

ARCHIVED IMAGERY: THE UNDISCOVERED COUNTRY AND THE NEED FOR EXPLORATION

Roger L. King

Department of Electrical and Computer Engineering
Box 9571

Mississippi State University
Mississippi State, MS 39762-9571 U.S.A.
Email: rking@ece.msstate.edu

1. ABSTRACT

To address the environmental and economic challenges of the 21st century it is incumbent upon the global community to evolve and sustain a global observation network. These observations serve as the foundation for the models that are used to describe Earth processes and can be used to predict the effect of different forcings on the planet (natural and anthropogenic). As this observational data accumulates in global archives new opportunities become available for knowledge discovery about the Earth system. However, access to these observational data is optimized for the science teams for whom the instruments were launched and access by operational users may be problematic. Also, sensor fusion and data mining algorithms are generally not considered as necessary exploration modes of the archives. This paper will address the need for global observations from a variety of vantage points, some descriptions of global observation archives within the United States, recommendations for better access to image archives to facilitate image mining, and the research agenda for data archiving and distribution systems being developed by the IEEE Geoscience and Remote Sensing Society.

2. NEED FOR GLOBAL OBSERVATIONS

In early March 2004, the U.S. Census Bureau's POPClock (population clock) (www.census.gov/main/www/popclock.html) estimated the US population at 292,726,343 and the world's population at 6,352,233,366. In mid 1999, the number of people on Earth was 6 billion, which means in the little over four intervening years, the world's population increased by an amount equivalent to 120% of the US population. World population should exceed 8 billion by 2025. As depicted in Fig. 1, the need for global observations of anthropogenic impacts on the planet is becoming more important.

Over the last decade, a series of international meetings addressed the challenges of managing the Earth's resources to meet the basic needs of increasing populations while preserving the environment. These included the Rio Earth Summit in 1992, the Johannesburg World Summit on Sustainable

Development in 2002, the Earth Observation Summit in July 2003, and the annual UN Framework Climate Convention (UNFCCC) Conference of the Parties (COP).

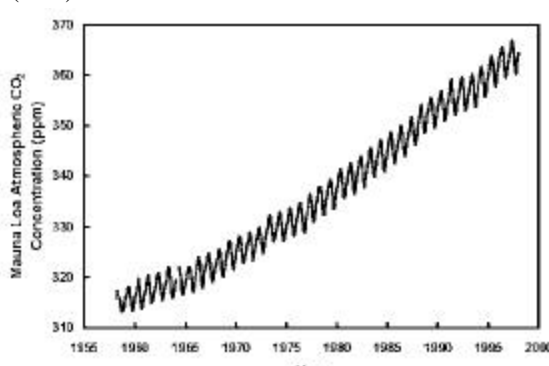


Fig. 1. 40-year trend in CO₂ observations made from an *in situ* sensor.

However, can the Earth system keep up with such great demands on its natural resources to accommodate future generations? Studies report the inherent difficulties associated with funding the growth in the Earth's physical infrastructure (transportation networks, desalination facilities, and so on) to keep up with the population expansion. Therefore, since it is not practical to expand the world's physical infrastructure at the rate required to keep pace with the needs of projected population growth, then is the solution to develop an effective global information infrastructure? It's a start. This infrastructure would consist of observation systems to collect data about natural resources and the means through which this data can be synthesized to generate information. Data from this infrastructure would then be interpreted to produce the knowledge necessary for optimizing resource management.

The challenge then is to establish an approach based on the capacity of Earth-science information to support decision makers in establishing policies and management solutions to better utilize Earth's resources (food, water, energy, and so on) for the global society's good [1]. The following assumptions underpin the need for a global observation and information infrastructure:

- The Earth's resources (for example, food, water, energy, and land) available to serve the growing population are limited [2,3,4,5].
- The Earth system is affected by anthropogenic influences [6].
- Our understanding of the Earth system as a dynamic set of interactive processes (involving oceans, land, and atmosphere) is increasing.
- As the world's population continues to increase, an increasing number of people will share the Earth's finite resources. The equitability of this division will be a driving force in world political stability.
- The countries of the world are challenged to balance economic and Earth resources' security and stewardship.
- Basic needs of the world's population are not effectively being served today (for example, water quality and quantity).

A major component of this challenge is to engage the information systems communities (computational, science, and engineering) to begin to address the need for accessibility to existing archived data sets and to the development of tools to mine the world's archives of measured data. Results from this activity can lead to discovery of new knowledge and understanding that will assist policy makers in meeting the needs of the Earth's burgeoning population. As illustrated in Fig. 2, the need for global observations is now being fulfilled by a myriad of sensors with multiple vantage points [7].



Fig. 2. Global observations from multiple vantage points.

3. GLOBAL OBSERVATION ARCHIVES IN THE UNITED STATES OF AMERICA

Civil global observations within the United States fall primarily under the auspices of three agencies – National Aeronautics and Space Administration (NASA), National Oceanic and Atmospheric Administration (NOAA), and the United States Geological Survey (USGS).

3.1 Archive Descriptions - NASA

The NASA Distributed Active Archive Centers (DAACs) are the data management and user services branches of NASA's Earth Observing System Data and Information System (EOSDIS). The DAACs process, archive, document, and distribute data from NASA's past and current Earth science research satellites and field measurement programs. The DAACs were established in the early 1990s, and each DAAC serves a specific science discipline (<http://nasadaacs.eos.nasa.gov/about.html>).

In 2001, in light of ongoing scientific, technological, and institutional changes, the DAACs created the DAAC Alliance in order to collaborate with other organizations managing science data. The Alliance includes the Alaska Satellite Facility (ASF) DAAC, GSFC Earth Sciences (GES) DAAC, Global Hydrology Resource Center (GHRC), Land Processes (LP) DAAC, Langley Atmospheric Sciences Data Center (LaRC) DAAC, National Snow and Ice Data Center (NSIDC) DAAC, Oak Ridge National Laboratory (ORNL) DAAC, Physical Oceanography (PO) DAAC, and Socioeconomic Data and Applications Center (SEDAC). Current holdings within each DAAC are shown in Fig. 3.

The Alaska Satellite Facility downlinks, processes, archives, and distributes SAR data from the European Space Agency's ERS1 and ERS2 satellites, NASA's JERS-1 satellite, and the Canadian Space Agency's RADARSAT-1 satellite. The GES DAAC is the archive for NASA's Earth Science Enterprise's Ocean Color, Hydrology, Atmospheric Chemistry and Dynamics, and Land Biosphere data and information, as well as data and information from other related disciplines. The Global Hydrology Resource Center (GHRC) provides both historical and current Earth science data, information, and products from satellite, airborne, and surface-based instruments. It encompasses the data and information system, supporting product generation, archival, and distribution of research quality and operational data sets for the Lightning Imaging Sensor (LIS), Optical Transient Detector (OTD), and a variety of passive microwave data sets.

The Land Processes DAAC processes, archives, and distributes land-related data collected by EOS sensors,

thereby promoting the inter-disciplinary study and understanding of the integrated Earth system. The role of the Land Processes DAAC includes the processing and distribution of ETM+ data acquired by Landsat 7, higher-level processing and distribution of ASTER data, and the distribution of MODIS land products derived from data acquired by the Terra and Aqua satellites.

The Langley Atmospheric Sciences Data Center is responsible for processing, archiving, and distributing Earth science data related to radiation budget, clouds, aerosols, and tropospheric chemistry. The National Snow and Ice Data Center serves as a cryospheric focal point for the scientific and educational communities. It includes resources such as, State of the Cryosphere, which provides an overview of the status of snow and ice as indicators of climate change.

The Oak Ridge National Laboratory DAAC is a source for biogeochemical and ecological data useful for studying environmental processes. These data have been collected on the ground, from aircraft, or by satellite or have been generated by computer models. The extent of data ranges from site-specific to global, and durations range from days to years. The Physical Oceanography DAAC is responsible for archiving and distributing data relevant to the physical state of the ocean.

The Socioeconomic Data and Applications Center focuses on human interactions in the environment. Its mission is to develop and operate applications that support the integration of socioeconomic and Earth science data and to serve as an "Information Gateway"

between the Earth and social sciences.

3.2 Archive Descriptions - NOAA

NOAA's imagery is managed through the National Environmental Satellite, Data, and Information Services (NESDIS) (<http://www.nesdis.noaa.gov/datainfo.html>). NESDIS operates four National Data Centers for Climate, Geophysics, Oceans, and Coasts (<http://nndc.noaa.gov/>).

The National Climatic Data Center (NCDC) archives 99 percent of all NOAA data, including over 320 million paper records; 2.5 million microfiche records; with over 1.2 petabytes of digital data residing in a mass storage environment. NCDC has satellite weather images back to 1960. NCDC annually publishes over 1.2 million copies of climate publications that are sent to individual users and 33,000 subscribers. NCDC maintains over 500 digital data sets, receives almost 2,000,000 requests each year, and records over 100 million hits per year on the website.

The National Geophysical Data Center (NGDC) provides scientific stewardship, products and services for geophysical data describing the solid earth, marine, and solar-terrestrial environment, as well as earth observations from space.

The National Oceanographic Data Center (NODC) serves the Nation with data and information for understanding the ocean and its role in our lives. NODC archives and provides public access to oceanographic

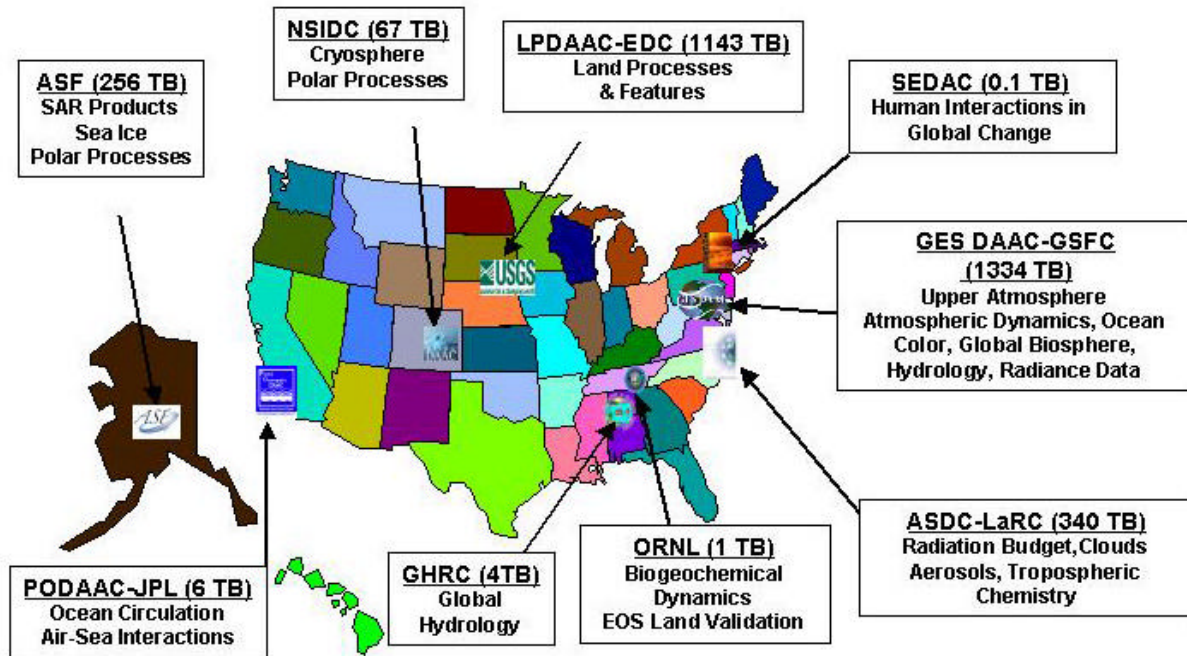


Fig. 3. Holdings in NASA and USGS archives.

observational data and products, provides scientific oceanographic data services, and conducts assessments of the ocean environment. The NODC manages the world's largest collection of publicly available oceanographic data. NODC holdings include in situ and remotely sensed physical, chemical, and biological oceanographic data from coastal and deep ocean areas. These were originally collected for a variety of operational and research missions by U.S. Federal agencies, including the Department of Defense (primarily the U.S. Navy); by State and local government agencies; by universities and research institutions; and private industry. NODC holdings currently contain in excess of one terabyte of data extending back over one hundred years, and the volume is expected to grow exponentially as new ocean observing systems are deployed.

The National Coastal Data Development Center (NCDDC) has a mission to access and integrate diverse coastal data distributed in multiple repositories and provide these data to users via the Internet using established and emerging technologies. They accomplish this by maintaining a searchable metadata catalog of coastal data, developing gateways to data repositories and using middleware technology that provides data in user specified formats.

3.3 Archive Descriptions - USGS

The Earth Resources Observation Systems (EROS) Data Center (EDC) of the USGS is a data management, systems development, and research field center (<http://edc.usgs.gov/>). The EDC is the national archive of remotely sensed images of the Earth's land surface (e.g., Landsat). These data are acquired by civilian satellites and aircraft and used to study a wide range of natural hazards, global environmental change, and economic development and conservation issues. It is co-located with the NASA Land Processes DAAC since they serve the same mission.

The EDC filled over 36,600 orders for imagery during the 2003 fiscal year (<http://edc.usgs.gov/about/reports/sales2003.pdf>). This was down from a peak of about 60,000 orders in fiscal year 2001. At the end of fiscal year 2003, the EDC had archived nearly 12.5 million frames of photographic data. This included nearly 2.9 million frames of Landsat photographic data. An interesting statistic is that 31 years of Landsat 1-5 data amounted to 165 terabytes, while only 4 years of Landsat 7 required 269 terabytes of storage. As can be seen from Fig. 3, over 1 petabyte of imagery related to global land processes and features are held in the co-located USGS and NASA archives. Fig. 4 shows the exponential growth of just the portion of the archive maintained by the USGS EDC.

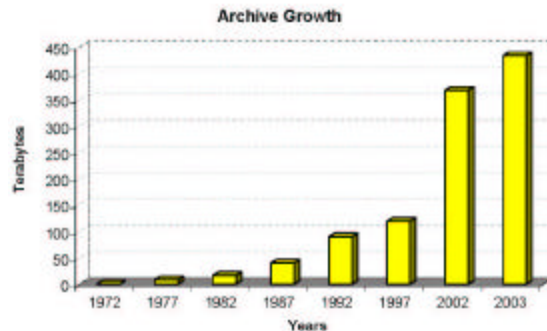


Fig. 4. Growth of USGS EDC archives since its inception.

4. IMAGE ARCHIVES – THE UNDISCOVERED COUNTRY

4.1 Data Discovery and Mining

Access to image and data archives related to the Earth system is becoming a more important research topic and, as previously described, there are several petabytes of imagery already in distributed US archives. With the enormity of imagery and datasets being held for the greater good, it is a non-trivial task to just discover and assemble the right data for knowledge discovery.

Historically, NASA's research satellites were used in gaining a better understanding of Earth system processes to enable the prediction of natural and anthropogenic forcings on the system. NOAA, on the other hand, has an operational mission and much of its data and imagery is used for real-time or near real-time predictions about oceanic and atmospheric phenomena. USGS hosts the premier long-term global land observations - Landsat series of satellites.

All three agencies provide access to imagery and other data they have in archive. Data and imagery is typically served to a user based on some set of parameters (location, time, sensor, etc.) specified by the user. Some datasets are free, while others are assessed a fee. One common characteristic of all the archives is that you are provided your data for off-archive exploitation. Exploitation of an archive's datasets for knowledge discovery through data mining algorithms is not permissible, nor does the technology exist to search the myriad of distributed archives in response to a user's query.

Recognizing the need for greater access to archived datasets and in preserving them until tools for knowledge discovery become available, the IEEE Geoscience and Remote Sensing Society's Technical Committee on Data Archiving and Distribution (DAD) is developing a research agenda. Also, the archive holders (e.g., NASA) are beginning to work toward more open access to datasets by operational users and to

identify approaches to facilitate access by the broader operational and educational communities.

4.1 IEEE GRSS DAD Research Agenda

Remotely sensed data streams or datasets often push available transmission and storage technology to extreme limits, and thus require special techniques in their handling, distribution, application, rendering, fusing, mining, and compression. The IEEE GRSS DAD (www.ieee.org/grss-dad) promotes the study of the manipulation and rendering of large data sets for geoscientific purposes.

The Technical Committee has the following charge from the IEEE GRSS:

To provide recommendations and responses to issues related to the archiving and distribution of remotely sensed geospatial and geotemporal data, and on how new media, transmission means, and networks will impact the archiving, distribution, and format of remotely sensed data. Also, to study the impact of media, channel, and network scaling on the archiving and distribution of data.

The DAD Technical Committee represents an international cross-section of industry, government, and academia (see Figs. 5 & 6).

The membership of the technical committee has been developing a research agenda to propose to the international Earth observation data archiving community. The following list represents the issues identified thus far. It is important to note that these topics have not been vetted in an open format yet, but

just represent a listing of current ideas.

- Archival media readability and integrity: How long can current media be expected to survive, and will it be readable after two or three generations of new media? How can we insure that critical data survive this process of technological evolution with integrity?
- Archival site stability: How can we insure that archival sites (especially those referenced in the archival literature) remain accessible for reasonable periods of time? How long a time period should be considered minimal for public access?
- Archival site size versus access bandwidth trends: Do the Moore's law exponents for storage media and internet bandwidth suggest a trend that capacity will eventually exceed ability to access data in reasonable time? If so, should compression, filtering, and estimation/detection processing algorithms along with data originator, data format, calibration issues, etc. which help researchers to experiment with raw data be located at the archival site?
- Data visualization: How can we provide on-line visualization tools that can assist users in identifying meaningful data subsets within large sets?
- An important theme for this research priority is the arrangement of archive data in a user-selected sequence (or hierarchy) in order to facilitate selecting coincident and co-located data. This is particularly important as work on data fusion is growing rapidly and the synergistic use of multi-platform data is so much encouraged. This proposal is for an off-line browsing tool. As for an on-line tool, it is certainly even more important and needed. It can be identified as a priority but any reasonable solution to bring on-line data from different sources into one system will be technically challenging and expensive.
- As we intend to provide data and information to the application communities, it might be useful to identify products of most socio/economic values along with the most recent and robust algorithms to generate those products. The list should be updated regularly as sensors and applications continue to evolve. We want, for example to make the user community aware that DEM can be produced accurately from SAR interferometry, or sea ice surface temperature has become available from IR channels, or a new vegetation index has been produced from MODIS data and so on.
- Data compression. Recently, there has been much work on this topic. Which kind of

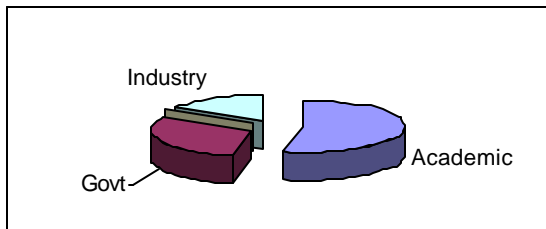


Fig. 5. IEEE GRSS DAD professional categorization.

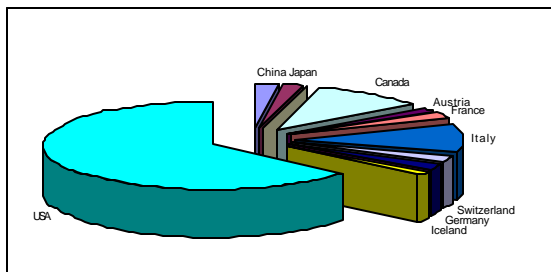


Fig. 6. IEEE GRSS DAD international representation.

compression do we need for DAD purposes? This point is also related to the network infrastructure (and bandwidth) through which users are going to access the archives. Can we use lossless compression? Do we need near-lossless or even lossy compression, and in the affirmative, what are the quality issues related to data alteration? Moreover, how do the archive access networks impact on the selection of data compression tools?

- Security (Authentication and Copyright Protection). These topics are hot ones in the multimedia field, and may apply to remote sensing data as well. One may want to provide a guarantee to the user, that the data they received are actually the ones they expected, i.e. nobody maliciously intercepted the data transmission and modified the data. Moreover, a data distributor may want to be able to enforce copyright protection on the distributed data, in order to check that licensing agreements are not violated by the users through improper use of the data (e.g., illegal copy and distribution). Which technologies can be used to this purpose? In the multimedia field digital watermarking is becoming the key technology, though it arises the same quality issues as compression. Otherwise, encryption may also be considered.
- Quality assessment. Every time an operation (for example lossy compression) alters the remote sensing data, e.g. in order to make the data transmission feasible, a user may be concerned on the quality of the altered data with respect to the original data or, more in general, on the possible decrease of added-value. So far, sufficiently general methodologies to assess data quality do not exist (or they are quite simplistic), though they are crucial to the deployment of signal processing techniques on remote sensing data.
- Data Archiving and Distribution will influence standards of relevance to GRSS.

4.2 NASA EOS-SWGD

NASA's Earth Observing System (EOS) Science Working Group on Data (SWGD) recently held a workshop on Data Access and Usability. The workshop included panel discussions by a range of users in a variety of discipline areas who described their objectives, the type of EOS data they used, how they got the data and how usable they found the data. The workshop panels represented:

- Climate Researchers
- Earth Science Researchers
- Applications and Operation Users

- Education and Outreach Users.

The author of this paper chaired the Applications and Operation Users Panel. The SWGD was especially interested in any barriers users encountered, how the barriers were overcome, and welcomed suggestions for improvements. The full workshop report is available at <http://swgd.gsfc.nasa.gov/>.

There were a number of common threads that crosscut the various user communities. They fell into the following categories: DAAC services, User Interface, Communication, Community-specific Custom Data Sets, Data Timeliness and Calibration.

There were three areas of DAAC services that were identified where improvements would help lower the barriers to the use of EOS data:

- Subsetting
- Tools
- Alternative Data Formats.

Four areas of the EOS user interface were identified where improvements would lower the barriers to using the data:

- Reducing the number of clicks
- Deploying portals for different user communities
- Enabling script-based queries
- Adding a semantic interface.

Better communication among holders of datasets and their users is an area that can always be improved. Also, it can become apparent quite rapidly that different users groups prefer data in a format that is "user-friendly" to them. Finally, timely access to datasets is an issue for operational users along with intercalibration of sensors (e.g., the Landsat series).

One of the recommendations from the Applications and Operation Users Panel was "systems and tools that permit image information mining need to be developed." A specific example cited for the report was that "The archives are rich with information that needs to be exploited to gain new insights for use in science understanding and operational decision support." The panel was asked to prioritize its recommendations and this one was graded as the 2nd highest priority.

The highest priority recommendation would also be of interest to this workshop. It recommended, "EOSDIS should reexamine how to better serve its many user communities (science, operational, and other distinct communities)." Some applicable recommendations of interest to the workshop included:

- Portals tailored to user communities.
- Tools that do simple things with the data (e.g., change detection for the conservation community).

5. CONCLUSIONS

In Act III, Scene 1 of Shakespeare's Hamlet, Hamlet gives his famous "To be, or not to be;" soliloquy. Within this discourse Hamlet also offers the following:

With a bare bodkin? who would fardels bear,
To grunt and sweat under a weary life,
But that the dread of something after death,
The *undiscover'd country* from whose bourn
No traveller returns, puzzles the will
And makes us rather bear those ills we have
Than fly to others that we know not of?

Here he expresses his dread of the *undiscovered country* that faces us all after death. In Star Trek VI: The Undiscovered Country, Chancellor Gorkon of the Klingon High Council expresses the fear that the people of the Klingon Empire and the United Federation of Planets have as they explore the notion of a universe where peace is the norm - their *undiscovered country*.

For those studying Earth system science seeking optimal solutions for the use of Earth's natural resources to meet the needs of the global community, we too have an *undiscovered country*. For us, it is the knowledge about the Earth system that lies hidden in the imagery and data archives of the global observation community. As echoed by Hamlet and Gorkon, there is trepidation in traveling into these unknown lands, but from them we can learn that the reward that lies beyond may greatly surpass our present state. Therefore, there exists a real need to develop the theory and applications of knowledge driven image information mining for exploiting Earth observation datasets.

6. REFERENCES

1. King, R.L. and R.J. Birk, Developing Earth System Science Knowledge to Manage Earth's Natural Resources, *IEEE Computing In Science and Engineering*, Vol. 6, No. 1, pp 45-51, January/February 2004.
2. Rosegrant, M.W., X. Cai, and S.A. Cline, Water and Food to 2025: Averting an Impending Crisis, Int'l Food Policy Research Inst. and Int'l Water Management Inst., 2002; www.ifpri.org.
3. Postel, S., *Last Oasis: Facing Water Scarcity*, W.W. Norton & Co., 1997.
4. Pimentel D., et al., Will Limits Of The Earth's Resources Control Human Numbers?, *Minnesotans For Sustainability*, Feb. 1999; www.mnforsustain.org/pimentel_d_limits_of_earth_s_resources.htm.
5. DOE/EIA -0383, Annual Energy Outlook 2003 with Projections to 2025, Energy Information Agency, U.S. Dep't. of Energy, 9 Jan. 2003.
6. Vörösmarty, C.J. et al., Global Water Resources: Vulnerability from Climate Change and Population Growth, *Science*, Vol. 289, July-Sept. 2000, pp. 284-288.
7. King, R.L., NASA's vision on monitoring natural resources in the 21st century, to be published in Proceedings Second International Workshop on the Analysis of Multi-Temporal Remote Sensing Images (Multitemp-2003), Ispra, Italy, July 16-18, 2003.