

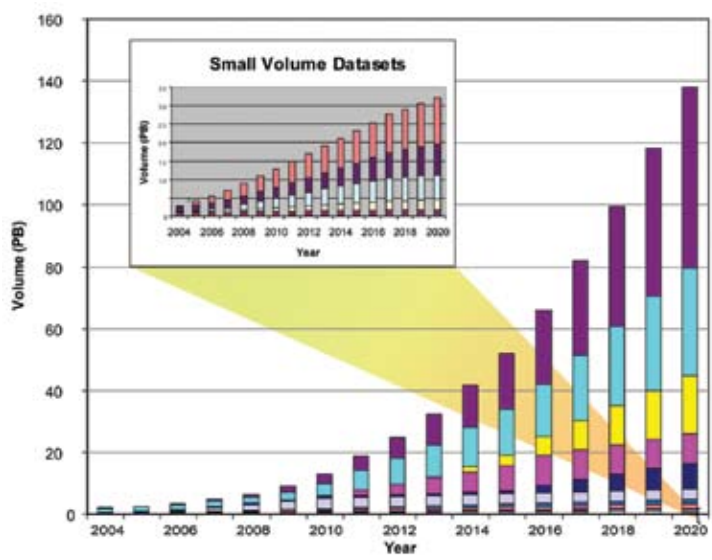
# Environmental Data Management at NOAA: Archiving, Stewardship, and Access

The National Oceanic and Atmospheric Administration (NOAA) collects, manages, and disseminates a wide range of climate, weather, ecosystem and other environmental data used by scientists, engineers, resource managers, policy makers, and others in the United States and around the world. The increasing volume and diversity of NOAA’s data holdings—which include everything from satellite images of clouds to the stomach contents of fish—and a large number of users present NOAA with substantial data management challenges. This report offers nine general principles for effective environmental data management, along with a number of guidelines on how the principles could be applied at NOAA.

The data NOAA collects from satellites, weather stations, computer models, and other sources stretch from the surface of the sun to the core of the Earth. These data are used for environmental monitoring, weather and climate forecasting, water resource management, and a number of other activities that provide a broad range of benefits to

society. Customers of this investment include other federal agencies, state and local governments, industry and business interests, scientists, educators, and the public, including the international community.

Historically, NOAA’s data management activities were driven primarily by the agency’s meteorological, oceanographic, and geophysical operational mission requirements. In recent years, the agency has taken on a much broader range of environmental and geospatial data, including data collected internationally and by other agencies. Currently, NOAA’s National Environmental Satellite, Data, and Information Service (NESDIS) operates three National Data Centers—the National Climatic Data Center, the National Geophysical Data Center, and the National Oceanographic Data Center—along with a number



## Large Volume Datasets

### Space Based Data

- NOAA Polar-orbiting Operational Environmental Satellite System (NPOESS)
- NPOESS Preparatory Project (NPP)
- Geostationary Operational Environmental Satellites (GOES)
- NASA Earth Observing System(Moderate Resolution Spectroradiometer) (EOS MODIS)
- Meteorological Operational Satellite Program (MetOp)

### Earth Based Data

- Weather Radar

### Model Data

- Atmosphere & Ocean

## Small Volume Datasets

### Space Based Data

- Polar Orbiting Earth Satellites (POES)
- Defense Meteorological Satellites Program (DMSP)

### Earth Based Data

- Atmosphere(Weather & Climate)
- Ocean (Weather & Climate)
- Continually Operating Reference Stations (CORS)
- Misc (Mesonets)

NOAA is projecting a sharp increase in the volume of archived environmental data, from 3.5 petabytes today to nearly 140 petabytes (140 billion megabytes) by 2020. SOURCE: NOAA, 2007.

of smaller “centers of data,” such as the National Snow and Ice Data Center, that house specialized data sets. Collectively, these centers are responsible for acquiring, integrating, managing, disseminating, and archiving environmental and geospatial data from worldwide sources to support NOAA’s mission.

Rapid increases in data volumes, data diversity, and data demands have created a significant management challenge for NOAA. Moreover, these challenges must be dealt with in an environment of rapidly evolving technologies and constant budgetary pressures. In response, NOAA asked the National Research Council for principles and guidelines to help identify the observations, model output, and other environmental information that must be preserved in perpetuity and made readily accessible, as opposed to data with more limited storage lifetime and accessibility requirements.

## **Principles for Effective Data Management**

The critical functions of an integrated, end-to-end data management system can be broken into three main elements: (1) archiving, which includes the acquisition, storage, and preservation of data; (2) access, which includes helping users find and obtain the data that meets their needs; and (3) stewardship, which spans archiving and access and includes all the activities that preserve and improve the information content of individual data sets and their associated metadata (information about the data, such as how they were collected).

Fortunately, NOAA has the opportunity to build on a number of successful data management activities that are already providing reliable data for many user communities, as well as several prototype projects designed to improve data discovery, utilization, and integration. NOAA also deserves praise for the steps that it has already taken to evaluate and improve its data management activities, such as its annual reports to Congress on its data management activities, the formation of the Data Archiving and Access Requirements Working Group, and the conceptual plan for the Global Earth Observation Integrated Data Environment.

This report proposes the following nine general principles for effective environmental data management, along with a number of more specific guidelines that explain and illustrate how the principles could be applied at NOAA. All nine principles should be considered equally important.

### ***PRINCIPLE #1: Environmental data should be archived and made accessible.***

The environmental data (including observations, model output, and products derived from these data)

collected by NOAA and its partners constitute an invaluable resource that should be securely archived and made broadly accessible so that a diverse group of users can conduct the analyses and generate the products necessary to describe, understand, and predict changes in the Earth’s environment.

- Although it is impossible to save everything, the goal of NOAA’s data management enterprise should be to ensure that the broadest possible collection of environmental data is archived and made discoverable and accessible to the widest possible range of users.
- Full and open access to data should be a fundamental tenet at all US agencies, including NOAA.

### ***PRINCIPLE #2: Data-generating activities should include adequate resources to support end-to-end data management.***

Data management planning should begin when plans are first made to acquire a new sensor system or other source of environmental data and continue until the archived data are no longer useful. Plans for data management should explicitly address data storage, preservation, and stewardship responsibilities, and sufficient funds should be reserved to acquire, process, store, maintain, update, and provide access to these data for extended periods of time. These activities require resources for both personnel and hardware.

- At the outset of any activity that will generate environmental data, end-to-end data management should be planned and budgeted.
- It is also essential to plan and secure adequate funding for the management of environmental data streams that were collected or generated without sufficient resources reserved for long-term data management.

### ***PRINCIPLE #3: Environmental data management activities should recognize user needs.***

The success of any data management enterprise is judged by its usefulness to current and future users. User involvement in data management planning is critical to make certain that essential data are properly archived and that data access activities are as practical, efficient, and cost-effective as possible.

- All environmental data management activities, including decisions regarding archiving and access of specific data sets as well as the development of the enterprise-wide data management plan, should incorporate substantial and ongoing user input.

***PRINCIPLE #4: Effective interagency and international partnerships are essential.***

Any environmental data management planning process or system needs to include substantial coordination and agreement among the relevant federal agencies and international partners in order to ensure proper data stewardship. Different agencies have different missions with respect to collecting, archiving, and providing access to different types of data, and these missions sometimes overlap or leave gaps in responsibility for critical data sets; for instance, the transition from research to operations has been a longstanding challenge.

- Although no agency can or should be expected to assume complete responsibility for all types of environmental data, NOAA should consider taking a leadership role in archiving and providing access to a broad range of environmental data, including data not traditionally regarded as falling under its operational mission.
- At the very least, NOAA should improve its relationship with several important partners to guarantee that all essential environmental data are archived and made available to users.

***PRINCIPLE #5: Metadata are essential for data management.***

To ensure the enhancement of knowledge for scientific and societal benefit, metadata—information about the data that helps facilitate its understanding and use—should be created and preserved.

- Metadata ideally should include not only descriptive information about the data and how they were originally collected or produced, but also information on the quality and appropriate uses of the data, how they have been managed and processed, and the relationship of the data to data in other collections.
- Whenever possible, NOAA's metadata protocols should be made compatible with international standards and should be expanded to ensure proper preservation and stewardship of data and to facilitate data discovery, access, and integration.

***PRINCIPLE #6: Data and metadata require expert stewardship.***

Scientific data stewardship encompasses all activities that preserve and improve the information content, accessibility, and usability of data and their associated metadata. These activities include maintaining a scalable and reliable infrastructure to support long-term access and preservation, preserving data

access and archive integrity during media migration and software evolution, providing effective data support services and tools for users, and enhancing data and metadata by adding information that is established throughout the data life cycle.

- Scientific data stewardship, with assigned organizational responsibility, should be applied to all environmental data sets managed by NOAA and to their associated metadata to ensure this information is preserved, remains continually accessible, and can be improved as future discoveries build understanding and knowledge.

***PRINCIPLE #7: A formal, ongoing process, with broad community input, is needed to decide what data to archive and what data not to archive.***

It is not possible to save everything, so at some point certain data will need to be designated for reduced archiving and/or access requirements. Original observations, being irreplaceable, are the most important type of data to preserve. Candidates for reduced archiving requirements include data that are obsolete, redundant, or clearly have only short-term uses.

- NOAA needs to establish a high-level, enterprise-wide approach to decide what data to include in their archives.
- In all cases, the decision to archive (or not to archive) data should follow established procedures, should be driven by scientific and societal benefits, and should explicitly incorporate broad community engagement and coordination with other agencies.

***PRINCIPLE #8: An effective data archive should provide for discovery, access, and integration.***

The ability to effectively search and retrieve data is dependent on good metadata and compatible formats. Accessibility to data is predicated on minimizing administrative, technical, and systematic barriers. In addition, all aspects of data access must take the needs of users into account. Since environmental data are generally managed in a distributed environment and user capabilities and needs vary widely, different access mechanisms are needed to accommodate different user requirements.

- Environmental data should be made easily discoverable and readily accessible to a broad range of users, and the data management system should be designed to allow the integrated exploitation of data from multiple sources.

- In general, all aspects of environmental data access would benefit from an enterprise-wide effort to improve and expand existing data access portals, increase the linkages between different portals, and create new portals targeted at particular applications or user groups.
- To the maximum extent possible, users should have full, open, and timely access to all of NOAA's environmental data and metadata.

***PRINCIPLE #9: Effective data management requires a formal, ongoing planning process.***

NOAA needs to continue to ensure the consistency and continuity of numerous data sets in order to meet its legal, mission, and administrative requirements and to fulfill its interagency and international agreements. However, the scale and complexity of NOAA's data archiving and access requirements have reached the point where the ad hoc, discipline-specific, instrument-oriented data management systems and decision-making processes relied on in the past are unlikely to be able to keep up with future demands.

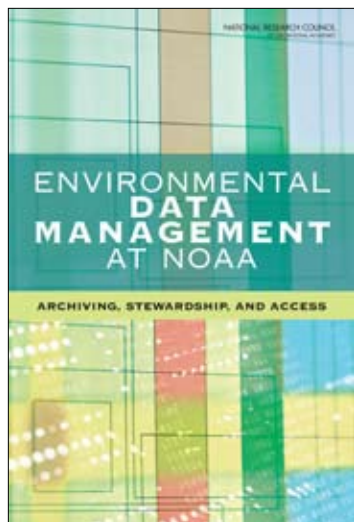
- To improve the coordination, integration, flexibility, adaptability, and cost-effectiveness of its future data management activities, NOAA

should establish and codify an enterprise-wide data management plan that explicitly incorporates all the principles in this report and is in alignment with NOAA's mission, vision, and goal statements.

- The plan should formally delineate individual and shared responsibilities for the archiving, stewardship, and access of all environmental data sets that fall under NOAA's purview.
- NOAA would be well advised to continue moving toward a federated but integrated "system-of-systems" to manage its environmental data.

## Conclusion

Understanding and predicting environmental change are likely to remain critically important activities throughout the 21st century. Hence, it is essential to archive and provide access to the environmental data needed to describe, understand, and predict changes in the Earth System as well as the impacts of these changes. NOAA is clearly dedicated to providing excellent service for its many user communities, and the agency is to be commended for soliciting external advice to make sure it continues to provide effective stewardship of important national assets.



**Committee on Archiving and Accessing Environmental and Geospatial Data at NOAA:** **David A. Robinson** (*Chair*), Rutgers University; **David C. Bader**, Lawrence Livermore National Laboratory; **Donald M. Burgess**, University of Oklahoma; **Kenneth E. Eis**, Colorado State University; **Sara J. Graves**, University of Alabama; **Ernest G. Hildner III**, NOAA Space Environment Center (retired); **Kenneth E. Kunkel**, Illinois State Water Survey; **Mark A. Parsons**, University of Colorado; **Mohan K. Ramamurthy**, University Corporation for Atmospheric Research; **Deborah K. Smith**, Woods Hole Oceanographic Institution; **John R. G. Townshend**, University of Maryland; **Paul D. Try**, Science and Technology Corporation; **Steven J. Worley**, National Center for Atmospheric Research; **Xubin Zeng**, University of Arizona; **Ian Kraucunas** (*Study Director*), National Research Council.



**This report brief was prepared by the National Research Council based on the committee's report.** For more information, contact the Board on Atmospheric Sciences and Climate at (202) 334-3512 or visit <http://nationalacademies.org/basc>. Copies of *Environmental Data Management at NOAA: Archiving, Stewardship, and Access* are available from the National Academies Press, 500 Fifth Street, NW, Washington, D.C. 20001; (800) 624-6242; [www.nap.edu](http://www.nap.edu). This report was sponsored by the National Oceanic and Atmospheric Administration (NOAA).

*Permission granted to reproduce this brief in its entirety with no additions or alterations.*