really in crop: corn, the mainstay of the region but being planted late still only about a foot or less high and hence not contributing much to the field signature. Two typical views of the countryside north of Bloomsburg, with a mix of farmlands and woodlands, appear below (these areas host a number of Amish farms).

BACK

1-3: Although the image gives little sense of verticality or relief, this is the Catawissa Mountain. Geologically it is a syncline. Its peak elevation is 641 m (1940 ft), making it the highest mountain in the TM subscene we are studying. It rises some 1400 ft above the Susquehanna River and about 900 ft above the flatlands just to its north. It is covered largely by deciduous trees, with scatterings of hemlocks. Here is a view of the mountain from about 10 km (6 miles) to the west:

BACK

1-4: You probably selected lumbering, that is, cutting down the trees for wood suited to various uses. But, there is no sign of any extensive clearcutting (see later question). Forest harvesting is a minor use, although locals chop down trees for wood-buring stoves. A few small companies take down certain tree species selectively in the woods to sell for pulp but this does not change the tree density enough to affect the vegetation signature. Lumbering is more active in the Appalachian Plateau to the north of this subscene, up to the New York state border. The major use in the Bloomsburg area is for hunting: this is a significant preoccupation with many of Pennsylvania's citizens - a tradition more than 200 years old. Deer, turkey, and ruffed grouse (the state bird) are the principal animals sought during hunting season. Much of the wooded areas are not only state forests but are part of the widespread State Game Lands system that is especially widspread in the northern half of the State. One popular hunting ground is Knob Mountain near Orangeville about 12 km (8 miles) northeast of Bloomsburg (long green "prong" in upper right of the full subscene; this is an anticlinal mountain).

BACK

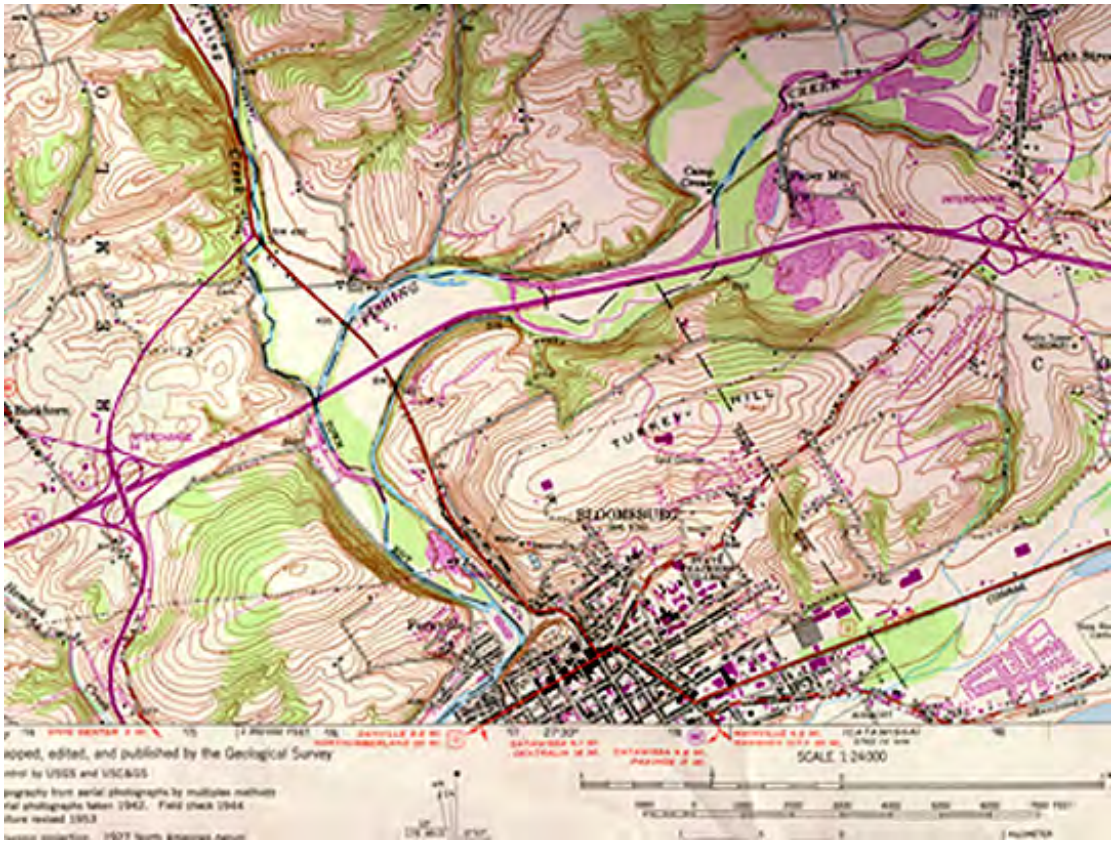
1-5:In the TM subscene, the area in question is a light tan. This is exposed soil or dirt. In 1987 this land was being cleared. By mid-1988, the aerial photo shows a very large building on this land, surrounded by a black top, typical of a parking lot. This is now the Columbia Mall, which opened for business in 1989. Note that there is a cloud just to the west in the TM subscene.**BACK**

1-6:There is a second cloverleaf where I-80 is crossed by State Route 487. Just to its north, about 1/2 mile is an elongate lake visible in the TM subscene and more so in the aerial photo. The photo shows a dam on its west (left) end. This is an artificial lake used as a water source and for fishing, mainly by the residents of the village of Lightstreet just to its east. In the aerial

photo there is also a light blue area less than a mile west. This is a sand excavation pit in which water has now formed a shallow lake.**BACK**

1-7: Bloomsburg lies astride a broad ridge known as Turkey Hill (I actually saw turkeys on it once!), an anticlinal structure. The Interstate runs through the small valley carved by Fishing Creek. Where Fishing Creek turns south along Route 42, the hill to its east is the highest locally. Most of the terrain north of the Interstate is rolling hills, much of which is open farmland. The dark spotches are isolated strands of hemlock trees; some woodlands (not leafed yet) to the north have more hemlocks. The bright surface on the Susquehanna is sun glint. There are no buildings tall enough to cast shadows that would suggest time-of-day, but shadows associated with the bridges at Bloomsburg and the smaller Catawissa to the southest (probably too thin to show on your screen or in the printout) are consistent with mid-morning.

This is a good time to reproduce here part of the 1:24000 7 1/2 minute topographic quadrangle map made in 1953 (hence the town is smaller) that covers part of Bloomsburg and areas to its north. Compare the contour patterns with your assessment of the topography derived from your stereo inspection.



BACK

1-8:At the juncture, the land is flat and open to the east and north of the Fishing Creek. This is partly why the Fairgrounds were located there but houses in this joint floodplain are occasionally flooded. To its south, there is a sharp cliff rising to uplands on the Susquenanna's west side. South of Bloomsburg the river abutts against a steep sloping bluff topped by farmed uplands towards Catawissa Mountain. But note that homes have been built along the side which have a striking view of Bloomsburg when trees are winter-bare (the whole setting reminds me of my visit to Heidelburg, Germany on the Neckar River). Farmland to the south is higher than that in much of the land a few miles north of Bloomsburg. The alternating light-dark patterns, found throughout the full subscene, are a prime example of contour farming, in which the field is plowed parallel to lines of equal elevation and commonly two crop types (one usually corn) are planted side by side. You will see an good example later on the ground and will be challenged to locate it in the image.

This next photo was taken from Route 487 on the northeast edge of Bloomsburg overlooking the valley of the Susquehanna, its south bluffs, and the mountain areas to the south which are largely game lands (my home is hidden beyond the foreground trees).

BACK

1-9:The cluster of white geometric patterns denotes the roofs (light gravel) of several of the factories (one makes rugs for all General Motors cars) in the Bloomsburg industrial area. The two ovals are race tracks (one for practice) connected with the Fairgrounds. The University is rather difficult to find: best to start with the false color aerial photo. Follow the prominent Route 11 into town, where it jogs abruptly at angle of about 120 °, then turns again as Main Street. Just to the right of that second turn is a cluster of buildings, some with light - others with dark roofs. This is the campus. In stereo you will see it as occupying a hill, so it overlooks town. Route 11 is now (and probably in the '80s) macadam (dark asphalt) with several short patches of concrete. The short strip to the south is the single runway at the local apart. The bridge is all concrete - from roadway to its underparts, as you saw in the view of Bloomsburg from across the river.

This aerial oblique photo (Courtesy: Bloomsburg University) shows the Main campus of BU looking northeast (the ridge ending abruptly in the background is Knob Mountain).

BACK

1-10: The blackish pattern to the left of A is Lake Chillisquaque (Indian name), the largest lake in the TM subscene. It has a rock-earth dam at its south end used to build up the lake from local stream inflow. Its curved length is nearly two miles. It was constructed by the Pennsylvania Power and Light Company (PP&L) as a storage area for water used elsewhere, but they have

turned it into a splendid recreational area for boating, fishing, and nature trails (it is the best birding spot for waterfowl). Here is a photo:

BACK

1-11: In the TM subscene, the area left of B shows what appears to be water, dark ground, perhaps a building or so, and smoke. You might have guessed it to be a power plant - an electricity generator which uses coal for fuel. This is one of the PP&L plants for central Pennsylvania. Here are two views:

BACK

1-11:The borough of Danville (it's not a town, remember), another old river hamlet that has grown to more than 6000 residents. Let's take a look at it:

You are seeing Danville from the south end of the brand new bridge (opened in July 2000) that leads to the residential community of Riverside. Next to it is the 100+ year steel bridge in process of being dismantled. The "Town" Hall like building is the county seat for Montour County. In the far background, against the hills, is the yellowish main building of the Geisinger Medical Center.

BACK

1-12: To try to locate the field, go just below the red D. There is a large elongate field with a darker slightly curved band, and then a broader field to its south, and a small grove of trees. This seems to be the area shown in the photo but in the year 2000, the planting pattern is somewhat different in that there are several narrow green fields (alfalfa?) alternating with the now dry stalks of corn. **BACK**

1-13: Several of these dark green patches are Christmas Tree Farms, as they are called locally. A farmer must plant them and then wait a number of years (at least 5) to cut them down for sale across the east. Starting in November, one might see a number of long flatbed trucks or ones with gatelike sides going south or east with a load of Xmas trees for the metropolitan areas of the East Coast. In both Landsat and aerial images, they have the dark signatures of evergreens but because of their regular spacing in rows, a fair amount of soil makes a mixed pixel so that their recognition is difficult. Here is a picture of one of these patches:

BACK

1-14:From F you will see some farms and fields and then a long ridge with about the same elevation. Thus:

At point G as you look west you should see a small lake. There is a second larger one further west. These are dammed to make reservoirs for water supply and for wildlife. This is the view of the closer, smaller lake; it is partially filled with aquatic plants:

This bears on the third question: where are the higher elevations. If you look at the lower of the two above pictures, you will see a ridge on the south side. The water is in a valley in the central part of the green pattern; it is most unlikely to have been developed at the top of a ridge, unless water were hauled up by thousands of trucks - for no apparent good purpose. What the TM subscene fails to show is the true nature of the relief in the green pattern. It is not a single ridge, highest in the center, but is two bounding ridges, with a valley in between. The north ridge is known as Little Mountain; the south ridge is Big Mountain. Geologically, this is a breached anticline - an upfold in which the resistant rock units (sandstones) at the top were eroded away, exposing softer rocks (shales) that were then rapidly eroded downward (breached) to form a valley with the ridges on either side staying higher because they are composed of the hard sandstone. This situation is a strong argument for having stereo capability to resolve topographic uncertainties.

BACK

1-15:This ground photo should tell you the answer:

These are, of course, power lines. When constructed, their right-of-way is cleared for about 100 feet to allow easy access for maintenance. This particular line's pathway is being somewhat overgrown and may require later re-clearing.**BACK**

1-16: You're in coal country now. The south side of Big Mountain is held up by the Devonian Oriskany sandstone that dips south. The rocks to its south are younger - Mississippian and Pennsylvanian. The latter contains the Coal Measures which because of low-grade metamorphism have converted to anthracite - the premium coal in this country because of its high BTU content. The area around I has been extensively strip mined, as seen here:

Spoil banks - a mix of coal and shale - cover the surface. There are many trees with whitish bark in the photo: these are birch, which are acid soil lovers and grow profusely where the strip wastes weather into acid (low pH) residue; they lose their leaves early in the Fall.

This coal waste area is locally "famous" (or perhaps "imfamous"). The nearest town is/was Centralia, about 1500 people whose lives depended on coal mining. There were a number of underground mines at one time but high costs of extraction led to their abandonment. Then, one year someone through waste down a mine shaft. This eventually started a fire by spontaneous combustion. The fire followed a thick coal seam, burning its way westward. As the coal turned to ash, the space created led to instability and collapse, affecting the surface and the row houses in the village. It got so bad that the Federal and State governments forced evacuation of Centralia (a few intrepid diehards remain) and resettlement of most of the population. Today, as you drive through this area, especially during a rain, there are lines of steam (groundwater reaching the fire) that mark the current advance of the burning. This may take hundreds of years to put out (attempts to douse the fires have failed) and threaten Mount Carmel to the west (at the very edge of the enlarged TM scene, below the large lake).**BACK**

1-17: This is a "gimme". It's another power line. You will notice other, more irregular lines in the forests associated with Catawissa Mountain and the Game Lands in the mountains attached to it. These are dirt roads used for access by hunters and forest rangers (to aid in fighting possible forest fires).**BACK**

1-18:Go to the red G and move diagonally (about 45°) to the upper right. There is an irregular polygon whose tone in the enlarged subscene is lighter green than its surroundings. Several logging roads lead up to it. At least one other similar road ends at a lighter green patch. These are clearcuts now undergoing reforesting. Look around also for several small areas of green that is darker than the general color of the forest lands. These are hemlock stands. The black specks throughout the green tones are slope shadows or even tree shadows. **BACK**

1-19: K marks the borough of Berwick, similar in size to Bloomsburg. It's economy is largely industrial - small factory operations. Many retired workers have stayed there. There is great civic pride in Berwick High's football team - it has been state champs in its division several times in recent years, in one of which it won USA Today's national championship. The bridge leads to Nescopeck across river. If you look northeast from the bridge, you will note two long, narrow uninhabited islands. The Susquehanna is a very shallow river over much of its length, until it nears the Chesapeake Bay. Boating is restricted to canoes and flat-bottomed row boats because propellers can be ruined when the river is low. Here is a ground photo of the island:

Further upstream, past L on the labeled TM image, the valley of the Susquehanna becomes a typical area of farms between ridges, as shown in this image:

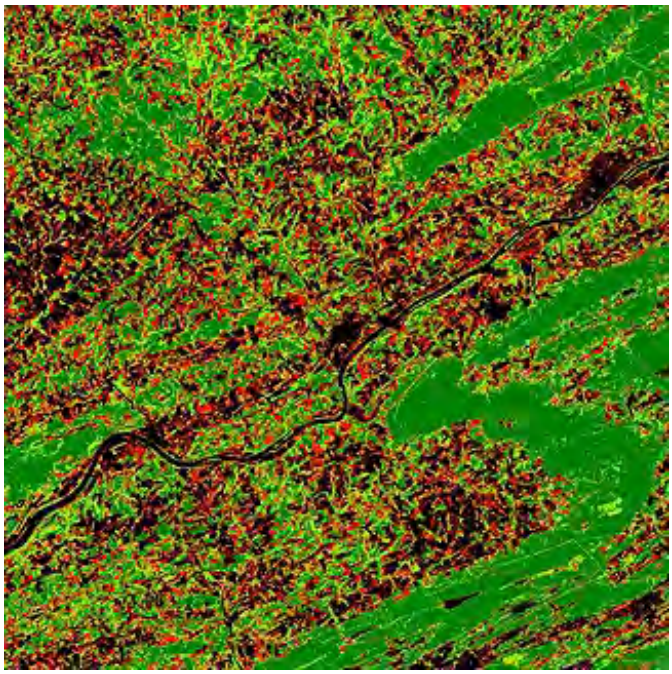
BACK

1-20: First, the municipalities are a richer blue than normal for a TM 2,3,4 = BGR version. This is because Bloomsburg, for example, is somewhat brighter in TM Band 1 than in Band 2, and thus contributes more to the blue band in the composite. Second, the Susquehanna River is red

instead of the usual blue or bluish-black. I've never seen this before. To be red, Band 4 should have a lightish tone in the river, which it doesn't. Sediments could to this, but they would be even more effect in Bands 1-3, which they aren't (if you look at the natural color rendition, the river is black). I don't have an explanation for this anomaly.**BACK**

1-21:Four general thermal patterns are of note: 1) the forested mountains are cooler (variably dark), but there is a little variation in tone which relates to differences in tree density and type (some areas have been reforested); no effect of differing elevations is obvious; the decreased temperature may well be from evapotranspiration cooling; some of the black fringe may be partial shadow; 2) most of the lowlands are comparatively warm, especially the "barren" fields; 3) the three municipalities - Danville, Bloomsburg, and Berwick - are bright, i.e., warm; this is the natural "heat island" characteristic of urban areas; and 4) the Susquehanna River is moderately cool (in the daytime) relative to the nearby land, as is typical of water (its temperature near the surface could be, say, 75° while the land in mid-morning might be 82° in mid-June. **BACK**

1-22: At first glance, the ratio 4/2 image seems almost like a density slice image with just a few gray levels. All vegetation - forests, grasses, crops - are very bright (the image was strongly stretched) and crop-early or barren fields and urban areas are dark. This is a good vegetation discriminator. But, there is a bit more variability in the scene which can be brought out by applying an IDRISI color palette to the image, as seen here:



In this version, there are four shades of green that roughly separate four types of vegetation, two or possibly three of which are on the mountains and one (or two) in the farmlands. The black tones are associated with bright fields and with urban areas. It is much harder to pinpoint the nature of the red tones. Many seem to be where fields or open areas with some vegetation are located. They generally correlate with a pinkish-gray color in the natural color renditions.**BACK**

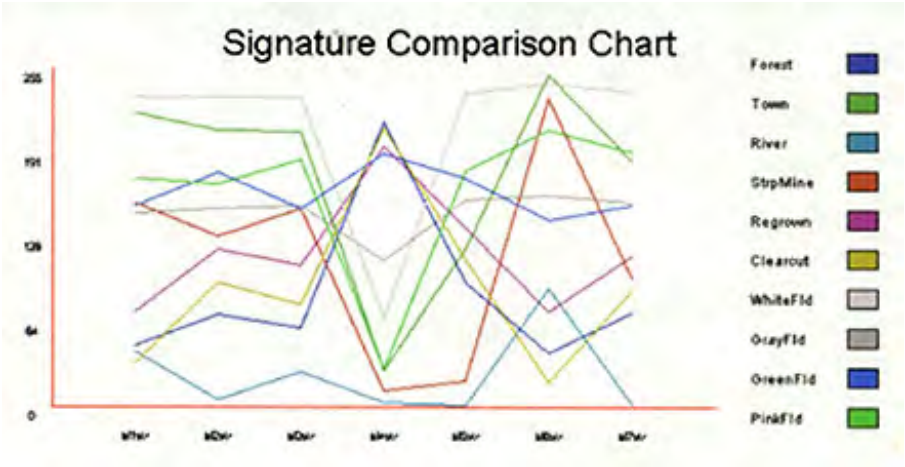
1-23: Green dominates the subscene, suggesting that the NDVI measure is very sensitive to vegetation. The lighter greens are scattered in the hill forests and more prevalent in the farmlands. Yellows correspond to open, near-barren fields. Towns are also yellow. The reds are found in the rivers and lakes.**BACK**

1-24: This image and its subset are a classic example of colorful remote sensing products that look at first glance as though they are rich in information. Yet, on closer scrutiny some of the

color patterns do not seem to correlate closely with either the raw imagery, i.e., before special enhancement procedures are applied, or with field experience. This PCA 142 composite does single out certain things effectively. The urban centers are distinctive. The vegetation on the mountains is well identified, and reds in the valley usually relate to woodlands and groves. The purples generally define the bare soils and the blue-greens associate with gray and/or (?) pinkish fields. The Susquehanna is red again but the small streams feeding it are in black. Inspection of the enlarged image shows that in a local area there may be a considerable patchwork of colors, including yellows and oranges, often with black borders, that may indicate the edges or boundaries (as between fields) are being highlighted. To sum up, this PC composite has notable correlations with reality in places but also seems to have a capricious or random component of patterns that defies easy interpretation. It is well to judge that remote sensing computer-processed products can assist in understanding surface, features, and conditions but often (especially with lower resolution) introduce or leave ambiguities. These produces are not necessarily a total solution.**BACK**

1-25: Several classes or categories are obvious. For instance, clouds and their shadows; the river; the forest; the municipalities and clusters of houses comprising villages. The forest appear to have three levels of green: 1) deeper green for thick forest growth (including hemlocks and maybe the Christmas Tree Farms but the latter are too small and isolated to be picked out reliably in the scene); 2) areas of medium green with a brownish tone - these could be second or multiple growth stands of trees developed long after they were timbered; 3) open areas of lighter green - most likely more recently cleared forests now undergoing regeneration. There seems to be four categories of fields (designated as such because of their location and shapes), based on color in the natural color subscene: 1) bright toned (whitish) fields, likely bare or with very early stage crops (e.g., corn); 2) gray fields, nature unknown but fairly common; 3) light green fields, presumably crops such as wheat or hay, that are at a more advanced stage of growth; and 4) pinkish-brown fields whose status is not deciphered (since this is a 1987 image, no supporting ground truth can be retrieved and thus no specific crop types can be chosen). The areas of strip mining in the lower right of the subscene are another discrete class. From the above, you probably can assemble 12 meaningful classes. But, note that all are land cover types rather than landforms (still, the mountains can be inferred from the knowledge that continuous forest cover occurs mainly on the larger structural units). Now return to see what the writer actually selected.**BACK**

1-26: 1) The natural color rendition of the TM subscene does a good job of displaying the different classes - those that are features larger than about 30 m in dimension. A skilled photointerpreter can generally identify and give names to most of the spatial/color patterns in this rendition. Classification is done mainly to specifically assign descriptive names to the classes, display each class in some separating color, and thus produce a map of the area shown; 2) Most of the classes chosen are straightforward, although some may seem a bit "artificial", i.e, the fields which are given names based on their image colors rather than what they are actually being used for or supporting (crops) at the time of image acquisition. One pair of classes should be renamed (the IDRISI program is unforgiving in this respect, not allowing editing of class names): the class called CLEARCUT is valid in the sense that once, a long time ago, it represented extensive forest removal that has since grown back to forest stands similar to the class FOREST; it probably should have been named REFOREST; the class REGROWN refers to more recent reforestation after later clearcutting; 3) Overall, the visual reaction one should get in looking at the classification (especially the printout version I made) is that the classes chosen and their locations make for a fairly successful classification (I would guess 80% or better accuracy. This is not surprising when the signatures are plotted out from IDRISI's SigComp program, as shown below (the legend colors are a default and cannot be rearranged to match the colors in the legend).



Most of the classes (Pond and Shadow were dropped in this plot) could be separated just from TM Bands 1-3 data; the remaining four improve the separability. None of the classes chosen could be mistaken for any other on the basis of signature. Four of the five vegetation classes all have similar TM4 values but WHITEFLD is much lower - its signature suggests it is truly barren (not yet planted or with just emergent crop). The towns - large and small - are well separated. RIVER is singled out; the large and small lakes in the valley between the mountain ridges near the image bottom classify as RIVER as well; the class POND was chosen from a single training site near Lightstreet and its color (yellow) shows only there. The various field classes seem correctly placed but note that the classification could not resolve those fields that consist of alternating strips following contours (they are plotted as a single color pattern so some misclassification occurs here). The reds of STRPMINE are properly placed in the Centralia area. But small patches of red occur where they shouldn't (e.g., around Elysburg; see map); this is probably caused by a land use class that resembles the wasted land of the strip mines. Note, too, that the PP&L power plant near Lake Chillaquaque is shown in brown and red: it, in fact, is similar to both a town and waste areas. Both small streams and some roads are shown in black (they were too thin to allow training sites to be drawn and are thus prone to misclassification). Shadows, in black, may be correct where associated with mountains but the few clouds in the scene are classed as UNKNOWN (first black in legend), since they were not selected as signature, and their shadows are blue (similar in signature to RIVER). To summarize, the classification came out better than I expected.

BACK

1-27: Just north of meeting of the two branches, and on the north side of the North Branch is Northumberland (whose greatest claim to fame is that James Priestly, the discoverer of oxygen as an element, lived there). Across river is the largest urban center, Sunbury, PA. On the west side of the combined branches is the borough of Shamokin Dam, mainly a very long strip of malls and businesses. Although its name is cut off the edge of the road map, the next town south along the river is Selinsgrove, home of Susquehanna University, a private liberal arts college. None of these really stands out in the L-band radar image, but all are in light tones (many building reflectors). Strangely, they are not clearly differentiated from their surroundings in the radar color composite, where they appear pink. Sunbury, however, shows little pink despite it being the largest community. Sometimes, radar composites can be misleading. The purples are mostly open fields. **BACK**

FINAL EXAM

It's been a long time between Exams - your first was at the end of Section 1. But, if you have worked your way through most (hopefully, all) of the Sections in between, your expertise has grown significantly. In this Final Exam, you will find similarities with the themes pursued in the first exam but with more sophisticated questions and interpretations to be handled. Once again, Landsat imagery will be the principal data type but with TM rather than MSS images to work on. Since you became familiar with one part of the U.S. - central Pennsylvania - in the earlier exam, we will make it easier for you by staying in the same geographic region. The emphasis in this Exam is, again, on visual interpretation of the features and contents of a scene.

To tie in with that first exam, recall your experiences with this MSS full scene, with Harrisburg near its left center.

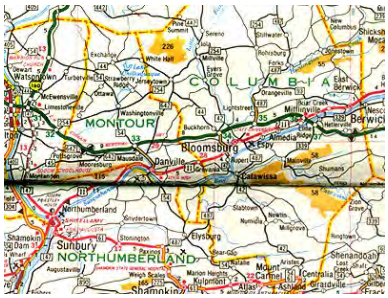
You should remember that the major river is the Susquehanna and that the upper part of the scene is dominated by the Valley and Ridge subprovince of the Appalachian Mountains. Now, as a transition to a TM scene, shown below is an enlargement (thus, somewhat fuzzy) of a subscene from the MSS image lying in the upper central area:

In the top center, where the Susquehanna bends sharply north, is the town of Bloomsburg, the writer's (NMS) present home, and the focal point of this second Exam. In the MSS subscene, it is very hard to pick out, being just a small bluish spot.

Now, lets introduce the TM scene on which this exam is based - it is extracted from a larger (full) image taken on June 17, 1987, comprising about 40% of the latter image. We will this time work mainly with the natural color version made by the IDRISI software program.

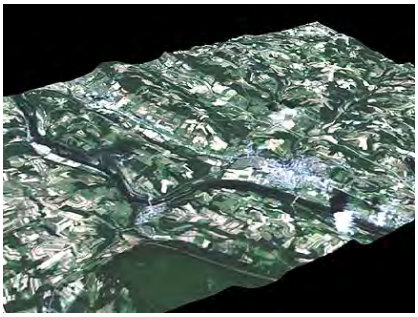
1-1: **Try to fit the TM subscene into the MSS subscene, using several landmarks. Roughly, what percentage of the TM subscene is present (overlaps) the MSS subscene?**
ANSWER

It should help, too, to get your bearings by relating the subscene to the Pennsylvania State Official Road Map. The area of that map reproduced below, and includes some locations outside the TM image but several of these will come into play near the end of the Exam. Use the Susquehanna River to fit the TM subscene to the map.



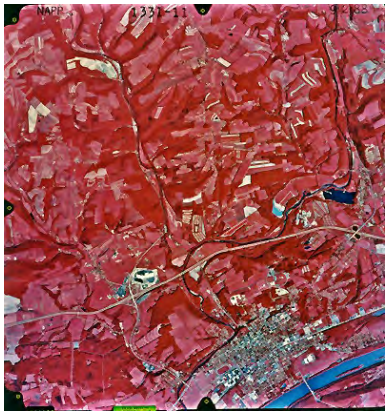
Bloomsburg is almost "dead center" within the above TM subscene image. Lets talk about this place for a bit. First, we show a further enlargement of it and surrounding countryside. Several small clouds are in the image; you should remember how to recognize them.

Digitized elevation data (DEM) have been combined with part of the Landsat subscene shown above to generate a perspective version of the terrain. You should now be familiar enough with the subscene and key landmarks to recognize the major features in this oblique view.



After this exam was first produced, a Quickbird image (Digital Globe; image acquired at Google Earth) of Bloomsburg and vicinity was downloaded off the Internet. It is shown here, with somewhat degraded sharpness:

And, now we supplement the above views with a higher resolution false color aerial photo, at a scale of approximately 1 inch to the mile (~1:62,500), taken in September of 1988. Bloomsburg is at the bottom.



Compare this with this aerial oblique black & white photo of Bloomsburg, looking north across the Susquehanna River

To feel more at home in this small urban area, here are two ground pictures: 1) looking at Bloomsburg from across the river, where we see that part of the town is perched on a hillside, and

2) a view of Main Street looking east, with the edge of the Bloomsburg University campus barely visible at the far end.

(NOTE: The Landsat subscene we are using became available to the writer in early October, 2000. A week later I made a tour of the region depicted in the subscene expressly to take ground photos. Fall was well underway. This is why so many of the photos you will see have the lovely yellows, reds, and browns in the deciduous foliage.)

Bloomsburg, the county seat of Columbia County, has a population of about 12000 within its city limits. Its surrounding "suburbs" raise the population by another 5000 or so. During the academic year, Bloomsburg University, a 4-year college that is one of the 11 State System of Higher Education (SSHE) establishments, adds another 6800+ to the residents. Bloomsburg has a local symphony orchestra (university students/staff plus town's people) and a resident drama theater of fine repute. Once a year, in late September, the town puts on the largest State Fair in Pennsylvania. Bloomsburg, for part of the 1830s, was the iron-producing center of the U.S., using iron extracted from sedimentary red beds plus coal from the southern part of the TM subscene, and limestone from beds along the river. Today, a number of industries (auto carpets; clothing; aircraft parts) have factories within the town or nearby within the river floodplain. One of two current centers of shops and services has shifted to a 5 mile strip along U.S. Highway 11 which runs through the town and then eastward as the main road, as shown below:

The other center is built around the Columbia Mall where Rt 42 passes over Interstate 80 (see '34' in above road map). There, a Walmart, Lowes, and Home Depot have, along with other stores, made this cluster the shopping center for Columbia and adjacent Momtour counties.

Bloomsburg is, officially, the only incorporated area in the State of Pennsylvania that is entitled to call itself a "TOWN". It was founded in 1802 so that 2002 is its Bicentennial Year. Settlements began to appear in Columbia County area in the mid- to late 1700s, established mostly by Quakers who then left for points west and north by the end of that century. Bloomsburg as a hamlet dates to about 1785 and its present street plan was laid out in 1805. The University was founded as a teachers college in 1838 and thus has a 150+ heritage. The county has one of the largest numbers of covered bridges in the U.S. Columbia County is traversed by Interstate 80, which runs from just west of New York City to just east of San Francisco; it passes Bloomsburg about a mile and a half north.

Two neighboring communities (boroughs) on the Susquehanna, similar in size, are Berwick (about 12 miles east) and Danville (10 miles west); the latter is the site of the Geisinger Medical Center. Find them on the map and then in the full subscene - they have a pinkish color. Catawissa and Elysburg are smaller settlements.

1-2: **Based on area within the subscene, what is the principal landuse and a major source of income? ANSWER**

1-3: **About 14 km (8.5 miles) south of Bloomsburg is a large green area shaped something like a "U" lying on its side; what do you think this is? ANSWER**

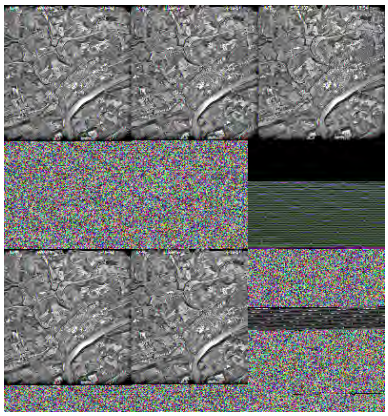
1-4: **The green area continues and expands to the south and east. Assuming you deduced that this is heavily forested hilly terrain, what do you think its main use is today? Careful, this is a tricky question. ANSWER**

1-5: **Go back to the enlarged area centered on Bloomsburg and to the aerial photo just below it. Leading out of the west side of Bloomsburg is a road (State Route 42) that eventually meets Interstate 80 at a cloverleaf intersection. Describe the area just to the northeast of the intersection as it appears in the TM enlargement, and again as it**

appears in the aerial photo. There is a difference. Speculate on what now occupies this area. ANSWER

1-6: **In the enlarged subscene there is some water besides the River. It appears again in the aerial photo. Where is it? With its odd shape, it does not appear to be natural. Any thoughts on its origin?** ANSWER

We will turn now to a trio of black and white aerial photos that pass over Bloomsburg. They were acquired on April 18, 1994, so the trees were in an early stage of leafing. The photos are adjacent, so stereo is possible. We reproduce one of these on this page, but since you will be asked to print these out, and the Exam page is long, a link to a separate page has just the pair on it, for easy printing. After the printout, just hit the browser BACK button to return to the Exam.



After cutting the stereo photos apart and aligning them properly, if you have a stereoscope or can see 3-D with the unaided eye examine the topography around and both north and south of Bloomsburg.

1-7: **Start with the pair: 5539-136 and 5539-137. Locate the major hills in the scene. Where are the high areas in the immediate Bloomsburg area? Are some of the farms located on rolling terrain? There is a wooded area just south of the Interstate and southeast of the Columbia Mall that is medium gray, with dark spotches within; any ideas as to what these specks are (other area show this feature, sometimes being larger and continuous)? Why is the Susquehanna in 136 so bright? Can you approximate the time of day when the photo run was made?** ANSWER

1-8: **Now examine 5539-137 paired with 5539-138. A major tributary to the Susquehanna - Fishing Creek - joins the big river west of Bloomsburg; near the juncture, describe the topography and also the topography along the west bank of the Susquehanna south of the confluence. What is the topography on the immediate south side of the Susquehanna across from Bloomsburg? Does the farmland further south seem higher than that north of Bloomsburg? Account for the pattern of alternating and curved light and dark patterns in several of the fields.** ANSWER

You should have picked out the topographic features as seen in stereo. Hills rising up to 300 ft above the Susquehanna flood plain will be obvious. Although hard to see the numbers, look for the brown elevation contours in this 1:100000 topographic map of the Bloomsburg area:

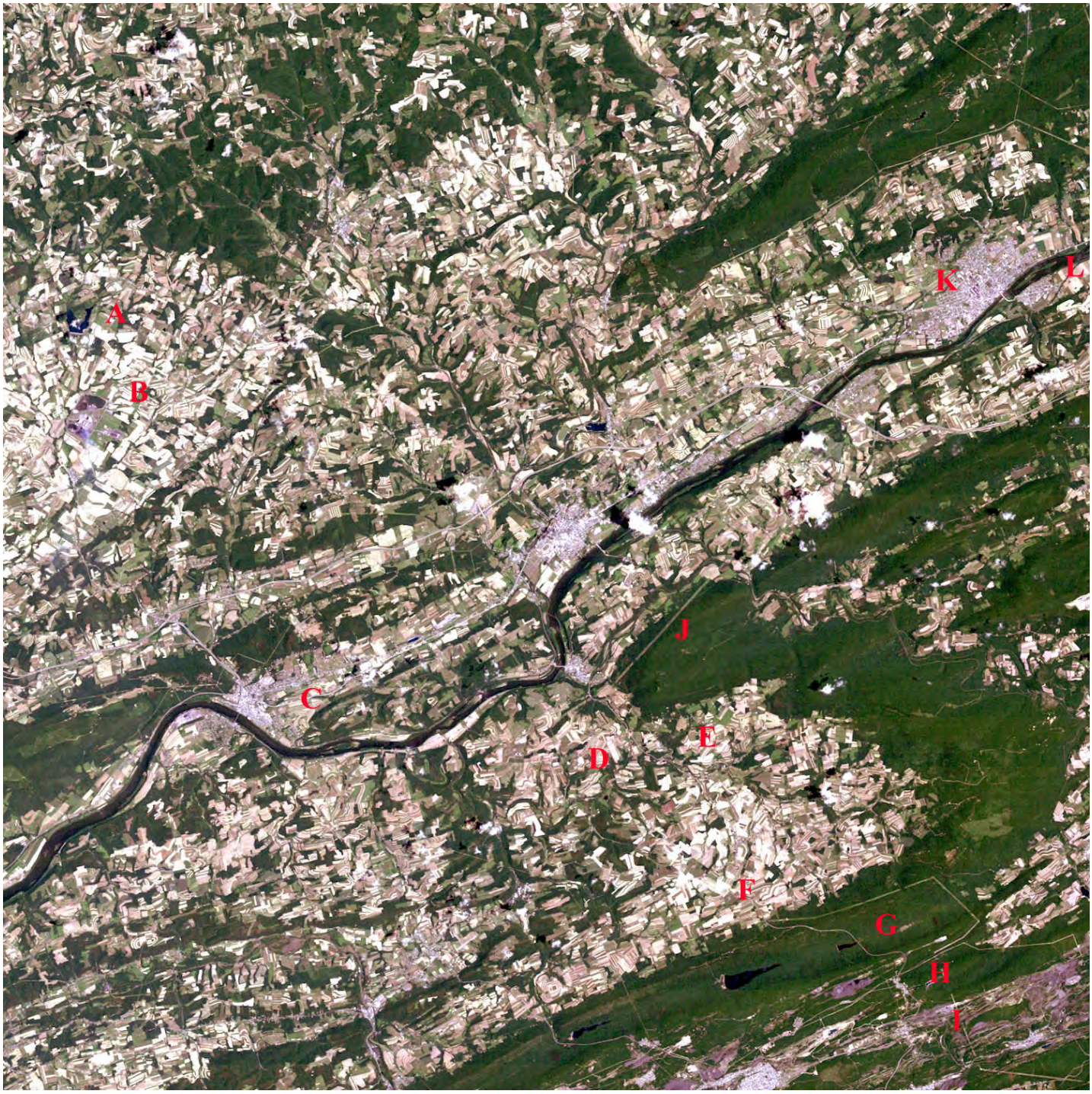
This map shows that the southwestern part of Bloomsburg is very flat and just above river level. This open area for a century and a half has been reserved for the famous Bloomsburg Fair. Ask yourself what would happen to this ground if a major flood spilled over the banks of the Susquehanna. Less than one week before the Sesquicentennial (150 year) opening of the Fair on September 24, the tropical storm remnants of the devastating Hurricane Ivan hit most of Pennsylvania. The Susquehanna crested briefly at 17 ft above flood stage, inundating the fairgrounds as seen below; but the water quickly receded, the ground was dried, and with herculean efforts the Fair opened on time.

Press-Enterprise photo

1-9: **Concentrating on the town of Bloomsburg itself as displayed in aerial photos (and, if you wish, checking also the false color aerial photo), what are the cluster of light patterns with square corners found in the western side of town? Why are there two ovals nearby? Can you find the University? Take a guess as to what is the covering material on Route 11 east of town and what is the identity of an isolated medium-dark straight line near the bridge across the river; what is the bridge surface itself?** ANSWER

Now we are going sightseeing around the full scene. But what you see below is greatly enlarged - you will have to scroll with your vertical and horizontal buttons to get to the different features, labelled A through L. You will be asked to identify each, starting with A, and then go to the answer page to learn if you are right, to see a scene or two that shows the feature, and to read a bit of descriptive background. Wander through the super-scene now and then find the first

question below it.



1-10:**Identify the features to the left of the red A, and then B.** ANSWER

1-11:**What is just next to C?**ANSWER

1-12:**There are some fields next to D. Here is one. See if you can locate it in the enlarged subscene.**ANSWER

1-13:**There are several patches of dark green to the left of E. We will tell you that they are not natural woods. They represent a seasonal product that is an important commodity produced in this part of Pennsylvania. Take a shot at saying what they contain.**ANSWER

1-14:**You are standing at point F looking south: What do you expect to see? Then, you are at point G looking west, and once more what do you see? Where is the high points (in elevation) in the green pattern in this vicinity?** ANSWER

1-15:**Looking around at point H in the subscene, there are several patterns of thin, regular lines that may be straight in segments. They don't appear to be roads since those are curved as they follow the topography, ranging between valleys and the need to climb over hills. What might these straight patterns be?**ANSWER

1-16:**All around the red I are irregular patterns that have a purplish color, mixed with**

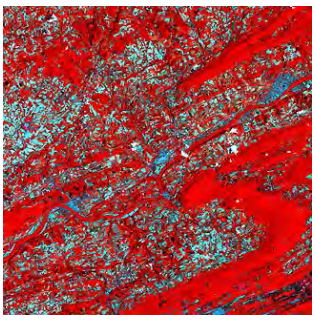
some black. Hark back to the first Exam during which this pattern was evident (look at the MSS image above) and was explained. With this hint, you should be able to identify these patterns.ANSWER

1-17:**Red J is on Catawissa Mountain. Here is a photo of it west end seen from the north. There is a long dark curved line about half way up its slope. What might it be?ANSWER**

1-18: **In the forests that extend from Catawissa Mountain to the east and south, there are several patches of lighter green. Find them and explain their nature.ANSWER**

1-19: **What's at red K? If you were standing on the bridge there looking northeast up the Susquehanna River, what special feature is in the river (near L)?ANSWER**

We now return to the original TM subset of data, which the writer (NMS) obtained after all of the above was completed. Using the 7 TM band images, various renditions were produced using the IDRISI for Windows package. The first product is a false color composite made from TM1 = Blue; TM3 = Green; TM4 = Red (not quite the usual false color rendition):



1-20: **While this appears to be a typical standard false color composite, it has two unusual or anomalous variations from the norm. Can you specify them?ANSWER**

Although we will not ask a formal question, look at just two of the bands from the TM set: Band 4 (left or top) and Band 7 (right or bottom). At first glance, they seem almost to have a negative-positive relation - what is light in one is dark in the other and vice versa. But a few features don't show that reversal. Look for them.

Band 6, the thermal band, is shown next. Its 120 m resolution was resampled to a repetitive 4 adjacent pixels of 30 m each to allow direct registration with the 30 m pixel size for the other bands.

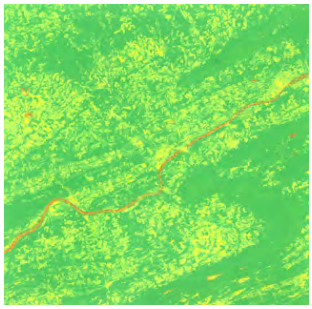
1-21: **Conduct a general analysis of this image in terms of relating warmer (light tones) and cooler (darker tones) areas to specific features.ANSWER**

It should be obvious to you that vegetation is one of the dominant features/materials in the subscene. By June 17 the deciduous trees, by far the most prevalent large vegetation, are fully leafed. Grasses are also growing well. In this part of Pennsylvania, in June many crops were just planted (commonly in May) and are not yet beyond their early stages of growth. So, when one drives across the countryside, these fields probably show some signs of the planting but from space the background soil occupies a higher percentage of a farmland pixel than the vegetative material it is nurturing.

One quick way to assess the relative proportions of vegetation-dominated terrain from other land uses is to ration the Near IR band to one of the visible bands, usually the red band. Here, we present a ratio of 4/2 (the highly reflective IR band to the green band):

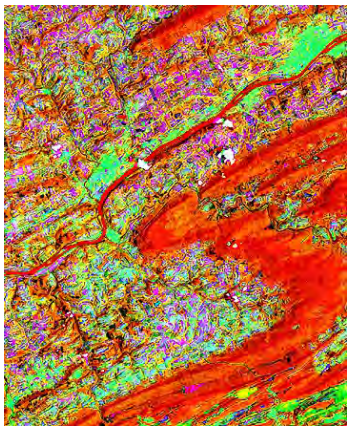
1-22: **Comment on the most readily apparent information in this image. A colorized rendition will appear in the answer. ANSWER**

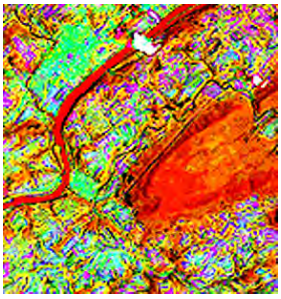
page 3-4), we introduced the NDVI as a sensitive means of recognizing vegetation at several levels of type, density, health, etc. The Index is, as you may recall, the ratio of Near Ir - Red to Near IR + Red (for Landsat, SPOT, AVHRR, etc.). IDRISI allow this calculation and has a special color palette to display the resulting image. Here it is:



1-23: **This a rather strange-looking (almost displeasing) appearance. There seem to be just four distinct colors. Comment on what you think this NDVI image is telling you. ANSWER**

IDRISI can rapidly calculate Principal Components. The resulting images can be informing or can offer little new or understandable insights into the scene's feature content. Using this image processing program, the 7 TM bands were incorporated into PCA and the seven resulting component images were examined. Individually, most revealed little added information. Various combinations of three PC images were tried - most failed to develop presentable composites when various IDRISI palettes were applied. Here is a colorful exception made by displaying PC1 = blue; PC4 = green; PC2 = red, applied to a large segment extracted from the full TM subscene. Below it is an enlargement of the area around Bloomsburg.



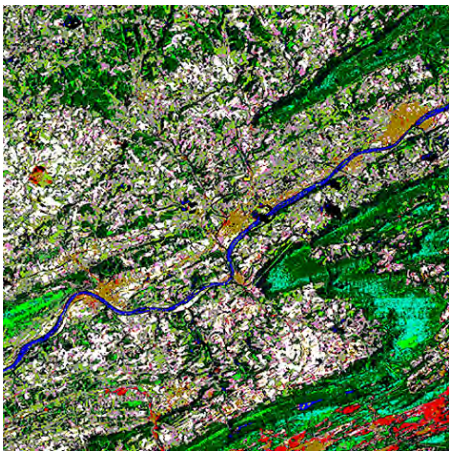


1-24: **This pair of PCA images reminds me of a cross between French Impressionist painting and modern abstract art (many space images have been collected into art shows!). But most of the color patterns seem equatable to specific scene features. Try your hand at interpretation. ANSWER**

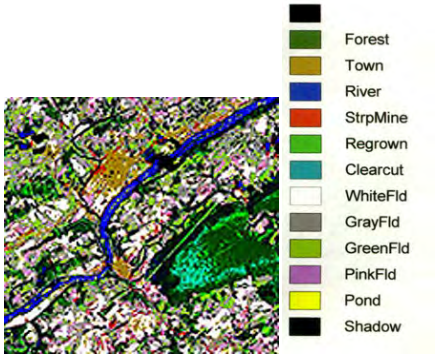
Assuming you read through the answer to this last question, you should be prepared mentally to tackle this next image. It is a Supervised Classification (IDRISI system; Maximum Likelihood) of the full subscene, using all 7 TM bands and attempting to discriminate 12 classes. The proper skeptical attitude to adopt after looking at it is: Are the classes meaningful and were those set up reasonably accurate in their distribution. Before proceeding to the classification, answer this question:

1-25: **Go back to the full screen plus enlargement of the TM subscene and scroll over different areas as you choose. Make a list of what you think are valid classes present in the areas represented. These could be completely different from one another or variants of the same class, such as various agricultural field uses or conditions. You should keep in mind that your selection ought to take into account the likelihood that the color variants will probably also have spectrally separable characteristics (thus, distinctive sets of DN values that will respond well to the statistical measures calculated in doing the classification). Make your list before going to the answer. ANSWER**

Without further ado, here is the Supervised Classification:

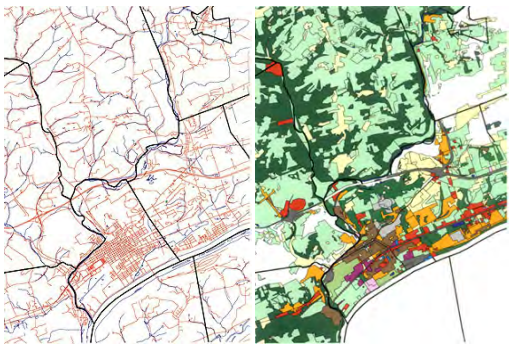


To help you in appraising details, there is an enlargement of the classification around Bloomsburg, and the legend (color scheme) for the 12 classes is shown to its right.



1-26: **Evaluate this classification (you may want to re-read pages 1-16 and 1-17 to remind yourself of the principles of classification) in terms of 1) the purpose behind it; 2) the wisdom in choosing the particular classes, and 3) an estimate (visual impression) of the accuracy in correctly picking the classes (errors of omission and commission.**
ANSWER

As this Exam was being conceived, I had hoped that the Geography program at Bloomsburg University had, by now, carried out a GIS study of the TM subscene (since there is a whole course on that subject that uses IDRISI's extensive GIS capability). Unfortunately, that hasn't happened as yet. But, I was steered by the course professor to Mr. Tim Murphy (a graduate of that program at BU) who now works for Columbia County as their GIS specialist. He, and the county, have kindly provided several data layers that are already in their system but the project is still continuing (a major GIS effort can take a long time), though not ready for the decision making phase. Digitized and captured into a data base so far are the roads and streams and township boundaries for the entire Columbia County, and land use maps for about half the area. Through their generosity, we can display here two thematic maps or data elements, both for the area around Bloomsburg and its immediate north. The left (or top) map is roads (red) and streams (blue) combined; the right (or bottom) map is land use categories:

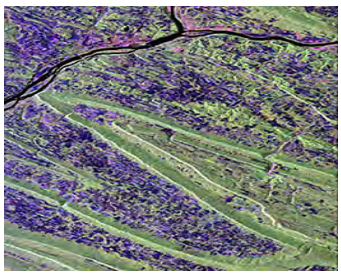


We will not ask you any questions directly pertaining to these maps. For your information, in the land use map, the land use categories are as follows: 1) Dark green = Forest; 2) Medium green = Open Field (at the Fairgrounds) or City park; 3) Light green = farm or grass lands; 4) Yellow = Low density housing; Orange = Medium density housing; 5) Brown = High density housing; 6) Pink = Schools; 7) Medium gray = Bloomsburg University; Dark gray = Roads or the Airport; 9) Red = Commercial; 10) Purple (in part) = Factories; Blue = Lake; Black = Township boundary lines; White = Not yet classified.

As we close out the Exam, we will swing westward to look at another enlargement from the full TM scene and also one covering some of the same area by the L and C band radar on SIR-C. Compare these three images:

Landsat TM Subscene

SIR-C L-band HH Subscene



SIR-C Color Composite using L and C band Images

1-27:In the Landsat subscene, use the map near the top of this page to identify the towns along the Susquehanna River (note the confluence of the West and North Branches [this second actually here comes in from the east but began in New York state as a southward-flowing stream]), using the road map above as a guide. Then, try to find them in the L-band HH middle image. Finally, how do they appear in the radar color composite? What does the purple color correspond to? ANSWER

It's time to close this rather lengthy and demanding Final Exam - almost as rigorous as a Ph.D. "written". You will have to grade yourself on how well you have done. Hopefully, good enough to feel mentally aroused and emotionally satisfied. Having worked through the main body of the Tutorial, if you assimilated the essentials you can add remote sensing to your skill mix. Let's believe that you will encounter opportunities to apply this knowledge. We close with a:

HALLELUJAH!!!

William E. Stoney, once with NASA and now with the Mitretek Systems, Inc., has recently been presenting annual reviews of the "big picture" for the future of remote sensing, especially as conducted from space vehicles. He has graciously permitted use of a paper given in 1997 which, along with a number of key illustrations, effectively summarized the Outlook for the Future (of Remote Sensing) at that time. This page is a copy of that paper. However, much has since become outdated. Efforts are being made to update some of his Tables and other figures. For information that goes to 2004, you can track down Stoney's article "A History of Civil Land Imaging Satellites" in the Encyclopedia of Space Science and Technology, 2005.

REMOTE SENSING IN THE 21st CENTURY:

Outlook for the Future

Land Sensing Satellites in the Year 2000

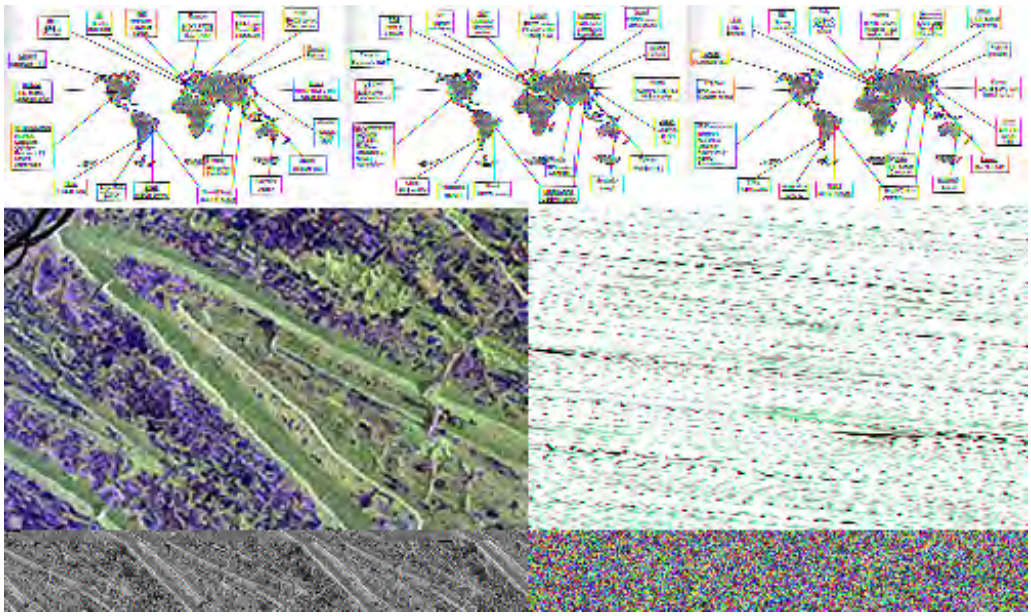
Based on a Paper presented at the
International Geoscience and Remote Sensing Symposium (IGARSS),
Singapore, August 7, 1997

(Updated through July, 2007)

By:**William E. Stoney**
Mitretek Systems
7525 Colshire Drive, McLean VA. 22102-7400
703 610 1768, (1767 fax), wstoney@mitretek.org

Editor's Note: The schedules stated on this page are inexact in that some launch dates have slipped and several satellites failed on launch or were cancelled. The survey includes mainly those satellites designed primarily for land coverage.

At the outset of this Tutorial, at the top of the Overview2 page, a graphic prepared by the author of this Section showed likely earth-observing remote sensing satellites expected to be operational between 1995 and 2005. That illustration is reproduced here:



While it is probable that the reality will not be as bountiful as the graphic indicates, it is certain that the number of high resolution satellites in operation will be significantly greater than before 1995. This Section builds upon that graphic by concentrating on the high resolution systems currently planned to be operating after the year 2000.

The systems can be usefully classified into four groups, (a) Broad area coverage, 5 to 30 meters resolution and multiple color bands, (b) Narrow swaths, 1 meter or less panchromatic resolution, and VNIR color only, (c) Hyperspectral sensors with 30 meter resolution and (d) Radar with 5 to 10 meter resolution. Their capabilities are described and compared in detail, including their spectral bands and resolutions, and their coverage capacity.

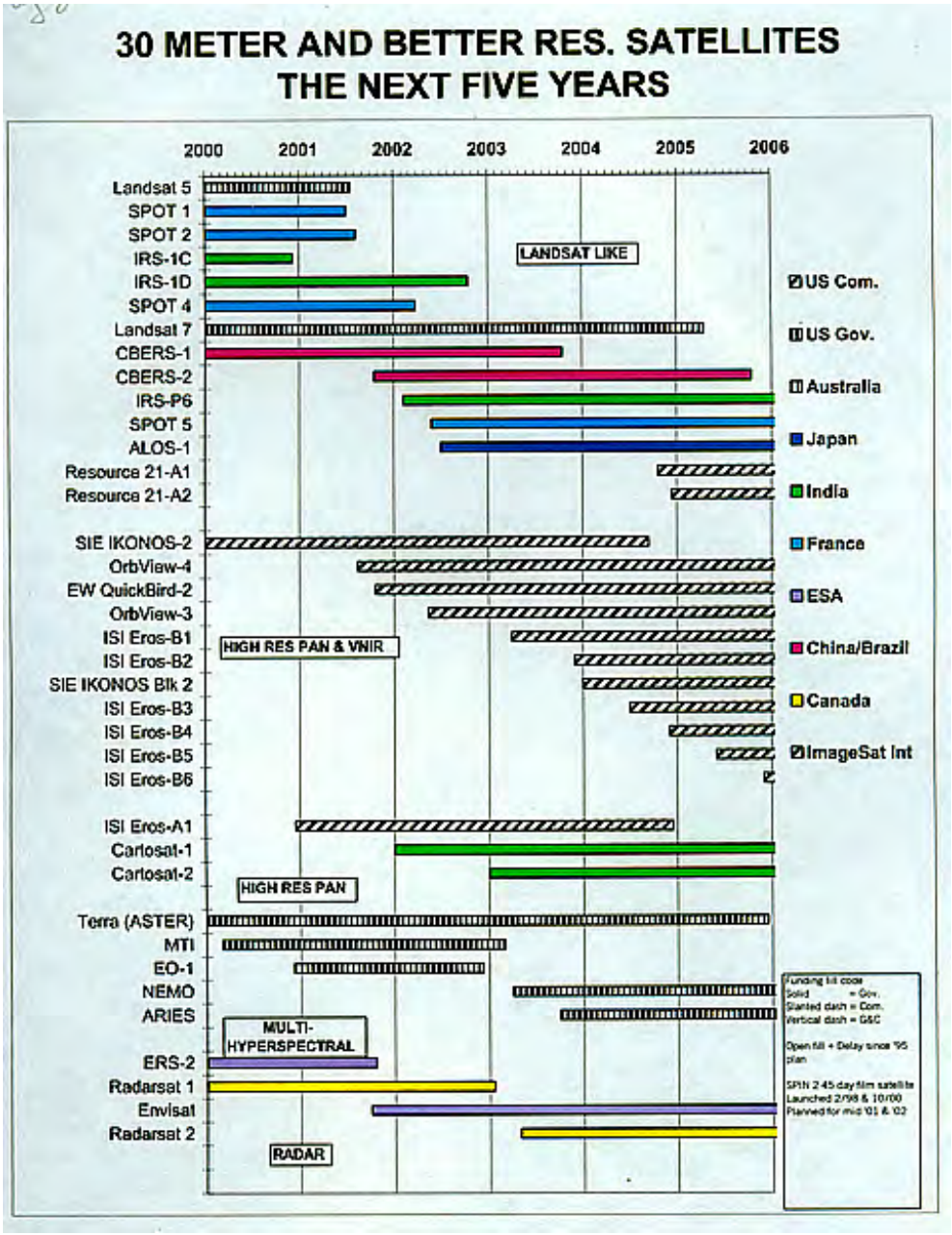
Purpose

Thus, this Section provides a brief overview of the explosion in land observing satellites planned for the first decade of the 21st Century and is intended as a wake-up call and planning tool for all who are interested in knowing and keeping track of the details of what is going on with the surface of our planet and in particular for those who are developing the skills to measure and understand the breadth and detail of the information that analysis of the satellite data could make available to us for the first time. The amount and quality of the land information data which the land observing satellite fleet in 2000 will be capable of providing could revolutionize both our scientific knowledge and our practical management of our Earth's resources. The satellites are however only the first step. Their value can only be realized through the ingenuity and efforts of the users.

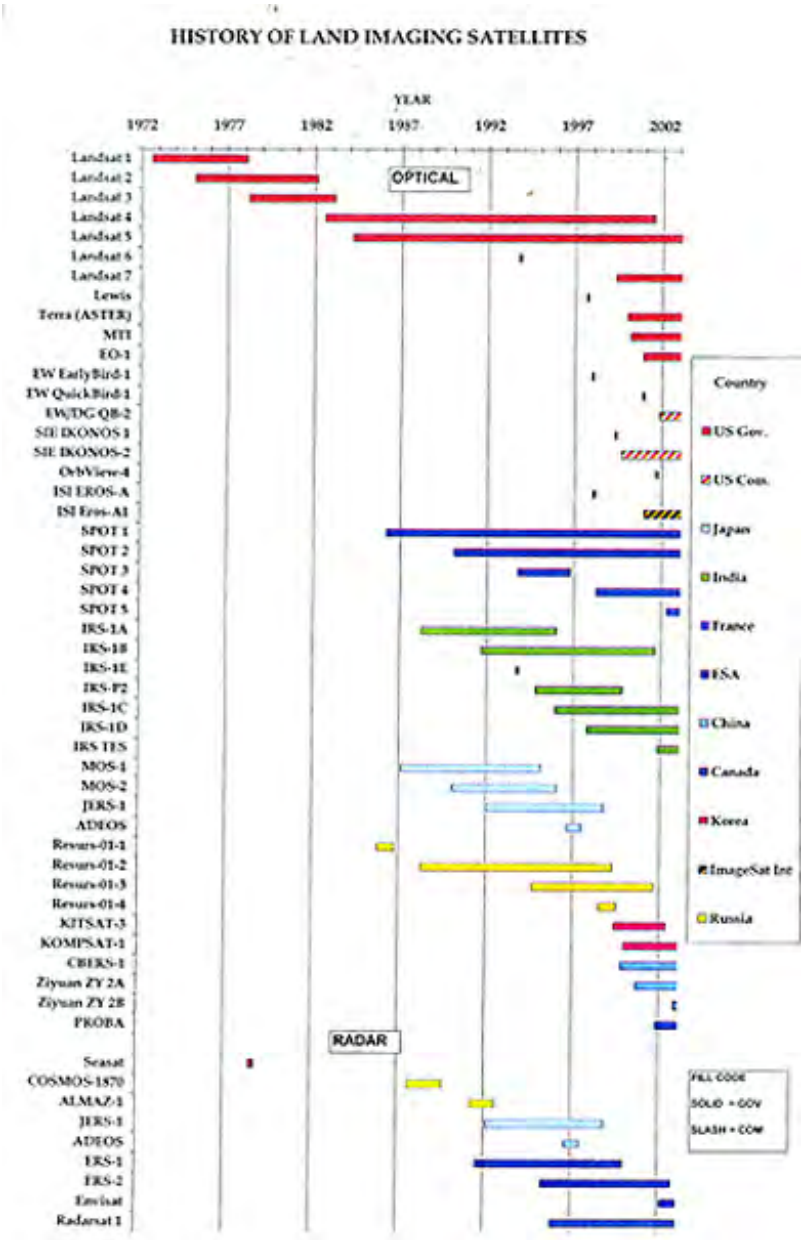
The Satellites

Among the many predictions for the new millennium are the orbiting of at least 31 satellites by the year 2000 in polar orbit providing land cover data at resolutions of one to thirty meters. These satellites are summarized in the first chart.

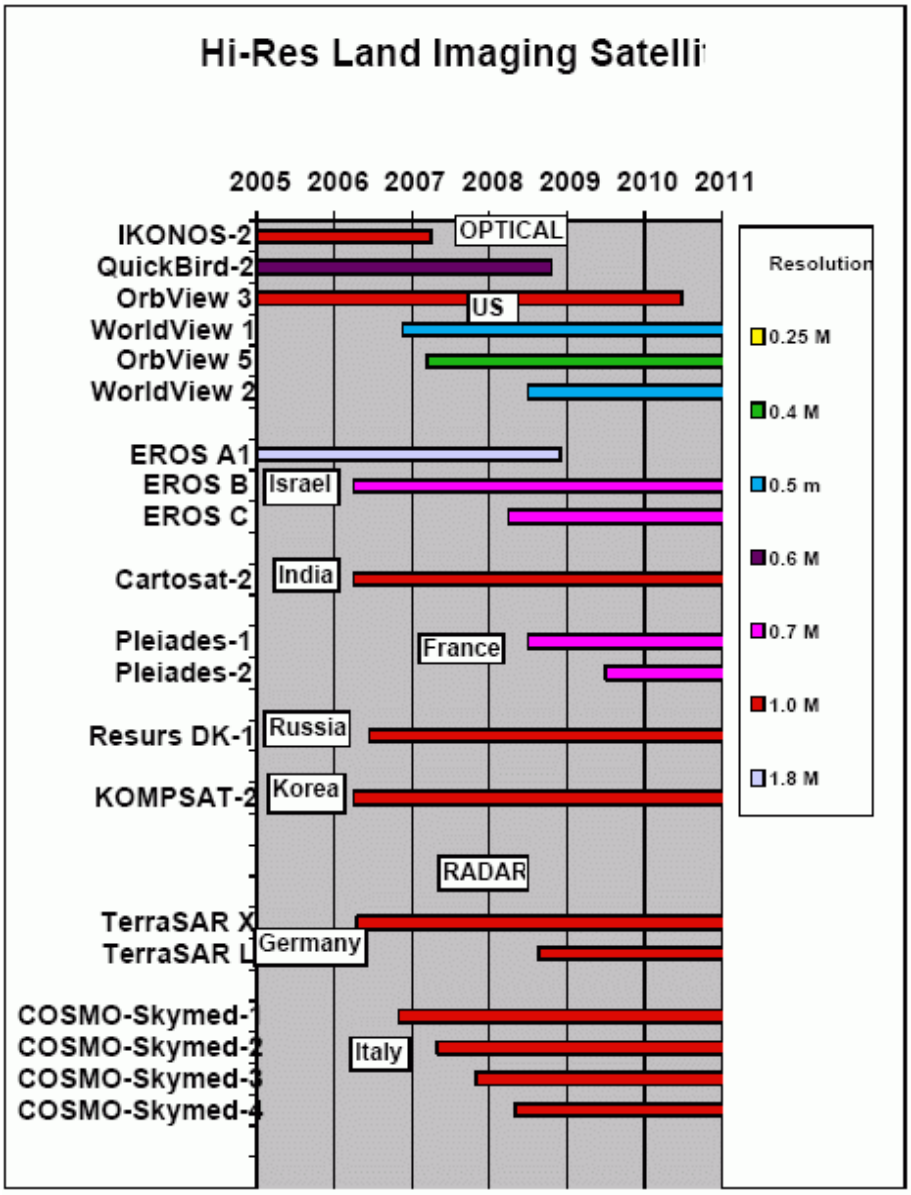
Satellite names of the principal vehicles actually in or planned for orbit and launch dates from 2000 through 2006 are summarized by the next graph.



A similar chart shows the historical span of actual satellites that are primarily land observers which are now in orbit at the end of 2002, or once were operational (failed or decommissioned):



This next chart, added by the principal writer (NMS), extends the time line to the year 2011:



Landsat-like: The Landsat-like satellites have the middle resolution, broad area and multispectral coverage characteristic of the current satellites, Landsat, SPOT and IRS. These current programs are being extended and expanded. As can be seen in the Indian program, plans for the flight of four satellites through this period are the most operationally robust of the government group. The group will be joined by two satellites created by a cooperative program between China and Brazil and one, four satellite, private system.

High Resolution: The twelve high resolution systems will provide an order of magnitude improvement in ground resolution, at the expense of less area and multispectral capability. With the exception of one Indian and one Russian satellite, these satellites are all funded and operated by private corporations. The almost exclusive interest of the private sector investors in the high resolution systems indicates their belief that this is the space capability required to create commercially valuable information products.

Experimental (Hyperspectral): The three government funded hyperspectral satellites and the proposed private system will explore the potential for the development of new multispectral analysis based applications by providing near continuous radiometry over the visible, near IR and short wave IR spectrum.

Radar: The current Canadian and ESA radar programs will be continued into this period as well. Radar's all weather capability makes it the instrument of necessity for many observational problems and it will become increasingly valuable for general problems as better techniques for analysis are developed, including the integration of radar and optical data.

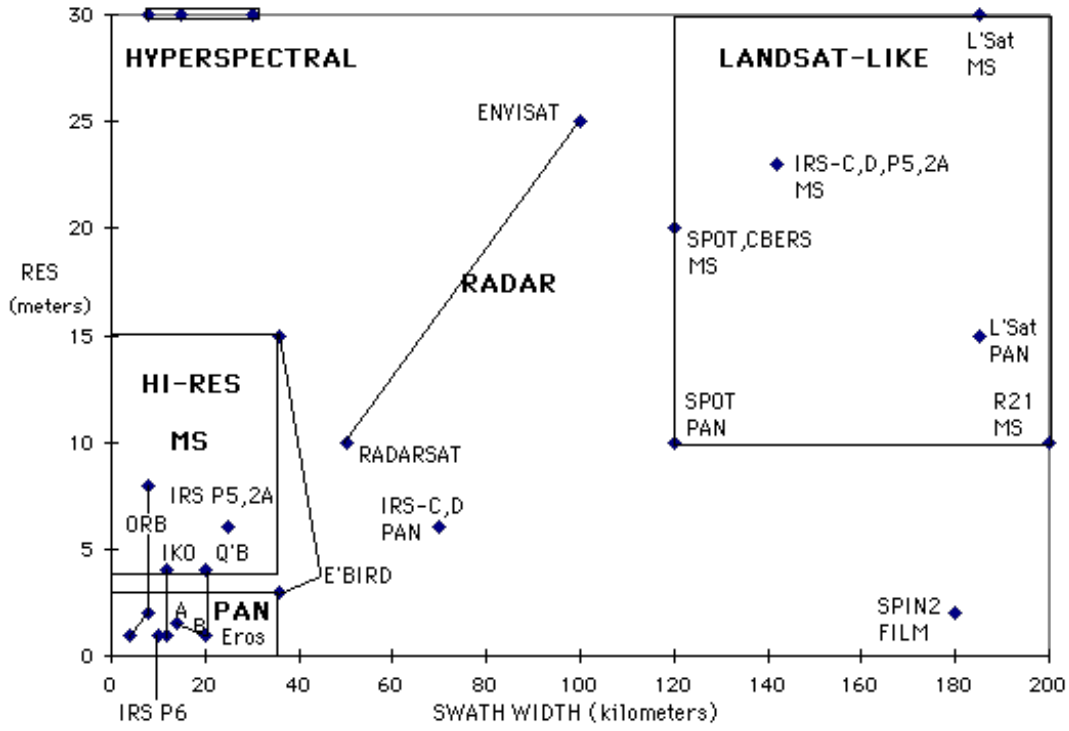
Technical Overview

The best way to understand the scope and variety of the data which will be available from the new millennium fleet is to look at the three principal observational dimensions of its data, ground resolution, land coverage frequency and spectral coverage. They are tied together in sometimes unfortunate ways, (from the user's point of view), by the laws of optics, orbital mechanics and the ultimate decision maker, economics. No one system can provide all the measurement features needed by the user community.

This review presents three summary maps of the data scope and variation which will be provided by the 31 satellites; land coverage and ground resolution, the spectral position of measured bands and the ground resolution of each band.

Land coverage and ground resolution: All but two of the satellites will cover the total land mass since they are in polar sun synchronous orbit. The two exceptions are SPIN-2 which is in a 65 degree orbit and QuickBird for which a 52 degree orbit inclination has been chosen.

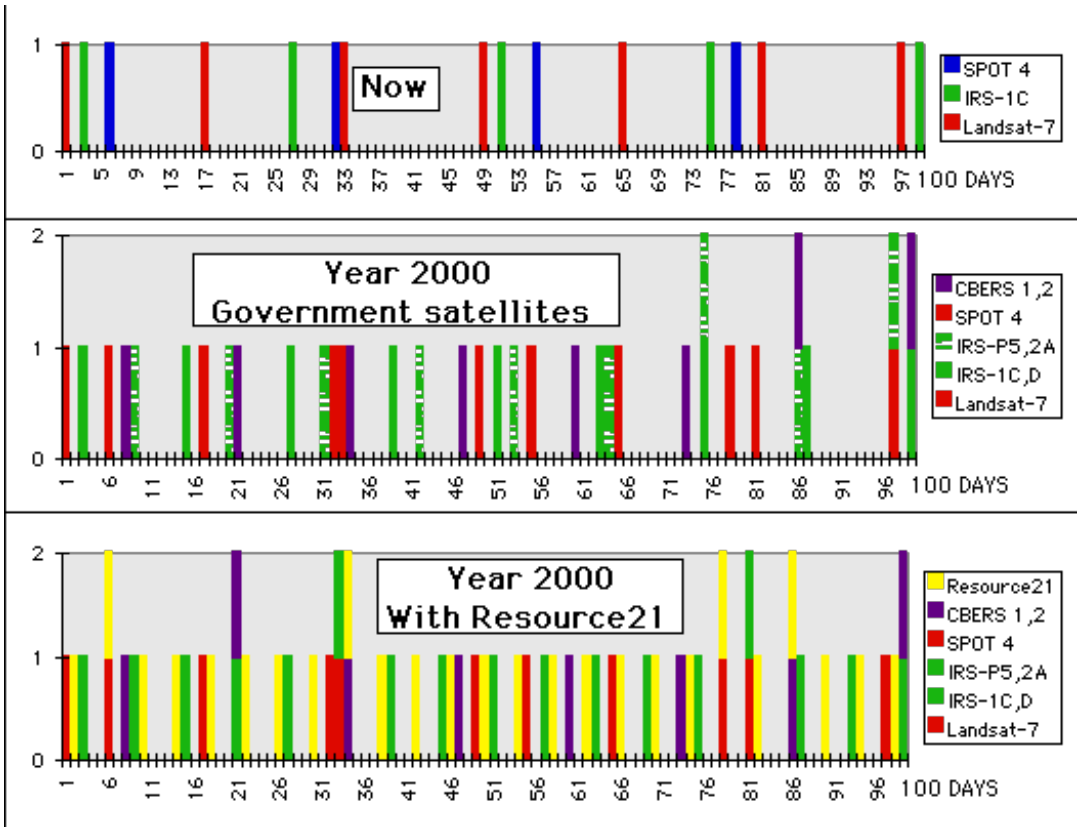
Land coverage frequency must be considered in two ways: the frequency with which the system can provide images of the total globe, and the time it takes to revisit a given site. Because global coverage frequency is inversely proportional to the sensor's ground field of view or swath width, this parameter will be presented as one measure of coverage capability in the following graph which presents the ground resolution and the ground swath width for all of the satellites noted above.



The above plot provides a graphic illustration of the difference in coverage and resolution between the four classes of satellites. Besides the radar satellites, two satellites fall outside the boxes: The IRS C,D Pan sensor flies on a satellite that is in the Landsat-like box, but lies away from that category because it sacrifices swath width for its higher resolution. However, it can be pointed off the orbit path which allows 2 to 4 day revisits to specific sites. SPIN-2's escape from its box is described below.

The next illustration triplet shows with color bars the repeat frequency of some of the major satellites. Now refers to 1997. These show the number of the Landsat-like satellites that you could see (or more precisely the number of satellites that could see you at an equatorial [1; refers to notes at bottom of this page] site) on any day over a randomly selected 100 day period

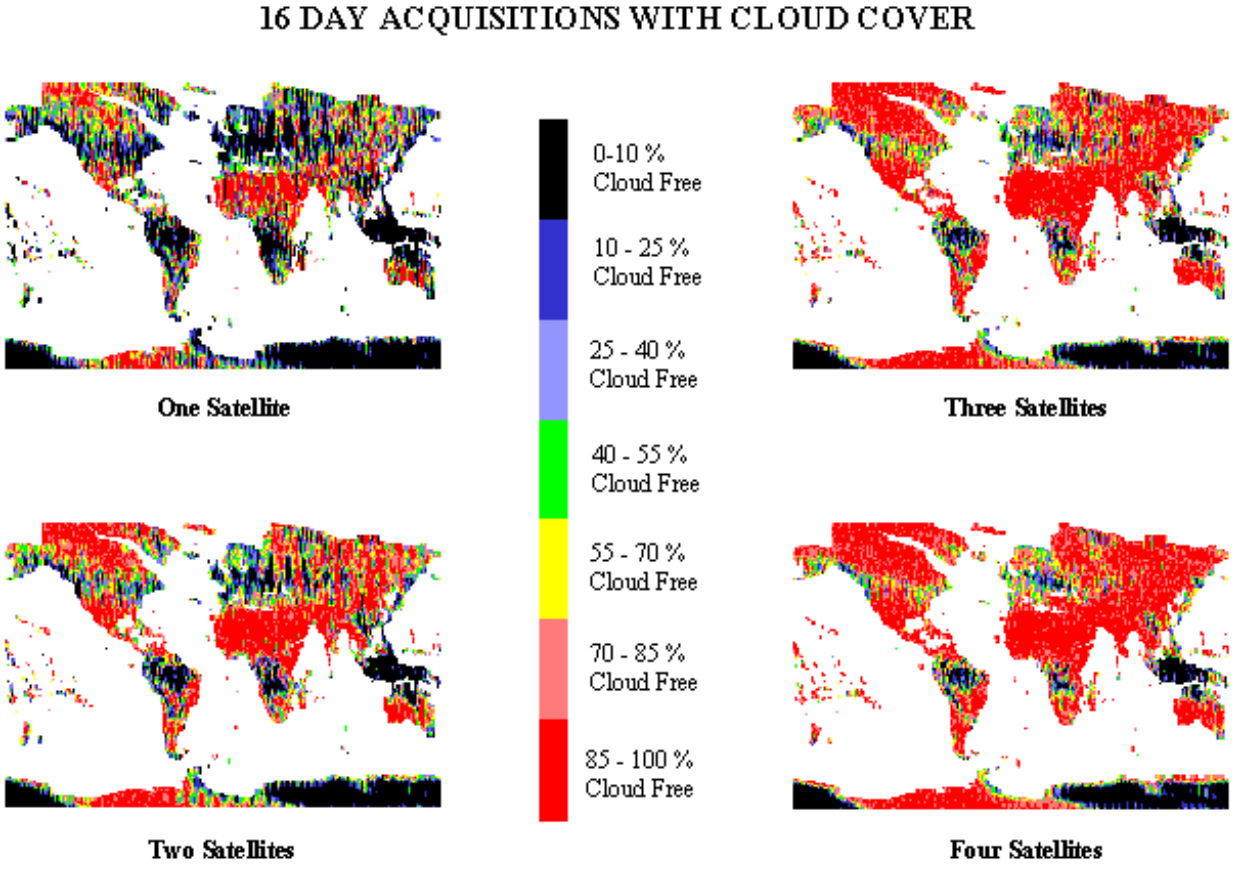
for the three satellites now in orbit, for the 8 government satellites in orbit in the year 2000, and for the 12 satellite fleet [2] resulting from adding the four Resource21 birds. The aperiodic nature of the second plot cries out for effective international cooperation to optimize the spacing of the coverage opportunities. (There is no indication that this is likely to happen).



The Landsat-type satellites are designed to provide fairly frequent global coverage by choosing the sensor ground swath and orbit parameters so that they will cover the complete equatorial surface each orbital repeat cycle. The current and planned satellites achieve this by having ground swaths between 120 and 200 kilometers. Their orbital periods and thus global coverage times, vary from 16 days for Landsat to 22, 24, and 26 days for the Indian, French and China/Brazilian satellites. Taken singly, even these repeat cycles are too long for many applications. However, the similarity of the sensor data from the 12 satellites in this group can, for many applications, make it possible to use the data from all of the satellites interchangeably and thus have available the one to seven day coverage rates illustrated in the figure above.

The High Resolution Group: In contrast to the Landsat-like group, half of this group has limited multispectral coverage, while the other half has none at all. It is obvious that as a group the critical measurement is the ground resolution which is essential for identifying man-made objects and for updating maps and GIS data bases. Whereas in the Landsat-like group the pan bands are used to sharpen the color bands, in this group the color bands will probably be used to add additional information to the pan band data. The much narrower ground swaths of the high resolution sensors, 4 to 36 kilometers, can only achieve total global coverage in periods ranging from 4 months to 2 years. Since the high resolution sensors being planned generate communication rates between 20 to 100 times that of Landsat this design limitation is caused by the practical and economic limits of the data collection systems. SPIN 2 avoids this problem since its data collection system is film return which places it in its unique position on the chart. However, for many users the good news is that the satellites are designed to be capable of quickly pointing off nadir and thus can see any given site in 2 to 4 days. Thus, even two high resolution satellites properly synchronized could provide daily repeat coverage nearly anywhere.

WARNING-Clouds severely effect the above quoted repeat times:

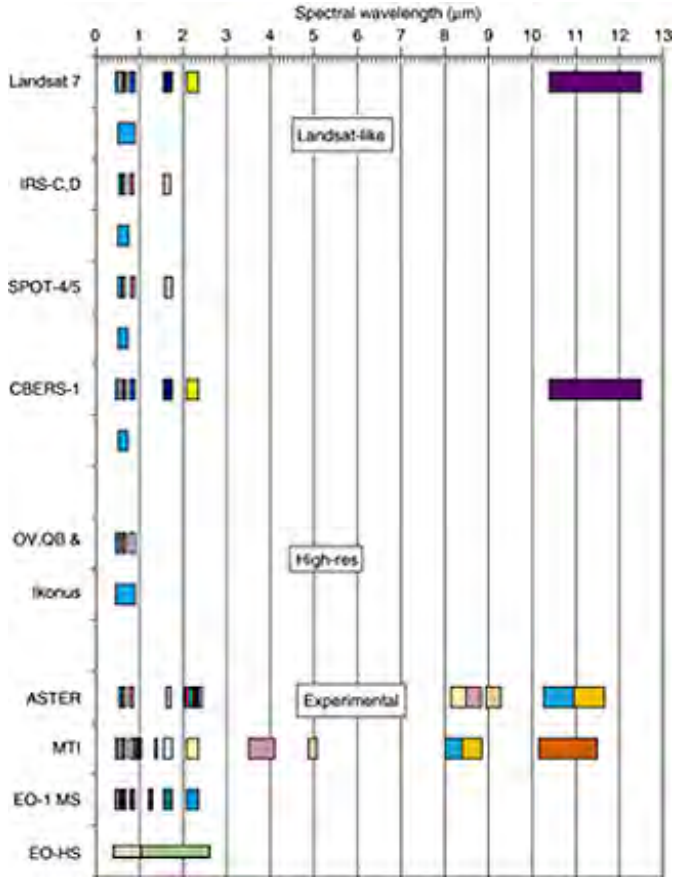


The above discussion of the repeat times should not be used without at least doubling the numbers quoted to provide some sense of the effect of clouds on the actual ability to get cloud free images, i.e. to actually see the desired targets. The above illustration presents the results of a simulation which recorded the best cloud free percentage images for each Landsat WRS site [3] collected over a 16 day period in early spring by using one, two, three and four satellites orbiting over a WRS grid containing the % of cloud cover in 5% increments for every day of the year [4].

The fact is obvious, our planet is habitually cloudy and the clouds obstruct satellite land views more than we would like for many of our time critical applications. The message is equally plain, multiple satellites (or radar) are required if we need assured land coverage in short time periods of weeks to months. (Note the four maps also represent very closely the collection capabilities of one satellite for 16, 32, 48 and 64 days.) As the maps make evident, the problem is geographically focused and as would be expected the agricultural belts, where frequent data are most required, are the cloudiest. It remains to be seen whether the small target areas and pointability of the high resolution systems will provide higher cloud-free data returns than those calculated for the large area non-pointing systems illustrated above.

Radar: The current and proposed radar satellites can provide data in a variety of resolution, swath combinations. The values on the figure represent their high resolution capabilities. Again, the practical limits on data rate have been an important factor in their resolution/swath tradeoffs.

Spectral Coverage and Ground Resolution: The illustration below provides the resolution of each band on a number of satellites (not all included in the previous illustration). (The bands are listed under their Landsat 7 band counterparts).



Many of the bands on the above satellites are close to those used by Landsat. This is of course because the Landsat bands were placed in nearly all the wavelength windows free of severe atmospheric absorption. The Landsat-like satellites emphasize multispectral coverage, all of them having at least the lower SWIR band and three including both the upper SWIR and TIR bands. ASTER will provide even greater spectral definition in the upper SWIR and TIR regions.

It is important to note that all multispectral data may not be equally usable for all applications even when the same bands are available. For analysis that are critically dependent on measuring the absolute reflected radiation over years to decades, sensor calibration becomes a critical parameter. Landsat 7 and Resource21 systems will have Sun and Moon based calibration capabilities while the other systems will rely on internal lamps and ground targets for their calibration. Equally important to such applications is the ability to adjust the measured radiation for the varying atmospheric conditions. NASA is planning to operate Landsat 7 and AM-1 (Terra) in very close proximity to measure the atmospheric input using an AM-1 sensor (MODIS).

As shown in second figure above the multispectral resolutions range from 10 to 30 meters with the exception of the 6 meter sensor on IRS-P5 and 2A which achieve their higher resolutions by reductions in swath width. The panchromatic sensors of interest in this group range from 6 to 20 meters. Experience with integrating the 10 meter panchromatic data and the 20 meter multispectral data from SPOT has shown the value for many applications of the use of the pan band in sharpening the color bands.

The Hyperspectral Group: The Hyperspectral Group: The US government is launching several satellites to test the full potential of multispectral analysis for the identification of both man-made and natural surface elements. Because of the very high data rates required by the hyperspectral sensors, the resolution of these systems has been restricted to 30 meters. There is also a sense that 30 meters may well be more than sufficient to characterize the majority of at least the natural targets, i.e. mineral and vegetative cover. The Australian government is stimulating interest in the private sector for the commercial development and operation of a near hyperspectral system, since the sensor uses two groups of 32 bands instead of the spectrometers of the other systems. The hyperspectral satellites are being flown to explore the potential of using the full spectral response over the VNIR and SWIR spectrum. Note on the last

figure above that hyperspectral is being defined as sensors with 32 to 256 bands per VNIR or SWIR range.

Radar: While the current and planned radar satellites will have only one frequency, they do have several polarization options and thus have a multidimensional analysis possibility analogous to the optical system's multispectral analysis.

Data Availability: The good news is that there are plans for more than 30 satellites capable of providing a wide range of land data information products. The amount, sophistication and variety of the land data that could be available for analysis is staggering. It is probably equally good news that all the data, government and private, except Landsat 7 data, will be available commercially at market determined prices. It's even better news that current US law requires that Landsat 7 data be made available to all at the " cost of furnishing user requests". However, to be available the data must be first acquired. Landsat 7 is the only system which plans to acquire and archive multiple total land cover data sets each year. The other government systems will be collecting for their own purposes and for orders acquired by their commercial sales outlets. The private systems' acquisition plans will be totally market driven.

Why So Many Satellites?: The large number of planned satellites may seem to be more than a few too many for needs of the Earth observing community. Before making that judgment however, it may be useful to consider the following points.

As noted above, none of the planned satellites will provide all of the data characteristics needed by the broad range of user requirements. Thus at least four systems would be needed to provide the different data types the fleet is currently planning. The day of the Battlestar Galactica - single satellites with suites of many instruments - appears to be over [5].

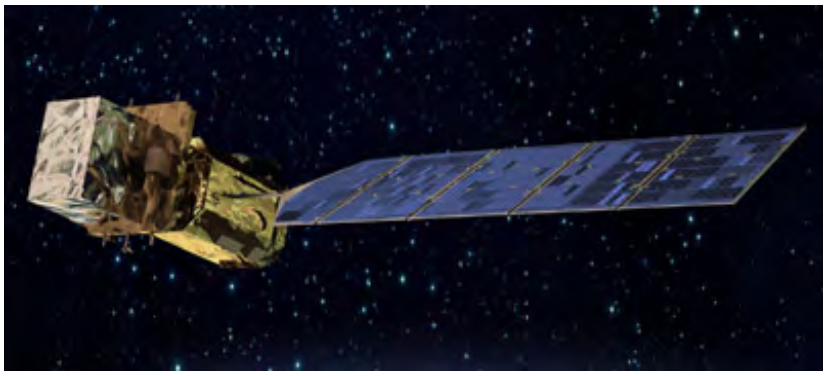
The need for multiple satellites was also discussed in the subsection on coverage frequency, which emphasized the negative effects of the world's (on average) 50% cloud cover. Resource21 is planning a four satellite system to meet their customers need for weekly observation of crop conditions. The Global Change Science goal of global of seasonal coverage requires a minimum of three to four satellites. The use of satellite data for disaster analysis and relief planning can be very effective but only if the satellite can acquire imagery almost immediately after the event, a possibility only if two or more pointable sensors are in orbit. For weather related disasters, radar is often the only system which can see the ground. Again multiple radar satellites would be required for sufficiently rapid coverage.

There is also the need to assure operational stability. Several land observation satellites have failed to make orbit - for example Landsat 6 and SPIN-2 - and two failed prematurely after attaining orbit - SPOT 3 and ADEOS. Obviously more than one system must be available to provide the operational assurance required if users are to be able to make the data a requirement for their activities. India, EarthWatch and Resource21 are planning operationally robust systems of four satellites each. CBERS and all of the high resolution satellite providers are planning two systems each.

As of late 2010, Landsat-5, 22 years in operation, is still functioning but is expected to cease usability at any time. Landsat-7 also is working but its images are somewhat compromised by scan line problems. In 2006, NASA announced an agreement with the U.S. Geological Survey to make certain U.S. scenes made by Landsat-7 available for download at no cost to the user public. Some information pertinent to this is to be found at this USGS website.

In retrospect, the Landsat series has proved to be and still remains the "creme de la creme" of earth-surface observing satellites. Despite the ever-growing competition, the Landsats are the prime systems for filling a vital niche: wide area coverage with multispectral sensors that provide data well suited to a variety of image processing procedures, including classification. The remote sensing community has repeatedly called for a policy that guarantees continuation of Landsat-style coverage. Commercial satellites are offering some appropriate data systems that provide

this. An unsuccessful attempt spanning several years to privatize Landsat forced a reversion to a joint operation by NASA and the USGS, as mandated by a 1992 congressional law. The archived data from all the Landsats, going back to 1973, constitutes a valuable data source for environmentalists, planners, scientists and others. The user community has called for the continuation of the Landsat program so that there will be an unbroken data set for this format. This has resulted in what is called the Landsat Data Continuity Mission (LDCM) which would involve some Landsat type system. Contracts have been let for Landsat-8, whose launch will be no earlier than December of 2012. Here is a picture of this satellite; below is a table that specifies its sensor capabilities:



Satellite	Band	Bandwidth, (nanometers)	Resolution, (meters)	Swath, (kilometers)
Landsat 8	1, coastal/aerosol	433 to 453	30	185
	2, blue	450 to 515	30	185
	3, green	525 to 600	30	185
	4, red	630 to 680	30	185
	5, NIR	845 to 885	30	185
	6, SWIR 1	1,560 to 1,660	30	185
	7, SWIR 2	2,100 to 2,300	30	185
	8, panchromatic	500 to 680	15	185
	9, cirrus	1,360 to 1,390	30	185

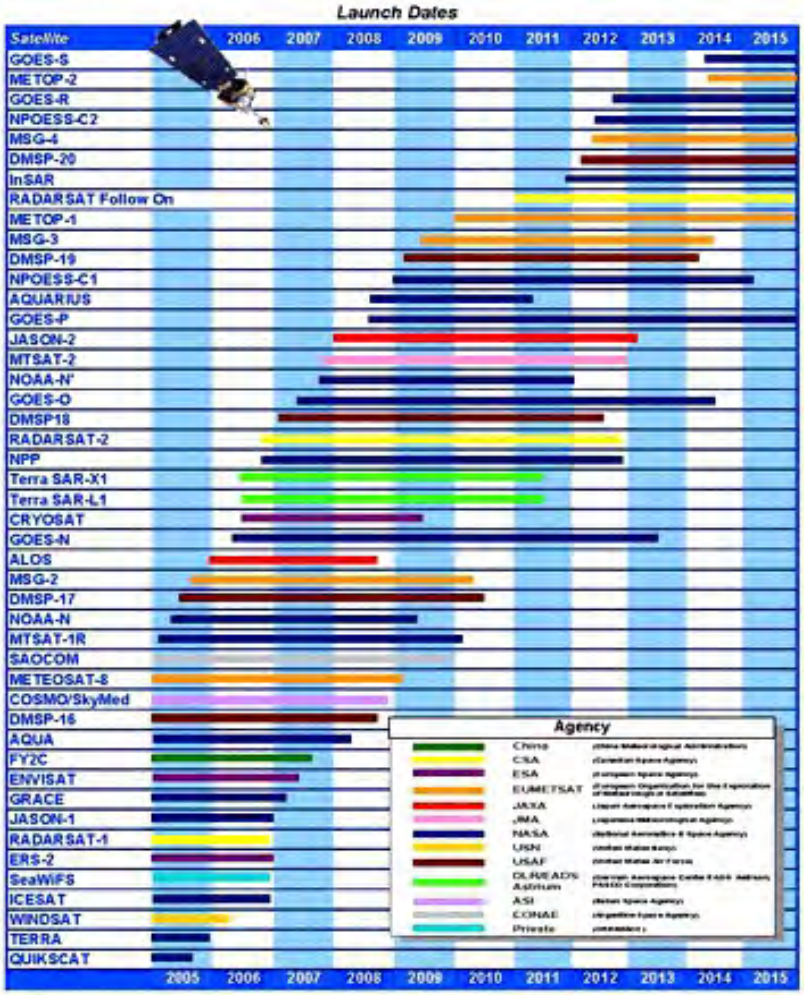
Some Further Thoughts

The goal of this Section is convince the land information user community and especially the so-called "value added" experts in industry and academia, that their cup is about to runneth over. The satellites are really coming, though probably not in the numbers presented in this Section. Roughly half are government funded and most of these are in or on the path toward construction. If only half the proposed commercial satellites make orbit there will be 20+ satellites in orbit by the mid-2000s.

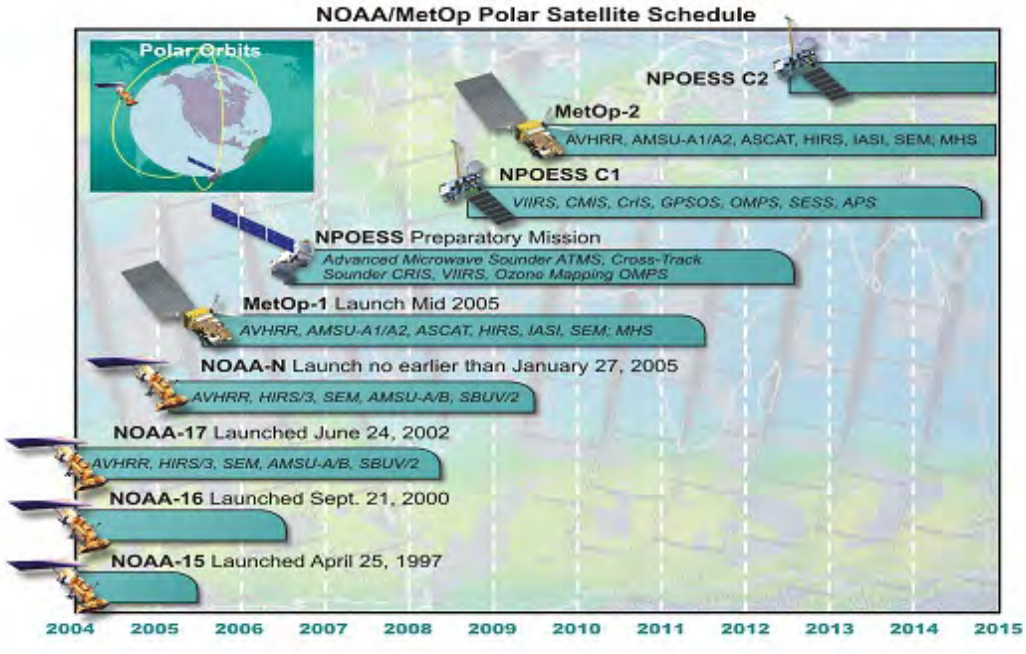
However, checking the above launch schedule diagram with actual history will reveal that most satellites, whether government-supported or privately funded, tend to slip their launch dates, a few fail to reach orbit or otherwise function, some may be renamed, some are abandoned or modified, and new satellites are proposed. It is hard to keep such charts current and accurate, as programs fluctuate or are revised. One way to stay updated is to visit the site maintained by SpaceFlightNow; this Tracking Schedule includes satellites of all kinds and purposes.

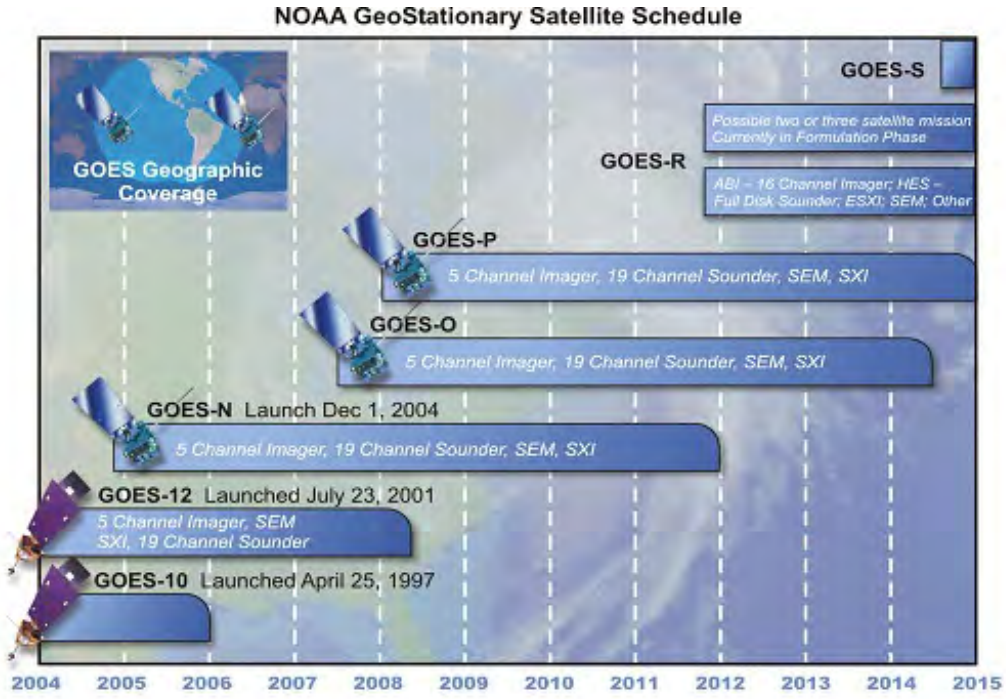
The primary writer (NMS) has attempted to keep track of current and planned missions throughout the Tutorial. This has proved a daunting task, largely because the sheer number of satellites being considered. In July 2007, the writer found this chart which, though comprehensive, is a good example of how obsolete and uncertain these overview summaries

can be:



This table includes some of the satellites used in meteorological applications. To highlight this group, here are two additional charts that specify operating and planned missions in this category:





The really big bucks, literally billions, required to create the earth-observing satellite systems are being spent by both the public and private sectors. It is now up to the users, public and private to invest in the development of the analysis technologies, the information products and the applications that will generate the dollars that will keep the new millennium satellites flying. The question is are you, they, anybody ready for the deluge?

Since the opening of the exploration of space by satellites, orbiting telescopes, and human-occupying spacecraft, hundreds upon hundreds of individual missions have been launched - most successfully. This will continue into the future. Elsewhere in the Tutorial, we have cited websites that describe past and current missions. One that summarizes many, but not all, NASA-related past, current and future missions, some pertinent to earth-observing satellites, can be found at this JPL website. The RST creator (NMS) has not found a single Internet site that gives a complete and comprehensive listing of **all** future planned (but not necessarily funded) missions. But here are three websites that each consists of a partial list: Welcome to the Future; RedOrbit (Space Missions in Development); and The Planetary Society.

Now, with this background move on to the next page which treats concisely what you have already been familiarized throughout the preceding Sections of the Tutorial: namely, the start of the Era of Commercialization of Space.

NOTES

1. Since polar orbits cross near the polls and have constant width ground swaths, the ground overlap between orbits increases with latitude. At 60 degrees the overlap is 100% and thus the equatorial coverage rate doubles.
2. I have taken the liberty in the third plot of assuming that the two Indian satellite series will eventually be planned to have the same orbital period in order to create the advantage of a total periodic period of 6 days in place of the aperiodic 11 and 12 day periods currently quoted (which also assumes that the two satellites of each series are placed in orbits halving their 22 and 24 day return periods.)
3. The maps are the world as seen by the constant swath Landsat image and thus are greatly distorted at the higher latitudes. The Landsat World Reference System (WRS) maps the world in 30,107 185x170 kilometer squares.
4. The WRS cloud data were created by the Air Force from a global data set for the year 1977 and represent the cloud coverage at the 9:30 AM local crossing time of the Landsat

satellite.

5. It is however worth noting that four of the Landsat-like systems, SPOT, IRS, CBERS and AM-1, do carry one or two other wide field of view sensors to provide daily to weekly coverage to supplement their main sensor data.

Primary Contact: Nicholas M. Short, Sr.

William E. Stoney (wstoney@mitretek.org)

Russian Space Program

Introduction:

During the late 1950s and 1960s, the dominant space power was the Soviet Union. Except for manned circumlunar, manned lunar landing, and exploration of the outer planets, all of the major space firsts were accomplished by Russia. These firsts included the first satellite (Sputnik 1, 1957), the first probes to hit and take pictures of the Moon (Luna 2 and 3, 1959), first interplanetary probe (Venera 1, 1961), first man in space (Vostok 1, 1961), first woman in space (Vostok 6, 1963), first multi-manned spacecraft (Voskhod 1, 1964), first walk in space (Voskhod 2, 1965), first softlanding on the Moon (Luna 9, 1966), first probe to orbit the Moon (Luna 10, 1966), first unmanned lunar soil sample return mission (Luna 16, 1970), first unmanned planetary rover (Luna 17/Lunokhod 1, 1970), first object on Mars (Mars 2, 1971), first Venus lander (Venera 7, 1970), and first Venus orbiter with lander and surface pictures (Venera 9, 1975). But this all came crashing down in January 1966 when the "Chief Designer", the soul of their space accomplishments, Sergei Korolev, died during an operation. Their space program never fully recovered, and they lost the race to put a man on the moon.

The problem of the Soviet space program was two-fold. First, whereas in the United States, NASA was the one central organization who was responsible for leading the race to the Moon, the Soviet Union had competing design bureaus which very rarely cooperated on such important ventures. Secondly, the Soviet Union did not have the financial resources to compete. The United States spent approximately $24 billion to get a man on the Moon. It has been speculated that Russia spent less than half that. When the race was lost, the Soviet Union denied that they had ever been in one. The truth came out during the era of perastroyka. But where there are endings, there are new beginnings. The Soviet program was re-oriented to further unmanned exploration of the Moon with their Luna series of planetary spacecraft, and the launching and maintaining of space stations in orbit. The 1970s would be a time of trial and error, but by the end of decade, two very successful space stations, Salyut 6 and 7, would be manned for long time periods. Russia would also continue its very successful exploration of Venus including the first soft landing on that planet.

Primary Contact: Nicholas M. Short, Sr.

The 1980's

Introduction

The theme of the 1980s could be "Shuttle to Space Command to Star Wars to Space Station (Part 1)." The 80s decade began with final ground and air testing of the Space Transportation System (STS), leading to the first manned shuttle launch in April 1981. But the high hopes of the American reusable manned space system were dashed in January 1986 with the destruction during launch of the Space Shuttle CHALLENGER. It took over two years for the civilian space program to recover from this disaster. But recover it did, and the shuttle system remained the heart and soul of the NASA program during this decade. Because of the budget drain to fund development of the shuttle, only two interplanetary probes were designed, built and launched in the 1980s. They were Galileo destined for Jupiter, and Magellan bound for Venus. NASA continued with its responsibilities for advanced spacecraft technology, space science and earth application satellites.

In the 1980s, American military launch vehicles fared no better than their civilian counterpart when within a six-month period, the Air Force experienced two Titan 34D launch failures between 1985 and 1986. But the 1980s could be called the beginning of the maturity of Air Force space system. With the ever growing capability of military spacecraft, though acquired and handled by Space Systems Division at Los Angeles Air Force Base (an acquisition command base), many of these space systems were operational controlled by many organizations. By the early 80s, the military establishment felt that a formal operational space organization was needed and should come into being. In 1982, Air Force Space Command was created. It was also during the 1980s in the Reagan administration that the Space Defense Initiative or SDI, commonly called Star Wars, began. It was to provide an anti-ballistic missile shield for the whole country. Many of the military space missions such as early warning, communications, nuclear detection, global positioning and weather either began, matured or were enhanced in this time period.

It was in the late 1970s and continuing into the 1980s that the Soviet Union found a focus for its space leadership which it has not relinquished to this day. This is in the area of extended stay manned space stations. Starting with Salyut 6 in 1977, continuing with Salyut 7 in 1986, and finalized with the MIR space complex, Russian cosmonauts have been continuously in orbit since then. Two more segments of the MIR space station were launched in the 80s, the Kvant and the Kvant 2 modules. Because the Americans had a reusable space shuttle, the Soviet Union attempted to design and build their own. It was launched twice in the late 80s, both unmanned. They continued their exploration of Venus (with their Venera series) and Mars (with the new Phobos series) and attempted a rather complex dual mission spacecraft, Vega, which went past Venus and then flew by Haley's Comet.

Europe continued to develop it own internal spacecraft and launch vehicle design and manufacturing capability. With the emergence of ESA, development of the Ariane family of boosters began. Launch of the first Ariane 1 was in 1979, and by the mid-1980s, ESA had an excellent foothold in the communications satellites launch business. With the CHALLENGER explosion and the hiatus of American boosters for two years, ESA and Ariane took control of the launch business and have not relinquished it. The mainstay of the present boosters, the Ariane-4, began design in the mid-80s with first launch in 1988. In the area of satellite design, ESA took the lead in developing a European weather system working in conjunction with the American

NOAA GEOS system. It also began design and manufacture of the French SPOT remote sensing satellite system, and took part in the Giotti interplanetary probe to Haley's Comet.

Asia continued its development of its own systems, which now included Japan, PRC and India. India began to flex its technological muscles and develop its own satellites and launch vehicles. Many Middle East countries also began purchasing commercial communications satellites from either the United States or Europe and operating them for the own regional use.

PREVIOUS RST NEXT

Primary Contact: Nicholas M. Short, Sr.

American Space Policy

Two Presidents, Ronald Reagan and George Bush, controlled American space policy in the 1980s. In 1982, President Reagan announced that there would be a new reinvigoration of the American space program. However, it would not be until 1984 that he would announce during the State of the Union Address that he supported the building of a new space station. He hoped that it would be finished by 1992, the 500th anniversary of Columbus coming to America. But the funding problems that continually plagued the Space Shuttle also followed the space station. NASA was prepared for this having learned from its STS lessons. The space station could not constitute a large percentage of its budget since the White House would not support that. Because of the lukewarm support from Congress and supposed indifference by the White House, the space station would not take root and begin its assembly process until 1998 under a Democratic President, Bill Clinton. President George Bush made an attempt to jump-start a manned mission to Mars. One can see the various stepping stones to the universe (reusable shuttle, space station, manned Mars missions) starting with Hermann Oberth's vision of the 1920s, and von Braun's articles in the 1950s, and NASA big push under President Nixon. But it would take time to convince a President and the Congress. In the case of President Bush's recommendations for Mars, they fell on deaf in the Congress. There just was not any political support for it at the time.

Primary Contact: Nicholas M. Short, Sr. email:

American Civilian Space Program (NASA)

Introduction

This was to be the era of the reusable Space Transportation System (STS), or Space Shuttle for short. Indeed, the Space Shuttle was to capture the American public attention for the whole decade for better or worse. But with the growing budgets demanding by the shuttle, NASA was able to build and launch two several interplanetary probes, but still missed an opportunity to fly-by Haley's Comet. Space science and planetary physics made strong showing during this decade.

Manned

On April 12, 1981, 20 years to the day since the first man (Yuri Gagarin) went into orbit, the Space Shuttle COLUMBIA made its orbital debut, orbiting two days and testing out all systems. It was commanded by an astronaut veteran, John Young, who had flown on 4 other missions including a walk on the Moon during Apollo 16, accompanied by Robert Crippen, a rookie astronaut. The flight made it through with flying colors except for a possible scare when it was noticed that several heat shield tiles had come loose from the tail. But this proved to be no problem, and the COLUMBIA landed at Edwards Air Force Base in California within out incident to a crowd estimated at 20,000 people. COLUMBIA became the first reusable Space Shuttle on November 12, 1981, when it was launched into space for a second time. NASA and its shuttle was on its way. Five shuttle spacecraft were built, one test vehicle (ENTERPRISE) and four flight vehicles (COLUMBIA, CHALLENGER, DISCOVERY and ATLANTIS). But the shuttle had been first sold to Congress and the White House with the premise of low operational cost and easy access to space. The shuttle system was envisioned to fly 24 times a year. By early 1986, the whole STS fleet had only flown 24 times, and then January 28, 1986 came. The Space Shuttle CHALLENGER blew up 72 seconds after launch due to a failed O-ring in one of the solid rocket boosters. What made the tragedy worse was that the crew represented such a cross-section of the American community of sex, race and religious background. It had also included the Teacher-In-Space, and the whole country mourned their loss. The Rogers Commission was set up to investigate the accident and recommend changes.

More than two years after the explosion of CHALLENGER, the Space Shuttle Discovery was launched to recertify the solid rocket boosters for manned flight and restart NASA manned spaceflight program. Another Space Shuttle was ordered, the ENDEAVOUR to replace the destroyed vehicle. NASA tried to expand the scientific return from the shuttle flights and included ESA as partners in the program. ESA's contribution to the shuttle system was the Spacelab module that could fit into the shuttle bay, and be used for scientific or business experimentation. The Long Duration Exposure Facility (LDEF) was another unmanned satellite let loose by the shuttle and returned several years later to determine the effects on materials of long exposure to the space environment. The decade ended with NASA still tied to the STS, and being forced to expand the expendable launch vehicle fleet which the STS was originally planned to replace. But the promise of late 1969 to 1971 with the vision of greater manned exploration was still pursued, and the second building block of a Space Station was now actively pushed. There were jump-starts to get Space Station Freedom going, but NASA received the reception it got with the shuttle when it came to getting money for the project.

Whereas the STS only went through several years of indecision due to the financial merry-go-

round between the White House and Congress, the Space Station project would be tossed around for over 10 years before a final design was accepted and metal being cut. NASA to its credit, got ESA, Russia and Japan to join in the program, making it an international project. However, in making it an international project, NASA would lose some overall control of it. NASA saw the Space Station as its program of the 90s as STS was its program for the 80s. Even at the end of the 1980s, President George H. W. Bush proposed a Space Exploration Initiative (SEI) that would get America back to the Moon by the year 2000 and to Mars by 2010. But the price of $500 billion over 20 years was just too great and Congress never gave it serious consideration.

PREVIOUS RST NEXT

Primary Contact: Nicholas M. Short, Sr.

THE MODERN HISTORY OF SPACE

Guest Writer: Jim Rosalanka

This long Appendix, without illustrations, was prepared at the U.S.Air Force Academy, in part to provide a background for the Cadet students in the role of space technology, and especially remote sensing, in various applications including military surveillance. It is also therefore a valuable survey and review for civilians interested in both U.S. and international space programs from the 1970s to the present day. It is very worthwhile (but optional) reading for any who wish to learn more about accomplishments in major aspects of space applications in the latter third of the 20th Century. Individual pages will not be summarized here. Read the Table of Contents for an outline of the principal topics covered in its subsections.

Note: Since this was written, the principal developer of this Tutorial, Dr. Nicholas M. Short, has prepared a single page, Introduction Page 26e, that deals (using imagery) with one of the main topics in this Appendix: Military and Intelligence Community use of satellites for reconnaissance and surveillance.

A HISTORY OF THE SPACE PROGRAM

The author of this Section chose to begin his review of the space program in 1970. However, much happened prior to that decade. Some specific events and programs are mentioned as background in the following pages. But, to set the stage for the 1970s, the writer (NMS) has found a web site that outlines the U.S. and Soviet activities in space from 1957 through 1969. Click on these links to access this coverage: 1957-1960, 1961-1965, 1966-1970.

The 1970s

Introduction

With the close of the 1960s, the American space program had made great strides with both its civilian and military components. The National Aeronautics and Space Administration (NASA) had reached its goal, established in 1961 by President John F. Kennedy, of "landing a man on the moon and returning him safely to earth." The landing of Apollo 11 in July 1969 has gone down as one of the landmark events of the 20th Century.

The U.S. military in the 1960s had developed the working technology with accompanying satellites, and required procedures to conduct its assigned space missions of reconnaissance, communication, and nuclear detection. It also conducted research and development programs (R&D) in the areas of navigation and early-warning. The USAF also established all the working infrastructure for launching spacecraft into either equatorial or polar orbit, and commanding and controlling these satellites once they were in orbit.

NASA continued into the 1970s with the launches of Apollo 13 through 17. The Apollo Applications program, later known as Skylab, established America's first space station and demonstrated the potential for man's continuous presence in space. The initial thawing of the Cold War gave the U.S. and the U.S.S.R. the opportunity for the first cooperative manned venture into space with the Apollo Soyuz Test Project (ASTP). But not all of NASA efforts were focused on manned spaceflight. This was also the great era of interplanetary space exploration to the outer planets with the flights of Pioneer 11 and 12, along with Voyager 1 and 2, and the landing of a spacecraft on Mars with Viking 1 and 2 in July 1976.

U.S. military spacecraft continued to be either developed or enhanced during the 1970s. Most military missions assigned to the USAF, Navy, or the Army were on their way to initial operations or at least were through final development. Two important points became quite clear during this time period. First, Los Angeles Air Force Base (part of Air Force Systems Command) became for all practical purposes a defacto military operational command. It controlled the Air Force Satellite Control Facility which handled many military satellites through its operations of Remote Tracking Stations, or RTSs, throughout the whole world. It continued this important function for almost 25 years, starting in 1959, until the formation of Air Force Space Command in 1982. Second, the Navy pioneered the military space missions of tactical communications and space based navigation through its operations of the Fleet Satellite Communications Systems, known as FltSatCom, and Transit satellite system.

For the Soviet Union, the 1970s were a period of disappointment, but also a time of new beginnings. Russia had lost the manned race to the moon, both circumlunar and lunar landing. Its national pride was hurt, and to the world, it declared that it never was in a space race at all. Twenty years would pass before the truth finally came out. In the 1970s, they continued their unmanned lunar program with the Luna series, which included the first remote control rover on another planet called Lunokhod. Their overall space program was redirected towards the building and operating of a space station, and the enhancement of their manned Soyuz spacecraft. The Russians launched and manned the first earth orbiting space station in 1971, Salyut 1, but tragically, the crew perished during reentry after living and performing in space for almost 24 days. By the end of the decade, five more stations would be launched, three military and two civilian, with the end result that on-going operations were established by the Salyut series. In their planetary program, the Soviet Union was the first nation to soft-land on the planet Venus with the Venera series.

PREVIOUS RST NEXT

Primary Contact: Nicholas M. Short, Sr. Appendix Author: Jim Rosalanka (jrosolanka@worldnet.att.net)

American Military Space Program: Initial Military Operations

Introduction

The two great activities of the 1980s for military space was the formation of Air Force Space Command (AFSPC) in 1982 and the initiation of the Strategic Defense Initiation (SDI) under President Ronald Reagan in 1984. Also with the advent of AFSPC, Air Force war planners began taking a more serious and thorough look at the tactical uses of space based systems.

Photo Reconnaissance

Intelligence missions continued into this decade but their missions remain classified to this present day.

Anti-Satellite Weapons (ASAT)

One type of weapon geared to negating the use of space systems is the anti-satellite weapon. During the 1960s, the USAF developed a ground based system called Program 437 which consisted of a Thor Intermediate Range Ballistic Missile (IRBM) with a nuclear warhead. The idea was to launch the missile in the direction of an enemy satellite and detonate the nuclear warhead. The system worked quite well except that although it would destroy an adversary's space system, it might also destroy your own systems. The Air Force deployed Program 437 on Johnston Island in the Pacific, and it remained operational until 1970. In 1975, the Air Force began development of an advanced anti-satellite system. The new system would use a two-stage rocket with a miniature infrared homing sensor device and be launched from an F-15. It would then home in on the target using the infrared sensor and destroy the enemy satellite by impacting it. The first free-flight test took place in 1984, and it succeeded in destroying an American military technology satellite in 1985. However, the program was canceled in 1988 due to budget constraints and Congressional restrictions.

Weather (Meteorology)

The military weather satellite system, DMSP, continued to evolve and improve. There were four launches of a new class of meteorological satellites, the 5D-2 model, during the 1980s using a new launch vehicle, the former Atlas E ICBM. These former ICBMs were refurbished to be medium weight launch vehicles.

Communications

The Navy continued to be the pioneers in tactical use of space based systems with its FltSatCom system. Eight satellites for this system were launched in the 1980s giving excellent coverage for tactical users. Because of the explosion for the need of military communications, the Navy realized that it own internal system was incapable of supporting all of its needs. It also needed a follow-on system to the FltSatCom satellites. Congress directed that the DoD should also make more use of leased commercial facilities and the Leasat satellites were designed. Five Leasats were launched between 1984 and 1990, and prepared the way for its own successor system, the UHF Follow-On, also called UFO. For the Air Force, they finished launching the last of its DSCS II satellites in 1982, and began deployment of its advanced DSCS III system. A total of

four DSCS III satellites were launched in the 1980s.

Primary Contact: Nicholas M. Short, Sr.

Russian Space Program

Introduction

During this period, the Soviet Union established it dominance in the areas of Space Stations and continuous manning of those stations. With the continuous manning of Salyut 6 launched in the 1970s, and the orbiting or Salyut 7 in 1982, and finally the establishment of the first segment of a long term space complex with the MIR system, Russia made the second building block to the stars - a Space Station. American had established the first building block with a reusable Space Shuttle. Russia attempted to fill the void of a reusable shuttle with its own Buran and Energia launch vehicle, but the collapsing Soviet Union could not provide the funds to complete that project. Russian continued with its own interplanetary program including sending more spacecraft to Venus and Mars, and also conducting a rather complex mission to Venus and Haley's Comet. Finally upgrades to various communications and weather systems continued.

Manned

Although Salyut 6 was launched in1977, the Space Station continued to be used through 1982 when it was deliberately deorbited. During that five years of operations, it was occupied by five long term crews, eleven visiting crews, and had been supplied by the Progress Transport spacecraft twelve times. It was also hooked to the Cosmos 1267 TKS module, the first time that two large structures had been remotely joined together in space. This was an envious record, but the best was yet to come. Russia wanted to establish a permanent presence in space, and this occurred with the Salyut 7 Space Station, and then establish a long term structure with the MIR complex. Salyut 7 was launched in early 1982 with the goal of permanent space presence and this was accomplished by changing crews. This crew replacement operation made continuous manned permanent orbiting Space Stations a reality. During Salyut 7's lifetime, it was occupied by six long term crews, and visited by four short-term crews. Also during this time period, it was hooked to Cosmos 1443 and 1686 TKS modules. The crowning jewel of Soviet Space Stations was the MIR complex. The base module was launched in 1986, and during the remaining part of the decade it was joined by two other modules, a Kvant and Kvant 2 segments. The complex was designed for up to three additional modules, and could be remotely supplied by Progress transport ships.

Also during this decade, Russia upgraded its basic Soyuz spacecraft from a "T" model to a "TM" model. It also upgraded the Progress spacecraft from the basic model to an "M" model. To complete the overall manned program, Russia wanted to replace its Soyuz manned spacecraft which had been in operations since 1967 with a reusable craft similar to the American Space Shuttle. They came up with the Buran shuttle attached to the Energia booster. There are definite similarities between the American and Russian reusable spacecraft, but there are definite differences. Both have delta wing orbiters, but the American STS has its main engines attached to the manned vehicle, while the Russian Buran does not. Both have a central core tank containing liquid oxygen and liquid hydrogen, but where the main engines are attached to the base of the American orbiter, the Russian engines are attached to the base of the core launch vehicle. Finally the boosters of the American STS are solid rockets, while the Russian Buran boosters are liquid fueled. This Russian boosters are dual purpose and can be used also as launch vehicles in their own right. So the Russian Buran does have some definite advantages that the American STS does not, even though they do look very much alike.

Unfortunately, the Energia / Buran combination attempted only two launches. The first launch in 1987 consisted of the Energia launch vehicle with a cargo compartment attached. The launch vehicle performed flawlessly but the cargo compartment separated badly. The second attempt in 1988 consisted of a Buran / Energia combination. The Buran was launched, circled the earth for one orbit, and returned and landed successfully by remote control. By this time, the Soviet Union economy was collapsing and could no longer support the development of Energia / Buran. The program was eventually canceled, and one of the Buran test vehicles can be seen in a Moscow park as a tourist attraction.

PREVIOUS rst NEXT

Primary Contact: Nicholas M. Short, Sr.

European, Asian, and Commercial Space Programs

Europe

During this decade, Europe broke away from its dependence on the United States for unmanned launchers, but still chose to cooperate with NASA on the STS system and future Space Station programs for most manned missions. After the CHALLENGER disaster, ESA with its Ariane family of booster vehicles took over the majority of the communication satellites launch business. ESA continued to design, develop and manufacture the boosters, but then gave it over to a new organization, Arianespace, to market its launch services. Europe showed its confidence in its capabilities by also launching its own designed and built interplanetary spacecraft when it sent the Giotti probe to flyby Haley's Comet in 1985. This was an event that NASA was unable to cover, and if it had not been for the Europeans, our understanding of comet activity would not have been as complete without the Giotti mission. France and Germany began to develop more sophisticated communications satellites, and France turned to earth remote-sensing as another application they could market. SPOT I was launched in 1985 in direct competition with the American Landsat series. At first, it did not have the resolution as Landsat, but as the SPOT system evolved, in many ways, it surpassed its American counterpart. Europe had definitely come of age in the 1980s, and a space power to reckon with.

Asia

Japan and the PRC continued to advance their own respective space industries. Two more Asian nations joined the space club by building their own satellites and launching them with their own booster, India in 1981, and Israel in 1988. Japan showed the maturity of its industry by building and launching its first interplanetary spacecraft, Sakigake, in 1985 and sending it to flyby Haley's Comet along with the Russian Vega series and the ESA Giotti probe.

Commercial

Evolution of the communications market continued, with many African and other Asian countries wanting to operate their own regional communication systems. There was a strong effort by the United States and France to develop the earth remote sensing market, but this did not take off as well as expected.

Primary Contact: Nicholas M. Short, Sr.

The 1990's

Introduction

The theme of the 1990s was and is the use of military space based assets in "the First Space War" and Space Station. The Kuwaiti War of 1990-1991 showed the force multiplier capability of military satellite systems to the world, and how space based combat support assets could help win a war in the area of early warning, global positioning and communications. Also during this time period, the International Space Station finally took root and began to be built. NASA's main manned program remained the shuttle, and with the collapse of the Soviet Union, their space program began to collapse as well for lack of funds. But the United States returned to Mars again with the Mars Rover and Surveyor, sent Cassini spacecraft to Saturn, and the Hubble Space Telescope continued to show the wonders of the heavens.

American Space Policy

A new administration took over in 1993, led by President William Jefferson Clinton, and showed their enthusiasm for the space program by quickly eliminating the National Space Council, an space advisory group to the President, which had been in existence since the Kennedy Administration. No new space initiatives were undertaken by the new President, but kept existing programs going, like the Space Station, planned planetary probes to Mars, and other space science and earth applications satellites.

Primary Contact: Nicholas M. Short, Sr.

American Civilian Space Program (NASA)

Manned

The heart of the NASA programs has remained its shuttle fleet of four manned orbiters (COLUMBIA, DISCOVERY, ATLANTIS, ENDEAVOUR). By the end of the decade, the fleet will be almost 20 years old, and have flown over 150 missions. Although the shuttle was never able to meet it planned expectations of low cost and ready access to space, it was and is a marvelous manned technological machine that proved that a vehicle could be reusable and is laying the ground work for follow on manned spacecraft. The building blocks to space have always been a reusable shuttle, a Space Station, and manned expeditions to the Moon, Mars and beyond. NASA was finally able to initiate, design, build and start assembly of its International Space Station.

Part of the agreement with Congress over beginning the Station program was to bring in the international community to help defray costs, and this NASA was able to do. There are almost a dozen international partners on the program including Japan, ESA and Russia. In fact, Russia is supplying two of the first three major components of the station based on MIR designs and technology and the country's experience in operating Salyut and MIR for over 20 years. Russia will also supply Soyuz spacecraft to act as emergency escape vehicles for the station early on in the program. Assembly began in 1998 with the launch of the first component from Russia, and then in 1999 with the first American component. Anticipating the international Space Station, the United States began sending American astronauts to the MIR Space Station complex.

Interplanetary

Interplanetary exploration had a mild resurgence in the 1990s. Two probes launched in 1989, Magellan to Venus and Galileo to Jupiter, reached their destinations in the 90s. Magellan arrived at Venus 1990 and began radar mapping operations soon thereafter. Magellan showed that Venus was more geologically active than previous thought. Galileo arrived at Jupiter in 1995, and began exploration of the Jovian system. This included release of a probe into Jupiter's atmosphere and with precise orbital maneuvers, exploration of the many moons of this large planet. In 1992, NASA launched its first probe to Mars in over 15 years, the Mars Observer. Unfortunately, two days before reaching Mars, the Observer went dead when controllers at the Jet Propulsion Lab (JPL) attempted some pressurization tests for the propulsion and attitude control systems. NASA also sent the last great interplanetary probe Cassini to Saturn in 1997. It will reach Saturn in the next century. Two new probes were designed and launched in 1996 on Delta expendable boosters, the Mars Pathfinder and the Mars Global Surveyor. The country watched with delight when the Mars Pathfinder successfully landed on the Red Planet, and released its Mars Rover to roam over the Martian terrain. Finally in 1997, NASA returned to the Moon with its Lunar Prospector.

Science, Technology, and Earth Applications Satellites

NASA maintained a healthy space science and earth applications programs during this time period. These satellites include: the Combined Release and Radiation Effects Satellite (CCRES) launched to perform a one-to-three year study of the earth's magnetic field; The Gamma Ray Observatory (GRO), the second in a series of four NASA Great Observatories, is a platform hosting four experiments that detect the gamma ray emissions of cataclysmic cosmic events

(black holes, pulsars, quasars) during a 15-month sky survey mission; the Upper Atmosphere Research Satellite (UARS) was the first in a series of NASA environmental monitoring satellites; Extreme Ultraviolet Explorer (EUVE) was designed to study the heavens through the extreme ultraviolet (EUV) region of the electromagnetic spectrum; and Geotail was a scientific mission to explore the earth's geomagnetic tail and was cosponsored by NASA and the Japanese Institute of Space and Astronautical Science (ISAS). NASA also continued its advances in spacecraft technology with the Advanced Communications Technology Satellites (ACTS) which was a joint NASA, Defense Department and private industry venture.

PREVIOUS RST NEXT

Primary Contact: Nicholas M. Short, Sr.

American Military Space Program: Initial Military Operations

Introduction

This period will be best remembered as the time when military space systems were put to the test in actual warfare. Most of the systems performed quite well, and some beyond expectations. Truly the Kuwaiti War can be considered the first Space War. Many systems were enhanced or being ready to be replaced. Also, some classified programs officially were declassified.

Photo Reconnaissance

In 1992, the existence of the National Reconnaissance Office (NRO) was officially announced. The NRO had been in existence since 1961 and the government determined that it was time to properly acknowledge this fact. Also in 1995, President Clinton declassified the CORONA program which was our nation's first photo reconnaissance satellite system.

Weather (Meteorology)

The DMSP 5D-2 model continued to fly during the 1990s but will be replaced by the 5D-3 model at the end of the decade. Also the government decided that the civilian and military weather satellite systems should be combined into one national resource. This is progressing on schedule.

Communications

The Navy began launching and operating the replacement for its FltSatCom satellite constellation, the Ultra-High Frequency Follow-On (UFO) satellites. A total of seven UFOs have been put into orbit to date. The Air Force was finally able to declare its DSCS III constellation fully operational with the launch of the fifth satellite in the series in 1993. As part of President Reagan's program to upgrade the country's strategic forces, the Air Force designed a satellite that would not suffer prolonged blackouts from high-altitude nuclear detonations using Extremely High Frequencies (EHF). This system was the Milstar communications satellite. Development of this system took over ten years, and the first launch was in 1994. The biggest supporters of the Milstar system are the tactical commanders in the field, our country's CINCs.

Early Missile Warning

The Air Force continued to launch the DSP-1 model of this early warning satellite, and it has performed beyond expectations. But due to changing technologies and changing threats, the mission for DSP was greatly expanded. The original mission for DSP in 1970 when the program was first started was the detection of strategic missile launches and nuclear explosions. By 1990, that mission now included the detection of tactical missile launches. The Air Force decided to acquire a totally new system with more capability. During the 1990s, the new proposed systems started out as the Follow-On Early Warning System (FEWS), but became the Space Based Infrared System -High (SBIR-Hi). This new satellite is planned for first launch in 2004.

Navigation (Global Positioning)

By the beginning of the Persian Gulf War, the planned GPS constellation was not totally operational. GPS attained initial operational capability in 1993, and full operational capability in 1994. From 1991 to 1994, there were 14 launches of the GPS II-A models navigation satellites because they had proven their worth in combat. During the latter part of the decade, the GPS II-R and II-F models were planned for and acquired. The first II-R model was launched in 1997.

Persian Gulf War

The war in Kuwait is called the first Space War. All space based assets were utilized tactically by all ground, naval and air troops. DMSP help in weather forecasting to determine ground sorties for planes, DSP detected the Scud missile launches and provided that information to Army Patriot batteries, and DSCS III communications satellites provided for 84% of the long-haul communications needs. But by far the satellite that proved itself in combat was the GPS satellites. They were used to guide Air Force and Navy aircraft to their targets, Army special forces helicopters depended on GPS to get them were they needed to go, and GPS receivers guided the movement of troops over hundreds of miles of featureless desert during the 100 hour war. From now on, no American military force could fight without the use of military space based assets.

PREVIOUS rst NEXT

Primary Contact: Nicholas M. Short, Sr.

Russian Space Program

Manned

With the demise of the Soviet Union, the Russian space program was now controlled by three Republics of the Commonwealth of Independent States (CIS), Russia, the Ukraine, and Kazakstan. Russia had all the training facilities and command and control centers, the Ukraine built many of the booster rockets, and Kazakstan contained the most important launch center, Baikanour. Russia had maintained the MIR Space Station complex attached to the Kvant and Kvant 2 modules through the 1980s. During the 1990s, even with limited funds, the total planned MIR complex was completed by launching the remaining modules, Kristall, Spektr, and Priroda. The Kristall expansion module is dedicated to materials processing, the Spektr modules is to be used for remote sensing, and the final module, Priroda, would be used for microgravity experiments and remote sensing for international users. MIR was suppose to last for 15 years, but due to budget limitations and Russia's financial obligations to the International Space Station, MIR may be deorbited before the end of the decade.

Interplanetary

Russia attempted one interplanetary spacecraft during the 1990s, Mars 8. Unfortunately, it ended up in the Pacific Ocean after an upper stage malfunction.

Science, Technology, and Earth Applications Satellites

Russia continued its programs for remote sensing with the Foton and Resurs satellites. It also launched a large scale commercial remote sensing spacecraft called Almaz 1 using radar imaging instead of photographic or infrared imaging similar to Landsat or SPOT. It had been constructed from the abandoned parts of a planned military Space Station. All of its communications and weather satellite program continue to be replenished. It has branched out into the unmanned materials processing arena with the Foton series. It has tried to mirror some of NASA Great Observatories series of satellites with its own spacecraft called Gamma used for observing gamma rays. Finally Russia joined the international weather monitoring group by supplying a geosynchronous Elektro 1 satellite.

Primary Contact: Nicholas M. Short, Sr.

American Space Policy

When the foundation of the infant American space program was first considered in the mid 1950s, President Eisenhower initially wanted control of all programs in the hands of the military. However, the chairman of the President's Science Advisory Committee, James R. Killian, and then Vice President Richard Nixon, strongly recommended the establishment of a separate, and open civilian space agency. With the legislative help of Senator Majority Leader Lyndon Johnson, the National Aeronautics and Space Act of 1958 was passed, and NASA officially opened its doors for business on 1 October 1958. Hence, all American space activities were divided into civilian and military programs. Since early American space policy was driven by military requirements, by the end of the Eisenhower Administration, these space activities were further split into civilian, military, and reconnaissance missions. Early in the Kennedy Presidency, his administration codified the Eisenhower organizations with the civilian program controlled by NASA, the majority of the military missions managed by the USAF, and the reconnaissance program in the hands of the newly established National Reconnaissance Office (NRO). Air Force Undersecretary Joseph Charyk was named as the first Director of the NRO. When President Nixon took office in 1969, he kept intact the same national space setup, and all subsequent Presidents in the 1970s (Ford and Carter), also maintained this organizational status-quo.

Two major space policy decisions came during the Nixon Presidency (1969-1974) . These were the go-ahead to begin development of the Space Transportation System (STS), or the Space Shuttle as it was commonly referred to, and the initiation for the flight of the first U.S./U.S.S.R cooperative manned space flight, which became know as the Apollo Soyuz Test Project (ASTP). The eminent NASA space historian, Eugene M. Emme, once observed that "No President was ever to be in office when major space missions happened as a result of his initiative or decision." President Eisenhower began the American space program, including the manned Project Mercury, but it was President Kennedy who was to bask in its glory of astronauts Shepard and Glenn. Presidents Kennedy and Johnson initiated and enthusiastically supported the race to the moon with the Apollo Project, but it was President Nixon who was to meet the Apollo 11 astronauts on the U.S.S. Kearsarge in July 1969. It was President Nixon who came closest to seeing the fruits of his space decision-making with the ASTP. The formulation of the Apollo-Soyuz Test Project began as one of President Nixon's global diplomacy initiatives starting in 1972 when he and Soviet Premier Alexi Kosygin signed a mutual agreement. ASTP flew in 1975 which Nixon would have seen during his second term in office, but for the blunder of Watergate.

It would be President Ford (1974-1977) who would see the first historic space handshake between Tom Stafford and Alexi Leonov and the other astronauts of Apollo 18 and cosmonaut of Soyuz 19 of Apollo-Soyuz. No major space decisions occurred during President Ford's period in office. He supported the continued funding of the development of STS, and if fact, was present when the first Shuttle, Enterprise, made its public debut. He also dedicated in 1976 the most visited museum in the Washington, D.C., the Smithsonian's National Air and Space Museum.

Although President Carter (1977-1981) initiated no major space starts, he did make a launch vehicle decision that would drastically affect the military in the mid-1980s. His Secretary of the Air Force, Dr. Hans Mark, ordered that all military spacecraft were to be launched from the Space Shuttle once it became operational. He curtailed and discouraged the manufacture of expendable launch vehicles, such as the Titan, Atlas, and Delta family of rockets. That poor decision would haunt the military when the Challenger exploded in 1986, along with several

Titan launch vehicles failures. This caused a major launch vehicle stand-down for almost two years, and no major military satellites were orbited during this time period. Also because of this, the French launch vehicle Ariane made significant inroads into the commercial launch vehicle business.

PREVIOUS RST NEXT

Primary Contact: Nicholas M. Short, Sr.

European, Asian, and Commercial Space Programs

Europe

Europe continued to master all aspects of advanced technology, space science and earth applications, and interplanetary satellite programs. Germany had launched a Roentgen Satellite (Rosat) carrying a large telescope/camera system to survey the X-ray sky. ESA built and launched from a shuttle the interplanetary probe Ulysses, a science mission to examine our Sun from an orbit that will take it over the north and south poles of the Sun. France orbited it several SPOT (SPOT 2 and 3) earth observation satellites to be used for commercial mapping and forestry, and agriculture and geological studies. Europe contributed to the world weather forecasting community by launching Meteosat 5 and 6 satellites, similar to NOAA's GOES system. Many second generation communications satellites were launched to support European commercial and military needs.

Asia

Japan, the PRC and India (IRS series) continue to expand their space capabilities in the areas of launch vehicles, satellite manufacturing, and satellite applications such as weather, communications, earth resources, and interplanetary probes. Two new nations began using communication satellites for regional use, Thailand and Pakistan. Israel continued to development its own space infrastructure for launch vehicles and satellite manufacturing. Satellites that Japan launched included Muses A on a lunar swingby mission with a lunar orbiter satellite attached (Hagoromo), Marine Observation Satellite IB (MOS 1B) to observe oceanic phenomena and earth terrain, Solar A which is a solar observation satellite carrying both soft and hard X-ray telescopes, Japanese Environmental Resources Satellite (JERS 1) which carries a synthetic aperture radar (SAR) and optical sensors, Astro-D which is an astronomy satellite carrying instruments to collect X-ray imagery, and finally the Geostationary Meteorological Satellite 5 (GMS-5) which joints weather satellites from NASA, ESA and Russia to form the worldwide weather forecasting network. The PRC capabilities continue to grow in the areas of communication satellites, earth resources, capsule recovery and launch services.

Commercial

Beside continuing into the commercial communications satellite market for other nations, business companies expanded into the commercial launch business. Such companies as Arianespace, PRC, Russia and the United States offer boosters for a price depending on the size of satellite that needs to orbited. Even an international consortium is developing a launch platform to be used at sea. A new field of expansion is the global positioning market which grew considerably after the Persian Gulf War. And finally, remote sensing is finally beginning to expand but its prime customers seem to be the intelligence services of various countries. Some of the American companies now actively funding and flying satellites include SpaceImaging, OrbImage, and EarthWatch.

WARNING: Appendix B is a very brief statement about the removal of the PIT software program. Clicking on the NEXT button below moves directly to Appendix C. To view the Appendix B page, access it through the Table of Contents.

Primary Contact: Nicholas M. Short, Sr.

American Civilian Space Program (NASA)

Introduction

During the first half of the 1970s, with public support waning, and cutbacks in funding and personnel, the American manned space program achieved great results. From Apollo 13 (April 1970) through Apollo 17 (December 1973), eight more astronauts walked on the Moon. Skylab began the U.S. venture into a permanent presence in space with three long-stay visits in 1973, which laid the foundation for the future International Space Station (ISS). Finally, international manned space cooperation started with ASTP, which would eventually lead to the mutually successful visits of the American Space Shuttle to the Russian space station Mir in the 1990s.

Manned

The 1970s quite literally started off with a bang for NASA with the now famous words "Houston, we've had a problem" from the Apollo 13 commander, James Lovell, the first man to go twice to the Moon. The launch of the planned third manned lunar landing began quite uneventfully on April 11, 1970. It was a sign of America's boredom with Moon flights that live transmissions from the Command Service Module (CSM) "Aquarius" were not carried live by any major TV networks. Less than one day from lunar orbit, an oxygen tank exploded blowing the side of the CSM, and causing loss of most electrical power and oxygen. The only thing that saved the astronauts was the oxygen and power reserves in the Lunar Module (LM) "Odyssey", now called a lifeboat. The lunar landing was canceled, the astronauts retreated to this lifeboat, rounded the Moon using the engines of the LM, and returned safely to earth on April 17, 1970. It would be almost a year before another lunar flight was attempted.

Apollo 14 was launched on January 31, 1971, commanded by America's first man into space, Alan Shepard, who had not flown since 1961 due to a inner ear problem. Because of his medical condition, Shepard was assigned as the chief of the astronauts corps under Deke Slayton. However, by 1969, he had returned to flight status, and using his powerful position, he was selected to command the Apollo 14 mission. This was much to the disappointment of many other experienced astronauts who had flown on previous Gemini and Apollo flights. His crew consisted of Stuart Roosa and Edgar Mitchell, both rookies. The assigned landing site was the Fra Mauro area, originally assigned to the Apollo 13 mission. Apollo 14 was the last "H" Apollo flights, which called for a limited stay on the Moon, and no excursions beyond visual range of the LM. Liftoff was uneventful, but after insertion into lunar orbit, the CSM had to make six attempts to dock with the LM. Once in lunar orbit, and after separation between the CSM and the LM occurred preparing for the lunar landing, a computer fault developed activating the abort system. After fixing that problem, the landing radar failed, but Stuart Roosa overcame that difficulty. Apollo 14 landed just 90 feet short of its intended target. All exploration was successful, and one of the more memorable events of the mission was a game of lunar golf by Shepard, hitting the world's longest drive in history. Apollo 14 returned to earth on February 9, 1971.

The last three missions of the Apollo lunar program (Apollo 15, 16, and 17) were called "J" missions. Upgraded spacecraft systems, improved life-support equipment and space suits, and the introduction of the Lunar Roving Vehicle (LRV), greatly enhanced the scientific return of all these flights. All of these improvements would permit lunar excursions of 20 miles versus 2-5 miles with the previous Apollo "H" missions. It was with Apollo 17 that the first Scientist-Astronaut, Harrison Schmitt, a trained geologist, finally went to the Moon. Apollo 17 was also the

only night launch of a Saturn V. Manned lunar exploration ended with the landing of Apollo 17 on December 19, 1972. The cost to send twelve men to the Moon was approximately $24 billion dollars. Some say that it was not worth the cost, but most Americans look back at this period of time with great and understandable pride in the great accomplishments of American know-how, ingenuity, and technology. It appears that it will be a long time before such a feat is duplicated.

America's first, and so far its only, space station was the Skylab missions of 1973. The prime idea behind Skylab was to utilize as much as possible hardware and technology from the Apollo program. Skylab had its early beginnings from an original proposal from Douglas Aircraft Corporation in 1962, calling for modifying a Saturn IB second stage into a orbital station. Once in orbit, the second stage would be drained, and made habitable. This was known as the wet concept. Another concept called for using a Apollo CSM / LM combination as a long-term orbital space station. In 1965, the Apollo Applications Program (AAP) was setup within NASA and in that same year, the go-ahead was given for the Orbital Workshop (OWS) as part of the AAP. In 1970 this OWS formally became Skylab, but with a new twist. Instead of using a spent Saturn IB second stage (wet concept), the space station would be launched by a Saturn V, and be launched dry. But all this seemed academic when the Skylab was launched on May 14, 1973, and problems occurred immediately. Apparently, 63 seconds after launch, the micrometeorite shield came loose, taking with it one of two solar panels. To make matters worse, the remnants of the micrometeorite shield had jammed the other solar panel. Skylab was a useless piece of space junk, orbiting without power or thermal protection. But the ingenuity of flight controllers on the ground, and the repair mission of the first Skylab crew saved the station. This provided for three successful long-duration space stays of 28, 59, and 84 days in space. The 84 day Skylab record was not broken until 1977 with the flight of Soyuz 26 to the Salyut 6 space station. The scientist-astronauts (chosen in 1977) proved themselves a extremely valuable and worthy partner in space exploration, because Skylab 3 Commander, Alan Bean, claimed "the flight was 50 per cent more productive having (them) aboard " in earth resources and space sciences research. A derelict Skylab made a final curtain call in 1979 when it reentered the earth’s atmosphere over Australia.

The call for international cooperation in space goes back to the earliest days of the space race. When the race was neck and neck, cooperative overtures were diplomatically rebuffed, but limited contact was maintained during the time period of 1962 through 1969. This was the political environment until Apollo 8 orbited the Moon in 1968 and Apollo 11 landed on the Moon in 1969. Again, President Nixon made overtures about a joint spaceflight in 1970 and the Soviet responded positively. A mutual agreement was signed in 1972. The original idea called for an Apollo spacecraft to dock with a Salyut space station, but this was changed to a linkup between a Soyuz and Apollo spacecrafts. NASA needed this flight desperately since no manned missions were planned for until 1978. The Skylab missions were scheduled for 1973 and the first Space Shuttle launches were not to occur until 1978-79 at the earliest (the first launch did not occur until 1981). Once the agreement was made, it took another three years (1972-1975) for the mutual docking device to be designed and manufactured (a Soviet design was accepted), and mutual operations to be agreed upon. It became know as the Apollo-Soyuz Test Project (ASTP). The American crew selected included a seasoned veteran, Tom Stafford, and two rookies, Vance Brand and Deke Slayton, although Slayton could hardly be called a rookie since he had been intimately connected with the American space program for 16 years when he finally flew on ASTP. This mission offered Deke, one of the original Mercury astronauts, the last opportunity to finally flew on a spacecraft. The Russians chose two veterans, Alexi Leonov, the world's first space walker, and Valery Kubasov. Soyuz 19 was launched first on July 15, 1975 and the Apollo 18 lifted off 7 hours later. The two spacecraft linked up two days later on July 17. They orbited linked together for three days before Soyuz 19 returned to earth. Apollo 18 stayed in orbit four more days. This was to be America’s last manned mission for over six years. For the rest of the 1970s decade, NASA's manned flight focus concentrated on designing, building, and testing the Space Shuttle system.

Primary Contact: Nicholas M. Short, Sr.

American Military Space Program: Initial Military Operations

Introduction

The practical beginnings of the total American space program had its roots in military space requirements. The first American military service to investigate the potential of satellites was the Navy in 1946. When the Air Force learned of the Navy activities, it asserted its rights to this new area as a logical extension of its air mission. During the late 1940s and 50s, the USAF directed the RAND Corporation to study the feasibility of satellite systems and associated launch vehicles. These studies identified the military missions of weather, communications and reconnaissance. Between 1953 and 1956, the Air Force initiated a "Satellite Component Study" giving it an official designation of Weapons System 117L or WS-117L, issued a system requirement and established a general operational requirement, and finally approved a development plan in 1956. The Air Force was on its way into space. WS-117L was envisioned to be a family of subsystems designed for different missions, including photo reconnaissance and missile warning. By the end of 1959, 117L had become three distinct programs - the Discoverer Program, SAMOS or Satellite and Missile Observation System, and finally MIDAS or Missile Detection Alarm System. The Discoverer Program and SAMOS were to be photo reconnaissance satellites, and MIDAS was for missile warning. In 1958, the Weapons System designation was dropped due to President Eisenhower's stated space policy of "freedom of the skies" and "the peaceful uses of space." Using WS nomenclature might be misinterpreted by the Soviets as a non-peaceful use of space. However, in the end, these three pioneering systems and associated launch/booster systems gave the American military space program its beginnings.

Photo Reconnaissance

The Discoverer Program officially had 38 launches, but in reality had many more. This program was a cover for the covert, intelligence-gathering reconnaissance satellite called CORONA, which was declassified in 1995. Its aim was to develop a film-return photo reconnaissance satellite using the Thor-Agena launch vehicle combination. There were a total of 134 launches between 1959 and 1972. Of these 134 launches, 102 were considered successful. In the 1970s, there were 8 CORONA missions, also known as KH-4Bs. These CORONA KH-4Bs each had two re-entry vehicles or RVs carrying photographic film, had an average mission life of 19 days, and carried out stereo photography of denied areas. In 1997, the Smithsonian's National Air and Space Museum opened up a display called the "Space Race" which showed a full scale CORONA KH-4B to the general public for the first time. However, the distinction of having the first ever public showing of an NRO system goes to the Schriever Exhibit at the Air Force Space and Missile Museum, Cape Canaveral, Florida in 1996.

Weather (Meteorology)

At the beginning of the American space program, NASA was given the overall responsibility of developing and operating a weather satellite system that could meet both civilian and military needs. NASA launched these first weather satellites, TIROS 1 and 2 in 1960. However, the NRO believed the TIROS system could not meet the strategic meteorological needs of the CORONA system. The first Director of the NRO (DNRO), Joseph Charyk, pushed for a separate weather satellite, and set up the Defense Meteorological Satellite Program or DMSP. The first DMSP

satellite was launched in 1962, and by 1964, four weather satellites were operational, still awaiting the full capability of the NASA system. In 1962, DMSP provided the National Command Authority (NCA) with critical cloud cover information on the Soviet Union and the Caribbean during the Cuban Missile Crisis. In 1965, the NRO gave total management responsibility of the DMSP program over to the Air Force, making the system an official military asset. The USAF has overseen the development of DSMP ever since. These satellites provided critical support to our troops during the Vietnam War, especially during air strike planning phases, saving lives and money. Most initial military satellites systems were designed for strategic use, but the DMSP system was the first spacecraft that both the Air Force and the Navy made tactical use of its weather data, primarily in strike aircraft mission planning. The Navy even put receivers on carriers to receive cloud cover information for its attacks in North and South Vietnam. These receivers have become standard equipment for the Navy since that time. The DMSP system was officially declassified in 1973. These satellites are polar orbiting, sun-synchronous spacecraft, with at least two in orbit at all times.

The DMSP system has gone through nine different design changes or upgrades through its operational lifetime, Blocks 1, 2, 3, 4, 5a, 5b, 5c, 5d-1 and 5d-2 - four during the 1970s. In fact, the last Block 4 satellite, Flight 23, was donated to the Museum of Science and Industry in Chicago, and is on display there. There were a total of 15 DSMP launches from 1970 to 1979, with 14 being successful. The launch vehicle for DSMP in the 70s was the Thor-Burner II and Burner IIA. In the beginning, these military weather satellites were designed to provide day and night visual cloud coverage. Now, other instrumentation now include finer resolution cloud cover sensors, temperature/moisture sounders, aurora detectors, and an infrared sounder. By the 1970s, the Air Force had two readout stations for the DMSP system in the Continental United States in the state of Washington and Maine, which feed this information to the Global Weather Central facility at Offutt Air Force Base in Omaha, Nebraska.

PREVIOUS RST NEXT

Primary Contact: Nicholas M. Short, Sr.

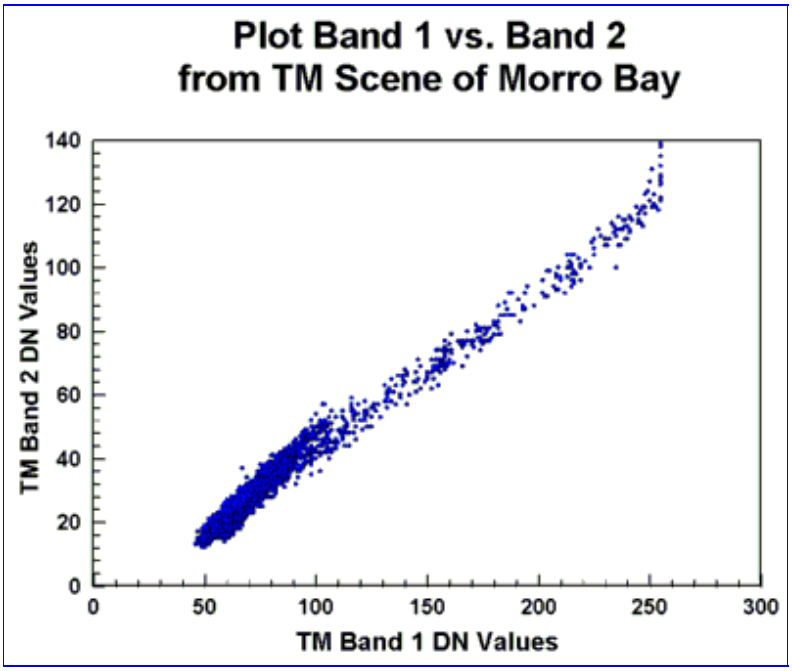
INTERACTIVE IMAGE PROCESSING

This Appendix, which dealt with Image Processing and introduced the PIT processing software, has now been removed from the Internet and CD versions of the Remote Sensing Tutorial. There were two reasons for this: 1) Feedback from users who tried to download and utilize the PIT program encountered difficulties with the process and failed to achieve a workable software system; no support was available from the NASA Goddard people, since their funding and authorization to work with the writer (NMS) had been withdrawn, so it proved impossible to correct the problems involved and develop a downloadable package; and 2) Over time, the Tutorial grew too big (in excess of 700 megabytes) to fit within the capacity of the CD writer that I use to make CD copies; by eliminating the copies of the PIT software and imagery in this Appendix, the size of the Tutorial has now dropped below 700 Mb.

PRINCIPAL COMPONENTS ANALYSIS: A BACKGROUND

Introduction

Principal Components Analysis, first introduced on page 1-14, is a procedure for transforming a set of correlated variables into a new set of uncorrelated variables. This transformation is a rotation of the original axes to new orientations that are orthogonal to each other and therefore there is no correlation between variables. The graph below shows a plot of band 2 versus band 1 of the Morro Bay TM scene. As you an see, the value of band 2 for a particular pixel is related to the value for band 1. The correlation is high.



Since the rotation is a linear combination of the original measurements, if all of the axes are included in the rotation, no information is lost. "No information is lost" means that the original measurements can be recovered from the principal components. If the original data set is singular, then principal components will produce a new representation that is not singular. There are several ways of viewing this transformation:

1. It can be viewed as a *rotation* of the existing axes to new positions in the space defined by the original variables. In this new rotation, there will be no correlation between the new variables defined by the rotation. The first new variable contains the maximum amount of variation, the second new variable contains the maximum amount of variation unexplained by the first and orthogonal to the first, etc...

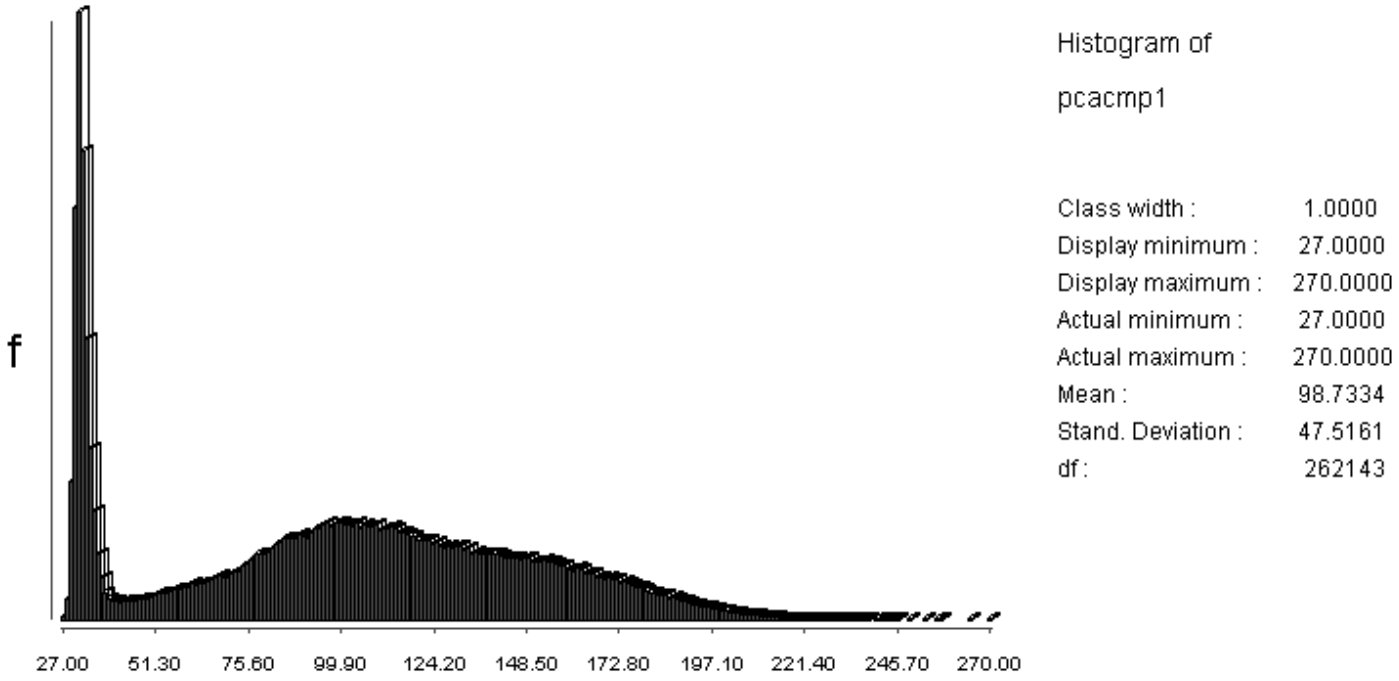
2. It can be viewed as finding a *projection* of the observations onto orthogonal axes contained in the space defined by the original variables. The criteria being that the first axis "contains" the maximum amount of variation, or "accounts" for the maximum amount of variation. The second axis contains the maximum amount of variation orthogonal to the first. The third axis contains the maximum amount of variation orthogonal to the first and second axis and so on until one has the last new axis which is the last amount of variation left. As you can see these are really two slightly different ways of saying the same thing!

There are several algorithms for calculating the Principal Components. Given the same starting

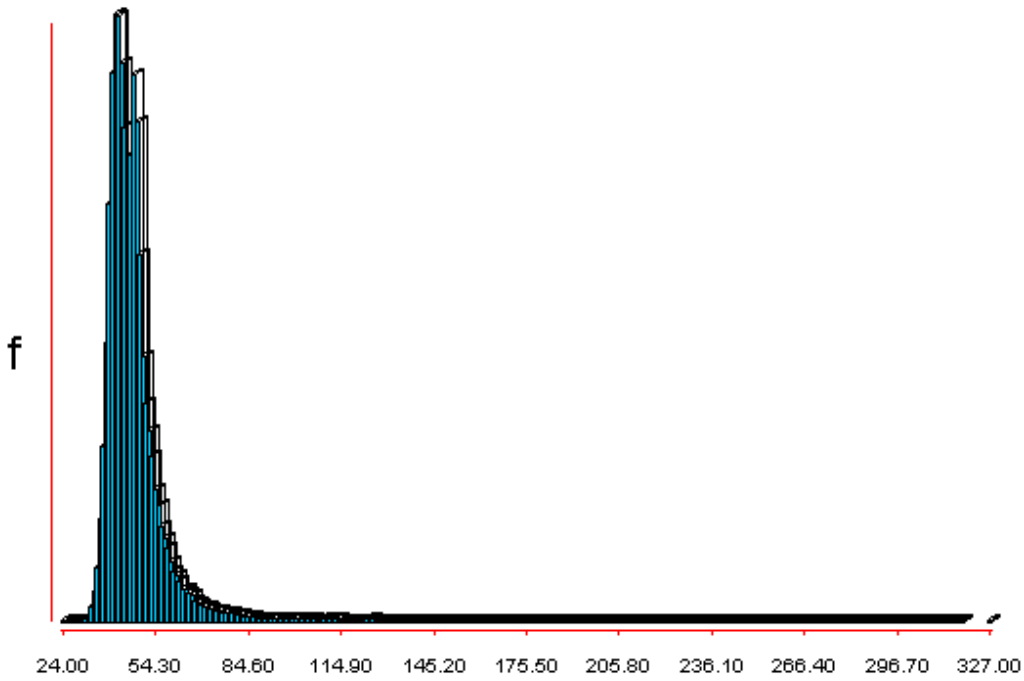
data they will produce the same results with the one exception (are you surprised?). This exception is that, if at some point, there are two or more possible rotations that contain the same "maximum" variation, then which one is used is indeterminate. In two dimensions the data cloud would look like a circle, instead of an ellipse. In a circle, any rotation would be equivalent. In an elliptical data cloud, the first component would be parallel to the major axis of the ellipse.

To calculate the rotation we can start with either a Variance-covariance Matrix or a Correlation Matrix. If one standardizes the data and calculates a Variance-covariance Matrix, then the result will be the same as a Correlation Matrix. Those that wish to practice their algebra can prove this by deriving the formula for the Variance-covariance Matrix and the Correlation Matrix calculated on "raw" data and then the Variance-covariance Matrix calculated on standardized data.

The histogram of the first Principal Component for the Morro Bay scene is:



The histogram for the second Principal Component of the Morro Bay scene is:



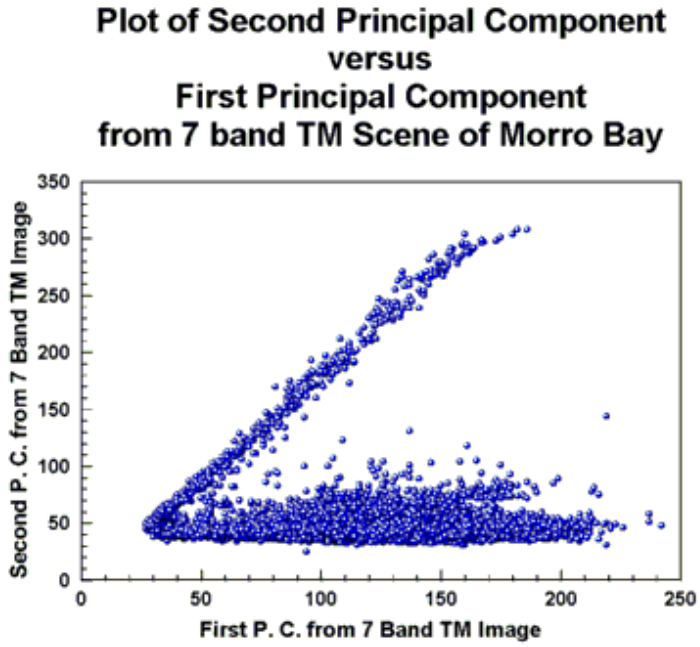
Histogram of

pcacmp2

Class width :	1.0000
Display minimum :	24.0000
Display maximum :	327.0000
Actual minimum :	24.0000
Actual maximum :	327.0000
Mean :	49.1695
Stand. Deviation :	23.0423
df :	262143

Compare these with the histograms of the original bands., which you can do by flipping back and forth to page 1-3 and then to this page.

We can plot the second principal component versus the first to get the 2D view that follows.



How do we get this figure? The elliptical cloud that lies parallel to the X axis is what we might expect. But we need to remember is that we are carrying out our rigid rotation of axes in a 7 dimensional space, one for each band (or variable). We can see here that the original data was not Multivariate Normal, an assumption that would need to be met if one wanted to carry out any parametric statistical tests. This non-normality is indicated the anomalous cloud of points going diagonally across the graph. If the data were multivariate normal in 7 dimensions, then the plot would only have a cloud like the horizontal one in the above plot.

Primary Contact: Nicholas M. Short, Sr.
Appendix C Author: Dr. Jon W. Robinson (robinson@ltpmail.gsfc.nasa.gov)

Linear Combination

For the Morro Bay TM scene there are 7 spectral bands. Thus each pixel has 7 values. The pixel in row i, column j of the image is a vector:

x(i,j,1) x(i,j,2) x(i,j,3) x(i,j,4) x(i,j,5) x(i,j,6) x(i,j,7)

x(i,j,1) is the value of band 1 in row i, column j, x(i,j,2) is the value of band 2 in row i, column j, etc.

A linear combination of these values, to calculate the first Principal Component, would look like:

$$z(i,j,1) = a(1,1)\cdot x(i,j,1) + a(1,2)\cdot x(i,j,2) + a(1,3)\cdot x(i,j,3) + a(1,4)\cdot x(i,j,4) + a(1,5)\cdot x(i,j,5) + a(1,6)\cdot x(i,j,6) + a(1,7)\cdot x(i,j,7)$$

This multiplication and addition is carried out for each of the picture elements, pixels, in the image. The Principal Components Analysis is the calculation of the values of the set of vectors a and then the multiplication of the image data by them to get the projections of the data points onto the Principal Components.

Primary Contact: Nicholas M. Short, Sr.

Singular Matrices

A singular matrix is one in which one or more of the rows or columns can be calculated as a linear combination of the other rows or columns. If one calculates the Variance-Covariance matrix of a singular data matrix, the determinant of that Variance-Covariance matrix will be 0.

For example consider the "data" matrix below with 4 variables and 5 observations.

3	9	11	2	5
5	3	4	3	1
2	7	5	5	11
17	42	41	22	44

If we call this matrix x, we can for example generate the fourth row as a linear combination of the other rows like this:

$$y = a^{t*}x'$$

Where x' is the data matrix without row 4

3	9	11	2	5
5	3	4	3	1
2	7	5	5	11

and a is a vector of 3 coefficients

2
1
3

that are used to pre multiply x' to produce y, the the fourth row. The mean vector is:

6	3.2	6	33.2

We then subtract the mean vector from each "observation" to shift the mean to zero

Matrix with mean vector shifted to Zero

-3	3	5	-4	-1
1.8	-0.2	0.8	-0.2	-2.2
-4	1	-1	-1	5
-16.2	8.8	7.8	-11.2	10.8

before calculating the Variance-Covariance matrix as **vcv = xm*xm**^{t}

The Variance-Covariance is:

60	1	9	148
1	8.8	-19	-46.2
9	-19	44	131
148	-46.2	131	642.8

and the determinant is: $3.699*10^{-11}$ which is within rounding error of 0

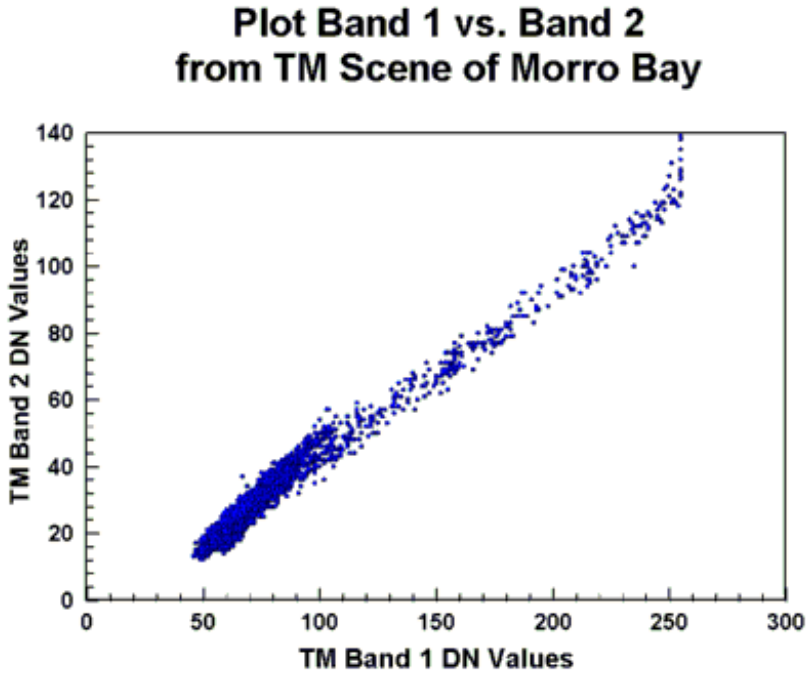
If we delete the 4 th variable and recalculate the determinat for the 3 variable data set, we get: 473.2 clearly much larger than 0! As an exercise, you can try calculating this value by hand, or with a matrix algebra package. Mathcad 5 plus was used to calculate this example.

PREVIOUS RST NEXT

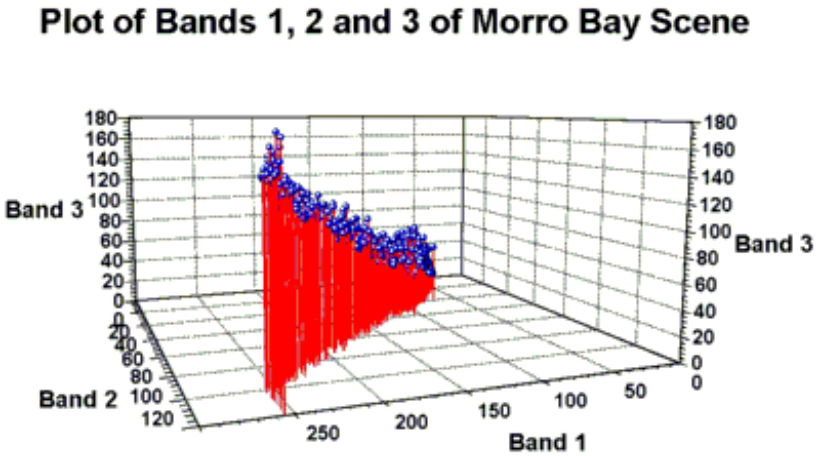
Primary Contact: Nicholas M. Short, Sr. email:

Space Dimensions

For the purposes of our discussion, a Space is defined by the image bands. Each band defines one dimension of the space. Thus if we have two spectral bands they define a two dimensional space. This space can be visualized by plotting the two spectral intensities for each pixel in a 2 dimensional x-y plot such as the figure below which is a plot of bands 1 and 2 from the Morro Bay scene.



If we add another band, then we will have a 3 dimensional space. We can still visualize this and even make a 3D model (as shown below with blue data points on the tips of red pins to show location on the Band 1 - Band 2 plane) but when we add the fourth band, we must resort to arguments by analogy to the lower dimensional 2D and 3D spaces.



With the 7 spectral bands available from Thematic Mapper, a 7 dimensional space is defined by the observations.

Primary Contact: Nicholas M. Short, Sr.

Standardizing Sets

Standardizing a set of observations means substracting the mean from each observation and dividing it by the standard deviation.

The mean is:

$$Ybar := \frac{\sum_{i=1}^{n} y(i)}{n}$$

The standard deviation is:

$$sd := \frac{\sqrt{\sum_{i=1}^{n} y(i) - \frac{\left(\sum_{i=1}^{n} y(i)\right)^2}{n}}}{n-1}$$

The standarized values of each observation are then:

$$z(i) := \frac{y(i) - \frac{\sum_{i=1}^{n} y(i)}{n}}{\frac{\sqrt{\sum_{i=1}^{n} y(i) - \frac{\left(\sum_{i=1}^{n} y(i)\right)^2}{n}}}{n-1}}$$

As an exercise, substitute Ybar and sd into the above equation.

Primary Contact: Nicholas M. Short, Sr.

GLOSSARY

This Glossary was extracted directly from an Internet Site prepared by Jeff Weissel and others at the Lamont-Doherty Earth Observatory of Columbia University and is reproduced here with their permission. Their sources for the definitions come principally from:

1) S.A. Drury, `A Guide to Remote Sensing', (Oxford), 199 pp., 1990.

2) F.F. Sabins, `Remote Sensing Principles and Interpretation', 2nd Edition, (W.H. Freeman & Co.), 449 pp., 1987

While the Columbia University Glossary is comprehensive, it is out of date. There is a newer version on the Web, which has proved difficult to download. We cite it here as Glossary 1. We also recommend these other glossaries: Glossary 2, Glossary 3, and Glossary 4, whose developers will be evident when you link onto each.

A

absolute temperature-Temperature measured on the Kelvin scale, whose base is absolute zero, i.e. -273 °C; 0 °C is expressed as 273 °K.
absorptance-A measure of the ability of a material to absorb EM energy at a specific wavelength.
absorption band-Wavelength interval within which electromagnetic radiation is absorbed by the atmosphere or by other substances.
absorptivity-Capacity of a material to absorb incident radiant energy.
achromatic vision-The perception by the human eye of changes in brightness, often used to describe the perception of monochrome or black and white scenes.
active remote sensing-Remote sensing methods that provide their own source of electromagnetic radiation to illuminate the terrain. Radar is one example.
acuity-A measure of human ability to perceive spatial variations in a scene. It varies with the spatial frequency, shape, and contrast of the variations, and depends on whether the scene is coloured or monochrome.
additive primary colors-Blue, green, and red. Filters of these colors transmit the primary color of the filter and absorb the other two colors.
adiabatic cooling-Refers to decrease in temperature with increasing altitude.
advanced very high resolution radiometer (AVHRR)-Crosstrack multispectral scanner on a NOAA polar-orbiting satellite that acquires five spectral bands of data (0.55 to 12.50 µm) with a ground resolution cell of 1.1 by 1.1 km.
aerial magnetic survey-Survey that records variations in the earth's magnetic field.
air base-Ground distance between optical centers of successive overlapping aerial photographs.
airborne imaging spectrometer (AIS)-Along-track multispectral scanner with spectral bandwidth of 0.01 µm.
airborne visible and infrared imaging spectrometer (AVIRIS)-Experimental airborne along-track multispectral scanner under development at JPL to acquire 224 images in the spectral region from 0.4 to 2.4 µm.
AIS-Airborne imaging spectrometer.

albedo (A)-Ratio of the amount of electromagnetic energy reflected by a surface to the amount of energy incident upon it.
along-track scanner-Scanner with a linear array of detectors oriented normal to flight path. The IFOV of each detector sweeps a path parallel with the flight direction.
alteration-Changes in color and mineralogy of rocks surrounding a mineral deposit that are caused by the solutions that formed the deposit. Suites of alteration minerals commonly occur in zones.
amplitude-For waves, the vertical distance from crest to trough.
Analog display-A form of data display in which values are shown in graphic form, such as curves. Differs from digital displays in which values are shown as arrays of numbers.
analogue image-An image where the continuous variation in the property being sensed is represented by a continuos variation in image tone. In a photograph this is achieved directly by the grains of photosensitive chemicals in the film; in an electronic scanner, the response in, say, millivolts is transformed to a display on a cathode-ray tube where it may be photographed.
angular beam width-In radar, the angle subtended in the horizontal plane by the radar beam.
angular field of view-Angle subtended by lines from a remote sensing system to the outer margins of the strip of terrain that is viewed by the system.
angular resolving power-Minimum separation between two resolvable targets, expressed as angular separation.
anomaly-An area on an image that differs from the surrounding, normal area. For example, a concentration of vegetation within a desert scene constitutes an anomaly.
antenna-Device that transmits and receives microwave and radio energy in radar systems.
aperture-Opening in a remote sensing system that admits electromagnetic radiation to the film in radar systems.
Apollo-U.S. lunar exploration program of satellites with crews of three astronauts.
apparent thermal inertia (ATI)-An approximation of thermal inertia calculated as one minus albedo divided by the difference between daytime and nighttime radiant temperatures.
artefact-A feature on an image which is produced by the optics of the system or by digital image processing, and sometimes masquerades as a real feature.
ASA index-Index of the American Standards Association designating film speed, or sensitivity to light. Higher values indicate higher sensitivity. The ASA index has been replaced by the ISO index.
ATI-Apparent thermal inertia.
atmosphere-Layer of gases that surrounds some planets.
atmospheric correction-Image-processing procedure that compensates for effects of selectivity scattered light in multispectral images.
atmospheric shimmer-An effect produced by the movement of masses of air with different refractive indices, which is most easily seen in the twinkling of stars. Shimmer results in blurring on remotely sensed images, and is the ultimate control over the resolution of any system.
atmospheric window-Wavelength interval within which the atmosphere readily transmits electromagnetic radiation.
attitude-Angular orientation of remote sensing system with respect to a geographic reference system.
AVHRR-Advanced Very High Resolution Radiometer, a multispectral imaging system carried by the TIROS-NOAA series of meteorological satellites.
AVIRIS-Airborne visible and infrared imaging spectrometer.
azimuth-Geographic orientation of a line given as an angle measured in degrees clockwise from north.
azimuth direction-In radar images, the direction in which the aircraft is heading. Also called flight direction.
azimuth resolution-In radar images, the spatial resolution in the azimuth direction.

B

background-Area on an image or the terrain that surrounds an area of interest, or target.
backscatter-In radar, the portion of the microwave energy scattered by the terrain surface directly back toward the antenna.
backscatter coefficient-A quantitative measure of the intensity of energy returned to a radar antenna from the terrain.
band-A wavelength interval in the electromagnetic spectrum. For example, in Landsat images the bands designate specific wavelength intervals at which images are acquired.
base-height ratio-Air base divided by aircraft height. This ratio determines vertical exaggeration on stereo models.
batch processing-Method of data processing in which data and programs are entered into a computer that carries out the entire processing operation with no further instructions.
bathymetry-Configuration of the seafloor.
beam-A focused pulse of energy.
bin-One of a series of equal intervals in a range of data, most commonly employed to describe the divisions in a histogram.
binary-Numerical system using the base 2.
bit-Contraction of binary digit, which in digital computing represents an exponent of the base 2.
blackbody-An ideal substance that absorbs all the radiant energy incident on it and emits radiant energy at the maximum possible rate per unit area at each wavelength for any given temperature. No actual substance is a true blackbody, although some substances, such as lampblack, approach its properties.
blind spot-The point of the optic nerve to the retina where no radiation is detected by the eye.
brightness-Magnitude of the response produced in the eye by light.
brute-force radar-See real-aperture radar.
byte-A group of eight bits of digital data.

C

calibration-Process of comparing an instrument's measurements with a standard.
calorie-Amount of heat required to raise the temperature of 1g of water by 1 °C.
camouflage detection photographs-Another term for IR color photograph.
cardinal point effect-In radar, very bright signatures caused by optimally oriented corner reflectors, such as buildings.
cathode ray tube (CRT)-A vacuum tube with a phosphorescent screen on which images are displayed by an electron beam.
CCD-Charge-coupled detector.
CCT-Computer-compatible tape.
cell assemblies-The linked receptors, retinal neurons, and neural cells in the visual cortex of the brain which enable interaction between perception and past experience.
centerpoint-The optical center of a photograph.
change-detection images-A difference image prepared by digitally comparing images acquired at different times. The gray tones or colors of each pixel record the amount of difference between the corresponding pixels of the original images.
charge-coupled detector (CCD)-A device in which electron are stored at the surface of a semiconductor.
chlorosis-Yellowing of plant leaves resulting from an imbalance in the iron metabolism caused by excess concentrations of copper, zinc, manganese, or other elements in the plant.

chromatic vision-The perception by the human eye of changes in hue.
circular scanner-Scanner in which a faceted mirror rotates about a vertical axis to sweep the detector IFOV in a series of circular scan lines on the terrain.
classification-Process of assigning individual pixels of an image to categories, generally on the basis of spectral reflectance characteristics.
coastal zone color scanner (CZCS)-A satellite-carried multi-spectral scanner designed to measure chlorophyll concentrations in the oceans.
coherent radiation-Electromagnetic radiation whose waves are equal in length and are in phase, so that waves at different points in space act in unison, as in laser and synthetic aperture radar.
color composite image-Color image prepared by projecting individual black-and-white multispectral images, each through a different color filter. When the projected images are superposed, a color composite image results.
color ratio composite image-Color composite image prepared by combining individual ratio images for a scene using a different color for each ratio image.
complementary colors-Two primary colors of light (one additive and the other subtractive) that produce white light when added together. Red and cyan are complimentary colors.
computer-compatible tape (CCT)-The magnetic tape on which the digital data for Landsat MSS and TM images are distributed.
conduction-Transfer of electromagnetic energy through a solid material by molecular interaction.
cones-Receptors in the retina which are sensitive to colour. There are cones sensitive to the red, green, and blue components of light.
contact print-A reproduction from a photographic negative in direct contact with photosensitive paper.
context-The known environment of a particular feature on an image.
contrast-The ratio between the energy emitted or reflected by an object and its immediate surroundings.
contrast enhancement-Image-processing procedure that improves the contrast ratio of images. The original narrow range of digital values is expanded to utilize the full range of available digital values.
contrast ratio-On an image, the ratio of reflectances between the brightest and darkest parts of an image.
contrast stretching-Expanding a measured range of digital numbers in an image to a larger range, to improve the contrast of the image and its component parts.
convection-Transfer of heat through the physical movement of heated matter.
corner reflector-Cavity formed by two or three smooth planar surfaces intersecting at right angles. Electromagnetic waves entering a corner reflector are reflected directly back toward the source.
COSMIC-Computer Software Management and Information Center, University of Georgia. This facility distributes computer programs developed by U.S. government-funded projects.
cross-polarized-Describes a radar pulse in which the polarization direction of the return is normal to the polarization direction of the transmission. Cross-polarized images may be HV (horizontal transmit, vertical return) or VH (vertical transmit, horizontal return).
cross-track scanner-Scanner in which a faceted mirror rotates about a horizontal axis to sweep the detector IFOV in a series of parallel scan lines oriented normal to the flight direction.
CRT-Cathode ray tube.
cut off-The digital number in the histogram of a digital image which is set to zero during contrast stretching. Usually this is a value below which atmospheric scattering makes a major contribution.
cycle-One complete oscillation of a wave.
CZCS-Coastal Zone color scanner.

D

data collection system (DCS)-On Landsats 1 and 2, the system that acquired information from seismometers, flood gauges, and other measuring devices. These data were relayed to ground receiving stations.
densitometer-Optical device for measuring the density of photographic transparencies.
density, of images-Measure of the opacity, or darkness, of a negative or positive transparency.
density, of materials (r)-Ratio of mass to volume of a material, typically expressed as grams per cubic centimeter.
density slicing-Process of converting the continuous gray tones of an image into a series of density intervals, or slices, each corresponding to a specific digital range. The density slices are then displayed either as uniform gray tones or as colors.
depolarized-Refers to a change in polarization of a transmitted radar pulse as a result of various interactions with the terrain surface.
depression angle (y)-In radar, the angle between the imaginary horizontal plane passing through the antenna and the line connecting the antenna and the target.
detectability-Measure of the smallest object that can be discerned on an image.
detector-Component of a remote sensing system that converts electromagnetic radiation into a recorded signal.
developing-Chemical processing of an exposed photographic emulsion to produce an image.
dielectric constant-Electrical property of matter that influences radar returns. Also referred to as complex dielectric constant.
difference image-Image prepared by subtracting the digital values of pixels in one image from those in a secon image to produce a third set of pixels. This third set is used to form the difference image.
diffuse reflector-Surface that reflects incident radiation nearly equally in all directions.
digital display-A form of data display in which values are shown as arrays of numbers.
digital image-An image where the property being measured has been converted from a continuous range of analogue values to a range expressed by a finite number of integers, usually recorded as binary codes from 0 to 255, or as one byte.
digital image processing-Computer manipulation of the digital-number values of an image.
digital number (DN)-Value assigned to a pixel in a digital image.
digitization-Process of converting an analog display into a digital display.
digitizer-Device for scanning an image and converting it into numerical format.
directional filter-Mathematical filter designed to enhance on an image those linear features oriented in a particular direction.
distortion-On an image, changes in shape and position of objects with respect to their true shape and position.
diurnal-Daily.
Doppler principle-Describes the change in observed frequency that electromagnetic or other waves undergo as a result of the movement of the source of waves relative to the observer.
Doppler shift-A change in the observed frequency of EM or other waves caused by the relative motion between source and detector. Used principally in the generation of synthetic-aperture radar images.
dwell time-Time required for a detector IFOV to sweep across a ground resolution cell.

E

EDC-EROS Data Center.
edge-A boundary in an image between areas with different tones.

edge enhancement-Image-processing technique that emphasizes the appearance of edges and lines.
Ektachrome-A Kodak color positive film.
electromagnetic radiation-Energy propagated in the form of and advancing interaction between electric and magnetic fields. All electromagnetic radiation moves at the speed of light.
electromagnetic spectrum-Continuous sequence of electromagnetic energy arranged according to wavelength or frequency.
emission-Process by which a body radiates electromagnetic energy. Emission is determined by kinetic temperature and emissivity.
emissivity (e)- Ratio of radiant flux from a body to that from a blackbody at the same kinetic temperature and emissivity.
emittance-A term for the radiant flux of energy per unit area emitted by a body. (Now obsolete).
emulsion-Suspension of photosensitive silver halide grains in gelatin that constitutes the image-forming layer on photographic film.
energy flux-Radiant flux.
enhancement-Process of altering the appearance of an image so that the interpreter can extract more information.
EOSAT-The commercial company that took over operations of the Landsat system in 1985.
ERBSS-Earth Radiation Budget Sensor System, carried by NOAA satellites.
EREP-Earth Resources Experiment Package, carried on Skylab and consisting of cameras and multispectral scanner.
EROS-Earth Resources Observation System.
EROS Data Center (EDC)-Facility of the U.S. Geological Survey at Sioux Falls, South Dakota, that archives, processes, and distributes images.
ERTS-Earth Resource Technology Satellite, now called Landsat.
ESA- European Space Agency, based in Paris. A consortium between several European states for the development of space science, including the launch of remote-sensing satellites.
ETC-Earth-terrain camera.
Evaporative cooling-Temperature drop caused by evaporation of water from a moist surface.

F

false colour image-A colour image where parts of the non-visible EM spectrum are expressed as one or more of the red, green, and blue components, so that the colours produced by the Earth's surface do not correspond to normal visual experience. Also called a false-colour composite (FCC). The most commonly seen false-colour images display the very-near infrared as red, red as green, and green as blue.
false color photograph-Another term for IR color photograph.
far range-The portion of a radar image farthest from the aircraft or spacecraft flight path.
film-Light-sensitive photographic emulsion and its base.
film speed-Measure of the sensitivity of photographic film to light. Larger numbers indicate higher sensitivity.
filter, digital-Mathematical procedure for modifying values of numerical data.
filter, optical-A material that, by absorption or reflection, selectivity modifies the radiation transmitted through an optical system.
flight path-Line on the ground directly beneath a remote sensing aircraft or space craft. Also called flight line.
fluorescence-Emission of light from a substance following exposure to radiation from an external source.
f-number-Representation of the speed of a lens determined by the focal length divided by

diameter of the lens. Smaller numbers indicate faster lenses.
focal length-In cameras, the distance from the optical center of the lens to the plane at which the image of a very distant object is brought into focus.
foreshortening-A distortion in radar images causing the lengths of slopes facing the antenna to appear shorter on the image than on the ground. It is produced when radar wavefronts are steeper than the topographic slope.
format-Size of an image.
forward overlap-The percent of duplication by successive photographs along a flight line.
fovea-The region around that point on the retina intersected by the eye's optic axis, where receptors are most densely packed. It is the most sensitive part of the retina.
frequency (v)-The number of wave oscillations per unit time or the number of wavelengths that pass a point per unit time.
f-stop-Focal length of a lens divided by the diameter of the len's adjustable diaphragm. Smaller numbers indicate larger openings, which admit more light to the film.

G

GCP-Ground-control point.
Gemini-U.S. program of two-man earth-orbiting spacecraft in 1965 and 1966.
geographic information system (GIS)-A data-handling and analysis system based on sets of data distributed spatially in two dimensions. The data sets may be map oriented, when they comprise qualitative attributes of an area recorded as lines, points, and areas often in vector format, or image oriented, when the data are quantitative attributes referring to cells in a rectangular grid usually in raster format. It is also known as a geobased or geocoded information system.
geometric correction-Image-processing procedure that corrects spatial distortions in an image.
geostationary-Refers to satellites traveling at the angular velocity at which the earth rotates; as a result, they remain above the same point on earth at all times.
Geostationary Operational Environmental Satellite-a NOAA satellite that acquires visible and thermal IR images for meteorologic purposes.
geostationary orbit-An orbit at 41 000 km in the direction of the Earth's rotation, which matches speed so that a satellite remains over a fixed point on the Earth's surface.
geothermal-Refers to heat from sources within the earth.
Goddard Space Flight Center-The NASA facility at Greenbelt, Maryland, that is also a Landsat ground receiving station.
GMT-Greenwich mean time. This international 24-h system is used to designate the time at which Landsat images are acquired.
GOES-Geostationary Operational Environmental Satellite.
gossan-Surface occurrence of iron oxide formed by the weathering of metallic sulfide ore minerals.
granularity-Graininess of developed photographic film that is determined by the texture of the silver grains.
gray scale-A sequence of gray tones ranging from black to white.
grid format-The result of interpolation from values of a variable measured at irregularly distributed points, or along survey lines, to values referring to square cells in a rectangular array. It forms a step in the process of contouring data, but can also be used as the basis for a raster format to be displayed and analyzed digitally after the values have been rescaled to the 0-255 range.
ground-control point-A geographic feature of known location that is recognizable on images and can be used to determine geometric corrections.
ground range-On radar images, the distance from the ground track to an object.
ground-range image-Radar image in which the scale in the range direction is constant.
ground receiving station-Facility that records data transmitted by a satellite, such as Landsat.

ground resolution cell-Area on the terrain that is covered by the IFOV of a detector.
ground swath-Width of the strip of terrain that is imaged by a scanner system.
GSFC-Goddard Space Flight Center

H

harmonic-Refers to waves in which the component frequencies are whole- number multiples of the fundamental frequency.
HCMM-Heat Capacity Mapping Mission, the NASA satellite launched in 1978 to observe thermal properties of rocks and soils. It remained in orbit for only a few months.
heat capacity-(c) Ratio of heat absorbed or released by a material to the corresponding temperature rise or fall. Expressed in calories per gram per degree centigrade. Also called thermal capacity.
Heat Capacity Mapping Mission (HCMM)-NASA satellite orbited in 1978 to record daytime and nighttime visible and thermal IR images of large areas.
highlights-Areas of bright tone on an image.
high-pass filter-A spatial filter which selectively enhances contrast variations with high spatial frequencies in an image. It improves the sharpness of images and is a method of edge enhancement.
HIRIS-High Resolution Imaging Spectrometer, possibly to be carried by the Space Shuttle.
HIRS-High Resolution Infrared Spectrometer, carried by NOAA satellites.
histogram-A means of expressing the frequency of occurrence of values in a data set within a series of equal ranges or bins, the height of each bin representing the frequency at which values in the data set fall within the chosen range. A cumulative histogram expresses the frequency of all values falling within a bin and lower in the range. A smooth curve derived mathematically from a histogram is termed the probability density function (PDF).
hue-In the IHS system, represents the dominant wavelength of a color.

I

IFOV-Instantaneous field of view.
IHS-Intensity, hue, and saturation system of colors.
image-pictorial representation of a scene recorded by a remote sensing system. Although image is a general term, it is commonly restricted to representations acquired by non-photographic methods.
image dissection-The breaking down of a continuous scene into discrete spatial elements, either by the receptors on the retina, or in the process of capturing the image artificially.
image striping-A defect produced in line scanner and pushbroom imaging devices produced by the non-uniform response of a single detector, or amongst a bank of detectors. In a line-scan image the stripes are perpendicular to flight direction, but parallel to it in a pushbroom image.
image swath-See ground swath.
incidence angle-In radar, the angle formed between an imaginary line normal to the surface and another connecting the antenna and the target.
incident energy-Electromagnetic radiation impinging on a surface.
index of refraction (n) -Ratio of the wavelength or velocity of electromagnetic radiation in a vacuum to that in a substance.
instantaneous field of view (IFOV)- Solid angle through which a detector is sensitive to radiation. In a scanning system, the solid angle subtended by the detector when the scanning motion is stopped.
intensity-In the IHS system, brightness ranging from black to white.
interactive processing-Method of image processing in which the operator views

preliminary results and can alter the instructions to the computer to achieve desired results.
interpretation-The process in which a person extracts information from an image.
interpretation key-Characteristic or combination of characteristics that enable an interpreter to identify an object on an image.
IR-Infrared region of the electromagnetic spectrum that includes wavelengths from 0.7µm to 1 mm.
IR color photograph-Color photograph in which the red-imaging layer is sensitive to photographic IR wavelengths, the green-imaging layer is sensitive to red light, and the blue-imaging layer is sensitive to green light. Also known as camouflage detection photographs and false-color photographs.
ISO index- Index of the International Standards Organization, designating film speed in photography. Higher values indicate higher sensitivity.
isotherm-Contour line connecting points of equal temperature. Isotherm maps are used to portray surface-temperature patterns of water bodies.

J

Johnson Space Flight Center-A NASA facility in Houston, Texas.
JPL-Jet Propulsion Laboratory, a NASA facility at Pasadena, California, operated under contract by the California Institute of Technology.

K

Ka band-Radar wavelength region from 0.8 to 1.1 cm.
kernel-Two-dimensional array of digital numbers used in digital filtering.
kinetic energy-The ability of a moving body to do work by virtue of its motion. The molecular motion of matter is a form of kinetic energy.
kinetic temperature-Internal temperature of an object determined by random molecular motion. Kinetic temperature is measured with a contact thermometer.
Kodachrome-A Kodak color positive film.

L

LACIE-Large Area Crop Inventory Experiment.
Landsat-A series of unnamed earth-orbiting NASA satellites that acquire multispectral images in various visible and IR bands.
Laplacian filter-A form of nondirectional digital filter.
large-format camera (LFC)-An experiment first carried on the Space Shuttle in October 1984.
laser-Light artificially stimulated electromagnetic radiation: a beam of coherent radiation with a single wavelength.
latent image- Invisible image produced by the photochemical effect of light on silver halide grains in the emulsion of film. The latent image is not visible until after photographic development.
layover-In radar images, the geometric displacement of the top of objects toward the near range relative to their base.
L band-Radar wavelength region from 15 to 30 cm.
lens-One or more pieces of glass or other transparent material shaped to form an image by refraction of light.
LFC-Large-format camera.
lidar-Light intensity detection and ranging, which uses lasers to stimulate fluorescence in

various compounds and to measure distances to reflecting surfaces.
light-Electromagnetic radiation ranging from 0.4 to 0.7µm in wavelength that is detectable by the human eye.
light meter-Device for measuring the intensity of visible radiation and determining the appropriate exposure of photographic film in a camera.
lineament-Linear topographic or tonal feature on the terrain and on images, maps, and photographs that may represent a zone of structural weakness.
linear-Adjective that describes the straight line-like nature of features on the terrain or on images and photographs.
lineation-The one-dimensional alignment of internal components of a rock that cannot be depicted as an individual feature on a map.
line drop out-The loss of data from a scan line caused by malfunction of one of the detectors in a line scanner.
line-pair-Pair of light and dark bars of equal widths. The number of such line-pairs aligned side by side that can be distinguished per unit distance expresses the resolving power of an imaging system.
line scanner-An imaging device which uses a mirror to sweep the ground surface normal to the flight path of the platform. An image is built up as a strip comprising lines of data.
look angle-The angle between the vertical plane containing a radar antenna and the direction of radar propagation. Complementary to the depression angle.
look direction-Direction in which pulses of microwave energy are transmitted by a radar system. The look direction is normal to the azimuth direction. Also called range direction.
look-up table (LUT)-A mathematical formula used to convert one distribution of data to another, most conveniently remembered as a conversion graph.
low-sun-angle photograph-Aerial photograph acquired in the morning, evening, or winter when the sun is at a low elevation above the horizon.
luminance-Quantitative measure of the intensity of light from a source.

M

Mach band-An optical illusion of dark and light fringes within adjacent areas of contrasted tone. It is a psychophysiological phenomenon which aids human detection of boundaries or edges.
median filter-A spatial filter, which substitutes the median value of DN from surrounding pixels for that recorded at an individual pixel. It is useful for removing random noise.
Mercury-U.S. program of one-man, earth-orbiting spacecraft in 1962 and 1963.
microwave-Region of the elctromagnetic spectrum in the wavelength range of 0.1 to 30 cm.
mid-infrared (MIR)-The range of EM wavelengths from 8 to 14 µm dominated by emission of thermally generated radiation from materials; also known as thermal infrared.
Mie scattering-The scattering of EM energy by particles in the atmosphere with comparable dimensions to the wavelength involved.
minimum ground separation-Minimum distance on the ground between two targets at which they can be resolved on an image.
minus-blue photographs-Black-and-white photographs acquired using a filter that removes blue wavelengths to produce higher spatial resolution.
mixed pixel-A pixel whose DN represents the average energy reflected or emitted by several types of surface present within the area that it represents on the ground; sometimes called a mixel.
modular optoelectric multispectral scanner (MOMS)-An along-track scanner carried on the Space Shuttle that recorded two bands of data.
modulate-To vary the frequency, phase, or amplitude of electromagnetic waves.
modulation transfer function (MTF)-A method of describing spatial resolution.
MOMS-Modular optoelectric multispectral scanner.
MOS-1-Marine Observation Satellite, launched by Japan in 1987.

mosaic-Composite image or photograph made by piecing together individual images or photographs covering adjacent areas.
MSS-Multispectral scanner system of Landsat that acquires images of four wavelength bands in the visible and reflected IR regions.
multiband camera-System that simultaneously acquires photographs of the same scene at different wavelengths.
multispectral classification-Identification of terrain categories by digital processing of data acquired by multispectral scanners.
multispectral scanner-Scanner system that simultaneously acquires images of the same scene at different wavelengths.

N

nadir-Point on the ground directly in line with the remote sensing system and the center of the earth.
NASA-National Aeronautical and Space Administration.
near infrared (NIR)-The shorter wavelength range of the infrared region of the EM spectrum, from 0.7 to 2.5 µm. It is often divided into very-near infrared (VNIR) covering the range accessible to photographic emulsions (0.7 to 1.0m), and the short-wavelength infrared (SWIR) covering the remainder of the NOR atmospheric window from 1.0 to 2.5m.
near range-Refers to the portion of a radar image closest to the aircraft or satellite flight path.
negative photograph-Photograph on film or paper in which the relationship between bright and dark tones is the reverse of that of the features on the terrain.
NHAP-National High Altitude Photography program of the U.S. Geological Survey.
NOAA-National Oceanic and Atmospheric Administration.
noise-Random or repetitive events that obscure or interfere with the desired information.
nondirectional filter-Mathematical filter that treats all orientations of linear features equally.
non-selective scattering-The scattering of EM energy by particles in the atmosphere which are much larger than the wavelengths of the energy, and which causes all wavelengths to be scattered equally.
non-spectral hue-A hue which is not present in the spectrum of colours produced by the analysis of white light by a prism of diffraction grating. Examples are brown, magenta, and pastel shades.
nonsystematic distortion-Geometric irregularities on images that are not constant and cannot be predicted from the characteristics of the imaging system.
normal color film-Film in which the colors are essentially true representations of the colors of the terrain.
NSSDC-National Space Science Data Center.

O

oblique photograph-Photograph acquired with the camera intentionally directed at some angle between horizontal and vertical orientations.
OMS-Orbital maneuvering system.
orbit-Path of a satellite around a body such as the earth, under the influence of gravity.
orthophotograph-A vertical aerial photograph from which the distortions due to varying elevation, tilt, and surface topography have been removed, so that it represents every object as if viewed directly from above.
orthophotoscope-An optical-electronic device which converts a normal vertical aerial photograph to an orthophotograph.
overlap-Extent to which adjacent images or photographs cover the same terrain, expressed as a percentage.

P

panchromatic film-Black and white film that is sensitive to all visible wavelengths.
parallax-Displacement of the position of a target in an image caused by a shift in the observation system.
parallax difference-The difference in the distance on overlapping vertical photographs between two points, which represent two locations on the ground with different elevations.
parallel-polarized-Describes a radar pulse in which the polarization of the return is the same as that of the transmission. Parallel-polarized images may be HH (horizontal transmit, horizontal return) or VV (vertical transmit, vertical return).
pass-In digital filters, refers to the spatial frequency of data transmitted by the filter. High-pass filters transmit high-frequency data; low-pass filters transmit low-frequency data.
passive microwaves-Radiation in the 1 mm to 1 m range emitted naturally by all materials above absolute zero.
passive remote sensing-Remote sensing of energy naturally reflected or radiated from the terrain.
path-and-row index-System for locating Landsat MSS and TM images.
pattern-Regular repetition of tonal variations on an image or photograph.
periodic line dropout-Defect on Landsat MSS or TM images in which no data are recorded for every sixth or sixteenth scan line, causing a black line on the image.
periodic line striping-Defect on Landsat MSS or TM images in which every sixth or sixteenth scan line is brighter or darker than the others. Caused by the sensitivity of one detector being higher or lower than the others.
photodetector-Device for measuring energy in the visible-light band.
photogeology-Mapping and interpretation of geologic features from aerial photographs.
photograph-Representation of targets on film that results from the action of light on silver halide grains in the film's emulsion.
photographic IR-Short-wavelength portion (0.7 to 0.9 µm) of the IR band that is detectable by IR color film or IR black-and-white film.
photographic UV-Long-wavelength portion of the UV band (0.3 to 0.4 µm) that is transmitted through the atmosphere and is detectable by film.
photomosaic-Mosaic composed of photographs.
photon-Minimum discrete quantity of radiant energy.
photopic vision-Vision under conditions of bright illumination.
picture element-In a digitized image, the area on the ground represented by each digital number. Commonly contracted to pixel.
pitch-Rotation of an aircraft about the horizontal axis normal to its longitudinal axis that causes a nose-up or nose-down attitude.
pixel-Contraction of picture element.
Planck's Law-An expression for the variation of emittance of a blackbody at a particular temperature as a function of wavelength.
point spread function (PSF)- The image of a point source of radiation, such as a star, collected by an imaging device. A measure of the spatial fidelity of the device.
polarization-The direction of orientation in which the electrical field vector of electromagnetic radiation vibrates.
polar orbit-An orbit that passes close to the poles, thereby enabling a satellite to pass over most of the surface, except the immediate vicinity of the poles themselves.
polarized radiation-Electromagnetic radiation in which the electrical field vector is contained in a single plane, instead of having random orientation relative to the propagation vector. Most commonly refers to radar images.
positive photograph-Photographic image in which the tomes are directly proportional to the terrain brightness.
previsual symptom-A vegetation anomaly that is recognizable on IR film before it is visible to the naked eye or on normal color photographs. It results when stressed

vegetation loses its ability to reflect photographic IR energy and its recognizable on IR color film by a decrease in brightness of the red hues.
primary colors-A set of three colors that in various combinations will produce the full range of colors in the visible spectrum. There are two sets of primary colors, additive and subtractive.
principal component analysis-The analysis of covariance in a multiple data set so that the data can be projected as additive combinations on to new axes, which express different kinds of correlation among the data.
principal-component (PC) image-Digitally processed image produced by a transformation that recognizes maximum variance in multispectral images.
principal point-Optical center of an aerial photograph.
printout-Display of computer data in alphanumeric format.
probability density function (PDF)-A function indicating the relative frequency with which any measurement may be expected to occur. In remote sensing it is represented by the histogram of DN in one band for a scene.
pulse-Short burst of electromagnetic radiation transmitted by a radar antenna.
pulse length-Duration of a burst of energy transmitted by a radar antenna, measured in microseconds.
pushbroom scanner-An alternate term for an along-track scanner
pushbroom system-An imaging device consisting of a fixed linear array of many sensors which is swept across an area by the motion of the platform, thereby building up an image. It relies on sensors whose response and reading is nearly instantaneous, so that the image swathe can be segmented into pixels representing small dimensions on the ground.

Q

quantum-The elementary quantity of EM energy that is transmitted by a particular wavelength. According to the quantum theory, EM radiation is emitted, transmitted, and absorbed as numbers of quanta, the energy of each quantum being a simple function of the frequency of the radiation.

R

radar-Acronym for radio detection and ranging. Radar is an active form of remote sensing that operates in the microwave and radio wavelength regions.
radar altimeter-A non-imaging device that records the time of radar returns from vertically beneath a platform to estimate the distance to and hence the elevation of the surface; carried by Seasat and the EAS-ERS-1 platforms.
radar cross section-A measure of the intensity of backscattered radar energy from a point target. Expressed as the area of a hypothetica surface which scatters radar equally in all directions and which would return the same energy to the antenna.
radar scattering coefficient-A measure of the back-scattered energy from a target with a large area. Expressed as the average radar cross section per unit area in decibels (dB). It is the fundamental measure of the radar properties of a surface.
radar scatterometer-A non-imaging device that records radar energy backscattered from terrain as a function of depression angle.
radar shadow-Dark signature on a radar image representing no signal return. A shadow extends in the far-range direction form an object that intercepts the radar beam.
radial relief displacement-The tendency of vertical objects to appear to learn radially away from the center of a vertical aerial photograph. Caused by the conical field of view of the camera lens.
radian-Angle subtended by an arc of a circle equal in length to the radius of the circle 1 rad = 57.3¡.
radiant energy peak-Wavelength at which the maximum electromagnetic energy is

radiated at a particular temperature.
radiant flux-Rate of flow of electromagnetic radiation measured in watts per square centimeter.
radiant temperature-Concentration of the radiant flux from a material. Radiant temperature is the kinetic temperature multiplied by the emissivity to the one-fourth power.
radiation-Propagation of energy in the form of electromagnetic waves.
radiometer-Device for quantitatively measuring radiant energy, especially thermal radiation.
random line dropout-In scanner images, the loss of data from individual scan lines in a nonsystematic fashion.
range-In radar usage this is the distance in the direction of radar propagation, usually to the side of the platform in an imaging radar system. The slant range is the direct distance from the antenna to the object, whereas the distance from the ground track of the platform to the object is termed the ground range.
range direction-See look direction.
range resolution-In radar images, the spatial resolution in the range direction, which is determined by the pulse length of the transmitted microwave energy.
raster-The scanned and illuminated area of a video display, produced by a modulated beam of electrons sweeping the phosphorescent screen line by line from top to bottom at a regular rate of repetition.
raster format-A means of representing spatial data in the from of a grid of DN, each line of which can be used to modulate the lines of a video raster.
raster pattern-Pattern of horizontal lines swept by an electron beam across the face of a CRT that constitute the image display.
ratio image-An image prepared by processing digital multi-spectral data as follows: for each pixel, the value for one band is divided by that of another. The resulting digital values are displayed as an image.
Rayleigh criterion-In radar, the relationship between surface roughness, depression angle, and wavelength that determines whether a surface will respond in a rough or smooth fashion to the radar pulse.
Rayleigh scattering-Selective scattering of light in the atmosphere by particle that are small compared with the wavelength of light.
RBV-Return-beam vidicon.
real-aperture radar-Radar system in which azimuth resolution is determined by the transmitted beam width, which is in turn determined by the physical length of the antenna and by the wavelength.
real time-Refers to images or data made available for inspection simultaneously with their acquisition.
recognizability-Ability to identify an object on an image.
rectilinear-Refers to images with no geometric distortion in which the scales in the horizontal and vertical directions are identical.
redundancy-Information on an image which is either not required for interpretation or cannot be seen. Redundancy may be spatial or spectral. The term also refers to multispectral data where the degree of correlation between bands is so high that one band contains virtually the same information as all the bands.
reflectance-Ratio of the radiant energy reflected by a body to the energy incident on it. Spectral reflectance is the reflectance measured within a specific wavelength interval.
reflected energy peak-Wavelength (0.5 µm) at which maximum amount of energy is reflected from the earth's surface.
reflected IR-Electromagnetic energy of wavelengths from 0.7 µm to about 3 µm that consists primarily of reflected solar radiation.
reflectivity-Ability of a surface to reflect incident energy.
refraction-Bending of electromagnetic rays as they pass from one medium into another when each medium has a different index of refraction.
registration-Process of superposing two or more images or photographs so that

equivalent geographic points coincide.
relief-Vertical irregularities of a surface.
relief displacement-Geometric distortion on vertical aerial photographs. The tops of objects appear in the photograph to be radially displaced from their bases outward from the photograph's centerpoint.
remote sensing-collection and interpretation of information about an object without being in physical contact with the object.
resampling-The calculation of new DN for pixels created during geometric correction of a digital scene, based on the values in the local area around the uncorrected pixels.
reseau marks-Pattern of small crosses added to photographs.
resolution-Ability to separate closely spaced objects on an image or photograph. Resolution is commonly expressed as the most closely spaced line-pairs per unit distance that can be distinguished. Also called spatial resolution.
resolution target-Series of regularly spaced alternating light and dark bars used to evaluate the resolution of images or photographs.
resolving power-A measure of the ability of individual components. and of remote sensing systems, to separate closely spaced targets.
reststrahlen band-In the IR region, refers to absorption of energy as a function of silica content.
return-In radar, a pulse of microwave energy reflected by the terrain and received at the radar antenna. The strength of a return is referred to as return intensity.
return-beam vidicon (RBV)- A system in which images are formed on the photosensitive surface o a vacuum tube; the image is scanned with an electron beam and transmitted or recorded. Landsat 3 used a pair of RBV's to acquire images.
ringing-Fringe-like artefacts produced at edges by some forms of spatial-frequency filtering.
rods-The receptors in the retina that are sensitive to brightness variations.
roll-Rotation of an aircraft that causes a wing-up or wing-down attitude.
roll compensation system-Component of an airborne scanner system that measures and records the roll of the aircraft. This information is used to correct the imagery for distortion due to roll.
rough criterion-In radar, the relationship between surface roughness, depression angle, and wavelength that determines whether a surface will scatter the incident radar pulse in a rough or intermediate fashion.
roughness-In radar, the average vertical relief of a small-scale irregularities of the terrain surface. Also called surface roughness.

S

SAMII-Stratospheric Aerosol Measurement experiment, carried by Nimbus-7.
SAMS-Stratospheric and Mesospheric Sounder, carried by Nimbus-7.
satellite-An object in orbit around a celestial body.
saturation-In the IHS system, represents the purity of color. Saturation is also the condition where energy flux exceeds the sensitivity range of a detector.
SBUV-Solar Back-scatter Ultraviolet Instrument, carried by NOAA satellites.
scale-Ratio of distance on an image to the equivalent distance on the ground.
scan line-Narrow strip on the ground that is swept by IFOV of a detector in a scanning system.
scanner-An imaging system in which the IFOV of one or more detectors is swept across the terrain.
scanner distortion-Geometric distortion that is characteristic of cross-track scanner images.
scan skew-Distortion of scanner images caused by forward motion of the aircraft or satellite during the time required to complete a scan.
scattering-Multiple reflections of electromagnetic waves by particles or surfaces.

scattering coefficient curves-Display of scatterometer data in which relative backscatter is shown as a function of incidence angle.
scatterometer-Nonimaging radar device that quantitatively records backscatter of terrain as a function of incidence angle.
scene-Area on the ground that is covered by an image or photograph.
scotopic vision-Vision under conditions of low illumination, when only the rods are sensitive to light. Visual acuity under these conditions is highest in the blue part of the spectrum.
Seasat-NASA unmanned satellite that acquired L-band radar images in 1978.
sensitivity-Degree to which a detector responds to electromagnetic energy incident on it.
sensor-Device that receives electromagnetic radiation and converts it into a signal that can be recorded and displayed as either numerical data or an image.
Shuttle imaging radar (SIR)-L-band radar system deployed on the Space Shuttle.
sidelap-Extent of lateral overlap between images acquired on adjacent flight lines.
side-looking airborne radar (SLAR)-An airborne side scanning system for acquiring radar images.
side-scanning sonar-Active system for acquiring images of the seafloor using pulsed sound waves.
side-scanning system-A system that acquires images of a strip of terrain parallel with the flight or orbit path but offset to one side.
signal-Information recorded by a remote sensing system.
signal to noise radio (S/N)-The ratio of the level of the signal carrying real information to that carrying spurious information as a result of defects in the system.
silver halide-Silver salts that are especially sensitive to visible light and convert to metallic silver when developed.
SIR-Shuttle Imaging Radar, synthetic-aperture radar experiments carried aboard the NASA Space Shuttle in 1981 and 1984.
Skylab-U.S.earth-orbiting workshop that housed three crews of three astronauts in 1973 and 1974.
skylight-Component of light that is strongly scattered by the atmosphere and consists predominantly of shorter wavelengths.
slant range-In radar, an imaginary line running between the antenna and the target.
slant-range distance-Distance measured along the slant range.
slant-range distortion-Geometric distortion of a slant-range image.
slant-range image-In radar, an image in which objects are located at positions corresponding to their slant-range distances from the aircraft path. On slant-range images, the scale in the range direction is compressed in the near-range region
SLAR-Side-looking airborne radar.
SMIRR-Shuttle Multispectral Infrared Radiometer, a non-imaging spectroradiometer carried by the NASA Space Shuttle covering ten narrow wavebands in the 0.5-2.4 m range.
SMMR-Scanning Multichannel Microwave Radiometer, carried by Nimbus-7.
smooth criterion-In radar, the relationship between surface roughness, depression angle, and wavelength that determines whether a surface will scatter the incident radar pulse in a smooth or intermediate fashion.
software-Programs that control computer operations.
sonar-Acronym for sound navigation ranging. Sonar is an active form of remote sensing that employs sonic energy to image the seafloor.
Space Shuttle-U.S. manned satellite program in the 1980s, officially called the Space Transportation System (STS).
Space Station-A planned series of three polar-orbiting, sun-synchronous satellites to be launched by NASA, the European Space Agency, and the Japanese Space Agency in the 1990s. They will carry a large range of remote-sensing devices.
spatial-frequency filtering-The analysis of the spatial variations in DN of an image and the separation or suppression of selected frequency ranges.
specific heat-The ratio of the heat capacity of unit mass of a material to the heat capacity of unit mass of water.

spectral hue-A hue which is present in the spectral range of white light analysed by a prism or diffraction grating.
spectral reflectance-Reflectance of electromagnetic energy at specified wavelength intervals.
spectral sensitivity-Response, or sensitivity, of a film or detector to radiation in different spectral regions.
spectral vegetation index-An index of relative amount and vigor of vegetation. The index is calculated from two spectral bands of AVHRR imagery.
spectrometer-Device for measuring intensity of radiation absorbed or reflected by a materiel as a function of wavelength.
spectroradiometer-A device which measures the energy reflected or radiated by materials in narrow EM wavebands.
spectrum-Continuous sequence of electromagnetic energy arranged according to wavelength or frequency.
specular-Refers to a surface that is smooth with respect to the wavelength of incident energy.
SPOT-Systeme Probatoire d'Observation del la Terre. Unmanned French remote sensing satellite orbiting in the late 1980s.
Stefan-Boltzmann constant- 5.68×10^{-12} W . cm^{-2} .K^{-4}.
Stefan-Boltzmann law-States that radiant flux of a blackbody is equal to the temperature to the fourth power times the Stefan-Boltzmann constant.
Stereo base-Distance between a pair of correlative points on a stereo pair that are oriented for stereo viewing.
stereo model-Three-dimensional visual impression produced by viewing a pair of overlapping images through a stereoscope.
stereo pair-Two overlapping images or photographs that may be viewed stereoscopically.
stereopsis-The ability for objects to be perceived in three dimensions as a result of the parallax differences produced by the eye base.
stereoscope-Binocular optical device for viewing overlapping images or diagrams. The left eye sees only the left image, and the right eye sees only the right image.
SSU-Stratosphere Sounding Unit, carried by NOAA-series satellites.
subscene-A portion of an image that is used for detailed analysis.
subtractive primary colors-Yellow, magenta, and cyan. When used as filters for white light, these colors remove blue, green and red light, respectively.
sunglint-Bright reflectance of sunlight caused by ripples on water
sun-synchronous-Earth satellite orbit in which the orbit plane is nearly polar and the altitude is such that the satellite passes over all places on earth having the same latitude twice daily at the same local sun time.
sun-synchronous orbit-a polar orbit where the satellite always crosses the Equator at the same local solar time.
supervised classification-Digital-information extraction technique in which the operator provides training-site information that the computer uses to assign pixels to categories.
surface phenomenon-Interaction between electromagnetic radiation and the surface of a material.
surface roughness-See roughness.
synthetic-aperture radar (SAR)-Radar system in which high azimuth resolution is achieved by storing and processing data on the Doppler shift of multiple return pulses in such a way as to give the effect of a much longer antenna.
synthetic stereo images-Stereo images constructed through digital processing of a single image. Topographic data are used to calculate parallax.
system-Combination of components that constitute an imaging device.
systematic distortion-Geometric irregularities on images that are caused by known and predictable characteristics.

T

target-Object on the terrain of specific interest in a remote sensing investigation.
TDRS-Tracking and Data Relay Satellite
telemeter-To transmit data by radio or microwave links.
terrain-Surface of the earth.
texture-Frequency of change and arrangement of tones on an image.
Thematic Mapper (TM)- A cross-track scanner deployed on Landsat that records seven bands of data from the visible through the thermal IR regions.
thermal capacity (c)- See heat capacity.
thermal conductivity (K)- Measure of the rate at which heat will pass through a material, expressed in calories per centimeter per second per degree Centigrade.
thermal crossover-On a plot of radiant temperature versus time, the point at which temperature curves for two different materials intersect.
thermal diffusivity (k)- Governs the rate at which temperature changes within a substance, expressed in centimeters squared per second.
thermal inertia (P)-Measure of the response of a material to temperature changes, expressed in calories per square centimeter per square root of second.
thermal IR-IR region from 3 to 14 µm that is employed in remote sensing. This spectral region spans the radiant power peak of the earth.
thermal IR image- Image acquired by a scanner that records radiation within the thermal IR band.
thermal IR multispectral scanner (TIMS)- Airborne scanner that acquires multispectral images within the 8-to-14mm band of the thermal IR region.
thermal model-Mathematical expression that relates thermal and other physical properties of a material to its temperature. Models may be used to predict temperature for given properties and conditions.
thermography-Medical applications of thermal IR images. Images of the body, called thermograms, have been used to detect tumors and monitor blood circulation.
THIR-Temperature-Humidity Infrared Radiometer, carried by Nimbus-7.
tie-point-A point on the ground which is common to two images. Several are used in the coregistration of images.
TIMS-Thermal IR multispectral scanner.
TM-Thematic mapper.
tone-Each distinguishable shade of gray from white to black on an image.
topographic inversion-An optical illusion that may occur on images with extensive shades. Ridges appear to be valleys, and valleys appear to be ridges. The illusion is corrected by orienting the image so that the shadows trend from the top margin of the image to the bottom.
topographic reversal-A geomorphic phenomenon in which topographic lows coincide with structural highs and vice versa. Valleys are eroded on crests of anticlines to cause topographic lows, and synclines form ridge, or topographic highs.
TOVS-TIROS Operational Vertical Sounder.
Tracking and Data Relay Satellite (TDRS)-Geostationary satellite used to communicate between ground receiving stations and satellite such as Landsat.
training area-A sample of the Earth's surface with known properties; the statistics of the imaged data within the area are used to determine decision boundaries in classification.
trade-off-As a result of changing one factor in a remote sensing system, there are compensating changes elsewhere in the system; such a compensating change is known as a trade-off.
training site-Area of terrain with known properties or characteristics that is used in supervised classification.
transmissivity-Property of a material that determines the amount of energy that can pass through the material.
transparency-Image on a transparent photographic material, normally a positive image.
transpiration-Expulsion of water vapor and oxygen by vegetation.
travel time-In radar, the time interval between the generation of a pulse of microwave

energy and its return from the terrain.
tristimulus colour theory-A theory of colour relating all hues to the combined effects of three additive primary colours corresponding to the sensitivities of the three types of cone on the retina.

U

unsupervised classification-Digital information extraction technique in which the computer assigns pixels to categories with no instructions from the operator.
UV-Ultraviolet region of the electromagnetic spectrum ranging in wavelengths from 0.01 to 0.4m.

V

variance-A measure of the dispersion of the actual values of a variable about its mean. It is the mean of the squares of all the deviations from the mean value of a range of data.
VAS- Atmospheric Sounder, carried by GEOS satellites
vector format-The expression of points, lines, and areas on a map by digitized Cartesian coordinates, directions, and values.
vegetation anomaly-Deviation from the normal distribution or properties of vegetation. Vegetation anomalies may be caused by faults, trace elements in soil, or other factors.
vertical exaggeration-In a stereo model, the extent to which the vertical scale appears larger than the horizontal scale.
vidicon-An imaging device based on a sheet of transparent material whose electrical conductivity increases with the intensity of EM radiation falling on it. The variation in conductivity across the plate is measured by a sweeping electron beam and converted into a video signal. Now largely replaced by cameras employing arrays of charge-coupled devices (CCDs).
vignetting-A gradual change in overall tone of an image from the centre outwards, caused by the imaging device gathering less radiation from the periphery of its field of view than from the centre. Most usually associated with the radially increasing angel between a lens and the Earth's surface, and the corresponding decrease in the light-gathering capacity of the lens.
visible radiation-Energy at wavelengths from 0.4 to 0.7mm that is detectable by the human eye.
visual dissonance-The disturbing effect of seeing a familiar object in an unfamiliar setting or in an unexpected colour.
VISSR- Visible Infrared Spin-Scan Radiometer carried by the GOES satellites.
volume scattering-In radar, interaction between electromagnetic radiation and the interior of a material.

W

watt (W)-Unit of electrical power equal to rate of work done by one ampere under a potential of one volt.
wavelength-Distance between successive wave crests or other equivalent points in a harmonic wave.
Wien's displacement law-Describes the shift of the radiant power peak to shorter wavelengths as temperature increases.

X

X band-Radar wavelength region from 2.4 to 3.8 cm.

TOP

Y

yaw-Rotation of an aircraft about its vertical axis so that the longitudinal axis deviates left or right from the flight line.

Z

Questions, Comments or Problems?

Please send e-mail to:

jeffw@ldeo.columbia.edu

PREVIOUS RST NEXT

Primary Contact: Nicholas M. Short, Sr.

Glossary Author: Jeff Weissel (jeffw@ldeo.columbia.edu)