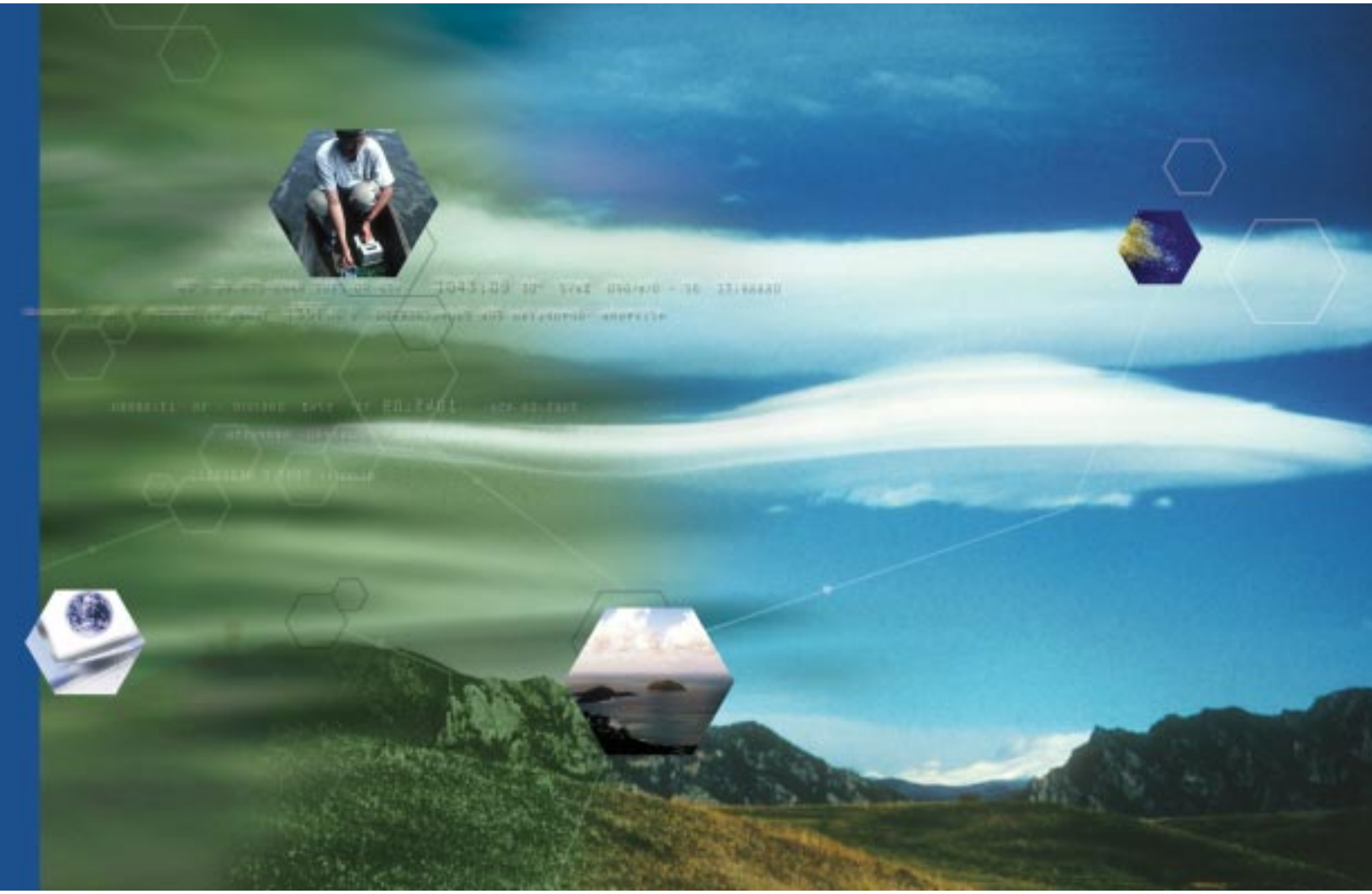
Cyberinfrastructure for Environmental Research and Education



Report from a workshop held at the
National Center for Atmospheric Research
October 30 - November 1, 2002

Sponsored by the National Science Foundation

Cyberinfrastructure for Environmental Research and Education

Report from a workshop held at the
National Center for Atmospheric Research
October 30 - November 1, 2002

Sponsored by the National Science Foundation
September 2003

Table of Contents

Introduction ..... pg 1

Organization and Methodology ..... pg 1

Background and Definitions ..... pg 2

Why Cyberinfrastructure Matters for Environmental Research & Education ..... pg 3

Challenges ..... pg 4

Near-term Opportunities ..... pg 5

Scientific, Educational, & Societal Benefits ..... pg 8

Coastal Ocean Prediction & Resource Management ..... pg 9

Regional Observation, Analysis, & Modeling of the Environment ..... pg 10

Data Assimilation for Carbon Cycle Science & Biogeochemistry ..... pg 11

Focused Integration of Geographic Information Systems in the Environmental Sciences ..... pg 12

Recommendations ..... pg 13

References ..... pg 14

Appendix 1—Workshop Participants ..... pg 15

Introduction

In November 2002, a group of university and public sector investigators, in collaboration with the National Center for Atmospheric Research (NCAR), hosted a workshop to identify key issues and opportunities in the systematic development and application of hardware and software (hereafter referred to as cyberinfrastructure) for environmental research and education over the next decade. Approximately 100 researchers, educators, industry representatives, and government program managers (listed in Appendix 1) participated in this three-day event. The participants represented many of the intersecting disciplines within the broad environmental sciences (ocean, solid earth, and atmospheric sciences; ecological science; remote sensing; biodiversity science; etc.), as well as the computer and computational sciences. The workshop had two main objectives:

- to provide advice and background information on how one might structure and implement future cyberinfrastructure programs,
- to facilitate communication and partnerships among interested environmental researchers and educators as they define and implement cyberinfrastructure projects.

Organization and Methodology

The workshop included a series of plenary presentations with associated discussions. Four interdisciplinary breakout groups focused on functional aspects of environmental research and education and the current and potential role of cyberinfrastructure for the broad environmental arena:

- collecting data and making it available,
- generating and using data: data assimilation, analysis, and modeling,
- collaboration tools and strategies,
- creating a new kind of environmental scientist.

On the third day, breakout groups representing disciplinary areas (ocean, atmosphere, solid Earth, and computer sciences) addressed the same topics from their perspectives. In this way, the primary issues were discussed from multiple vantage points. Each group was asked to identify cyberinfrastructure needs, major challenges, broad principles to guide implementation, near-term opportunities for progress, issues requiring further investigation, and the best means of fostering cooperation between computer scientists and environmental scientists.

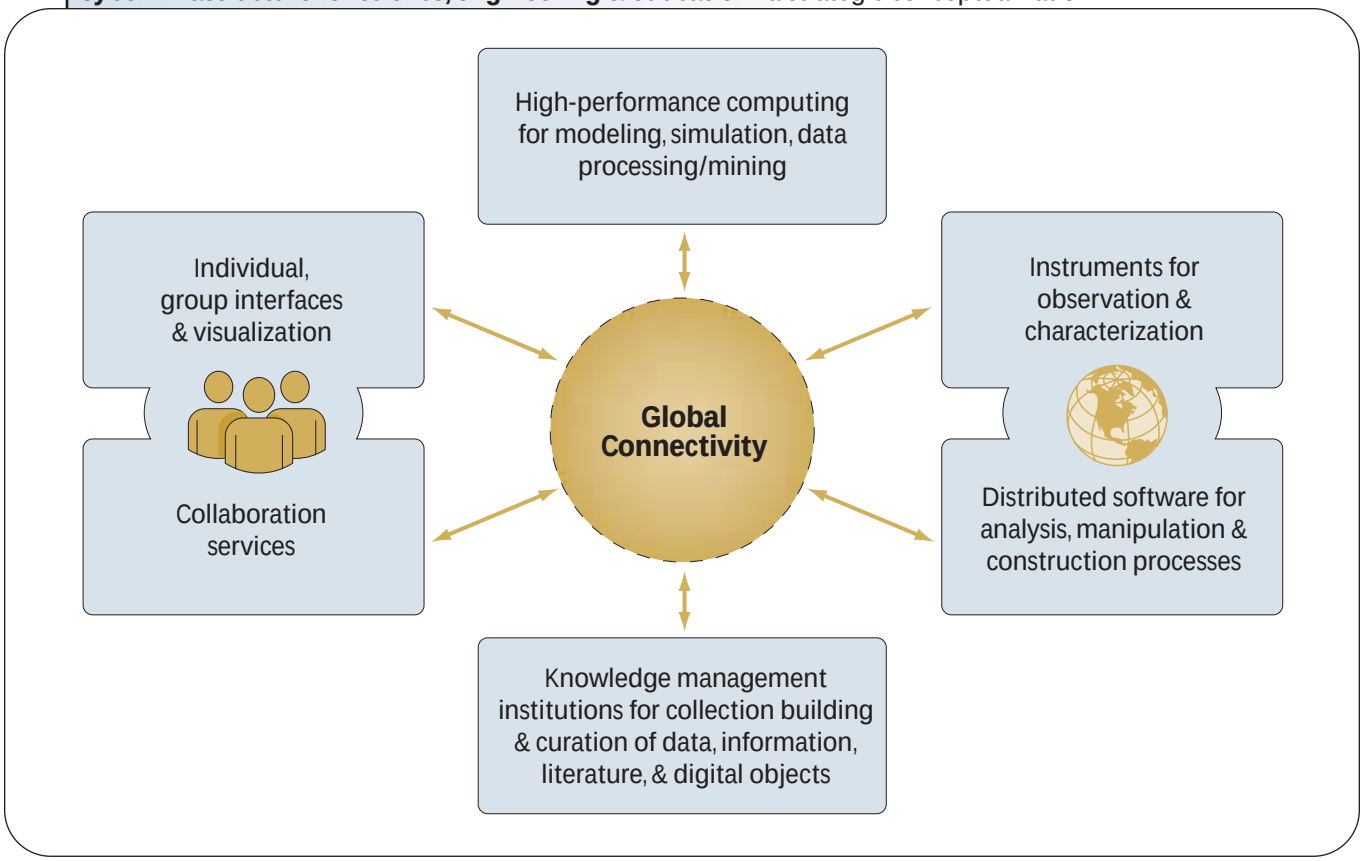
The workshop resulted in two main products. This document is intended as a high-level synthesis of the workshop results. It was authored by the workshop steering committee (listed in Appendix 1) and focuses on those issues identified during the workshop that the committee sees as most important. The conference Web site (http://www.ncar.ucar.edu/cyber/index.htm) includes more detailed information about the workshop breakout sessions, copies of the presentations made during the workshop, and links to other relevant sites. The steering committee emphasizes that [we do not see these reports as providing definitive community consensus]. We rather see this work as an initial step in a complex process of identifying and agreeing on cyberinfrastructure priorities for environmental research and education.

Background and Definitions

There are many overlapping definitions of cyberinfrastructure. For the purposes of this report, we have used the term to mean the set of reliable, well-specified, and interoperable connections of electronic hardware and software that allows people to discover, learn, teach, collaborate, disseminate, access, and preserve knowledge in their domain. Cyberinfrastructure extends from the scientific instrument to the desktop of the working scientist and encompasses networks, models, data sets, metadata, data archives, data analysis and manipulation tools, as well as communication and collaboration tools and environments. The successful development and use of cyberinfrastructure are largely dependent on successful integration of the work of computer and computational scientists, domain scientists, social and behavioral scientists, engineers, and information technologists.

We use the term environmental research and education in the broadest possible sense to mean activities intended to understand, document, and explain the Earth system. Environmental research and education thus embrace the entire suite of environmental disciplines, including the oceanic, atmospheric, biological, hydrological, ecological, geographical, geologic, and solar-terrestrial sciences, environmental engineering, and social science and computer science focused on or relevant to environmental issues. It also encompasses interdisciplinary activities such as research and education focused on climate change, coupled human and natural systems, coupled biological and physical systems, the relationship of people and technology, and other areas.

Cyberinfrastructure for science, engineering & education: a strategic conceptualization



Why Cyberinfrastructure Matters for Environmental Research & Education

Environmental research and education are characterized by a number of attributes that make cyberinfrastructure especially important for this field of scientific endeavor. Many environmental research activities are observationally oriented, rely on the integration and analysis of many kinds of data, and are highly collaborative and interdisciplinary. Much of the relevant data needs to be geospatially indexed and referenced, and there is a host of currently noninteroperable data formats and data manipulation approaches. Spatial scales vary from microns to thousands of kilometers; time scales range from microseconds (for some fast photochemical reactions) to centuries or millennia (for paleoclimate and Earth evolution studies); and data types range from written records and physical samples to long-term instrumental data or simulation model outputs.

The successful development and deployment of a new generation of satellite, ground-based, and aircraft observing systems over the last decade have resulted in the ongoing production of very large data sets that must be ingested, analyzed, and archived in an easily accessible fashion. Further increases in data volumes are expected. For instance, NASA's Terra satellite produces about 194 gigabytes (GB) of raw data per day; this total rises to 850 GB per day when processed to higher levels. At the other end of the range are individual scientists (and students) gathering data one point at a time who want to collect, archive, and analyze their results with sophisticated tools and integrate their results into the global knowledge base. The societal and policy relevance of much environmental research has resulted in significant funding but is also resulting in demand for analysis and production of results on time scales of days to months rather than years to decades. In some cases, real-time or near-real-time data collection, distribution, use, and response are required for decision support. Many scientific investigations and policy and management applications rely on geospatially referenced data, and the effective integration of these data with data that are not geospatially referenced is increasingly important.

These and other factors have led to the ongoing development of environmental research cyberinfrastructure. There is a strong consensus in the broad environmental science and education community that (1) increased support for and (2) more effective application of such cyberinfrastructure will

- accelerate discovery,
- increase scientific productivity and educational effectiveness,
- revolutionize existing disciplines,
- foster new interdisciplinary science and create new fields of scientific endeavor,
- democratize science—education and research—by making a full range of capabilities available to anyone with Web access, and
- enable new connections between science and society (government, communities, businesses) that could lead to significant societal and economic benefits.

Comprehensive development and deployment of modern cyberinfrastructure for environmental science and education will revolutionize the individual disciplines. It will also create new hybrid disciplines that will fundamentally change the way that science is taught and transform the relationship between scientific research and societal decision making. It was the strong sense of the participants in the workshop that a full commitment to the development of environmental cyberinfrastructure is a *sine qua non* for the success of planned major research programs within the National Science Foundation and other agencies, as well as for multiagency initiatives in Earth system science. The question therefore is not whether to embark on an aggressive program but rather how to shape it to optimize the effective use of resources and augment its positive impact on science and society.

Challenges

Enhanced communication and collaboration among heretofore separate efforts, and planning that considers environmental cyberinfrastructure as a whole, are needed to assure that the potential benefits of environmental research are realized. For instance, continued improvements in high-end supercomputing capacity, accessibility, and networking are considered necessary to meet growing scientific requirements, but these improvements must be accompanied by software enhancements and improved collaboration tools in order maximize scientific benefits. Improvement in tools and capabilities must in turn be accompanied by greater attention to educational and workforce issues, with particular attention to more effective teaming of computer scientists, software engineers, and environmental scientists and better integration of computer and computational science in environmental coursework. Development of such a holistic approach will be challenging. Duplication of effort, wasted resources, and missed opportunities are real dangers. "Last mile" issues (extending network capabilities to individual end users) and legacy data issues, as well as support for usable permanent data archives, have proven to be persistent problems.

There was widespread agreement that many of the most fundamental challenges are cultural and sociological. The successful development, deployment, and application of cyberinfrastructure are fundamentally about making connections and fostering a new level of collaboration among scientists, educators, and students. The most obvious example is the need for deeper collaboration between environmental domain researchers and computer and computational scientists. Agreement on this point recurred across presentations and breakout group reports—all four breakout groups listed it as a major challenge—along with recognition that the prevailing cultures in environmental and computer sciences do not fully support such collaboration. Matching needs and capabilities is difficult. When it comes to cyberinfrastructure, most environmental scientists and educators value reliability and usability over cutting-edge innovation, while many computer scientists value innovation more. Another aspect of this problem is reward structures. Even within the broad environmental domain, cross-disciplinary and interdisciplinary activity tend not be valued as highly as work within disciplines in the granting of promotions, tenure, and awards. Cooperation with the computer sciences is even less important in building professional credentials and reputation. The same holds true for infrastructure work, such as the assembly and curation of large data sets and the development and management of software.

The workshop also identified programmatic challenges. Interagency coordination was seen as a major issue. In the environmental sciences, the Department of Energy, the National Aeronautics and Space Administration, and the National Oceanic and Atmospheric Administration provide major support to the academic community for activities that could be characterized as environmental cyberinfrastructure. The federally mandated National Spatial Data Infrastructure and Geospatial Data One-stop initiatives are run through the U.S. Geological Survey in the Department of the Interior. Enhanced coordination among these existing efforts and any new National Science Foundation program in cyberinfrastructure is needed to minimize overlap and duplication, ensure that there are no major gaps, and enable NSF to profit from lessons learned in other programs. It is not clear that any of the existing coordination mechanisms in IT or environmental research are adequate for this task.

Table 1: **Environmental Research & Education Cyberinfrastructure Needs**

Hardware and Systems
- High-speed networking
- Last mile solution
- More, and more available, computing cycles
- Environmental research synthesis centers

Software and Services
- Interoperability
- Agent technology
- Tool boxes and tutorials
- Development and deployment of interaction tools

Data
- Analog-to-digital conversion for preservation of legacy data sets
- Data integration and representation
- Metadata standards, linkages to data
- Documentation and quality control

Sociocultural
- Community building
- Reward systems and incentives

Long-term, sustained support for cyberinfrastructure development is itself another key programmatic challenge. There is no clear end point for such an effort, and some aspects of cyberinfrastructure development are unlikely to be effectively supported through three- to five-year grants. Sustaining cyberinfrastructure is likely to require a new business model and adequate expertise in running production and perhaps customized systems and services. Successful implementation of environmental cyberinfrastructure programs will depend on

- a process for building and sustaining shared infrastructure within and across communities,
- application of best practices and lessons learned so that the current cyberinfrastructure can become as reliable as the physical infrastructure we all rely on in our daily lives,
- program structures that substantially increase community buy-in for metrics, interoperability, and standardization.

As noted throughout this document, environmental cyberinfrastructure falls at the intersection of the environmental and computer science domains, making the structuring of competitive opportunities and the effective review of proposals quite difficult. In the current system, innovative proposals risk having too much computer science for environmentally oriented reviewers and too much environmental science for computer science–oriented reviewers.

Even the NSF Information Technology Research (ITR) program, which requires integration of domain and computer sciences, routinely denies funding for cyberinfrastructure projects because reviewers usually prefer proposals in which both components are innovative over proposals that are primarily development or application oriented. The former often consume most of the limited funds available.

Several scientific and technical challenges were also seen as significant. The rapid pace of technological change and its mismatch with the timelines of major environmental research projects are seen as a major issue. The ongoing evolution of computer hardware, driven largely by commercial forces, has reduced the cost of high-end computing but has forced environmental scientists to frequently adapt codes and algorithms to new architectures. The problem of optimizing high-end climate and weather models for parallel architectures is one example, but similar difficulties afflict many smaller-scale modeling, analysis, and visualization activities, including the migration of advanced programming languages to high-end platforms. Another important technical challenge is the very rapid and ongoing increase in the volume of data from large-scale ground- and space-based measurement systems and simulation models. It is becoming difficult to continue to archive all raw data produced by such systems, leading to formidable questions about which data should be preserved and how to provide data access, discovery, and retrieval.

Near-term Opportunities

The workshop identified a wide variety of environmental cyberinfrastructure needs (see Table 1) and opportunities. The following discussion focuses on five areas where potential benefits seem most widespread and near-term scientific payoffs most likely.

1. Development and deployment of more effective and economical collaboration tools to enable deeper research cooperation among researchers at widely dispersed locations. Especially important are basic, affordable, and ubiquitous videoconferencing capabilities and more sophisticated means of collectively visualizing and manipulating data in real time. Ideally, multiple entry points with varying capabilities would be available, and the ongoing design and upgrade of interfaces would be informed by behavioral research to ensure that they are as usable as possible. The need for improved interaction tools cuts across research and education and was seen as fundamental to enabling deeper collaboration among disciplines and to facilitating access to information and analytical capabilities, thus stimulating democratization of science and helping to remove barriers for underrepresented groups.

Opportunities for near-term progress include

> Formation of new environmental research and education collaboratories that would encompass scientific (disciplinary or interdisciplinary) research programs, interlinked observational networks, data analysis and modeling facilities, and educational activities. These could be regional or could be organized around the cross-cutting research themes of *Complex Environmental Systems: Synthesis for Earth, Life, and Society in the 21st Century* (AC - ERE, 2003). They could help galvanize the development and application of new collaboration tools while promoting the interdisciplinary approaches needed to improve understanding.

> A major expansion of the AccessGrid, or similar technologies, across the academic community. Such an expansion would enable more routine interactions of investigators at dispersed locations and should be coupled with funding opportunities to build on, improve, and generalize the current generation of collaboratory technologies.

2. Improvement in the accessibility and usability of computational tools and the interactive capabilities of data systems, data archives, and repositories through better documentation and quality control, more sophisticated data portals and related methodologies, and development of cross-disciplinary solutions and tools, including discipline-neutral standards for interoperability. Near-term opportunities (that could perhaps be supported through a solicitation for innovative methods for data analysis, storage, and manipulation) include

> Development of data grids and federation of all classes of databases for community sharing, based on best practices for the systems and services and their federation. The development of tools for handling special data types is another area that is ripe for progress (examples: True 3D data, structured vs. unstructured data, relational vs. "flat" data sets), as are data interface standardization (or a process for standards consensus), availability, and preservation. Analysis and visualization tools are crucial for advancing scientific knowledge and hinge directly on cyberinfrastructure that supports transparent distributed data access and flexible coupling with collaboration environments.

> The creation of a roadmap for the application of cyberinfrastructure that could include information on how to develop and/or adapt and use tools such as ontologies, data mining, and data cleansing. Ideally, this could be coupled with the development of an environmental research and education cyberinfrastructure portal or interface that could be called up at any site and would make the roadmap, tools, and toolboxes easily available.

> Further development of digital libraries with a distributed system of permanent archives, including deeper links among the existing digital library projects and federations (which umbrella the projects) that would respond to the education needs of environmental research beyond their current discipline-specific focuses. (It is possible that this can be done for collaboratories as well.) Formalizing teaching modules for environmental science and engineering, with mechanisms for review, feedback, etc., and insertion of these into a digital library, should be part of this effort.

3. Model simulation, where the need for more capable and accurate models is coupled with requirements for better interoperability among models and model components, more sophisticated data assimilation techniques and systems, more capable hardware, and improved access to computing capabilities. Near-term opportunities for progress include

> The development of large-scale community models of Earth system processes that integrate the efforts of many investigators in community code development, dissemination, and intercomparison. Such efforts have resulted in significant scientific progress and community building in climate and weather modeling, and are clearly applicable in oceanography, solar-terrestrial physics, and other fields.

> Interoperation of model and experimental data sets such as in the Network for Earthquake Engineering Simulation grid (NEESgrid), and the development of data assimilation methodologies that more effectively use observational data streams in simulation activities. Another area of particular promise is the interoperability of atmospheric data sets and models with Geographic Information Systems.

> Application of modern software engineering practices to small- and medium-scale model development, including the definition of software frameworks, analysis and evaluation of model component interoperability (similar to the current Earth System Modeling Framework (ESMF) effort focused on climate and weather models) and creation of flexible codes that are easily adaptable to new architectures. Such activities hold significant promise for easing model development by increasing the availability of well-documented software components that could be applied in diverse development and application efforts.

4. Continued improvement in scalable computing capabilities and access to such capabilities. It was generally agreed that major progress has been made in establishing widespread high-speed networks in the United States, but that requirements are growing faster than capabilities. The major needs are for extending networking capability and access (the "last mile" problem and global connectivity). Further improvements in computing power and accessibility of high-end computing systems are also critical. Workshop participants concurred that computing requirements in environmental science overall continue to far exceed available resources, especially in the areas of climate and ocean modeling. In other areas, existing computational resources are adequate but not accessible, efficacious, or integrated. Significant investment and progress in networking and, even more importantly, in software and services, are needed to improve access and usability. Near-term opportunities include

> Increasing the access of small and medium-sized centers and individual investigators to computational resources and improving the use of these resources. Coupled with software improvements, improved connectivity among these groups and between them and major computing centers would help to bring computing resources and services to a wider range of investigators.

> Improving the connectivity among major research centers (i.e., through expansion of the TeraGrid and other grid programs and networks). Such improvements would enable deeper connections among the high-end National Partnership for Advanced Computational Infrastructure (NPACI) centers and the researchers in and communities served by environmental research centers, and would facilitate new software and data management collaborations and more-rapid transfer of new computer science capabilities into the environmental science domain.

5. Cyberinfrastructure community building. There is a strong need to enhance relationships and team building between the computer science and environmental domains. Potential innovations include educational modifications aimed at nurturing a new generation of scientists and educators who are more informed by both domains, improving support services for, and thus the inclusion of, lower-end users, and forging new links among activities that are now pioneering environmental cyberinfrastructure development and application in relative isolation. One of the more interesting aspects of the workshop was the widespread recognition of common cyberinfrastructure challenges in different environmental disciplines and the potential for significant benefits and efficiencies through more-regular exchange of information and the development of new partnerships and collaborations. Near-term opportunities include

> Support for interdisciplinary team proposals and, in particular, for those that integrate the efforts of software engineers and environmental domain scientists in the definition and development of new applications. Such interdisciplinary activities will build the cyberinfrastructure community and accelerate progress towards a robust, reliable, and usable cyberinfrastructure that serves a wide range of users and enhances environmental research and education.

> Leveraging the collaboratories. The existing suite of collaboratories offers a strong base to build on. We suggest taking advantage of the cyberinfrastructure within existing collaboratories and educational systems that provide immediate, exciting, and relevant science/engineering, and establishing a new class of "production" collaboratories with dedicated cyberinfrastructure funding for coordinating activities among distributed scientific communities.

> Continued workshops. An ongoing series of cyberinfrastructure workshops would facilitate discussion, cooperation, and community building. An annual or biennial general conference on environmental cyberinfrastructure could serve as an anchor for the developing community. More specialized topical workshops that involve scientists, educators, and practitioners would be an effective means of disseminating education and training materials and instruction in how to use cyberinfrastructure-based tools. Workshops organized around the interdisciplinary research areas outlined in *Complex Environmental Systems* would help with defining specific requirements. Workshops are also an effective means of gathering various environmental research and education communities together to identify and describe the grand challenges that could be facilitated by cyberinfrastructure (examples: documentation of biodiversity, natural hazard and disaster prediction and response, human health and the environment).

> On-line journals. Electronic journals have helped revolutionize some areas within the physics community but have not yet had the same success in the environmental sciences. On-line journals such as the American Geophysical Union's *Geochemistry, Geophysics, Geosystems* (G3) may be the prototype for the rest of environmental cyberinfrastructure. Methods to support and expand e-journals ought to be pursued within and across disciplinary lines.

> New reward systems. Finally, the development of cyberinfrastructure reward systems using new metrics (data citations, interdisciplinary collaborations, education, etc.), and possibly including the establishment of a new cyberinfrastructure journal, and the integration of computer science and information technology into environmental graduate education provide a means of building the environmental cyberinfrastructure community.

Table 2: **General Principles to Guide Implementation of Environmental Cyberinfrastructure**

Effective new environmental cyberinfrastructure programs and projects should evolve from and take full advantage of current efforts; be responsive to evolving technology and changing scientific, educational, and societal requirements; and address the needs of the full scientific community, not just those engaged in cutting-edge technological areas.

- Open systems and free and open exchange of data, software, and results are critical to enabling collaboration and ongoing assessment of techniques and practices.
- Environmental research and education needs should drive the development of environmental cyberinfrastructure. Environmental cyberinfrastructure should be based on a marriage of existing and new technologies, responsive to user needs, robust, stable, reliable, and adjustable in response to lessons learned and feedback from users.
- Environmental cyberinfrastructure programs require constant balancing of top-down guidance and bottom-up self-direction, innovation and reliability, and disciplinary and interdisciplinary priorities.
- Creating and maintaining cyberinfrastructure requires an ongoing commitment on the part of sponsors. Many of the necessary projects are inherently long-term, and success will create and increase user demand for further advances.
- Collaborations, linkages, and portals to the broader community should be part of environmental cyberinfrastructure programs and projects. A new level of cooperation among agencies, environmental science disciplines, computer scientists, and environmental domain scientists, and among researchers, technologists, and educators is needed to realize the promise of environmental cyberinfrastructure.
- Making environmental cyberinfrastructure work for education and vice versa requires that educational considerations be part of initial design and definition of programs, including processes for review, quality control, validation, and credentialing for all education and training materials. Environmental education must include a greater emphasis on IT and computer science.

IMPLEMENTATION MUST BE EFFECTIVE, LEVERAGEABLE, DISTINCT, AND HIGHLY VISIBLE.

Scientific, Educational, & Societal Benefits

Workshop participants identified a wide range of benefits that are likely to accrue from enhanced support for environmental cyberinfrastructure. The following examples illustrate the kind of innovative activities that could be facilitated.

1. ***Coastal Ocean Prediction & Resource Management***
2. ***Regional Observation, Analysis, & Modeling of the Environment***
3. ***Data Assimilation for Carbon Cycle Science & Biogeochemistry***
4. ***Focused Integration of Geographic Information Systems in the Environmental Sciences***

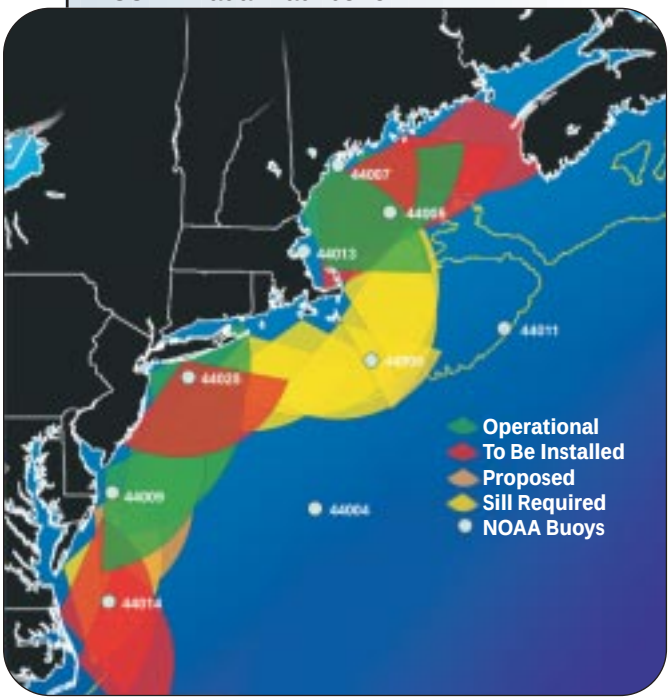
1 Coastal Ocean Prediction & Resource Management

The coastal ocean is of fundamental importance to many activities—including fisheries, homeland defense, recreation, and human health. It is also the region of the ocean in which human activity has perhaps the strongest impact through the modification of freshwater runoff, the introduction of large inputs of nitrogen and other nutrients, the episodic release of pollutants, the physical modification of the seafloor by dredging, and the alteration of ecosystems by fishing. In the coastal ocean, the physical circulation is extremely complicated, the seafloor morphology itself evolves as a result of sediment transport, and the structure of biological communities varies rapidly on relatively short space and time scales.

Studies of the coastal ocean are an area where many of the pressures on cyberinfrastructure resources are acutely felt. The potential payoff from making additional cyberinfrastructure investments is therefore large. "Recent advances in computational capabilities, combined with increasingly sophisticated observational technologies (e.g., remote sensing, telemetry, networking, autonomous underwater vehicles, long-term monitoring systems) present unprecedented opportunity to advance understanding of shelf and estuarine systems and their management" (OITI, 2002). The challenges presented by the analysis of coastal data, the assimilation of coastal biogeochemical and physical data, and the development of whole-system models that integrate physics and biology and combine small scales (e.g., river estuaries and coastal inlets) with large scales (e.g., continental shelves of the United States and the basins to which they are attached) are profound.

The coastal ocean continues to be a region of intense observational activity with a growing network of coastal observatories, a spreading network of coastal radars that map surface currents (see below), the development and testing of autonomous underwater vehicles to remotely sample interior property fields, video recorders that monitor the nearshore wave field, and other instruments that provide routine physical, chemical, and fisheries observations. A particular challenge to the coastal ocean science community will be the integration and interoperation of these disparate data systems and their use to produce hindcast and operational (forecast) products for research and coastal ocean management.

NEOS HR-Radar Backbone



Coastal radar backbone for the proposed NorthEast Observing System (NEOS). Systems shown in green are currently operational; those shown in red are planned for future installation. See http://marine.rutgers.edu/neos/.

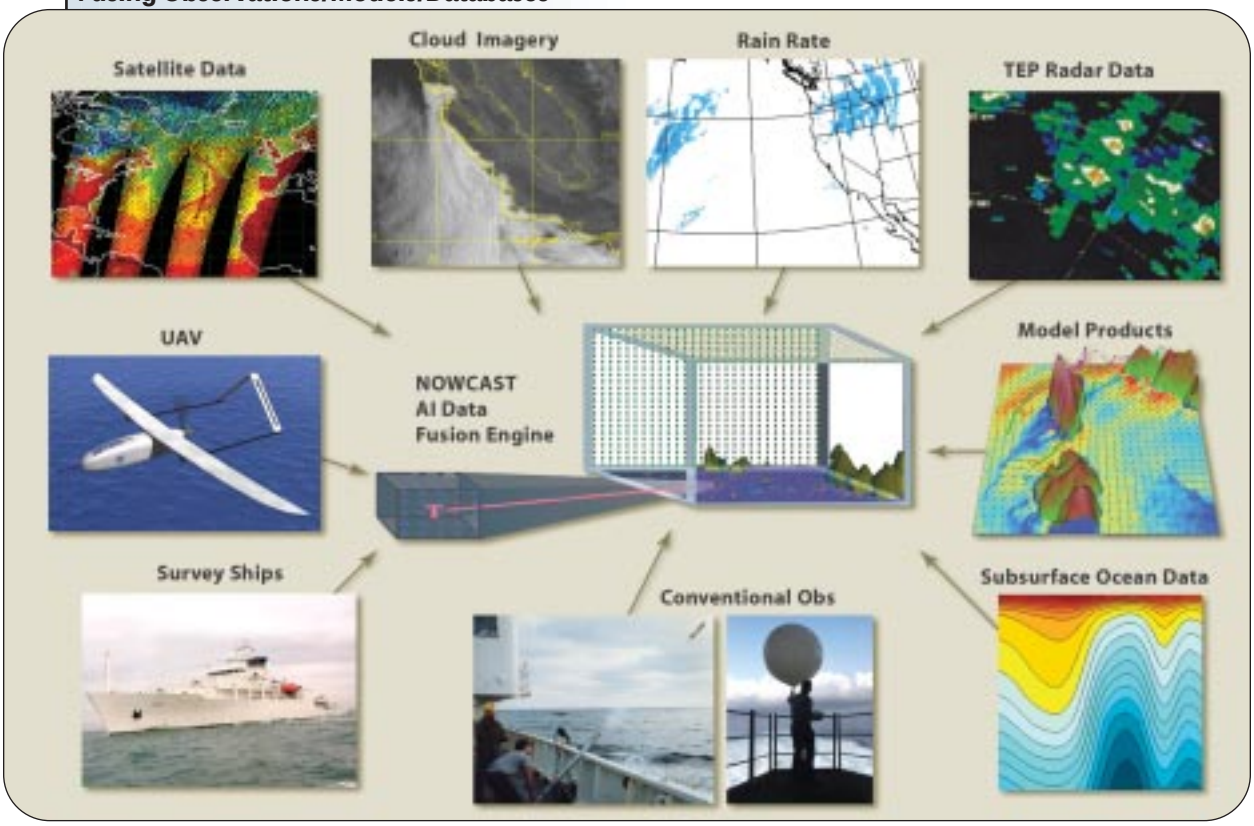
2 Regional Observation, Analysis, & Modeling of the Environment

Over the last decade, significant effort and funding have been applied to the global-scale observation, simulation, analysis, and assessment of climate change and other environmental stresses, leading to many impressive results, including the creation of global baseline data sets in some key areas. It is becoming clear, however, that global-scale activities—whether in observing systems, data systems, or modeling—are not sufficient as a scientific approach to understanding the consequences of global change. There is an increasingly evident need for regional-scale interdisciplinary investigations that integrate observations, data and information systems, process studies, and models with the process detail and spatial resolution necessary to advance scientific understanding of ecosystem processes and the impacts of multiple interlinked environmental stresses. Ongoing improvements in cyberinfrastructure (computational capabilities, networking, data systems, software, observational technologies, and data assimilation methods) are coming together to make the implementation of such projects feasible.

Weather and climate models can provide the core for a broader regional environmental forecasting capability, to which models of air quality, river flows, ecosystems, and other variables can be added. Linking such modeling efforts to advanced sensor webs and information systems would provide the opportunity to initiate experimental forecasts of new variables, assess impacts and responses, and advance scientific knowledge. Cyberinfrastructure is the key to developing such efforts because of the need to integrate the work of regional networks of experts, distributed over institutions and locales. Typically, regional projects also must integrate a diversity of data types, from continuous operational measurements to episodic research-oriented process studies. There is an increasing need for integration of human dimensions data with geophysical information and the combination of both within some sort of geospatial framework. Models used at the regional scale cover smaller domains than global models but typically have extremely high spatial resolution and so require computing resources comparable to global models. Overall, regional networks generate a need for cyberinfrastructure that, in terms of spatial resolution and the diversity and detail of data required, poses great challenges and represents an opportunity for creative large-scale developments.

Such activities would advance scientific understanding, help inform regional and national decisions, provide many opportunities for graduate and undergraduate education, and complement and enhance the ongoing suite of global-scale investigations of environmental processes and change. They could draw on larger- or global-scale data and models to establish boundary conditions, and the results and insights from regional projects could in turn be aggregated to synthesize larger-scale results that would assist with diagnosis and verification of global-scale activities. Moreover, once such activities are in operation, demonstration of their benefits would likely generate interest in replicating them with linked activities in other regions, perhaps yielding progress toward the unmet goal of integrated national or global observing systems.

Fusing Observations/Models/Databases



3 Data Assimilation for Carbon Cycle Science & Biogeochemistry

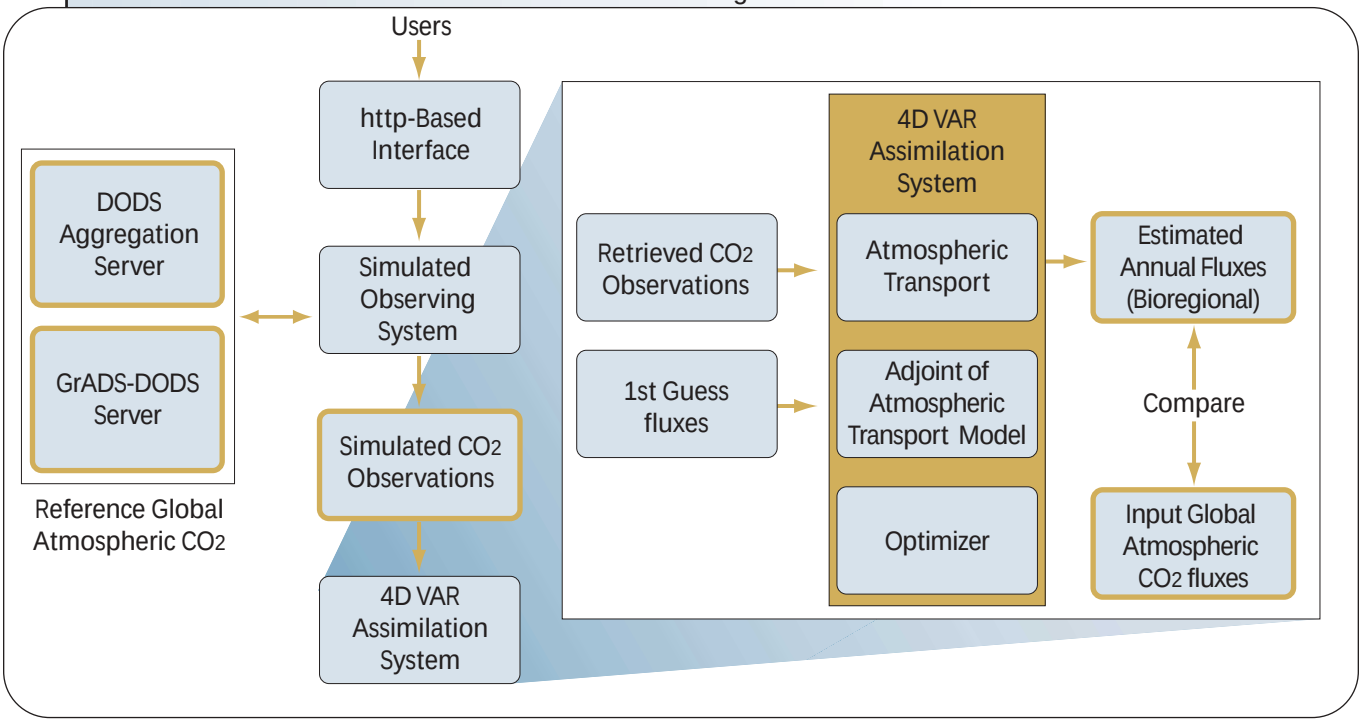
Improving our understanding of the exchange of carbon and other key species among the Earth's atmosphere, oceans, soils, and living vegetation is one of the highest priorities in climate change science. More accurate projections of long-term climate change resulting from emissions of carbon dioxide and other trace gases and aerosols and effective evaluation of mitigation and adaptation options are both dependent on such improvement. One of the key impediments is the difficulty and complexity of assembling and analyzing a suite of observations of the biogeochemical cycles as an integrated whole. There are many measurement programs producing relevant data, but these tend to focus on parameters within individual reservoirs or subsets of biogeochemical species. Observing and understanding a single reservoir (land, atmosphere, or oceans) or cycle (carbon, nitrogen, ozone, dust) does not translate to understanding that reservoir or cycle within the coupled system.

The development and application of new data assimilation tools hold great promise for meeting this challenge. Data assimilation is a family of techniques for improving estimates of geophysical quantities combining models and observations. Although best known as a tool in weather forecasting, data assimilation is also used in analysis of complex ocean and chemistry data sets, and in estimation of parameters in models. It is well suited to addressing the challenge of integrating atmospheric, terrestrial, and oceanic data into a common analysis framework for carbon cycle science. Data assimilation techniques can also be used to simulate the impact of new observations on analyses of carbon sources and sinks.

The state of the art in data assimilation is relatively mature, and experience gained in ocean, chemical, and meteorological data assimilation can provide a head start for application of this suite of techniques to the carbon cycle problem. Initial work on defining a research agenda suggests an approach including assimilation model development, reanalysis of key phenomena such as the carbon system response to El Niño, education and training of a new generation of carbon cycle modelers in assimilation techniques, and the creation of a data clearinghouse and prototype data assimilation teams.

A robust carbon cycle data assimilation effort would represent a creative and timely application of cyberinfrastructure to accelerate progress in an important scientific area. Specific requirements include data interoperability and availability, so that research and operational data on the cycles and domains can be integrated. The massive computing requirements for biogeochemical data assimilation will lend themselves to advanced technologies such as supercomputing and grid computing. The diversity of the cooperating communities motivates researchers to use collaboration technologies including advanced networking, conferencing, and shared computing and data. In addition, developing assimilation models for biogeochemistry will of necessity require specialized groups distributed around the nation and world to develop component models and modules and an infrastructure to unify these components into a system. The resulting advanced techniques and methods would in turn be applicable to the study and explanation of other important biogeochemical cycles.

Carbon Data-Model Assimilation: Retrieval of fluxes using data assimilation



4 Focused Integration of Geographic Information Systems in the Environmental Sciences

Geographic Information Systems are a class of information systems used for visualizing, analyzing, and managing spatial data. GIS provides domain-neutral, geospatial frameworks that naturally lead to the amalgamation of cross-disciplinary data and analysis. Many of the cross-cutting needs and issues identified during the workshop have been at the core of GIS development and use over the last decade:

- data and component interoperability,
- metadata standards and data discovery services,
- development of common data models and ontologies,
- federation of data stores,
- distributed networks for data delivery,
- analysis and visualization tools,
- coupling of computer, information, and environmental sciences,
- educational ties to K-12.

The integration of GIS into the environmental sciences involves three broad types of activities: (1) the use of existing GIS tools for scientific endeavor, (2) the modification of GIS architectures to better address scientific needs, and (3) the modification of environmental information systems to incorporate advances in GIS development (listed above). These integration activities provide new collaboration opportunities in many areas.

For example, addressing the data modeling complexities introduced by the need to manage four-dimensional environmental data sets (which can be very large and may change in real time) will drive deeper linkages among GIS developers, environmental scientists, and computer and information scientists. The application of GIS to studying the complex interactions between the built and natural environments will add social scientists to this mix and require an unprecedented level of integration of data sets from social science and physical science.

The technological integration that is needed to merge environmental information systems with GIS also offers substantial opportunities for new partnerships with industry. By advancing the awareness of industry partners and working with them to implement information systems that address the needs of scientific research, we can lighten some of the financial burden on the scientific community for developing its own customized technology. By infusing the environmental science community with a deeper understanding of the industry processes required to produce high-quality, production-grade systems, we can expect to improve the quality of software and information systems emerging from the environmental disciplines. Finally, developing interfaces with, and in some cases adapting and using, common industry supported tools, rather than continuing to rely on highly sophisticated custom tools, should result in easier public access to scientific data and information.

Recommendations

Workshop participants agreed that increased and sustained support of cyberinfrastructure for environmental research and education represents a significant opportunity for scientific progress and that an agency like the National Science Foundation (NSF) is well positioned to lead an effort to enhance such support. Participation in the workshop and review of the materials presented have led the steering committee to propose the following recommendations.

NSF should establish a cross-cutting program to support the development and deployment of cyberinfrastructure for environmental research and education (see Table 2).

As part of such an effort, NSF should

- Establish an "Environmental Research and Education Cyberinfrastructure Federation." Participants should commit to achieving interoperability and contribute access to data, results, model components, methods, lessons learned, and technology adaptations. They should receive funding to cover nominal costs of sharing their data and documenting their projects, collaborative opportunities, and lessons learned.
- Initiate a new suite of environmental science collaboratory efforts that could include centers focused on environmental synthesis, environmental research grand challenges, and regional environmental observations, modeling, and analysis, as well as "production" collaboratories with explicit funding for infrastructure and clearly articulated missions and roles for coordinating activities among distributed scientific communities.
- Establish ongoing competitive funding for cyberinfrastructure as a complement to the ongoing Information Technology Research (ITR) competition. Proposals should place significant weight on cyberinfrastructure capacity building in domain sciences (similar to existing NSF programs for physical infrastructure) and broader societal impact, as opposed to the primary emphasis on innovation in the existing ITR program.
- Fund workshops focused on educational applications of cyberinfrastructure and the cyberinfrastructure aspects of the environmental grand challenges outlined in *Complex Environmental Systems*.
- Ensure that NSF solicitations in environmental research and education have a specific cyberinfrastructure element to promote further development and documentation of innovative techniques and methods and engender new collaborations.
- Explore means of further enhancing computational capacity and connectivity, including improving links between and among individual researchers, small and medium-sized centers, and large centers in the environmental domain, and improving cooperation between computational science centers, environmental domain centers, and individual researchers.
- Support the development and deployment of new collaborative technologies, including improvement and expansion of the existing AccessGrid (or equivalent) networks to include multiple nodes in all major environmental research universities and means of linking more widely dispersed individual investigators into access grid sessions.

NSF should ensure that this new program is coordinated with other relevant activities, including the National Aeronautics and Space Administration's Earth Science Enterprise; the Department of Energy's Office of Biological and Environmental Research , Office of Fossil Energy, and labs; the National Oceanic and Atmospheric Administration's Office of Oceanic and Atmospheric Research, National Environmental Satellite Data and Information Service, and labs; and the U.S. Geological Survey's Federal Geographic Data Committee and National Spatial Data Infrastructure.

NSF should work with its partner agencies to

- Stimulate cooperation between environmental and computer sciences, perhaps through designating funding to support projects that include PIs from both domains and that do not require simultaneous high-risk innovation in both arenas.
- Support and implement cooperative environmental cyberinfrastructure programs and projects in accordance with a set of principles established by the community (e.g., maintaining open access to data and integrating educational and research goals). A first-order draft of such principles is proposed in Table 2 of this document.
- Eliminate exclusive long-term data rights for PIs and require that all data collected and/or generated with federal funds be made openly available using cyberinfrastructure methods.

References

NSF Advisory Committee for Environmental Research and Education. 2003.
Complex Environmental Systems:
Synthesis for Earth, Life, and Society in the 21st Century.
A report summarizing a ten-year outlook in environmental
research and education.
Arlington, Va.: National Science Foundation. 68 pp.
http://www.nsf.gov/geo/ere/ereweb/acere_synthesis_rpt.cfm

NSF Blue Ribbon Advisory Panel on Cyberinfrastructure. 2003.
Revolutionizing Science and Engineering through Cyberinfrastructure.
Arlington, Va.: National Science Foundation.
http://www.cise.nsf.gov/evnt/reports/toc.htm

Ocean Information Technology Infrastructure Steering Committee. 2002.
An Information Technology Infrastructure Plan to Advance Ocean Sciences.
Washington, D.C.: National Oceanographic Partnership Program. 80 pp.
http://www.geo-prose.com/oiti/report.html

Appendix 1

Workshop Participants

Michael Aitken	UNC Chapel Hill
Guy Almes	Abilene/Internet2
Dan Atkins	University of Michigan
Chaitan Baru	UCSD/GEON
Andrew Bennett	Oregon State University
Terri Betancourt	NCAR
Rosina Bierbaum	University of Michigan
Ross Black	University of Kansas
Maurice Blackmon	NCAR
Lawrence Buja	NCAR
Karen Cambell	University of Minnesota
Greg Carmichael	University of Iowa
Bonnie Carroll	Information International, NBII
Kathy Carusone	MIT Lincoln Laboratory
Theresa Crooks	University of Kansas
Lois Delcambre	Oregon Graduate Institute
Cecelia DeLuca	NCAR
Rudy Dichtl	University of Colorado, NSIDC
Ben Domenico	UCAR
Mark Eakin	NOAA NCDC
Daphne G. Fautin	University of Kansas
Stuart Feldman	IBM
Mike Folk	UIUC/NCSA
Jim French	NSF/CISE
Catherine Gautier	UCSB
Steve Hankin	NOAA PMEL
Joseph Hardin	University of Michigan
James Harrington	NASA
Alan Hastings	UC/Davis
John Helly	SDSC
Chris Hill	MIT
David Holland	NYU
Alexandra Isern	NSF/GEO
Cliff Jacobs	NSF/GEO
Greg Jones	University of Utah
Anke Kamrath	SDSC
Mitsuhiro Kawase	University of Washington
Kevin Kelleher	NOAA NSSL
Randy Keller	GEON
Al Kellie	NCAR
Peter Knoop	University of Michigan
Venkat Lakshmi	University of South Carolina
Meredith Lane	GBIF
Margaret Leinen	NSF/GEO
Xu Liang	UC/Berkeley
David Maidment	University of Texas
Bill Matthaeus	University of Delaware
Scott Matthews	Cornell
Chris Maples	University of Indiana
Charles Marshall	Harvard
Peter McCartney	Arizona State University
Julie McClean	Naval Postgraduate School
Ken McDonald	NASA
Gail McConaughy	NASA
Deborah McGuinness	Stanford
Steve Meacham	NSF/GEO
Marla Meehl	NCAR
Don Middleton	NCAR
Barbara Minsker	University of Illinois
Jim Ogg	Purdue
Kim Olsen	UCSB
Peter O'Neil	NCAR
Kurtis Paterson	Michigan Tech University
William Piel	SUNY Buffalo
Mohan Ramamurthy	University of Illinois
Glen Rutledge	NOAA NCDC
Mark Schildhauer	UCSB
Paul Schopf	COLA
Dogan Seber	Cornell
Ashbindu Singh	UNEP GRID
David Skole	Michigan State University
Detlef Stammer	UCSD, SIO
Hubert Staudigel	Geochem
Steve Tanner	University of Alabama/Huntsville
Gregory van der Vink	IRIS
Frank Vernon	UCSD, SIO
David Vieglais	University of Kansas
Doug Walker	University of Kansas
Bruce Wardlaw	USGS
Patricia Waukau	NCAR
Mike Wright	NCAR
May Yuan	University of Oklahoma

Steering Committee

Lee Allison	Kansas Geological Survey
Peter Backlund	NCAR (Executive secretary)
James H. Beach	University of Kansas
Peter Cornillon	University of Rhode Island
Kelvin Droegemeier	University of Oklahoma
Peter Fox	NCAR (NCAR liaison)
Dale Haidvogel	Rutgers
Thomas Jordan	USC
Tim Killeen	NCAR
Jerald Lee Schnoor	University of Iowa

National Center for Atmospheric Research
P.O. Box 3000
Boulder CO USA
80307-3000
303-497-1000

www.ncar.ucar.edu

The National Center for Atmospheric Research is sponsored by the National Science Foundation. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the National Science Foundation.