# Archive Management: The Missing Component

Howard DIAMOND[1], John BATES[2], David CLARK[3], Robert MAIRS[4], George SHARMAN[5]

[1]National Oceanic and Atmospheric Administration (NOAA)
National Environmental Satellite, Data, and Information Service (NESDIS)
1335 East-West Highway
Silver Spring, Maryland USA 20910

[2]NOAA/NESDIS
National Climatic Data Center (NCDC)
151 Patton Avenue
Asheville, North Carolina USA 28801

[3]NOAA/NESDIS
National Geophysical Data Center (NGDC)
325 Broadway
Boulder, Colorado USA 80305

[4]NOAA/NESDIS
1335 East-West Highway
Silver Spring, Maryland USA 20910

[5]NOAA/NESDIS/NGDC
325 Broadway
Boulder, Colorado USA 80305

howard.diamond@noaa.gov

**Abstract -** The U.S. National Oceanic and Atmospheric Administration's (NOAA) National Environmental Satellite, Data, and Information Service (NESDIS) views the area of Archive Management, as comprising three components: Information Technology Infrastructure, Customer Service, and Scientific Stewardship. This last component, Scientific Stewardship, can be characterized as the long-term preservation of the scientific integrity, monitoring and improving the quality, and the extraction of further knowledge from the data. As our data volumes and complexities have increased, this component, while being recognized as being important, has suffered. Without this component, Archive Management is static, sterile, and lacks the true ability to provide meaningful information and knowledge derived from the archived data. Proper Scientific Stewardship is performed by scientists and data managers knowledgeable in the scientific assessment of a particular data type, and the practice will ensure effective data management where data are 1) used, reprocessed, and reapplied for purposes both intended and newly discovered; 2) improved through repeated analysis and evaluation; and 3) made more accessible with new technologies or innovative structures. This paper will address the importance of Scientific Stewardship in the process of going from data to information; information to knowledge; and ultimately knowledge to wisdom in which environmental data can more easily be used by policy makers in a variety of decision processes. We believe that this practice is a valuable component for ensuring the long-term preservation of data archives while also adding significant value to scientific and technical data.

## 1 - Introduction

Archive management as viewed from a data center has three basic components: (1) the information infrastructure necessary to the store, process, and access the data; (2) the customer service features required to get users efficient and effective access to the data; and (3) finally, what we view as the missing yet most critical area of archive management, Scientific Stewardship. Essentially, Scientific Stewardship is a data management discipline that seeks to ensure the quality, calibration, and use in and for scientific applications beyond the initial use of the data of scientific information. It seeks to provide meaningful and derived information, knowledge, and ultimately wisdom from the archived data that can be applied to solving real-world scientific problems.

As a result of increasing amounts and complexities of environmental data (e.g., satellite and *in-situ* observations), improvements in observational instrumentation, the development of new environmental products and datasets, and the growth of a more sophisticated and knowledgeable user community, NOAA faces major challenges in providing an ever-increasing number of environmental data products to a broader and more demanding user base. NOAA processes and distributes data from its own observing systems as well as those from an increasing number of other environmental agencies. They include the National

Aeronautics and Space Administration (NASA), Department of Defense (DoD), and international environmental satellite operators. These data are used to track hurricanes, tornadoes, and other severe storms; predict future weather and El Niño-Southern Oscillation (ENSO) climate events; monitor numerous ocean phenomena ranging from global coral reef bleaching to the level of the global seas to the quality of the coastal water environment; measure world-wide climate change and the state of the Earth's ozone layer; and detect and monitor a wide range of environmental hazards, including fires, ash from volcanic eruptions, droughts, sea ice, and conditions conducive to flash floods and aircraft icing. Although much progress has been made in recent years, the fact remains that environmental data are underutilized. This situation is being exacerbated as a new generation of more advanced instruments with finer spatial and spectral resolutions producing orders of magnitude more data come on-line. Furthermore, while applications such as weather prediction have reaped remarkable benefits from new satellite and *in-situ* data, newly evolving climate and ocean prediction models are just beginning to exploit these observations. Accelerating the use of a diverse set of environmental observation data in numerical weather, climate, and ocean forecasts is a major challenge. Another challenge is transitioning experimental products from research into operations. An excellent example is that of NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) polar wind retrievals; reference <http://modis.gsfc.nasa.gov/>. These products have been shown to significantly improve weather forecasting by adding wind observations from the data sparse polar areas.

# 2 – Challenges

## 2.1 Accelerating Pace of Change

NOAA spends over a billion dollars each year on the observing systems that collect environmental data from all over the world. These data are used for a wide range of environment prediction programs – from severe weather forecasting in which data must be used within minutes of collection, to climate prediction programs that use data from the past in order to project the climate over the next 100 years. NOAA continues to move forward with the development of new observing systems and initiatives in response to these needs. The enhanced systems and initiatives are producing new data, and are driving the need to develop and maintain new and more accessible data sets to be used in answering important questions and solving real-world problems.

In the past decade, it has become apparent that the Earth's environment is a fully coupled, complementary system. Therefore, to adequately understand what is occurring in the U.S. and our adjacent waters, we must be aware of, and understand, phenomena that are occur globally. Regional or national environmental data are no longer sufficient to fulfill NOAA's mission. In order to meet the needs of modern environmental forecasters, NOAA must acquire near-real-time global observations for local, regional, national, and international analyses and predictions. As more and more nations understand that environmental phenomena are not purely local in cause and effect, there are greater demands for cooperation in the collection and sharing of environmental data, on-line data exchange via the Internet, and the creation of global databases that can be searched via the Internet.

Environmental observation platforms including, but not limited to satellites, have become key to the success of many components of NOAA's missions. However, no one nation – let alone one operation – can provide complete global observational coverage. Therefore, it has become critical for NOAA to negotiate and acquire space based data from a growing number of national and international missions. NOAA must partner with other U.S. Federal government agencies and international consumers/operators of environmental data and information. However, the increasing numbers of these missions are placing ever-increasing demands on NOAA's data processing, distribution, archiving and access systems.

No previous decade has seen the magnitude of changes in the volume of data coming into NOAA for processing and archiving as those experienced in the 1990s. However, that explosive growth is nothing compared to what is going to happen between now and the year 2015. We expect a growth in our data holdings to increase exponentially to over 20 petabytes of archived data. Even as current observing systems continue to provide data, new satellite systems such as NASA's series of Earth Observing Satellites (EOS), the National Polar-orbiting Operational Environmental Satellite System (NPOESS), and the NPOESS Preparatory Project (NPP) will be, or are, going into operations. These systems will provide massive amounts of new data, which will present formidable challenges for NOAA.

A critical function performed by NOAA is to ensure continuity of the Nation's environmental record across heterogeneous observing systems, through changes in sensor technologies over time, through re-calibration of instruments, through changes in instrument site, etc. With the modernization of observing systems and the increasing deployment of new observing systems, the performance of this function is becoming even more critical to solutions for today's climate change issues. The capability to create an accurate, continuous record has enormous consequences from an economic standpoint. For example, capacity planning by local power utilities is heavily dependent upon average temperature conditions derived from the data in NOAA's archives. A decision on building a new power plant at the potential cost of hundreds of millions of dollars may hinge on a fraction of a degree difference in the climatological average. Therefore, a smooth and continuous record reflecting true condition, and not artifacts of the observation process, is critical in making such decisions. NOAA is in a race to preserve historical data before they are lost or have become unrecoverable. Decision makers need this data in order to find solutions to environmental problems both for the present and into the future. These activities involve rescuing data from around the world, and converting data from deteriorating media to modern media.

New users groups are evolving. One such group, weather derivatives, is potentially a $75 billion international industry. This vital new industry includes financial management companies, insurance and reinsurance companies, energy companies, and other industries whose costs are affected by weather and climate extremes in the environment. This new risk management tool uses financial instruments rather than traditional insurance policies to manage the risk of losses due to extremes in weather and climate. Both the development and settlement of these financial instruments are based upon access to accurate, objective, and very timely national and international data from a source that is considered to be accurate and unbiased. NOAA operates the networks that collect these data, and NOAA archives these data; therefore, NOAA is the only source for the data in the U.S. that are the basis for this new industry.

In general, user requests increased throughout the 1990s. Although off-line data user requests doubled, the truly exponential growth has been in the number of on-line users. This currently averages nearly 900,000 per month and is increasing. While on-line requests have increased, it is important to realize that only a small portion of NOAA's data archive is available to the user on-line. As on-line access to NOAA's data expands, the user's average level of technical sophistication and scientific expertise is changing. On-line users are searching for information and answers to specific questions rather than for access to data.

## 3 - Addressing the Challenges

There is reason to expect that the information technology advances we have seen in the last ten years will continue for the foreseeable future. With these advances, NOAA has made significant progress in its ability to archive and provide access, and will continue to leverage on these advancing technologies. Management of these data can be accomplished only through a rapid expansion in storage capacity, increased communications bandwidth, and automation of the means of data ingest, quality control, and access. The Comprehensive Large Array-data Stewardship System (CLASS) program will act as the connection in NOAA's effort to meet these challenges and pave the way to accommodate the additional massive data volumes expected over the next several years.

There are a number of aspects in the successful implementation of CLASS. First, there are those aspects that are purely mechanical or technical in nature. They include, for example, communicating the data from the source to the primary and backup storage locations; quality control and pre-processing of the data; storage of the data on media such as tapes and disk; and, post-processing the data to extract information. In addition, there are those issues concerning the virtual on-line search, retrieval, display, and customer order processing capabilities for the user community. All these various tasks must be accomplished securely, quickly, and efficiently to meet the needs of NOAA's user community. Placing data on-line for access via the World Wide Web is a high priority in accordance with the U.S. Federal Government's overall electronic Government initiative. Data storage and retrieval systems will continue to be upgraded to support effective and efficient access with special focus on World Wide Web interfaces, emerging telephony technologies, and on-line data that support the objectives of the CLASS concept of operations and ensure that the U.S. has access to their data and information.

The ability to ensure on-going scientific stewardship for NOAA's environmental data and information will only be possible through extensive enhancement of NOAA's current data ingest, quality assurance, storage, retrieval, access, and migration capabilities. This goal will be met through the development and implementation of a standardized archive management system. Such an archive management system will have to be integrated with a robust, large-volume, rapid-access storage, and retrieval system that is capable of a number of functions. These functions include the ability to store large volumes of incoming large-array environmental data and operational products, automatically process on-line data requests from users, and provide the requested data on the most appropriate media. This system will provide standardization in media, interfaces, formats, and processes for the very large datasets produced by a diverse array of environmental observing systems. Additionally, the system must facilitate the on-going migration, preservation, and validation of data to new storage media technologies. This system will be modular in design, and built to be integrated with automated real-time or near-real-time systems that deliver data. Transaction processing will be implemented to enable an essentially "hands-off" operation and, where appropriate, the system will allow users to pay for data or services through credit card or automated billing. The system will be able to handle the data flow from current satellite-based (e.g., GOES and POES) as well as ground-based in-situ observing systems (e.g., surface, upper air, and radar observing), and be structured to handle the large increases in data that will come from planned satellite launches, including the MetOp, NPOESS, NPP, and the EOS series.

The target architecture goal will be one which will, through life cycle replacements and upgrades, bring the current NOAA National Data Centers under a single archive and access architecture that will be under formal configuration management control. This will allow elimination of duplication of effort; minimize stand-alone systems, establishment of an infrastructure to accommodate the large-array data sets, and a reduction in the overall operational and system maintenance costs. The foundation system that is being used is NOAA's highly successful and stable Satellite Active Archive (SAA) <http://www.saa.noaa.gov/>. Recognized as a stable, modular, well-built system, the SAA approach provides the maximum flexibility while minimizing development work and costs. The heart of the development centers on the upgrade of communications capabilities, an increase in computer storage and power, the use of commercially available modular hardware and software, and the expansion of the World Wide Web access to the data and information through new or enhanced database management, search, order, browse, and sub-setting techniques.

## 4 - Scientific Stewardship

Much progress has been made in the utilization of environmental data since the first meteorological satellite was launched in 1960. During the 1960s and 1970s, the advances in satellite instruments raced ahead of computing capacity and analysis techniques required to use these data effectively in weather forecasts. Satellite imagery analysis by local weather forecasters was an immediate success, but the use of quantitative satellite data and products in computerized weather forecasting lagged behind. By the late 1990s, however, very fast computers and sophisticated methods of merging vast amounts of satellite and *in-situ* data with numerical forecast models were becoming available.

Significant improvements in weather forecast accuracy came in the last half of the 1990s when, for the first time, satellite sounder data were directly assimilated into U.S. operational weather forecast models. Data assimilation is the process by which weather observations and a short-term (e.g., 6-hour) forecast of the weather variables are merged to obtain the initial conditions needed to make a numerical weather forecast. The important scientific advances that made direct data assimilation possible were: (1) the development of fast radiative transfer models that allow transformation of weather forecast model variables, such as atmospheric temperature and humidity, into radiance, the quantity measured by weather satellites, and (2) techniques by which the model variables are modified so that the radiance out of the models matches the data from the actual observations.

The evolution of satellite capabilities is imposing a requirement for significant scientific effort to accommodate the new data. Between now and the year 2010, the volume of potentially useful and routinely available environmental observation data will grow by a factor of approximately 100,000. This includes data from new operational NOAA and NASA, as well as missions from other countries and agencies. To accelerate the use of this future environmental data in operational weather forecasts, NASA and NOAA have

formed a collaborative Joint Center for Satellite Data Assimilation (JCSDA) in order to develop an end-to-end process for the operational utilization of satellite observations. The JCSDA will be a center distributed among several centers of expertise. These NOAA centers will include NESDIS, the National Centers for Environmental Prediction (NCEP), NOAA's Oceanic and Atmospheric Research (OAR) office, and the NASA Data Assimilation Office (DAO). Each will bring its own area of expertise to the joint effort and by collaborative efforts will make efficient and rapid advances in the use of satellite data in weather forecast models.

The JCSDA will promote the development of common weather forecast models for research and operations. At the present time each U.S. forecasting center runs its own models, and data assimilation advances made at one center are not easily transferred to other centers. Common models will make this process efficient, and components required by data assimilation will be developed for wider community use. This will include community radiative transfer models, surface emissivity models, and surface physics models.

## 4.1 Principles of Scientific Stewardship

The concept of scientific stewardship within NOAA means providing the data and information services necessary to answer the global change scientific questions of highest priority, both now and in the future. The NOAA scientific stewardship program has five principles as follows:

- ***Ensure Observing System Quality***. To provide for the real-time monitoring of climate-scale biases in the global suite of satellite and *in-situ* observing systems by monitoring observing system performance. Since subtle spatial and temporal biases can create serious problems in future use of the data, we must develop the tracking tools necessary for detection of biases in the climate record. These biases can then be minimized or eliminated through efficient communication and coordination of information related to network performance using both *in-situ* and satellite observations.

- ***Provide Basic Information Technology (IT) Support***. To document Earth system variability and change on global, regional, and local scales. This will be accomplished by building and maintaining a high quality base of data and information and establishing the best possible historical perspective critical to effective analysis and prediction. This requires a flexible and efficient use of IT resources in order to quickly adapt to revolutionary IT changes (e.g., telecommunications, commercial off-the-shelf software, and interoperable hardware). The creation of long-term, consistent records requires a long-term commitment of resources to accomplish these tasks.

- ***Develop a Climate Processing System***. To provide the necessary algorithms to ensure that understanding of key climate processes can be derived from space-based systems and the combination of space-based and *in-situ* systems. The best possible scientific understanding of critical climate and global change issues can only be reached when all opinions and ideas can be explored. Thus, an active program of engaging the research community, establishing partnerships with industry, and increasing interactions with local and regional governments to develop a processing system for satellite and *in-situ* observations are envisioned.

- ***Document Earth System Variability***. In order to better document the overall Earth system we need to build and maintain the highest quality climate database and establish the best historical perspective. This will optimize data and information services in order to make research easier and more effective by ensuring that those services are simple, straightforward, direct, and responsive. This will be achieved by establishing end-to-end accountability for establishing long-term, scientifically valid, and consistent records for global change studies. This will ensure that our data and information are available to the maximum amount of users

- ***Enable and Facilitate Future Research***. Because action is required now, and climate and global change societal imperative questions may not come into focus for many years, we must invoke the concept of stewardship to justify this effort. This aspect of stewardship involves providing the basic information technology, hardware, telecommunications, and software support to guarantee that the data can be safeguarded and communicated both within NOAA and to outside users for generations in the future. As new global change imperative questions arise, and in order to safeguard the interests of future

generations, we must make data sets easily available on the Internet and emerging Grid technology outlets. These data sets will be used to update scenarios and assessments, and to identify and respond to emerging questions that the scientific community will be looked to for providing answers.

## 4.2 Building the Climate Record

Over 20 years of satellite and *in-situ* data are now available for climate analysis and detection of climate trends. However, a completely new polar-orbiting environmental satellite system, NPOESS, will begin operation in 2008. The NPOESS system will have new weather and climate monitoring instruments as well as new instruments for monitoring ozone. A prototype of the NPOESS spacecraft, the NPP, will be launched in 2006. A substantial research effort will be required in order to ensure continuity in the earth's climate record between the current operational system that has been in operation over the past 20 years and NPOESS. The proper time and data sets for this research effort will be provided by NPP, which will provide the bridge to NPOESS. The NPP will allow coincident climate observations between the old satellite instruments and the new ones that will begin operation with NPOESS and continue for many years. Construction of a seamless climate record between the current satellites and NPOESS, as well as other instrumented climate data, is a very important yet difficult challenge.

As the concept of Scientific Stewardship continues to evolve within NOAA, a few important notes must be made as, to where it is headed with building the climate record. With respect to observing system performance, by placing an emphasis on this area, we can identify small problems early before they get too big and systematically improve observing system quality. Another area of emphasis deals with the idea of data character. Data character is a philosophy that involves the performance of calibration and inter-calibration while using additional analysis that up until now had only been on the retrieved products. For example, with respect to climate, retrieval techniques, [e.g., Hovmoller diagrams, Empirical Orthogonal Functions (EOF), and time series] have a tendency to amplify any systematic errors, and so we must strive to reduce such errors to the greatest extent possible prior to the application of retrieval algorithms. This also has multiple benefits in that (1) the raw quality of the data is improved for all users who work on the data; (2) the amount of reprocessing the data for new algorithms is minimized; and (3) the overall quality of climate products is improved. We believe that achieving these priorities will result in a long-term archive that is flexible and innovative, appropriately focuses responsibility on NOAA and NESDIS for the preservation of optimal data character, provides for open access to the data by the scientific community and the public, and rapidly tracks technological developments.

## 4.2 Training and Education of the Users

A major challenge is training and educating the large and varied user community on the nature and use of the large selection of environmental data and products. Improving data utilization is planned by increasing training, user interactions and also by data access through web sites. Currently users can access data and many of our operational products through the Internet. Examples include imagery from GOES, ocean color, coral reef bleaching, ocean winds and altimetry, atmospheric temperatures, surface temperature, snow cover and ice, aerosols, fires, clouds, and the vegetation index.

NOAA invests in classroom and computer-based training through the Cooperative Program for Operational Meteorology, Education and Training (COMET), as well as a joint NESDIS/National Weather Service (NWS) cooperative institute [the Cooperative Institute for Research in the Atmosphere (CIRA) and Cooperative Institute for Mesoscale Meteorological Studies (CIMMS)] program called the Virtual Institute for Satellite Integration Training (VISIT). Attendees of COMET are primarily NWS personnel, although representatives from other agencies (for example, DoD forecasters, university faculty and students, etc.) also participate. The VISIT program uses Internet technology to provide distance learning that allows interaction between students and instructors similar to a classroom situation. VISIT provides concurrent instruction to multiple sites. From April 1999 through April 2001 there were 245 training sessions with a total of 4,585 students. In addition to the VISIT interactive classroom, NESDIS maintains, through its web pages, a Virtual Institute with satellite tutorials, and a wide assortment of case studies, and technical information and documents for users. This concept of remote training has been expanding well beyond the U.S. Through the auspices of the World Meteorological Organization, NESDIS scientists are bringing real-time geostationary data and interpretative assistance to Central and South America, and the Caribbean Regional and

Meteorological Training Centers. On the horizon are plans to expand such collaboration and training programs to Europe, Africa, and Asia.

In addition to classroom efforts, NESDIS scientists participate in traditional workshops, professional conferences, and the publication of articles in scientific literature. NOAA also established the CoastWatch program, which is a national network of eight regional offices that serves to provide assistance and data to coastal managers, forecasters, and researchers. Each office is located within an appropriate NOAA line office activity in the region (for example, National Marine Fisheries Laboratories, NWS offices, etc.). By virtue of the fact that CoastWatch regional offices are located "on-site" along the U.S. coasts, we have been able to establish more intimate contact with our users and have therefore become more familiar with their particular resource management issues. CoastWatch has helped users become more aware of the availability of the diverse array of available environmental data, and this has enabled them to make better use of the data as it applies to their local and regional coastal and ocean resource management activities. Due to the projected exponential growth of environmental observation data and products, NOAA will need to continue and expand its user interaction and training to accommodate this broadening user base.

## 4.3 Examples of Scientific Stewardship

### 4.3.1 New Uses and Data Fusion

In parallel with an increasing number of satellite and *in-situ* products and applications, there are a growing number of user communities that NESDIS serves and interacts with. These include weather forecast offices; NOAA's environmental prediction centers specializing in hurricane, severe storm, aviation, space environment, hydrological, ocean, and climate forecasts; the national and international climate community concerned with climate change; federal, state, and local resource managers responsible for coastal environmental monitoring; officials responsible for reacting to environmental hazards (e.g., fires); the aviation and shipping communities; those involved in drought mitigation activities; and local and international groups assessing the health of coral reefs. A major challenge is training and educating this large and varied user community on the nature and use of environmental observation data and products. In addition to the real-time user base, there are a large and growing number of users who work with retrospective archived data largely in support of research on the Earth's weather, climate, and oceans. The challenge of how best to make these vast quantities of satellite and *in-situ* data and products accessible to these communities quickly and at minimum cost will be continuing. This challenge will only grow in along with the exponential increases in data volume and complexity.

### 4.3.2 Data Quality

NOAA's current and near-future environmental instruments have been designed to measure properties of the Earth and its atmosphere for application primarily to weather forecasting. The basic measurement of the satellite instruments is of the intensity of the radiation upwelling from the Earth-atmosphere system, from which geophysical properties are derived mathematically. It goes without saying that the basic radiation measurements need to be highly accurate. This means that the instruments need to be calibrated accurately on orbit. NOAA has a vigorous program to achieve this. The instruments are carefully calibrated before launch, and NASA and NOAA conduct extensive checkouts of the instruments' performance after launch. Finally, the data products are continually validated after the instruments become operational. Yet there is room for improvement. For example, data from the current generation of atmospheric sounders have begun to make an impact on numerical weather forecasts. It is believed that the impact could be enhanced significantly if the accuracy and vertical resolution of the sounder observations were increased. Furthermore, as forecast models become better and computer capabilities increase, the impact of satellite and *in-situ* data will lessen unless their quality increases. The challenge, then, is to develop and deploy operational sounders that provide higher accuracy and vertical resolution. Such sounders are being developed for the NPOESS and GOES-R systems. In addition, NOAA is beginning to enlist the services of the U.S. National Institute of Standards and Technology in order to bring the calibrations of its various observing instruments in line with international standards. The goal is to assure the accuracy of the measurements and to achieve consistency among the measurements made by the satellite and *in-situ* instruments of NOAA as well as those from other agencies.

There is a growing demand to detect climate change from satellite observations. The instruments and NOAA's observing strategies were not designed to provide the long-term continuity, stability, and consistency in the observations that this application requires. Furthermore, NOAA occasionally makes changes in sensor characteristics or measurement techniques, and these present serious problems, as do time gaps in the observing period. For constructing climate-quality data sets, it is imperative that we include supporting ground-based *in-situ* observations, which validate and complement the satellite observations. In 2000, the National Research Council (NRC) responded to a request from the Assistant Administrator of NOAA/NESDIS with a letter report containing recommended ways to improve NOAA's observing strategies in order to facilitate the construction of climate-quality temperature records from data of the microwave sounding unit on the NOAA series of polar-orbiting satellites. The NRC's recommendations also have merit for construction of climate-quality records from other sensors and for other environmental variables. The NRC recommendations were classified in three categories, (1) the satellite observing system, (2) the ground-based observing system, and (3) issues associated with the climate data record.

In the first category, the recommendations included the following elements: (a) maintaining constant local observing times for polar satellites; (b) obtaining continuity or, even better, one-year overlap between observations of successive satellites; (c) improving instrument calibration systems to account for on-orbit drifts in radiometer gain; and (d) making information available on instrument performance status and changes that might affect the observations. Although current satellite systems do not comply with many of these recommendations, NOAA intends to work towards compliance in its future systems, e.g., NPOESS.

In the second category, the recommendations included: (a) maintaining the current observational and primarily radiosonde-based upper air observing system; (b) improving the accuracy and reliability of observations of the stratosphere; (c) assuring temporal continuity and consistency of the data record; (d) generating and making available a "climate data record" of the observations and associated metadata; (e) upgrading the radiosonde temperature and humidity sensors while maintaining continuity of calibration; and (f) exploring options for a significantly improved next-generation atmospheric sounding system. NOAA plans to comply with all of these.

The recommendations from the third category dealing with the climate data record included: (a) reinvigorating the full range of activities within NOAA to ensure a long-term climate record; (b) monitoring and making available the performance and error characteristics of the space-based and *in-situ* networks; and (c) establishing a dialogue and information exchange on the climate data records and the sensors that are their source. Here too, NOAA expects to comply.

### 4.3.3 Long-term Stewardship

The use of a variety of environmental data for climate studies has progressed from being experimental to routine. These data sets have proven to be of high value for climate studies. They have been used in regional and global temperature and upper tropospheric humidity trend studies, in studies of the ozone hole, and in studies of clouds and rainfall. The Intergovernmental Panel on Climate Change (IPCC) and various World Climate Research Programs (WCRP) have in turn, used these products in assessments. Further applications of satellite data to climate studies, particularly for retrieval of column $CO_2$, are currently under development and appear promising.

As we eagerly move into the next generation of instrumentation, it is now clear that the constellation of operational satellites will be the backbone of the long-term global climate observing system. There are major challenges we face, however, in moving to the next generation systems and in preserving and using the past data. How do we maintain a seamless time series of fundamental observations during this transition? How do we deal with the quantum jump in data volume as well as for decades of data? How do we ensure that user access to these data archives is simple, straightforward, direct, and responsive?

Answering these challenges will take a concerted effort by instrument scientists, climate scientists and computer scientists. There must be extensive collaboration between the research and operational climate communities. Computer scientists and climatologists will need to provide a sound and effective means to ensure that all necessary data is preserved and remains accessible in easy-to-use formats. It will also take a long-term commitment to provide resources to enable preservation of the climate archive from the first

generation satellite systems of the 1980s and 1990s, through the transition satellite systems of the EOS era, to the second generation of operational systems in the future NPOESS

### 4.3.4 Target Architecture

The target architecture goal will be one which will, through life cycle replacements and upgrades, bring the current NOAA National Data Centers under a single archive and access architecture that will be under formal configuration management control. This will eliminate the duplication of effort, minimize stand-alone systems, build the infrastructure to accommodate the large-array data sets, and reduce the overall operational and system maintenance costs. The foundation system that is being used is the highly successful and stable Satellite Active Archive (SAA). Recognized as a stable, modular, well-built system, the SAA approach provides the maximum flexibility while minimizing development work and costs. The heart of the development centers on the upgrading of communications capabilities, increasing computer storage and power, use of commercially available modular hardware and software, and expansion of the World Wide Web access to the data and information through new or enhanced database management, search, order, browse, and sub-setting techniques.

## 4.4 Future Developments

As new technologies are developed they will also have to be factored in. Recent developments in Grid technologies, viewed as the Internet infrastructure for the 21st century, is designed to facilitate collaboration in a more seamless distributed processing environment with the intent of having a more uniform data access. Grids, as they are known, are persistent environments that enable software applications to integrate instruments, displays, computational and information resources that are managed by diverse organizations in widespread locations. While this in many cases gets associated with high-end super computing, it is probably the next step in the evolution of the Internet for the dissemination of information in a more seamless manner. A popular example of Grid technology is the SETI@home screensaver that allows the unused computer resource time of PCs to participate in the ongoing Search for Extraterrestrial Intelligence by analyzing data specially captured by the world's various radio telescopes. The Committee on Earth Observing Satellites' (CEOS) Working Group on Information Systems and Services (WGISS), of which NOAA is a member, has recently begun investigating the use of Grid technology in order to facilitate a more seamless exchange of data among satellite data agencies. With the exponential growth in data archives referenced earlier, the ability to "gridify" (for lack of a better term) these extensive archives will make them much more accessible to a wide variety of users. The ultimate goal for WGISS is to have all the CEOS Data Centers on an interconnected Grid with seamless data access with consistent access authentication by the year 2010.

## 4.5 Implementing Scientific Data Stewardship

The actual implementation of scientific data stewardship covers not only the archiving plans for all the various satellite and *in-situ* data sources, but it also involves applications with a number of groups and activities as follows: (1) data character; (2) mission groups; (3) interdisciplinary groups; and (4) external grants. The data character group has the mandate for long-term calibration, inter-calibration, and validation of all sensors; collaborates with existing national and international observing system groups; and assures that customers get the highest quality basic data while also responding to data quality questions. The mission groups are specific to each observing platform (e.g., NPOESS), ramp up during the implementation of the platform and then transition to the data character group during stable operations; these groups have the competency in the specifics of each mission along with complete documented metadata. The interdisciplinary groups address major theme areas (e.g., water and energy cycles) and use all instruments and blend with all data sources to solve climate and global change science questions in order to help provide data and information assessments and options. Finally, the external grants program uses expertise from existing NOAA grants and contracts to assure the involvement of academia and industry; and works with other Scientific Data Stewardship groups to take advantage of directed research with cooperative institutes.

## 5 - Conclusion

As we have shown, the missing element of archiving is Scientific Stewardship. In order to build consistent and high-quality records of environmental observations and produce comprehensive analyses of environmental change, it is imperative for data providers such as NOAA to partner with the scientific community by ensuring the provision of high quality data and services as well as the generation of useful and understandable products that can be easily accessed. This will include the provision of value-added products through the development of new algorithms in order to derive environmental parameters from the instrumental record, the generation and validation of environmental data records from satellite and *in-situ* products that are calibrated by in-situ measurements, and the generation and analysis of products that more accurately identify environmental change. This will require increased access and utility to vast archives of data that will be built over the next 15 years and will require the refinement and development of tools and techniques (e.g., Geographic Information System, Grid Technology, data mining) in order to effectively take advantage of the avalanche of more complex environmental data that is to come.

## References

[1] Clark, D.M. *Science Data Stewardship Through a Global Science Data Network*. Presented at Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data symposium held at Institut Aeronautique et Spatial Complexe Scientifque de Rangueil, Toulouse, France, 5-7 November, 2002. In press.

[2] NOAA. *The Nation's Environmental Data: Treasures at Risk, Report to Congress on the Status and Challenges for NOAA's Environmental Data Systems*, Washington, DC, 2001.