Preliminary Principles and Guidelines for Archiving Environmental and Geospatial Data at NOAA: Interim Report

Committee on Archiving and Accessing Environmental and Geospatial Data at NOAA, National Research Council

ISBN: 0-309-66153-6, 37 pages, 8 1/2 x 11, (2006)

Preliminary Principles and Guidelines for Archiving Environmental and Geospatial Data at NOAA: Interim Report

Committee on Archiving and Accessing Environmental and Geospatial Data at NOAA

Board on Atmospheric Sciences and Climate

Division on Earth and Life Studies

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

Support for this project was provided by the National Oceanic and Atmospheric Administration under Contract No. DG133RO4CQ0009. Any opinions, findings, and conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number 0-309-10227-8 (Book)

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, D.C. 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

COMMITTEE ON ARCHIVING AND ACCESSING ENVIRONMENTAL AND GEOSPATIAL DATA AT NOAA

DAVID A. ROBINSON *(Chair)*, Rutgers University, New Jersey
DAVID C. BADER, Lawrence Livermore National Laboratory, California
DONALD W. BURGESS, University of Oklahoma, Norman
KENNETH E. EIS, Colorado State University, Fort Collins
SARA J. GRAVES, University of Alabama, Huntsville
ERNEST G. HILDNER, NOAA Space Environment Center (retired)
KENNETH E. KUNKEL, Illinois State Water Survey, Champaign
MARK A. PARSONS, University of Colorado, Boulder
MOHAN K. RAMAMURTHY, University Corporation for Atmospheric Research, Boulder, Colorado
DEBORAH K. SMITH, Woods Hole Oceanographic Institution, Massachusetts
JOHN R. G. TOWNSHEND, University of Maryland, College Park
PAUL D. TRY, Science and Technology Corporation, Williamsburg, Virginia
STEVEN J. WORLEY, National Center for Atmospheric Research, Boulder, Colorado
XUBIN ZENG, The University of Arizona, Tucson

NRC Staff

IAN KRAUCUNAS, Study Director
CHRIS ELFRING, Board Director
ELIZABETH A. GALINIS, Research Associate
ROB GREENWAY, Senior Program Assistant

BOARD ON ATMOSPHERIC SCIENCES AND CLIMATE

ROBERT J. SERAFIN *(Chair)*, National Center for Atmospheric Research, Boulder, Colorado
M. JOAN ALEXANDER, NorthWest Research Associates/CORA, Boulder, Colorado
FREDERICK R. ANDERSON, McKenna Long & Aldridge LLP, Washington, D.C.
MICHAEL L. BENDER, Princeton University, New Jersey
ROSINA M. BIERBAUM, University of Michigan, Ann Arbor
MARY ANNE CARROLL, University of Michigan, Ann Arbor
CAROL ANNE CLAYSON, Florida State University, Tallahassee
WALTER F. DABBERDT, Vaisala Inc., Boulder, Colorado
KERRY A. EMANUEL, Massachusetts Institute of Technology, Cambridge
DENNIS L. HARTMANN, University of Washington, Seattle
PETER R. LEAVITT, Weather Information Inc., Newton, Massachusetts
JENNIFER A. LOGAN, Harvard University, Cambridge, Massachusetts
VERNON R. MORRIS, Howard University, Washington, D.C.
F. SHERWOOD ROWLAND, University of California, Irvine
THOMAS H. VONDER HAAR, Colorado State University/CIRA, Fort Collins
ROGER M. WAKIMOTO, National Center for Atmospheric Research, Boulder, Colorado

Ex Officio Members

ANTONIO J. BUSALACCHI, JR., University of Maryland, College Park
ERIC F. WOOD, Princeton University, New Jersey

NRC Staff

CHRIS ELFRING, Director
PAUL CUTLER, Senior Program Officer
AMANDA STAUDT, Senior Program Officer
IAN KRAUCUNAS, Associate Program Officer
CLAUDIA MENGELT, Associate Program Officer
ELIZABETH A. GALINIS, Research Associate
LEAH PROBST, Research Associate
ROB GREENWAY, Senior Program Assistant
DIANE GUSTAFSON, Administrative Coordinator
ANDREAS SOHRE, Financial Associate

Preface

The National Research Council of the National Academies empanelled this Committee in response to a request from the National Oceanic and Atmospheric Administration-National Environmental Satellite, Data, and Information Service (NOAA-NESDIS) to provide advice on how to archive and provide access to its environmental and geospatial data. The Committee's full statement of task is reproduced in Appendix B.

NOAA is to be commended for addressing data archiving and access, and for seeking external advice for such critical endeavors. In this interim report, the Committee proposes principles and guidelines that NOAA can use to begin planning specific archiving strategies for the environmental and geospatial data it currently collects. It is important to emphasize that these preliminary principles and guidelines are intended to provide a foundation for further discussions with NOAA and its community of data users on the topic of data archiving; they will be further developed and expanded by this Committee, incorporating community input, in a final report that also addresses data access issues.

As part of its deliberations in preparing this report, the Committee met twice, received briefings from representatives of each of the NOAA divisions involved in data archiving activities, and reviewed the existing and planned observational and derived data streams collected by NOAA and its partners, along with some of its current data management procedures and legal requirements. The Committee also considered the relative costs of saving certain types of derived data products versus regenerating these data from archived first-stream input, as well as the current and potential value to society of archiving a broad variety of environmental and geospatial data. We look forward to the second phase of our activities, which will focus on expanding these preliminary principles and guidelines to include guidance on the extent to which a wide variety of data sets and derived products should be made available.

The Committee would like to thank the following invited speakers for their input: Thomas Karl, Christopher Fox, John Bates, Richard Brooks, Kurt Schnebele, Bonnie Ponwith, Susan McLean and Richard Beeler. The insights of Robert Serafin, Chair of the National Research Council's Board on Atmospheric Sciences and Climate (NRC BASC) were also appreciated. Lastly, on behalf of the entire Committee, I want to express gratitude to Chris Elfring, Ian Kraucunas, Elizabeth Galinis, and Rob Greenway of the NRC BASC for their excellent support of this initial Committee effort.

David A. Robinson, *Chair*
Committee on Archiving and Accessing
Environmental and Geospatial Data at NOAA

Acknowledgments

This report has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the National Research Council's Report Review Committee. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process. We wish to thank the following individuals for their review of this report:

William J. Emery, University of Colorado, Boulder
Kelvin K. Droegemeier, University of Oklahoma, Norman
Vincent V. Salomonson, University of Utah, South Jordan
Steven F. Clifford, University of Colorado, Boulder
Robert Chadduck, Electronic Records Archives Program, College Park, Maryland
Suzanne Carbotte, Columbia University, Palisades, New York
Steve Williams, University Corporation for Atmospheric Research, Boulder, Colorado
Scott Kirkpatrick, The Hebrew University of Jerusalem, Israel
Margaret A. LeMone, National Center for Atmospheric Research, Boulder, Colorado
Lee R. Raymond, Exxon Mobil Corporation, Irving, Texas

Although the reviewers listed above have provided constructive comments and suggestions, they were not asked to endorse the report's conclusions or recommendations, nor did they see the final draft of the report before its release. The review of this report was overseen by Sheldon S. Alexander, Pennsylvania State University, and Carl Wunsch, Massachusetts Institute of Technology. Appointed by the National Research Council, they were responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the authoring panel and the institution.

Contents

Summary

The Committee that assembled this report was empanelled by the National Research Council (NRC) at the request of the National Oceanic and Atmospheric Administration-National Environmental Satellite, Data, and Information Service (NOAA-NESDIS) to provide advice on archiving and providing access to the broad range of environmental and geospatial data collected by NOAA and its partners. With limited resources and enormous growth in data volumes, NOAA seeks input on how to identify the observations, data and derived products that should be preserved in perpetuity and made readily accessible versus those that require limited access and storage lifetimes. Pursuant to its statement of task (Appendix B), and based on its collective experience, the materials it has reviewed to date, and its initial deliberations, the Committee proposes the following preliminary principles and guidelines for data archiving activities at NOAA. These preliminary principles and guidelines are not ranked, and should be regarded as a framework for further discussion; they will be further developed and expanded by this Committee in a final report that also addresses data access at NOAA.

- The environmental and geospatial data collected by NOAA and its partners, including model output, are an invaluable resource that should be archived and made accessible in a form that allows researchers and educators to conduct analyses and generate products necessary to accurately describe the Earth System.
- The decision to archive or continue to archive data or model output should be driven by its current or future value to society. The decision will need to take into account the cost to archive versus the cost to regenerate, as well as the costs of providing access to the data.
- Funding for Earth System measurements should include sufficient resources to archive and provide ready and easy access to these data for an extended period of time. In particular, at the outset of undertaking an activity which will generate data or model output, end-to-end data management needs to be planned and budgeted.
- All data that are well documented are of known quality, and represent systematic collections or characterizations of the state of the environment should be archived in their most primitive useful form.
- Decisions not to archive data permanently should only occur when the original and predicted purpose of the data has been satisfied, or when the cost of storing the data exceeds the cost of regeneration,[1] and should be made in collaboration with the appropriate user communities.
- Metadata that completely document and describe archived data should be created and preserved to ensure the enhancement of knowledge for scientific and societal benefit.
- NOAA's archival process should be designed to allow the integrated exploitation of data from multiple sources to answer environmental questions and support the total life-cycle

[1] Reproduce from archived first-stream input (especially in the case of model output)

aspects of individual data sets. This could potentially be accomplished through a distributed but federated archival system facilitated via a single user portal.

- Broad community representation is essential to establish the process whereby data proposed for archiving can be evaluated and prioritized in terms of scientific and societal benefits.
- Scientific data stewardship should be applied to all archived information so it is preserved, continually accessible, and can be supplemented with additional data as discoveries build understanding and knowledge.

These preliminary principles and guidelines are intended to help NOAA and its partners begin planning specific archiving strategies for the data streams they currently collect as the Committee prepares for the next phase of its work. To effectively implement these ideas, NOAA will need to collaborate closely with its user communities and with other government agencies (e.g., the National Aeronautics and Space Administration (NASA) and the United States Geological Survey (USGS)). NOAA will also need to work with international partners through organizations such as the Group on Earth Observations, Committee on Earth Observation Satellites, International Council for Science, World Meteorological Organization, International Oceanographic Commission, and International Hydrographic Organization to develop internationally agreed standards and protocols to ensure that key data sets can be accessed and exchanged.

During the second phase of its activities, the Committee plans to convene a user forum designed to engage both NOAA data managers and NOAA's user community in order to gain additional information and insights on effective data access strategies. Following the forum, and after considering additional materials and extensive deliberation, the Committee will release a final report. This document will include an expanded set of principles and guidelines, illustrated with examples, that NOAA and its partners can use to identify the observational and generated data that should be preserved indefinitely versus those that require only limited storage lifetimes or can be readily regenerated from archived first-stream input, and also the extent to which a wide variety should be made available. A more extensive discussion of the specific scientific requirements for data access and data stewardship, including climate change detection and analysis, will also be included, as will further discussion of funding issues, both in general and in the context of specific archiving and access strategies for individual data streams.

1
Introduction

NOAA is a mission agency and historically it has collected environmental and geospatial data of many types to meet its primary meteorological, oceanographic, and geophysical operational mission requirements. The data managed by NOAA stretch from the surface of the sun to the core of the earth, and affect every aspect of society. Although it is difficult and beyond the charge of this committee to assess the monetary value of environmental data, the importance can be inferred: For example, the Department of Commerce's Bureau of Economic Analysis estimates that 42 percent of U.S. Gross Domestic Product is sensitive to weather and climate.[2] Customers of this investment include NOAA, other Federal agencies, state and local governments, industry, business interests, scientists, educators, the general public, and the international community. The needs of these customers are diverse, making it difficult to assess the value of any particular environmental data stream. For instance, while outdated weather forecasts are of little use to most customers, they may be crucial for the legal and research communities.

NOAA's National Environmental Satellite, Data, and Information Service (NESDIS) operates three national data centers (the National Climate Data Center, National Geophysical Data Center, and National Ocean Data Center) and over thirty centers of data (e.g., the National Ice Center); collectively these entities are responsible for "acquiring, integrating, managing, disseminating, and archiving environmental and geospatial data and information obtained from worldwide sources to support NOAA's mission".[3] For the purposes of this report, the term "data" will be taken to mean both environmental or geospatial observations, including physical samples, and also model output. Equally important are *metadata*, which are all the information necessary for data to be independently understood by users, to ensure proper stewardship of the data, and to allow for future discovery.

In the NOAA 2007 budget request[4] it is noted that "Collectively, the three national data centers acquire over one petabyte (10^{15} bytes) of new data annually, provide access to an archive exceeding 3.5 petabytes, and support over 100 million worldwide queries per year, providing data transfers to over two million customers." The rapid increase in the volume of data distribution (Figure 1) and its associated stewardship and management activities is a significant concern. Furthermore, the challenges associated with managing NOAA's data are only expected to increase with the anticipated explosion in model output and new satellite systems in the years ahead (Figure 2). Even though the launch dates of some assets such as the National Polar-orbiting Operational Environmental Satellite System (NPOESS) are uncertain, NOAA has agreed to archive certain data collected by other agencies, such as MODIS (Moderate Resolution

---

[2] Bureau of Economic Analysis figures reported in National Research Council, 1998, *The Atmospheric Sciences Entering the Twenty-First Century*, National Academy Press, Washington, D.C., page 25.
[3] NOAA Administrative Order 212-15, effective 22 December 2003, available at http://www.corporateservices.noaa.gov/~ames/NAOs/Chap_212/naos_212_15.html
[4] NOAA 2007 budget request "blue book," available at http://www.corporateservices.noaa.gov/~nbo/07bluebook_highlights.html

Imaging Spectroradiometer) data from the EOS (Earth Observing System) satellites operated by NASA. This large and exponentially growing data volume indicates an urgent need for NOAA to address its ability to handle the current and future needs of NOAA archive users, and in fact it has already begun to do so, but significant work remains.

In addition to data volume, data diversity is another challenge; NOAA's consolidated observation requirements include over 2000 diverse variables ranging from hyperspectral satellite imagery to the stomach contents of fish (McLean S., 2006). These data come from a broad range of platforms including (but not limited to) satellites, fixed and mobile radars, research aircraft, buoys, and ships of opportunity, and may be derived from such diverse sources as embedded sensors, models, physical samples, and self-organizing networks, each of which are associated with unique challenges in organizing, cataloguing, archiving, and providing access to the data they collect or generate.

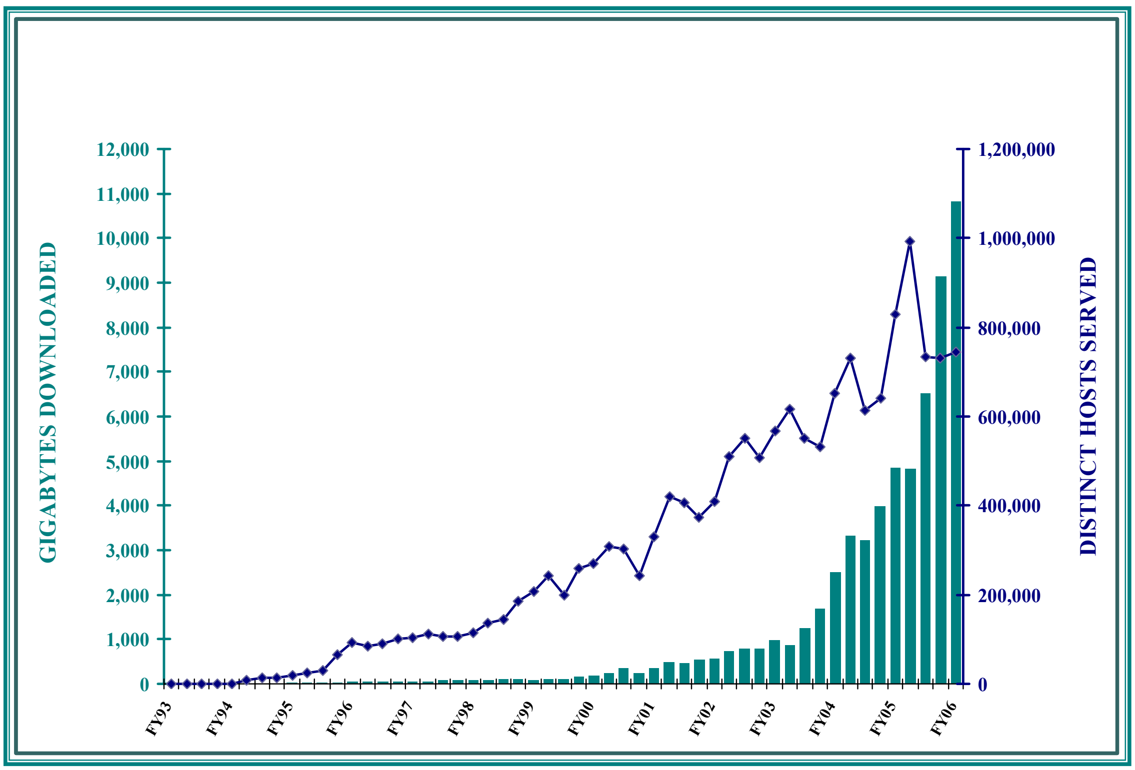

Figure 1: Quarterly data downloads from NOAA's National Geophysical Data Center (NGDC), in gigabytes (line plot and left axis), and number of distinct hosts served (bars and right axis) for fiscal years 1993-2006 (Source: Fox C., 2006)

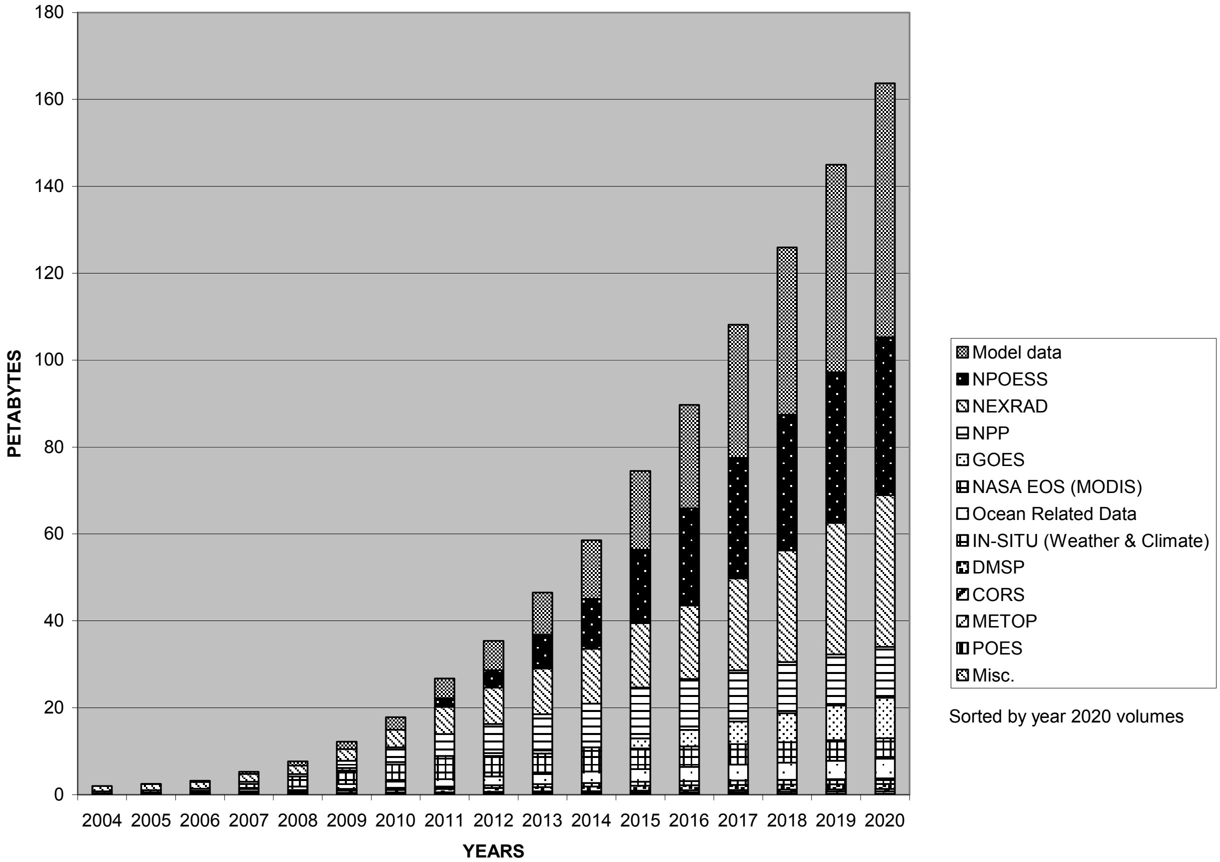

Figure 2: Current NOAA-NESDIS data archive volume projections under the Comprehensive Large Array-data Stewardship System (CLASS), in petabytes (Source: Updated May 4, 2006 from NOAA, 2003)

NOAA deserves praise for the steps it has taken and is taking to address its considerable and growing data management challenges. For instance, the recently completed Assessment of NOAA's Environmental Data and Information Management report (NOAA, 2006) includes a comprehensive, NOAA-wide assessment of data management capabilities organized by its mission goals. This effort, along with the establishment of the NOAA Observing Systems Council (NOSC) and its components (NOAA Observing System Architecture (NOSA), Data Management Committee (DMC), and Chief Information Officer (CIO) efforts) will eventually bear fruit in an appropriate enterprise-wide culture and coordinated best processes, if they can be made effective.

2
Context

The challenges of data archiving and access have been discussed in many settings for a number of years. Previous reports from the National Research Council, in particular, have helped set the stage for this report by discussing data collection, management, archiving, and dissemination from various perspectives. For example, the NRC has advised the U.S. National Archives and Records Administration and other federal agencies on the long-term retention of scientific and technical data, particularly in electronic formats (NRC, 1995). A more recent report, *Climate Data Records from Environmental Satellites*, (NRC, 2004a) focuses more specifically on generating, analyzing, and archiving the records that are most useful for understanding climate variability and change. This latter report, although focusing almost exclusively on climate data, offers many insights that apply to archiving other kinds of environmental and geospatial data, including the identification of 14 essential elements of successful climate data record generation programs. Many of its findings and recommendations could also be applied more broadly to environmental and geospatial data from diverse sources with a simple change of wording, such as this paragraph from the executive summary:

> Underlying many of the elements of success is early attention to data stewardship, management, access and dissemination policies, and the actual practices implemented. Because a successful [climate data record] program will ultimately require reprocessing, dataset used in their creation, such as metadata, should be preserved indefinitely in formats that promote easy access. The ultimate legacy of long-term [climate data record] programs is the data left to the next generation, and the cost of data management and archiving must be considered as an integral part of every [climate data record] program.

Another relevant report, *Government Data Centers: Meeting Increasing Demands* (NRC, 2003a), summarizes a workshop exploring how the increasing volume and number of data sets, coupled with greater demands from more diverse users, are making it difficult for data centers to maintain records of environmental change. The report focuses on technological approaches that could enhance the ability of environmental data centers to deal with these challenges, and improve the ability of users to find and use information held in data centers. The NRC has also provided more focused reviews of NOAA's National Geophysical Data Center (NRC, 2003b), NOAA's National Climatic Data Center (NRC, 1993), and NASA's Distributed Access Archive Centers (NRC, 1999), which offer a variety of insights on how to best archive and provide access to environmental and geospatial data.

Perhaps the most relevant and still timely supporting NRC report is *Utilization of Operational Environmental Satellite Data: Ensuring Readiness for 2010 and Beyond* (2004b), which offers findings and recommendations aimed at defining specific approaches to resolving the potential overload faced by two agencies—NOAA and NASA—responsible for satellite data. The report focuses on the end-to-end utilization of environmental satellite data by characterizing the links from the sources of raw data to the end requirements of various user groups. The "Utilization Report" is an important foundation document because it addresses three stillchallenging areas: (1) the value of and need for environmental satellite data, (2) the distribution of environmental satellite data, and (3) data access and utilization; it also includes still-pertinent findings and recommendations that will be considered and updated in this committee's final report. However, some of the findings are particularly relevant to set the stage for this report, including (in brief):

- Improved and continuous access to environmental satellite data is of the highest priority for an increasingly broad and diverse range of users.
- The national and individual user requirements for multiyear climate system data sets from operational satellites are placing special demands on current and future data archiving and utilization systems.
- The Comprehensive Large Array-data Stewardship System (CLASS) is being designed[5] by NOAA to catalog, archive, and disseminate NOAA environmental satellite data. Given the magnitude of this effort—and considering the growing volume, types, and complexity of environmental satellite data; the increasingly large and diverse user base; and expectations for wider and more effective use of data—NOAA needs to have a comprehensive understanding of the full scope of the technical requirements for data cataloging, archiving, and dissemination and needs to ensure that implementation is based on that knowledge. Key to successful implementation of a strong system that will serve operational users and the nation well are detailed planning, proactive follow-through, and NOAA's incorporation of lessons learned from previously developed, similarly scaled initiatives with similar system requirements.
- Data from diverse satellite platforms and for different environmental variables must often be retrieved from different sources, and these retrievals often yield data sets in different formats with different resolution and gridding. The multiple steps currently required to retrieve and manipulate environmental satellite data sets are an impediment to their use.
- Early and ongoing cooperation and dialog among users, developers of satellite remote sensing hardware and software, and U.S. and international research and operational satellite data providers is essential for the rapid and successful utilization of environmental satellite data. Many of the greatest environmental satellite data utilization success stories have a common theme: the treatment of research and operations as a continuum, with a relentless team focus on excellence with the freedom to continuously improve and evolve.

A number of other reports were of particular importance in helping this committee begin its work. *Global Change Science Requirements for Long-Term Archiving: Report of the Workshop, October 28-30, 1998* (USGCRP, 1999), provides a good example of preliminary high-level guidance on long-term archiving of Earth observation data and derived products, lessons learned from current and past experiences, and the guiding principles and essential function necessary for any program to be successful. The report *Recommendation for Space Data System Standards: Reference Model for an Open Archival Information a System*, (Consultative Committee for Space Data Systems, 2002) is an internationally developed set of

[5] CLASS is still being developed; the Committee will address this effort more explicitly in its final report.

recommendations about how open archival information systems should be structured. Among other tasks, the "OAIS" document defines a model system that:

- Provides a framework for the understanding and increased awareness of archival concepts needed for long term digital information preservation and access.
- Provides the concepts needed by non-archival organizations to be effective participants in the preservation process.
- Provides a framework, including terminology and concepts, for describing and comparing architectures and operations of existing and future archives.
- Provides a basis for comparing the data models of digital information preserved by archives and for discussing how data models and the underlying information may change over times.
- Provides a foundation that may be expanded by other efforts to cover long-term preservation of information that is not in digital form.
- Expands consensus on the elements and processes for long-term digital information preservation and access.

The most recent report from the NRC's Committee on Earth Science and Applications from Space (NRC, 2005) has also highlighted the critical need for the archival, access and stewardship of climate data records, focusing on satellites, as stated in one of their recommendations:

> The committee recommends that NOAA, working with the Climate Change Science Program and the international Group on Earth Observations, create a climate data and information system to meet the challenge of ensuring the production, distribution, and stewardship of high-accuracy climate records from NPOESS and other relevant observational platforms.

Similarly, the National Science Board at the request of the National Science Foundation has also reported on the importance of long-term archives in "Long-lived Digital Data Collections: Enabling Research and Education for the 21st Century" (NSF, 2005). Here there is recognition that there is need for a broad dialogue between agencies that collect data, and a clear technical and financial strategy along with support data policies are required to preserve the valuable data resources. Data archiving and access are also identified as critical components in the multiple agency Data Management and Communications (DMAC) Subsystem of the Integrated Ocean Observing System (IOOS) plan (Hankin, S. and the DMAC Steering Committee, 2005). As implementation of IOOS/DMAC takes shape with accountability for data preservations two things are obvious, data archiving and accesses are required components in end-to-end data systems, and are necessary focus for all agencies that collect data. Other relevant reports pertaining to the challenges of data archiving include the *International Council for Science (ICSU) Report of the CSPR Assessment Panel on Scientific Data and Information* (ICSU, 2004) and the *Final Report from the Workshop on Research Challenges in Digital Archiving and Long-Term Preservation* (NSF, 2003).

Together these documents and many others, often available on the websites of facilities responsible for data management,[6] make it clear that the development of principles and guidelines for archiving environmental and geospatial data need not be developed from scratch.

[6] e.g., World Data Center System, http://www.ngdc.noaa.gov/wdc/

There is a wealth of information available that waits to be focused on NOAA's upcoming challenge—how to be a wise steward of such a broad range of data, given both the reality of limited resources and the enormous growth in data volume that is both inevitable and invaluable.

The importance of long-term data collection, the need to archive environmental and geospatial data for maximal societal benefit, and the recognition of the significant costs involved are all issues the Earth System science and data management communities have long discussed, albeit not always focused on the specific challenges faced by NOAA. In this interim report, the committee is thus attempting to synthesize and apply concepts that have gained broad acceptance in the data management community as preliminary principles and guidelines for archiving environmental and geospatial data in particular.

3
Preliminary Principles and Guidelines

As is explicitly requested in its statement of task, the Committee proposes the following preliminary principles and guidelines to help NOAA begin planning specific archiving strategies for the data streams it currently collects and manages. Each proposed principle is followed by some explanatory text to define the terms used and to put the suggestion in the proper context. The Committee developed this guidance based on its collective experience, the materials it has reviewed to date, and its initial deliberations. It is important to note that these ideas, while phrased as recommendations, are not ranked in order of importance, and are intended to serve as a framework for further discussions with NOAA and NOAA's user community.

Following additional data gathering, including a user forum, and additional Committee deliberations, a final report will be generated. This report will include an expanded set of principles and guidelines, illustrated with examples, that can be used to identify the observational and generated data that should be preserved indefinitely versus those that require only limited storage lifetimes or can be readily regenerated from archived first-stream input, as well as the degree to which a wide variety of data should be made available. A more extensive discussion of the specific scientific requirements for data access and data stewardship for a range of applications, including climate change detection and analysis, will also be included, as will further discussion of funding issues, both in general and in the context of specific archiving and access strategies for individual data streams.

The environmental and geospatial data collected by NOAA and its partners, including model output, are an invaluable resource that should be archived and made accessible in a form that allows researchers and educators to conduct analyses and generate products necessary to accurately describe the Earth System.

The Earth System is a complex, interactive biogeochemical system that requires a large number of environmental variables for an accurate description. Any data stream, data set, or model output array that contributes to the understanding, prediction, or long term description of the Earth System should be considered for permanent archiving by NOAA. Data considered for archiving need to have sufficient and understood accuracy and temporal and spatial resolution to increase our understanding of the System, improve our characterization and predictability of the System, and/or allow required analyses for determining the past, current, and possible future states of the System along with its past variability.

The decision to archive or continue to archive data or model output should be driven by its current or future value to society. The decision will need to take into account the cost to archive versus the cost to regenerate, as well as the costs of providing access to the data.

The maximum benefit to society ultimately defines the rationale for government data collection (and its funding) that is to support broad government and private sector decision-making abilities. The collection, archiving and accessing of Earth System data needs to support the analyses and products necessary to make these decisions. In general, the cost of archiving and providing access to data represents only a small fraction of the total resources invested in collecting or generating data.[7] In addition, it is extremely difficult to estimate the present and future value of environmental or geospatial data to society. For instance, recently global atmospheric reanalysis of observations beginning about 1950 have proven to be an important climate assessment archive (Kalnay et al., 1996). It is difficult to appreciate the specific contributions that any data stream or data set makes, or might make, to long term climate monitoring or other environmental research requirements. In addition, data perceived to be of little use in the present could become quite valuable in the future as advancing technology increases our ability to make use of data. As discussed in more detail below, it is essential to actively engage the user community to help make these decisions. Effective stewardship of the nation's investments requires preservation of what has taken substantial resources to collect.

Funding for Earth System measurements should include sufficient resources to archive and provide ready and easy access to these data for extended periods of time. In particular, at the outset of undertaking an activity which will generate data or model output, end-to-end data management needs to be planned and budgeted.

Many in situ data sets and the highest volume data sets (especially radar and satellite data) have been collected and funded in support of NOAA's operational missions, with little initial provision made for long term preservation to support Earth system research, weather and climate prediction, and other societal benefits. Also, in the past, many data sets often had little use past their operational needs, while others were little used because of their spatial and temporal deficiencies. However, with new data assimilation methods and systems, most meteorological and oceanographic data can now be ingested by numerical models and used, for example, for model initialization and verification. These circumstances emphasize the vital importance of establishing an enduring, long term archive of environmental and geospatial data that is supported by the resources necessary to effectively meet these requirements, including resources for hardware and data managers. Assuring adequate and sustained levels of funding to archive and provide access to data remains a major ongoing challenge.

All data that are welldocumented, are of known quality, and represent systematic collections or characterizations of the state of the environment should be archived in their most primitive useful form.

Several of the considerations noted in the preceding section support this principle. First, observations of the state of the environment are generally expensive to obtain, and these costs far exceed archival costs. Second, it is impossible to anticipate all future applications of a data set, so a data set of uncertain present value may provide the key to some future scientific issue. Third, well calibrated data sets, or at least data sets with well defined error characteristics, are essential to long term climate monitoring and many other Earth System research requirements; this requirement implies a commitment to maintain records of successive improvements,

---

[7] As an illustration, in NOAA's FY2007 budget request, $994 million is requested just for satellite acquisition and satellite observing services, while only $51 million is requested for all of NOAA's Data Centers and information services.

recalibrations, reprocessing, or other changes (including, in the case of model output, records of changes to the model), in addition to maintaining an archive of the original first-stream data. Finally, Earth System observations are unique; it is usually not possible to go back in time and resample, although if proper care is taken to document the data it is often possible to regenerate them from archived first-stream input (especially in the case of model output).

Original Data are one of the seven types of data defined by NOAA (Appendix C), and represent the most obvious data type to consider for long-term archiving based on the principle above. Some of the other data types defined by NOAA are also of long-lasting value and deserving of consideration for permanent archiving (e.g., certain *Synthesized Products* and *Experimental Products*). In addition, certain types of at-risk data, such as data on deteriorating or substandard media, are critical to NOAA's mission. NOAA Data Centers should be encouraged to be proactive and, where appropriate, obtain and archive these data. To make the most judicious data rescue decisions, it would be useful to obtain recommendations from an advisory panel made up of data users, as discussed in more detail below.

Decisions not to archive data permanently should only occur when the original and predicted purpose of the data has been satisfied, or when the cost of storing the data exceeds the cost of regeneration, and should be made in collaboration with the appropriate user communities.

Since estimating the present and future value of data is extremely difficult, it may be of greater use to first define those data that clearly have only short-term uses. For example, derived analyses or products for specific decisions may have little value after the decision has been made, and could be reproduced if necessary. These data could include long range and intermediate model and forecast output and/or high volume data used to generate specified parameters such as sub-second wind or radar returns used to produce several-minute averages for operational reporting purposes. Data collected for specific short term research (e.g., *Experimental Data*, as described in Appendix C) or near term operational requirements may also fit some of these criteria and therefore not meet NOAA's mission requirements for archiving. As discussed in more detail below, it is essential to actively engage the user community to help make these difficult decisions.

Metadata that completely document and describe archived data should be created and preserved to ensure the enhancement of knowledge for scientific and societal benefit.

Metadata are all the information necessary for data to be independently understood by users, to ensure proper stewardship of the data, and to allow for future discovery.[8] Metadata are essential for effective archive management throughout the entire data lifecycle, and are the essential data management component that makes an archive useful for data discovery and integration, through data mining and other techniques. Where possible and practical, it is advisable to use established metadata standards. As necessary these standards should be further developed, in coordination with other federal agencies and international entities, to take advantage of current and future technologies needed for data search, mining, discovery, and integration capabilities. Effective metadata management will help meet the challenge of the increase in data volumes, enable better integration of information across data sources and

---

[8] See the information package definition on page 2-5 of the Open Archival and Information System reference model (Consultative Committee for Space Data Systems, 2002) for details.

disciplines, and improve understanding and usability of the data. As part of this emphasis, data systems should be designed to share metadata and build catalogs to enable data discovery across systems, disciplines, and programs. Further, since data systems will evolve to incorporate new information and to take advantage of technological improvements, the data system philosophy needs to account for the reality that metadata continually evolve, expand, and mature. This necessitates the use of existing standards and the adoption of new ones, when appropriate, to greater facilitate services and integration across data sources and disciplines; as such, metadata management will need to expand beyond mandatory requirements. For example, information should include context (relation to other information, appropriate application, and limitations) and information necessary for tools and interoperable services.

NOAA's archival process should be designed to allow the integrated exploitation of data from multiple sources to answer environmental questions and support the total life-cycle aspects of individual data sets. This could potentially be accomplished through a distributed but federated archival system facilitated via a single user portal.

NOAA's mission represents a wide breadth of responsibilities and disciplines related to understanding, describing and predicting various aspects of the highly complex Earth System. This mission requires the integration and synthesis of disparate data holdings both within and outside NOAA and, often, assimilation of multiple sets of environmental variables. NOAA's many data streams reflect this diversity, as does the system-of-systems architecture of its observing mission. NOAA's archival activities should recognize and reflect this same diversity in data sources and needs of its users. In addition, the data archival process needs to be rooted in an evolutionary framework, and to be flexible, reliable, extensible, and scalable in order to accommodate the increasing volume and complexity of data holdings and evolving needs of NOAA's customers. NOAA therefore needs to ensure that they develop a scalable, extensible, and reliable infrastructure to ensure the long-term access and preservation of digital assets.

Given the widely distributed nature of its data activities and holdings, NOAA could consider its archival process as part of a federation of distributed data sources and archival delivery partners. One particularly promising framework is a decentralized approach to archiving and data provision, enabled by a *centralized* corporate-level portal that facilitates discovery and access of integrated data sets tailored to user needs. This portal concept does not preclude the existence of other portals, but would allow NOAA to facilitate a "one-stop-shopping" capability and a more recognizable web presence to current and potential users. Within this framework, data could be made available and discoverable to all users in reliable, standardized formats for easy use and integration with other related data. A coordinated integration process would also allow NOAA to easily collaborate with non-NOAA data holders for the benefit of users. Since NOAA's internal data stewards include many data managers, centers of data and Data Centers, integration of these various centers' data holdings, and the migration of data through the chain from individual researchers to centers of data to formal Data Centers needs to be recognized and more explicitly formalized.[9]

[9] According to NOAA Administrative Order 212-15, "Centers of Data transfer their data holdings to the NOAA National Data Centers for permanent archiving when continued storage at the Center of Data is no longer appropriate."

Broad community representation is essential to establish the process whereby data proposed for archiving can be evaluated and prioritized in terms of scientific and societal benefits.

It is difficult to implement archiving principles that can be applied to all of the diverse data types managed by NOAA. Thus, user input is critical. For example, for data sampled at very high temporal, spatial, and/or spectral resolutions, advice from relevant users could be sought regarding what level and at what resolution the data should be archived. For instance, NEXRAD (Next Generation Weather Radar) Level II data from Alaska, Hawaii, and a few other places is not currently archived. If present funding levels and technology do not support the archiving of all Level II data from the WSR-88D radars in the United States (and definitely not the voluminous Level I data), it is not clear if it will be possible to archive all channels of GOES-NEXT (Geostationary Operational Environmental Satellites-Next Generation) output. Important decisions on what types and amounts of data to archive should be made with the broadest possible cross section of inputs. Data users can also help identify at-risk data streams.

Other federal agencies, non-federal government agencies, and non-government organizations are heavily involved in collection and archiving of data that are relevant to NOAA's mission. Therefore, NOAA will need to coordinate with these partners to establish the criteria for archiving and providing access to these data at NOAA versus at other organizations. The criteria for archiving by NOAA versus other organizations will need to be agreed to among the governmental agencies supporting the collection, since all data produced by governmental resources should be considered for retention and archival. NOAA will also need to work with international partners through organizations such as the Group on Earth Observations, Committee on Earth Observation Satellites, International Council for Science, World Meteorological Organization, International Oceanographic Commission, and International Hydrographic Organization to develop internationally agreed-upon standards and protocols to ensure that key data sets can be accessed and exchanged.

The Committee's final report will discuss in greater depth the factors leading to the decision whether to archive data or not, as well as the potential consequences and tradeoffs of these decisions, in addition to who might be in the best position to make them.

Scientific data stewardship should be applied to all archived information so it is preserved, continually accessible, and can be supplemented with additional data as discoveries build understanding and knowledge.

Stewardship is vital to maintaining a long-term archive. It includes preserving archive integrity, assuring accurate media and format migration, maintaining data access[10] and integrity during technology and software evolution, and enhancing the archive by adding information that is established throughout the data lifecycle.

Archived information encompasses individual data sets and all their associated metadata. Ideally, metadata fully describe and document the data as well as the relationships between collections within a NOAA center and other collections at other centers.

The primary stewardship activities are to:

[10] Expert stewardship practices enable improved information access. Access functions are not discussed here, but are critical and a topic that will be addressed in the Committee's final report.

- Assure data backup strategies that guard against any loss of irreplaceable data (e.g. *Original* or *Synthesized Product* data types that cannot be easily and economically regenerated).
- Assure that when the data are migrated to new media or translated to another form (changing the original data format), no information is lost and recovery from errors is always possible.
- Assure that information integrity and access are not compromised during software and technology evolution.
- Assure that data management systems can support the growth of archived information that occurs during the data lifecycle. During a data lifecycle the archived information may be enhanced by reprocessing, error correction, and with the addition of supplemental data, quality control assessments, and other metadata derived from the scientific research and knowledge building processes.
- Maintain close interactions with the scientific community and evolving user base to capture information about data use and limitations.

References

Consultative Committee for Space Data Systems, 2002. Reference Model for an Open Archival Information System (OAIS) CCDSD 650.0-B-1. Blue Book. Issue 1. CCSDS Secretariat, Washington, DC. Available at *http://public.ccsds.org/publications/archive/650x0b1.pdf*.

Fox, C. 2006. National Geophysical Data Center Presentation to the NRC Committee on Archiving and Assessing Environmental and Geospatial Data at NOAA, Washington, D.C., February 23.

Hankin, S. and the DMAC (Data Management and Communications) Steering Committee, 2005. Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems: I. Interoperable Data Discovery, Access, and Archive, Ocean. Arlington, VA. Available at *http//dmac.ocean.us/dacsc/imp_plan.jsp*.

International Council for Science (ICSU), 2004. ICSU Report of the CSPR Assessment Panel on Scientific Data and Information.

Kalnay, E., M. Kanamitsu, R. Kistler, W. Collins, D. Deaven, L. Gandin, M. Iredell, S. Saha, G. White, J. Woollen, Y. Zhu, M. Chelliah, W. Ebisuzaki, W. Higgins, J. Janowiak, K. C. Mo, C. Ropelewski, J. Wang, A. Leetmaa, R. Reynolds, R. Jenne, and D. Joseph, 1996. The NMC/NCAR 40-Year Reanalysis Project. Bulletin of the American Meteorological Society, 77, 437-471.

McLean, S. 2006. E-mail to the NRC Committee on Archiving and Assessing Environmental and Geospatial Data at NOAA, March 22.

NRC (National Research Council), 2005. Committee on Earth Science and Applications from Space: A Community Assessment and Strategy for the Future. National Academies Press. Washington, D.C.

NRC, 2004a. Climate Data Records from Environmental Satellites. National Academies Press. Washington, D.C.

NRC, 2004b. Utilization of Operational Environmental Satellite Data: Ensuring Readiness for 2010 and Beyond. National Academies Press. Washington, D.C.

NRC, 2003a. Government Data Centers: Meeting Increasing Demands. National Academies Press. Washington, D.C.

NRC, 2003b. Review of NOAA's National Geophysical Data Center. National Academies Press. Washington, D.C.

NRC, 1999. Review of NASA's Distributed Active Archive Centers. National Academy Press. Washington, D.C.

NRC, 1995a. Study on the Long-term Retention of Selected Scientific and Technical Records of the Federal Government: Working Papers. National Academy Press. Washington, D.C.

NRC, 1995b. Finding the Forest in the Trees: The Challenge of Combining Diverse Environmental Data. National Academy Press. Washington, D.C.

NRC, 1993. Review of the World Center-A for Meteorology and the National Climatic Data Center. National Academy Press. Washington, D.C.

NOAA (National Oceanic and Atmospheric Administration), 2003. NOAA Environmental Data: Foundations for Earth Observations and Data Management System. U.S. Department of Commerce, National Oceanic and Atmospheric Administration, July 2003.

NOAA, 2006. NOAA's Environmental Data Management: Integrating the Pieces. Department of Commerce Technical Report. *In press*.

NSF (National Science Foundation), 2005. Long-lived Digital Data Collections: Enabling Research and Education for the 21st Century, National Science Board. Available at *http//www.nsf.gov/pubs/2005/nsb0540/start.htm*.

NSF, 2003. Digital Government Program and Digital Libraries Program, Directorate for Computing and Information Sciences and Engineering, and the Library of Congress, National Digital Information Infrastructure and Preservation Program. It's About Time: Research Challenges in Digital Archiving and Long-Term Preservation, Final Report, Workshop on Research Challenges in Digital Archiving and Long-Term Preservation. August 2003. National Science Board.

USGCRP (U.S. Global Change Research Program), 1999. Global Change Science Requirements for Long-Term Archiving: Report of the Workshop, October 28-30, 1998. National Aeronautics and Space Administration – National Oceanic and Atmospheric Administration.

Appendix A: Biographical Sketches of Committee Members

David A. Robinson *(Chair)* is the Chairman of the Geography Department at Rutgers University. He has expertise in the collection and archiving of accurate climatic data, and is interested in climate change (particularly state and regional climate issues), hemispheric and regional snow cover dynamics, interactions of snow cover with other climate elements, and the dynamics of solar and terrestrial radiative fluxes at and close to the surface of the earth. Robinson is the author or co-author of approximately 130 articles, over half in peer-reviewed journals and book chapters. He also is the State Climatologist for New Jersey. Robinson has served on several NRC committees, and served as the chairman of the Committee on Climate Data Records from Operational Satellites: Development of a NOAA Satellite Data Utilization Plan. He received his Ph.D. from Columbia University.

David C. Bader is the Director for the Program for Climate Model Diagnosis and Intercomparison (PCMDI), a scientist at Lawrence Livermore National Laboratory, and Chief Scientist for the DOE Climate Change Prediction Program (CCPP). Before holding his current positions, he held the positions of project manager, research scientist and senior research scientist at Pacific Northwest National Laboratory (PNNL). Bader has also been a visiting scientist at the National Center for Atmospheric Research and has had an intergovernmental program act (IPA) assignment as program manager for the U.S. Department of Energy (DOE) Office of Science's Climate Change Prediction Program. He has held the position of director of the DOE Office of Science Scientific Discovery through Advanced Computing (SciDAC) program and the DOE's Computer Hardware, Advanced Mathematics and Model Physics (CHAMMP) Program. Bader received his Ph.D. in Atmospheric Science in 1985 from Colorado State University.

Donald W. Burgess is a research fellow at the Cooperative Institute for Mesoscale Meteorological Studies (CIMMS) at the University of Oklahoma. He has served as the Chief of the Warning Research and Development Division of the National Severe Storms Laboratory (NSSL). He has also served as Acting Director and Chief of Operations at the National Weather Service Radar Operations Center (ROC), and Chief of the National Weather Service Radar Training Branch. His research interests lie in the areas of severe weather and on techniques for improving warnings of weather hazards, particularly techniques using Doppler radar to detect and warn of tornadoes. In 1979, he was a co-winner of the Department of Commerce Silver Medal, and in 2003, he was a co-winner of the Department of Commerce Bronze Medal. He is a fellow of the American Meteorological Society. Burgess received his M.S. from the University of Oklahoma in 1974.

Kenneth E. Eis is Deputy Director of Colorado State University's Cooperative Institute for Research in the Atmosphere (CIRA). He is in charge of CIRA's infrastructure and Earth Station, and data holdings and oversees collaborative research with CIRA collaborators working within

NOAA's Earth System Research Laboratory (ESRL). In addition to this ESRL-oriented work, Eis has helped directed the research requirements for the CSU Center for Geosciences Atmospheric Research (CG/AR) Phase II-IV. Eis is also the director of the CLOUDSAT Data Processing Center (DPC). The DPC is responsible for the ingest of all CloudSat data and the production of science products and their distribution to the science community. Prior to his current job and retirement from the U.S. Air Force, Lt. Col. Eis was commander of the Environmental Technical Applications Center (USAFETAC) now called the Air Force Combat Climatology Center (AFCCC). As USAFETAC/CC he was responsible for all Air Force climate support and data holdings. Next he was Director of Environmental Sciences at HQ Air Weather Service, and Chief Staff Meteorologist for the Air Force Systems Command. In these jobs he developed aerospace requirements in support of the Air Weather Service future systems as well as heading the analysis and mitigation efforts of Air Force advanced weapons systems environmental limitations. He also managed weather related launch support requirements at Vandenberg and Patrick AFBs.

Sara J. Graves is the director of the Information Technology and Systems Center and University Professor of Computer Science at the University of Alabama in Huntsville. She is also the Director of the Information Technology Research Center at the National Space Science and Technology Center. Graves is on the National Academies U.S. National Committee on CODATA and the National Biological Information Infrastructure Science Committee. She has served as a member of the NASA Headquarters Earth System Science and Applications Advisory Committee (ESSAAC) and Chair of the ESSAAC Subcommittee on Information Systems and Services (ESISS). Graves directs research and development in large-scale distributed information systems, data mining and knowledge discovery, high performance computing and networking, grid technologies and services, geospatial data analysis and visualization and bioinformatics. She received her Ph.D. in Computer Science from the University of Alabama in Huntsville in 1984.

Ernest G. Hildner was the Director of NOAA's Space Environment Center (SEC), the nation's official source of information about space weather storms, from 1986 until 2005. Under his direction, SEC conducted research and consulted on space weather instrument development for NOAA, NASA, and the Air Force. Hildner is a solar physicist who has worked for the High Altitude Observatory, NCAR, and for NASA Marshall Space Flight Center as Chief of its Solar Physics Branch. He was an experiment scientist for both the Skylab and the Solar Maximum Missions during the 70's. Hildner has published dozens of papers in coronal and interplanetary physics and co-holds a patent for a variable-magnification x-ray telescope. In addition to his administrative responsibilities with NOAA, including being NOAA's Program Manager for Space Weather, Hildner was a Co-chair of the Committee on Space Weather for the National Space Weather Program, a member of the advisory committees for the NOAO National Solar Observatory and NCAR High Altitude Observatory, and served on review panels for NASA and DoD projects. In December 2003 he received the Department of Commerce Gold Medal for advancing the nation's space weather services through the conception, funding, and development of the first-operational Solar X-ray Imager (SXI). He has twice received the Presidential Rank Award for Senior Executive Service managers.

Kenneth E. Kunkel is the director for the Center for Atmospheric Science at the Illinois State Water Survey. He has held a variety of positions at the Illinois State Water Survey since 1988, including Director of the Midwestern Regional Climate Center and Director of the Office of Applied Climatology. During 1982-1988, he served as the New Mexico State Climatologist and had a research and teaching appointment as an Associate Professor at New Mexico State University. Kunkel also studied atmospheric optical phenomena as a research meteorologist with the Atmospheric Sciences Laboratory at White Sands Missile Range. His recent research has focused on climate variability, extremes, and change. He has managed several projects with goals to expand, quality control, and analyze surface climate data sets, a specific emphasis being on the early cooperative observer network data. He has published numerous articles in peer-reviewed scientific journals and has written three book chapters. Kunkel received his Ph.D. in 1978 in meteorology from the University of Wisconsin-Madison.

Mark A. Parsons is an Associate Scientist III at the National Snow and Ice Data Center at the Cooperative Institute for Research in Environmental Sciences of the University of Colorado where he is the data set manager and the project manager for the Cold Land Processes Experiment Data Management. Parsons is also the program manager for the International Polar Year Data Management and the Frozen Ground Data Center at the World Data Center for Glaciology. He has over 15 years of data management experience, including appointments as an environmental scientist at AScI Corporation and Coe-Truman Technologies as well as research associate and research assistant at University of Edinburgh and Cornell University respectively. Parsons earned his B.Sc. in Natural Resources and Communications from Cornell University in 1988.

Mohan K. Ramamurthy is the Director of UCAR's Unidata Program and is a scientist in NCAR's Mesoscale and Microscale Meteorology Division. Unidata provides a broad array of data for use in geosciences education and research. In addition to providing data, Unidata also develop software for data access, processing, management, analysis, and visualization and provides support to a diverse community of over 160 institutions vested in the common goal of sharing data. As a scientist, Ramamurthy studies weather processes and prediction, including mesoscale phenomena such as gravity waves, precipitation band, hurricanes, and ensemble forecasting. His other research interests include information technology, interactive-multimedia instruction and learning, and end-to-end data services. Ramamurthy joined UCAR in 2003 after spending nearly 16 years on the faculty in the Department of Atmospheric Sciences at the University of Illinois at Urbana-Champaign. He earned his Ph.D. from the University of Oklahoma, where his doctoral research dealt with the four-dimensional assimilation of data and modeling of disturbances associated with monsoons.

Deborah K. Smith is a senior scientist of geology and geophysics at the Woods Hole Oceanographic Institution. Her research focuses on large scale plate tectonics as well as the dynamics of submarine and subaerial rift zones. She routinely goes to sea and collects bathymetry, side-scan sonar, gravity, magnetic, and photo-imagery data. She has strong interests in data quality and preservation and has organized workshops about these topics. She also has interests in education and outreach and has written for popular magazines. She has designed a website permitting school children and the public to participate in a virtual research expedition and one highlighting the day-to-day lives of research scientists. Smith has recently served on two advisory committees: The RIDGE 2000 Executive Committee and the U.S. Science Advisory Committee. She earned her Ph.D. in Earth Sciences from Scripps Institution of Oceanography at the University of California in San Diego.

John R. G. Townshend holds a joint appointment as Professor in the Institute for Advanced Computing Studies and the Department of Geography at the University of Maryland. He is also a member of the Department of Geography's Laboratory for Global Remote Sensing Studies. Townshend's research centers on the use of remote sensing and advanced computing methods for improvements in the characterization of regional and global land cover. He has been a member of the NASA's MODIS Science Team, and he is a PI on the Landsat Pathfinder Project for monitoring the Earth's Tropical Moist Forests. Townshend has also been chairman of the Joint Scientific and Technical Committee of the Global Climate Observing System. His previous NRC service includes membership on the Committee on Geophysical and Environmental Data and the Board on Earth Sciences and Resources. He served as a member of the NRC Committee for Review of the Science Implementation Plan of the NASA Office of Earth Science. Townshend earned his Ph.D. in Geography (Geomorphology) in 1971 from the University College in London.

Paul D. Try is a Senior Vice President and Program Manager at Science and Technology Corporation (STC) and Former Director of the WMO/WCRP International Global Energy and Water Cycle Experiment (GEWEX) Project Office. Try has expertise in meteorological in-situ and remote sensors (satellite and radar), as well as data collection, processing, exchange and archival activities. His recent STC management activities include meteorological satellite processing and application support activities for NOAA's National Environmental Satellite Data and Information Service (NESDIS), and management of three research support efforts at laboratories of NOAA's Office of Atmospheric and Oceanic Research Environmental Technology Laboratory, Air Resources Laboratory, and Forecast Systems Laboratory. Prior to joining STC, he served in the U.S. Air Force Air Weather Service where his responsibilities included oversight of the Automated Weather Distribution System (AWDS), and in the Office of the Secretary of Defense, with oversight of all DoD research and development in environmental sciences. Try is a fellow of the American Meteorological Society (AMS) and was president of the AMS from 1996 to 1997. He received his Ph.D. in atmospheric sciences from the University of Washington.

Steven J. Worley is the manager of the Data Support Section, Scientific Computing Division, Computational and Information Systems Laboratory at the National Center for Atmospheric Research (NCAR) where he has also work as a programmer IV and a programmer III. Before his work at NCAR, Worley worked at Texas A&M University as a research assistant and a research associate. He is involved in activities such as the lead for the U.S. data management data center for WWRP THORPEX (The Observing system Research and Predictability Experiment) TIGGE (THORPEX Interactive Grand Global Ensemble), the Users Working Group advisory panel for the Jet Propulsion Laboratory at NASA, the Integrated Ocean Observing System (IOOS) Data Management and Communications (DMAC) Steering Team and Expert Team Chair on data archiving. Worley is a member of the American Geophysical Union and the American Meteorological Society. He received his M.S. in Oceanography from the University of Alaska, Fairbanks in 1977.

Xubin Zeng is a professor at the University of Arizona in Tucson. Zeng's research in the past twenty years, through over 80 peer-reviewed publications, has covered atmospheric turbulence (theory, parameterization, its interaction with clouds and radiation, and large-eddy simulations), mesoscale modeling of atmospheric flow over complex terrain, chaos theory and its applications to the atmosphere, global land-atmosphere interactions, ocean-atmosphere interactions, sea ice-atmosphere interactions, monsoon dynamics, remote sensing, and most recently, nonlinear dynamics of vegetation. In the past ten years, he has focused on the land-atmosphere-ocean-sea-ice interface processes of the earth's climate system by integrating global modeling with remote sensing and in-situ data. He has acted as a bridge linking the remote sensing and field experiment community to the weather and climate modeling community. His research products (including models, algorithms, and value-added datasets) have been used worldwide by numerous groups (including NCAR, NCEP, ECMWF). He has extensive experience with most satellite land products and some experience with satellite atmosphere and ocean products. Zeng earned his Ph.D. in atmospheric sciences from Colorado State University in 1992.

Appendix B: Statement of Task

The ad hoc committee charged to conduct this study will assist NOAA as it develops plans to meet its data archiving and data access requirements. The committee will first produce a letter report that includes a preliminary list of principles and guidelines that NOAA can use to begin planning specific archiving strategies for the data streams it currently collects. This preliminary set of principles and guidelines for data archiving will be refined and expanded in a final report that also addresses the extent to which a wide variety of data sets should be made available. The final committee report will also include specific examples of how these principles and guidelines could be applied to existing and planned data streams across NOAA.

In summary, the committee will:

1. Consider the existing and planned observational and derived data streams collected by NOAA, along with current data management procedures, in the context of existing NOAA legal requirements.
2. Consider the relative costs of preserving and providing access to certain types of derived data products versus regenerating these data from archived first-stream input.
3. Develop a list of principles and guidelines regarding what types of data should be archived indefinitely (for at least 75 years), and what types of data could be stored for shorter durations under budgetary constraints.
4. Develop a list of principles and guidelines on how best to provide access to different variables, data sets and derived products, illustrated with a limited set of examples and case studies.

Appendix C: NOAA Data, Information, and Products

Guidance for Developing Principles and Guidelines for NOAA Data Archive and Access[11]

NOAA archives and provides access to a wide variety of data and products. These activities are based upon numerous legislative mandates. A few of many examples include:

1. 1 The National Climate Program Act of 1978 (15 USC CH29 PL 95-357) which calls for "… management & active dissemination of climatological data …"
2. 2 Magnuson-Stevens Fishery Conservation and Management Act (Public Law 94-265) which states that "The collection of reliable data is essential to the effective conservation, management, and scientific understanding of the fishery resources of the United States."

NOAA must continue to archive and provide access to all data as required by law. This is a fundamental principle that NOAA will strictly adhere to, so this NRC study can be of most value to NOAA by focusing on the scientific and societal value of NOAA's data and not the legislative mandates.

NOAA's data, information, and products have previously been grouped into seven broad categories.[12] Five of the seven categories are relevant to archived data and this NRC study. They appear bold text in the following list: **1) Original Data**; **2) Synthesized Products**; 3) Interpreted Products; **4) Hydrometeorological, Hazardous Chemical Spill, and Space Weather Warnings, Forecasts, and Advisories**; **5) Natural Resource Plans**; **6) Experimental Products**; and 7) Corporate and General Information. These seven categories are described in more detail below.

Original Data are data in their most basic useful form. These are data from individual times and locations that have not been summarized or processed to higher levels of analysis. While these data are often derived from other direct measurements (e.g. spectral signatures from a chemical analyzer, electronic signals from current meters), they represent properties of the environment. These data can be disseminated in both real time and retrospectively. Examples of original data include buoy data, survey data (e.g., living marine resource and hydrographic surveys), biological and chemical properties, weather observations, and satellite data.

Synthesized Products are those that have been developed through analysis of original data. This includes analysis through statistical methods; model interpolations, extrapolations, and simulations; and combinations of multiple sets of original data. While some scientific evaluation and judgment is needed, the methods of analysis are well documented and relatively routine. Examples of synthesized products include summaries of fisheries landings statistics, weather

---

[11] Submitted by NOAA-NESDIS to the NRC Committee on Archiving and Accessing Environmental and Geospatial Data at NOAA

[12] National Oceanic and Atmospheric Administration Information Quality Guidelines, 2002: http://www.noaanews.noaa.gov/stories/iq.htm

statistics, model outputs, data display through Geographical Information System techniques, and satellite-derived maps.

Interpreted Products are those that have been developed through interpretation of original data and synthesized products. In many cases, this information incorporates additional contextual and/or normative data, standards, or information that puts original data and synthesized products into larger spatial, temporal, or issue-oriented contexts. This information is derived through scientific interpretation, evaluation, and judgment. Examples of interpreted products include journal articles, scientific papers, technical reports, and production of and contributions to integrated assessments, n.b., this does not include the data and products used to make these interpretations. **These data are not applicable to the NRC study.**

Hydrometeorological, Hazardous Chemical Spill, and Space Weather Warnings, Forecasts, and Advisories are time-critical interpretations of original data and synthesized products, prepared under tight time constraints and covering relatively short, discrete time periods. As such, these warnings, forecasts, and advisories represent the best possible information in given circumstances. They are derived through scientific interpretation, evaluation, and judgment. Some products in this category, such as weather forecasts, are routinely prepared. Other products, such as tornado warnings, hazardous chemical spill trajectories, and solar flare alerts, are of an urgent nature and are prepared for unique circumstances.

Natural Resource Plans are information products that are prescribed by law and have content, structure, and public review processes (where applicable) that are based upon published standards (e.g. , statutory or regulatory guidelines which are not applicable to the NRC study). These plans are a composite of several types of information (e.g.,, scientific, management, stakeholder input, policy) from a variety of internal and external sources. Examples of Natural Resource Plans include fishery, protected resource, and sanctuary management plans and regulations, and natural resource restoration plans.

Experimental Products are products that are experimental (in the sense that their quality has not yet been fully determined) in nature, or are products that are based in part on experimental capabilities or algorithms. Experimental products fall into two classes. They are either 1) disseminated for experimental use, evaluation or feedback, or 2) used in cases where, in the view of qualified scientists who are operating in an urgent situation in which the timely flow of vital information is crucial to human health, safety, or the environment, the danger to human health, safety, or the environment will be lessened if every tool available is used. Examples of experimental products include imagery or data from non-NOAA sources, algorithms currently being tested and evaluated, experimental climate forecasts, and satellite imagery processed with developmental algorithms for urgent needs (e.g., wildfire detection).

Corporate or general information includes all non-scientific, non-financial, non-statistical information. Examples include program and organizational descriptions, brochures, pamphlets, education and outreach materials, newsletters, and other general descriptions of NOAA operations and capabilities. **These data are not applicable to the NRC study.**