

**Cyberinfrastructure  
for the Atmospheric Sciences  
in the 21<sup>st</sup> Century**

A report from the

Ad Hoc Committee for Cyberinfrastructure  
Research, Development and Education  
in the Atmospheric Sciences

June 2004

# Table of Contents

Executive Summary	1
1. Introduction	5
2. Background	6
2.1 Cyberinfrastructure – Ideas, Tools and People	
2.2 Scientific Challenges and Opportunities	
2.3 Educational Requirements	
2.4 Frustrations with Cyberinfrastructure	
3. Issues	13
<u>Thematic Perspectives</u>	
3.1 Social and Cultural Issues	
3.2 Data Issues	
3.3 Computing Infrastructure and Capacity Building	
3.4 Software Issues	
<u>Cross-Cutting Issues</u>	
3.5 Barriers to Entry	
3.6 Operational and Real-Time Needs	
3.7 Software Life Cycle	
3.8 Coordination	
4. Recommendations	33
References	40
Appendices	
Appendix A – CyRDAS Charge, Membership and Procedures	41
Appendix B – Software Projects	43
Appendix C – High-End Computing	46
Appendix D – Examples of Computational Issues	48
Appendix E – Community Models	51

# Executive Summary

The convergence of computation, information management, networking and intelligent sensing is poised to transform the conduct of science. This emerging **cyberinfrastructure** (CI) has the potential to provide powerful new tools, research methodologies, and processes that will enable scientists to investigate the atmosphere in new and previously unimaginable ways. To ensure rapid progress in the atmospheric sciences, it is essential that the technology that enables frontier research and effective education continues to be developed, deployed and maintained. Given the computing-intensive and data-intensive nature of the atmospheric sciences, an important component of the enabling technology is the CI that undergirds computation and data management and analysis. Development, deployment and evolution of the most effective CI must be driven by science priorities, and the science community must be engaged and remain involved in the planning process.

The planning process for how best to implement a CI, that serves the research and educational needs of the atmospheric sciences, can be viewed from four interrelated perspectives: human resources, data, computing infrastructure and capacity building, and software life cycle. Human resources has the highest importance among these perspectives in terms of the development and widespread use of CI within the atmospheric sciences, improving diversity, and improving the access to research and educational tools and capabilities by traditionally underserved populations.

This report describes the issues associated with each of the perspectives and recommendations for how to address these issues are listed. In addition to the specific issues and recommendations listed for each perspective, there are several issues that cut across these areas and stand out as problems that must be addressed in any CI initiative.

## Cross-Cutting Issues

**Barriers to entry:** There are real barriers to effective development and use of CI in the atmospheric sciences. Among the barriers are the following:

- Human resources – An essential and underemphasized element for a successful CI is the need for atmospheric sciences faculty and students, as well as computing professionals who are well prepared to take advantage of or contribute to emerging capabilities in the computing, data and software. A successful CI initiative will address this high priority need in order to facilitate both frontier research and education. Attention should be given to the demographics and issues of the next generation workforce, and strategies should be developed to insure that the atmospheric sciences are competitive in attracting the best and most promising talent.
- Lack of awareness – Too many researchers and educators in the atmospheric sciences are unaware of computing, data and software resources that are already available, either in the academic community or in the commercial marketplace. There is a need for better communication of, and dissemination to, the atmospheric sciences community of such resources and how they can be used.
- Uneven and inequitable access to or distribution of resources – The ownership of and access to computing, networking, data, and computing professional resources varies greatly from department to department. A response should be developed to provide better access to and a more equitable distribution of resources, based on NSF's mission, through intelligent and well-planned investments.
- Opaqueness – The CI that has been developed and put in place over the last several years has, at times, made computing and scientific use of data more difficult. This is a trend

that must be reversed, through hardening of middleware and application software, through increased transparency of the computing infrastructure, through development of new software tools, and through increased discoverability and interoperability of the diverse data sets that are becoming available, both within the atmospheric sciences and in related disciplines.

- Computing infrastructure – There continues to be a large gap between available high-end computing capabilities and the computing capabilities needed to address important problems in the atmospheric sciences, including studies of the Earth system. This large gap should be addressed in the context of the evolving computing infrastructure.

***Operational and real time needs:*** A unique aspect of atmospheric sciences is the close relationship among research, education, and operational meteorology – pollution alerts, weather prediction, climate prediction, space weather alerts, etc. The CI that is developed for atmospheric sciences should, at all stages of design and implementation, support this relationship, whenever practical. This has implications for the design of networks (e.g., high-speed connections between research laboratories and operational forecasting centers), data distribution (e.g., real-time meteorological observations in the classroom), computing infrastructure (e.g., on-demand computing), and management of and access to data sets and community models.

***Software life cycle:*** The high concentration of research resources on innovative development has led to neglect of funding for the full software life cycle, including hardening and deployment. Further, there is a critical need for funding for training, support, and maintenance of the underlying infrastructure. Without viable options for supporting software throughout its useful lifetime, investments in software infrastructure and tools will be lost.

***Coordination:*** There are already many diverse CI planning and implementation efforts within different divisions of the Geosciences, different directorates of the NSF and among the various federal agencies with sizable portfolios in research and development. These activities should be coordinated, whenever possible, to optimize the return on investment and to minimize the duplication of effort.

## Challenges in the Atmospheric Sciences

The atmospheric sciences research agenda is an ambitious one, whose individual elements require a significant enhancement of and substantial investment in the national computing and data infrastructure. The questions that have been articulated in several venues and reports include the following representative examples:

- Weather: What is the limit of predictability of the atmosphere and how does it change as a function of location, spatial scale or temporal scale? For example, to what extent can tornadoes be predicted and what will it take to make skillful predictions with lead times that can reduce the associated loss of human life and property?
- Climate: What are the global, regional and local effects of climate variability and change on the world's production of food and use of energy, on local ecosystems, and on the spread of disease? What will it take to build integrated Earth system models that include the treatment of atmospheric chemistry, dynamical vegetation and aquatic and terrestrial biogeochemical cycles deemed necessary to realize the predictability of the Earth's climate and make reliable estimates of regional climate change?

- Air Quality: What emission control strategies can effectively reduce the health and ecosystem impacts associated with air pollution? How can we optimally combine measurements with models to provide operational forecasts of air quality and chemical weather for use in emergency response and public health protection? How can we better quantify the role of aerosols in the atmosphere, and their specific connection to health effects and their role in cloud and climate systems?
- Space Physics and Space Weather: Why does solar activity vary regularly in 11-year, 22-year and other cycles? How does the geospace environment respond to solar variations? What is the probability that specific types of space weather phenomena will occur over periods from hours to days? Can these changes be forecast?

Addressing these questions, as well as other challenging scientific problems that cut across discipline boundaries within the atmospheric sciences and across the geosciences as a whole, will require substantial new research that will push the limits of the observing systems and the computing and data infrastructure that is currently in place. In particular, these questions will require

- massive ensembles of cloud-resolving and eddy-resolving weather, climate, and space physics models;
- advanced data assimilation systems capable of accurately analyzing immense volumes of observational data from radar and satellites;
- observing and analysis systems capable of delivering accurate and adequately-resolved relevant measurements from the Sun to the Earth's surface;
- data acquisition, management, integration, analysis, mining, and visualization tools to enable advances in scientific understanding and forecasting in a distributed environment, and to support on-demand and real-time activities, employing both small and very large data streams;
- configurable workflow orchestration and web services capable of coupling observations, modeling, and data handling, analysis, publishing and visualization, through grid-enabled facilities.

To provide these capabilities requires computing infrastructure, data systems, software and, importantly, human resources, that go far beyond what is currently available. Substantial investments in CI will be needed to develop, enhance and sustain the systems that can meet these challenges.

### General Recommendations

In order to address these pressing issues, it is recommended that the NSF change the way in which it invests in CI for the geosciences in the following ways:

- Give human resource issues top priority in CI development, hardening, maintenance and distribution. The academic reward structure should recognize CI-related activities, and NSF should seek additional ways to support investments in CI personnel at all levels, including departments, campuses and centers.
- Encourage the education of students, educators, support staff and scientists through exposure to the computational environment, including tools, data sets, and models.
- Provide the necessary support for communication and dissemination of ideas, technologies, and processes to promote the interdisciplinary understanding needed to successfully use CI in the atmospheric sciences.

- Help organize and coordinate CI activities across the geosciences and provide forums for innovation. The nascent activities of the Geoinformatics Steering Committee and the Geosciences Technology Forum are good first steps in this direction.
- Initiate a coordinated effort with NASA, NOAA and other agencies to provide geoscientists with the ways and means to find, effectively use, publish, and distribute geoscience data, including the development of appropriate standards for metadata to which grant recipients should be encouraged to conform.
- Provide funding for the entire CI software life cycle, including development, testing, hardening, deployment, training, support and maintenance, subject to ongoing review and demonstrated utility and sustainability.
- Invest in computing infrastructure and capacity building at all levels, including centers, campuses, and departments, in support of atmospheric science research and education.
- Support the development of geosciences cyber-environments that allow the seamless transport of work from the desktop to supercomputers to Grid-based computing platforms.

### Expected Outcomes

If recommendations for large investments by the NSF in CI are adopted, we foresee that the Nation in general, and the atmospheric sciences in particular, will gain both scientifically and more broadly in several ways. The potential outcomes include:

- A diverse, informed, educated and well-connected work force that is better able to meet the emerging complex scientific challenges, which inherently lie at the boundaries among disciplines, to determine and quantify the impacts of natural variations in the environment on society, economic interests and ecosystems;
- A Nation that is more resilient to disruption by intentional or unintentional influences and better equipped to assist other nations through accelerated progress in forecasting to reduce the loss of life and property and the adverse effects on ecosystems associated with severe weather, hurricanes, toxic or lethal pollutant dispersal, floods, droughts, solar flares and other space weather events, and regional climate change;
- A Nation that maintains its international competitiveness in the atmospheric sciences;
- A more effective and efficient environment for accelerating scientific discovery, in which researchers spend more time conducting their scientific investigations and less time addressing information technology issues; and
- Better understanding of the Earth as a complex environmental system, through increased use of atmospheric data and improved communication and information sharing among researchers and educators.

# 1. INTRODUCTION

As described in *NSF Geosciences Beyond 2000: Understanding and Predicting Earth's Environment and Habitability* [1], a strategy for accelerating progress in research and education in the geosciences will require "...a commitment to improve and extend facilities to collect and analyze data on local, regional, and global spatial scales and appropriate temporal scales," including real-time observing systems, and modern computational facilities to support rapid computation, massive data archiving, distribution, analysis and management.

Recognizing that this acceleration of progress requires a substantial change in the way that infrastructure supporting research and education is viewed, in January 2003, the National Science Foundation (NSF) Division of Atmospheric Sciences formed an ad hoc committee for Cyberinfrastructure Research and Development in the Atmospheric Sciences (CyRDAS) to promote discussion within the community and develop a strategic planning process for the development of cyberinfrastructure (CI) for the atmospheric sciences. The committee was charged with determining the needs of the atmospheric science research community and assessing the opportunities for advances in atmospheric science research that (1) are made possible by current or anticipated advances in information technology and computer science, and (2) might result from collaborative research between atmospheric scientists and computer scientists. The committee was also charged with determining the ways in which CI can contribute to formal and informal education in atmospheric science. The full charge to the committee, the membership of the committee and the procedures and activities of the committee are described in Appendix A.

*This report recommends strategies that will help the academic community to exploit the opportunities afforded by CI for advancing atmospheric science research and education.*

This assessment and planning exercise is not being undertaken uniquely in the atmospheric sciences. In February 2003, the report of the NSF Blue Ribbon Advisory Panel on CI [2] recommended that NSF lead an interagency advanced CI initiative to "create, deploy and apply CI in ways that radically empower all scientific and engineering research and allied education." There are planning efforts underway in each of the divisions of the NSF Directorate for Geosciences, including atmospheric sciences, oceanic sciences and earth sciences. In addition, there is an overarching environmental sciences CI planning activity, and the various assessment and planning groups are closely coordinating these activities. The issues relevant to the atmospheric sciences community, and the strategies described in this report for developing a distributed CI that will meet the needs of the academic atmospheric science research and education community and which includes the flexibility to grow smoothly as atmospheric science advances and CI needs grow, will be contributed to the evolving NSF Geosciences strategic planning.

The report is organized as follows. Chapter 2 provides background on the technical and institutional framework in which planning for CI in the atmospheric sciences is evolving. The issues of relevance to the atmospheric sciences, both thematic and cross-cutting, are described in Chapter 3. The CyRDAS committee's recommendations for both the short term (next 2-5 years) and the long term (5-10 years) are provided in Chapter 4.

## 2. BACKGROUND

The technical and institutional framework in which planning for the development of the cyberinfrastructure (CI) suitable for research, development and education in the atmospheric sciences is a complex one in which government, academia and the private sector all play roles.

There is a rich set of computer science and computational science methodologies and tools already in place in the atmospheric sciences, including high-end computing, data analysis and management and software tools for modeling and analyzing the Earth environment. In parallel, great progress has been made in the computer sciences to advance ways and means of tackling the computational, programming and data management problems encountered across the technical disciplines. To some extent, these parallel paths have been interconnected through the communication of requirements from the domain sciences to the computer sciences in targeted collaborative projects and through programs intended to apply recent computer science advances in the domain sciences. A fundamental aspect of these cross-disciplinary issues and programs is the need to communicate, inform, educate and train students, faculty and research scientists in both computer sciences and domain sciences.

The atmospheric sciences have undergone dramatic changes in the past 50 years – numerical weather prediction with high-speed computers, satellite-based sensors providing global coverage at high resolution, centralized and collaborative distributed facilities for observing and modeling the atmosphere, near-Earth environment and the solar-terrestrial connection. The field is continuing to evolve with the expanding prospects for data assimilation, climate modeling and prediction and real-time observation and prediction of space weather events. The evolutionary and revolutionary changes in the atmospheric sciences have required an ever-growing use of and dependence on computers, networks and software. There are serious challenges to maintaining that progress.

In this chapter, a summary of the framework in which planning for CI for the atmospheric sciences is embedded is provided. While not intended to be comprehensive, it describes a broad range of activities, programs and challenges for CI planning.

### 2.1 Cyberinfrastructure – Ideas, Tools and People

As stated in the National Science Foundation strategic plan, the Nation's science and engineering research agenda is focused on ideas, tools and people.

*Ideas:* An essential element of any scientific or engineering enterprise is the ability to discover and capitalize on new ideas. The trend in both science and engineering is an increasing proportion of the research that is conducted at the seams between scientific disciplines. Also, while innovation has always been the main ingredient of scientific research, there is a trend toward tighter integration of research with education, not only at the post-graduate level, but also in the undergraduate and K-12 curricula, and a tighter linkage of research goals to the needs of society.

In the atmospheric sciences, innovations in several areas are putting stress on existing research, education and service strategies. There has been a rapid improvement in and adoption of computer-based models for many aspects of the atmospheric sciences from numerical weather prediction to modeling the Earth-Sun relationship. The need to relate models to measurements has led to a dramatic improvement in the methods of analyzing data, both from observing systems and from sophisticated models. The deployment of space-based observing systems, high-resolution radars, autonomous observing systems and the steady upward trend in the spatio-temporal resolution of numerical models has motivated new methodologies in managing and manipulating previously unimaginable data volumes.

In parallel with these developments in the atmospheric sciences, there have been huge advances in computer sciences and technology. Now almost taken for granted as a natural law, the advances over the past three decades in chip design, large-scale integration of logic circuitry and chip fabrication have supported exponential growth in computer capacity and speed. The introduction of fourth generation programming languages and structured programming techniques has made it possible to manage this vast increase in power. To some extent, many of the activities previously handled by interaction between humans and computers have been automated using intelligent systems in, for example, data and metadata management.

*Tools:*

It is widely recognized that broadly accessible, state-of-the-art tools are necessary to advance scientific and engineering research and education. To be effective, such tools must be easily shared among researchers with a diversity of backgrounds and capabilities. There is a spectrum of research and education tools from the most basic operating system instruction sets to the most sophisticated domain-specific application programs. The low-level tools have facilitated efficient data file management and data movement, effective use of large numbers of processors all working on the same application program, and protection from security, maintenance and upgrade disruptions. At the higher level, research tools have been developed both by the computer science community and by the domain sciences to enable multi-model ensembles for numerical weather forecasting and climate prediction, analysis and visualization of very large data sets that are distributed over the wide area networks, and the objective combination of models and measurements of the atmosphere, upper ocean and Earth's space environment.

*People:* One of the fundamental requirements set forth in the National Science Foundation strategic plan is that the Nation requires a diverse, internationally competitive and globally-engaged workforce of scientists, engineers and well-prepared citizens in order to make progress, address societal issues and maintain its preeminence in the world. In the latter half of the twentieth century, it became clear that competitiveness and preparedness are closely linked with effective use of computers and computational technology.

The ability to make effective use of computing resources, data resources, communication and collaboration tools and other elements of CI likewise depends on a cadre of highly-trained and capable people. Outside the atmospheric sciences, such people include system administrators, network administrators, webmasters, facility operators and others who provide services and expertise to maintain and operate the various elements of CI. There are also computer science experts – innovators, designers and programmers who work toward upgrading and improving existing CI and the development of new CI to improve services or introduce new capabilities. Within the atmospheric sciences, there are several levels of CI activity, including developers, early adopters, mainstream users and those without specific atmospheric science CI requirements who use commodity systems such as office tools and Internet browsers.

## **2.2 Scientific Challenges and Opportunities**

Cyberinfrastructure (CI) occupies an essential role in Atmospheric Sciences research, and the internet, e-mail, the WWW, supercomputers, and powerful desktop computers evidence the astounding impact that CI has had in the past two decades. Atmospheric Sciences have benefited accordingly, in the broad sense because of the way that science is carried out, but even more emphatically because of the reliance of this discipline upon extensive modeling and analysis of large datasets. The *Blue Ribbon Advisory Panel on Cyberinfrastructure* asked

## How will science and engineering research be changed through implementation of an *Advanced Cyberinfrastructure Program*?

In coming decades we envision further revolutionary developments in the tools available to the broad research efforts supported by NSF Atmospheric Sciences programs. These will be based upon use of the emerging advances in information science and technology (discussed in Section 2) that will lead to new ways to produce, organize, analyze and disseminate scientific knowledge.

There is a great deal of common ground in the scientific challenges of the 21<sup>st</sup> century for the various disciplines. Atmospheric Sciences challenges are related<sup>1</sup> to those that face Ocean Science, Space Weather, Physics, Astronomy, and Engineering. Similar fundamental scientific themes are encountered, such as cross scale coupling in dynamical models, and the unifying idea that nonlinear dynamics can produce complexity in observed behavior. Developments in CI will be essential in addressing these problems.<sup>2</sup>

### 2.2.1 Questions Arising in Computational Modeling and Data Analysis.

A continuing challenge in Atmospheric Sciences is the direct numerical simulation of a three dimensional (3D) fluid model, such as turbulence on an ordinary fluid. In many ways this is a standard problem in terms of required computing power, storage, network bandwidth, visualization and other CI resources. Similar demands emerge in climate, weather and severe weather modeling, 3D magnetohydrodynamics, solar and Geospace modeling, and various multi-component flow models, in which both methodological and physical similarity give rise to commonality in computational scaling. Consequently the projected CI requirements for both production and grand challenge problems are comparable across many Atmospheric Sciences subdisciplines. The goal of complete “surface to stratosphere” modeling in climate is analogous to “core to crust.” modeling in geophysics, or the Space Weather goal of “Sun-to-mud” modeling of Geospace couplings. In each case, one can ask the questions:

- What are the requirements in the next decade for implementing more accurate, and better spatially resolved (“larger”) problems, and problems with more realistic physical parameters?
- What are the requirements for longer duration models, and ensembles of model calculations needed to carry out key statistical studies?
- What are the requirements for predictive computing model results in real time or near real time?
- What kinds of demands will there be for computing facilities, including large shared memory multiprocessor supercomputers, local Beowulf clusters, systems based upon high performance networks, and distributed grid computing capabilities?

---

<sup>1</sup> The *Blue Ribbon Advisory Panel's* distinction between disciplines only partially corresponds to NSF organization. Space Weather is included in the Atmospheric Sciences Division of GEO. In the NSF, Solar and Space Plasma physics are part of the upper atmosphere, and astrophysics and astronomy are associated with Physics and Mathematics. The Sun resides in the Earth's upper atmosphere; the Earth is a part of the Sun's extended atmosphere. There are many research interests that cross traditional boundaries. Atmospheric Sciences conceptually borders on and overlaps with solid earth geophysics, plasma physics, fluid dynamics, chemistry, nonlinear dynamics, solar and cosmic ray physics and astrophysics.

<sup>2</sup> The *Panel on Theory Modeling and Data Exploration*, in their Report in support of the Solar and Space Physics Decadal Survey Committee commissioned by the National Research Council, emphasizes that computational (CI) issues would provide important capabilities in the coming decade, but that “it is essential that the focus remain on basic science and the computational tools needed to advance this science.” (see <http://books.nap.edu/catalog/10860.html>)

Some answers vary in their particulars. In meteorological forecasting or space weather modeling, near real time response is essential. But turbulence theory computations that strain the largest computers lack time criticality.

Complex research efforts require linking together modules with different physical content, requiring that the components interact in a seamless way. This requirement for modular “interoperability” raises additional questions such as:

- What additional CPU, memory, network bandwidth, storage, and software requirements are associated with implementation of interoperable systems of codes that describe complex couplings that occur in Atmospheric Sciences?
- How can Computer Science assist the physical scientist in producing seamless connections between different models, having different variables, different grids, different timescales, and possibly even residing at different locations?

CI in the next decade will need to support the enormous data archiving and retrieval requirements that go hand-in-hand with the ability to carry out these state of the art numerical modeling efforts. New techniques will need to be developed and implemented for dealing with vast numerical databases, both from numerical modeling and from new generation efforts in multipoint observations. More broadly there will be opportunities for taking advantage of virtual observatories, distributed databases and archives, and new network-based approaches to collaboration.

- What are the requirements for the subdisciplines of Atmospheric Sciences for data archiving and dataset organization and access for vast datasets that will be accumulated using multi-point measurements and multi-station observational datasets. (weather stations, spacecraft, etc).
- Can developing methods for linear and nonlinear data assimilation lead to improved fundamental physics models and improved predictions?

### 2.2.2 Scaling of Requirements for Computational Resources

A typical model computation in Atmospheric Sciences involves a set of  $M$  scalar variables defined on a three dimensional grid<sup>3</sup> with a spatial resolution of  $N^3$  points. Usually in fluid flow problems there is some scale at which the systems is stirred, or equivalently a scale that contains the most energetic motions. Suppose this scale is  $L$ , occupying a number of grid points  $J = L/dx$  relative to the grid resolution  $dx$ . For simplicity we assume here that  $J \sim N$ , and that the energy containing structures are comparable to the box size. Let the characteristic speed  $U$  in the problem be that of the motion of the energy containing eddies of size  $L$ , and all other speeds are measured in these units. This sets the main “clock” for this dynamical system.

The fastest time scale may be estimated as the sweeping of the smallest resolved structures by the energy-containing eddies. This sets the time stepping scale to be of order  $dt = 1/N$ , implying that  $N$  time iterations are needed to compute evolution of a single characteristic time scale. For each iteration, we estimate that a high order finite difference or pseudospectral calculation will require on the order of  $N^3 \log N$  arithmetic operations. Assembling the results so far we can see that computational work<sup>4</sup> for a one characteristic time  $T = L/U$  is  $N^4 \log N$  for

---

<sup>3</sup> A symmetric representation is assumed for convenience. Often it is possible to decrease the resolution in one or two directions. An example is a global climate model which typically permits tenfold less grid point (or levels) in the vertical direction.

<sup>4</sup> We leave out various multiplicative and additive constants in this estimate, such as number of distinct nonlinear calculations to be done at each step for each equation. Therefore the estimate gives scaling for the relative work, which is all we need to compute ratios of work for different problem sizes

each variable. Predictive forecasting computations may proceed for just a few times  $T$ , with parameters chosen for maximum possible accuracy. We should think of this amount of work, say  $W_N$ , as a useful unit of computational work in Atmospheric Sciences calculations of a given type. For longer term studies of statistical ensemble properties (such as in climate modeling or turbulence theory calculations) the duration of the calculation may be  $1000 W_N$ .

At present the largest routinely attainable hydrodynamics turbulence computations have approximately  $N=1024$ . This is largely due to running times (approximately 2 weeks on a Beowulf cluster with high performance interconnect) and large memory requirements. The nominal requirements for such a simulation, as well as the scaling with problem size, are illustrated in Table 1, in which the relative work is scaled to an  $N=1024$  turbulence calculation. Since the latter is presently on the verge of feasibility, it is clear that the requirements for each of these problems is extremely daunting when all scales are represented. Typically what is done to account for dynamical scales that cannot be resolved is some form of parameterization. For unresolved turbulence, this may take the form of some type of subgrid scale modeling. For climate modeling it may involve parameterization of cloud dynamics or of the interaction of the atmospheric flow with ocean surface eddies. Models improve, sometimes dramatically, when subgrid parameterizations are replaced by a more refined direct model. In Space Weather modeling the unresolved scales include both turbulence and kinetic plasma effects, and, inspecting the estimates in Table 1, we see again that vast refinements are needed before attaining resolution of all dynamically important scales.

	<b>System Size <math>L</math></b>	<b>Outer Scale <math>l</math></b>	<b>Smallest Resolved Scale (current)</b>	<b>Inner Scale <math>l_{di}</math></b>	<b>Degrees of Freedom</b>	<b>Computational Work</b>
<b>Weather</b>	10,000 km	5000km	20km	<1 km	$10^{15}$	$10^8$
<b>Climate</b>	30,000 km	5000km	100km	<1 km	$3 \times 10^{16}$	$10^{10}$
<b>Magnetosphere</b>	600,000 km	60,000 km	3000 km	200 km	$3 \times 10^{10}$	100
<b>Solar Wind (weather)</b>	$1.5 \times 10^8$ km	$2 \times 10^6$ km	$2 \times 10^6$ km	400 km	$10^{17}$	$5 \times 10^{10}$
<b>Global Heliosphere (climate)</b>	$2 \times 10^{10}$	$2 \times 10^6$ km	$2 \times 10^8$ km	400 km	$10^{22}$	$5 \times 10^{14}$
<b>Solar Corona</b>	700,000 km	30,000 km	70,000 km	10 m	$10^{15}$	$10^8$
<b>Earth's Dynamo</b>	5000 km	2000 km	10 km	2 km	$10^{10}$	100
<b>Homogeneous Turbulence</b>	10 L	L	$L/100 - L/1000$	$L/R^{3/4}$	$1000 R^{9/4}$	$\sim(R/1000)^3$
<b>“1000-Cubed” Simulation</b>	10	1	0.01	0.01	$10^9$	1

Table 1. Some scale sizes and computational work estimates relevant to atmospheric and space modeling and research areas. Memory and storage requirements scale with number of degrees of freedom. Computational work estimate is given in the text. The full number of degrees of freedom in each system far exceeds current CI capacity. Computational work at full resolution is given in the last column, relative to the “1000-cubed” simulation of turbulence. Order unity factors are ignored for simplicity.  $R$  is the Reynolds number. Models improve greatly as they include more of the relevant physics and more of the dynamically relevant scales.

### Challenges for the Coming Decade

During the next decade, advances will be made in closing the gaps between currently achievable modeling and full realization of all relevant physical processes and scales.

Some of the specific challenges in the next decade will be:

- Understanding natural and anthropogenically-influenced climate variability
- Developing understanding of regional impacts of global climate fluctuations
- Understanding the interaction between climate dynamics, ecosystem dynamics, and biologically active constituents such as carbon, nitrogen and water.
- Developing a better understanding of the causes, distribution and intensity of extreme weather events.
- Understanding the causal chain that begins with solar eruptions and coronal mass ejections, ultimately impacting human and technological assets in the Geospace environment.
- To understand the long term influence of solar activity, and the flow of energy and particles in the heliosphere, upon the climate on and near Earth

In addition to gaining a fundamental understanding of the Earth's climate system, research is necessary to develop the capacity to predict:

- Extreme climate fluctuations and abrupt climate change
- The response of climate and ecosystems to different scenarios for anthropogenic climate forcing over the coming decades
- Space Weather effects on orbiting and ground level assets, and on safety of human space flight.

### **2.2 Educational Requirements**

Cyberinfrastructure will undoubtedly bring new opportunities and challenges with respect to education as the CI grows to become ever more pervasive, accessible, and robust. We can view the impacts in two dimensions; the impact of CI on the education process and outcomes, and the education requirements of providing the future scientists, educators and engineers who work in the atmospheric sciences with the computer technology foundations they need to work with the emerging CI.

The various technologies of the Internet (e.g. email, World Wide Web, Chat) have gained ground as tools used in all levels of our education community. There are still issues in terms of the level of access to networks and tools in the various sectors of education from K-12 through university, and between institutions within a sector. However, we have seen these technologies become not only mechanisms for interaction between instructors or colleagues, but also in the classroom between teacher and students and between professor and student. These technologies are becoming integral parts of the learning environment in both the instruction and testing stages of education. We are also seeing changes in the type of material served through such technologies. In the early period of the Web, information was often static sets of text, graphics etc., if you will, textbooks transferred to the new Web medium.

Today, we are beginning to see a great deal more interactive engagement of resources within learning contexts across the breadth of the education realm, e.g. access to real-time meteorology data for use in undergraduate atmospheric programs, or in the K-12 math classroom in support of understanding math principles such as statistics and graphing. The future ideas that the CI will develop will add to the mix as we see more pervasive access to digital libraries (some specifically supporting education, others the broader archive of the intellectual output of the disciplines), access to broader datasets themed around earth system events. We may see a future where the education and research agendas are highly intertwined where scientists have periods of

intense learning along the boundaries of their area of research, and where educators and students can use authentic research problems and data in their learning environment.

Key to the success of these possibilities will be whether the participants in the new CI-enabled atmospheric sciences have the necessary knowledge and training in CI technologies to take full advantage of new possibilities, and the knowledge and training where necessary to allow them to develop new atmospheric-science related CI technologies, e.g. engineers and programmers who can develop the implementations of models on the computational hardware and operating systems. This will require the cooperation between the atmospheric sciences and the computer sciences in understanding each others domains, directions, and driving forces in order to develop the technical curricula needed by our future scientists and educators. It is interesting to note that the K-12 schools are now sending on students who are very comfortable with computer technology, the Internet and the web, yet not enough of these same computer-savvy students are interested in computational science to push computer science departments to offer such courses.

Because the CI will evolve with the ever-changing technology arena, we must recognize that ongoing training of the scientific and education members of the atmospheric science community will be important to success. As new computer science approaches and paradigms evolve, they will need to be inculcated into the broader scientific community in order to enable the science to be performed at the best possible level in utilizing CI innovations.

### **2.3 Frustrations with Cyberinfrastructure: Barriers to Progress in Atmospheric Sciences**

Advances in workstation and networking technology over the past decade brought unprecedented computing power and information to the desktop. This new power produced major advances in modeling phenomena as well as in the analysis and visualization of data. Unfortunately, advances at the high end have not been as dramatic. The primary unmet needs of the atmospheric science community are twofold. State-of-the science modeling, simulation and prediction require exponentially increasing high-end computing resources for problems ranging from basic research in space physics, weather and climate to real-time forecasting to protect lives and property. New surface and spaced-based observational platforms are now providing a wealth of information with unprecedented spatial and temporal resolution for many atmospheric applications, including weather and climate. In the future, the volume of these data streams will increase as new instruments are deployed. The availability of high-end data storage technology that can keep pace to archive these observations and model-generated output, as well as rapidly retrieve them on demand, is essential.

With the Information Technology (IT) explosion of the late 1980s and early 1990s, the focus of computing technology quickly shifted away from science and engineering applications. Hardware manufacturers found new and lucrative markets in the expanding field of business enterprise computing enabled by the dissemination of local network technology. Meanwhile, computer science researchers began to emphasize the seminal development of high-risk, high-payoff technologies, needed to sustain the “IT Revolution.” The types of systems needed by the atmospheric science community were quickly relegated to much lower priorities by market forces.

The complexity and diversity of available technologies posed new problems for the community, as it struggled to assemble the systems it required from components typically designed for other applications. Moreover, the software infrastructure to make these components work together as an integrated system was either extremely primitive, or non-existent. Adaptation to this new environment was inhibited because of the shortage of qualified people who understood both the new technology and the needs of the science. Although these problems are not as severe today as they were several years ago, they still persist and require long-term solutions. They are explored in the following sections in greater detail.

### 3. ISSUES

The general question of how the academic atmospheric sciences research and education community can best exploit opportunities that are made possible by advances in information technology and computer science is complicated by the fact that several dimensions of the problem are on exponential growth curves. The capacity and speed of computer hardware is doubling every 18-24 months. The number processors routinely available in a single computing platform is doubling every year or two, and the number of processors routinely used by a cutting-edge modeling application is increasing similarly. The volumes of data that are collected from measurements, generated by models, stored and analyzed are growing at various exponential rates, with the net growth stressing the Nation's data archival and management facilities. Software is becoming more complex, growing quasi-linearly in the number of functions it performs and possibly exponentially in the number of lines of code. Communication switches, local area networks and wide area networks, particularly those available for special-purpose scientific and engineering problems, are increasing their speed at various rates of growth, although networking technology has advanced more slowly than other aspects of CI. This mismatch in paces of improvement represents a potential bottleneck in harnessing the combined computing power of many processors. Planning a CI that addresses the diversity of requirements across the atmospheric sciences, let alone the geosciences and beyond, becomes all the more difficult in an environment dominated by rapid change.

Further complicating the question is the fact that computers, networks, software and data are used by people. This introduces a wide variety of factors – access, preparedness, career path, cooperation vs. turf struggles, jargon normalization, etc. – that serve to make a complex problem more difficult. To begin to organize the problem, the following subsection describes the relevant issues in terms of four perspectives that touch upon the themes mentioned above, namely, social and cultural issues, data issues, computing infrastructure and capacity-building and software issues. Each thematic perspective is broken down into several components to help define the scope of the problem. For many of the components, a series of questions are posed that further develop aspects of the issue.

In the course of examining this question, it was found that there are several issues that have elements that touch upon several or all of the thematic perspectives. These are described separately in a subsection on cross-cutting issues.

#### **Thematic Perspectives**

##### **3.1 Social and Cultural Issues**

###### **3.1.1 Human Resources**

People are a critical aspect of CI. Infrastructure such as powerful computers, fast networking, sophisticated programming, and extensive databases offer little value without people who know how to implement, maintain, and use them. Indeed, one can imagine situations in which productive science is not limited by insufficient technology, but rather by the inability to implement existing technology. Students must be exposed to the relevant careers at an early age, and these career paths must be seen to have long-term growth potential. Computationally-oriented career paths in the sciences must be developed and supported primarily by the scientific community itself in cooperation with its government sponsors such as NSF.

It is useful to partition human resource issues into three broad areas: personnel, partnerships, and outreach. The personnel area addresses the need to attract, train, reward, and fund people to develop and utilize CI. Partnership issues involve how to best bring together people from relevant departments and sectors, how best to keep them together, and how CI may

contribute to enhanced collaboration among atmospheric scientists. Outreach addresses using CI to encourage pre-college students to consider careers in science and engineering and to attract undergraduate and graduate students to the atmospheric sciences and to relevant areas of computer science.

### Personnel

Atmospheric scientists should not all be expected to be computer science and programming experts. It is reasonable to expect significant productivity gains by enabling atmospheric scientists to easily fund needed computer science professionals. Nevertheless, there also remains a need to train atmospheric science students to use both existing and new CI tools.

A well defined career path and reward structure for scientific computing is required both for support staff and primary researchers. Today's computer science graduates are usually ill-prepared for scientific computing, but well-prepared for more remunerative jobs in the private sector. Academic salaries cannot be comparable to those in industry, but the advantages of non-commercial careers must be emphasized. Mid-career professionals who have come to see the disadvantages of the commercial world may find the stability and relatively less-harried pace of academia attractive, and can be an important source of expertise. In addition, young faculty members spend more and more time on computer science-related issues. The community needs to realize that this is becoming an important aspect of atmospheric science, and allowances must be made in the reward structure to attract, retain, and promote these faculty.

#### QUESTIONS

- Is CI support best facilitated by a departmental pool of computer scientists/programmers, by providing individual PIs with the funding to support their own computer scientist/programmer, or something else?
- Should funding agencies earmark funds for computer scientists/programmers using existing grant structures, or provide a separate source of computer scientist/programmer funding?
- What are the CI-related atmospheric science work force needs of the nation? Will different sectors (academia, government, private sector) have different needs?
- Should we promote a training infrastructure to provide interested computer scientists and programmers with the necessary background and tools?
- How can we ensure that students pursuing careers in atmospheric sciences know how to use the necessary hardware and software tools?
- How can the atmospheric community best attract appropriately qualified people? Do we take an "if we promote CI they will come" approach, or should we be more proactive?
- What changes, if any, should be made to the academic reward structure (tenure) to appropriately recognize young faculty members who are actively involved in scientific computing?

### Partnerships:

Partnerships can range from informal collaborations to high-profile, multi-national relationships. Regardless of the level of partnership, the issues remain the same: partnerships must be formed, facilitated, and maintained. Few scientists would turn down the opportunity to form a partnership if it provides them with funding, but that does not mean that true collaborative research will be carried out. Computer science advances increase the need for collaboration, for example, by enabling the solution of complex inverse problems through the use of sophisticated first-principle models together data from many of instruments.

#### QUESTIONS

- How best can computer scientists and atmospheric scientists be encouraged to work together, rather than just providing one another with tools (CS) and problems (ATM)?
- Partnerships can exist on many levels: between individual scientists, between departments, between institutions, and between nations. What aspects of formation, facilitation, and maintenance are common across levels, and what aspects are level specific?
- Providing funding for specific collaborative efforts can be more effective than providing funding for general collaborative efforts. But the general efforts, if successful, are more likely to provide long term benefits. How can general collaborations be formed, facilitated, and maintained (if at all)?
- There have been significant attempts at building collaborative environments for scientists. A well-know example, often cited as a success story, is SPARC (Space Science and Aeronomy Research Collaboratory). Yet few would say that SPARC has lived up to its potential. What lessons can be learned from this and other similar experiences?

### Outreach

As discussed in a new report by the National Science Board ([www.nsf.gov/nsb](http://www.nsf.gov/nsb)), the federal government must take action to guarantee that a sufficient number of students choose careers in science and engineering. Action must take place on many fronts, and CI may be able to play an important supporting role. An example of particular relevance to the atmospheric sciences is the educational utilization of dense, networked arrays of simple instruments such as weather stations and GPS satellite receivers. A networked GPS receiver at every at every middle school, high school and junior college would enable students and their teachers to participate in real scientific research.

#### QUESTIONS

- How can CI be best leveraged to attract undergraduates and graduates to the atmospheric sciences?
- How can CI contribute to guiding pre-college students toward careers in the atmospheric sciences?

### **3.1.2 Education**

As the CI becomes more pervasive and accessible, it will bring new opportunities and challenges to research and education practice. The CI should enable broader participation in research and education by offering technologies that overcome challenges of isolation through

institutional size, geographic distance, economic distance, and socio-cultural boundaries. Overcoming these challenges will be important in developing broader access to science for minorities and disadvantaged economic groups. A challenge is to facilitate individuals progressing all the way through to graduate study, in line with NSF goals to improve diversity in the scientific community.

Utilization of CI in the atmospheric sciences will depend upon the preparedness of researchers, educators and learners. CI also needs to capitalize on the benefits of reciprocal influence, i.e. that educational practices enabled by the CI will evolve, producing impacts that can, in turn, positively influence the direction of future CI development both in the Atmospheric Sciences, and in collaboration with other disciplines such as Computer Science. These requirements of preparedness and reciprocal influence were two common themes of the discussions and focus groups that occurred in preparation of this report.

The following sub-sections provide comments on four areas considered key to providing an effective CI-enabled education environment for the Atmospheric Sciences. The first covers the needs for communication to increase awareness of CI in the domain. The second looks at the requirements for preparing researchers and educators to effectively utilize CI. The third discusses the need to bring together domain experts and CI technology experts. The fourth looks at the need to provide CI services and data into a broad range of learning environments.

#### Communication: Awareness of CI

The need to be aware of the capabilities and potential use of CI technologies was a common thread in discussions and focus groups. Solutions included the idea of supporting various information outlets from newsletters, electronic magazines and journals (with a range of peer-review mechanisms), catalogs or registries of tools and services, and examples of implementation and use. Such outlets can be customized to particular communities such as K12, community college, or undergraduate education in addition to those aimed for the researchers. Such communication channels would provide an infrastructure for the timely distribution of CI-related information, and could be developed and supported through a selection of organizational mechanisms such as professional societies (e.g. AMS or AGU), or through digital collections and libraries. Another avenue is to include CI activities in permanent sessions at domain meetings such as AMS and AGU, and for atmospheric science-based CI approaches and solutions to be included in the relevant meetings and conferences of the computer sciences (e.g. ACM special interest groups and IEEE technical groups).

A second important mechanism for promoting awareness of CI was articulated as the ability of scientists, educators and CI technologists (such as computer scientists) to have opportunities for closer collaboration and cooperation. This can be achieved through sabbaticals for domain scientists in technology groups, and for technologists into domain science groups. The idea of promoting more collaboration and cooperation underpins the concept of reciprocal influence. Other approaches to this could be development of short workshops on particular CI topics for scientists and engineers, along with building in some CI related topics into education programs, e.g. including software engineering/programming components in an atmospheric degree program, or enhancing availability of computational science programs. It is also important that the atmospheric science research and education community articulates challenges and problems within the domain for which computational and network solutions would be of great benefit.

#### Preparing the Workforce

The need for preparedness of the domain's researchers, educators and learners to understand and incorporate the CI in their activities is particularly important and requires the preparation of future scientists coming up through the education system (K12, community college, undergraduate Earth sciences and computer sciences). The need for better preparation in

the area of computational science, programming languages, problem solving, and data analysis were highlighted in many of the preparatory meetings for this report. Such preparation will require more collaboration with computer science and engineering programs to be able to identify and put in place the courses required, and to evaluate programs to confirm that they do indeed provide the requisite skills.

We need more computer scientists and engineers to become partners in the domain, working in collaboration with the domain scientists to develop the required computational infrastructure, services and tools. In the same way we would like to see more computational science in the atmospheric science programs, we will need to provide interesting problems and challenges to the computer science and engineering community in order to encourage their greater participation in atmospheric science. This could be achieved through more degree programs, including technical programs with geosciences emphasis, that bring together domain science, computer science and engineering, and technical management (needed to manage future operational infrastructure). The availability of a geosciences computing environment that will provide access to research-quality datasets would be a major benefit in providing a means for computer scientists and engineers to work on real science problems.

Preparation will also be required at the pre-college level. The strength of preparation in math and science of K12 students is one of the critical areas to their success in any science program. There is also the need to develop an interest in science in the early years and to engage students in understanding more about their environment. Providing access to tools and data in the K12 environment, can help to develop an interest in science. However, there are significant hurdles to overcome in terms of available computational and network resources, and in view of the many disparities among socio-economic groups and schools. Further, success will also depend upon the level of training, or preparation, teachers receive either in their formal training, or in later professional development. The atmospheric sciences will need to be engaged in supporting these efforts through courses and workshops organized around utilization of atmospheric science related materials, services, tools and data provided through the CI. Partnering with teachers and education technology specialists will help in determining how teachers will meet the educational goals and standards required of the K12 environment while utilizing the CI services, tools and data. The ability of the K12 community to develop educational resources that build upon CI, and their ability to share such resources through avenues such as digital libraries will help to bring atmospheric science into their classrooms, and will nurture interest and involvement in the science.

#### Collaboration: Domain Experts and CI Experts Together

Previous sections have highlighted the need for collaboration and cooperation between the science domain experts and those from the computer science and engineering areas who are involved in developing the technologies that underpin the CI. The need for successful partnerships between domain experts and CI experts was a common thread in many of the report's preparatory discussions. Some of the activities toward this end would be to allow for more sabbaticals that have CI components for the domain expert and domain components for the CI experts. The academic reward system would need to recognize such cross-disciplinary work and contributions to CI at both the institutional level and domain level (e.g. through recognition at AMS or AGU), and by recognizing other works that are of intellectual merit (e.g. "published" code libraries, technical implementations and usage). As partnerships develop, then the domain experts will be able to contribute to future CI technological development.

Consideration of career advancement is also important at the department or institution level where a science project may include both domain experts and computer scientists, or software engineers. While a domain expert may have advancement through tenure-track advancement or other discipline duties, the technical specialists will need similar career progression support. The same would apply to a domain expert working in a technical group.

Providing a reciprocal career advancement mechanism will undoubtedly require collaboration between departments in an institution, or between institutions, to provide the professional rewards for all staff – domain or technical. This will be an important area to address, especially for small departments, in order to retain the quality staff that will be required to implement and maintain any CI.

#### CI Services and Data in the Classroom

A key requirement in the vision of a CI-enabled science education is the availability of CI services and data to the range of educational environments we noted in the section on preparing the workforce. The range includes universities, colleges, and K12, and potentially the growing numbers of distance and home schooled students. The ability to provide such ubiquitous access will be a huge challenge for the CI in general, not just to that part supporting the atmospheric sciences. If we can meet this challenge, we then have a way to broaden the audience for the science, both in the education of the student body, and in better developing a public interest and understanding in the science. Inadequate access to computational infrastructure and network bandwidth at many schools and colleges will reduce the size of the pool of students from which to draw the next generation of scientists and technologists. Disparities in access to CI often run along the economic and cultural boundaries in our society. We may not be able to directly solve the disparities, but these factors should be considered in development of CI services so that they have the broadest use possible.

In bringing CI services and data into college and school classrooms, and even to those learners in unconventional environments such as home schooling, we also need to develop the partnerships with the teachers, instructors and education specialists who can provide input on how to use data and services within well designed educational resources that meet educational goals, whether those goals are degree level domain specific, or K12 science and math standards. While the educational resources utilizing CI services and data will be different based on the goals (e.g. resources developed to challenge upper-level undergraduates, or those developed to support 6<sup>th</sup> grade science and math), we would like to ensure that the underlying services can provide the interfaces needed to support the variety of education goals.

Tools to analyze and visualize data derived from CI data services will need to be developed to support the education environments as end users. This will present a challenge to tool developers to understand the requirements of different educational and research tasks, and from that develop application frameworks upon which interfaces can be built to support the particular task environment. Collaboration with experts in the fields of human-computer interaction, cognitive science and psychology in addition to education technology specialists and evaluators can help in understanding these types of requirements and developing solutions. Not only can these approaches be used for tools, but they can also be applied to CI services in helping to support the range of CI users from novices through experts.

#### QUESTIONS:

- Is it reasonable to assume that the CI can stimulate closer relationships between research and education communities?
- While a digital library can be an environment for bringing together interactive resources, data, intellectual ideas and knowledge, is the atmospheric science community prepared to produce the needed materials and resources needed for the library contents?
- What changes, if any, are needed to the academic reward structure to allow for recognition of outreach efforts in education?
- How can the atmospheric science community use the CI to help prepare, or improve, the science training of the K-12 teaching community?

### 3.1.3 Intellectual Property

We are seeing an explosion of online availability to digital information where the World Wide Web has provided a platform for ease of information sharing. The CI initiative calls for more collaboration in the scientific community, not just in sharing information over the web (or its future incarnations), but also in collaboration on experiments, systems and knowledge development across discipline and geographic boundaries. This expansion of access and collaboration can cause concern when considering the ownership of information, data or algorithms, and how the wishes of the owner can be maintained in the broader dissemination and use CI envisions.

Intellectual property is protected under laws such as patent and copyright where rights (subject to time limits) have been given to the original author/creator/inventor to decide who should have access to, and make use of their intellectual output. They also have the right to transfer those rights to others, or remove all limitations (release to the public domain). An important aspect to a successful CI and the culture of collaboration envisioned will be the extent to which the wishes of the owner of intellectual property (derived data, algorithms, papers, reports, instrumentation innovations, etc) are considered and incorporated into the scientific culture that uses the CI as a fundamental tool. The wishes of an owner are usually expressed through licensing mechanisms (licensing of copyright or patent rights). These provide the guidance of how a work can be used by others e.g. what sorts of use are allowed, what type of reuse is allowed, to what extent derived products can be developed, to what extent can the work be used in other work, inclusion of attributions, citations etc. that provide the intellectual history of a work. There will be the need for tools and processes to allow these “license” terms to be visible, transportable, maintainable, and protected in the community.

Within the atmospheric sciences research and education communities there is a tradition of sharing of data based on the perspective that climate has no political or geographic boundaries. The use of “proprietary holds” on data is not too prevalent. There is also a strong notion of providing back to the public arena those data collected at the public expense. In terms of the tools used, there is a broad collection of open source systems and proprietary systems in use in our research and education institutions. There are questions related to using proprietary systems, especially if they do not conform to openly known standards, or allow for inspection of how they work, both of which can be important to supporting the scientific reporting process (publishing results). When such proprietary systems are opened for access to the community there is the need to protect the intellectual property rights of such systems. Conversely, a researcher may want to protect their approach and findings to allow for future private development. The culture of the scientific domain also sets a general tone of what is considered acceptable in retaining control, and that cultural tone may be tested, or require changes, to allow the envisioned cyberinfrastructure to succeed.

#### QUESTIONS:

- What should the intellectual property rights culture in the atmospheric research and education community look like in order for the community to succeed in utilizing cyberinfrastructure?
- While there is a strong ethic of sharing in the atmospheric research and education community, what are the potential benefits, and pitfalls, of managing intellectual property rights in the expanded collaboration space envisioned by the cyberinfrastructure? Consider that in protecting IP rights, various protection and security tools and mechanisms will obviously be needed which can impact efficiency, privacy, and speed of dissemination of ideas and results.

### 3.1.4 Integrating Scientific Information Systems and Geographic Information Systems

Geographic Information Systems (GIS) have become indispensable tools for exploratory geoscience, commerce, and for decision making in the environmental and social sciences. In general, GIS involves a broad array of computer tools for mapping, managing geographically referenced data, and spatial analysis. The extraordinary gains in computer performance over the past two decades have seen a parallel growth in GIS applications. These applications have not only been growing in number but also in their diversity. An important reason for the proliferation of GIS use is that it provides a convenient framework for multidisciplinary analysis and synthesis, which is becoming increasingly important as researchers explore the frontiers of science. As the demand for innovative GIS applications, services, and know-how grows, GIS is expected to play a pivotal role in shaping the CI development in the geosciences.

Due to its powerful capabilities, GIS enables spatially referenced data from disparate sources to be integrated into a single environment. An important application of GIS is to link remotely sensed data from a variety of instruments with various socio-economic data and biophysical datasets in a common framework. For example, GIS can be used to integrate radar, lightning, and satellite data with land use and population data to study how deforestation and urban development affects the occurrence and frequency of forest fires.

While the atmospheric science community has a rich tradition in developing specialized scientific analysis and visualization tools, which can be loosely characterized as scientific information systems (SIS), to process, analyze and display atmospheric data, the field has only recently begun to embrace the use of GIS in education and research. One of the reasons for the slow adoption of GIS by the atmospheric science community is that current GIS frameworks, due to their limitations in data models and lack of conceptual and physical interoperability, are not suited to the management and analysis of dynamic, multidimensional atmospheric datasets. Research is needed to identify new frameworks and methodologies to integrate database constructs of GIS with SIS datasets. For instance, combining real-time and forecast weather information with GIS databases of population and infrastructure has significant potential for greatly improving weather related decision support systems. The utility and integration of “off-the-shelf” GIS tools and scientific applications in the context of climate change and disaster research, mitigation and response should also be of high priority in the development of future cyberinfrastructure in the atmospheric sciences.

The trend of GIS application shifting from an operational support tool to a strategic decision support system, as described above, is followed by the demand for the incorporation of more powerful analysis techniques. It is into this context that the need for basic research in the operation, development and use of spatial data analytical techniques and spatial modeling should be identified as an important focus in the future CI development.

Another recent trend of GIS is to make it accessible via the Internet, allowing more easy exchange of data and functionality. GIS applications were historically built as stand-alone tools, often based on proprietary architectures. However, with the increasing availability of geospatial information from diverse sources and their disparate application in different settings, there is a growing recognition that standards are the key to the interoperability and wider use of GIS tools and services. Organizations like the Open GIS Consortium and International Standards Organization are aiming to address these interoperability and connectivity issues based on open standards, interfaces and protocols. In fact, one of the stated goals of the OpenGIS specifications is to make it easy to integrate, superimpose and render for display geospatial information from different sources and perform in spatial analyses even when those sources contain dissimilar types of data.

### **3.1.5 Intra-Agency Coordination**

The report of the NSF Cyberinfrastructure Advisory Committee recommends that “the NSF should establish and lead a large-scale, interagency, and internationally coordinated Advanced Cyberinfrastructure Program (ACP) to create, deploy, and apply cyberinfrastructure in ways that radically empower all scientific and engineering research and allied education.” It further recommends the establishment of an ACP Program Office (ACPO), the director of which would be adequately positioned within the NSF (e.g., at the Assistant Director, or AD level) to effectuate leadership at a national level. More locally within the NSF, the ACPO is envisioned as an entity that does not fund research directly, but rather sets priorities based upon input from, and distributes funding to, the directorates for subsequent distribution to the community via established mechanisms. A Steering Committee, consisting of the ADs of all directorates and chaired by the AD of CISE, would collectively have responsibility for the overall success of the ACP.

This matrix management model will require a new, much greater level of interaction among all ADs and directorates given the crosscutting nature of CI and the potentially significant amount of funding involved. Although each of the directorate Advisory Committees (AC) is envisioned to play an important role, an additional group within each directorate, focused exclusively on CI and perhaps operating as a formal sub-committee of the AC, with membership not restricted to the AC, likely will be needed. Such groups, the CyRDAS being one such example, are now planning for the future; however, a sustained presence will be needed to ensure the continued health of the ACP. Membership on these committees should not be restricted to disciplinary experts, but should involve individuals from other relevant disciplines (and thus directorates) to ensure an adequately broad vision of CI, to identify areas of commonality and distinctness, and to effectuate interdisciplinary activities that might be brought forward as opportunities for funding. In the area of atmospheric sciences, virtually every Directorate is relevant, with the following particularly so: Education and Human Resources (HER); Social, Behavioral, and Economic Sciences (SBE); Computer Information Sciences and Engineering (CISE); Biology (BIO); and Mathematics and Physical Sciences (MPS).

### **3.1.6 Interagency Coordination**

There is a long and successful history of cooperation and coordination among program leaders from the many Federal agencies supporting research in both the atmospheric and computational sciences. Most notable are the Global Change Research Program and the High-Performance Computing and Communications Program, both of which were started over ten years ago. The interagency coordination process provides a diversity of perspectives and missions that can be used to develop a set of research priorities to address the most critical research problems. Often interagency support can enable multi-institutional collaboration among research centers to bring together the multi-disciplinary teams required to address complex problems.

Nevertheless, such coordination requires dedicated effort by Federal managers to be successful. Interagency committee and working group participation is often an addition to the workload of an already overburdened manager. Agencies must also develop a mechanism whereby conflicts between agency and interagency programs can be resolved optimally.

In the case of CI, the NSF must coordinate the program with other Federal IT and computational science initiatives, such as the Department of Energy SciDAC and NASA ??? programs. Further, links between these programs and atmospheric science research programs must be established both between and within the participating agencies. This will result in a somewhat complex three-dimensional matrix of program goals and agency responsibilities that will take dedicated effort to maintain. The potential benefit from coordination fully justifies the investment in time and energy. The rapid transfer of knowledge and technology among disciplines and between agencies would accelerate research progress in many fields. In many

cases, cooperation will provide a critical mass of needs and experience that can provide IT solutions to many research problems that are poorly served by commercially available technology.

## 3.2 Data

Scientific research and education are based upon inferences that can be drawn from measurements to either support or reject hypotheses. Such measurements may be made directly on a physical or biological system or may be taken from a remote platform, which is displaced at some distance from both the scientist and the object being observed.

Within the past 50 years, in addition to theory, experimentation and measurement, a fourth scientific methodology has emerged, namely computer simulation, that has taken its place as a legitimate way of conducting an inquiry. Data are at the heart of the scientific method, whether in the form of measurements taken in the course of an experiment to test a theory or generated in a computer simulation.

The atmospheric sciences, typically earlier than most other disciplines, have taken great advantage of remote sensing and computer simulation to advance understanding of phenomena in the atmosphere. As a result, atmospheric scientists have often been the first to experience the difficulties of handling very large volumes of data from diverse observing systems or numerical models. Because data are the raw material on which computers operate, it is natural that the CI that will be developed for the atmospheric sciences will have to respond to significant demands for data and data services. The following describes some of the issues that must be addressed in this regard.

### 3.2.1 Volume

One of the exponential growth curves that makes planning for a CI that enables atmospheric science to advance most rapidly a challenge is the volume of data and metadata. As an example, the repository of atmospheric science data at the National Center for Atmospheric Research has experienced a net annual growth of 33% for at least the past decade, growing to over a petabyte ( $10^{15}$  bytes) within the past year. At any given time on this exponential curve, half of all the data in the archive was added within the last 1.5 years. This rate of growth puts stress on all the systems that are used to support the data archive, including

- dedicated processing power,
- the sophistication of the machinery (e.g. robotic tape archives systems),
- technological development of media to contain the data,
- migration of media to higher density and transfer speed over time, and
- the network over which the data are delivered to scientists.

For a typical research project whose lifetime is 3-5 years, the most interesting portion of the archive is the data added to the system within that period, i.e., more than 75% of the total, which means that speed of access to nearly every bit of data has to grow at least as fast as the archive.

### 3.2.2 Complexity

For those who have been working with computers for many years, the evolution of the complexity of data is perhaps the most jarring. Several decades ago, data were typically stored in a central facility and written using a simple encoding that could be decoded with a very short computer program and that would maximize efficient usage of the scarce resource – media capacity. The encoding typically included a format that could be printed (on paper or punch cards!) with a packing algorithm to save precious space on media. Metadata, the data that describe the measurements or computer model output, were kept very simple, including primarily location and time information. Ancillary information such as units, observing system description, and observer (or model) identification, were typically stored in separate manuals or were included as part of the encoding algorithm.

Today, the situation is far different. For example, data may be obtained from remote sensing platforms for which the location and time of the observation is a complicated

mathematical calculation that is considered part of the “post-processing” of the data. Data are often stored in a format that is not described in simple terms but is accessible by an applications programming interface (API). Extremely rich metadata that describe not only the data themselves but also include copious ancillary information items may be stored directly with the data as part of the same file. The data may be distributed to several data centers or even more heterogeneously among a large group of collaborating scientists at multiple distant research organizations.

This significantly increased level of complexity introduces several challenges for the CI. For example, increased complexity increases the time required to be able to begin using a data set for the first time. While this is an annoyance to those scientists who plan to make extensive use of a given data set and who will reap the rewards of the sophistication built into the data system, it represents a barrier to entry for casual users or, worse, for educational users. Increased complexity also adds to the data management burden both for those who create a data set and for those who need to maintain it. For example, scientists who are engaged in a field campaign to make a large number of measurements may not be able to populate all the fields of a rich metadata base, so that fields go unfilled that might be useful to later users.

### **3.2.3 Management**

There are many aspects to data management, but three are of special importance for the atmospheric sciences. First, the fact that data sets are now more likely to be distributed than centrally managed introduces several problems such as keeping track of the various locations and status of each part of the data set, maintaining consistency between the various parts of the data set (this is especially critical for real-time data sets), maintaining security and reliability of the data sets, providing comparable levels of redundancy and backup for data sets that are spread over multiple storage centers.

Second, there are now multiple data sets that represent measurements or simulations of the same four dimensional volume of the atmosphere that should be interoperable in some way. For example, when multiple satellites pass over the same portion of the globe at the same instant, recording the characteristics of the atmosphere at that point, how can the disparate measurements be brought together for making more robust estimates of the measured quantities or the errors in those measurements? Similarly, how can a model simulation be compared with remote sensing measurements of the quantities being simulated? These questions are at once scientific (for example, how does one compare point measurements, swath measurements and simulated volume integrals?) and technological (for example, how does one compare data maintained by different government agencies that measure or simulate the same thing?). The technological aspects of the data management problem of linking diverse data sets can be addressed by the CI.

A third important data management issue is the handling of historical data. This problem has two aspects. First, there is the need for digitizing old data currently available only in analog format, for example magnetometer strip chart data. Second, there is a need for archiving and maintaining old and current data. Data archives must regularly be updated because data stored on existing media have a finite lifetime and storage technologies change rapidly, rendering data unreadable. The problem exist both for large archives and the individual PI who needs funding for such upgrades.

### **3.2.4 Acquisition and Instrumentation**

The vast domain that is addressed in the atmospheric sciences, extending from the upper ocean and lower atmosphere through the upper atmosphere and the near-Earth environment to the surface of the Sun, represents a huge measurement challenge. Nearly all of this domain is inaccessible to direct measurement, either because it is too far away to reach by human observers or the environment is too hostile to envision sending human observers to make *in situ* measurements. In fact, meteorologists were among the first to make use of remote sensing by

sending measuring devices aloft into the atmosphere aboard kites, balloons and rockets. The atmospheric sciences are therefore uniquely dependent on remote-sensing apparatus. Synoptic meteorology and weather prediction make heavy use of operational weather satellites and radars, climate studies use autonomous floats and drifters that can telemeter measurements to receiving stations on land, and upper atmospheric research makes use of satellites and radio telescopes to infer the structure and composition of the atmosphere beyond where *in situ* measurement is possible. The heavy dependence on remote sensing introduces a data acquisition and instrumentation challenge that has, in the past, been met on a case-by-case basis.

To provide a concrete example, the following relates to upper-atmospheric research, and radiowave measurements in particular. This example serves to illustrate how important remote sensing has become for the atmospheric sciences and also to indicate how crucial it will be that the CI addresses this challenge.

NSF-supported observations of the Earth's upper atmosphere, from the mesosphere to the Sun, are carried out using remote-sensing techniques. Most of these observations utilize radiowave measurements. These observational techniques are already undergoing revolutionary changes as a result of advances in computational power, data storage capacity and computer networking, primarily because it is now possible to implement a radio receiver in software. Yet the potential for these techniques has barely begun to be tapped because digital receivers produce data at far greater rates than can be saved, transported or processed with current technology.

As an example, incoherent scatter radars, the most powerful ground-based technique for probing the Earth's upper atmosphere can now collect raw data at rates on the order of 10 T/bytes per day. In the past these data have always been processed in real-time to reduce the data volume. However this entails significant information loss, and the trend is now to save as much of the raw data as possible. This is almost possible now, but several trends will increase the data-processing requirements by orders of magnitude. First, incoherent scatter radars are beginning to move away from their traditional research-oriented campaign-based mode of operation to much longer runs. To fully support real-time space weather forecasting, continuous operations will eventually be required. Second, the number of radars is increasing, for example the new Advanced Modular Incoherent Scatter Radar (AMISR). The three AMISR faces now being built will collect data at prodigious rates, and typically will be located in regions with poor network infrastructure, such as the Arctic and Antarctic. Third, recent technological advances and the requirements of space weather forecasting are likely to lead to a new generation of less inexpensive radars which will be widely distributed around the globe.

Meanwhile, the digital revolution in radio technology has led to entirely new classes of instrument which will present even greater cyber challenges. For example, the LOFAR radio telescope, which will be a powerful tool for atmospheric science as well as astronomy, will collect data at an aggregate rate of ~25 Tb/s. Distributed arrays of hundreds of passive instruments requiring a minimum of 100 Mb/s Internet connection per site, and ideally orders of magnitude more, are also under consideration.

An important aspect of the switch to digital instrumentation is the change in the workforce needed to design, implement and maintain this instrumentation. An increasingly higher proportion of the hardware is off-the shelf commodity hardware or hardware produced by industry in moderate to large numbers for distributed networks of instruments. There will be less need for hardware engineers and electronics technicians, and a much greater need for software engineers.

#### QUESTIONS:

- What is the proper balance between archiving data at a central location, such as NCAR, and at distributed sites such as observatories or PI institutions?
- Can anticipated developments in networking, e.g. teledesic, meet the needs of globally-distributed, high-data-rate atmospheric science instrumentation?
- What can we do to make old analog data available in digital form, and to ensure that both old and new digital data are preserved on current media?

### 3.2.5 Knowledge

#### Metadata

The World Wide Web provides a vast potential for the efficient access to and exchange of educational materials, data and tools. But while the web offers an enormous and fast growing amount of information, this information is largely unstructured and not organized efficiently. Searching for any specific information and data is often cumbersome.

Metadata or "data about data" describe the content and other characteristics like quality, lineage and condition of data. The primary use of metadata is resource discovery, making it possible for users to search, retrieve, filter, analyze, classify and evaluate datasets of interest without needing to actually process the content. Also, distinction has to be made between resource discovery and knowledge discovery or knowledge extraction. Metadata facilitates the former, where as data mining is a vehicle for the latter. The atmospheric science community like any other scientific community needs better tools for both resource discovery and knowledge extraction.

To reach the broadest community of users, metadata must be made available in accordance with a number of popular metadata standards. Moreover, metadata maintained in one standard should be accessible in alternate standards via so-called crosswalk transformations. In recent years, a large number of metadata standards have emerged, and many more are currently under development. Examples of popular standards include Dublin Core, Federal Geographic Data Committee (FGDC), and Global/Government Information Locator Service (GILS). Despite the growing interest in metadata within education and scientific communities, however, the quality and quantity of metadata that is associated with digital content varies greatly. The development of interoperable online metadata standards and third-party aggregation services that support a broad range of applications and activities is an important requirement for any online discovery system. To that end, new middleware is needed to overcome inherent difficulties of autonomy, heterogeneity, and interoperability.

An often overlooked aspect of knowledge management in the atmospheric and related sciences is the metadata describing models and model runs. We are moving into an era in which there is a growing desire and ability to run models flexibly: in different configurations, in ensembles, with a variety of components in multi-component simulations. It is useful, when assembling applications and comparing results, for researchers to be able to readily identify the types of parameterizations a model contains, the type of dynamical core, what platforms it runs on, what data it requires to run, what data it can make available, and other basic model-related information. Groups such as the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and model developers at GFDL have begun to archive and categorize such information in a systematic way.

#### Semantic Web and Controlled Vocabulary

Automating resource discovery, retrieval, and information synthesis is difficult because the process often involves human interpretation. To address this problem, a new architecture for the Web, called the "Semantic Web" is emerging. It is important to note that the semantic web is not a separate Web but a natural [language] extension of the popular World Wide Web. In

broadest terms, semantic web encompasses efforts to create mechanisms that augment digital materials with their formal semantics, yielding information that can be used by automated, intelligent agents. The semantic web allows more automated functions and paves the way for true customization of information for individual users. For example, information on the Web can exist in their native form and any context-dependent presentation can be rendered, on demand, to meet the needs of end users.

Retrieval based on controlled vocabulary is step toward and one way of exploiting content-based metadata. A controlled vocabulary provides an interpretive layer of semantics between the term used by the user and the underlying database to better represent the original intention of the terms of the user. By examining metadata from across a collection of documents, and combining it with data from other sources one can begin to discover information that is not available in any single document.

#### Knowledge Discovery and Data Mining

Given the explosive increase in the volume of data generated by modern observing systems and atmospheric prediction models, Atmospheric Scientists are increasingly finding the need for not just resource discovery but also knowledge discovery systems to extract “knowledge” from the vast volume of data that they deal with. Knowledge discovery in scientific databases, often referred to as Scientific Data Mining, is a relatively new area in the application of artificial intelligence concepts to data analysis and interpretation. Data mining has already had considerable success when dealing with numerical and other structured relational databases in areas like consumer behavior and purchasing analysis and fraud detection.

The practical use of data mining is best illustrated using a few concrete examples. For example, one can query the large WSR-88D radar archive at the National Climatic Data Center and retrieve all radar datasets associated with tornadic vortex signatures over a geographically bounded region during a given period. Another example would be to automatically extract, composite and visualize simulated Category 3 hurricanes for an ensemble prediction system of hundred members. However, the extension of data mining concepts from today’s relational databases that are widely used in the business applications to scientific datasets is not straightforward. Traditional data mining techniques and algorithms are easily overwhelmed by the sheer volume and complexity of information in scientific datasets. The fundamental problem of extracting knowledge and insight from massive datasets will continue to be a challenging computational problem. A new, interdisciplinary field of Data Mining and Knowledge Discovery in Databases has in fact arisen in recent years to confront this challenge, merging ideas drawn from statistics, artificial intelligence, machine learning, database research, and high performance computing.

The CI development for implementing data mining ideas in high-performance scientific computing requires a fundamental shift in how computing resources are allocated. Dedicated access to massively parallel and distributed systems is needed to ensure system scalability as datasets grow in size and scope. Some level of interactivity will also likely be needed to be built into these systems to train machine learning algorithms.

#### QUESTIONS:

- How is metadata relevant to your educational and research needs?
- Are the ongoing efforts in metadata generation by digital library communities sufficient to address your specific needs?
- Does the atmospheric science community need a concerted effort to produce metadata for various digital collections in our field?
- Given the complexities of generating metadata, what barriers you foresee in generating metadata for your own datasets?
- Should future CI investments be made to produce ontologies for atmospheric and related sciences?
- How do you envision using scientific data mining tools in your research?
- What scientific and computational advancements need to be made to enhance the current generation of scientific data mining tools?

### **3.3 Computing Infrastructure and Capacity Building**

In the early 1990s, We now have evidence that there are applications of national importance that would benefit significantly from an alternative to COTS-based solutions. Therefore, research and development efforts in alternative architectures and enabling technologies are needed to ensure U.S. leadership in high-end computing (HEC). Further, high procurement and operating costs of HEC systems limit access by many agencies to these resources. A strategy of coordinated acquisition processes coupled with new mechanisms for coordinated access to unique high-end computing facilities is important to ensure the effectiveness and efficiency with which the government acquires these resources.

The atmospheric sciences community (and other science fields) has, in the past, been able to address its computing, networking and data storage needs through the selection of products from a small number of vendors who specialized in serving the science and engineering research communities. Shared-memory, vector-supercomputers, such as those manufactured by Cray Research Inc., were the most obvious examples. That situation changed dramatically in the early 1990s as high-bandwidth networking technology coupled with the exponentially increasing speed of commodity microprocessors and other IT components (Moore's Law) enabled the development of what became known as "enterprise computing." Large systems consisted of many smaller components, such as data servers, workstations, compute servers and disk arrays, which were all connected by a network. Vendors developed their products around building-block components connected through a software and hardware infrastructure.

Market forces determined the types and availability of products offered by the vendors. High-end workstations and inexpensive data storage technology enabled the individual researcher to perform computationally intensive applications locally that required a major supercomputing facility just a few years before. At the high-end, massively parallel high-end computers, built by coupling hundreds of fast commodity microprocessors into a single system replaced shared-memory vector supercomputers. The Federal government adopted a strategy of pursuing computing capability based on systems built from commercial-off-the-shelf (COTS) components. The promise of high aggregate performance at relatively low cost made procurement of COTS-based systems attractive; however, a capability and usability gap developed between what was available and what was needed. These new systems did not adequately solve the computational requirements for numerical model development and application. Although the machines had a large amount of potential power, this power was increasingly difficult to realize. Uncertainty in

the vendor community, as well as a surprisingly high degree of diversity among high-end designs has made developing models that are flexible, portable and easily maintained nearly impossible. The major problems encountered in developing models on large, distributed-memory, massively-parallel systems are explored in more detail in the following section. The difficulties of the past several years, as well as the continued success of shared-memory vector supercomputers internationally, has renewed interest in the development of shared-memory vector systems, although their cost and availability to the atmospheric science community remains in question.

#### QUESTIONS

- Are the high-end computing solutions that are currently available commercially, and those that may reasonably be anticipated to be available in the near future, meet the needs of atmospheric sciences research?
- Is special-purpose design and development of computational systems required for atmospheric science research projects that may arise over the next ten years?

### 3.3.1 Computation

In the atmospheric sciences, historically at the forefront of the application of new computing technology, the introduction and wholesale adoption of advances in high-end computing is inevitable. Similarly, the acceleration of development and deployment of computer hardware technology is pushing innovation in atmospheric science research. For example, the rapid increase in processor speed – Moore’s Law –has enabled previously unimaginable increases in the resolution and complexity of numerical weather prediction, climate simulation, space weather simulations and near-Earth models.

Numerical simulation and prediction poses two major challenges. First, numerical models typically solve coupled partial differential equations in four dimensions, so that doubling the resolution of the model results in a factor of 10-16 increase in computational burden, depending on how tightly the temporal and spatial integrations are coupled. Thus, a model with twice the resolution of its predecessor requires as much as four doublings of processor speed, or 4-6 years of chip advances, just to complete the same amount of simulated time in a given interval of wall-clock time. Second, over the course of 4-6 years, models seldom remain static – new observations, improved parameterizations or advances in numerical methods are incorporated into models, usually making them more complex and, hence, computationally costlier. In order to realize the increases in resolution and complexity that faster processors enable, therefore, it is essential to use more and more processors simultaneously in a maximally efficient manner. Performing ever larger and more complex computations requires a very large number of processors to be applied to each problem, and it requires a very efficient parallelization of codes to efficiently use that multitude of processors. The challenge of parallel computing is described in Appendix C.

#### QUESTIONS

- What computing capability and capacity requirements are anticipated for the next five years? 10 years?
- Is there a particular architecture or programming model that should get priority in the atmospheric sciences?
- In what way can progress be made most efficiently to take advantage of parallel computing resources for the atmospheric sciences?
- What level of support should be provided for the various programming models?

One critical decision that is engendered by these changes is whether computing should be centralized or distributed. In this context, “centralized” computing refers to the very few centers with budgets large enough to acquire, maintain and support the largest clusters of computing nodes to provide computing capacity and capability for atmospheric science researchers. “Distributed” computing refers to the deployment of small to mid-size computing platforms in university departments and laboratories, being used collectively or even by individual principal investigators (PIs) throughout the U.S. A third form of computing that is being vigorously investigated as a computer science research question is “Grid” computing, in which the distributed computing assets are treated as a logical “center” in the sense that heterogeneous, physically remote computing resources can be made available to a single code. The advantages and disadvantages of centralized and distributed computing are enumerated in Appendix C.

To provide some concrete examples, the computation issues in three areas of the atmospheric sciences: weather prediction, climate modeling, and space weather are described in Appendix D. Climate and weather research have traditionally represented one of the Grand Challenge problems for high-end computing. This is due in large part to the enormous range of spatial and temporal scales that characterize the Earth's physical and biological processes. Spatial scales range from the molecular scales that characterize many chemical processes in the atmosphere and ocean to the large-scale dynamics that determine the Earth's climate. Temporal scales range from the minutes required to forecast a tornado to multi-century model integrations used to investigate global and regional climate change. In recent years, weather modeling has broadened to include space weather. This not only brings in additional physics (magneto-hydrodynamics and charged particles) but also an even larger range of spatiotemporal scales.

#### QUESTIONS

- What are the requirements of the atmospheric sciences research community for centralized and distributed computing resources?
- How should resources be distributed among individual PIs, small centers and large centers?
- Which computing resources should be centralized and which should be distributed?
- Can the costs and benefits of centralized and distributed computing be quantified?

These types of research typically require very large bandwidths between memory and processors, and between computational nodes. The unavoidably global nature of some steps in these calculations also introduces a requirement for a high degree of inter-nodal connectivity. These requirements indicate a need for architectural innovations, and software engineering resources are required to convert the very complex weather and climate codes to run on each new architecture. High-resolution climate and weather models generate a very large amount of data (on the order of 1TB/model-day) and hence require a machine with very efficient parallel I/O and a large local storage system. Scientific analysis of the model output would benefit from very capable visualization and data-mining tools. When used in an operational mode, high-resolution weather models must also ingest large amounts of real-time data and return large quantities of

output to the user (possibly another model running on a different system), necessitating high bandwidth network links.

### 3.4 Software Engineering

There is no single characterization of the nature of software engineering in the atmospheric sciences. Software engineering may be undertaken by people with a broad spectrum of skills and training, working on a variety of problems, on projects ranging in scale from individual to major national efforts. Any of the following activities can reasonably be thought of as falling in the realm of software engineering:

- a seasoned support scientist coding a complex model that runs on parallel supercomputers
- an applied mathematician developing high-performance parallel numerical libraries
- a graduate student new to programming developing a tidal model on a workstation
- a commercial programmer working with a team on a networking project
- a software project manager responsible for structuring the technical aspects of a large modeling or infrastructure project, overseeing the development team, and coordinating with management, scientists, and program managers

Depending on the task required, expertise can be a combination of science, mathematics, computer science, software design, software process, and management. Given this diversity, how do we approach the challenge of advancing software engineering in the atmospheric sciences? We can begin by identifying a number of basic categories describing software engineering projects in our field, and key software engineering issues. These are the scale of the project, its nature – scientific or engineering, its requirements and appropriate management. These characteristics are further explained in the Appendix B.

#### QUESTIONS:

- What software practices and design methods are most critical for smaller projects in the atmospheric sciences, and how can those skills be taught?
- What software practices and design methods are most critical for large efforts? When is it appropriate to train individuals within the Earth science community with the required skills and when is it appropriate to seek expertise from outside our community?
- What can we learn from the commercial domain and other fields about the management of software projects?
- How can software processes help to bridge the gap between the computer scientists and software engineers who develop software and the Earth scientists who use software?
- How can we identify and train software project managers?
- How can we best mentor and develop young software engineers?

#### 3.4.1 Modeling Frameworks

Two decades ago, an atmospheric simulation was generally a stand-alone application forced with data sets. Today, while an atmospheric simulation may still be a stand-alone application, it may also be one of multiple interacting components in a climate model, it may combine with a data assimilation component to form a forecasting system, it may drive a chemistry model, or it may be one of many instances in an ensemble. The list of possible configurations goes on, and continues to grow as scientific exploration motivates new experiments and technology offers greater computational capabilities.

While the coupling of components developed by independent teams can be done in an *ad hoc* manner, this can lead to applications that are difficult to understand, modify, and extend. Definition of a software architecture – a set of clearly defined rules for how components can interact – allows an atmospheric model to combine with other components in a systematic way to form an application. Ideally the architectural rules are strict enough to make the integration of model components into composite applications quick and easy from a technical standpoint, and are flexible enough not to impose restrictions on scientific development. Modeling frameworks, which may be viewed as integrated sets of customizable software building blocks, can offer interoperability of components through the careful definition of architectural rules, together with underlying toolkits for data regridding and redistribution. The ways in which modeling frameworks can help the atmospheric sciences are described in Appendix B.

### 3.4.2 Community Models

The growing sophistication and complexity of numerical simulation/prediction systems – coupled with increasingly limited support for developing and maintaining them, the desire to avoid duplication of effort and focus intellectual resources, and the need to establish more effective pathways from basic research to operational implementation – have resulted in the establishment of several community models. From these systems are emerging exciting new discoveries and capabilities, as well as standards for software interoperability, output sharing, and full model coupling. A full description of the nature of community models is given in Appendix E.

## Cross-Cutting Issues

While it is very useful to examine the problem of developing an effective CI that meets the needs of the atmospheric sciences in terms of the four thematic perspectives above – people, data, hardware and software – it is also useful to consider several issues that have elements in each of these perspectives.

We have identified four cross-cutting issues that must be addressed in the development of a CI for the atmospheric sciences: barriers to entry, operational and real-time needs, software life cycle, and coordination.

### 3.5 Barriers to Entry

There are real barriers to effective development and use of CI in the atmospheric sciences. Among the barriers are the following:

- Human resources – An essential and underemphasized element for a successful CI is the need for atmospheric sciences faculty and students, as well as computing professionals with some atmospheric science knowledge, who are well prepared to take advantage of or contribute to the computing, data and software capabilities that are emerging. A successful CI initiative will address this high priority need in order to facilitate both frontier research and education. Attention should be given to the demographics and issues of the next generation workforce, and strategies should be developed to insure that the atmospheric sciences are competitive in attracting the best and most promising talent.
- Lack of awareness – Too many researchers and educators in the atmospheric sciences are unaware of computing, data and software resources that are already available, either in the academic community or in the commercial marketplace. There is a need for better communication of, and dissemination to, the atmospheric sciences community of such resources and how they can be used.

- Uneven and inequitable access to or distribution of resources – The ownership and access to computing, networking, data, and computing professional resources varies greatly from department to department. A review of this situation should be undertaken and a response developed to provide better access to and a more equitable distribution of resources, based on NSF’s mission, through intelligent and well-planned investments.
- Opaqueness – The CI that has been developed and put in place over the last several years has, at times, made computing and scientific use of data more difficult. This is a trend that must be reversed, through hardening of middleware and application software, through increased transparency of the computing infrastructure, through development of new software tools, and through increased discoverability and interoperability of the diverse data sets that are becoming available, both within the atmospheric sciences and in other related disciplines.
- Computing infrastructure – There continues to be a large gap between available high-end computing capabilities and the computing capabilities needed to address important problems in the atmospheric sciences, including studies of the Earth system. This large gap should be addressed in the context of the evolving computing infrastructure.

### **3.6 Operational and Real-Time Needs**

A unique aspect of atmospheric sciences is the close relationship among research, education, and operational meteorology – pollution alerts, weather prediction, climate prediction, space weather alerts, etc. The CI that is developed for atmospheric sciences should, at all stages of design and implementation, support this relationship, whenever practical. This has implications for the design of networks (e.g., high-speed connections between research laboratories and operational forecasting centers), data distribution (e.g., real-time meteorological observations in the classroom), computing infrastructure (e.g., on-demand computing), and management and access to data sets and community models.

### **3.7 Software Life Cycle**

The high concentration of research resources on innovative development has led to neglect of funding for the full software life cycle, including hardening and deployment. Further, there is a critical need for funding for training, support, and maintenance of the underlying infrastructure. Without viable options for supporting software throughout its useful lifetime, investments in software infrastructure and tools will be lost.

### **3.8 Coordination**

There are already many diverse CI planning and implementation efforts within different divisions of the Geosciences, different directorates of the NSF and among the various federal agencies with sizable portfolios in research and development. These activities should be coordinated, whenever possible, to optimize the return on investment and to minimize the duplication of effort.

## 4. Recommendations

In order to address these pressing issues, it is recommended that the NSF change the way in which it invests in CI for the geosciences in the following ways:

- Give human resource issues top priority in CI development, hardening, maintenance and distribution. The academic reward structure should recognize CI-related activities, and NSF should seek additional ways to support investments in CI personnel at all levels, including departments, campuses and centers.
- Encourage the education of students, educators, support staff and scientists through exposure to the computational environment, including tools, data sets, and models.
- Provide the necessary support for mechanisms of communication and dissemination of ideas, technologies, and processes to promote the interdisciplinary understanding needed to successfully use CI in the atmospheric sciences.
- Help organize and coordinate CI activities across the geosciences and provide forums for innovation. The nascent activities of the Geoinformatics Steering Committee and the Geosciences Technology Forum are good first steps in this direction.
- Initiate a coordinated effort with NASA, NOAA and other agencies to provide geoscientists with the ways and means to find, effectively use, publish, and distribute geoscience data, including the development of appropriate standards for metadata to which grant recipients should be encouraged to conform.
- Provide funding for the entire CI software life cycle, including development, testing, hardening, deployment, training, support and maintenance, subject to ongoing review and demonstrated utility and sustainability.
- Invest in computing infrastructure and capacity building at all levels, including centers, campuses, and departments, in support of atmospheric science research and education.
- Support the development of geosciences cyber-environments that allow the seamless transport of work from the desktop to supercomputers and other Grid-based computing hardware and software.

While these recommendations are very broad, they are deemed necessary to accelerate progress in the atmospheric sciences. To further articulate the nature of changes that are recommended, the table below provides more details on both the issues and the short term (2-5 years) and long term (5-10 years) recommendations, organized according to the perspectives described above.

Issues	Short-Term Recommendations	Long-Term Recommendations
<i>Cross-Cutting Issues</i>		
1. <u>Barriers to entry</u> : Human resources, an essential element of CI, have been significantly underemphasized	Recognize “duality” in CI – physical and human resources	Give human resources issues top priority in CI development and investment
<u>Barriers to entry</u> : Lack of awareness	* Support establishment of a refereed CI journal for AS. * Include a permanent schedule of CI activities (public forum, student sessions) at the major AMS and AGU meetings. * Support AS faculty in technical sabbaticals to IT groups	* Establish a “clearing house” organization for promotion and dissemination of CI information to AS community * Support CI involvement of domain scientists at all levels via targeted solicitations
<u>Barriers to entry</u> : Uneven distribution of CI resources	Assess the types and distribution of resources across the AS in relation to AS objectives in research, education and operational meteorology	* Distribute additional resources to medium-sized projects, which have best chance of producing sustained benefits * Seek broader distribution of resources to inform the AS community and empower scientists to select and use the tools that are most effective
<u>Barriers to entry</u> : Opaqueness of CI to domain scientists and students	Establish a clearinghouse for CI information (e.g., digital library)	Concentrate notable expertise at each large center in a small number of domains
<u>Barrier to entry</u> : Computing infrastructure is not deployed or managed such that major AS challenges can be addressed	This report includes a short list of AS problems that can be addressed with a modest, $O(10^2)$ increase in computing capability	Develop and maintain a catalog of the largest AS challenges, including expected outcomes and broader impacts, whose large computational requirements are quantified and monitored
2 Two unique aspects of AS are its operational and real-time elements	Encourage more interaction between operations and educational and research activities in the AS	Support development of Geosciences computing environment that supports transition from research to operations
3 The emphasis on investments in software R&D have led to neglect of the full software life cycle that also includes training, support and maintenance	Begin to develop methods for determining applicability and longevity of given software projects	Develop set of options and criteria for renewal of successful software projects that provide adequate funding to ensure usefulness to AS community
4 A wide and complex diversity of CI planning activities threatens to lead to inefficiency, duplication of effort, and retardation of progress	Continue CyRDAS-like planning activity and ensure that CI planning activities across the Geosciences continue to communicate and collaborate on recommendations and implementation plans	Coordinate with other Geosciences, remainder of NSF, and other federal agencies via cross-cutting panels with budget authority

Issues	Short-Term Recommendations	Long-Term Recommendations
<i>Social and Cultural Issues</i>		
1. The academic culture inhibits development/application of CI.	<ul style="list-style-type: none"> <li>* Encourage NCAR to take lead in rewarding scientists who engage in CI contributions</li> <li>* Encourage short-term exchanges of technologists among labs</li> <li>* Establish a peer-reviewed journal for CI in the geosciences</li> </ul>	<ul style="list-style-type: none"> <li>* Encourage sabbaticals that includes a major CI component</li> <li>* Encourage academic reward process to recognize “published” code and code usage metrics</li> <li>* Work with AMS Heads and Chairs to develop suitable procedures to reward CI work</li> </ul>
2. Many departments are inadequately provided with computing professionals.	Survey AS departments to determine the level of computing support they have and they need	Support departmental level investments in CI personnel.
3. <u>Inadequate education and training</u> : AS education needs to be closely associated with tools and data sets that are used in research	Support workshops and collaboration opportunities to bring together AS educators, tool developers, and data providers with educational technologists, instructional designers and educational evaluators	<ul style="list-style-type: none"> <li>* Support development of a Geosciences computing environment that delivers research-quality data sets to graduate classrooms and substantial data and analysis capabilities to K-12 classrooms</li> <li>* Encourage departments to train next generation of technologists to work in AS environments</li> <li>* Encourage degree programs in technical management and development in the Geosciences.</li> <li>* Encourage interdisciplinary programs and degrees that provide a geosciences emphasis in computer science and engineering departments</li> </ul>
<u>Inadequate education and training</u> : Programming languages and networks change faster than domain sciences.	Organize “workplace of the future” workshops	Support software development and open software practices (including verification) to allow researchers to concentrate on science rather than programming
<u>Inadequate education and training</u> : Students are unprepared to understand computational physics, work with complex programming environments and interpret large volumes of data.	<ul style="list-style-type: none"> <li>* Encourage undergraduate courses in programming for scientists and engineers, including exposure to practical tools (e.g. debuggers)</li> <li>* Work with AMS Heads and Chairs to develop undergraduate sequence in computational science (also coordinate with other disciplines through their professional societies)</li> </ul>	<ul style="list-style-type: none"> <li>* Require students to demonstrate competency in computational physics, programming and data analysis (analogous to math proficiency)</li> <li>* Develop computational courses</li> <li>* Encourage AS/IT courses intended for CS students</li> </ul>
4. Culture clash between IT and applications, especially in AS – applications are not attractive for CS researchers; AS does not influence technology trends	<ul style="list-style-type: none"> <li>* Encourage AS researchers to visit, teach in or work in IT groups</li> <li>* Encourage connections among professional societies, e.g., AMS,</li> </ul>	<ul style="list-style-type: none"> <li>* Encouraged a requirements process to bridge the gap between technical collaborators and Geoscientists</li> <li>* Support small teams of</li> </ul>

	ACM and IEEE	software “craftspersons”
5. AS research and education is increasingly difficult with too many data streams and sources	Deepen coordination with NOAA and NASA	Encourage development of data grids and consider virtual observatories (metadata standards, common portal design etc.) in AS and Geosciences
6. Collaboratories have not fulfilled their potential	* Consider sociological studies to determine how best to structure collaborative facilities. * Improve awareness in AS of computer-supported collaborative work and computer-supported collaborative learning.	Support collaboratories only when success depends on collaboration

Issues	Short-Term Recommendations	Long-Term Recommendations
<i>Data Issues</i>		
1. In AS, data and metadata from diverse geography- and sub-discipline-specific sources must be universally available in a timely manner and seamlessly interoperable	Educate the community about availability of data across the Geosciences.	Support a Geosciences computing environment that enables the development, documentation and distribution of diverse data sets as widely as possible
2. <u>Improve data utility</u> : The value of individual data sets can be greatly enhanced when used with other data sets.	Encourage sharing existing data sets via distributed servers.	Support development of many interoperable data sets and catalogs of data sets
<u>Improve data utility</u> : Finding and using data that may be relevant in AS is very difficult, primarily due to lack of metadata standards.	* Develop Geosciences data catalogs * Encourage development of innovative information retrieval applications for geosciences data sets	* Encourage development of standards for metadata, mass storage and data transport among large centers * Encourage development of tools for data integration and data assimilation
<u>Improve data utility</u> : Data from different sources must be rigorously compared and validated.	Encourage thematic data grids on specific science problems	Encourage adoption of standards and tools for rigorous comparison of models and observations
3. Scientists must be able to publish and distribute (both formally and informally) their analysis, data mining and machine learning results, visualizations, etc. and tie them to the underlying data.	* Work with AMS to ensure that planning is cognizant of development of online journals etc. * Ensure that planning is cognizant of intellectual property rights issues	Support development of tools and capabilities for linking publication to underlying data and entire processing stream
4. Valuable, unique data that cannot be recreated must be sustained for the long term.	Provide support for curating AS data sets in multiple locations with on-site data stewards	* Support planning for ongoing media migration, ongoing testing of back-up and recovery plans * Support the development of strategies for curating and reviewing data archives

Issues	Short-Term Recommendations	Long-Term Recommendations
<i>Computing Infrastructure and Capacity Building</i>		
1. AS will continue to drive the high-end computing requirements of the Nation for the foreseeable future, but it will not necessarily drive the computing infrastructure market	* Begin to forge an effective linkage between the sciences and other drivers of high-end computing and CI * Assess common requirements among AS and other high-end technology drivers	Continuous, sustained improvements in the capacity and capability computing in the US must be supported and made universally available to AS researchers
2. Distribution of computational resources – individuals, departments, campuses, national centers – is not adequate, balanced, or seamless	Survey AS practitioners in education, research and operations to determine how they do their computing today	* Invest in computing infrastructure and capacity building at all levels (centers, campuses and departments) * Prepare for Grid computing
3. The proliferation of architectures and computing paradigms and the related lack of effective systems-level tools, e.g., compilers presents difficulties in tracking and using evolving technology in the AS; opinion sharply divided on best architecture for AS	Survey AS practitioners in education, research and operations to determine range of architectures and priorities for investment in system-level tools	* Support the development of a universal Geosciences computing environment that allows the seamless transport of codes from desktop to supercomputers and the Grid * Conduct in-depth survey to determine how best to invest scarce resources

Issues	Short-Term Recommendations	Long-Term Recommendations
<i>Software Issues</i>		
1. A proliferation of specialized software tools limits productivity with steep learning curves and a fog of ignorance of what is available	Develop a catalog of available tools and capabilities	Support the development of a Geosciences computing environment that provides for universal interoperability of tools for working with data and metadata
2. Many community codes suffer from poor performance and a high error rate due to slow adoption of the “open source” model of software development	Encourage academic recognition for open development activities	Support broader adoption of open source practices in the AS, especially for community codes, to ensure optimally effective code development
3. Students do not have adequate knowledge of operating environments used in AS research and education		Encourage training and educational programs that support students transitioning from Windows to AS operating environments
4. <u>Software life cycle</u> : * Software frameworks and community models suffer from the absence of sustained funding for training, support and maintenance * Large investments in community models has led to neglect of model development in the academic community	Fund applications groups and IT development groups jointly to motivate users to adopt new frameworks	* Integrate data and software tools at all stages for science and decision making through development of standards and frameworks  * Plan investments in software development with long-term support in mind * Invest in a wider variety of models targeted at various AS sub-disciplines
5. Steep learning curve of many tools makes them impractical in educational setting	Support development of training modules and documentation for most widely-used tools	Encourage partnerships with specialists in educational technology, cognitive science, and human-computer interaction
6. GIS software has the potential to offer a computational framework for integrating weather, climate, environmental, and socio-economic data	Encourage projects that seek greater levels of integration and interoperability between scientific data systems and GIS data sets and tools	Encourage participation by AS community in OpenGIS and other organizations to promote seamless integration of scientific information systems and GIS

## REFERENCES

1. [http://www.geo.nsf.gov/adgeo/geo2000/geo\\_2000\\_summary\\_report.htm](http://www.geo.nsf.gov/adgeo/geo2000/geo_2000_summary_report.htm)
2. Atkins Report.
3. Rew, R.K., and G.P. Davis, NetCDF: An Interface for Scientific Data Access, Computer Graphics and Applications, 76-82, IEEE, Los Alamitos, CA, 1990.
4. Message Passing Interface standard, <http://www.mcs.anl.gov/mpi>
5. Geophysical Fluid Dynamics Laboratory. Flexible Modeling System, <http://www.gfdl.gov/fms>
6. Office Note Series on Global Modeling and Data Assimilation, GEOS-3 Data Assimilation System, NASA Goddard Space Flight Center, DAO Office Note 97-06, 1997.
7. The CCSM Coupler – Version 6, <http://www.cesm.ucar.edu/models/cpl6>
8. Michalakes, J.G., S. Chen, J. Dudhia, L. Hart, J. Klemp, J Middlecoff, and W. Skamarock. Development of a Next Generation Regional Weather Research and Forecast Model. In W. Zwiefelhofer and Norbert Kreitz, editors, *Developments in Teracomputing: Proceedings of the Ninth ECMWF Workshop on the Use of High Performance Computing in Meteorology*, pages 269 - 276. World Scientific Press, 2001.
9. Earth System Modeling Framework, <http://www.esmf.ucar.edu>
10. Armstrong, R., D. Gannon, A. Geist, K. Keahey, S. Kohn, L. McInnes, S. Parker, and B. Smolinski. Toward a Common Component Architecture for High Performance Scientific Computing. In *Proceedings of the Conference on High Performance Distributed Computing*, 1999.
11. Jones, C., Patterns of Software Failure and Success. International Thomson Computer Press, 1996.
12. McConnell, S. Software Project Survival Guide, Microsoft Press, 1997.

## APPENDIX A – CyRDAS CHARGE, MEMBERSHIP AND PROCEDURES

### Charge (February 2003)

The ad hoc committee for Cyberinfrastructure Research and Development in the Atmospheric Sciences (CyRDAS) is charged with

#### A. Assessing:

1. The opportunities for advances in atmospheric science research that are made possible by current or anticipated advances in information technology and computer science
2. The opportunities for advances in science that might result from of collaborative research between atmospheric scientists and computer scientists
3. Ways in which cyberinfrastructure can contribute to formal and informal education in atmospheric science
4. The cyberinfrastructure needs of the NSF-funded atmospheric science research community

#### B. Recommending:

1. strategies for NSF that will help the academic research community to exploit the opportunities identified in the assessments A.1, A.2 and A.3 above.

#### C. Developing:

1. an implementation plan for a distributed cyberinfrastructure that will meet the needs of the academic atmospheric science research community and which includes the flexibility to grow smoothly as that research advances and CI needs grow.

### Membership

The membership of the committee was drawn from academic, government and non-profit departments and laboratories. Members represented many sub-disciplines of the atmospheric sciences, many regions of the country and several agencies within the Federal government. The list of members and their affiliations is as follows:

Dave Bader	Lawrence Livermore National Laboratory
Eric Barron	Penn State University
Joe Bredekamp	NASA
Greg Carmichael	Univ. of Iowa
Cecilia DeLuca	NCAR/UCAR
Kelvin Droegemeier	Univ. of Oklahoma
Tamas Gombosi	Univ. of Michigan
Jim Hansen	Mass. Institute of Technology
John Holt	Mass. Institute of Technology
James Kinter (chair)	COLA/IGES
Bill Matthaeus	Univ. of Delaware
Mary Marlino	DLESE/UCAR
Marla Meehl	NCAR/UCAR
Mohan Ramamurthy	Unidata/UCAR
Bob Wilhelmson	Univ. of Illinois
<i>National Science Foundation observers:</i>	
Stephen Meacham	Directorate for Geosciences
Kile Baker	Division of Atmospheric Sciences

## Procedures

In order to carry out its charge, the CyRDAS committee established the following principles:

- Seek broadest possible representation of the atmospheric sciences community
- Begin process for integrating CI planning for the atmospheric sciences into the larger planning processes for geosciences and environmental sciences
- Seek broadest possible dissemination of findings and recommendations

The NSF provided a grant for one-year to support the fact-finding and report generation activities.

The committee established a web site (hosted at UCAR) for information gathering and dissemination ([www.cyrdas.org](http://www.cyrdas.org)) as well as an internal web site for exchanges of information and documents among committee members. The public web site continues to be a “live” repository for updated information on CI in the atmospheric sciences.

Developed set of questions in four categories – social and cultural issues, data issues, computational infrastructure and capacity issues and software issues – to draw out opinions about how to develop and implement CI for the AS.

The committee held a series of regionally-distributed focus group meetings with atmospheric scientists from academic, government and private sector organizations. Email was sent to several thousand people whose email addresses were drawn from UCAR, Unidata, NSF and other CI-related mailing lists. The focus group meetings were organized in such a way as to provide the maximum number of people with access, either by close (driving) proximity to their organizations or by means of the Access Grid. The regional meetings were held in the following places on the indicated dates:

- Midwest region, NCSA, 13 October 2003
- Northeast region, ACCESS (DC), 15 October 2003
- Southwest region, SDSC, 21 October 2003
- Southeast region, Georgia Tech, 28 October 2003
- Northwest region, Univ. Washington, 30 October 2003

The focus groups were directed by members of the CyRDAS committee on site at each location. In addition, several of the locations were connected to the Access Grid for Internet-based video-teleconferencing participants. To stimulate and organize the focus group discussions, the CyRDAS committee developed several thematic questions, summarized in this report, as well as some general questions intended to draw out opinions on CI in the atmospheric sciences. The focus group general questions were:

- How can cyberinfrastructure lead to more rapid and more substantial progress in research and more efficient and effective education?
- What cyberinfrastructure barriers are impeding progress?
- What are the central issues that atmospheric scientists, educators and technologists consider most important, from their individual perspectives, to help them achieve what they hope to accomplish?

After the focus groups had met, the CyRDAS committee held weekly teleconferences for two months to evaluate the input received, develop a set of findings and make recommendations on how to go forward. This report documents the findings and recommendations.

## APPENDIX B - SOFTWARE PROJECTS

### Software Engineering

Perhaps the most useful categorization of software projects is by scale. Individuals involved with smaller software projects, like the new graduate student struggling with a computational task, are likely to benefit from basic training in software structure and development processes, and perhaps, depending on the project, high performance computing. Large software projects, whether focused on research or production software, require considerably more organization and staff expertise, in both software structure and process, to function effectively. For example, major climate and weather modeling efforts in the US and Europe, motivated by the need to manage a large set of contributors, have found it advantageous to create software Change Review Boards, to systematically identify, prioritize, and schedule changes to model software. Such formality would be out of place for an individual working on independent research. Software structure, in the form of a clearly defined architecture, is critical for a larger projects if the code is to be understandable and extensible. The new architectures emerging in the Earth sciences are being developed by computer scientists and software engineers with expertise in object-oriented design, component-based design, and multi-programming-language development. For an Earth science graduate student, exhaustive training in these areas would be both impractical and of dubious value.

Another useful categorization is whether the software project is essentially scientific or essentially technical in nature. In general, modeling efforts fall into the first category and CI efforts into the second. The expected background and training for a software engineer coding parameterizations for a model will be fundamentally different than that for a software engineer developing data management utilities. In the former case, the question may be how to develop computing skills in a meteorologist; in the latter, how to attract and assimilate software engineers from the commercial domain into the academic or national laboratory environment.

In the software engineering literature, two key factors are repeatedly called out as critical to the success of software projects.[e.g. 11,12] These are an adequate understanding of requirements and competent project management. The first key factor, understanding requirements, is crucial because early in a project a statement of requirements bridges the gap between developers of the software and future users of the software. For example, requirements analysis can encourage communication among a group of computer scientists who are interested in developing CI for modeling, but have limited knowledge about Earth science, and a group of Earth scientists who are interested in improving the capabilities of their model but are unfamiliar with the technology and terminology of computer science. Requirements can help to clear away jargon, identify the scope of the project, draw out assumptions, and set expectations.

The second key factor is management of software engineers and software projects. For larger projects, the collection of requirements, hiring and coordination of staff, oversight of software design, sequencing of development, and management of relationships with customers require attentive, dedicated management. The trend, in both infrastructure and modeling projects, has been to shift this responsibility away from project PIs and lead scientists to a full-time software project manager. As software projects grow ever more complex and critical for the advancement of atmospheric science, identifying and training software project managers is a critical activity area.

### Modeling Frameworks

Frameworks offer solutions to the problems of software redundancy, consistency, robustness, and performance portability that arise in atmospheric and Earth system modeling.

There are some well supported and widely used infrastructure libraries, such as netCDF[3] and MPI[4] but these represent a small fraction of the modeling infrastructure that may usefully be prefabricated. Many large modeling groups across the U.S. have developed their own infrastructure libraries, and a climate application that combines four independently developed geophysical components may, for example, find itself with four different I/O, time management, performance profiling and error handling systems. This leads to inconsistencies across components, and the redundancy in development and maintenance is a drain on resources. Further, since infrastructure is frequently a lower priority in atmospheric science projects, the utilities developed can often be improved in robustness, generality, and performance portability.

It is important to recognize that a framework can only offer widespread interoperability and software reuse if the framework is adopted throughout the community. The continued development of new frameworks and alternative implementations of established libraries and frameworks is essential to technical innovation. The atmospheric science community will benefit most by finding a productive balance between the promotion of community frameworks and the initiation of new, innovative infrastructure projects.

The current state of the art in frameworks for atmospheric and Earth system modeling is represented by efforts such as the Flexible Modeling System (FMS)[5] from GFDL, the NASA Goddard Earth Modeling System (GEMS)[6] the modular NCAR Community Climate System Model (CCSM) coupler (CPL6)[7] and the Weather Research and Forecast (WRF) Model Advanced Software Framework[8]. Each of these packages is a modular, portable, high-performance system for creating a variety of multi-component applications on parallel computing platforms.

Existing frameworks utilize modern data structures along with object-oriented and component-based design concepts. For example, in GEMS encapsulation via derived types and modules increases portability, hides details of the computing platform from the user, and results in a very simple and natural interface for creating coupled applications. In FMS, polymorphism and operator overloading simplify interfaces. Coupling systems such as CPL6 employ the concept of generic components to increase interoperability. The WRF code effectively and simply allows users to create data structures designed for a cluster architecture, and it has been carefully implemented for good performance on both vector supercomputers and microprocessor-based platforms.

These frameworks address the issues of software interoperability, reuse, and quality within a single organization or modeling project, but not across the broader Earth science community.

Perhaps the most significant trend in modeling frameworks has been the consolidation of smaller, institutional efforts into coordinated national efforts. Many of the groups who developed the first generation of modeling frameworks are participants in the NASA-funded Earth System Modeling Framework (ESMF) project [9]. With the ESMF, the creators of these successful smaller frameworks are collaboratively creating an entirely new system that extends the features of the original frameworks and provides enhanced cross-domain interoperability and reuse.

A second major and recently funded modeling infrastructure effort is the DOE's Common Component Architecture (CCA) project [10]. The CCA is developing a software specification that defines component interactions and interfaces, along with a set of toolkits. Although the CCA targets a broad set of scientific domains, climate is one of the project's initial focus areas.

Both the ESMF and the CCA are currently in a post-prototype stage of development. They share similar architectural concepts, and have demonstrated an ability to interoperate with each other in joint experiments. They also share the distinction of being implemented by technical and not scientific teams. As such they are able and expected to devote considerable resources to support, documentation, tutorials, and other ease of use and outreach issues. If successful, these

large frameworks have a potential to provide modeling infrastructure access and training to university researchers and students in a way that smaller efforts would find difficult.

The next generation of modeling framework projects is likely to involve the extension of existing frameworks into comprehensive, integrated modeling environments. Using a graphical interface, one can imagine that scientists will be able to research and access model components; configure selected components into an application; execute the application; archive data, perhaps remotely; locate and access observational data and results from other applications for comparison; analyze the data; and visualize results, all seamlessly. Each of these capabilities already exists in some form; the challenge is in coordinating the integration of a number of massive, distributed, multi-agency, software-intensive projects involving data and Grid services, programs for model comparison and evaluation, and modeling infrastructure, in a synergistic fashion.

## APPENDIX C – HIGH-END COMPUTING

According to the “Federal Plan for High-End Computing (Report of the High End Computing Revitalization Task Force, May 2004), “... pioneering computing capabilities have been a principal foundation of the Nation’s technological and economic strength.” The word “supercomputer” means the fastest available computing system at any given time. However, the simplicity of the term prior to the 1990s has given way to a fairly complex high-end computing environment that relies on commercial off-the-shelf (COTS) processors arrayed in massively parallel, tightly coupled systems. The widespread adoption of parallel processing systems made up of many hundreds or thousands of commodity chips has made it possible to consider intermediate scale computing facilities that are more distributed, i.e., closer to researchers, than the centralized facilities normally associated with supercomputing.

### Parallel Processing

Parallel computing hardware has been taking the place of vector supercomputers in the U.S. for several years. A spectrum of computer architectures with a wide range of costs is now available in the marketplace from relatively inexpensive, small-scale computing nodes with a small number (2-4) of fast processors that share memory, to more expensive larger shared-memory systems that have up to 1,000 or more processors in a single node, to hybrid architectures with clusters of small-scale shared-memory nodes arranged on a high-speed interconnecting network that can distribute a computation (and the memory requirement) over the nodes. Clusters are available over a fairly wide range of costs. Outside the U.S., vector supercomputers and clusters of vector supercomputers are still available, the fastest of which are substantially higher performance than anything currently in use in the U.S.

Programming in a parallel computing environment has introduced several challenges that are slowing progress in numerical modeling. A very large fraction of the programs being used for simulation are serial codes that employ a single processor. Converting such codes to run in a multi-task or multi-thread mode requires a very large investment of effort and resources. An even larger investment is frequently required to achieve a reasonable degree of parallelization in a distributed processing environment, due to the large overhead associated with inter-processor communication. Typically, processors pass small parcels of data, called messages, to each other across the network, so parameters such as network bandwidth and latency can have a large effect on performance. Often, once a code has been converted or rewritten for a multi-processor computing environment, its performance is inadequate, and substantial further investment of resources is required to optimize the code for a given computing architecture or platform.

A critical element of the parallel programming paradigm is scalability – that is the rate at which additional processors applied to a problem can reduce the wall-clock time of solving that problem as the problem size is increased. A further complication frequently encountered in geophysical codes is the necessity to balance the load across processors or computational nodes. For example, a numerical weather prediction model must treat areas with clouds quite differently than it handles cloud-free areas, with a large differential in the computational load. In a parallel computing environment, efficiency is significantly degraded if processing load is not well balanced across the system. Finally, the fact that most high-end computing platforms are combinations of shared-memory nodes in communication-intensive networks, the algorithms that are employed on such systems must be coded in such a way as to exploit both the advantages of multiple processors sharing memory (multi-tasking) and the advantages of message passing. Such hybrid algorithms can be extremely complex and highly machine-dependent.

It should be noted that the recently released Federal High-End Computing Plan recommends in favor of an aggressive research and development effort that will lead to the next

generation of high-end computing that will overcome the issues and bottlenecks currently being experienced.

### Centralized vs. Distributing Computing

There are several advantages to centralized computing. First, the budgets can be sizable enough to provide for the acquisition of tightly coupled computing resources that are suitable for the largest computational problems – often referred to as computing capability. Second, the total cost of ownership of computing resources, including acquisition, maintenance, and support both for the hardware and software and for the users, can, in principle, be better managed and resources can be used more efficiently. Third, the peripheral resources that high-end computing requires – online, near-online, and archival services for data; local area and wide area networking support; algorithm development; testing of future hardware and software systems; user support etc. – are more likely to be assembled in a center than in any one of many distributed computing sites.

Centralized computing has its disadvantages too. First, the very act of establishing a center usually involves institutional commitments and constructions that tend to be self-perpetuating. The flexibility that is needed in the rapidly evolving computing environment is difficult to maintain in a center. Second, because the human element of computational science is not necessarily scalable, centralized computing is not necessarily the best way to provide computing capacity for the bulk of the atmospheric sciences. For example, the efficient management of a large system with thousands of users is a very difficult problem with so many constraints that it may be ill-posed. As a result, the individual wait times for relatively small amounts of resources can be much larger than they would be in a distributed computing model. Also for example, the support that is required by the thousands of users varies immensely across the user population, and it frequently occurs that support resources are concentrated on either those with very elementary support requirements or the so-called “power-users” who can dominate any facility.

Distributed computing has several attractive features that have inspired many atmospheric sciences departments to invest in such facilities. First, departments gain control and reduce the administrative overhead of managing computing resources by placing those resources within the department. Second, an obvious benefit to users is the reduction in turnaround time of computational tasks – competing with tens of others instead of thousands of others substantially reduces the waiting time that individuals experience. Third, having computing resources inside the department enables the integration of parallel programming practices into the atmospheric sciences classrooms.

The disadvantages of distributed computing are very much the inverse of the advantages of centralized computing. First, there is a loss of efficiency in the distributed model, because resources are more easily wasted when less is at stake. That is, the incentive to invest in efficiently time-sharing a given system increases with the cost of the system itself. Second, even though acquisition costs are becoming quite reasonable, the total cost of ownership of a distributed computing facility may be too large for a single department to bear. Third, the investment required to parallelize the serial codes in a given department and to provide support for distributed computing at the departmental level may be prohibitively large.

The Federal High-End Computing Plan includes a multi-agency strategy for making resources available to government and government-sponsored researchers.

## APPENDIX D - EXAMPLES OF COMPUTATIONAL ISSUES

### Numerical Weather Prediction

Improving forecasts of severe weather events such as hurricanes, tornadoes, and hailstorms has long been a focus of weather research because of the need to reduce the loss of life and property. Accurate simulations of these mesoscale phenomena will assist in better fundamental understanding their structures and the limits to predictability of such events.

Nearly all current numerical weather prediction models utilize limited-area domains and grid spacing of order 10 km. Unresolved phenomena such as convective clouds (at the coarser end of the resolution range), turbulence, and other sub-grid scale processes are parameterized. Additionally, grid nesting is frequently employed to represent the structure of the large-scale environment while also capturing the fine-scale structures of localized phenomena such as thunderstorms.

Data assimilation is an extremely important component of a complete weather prediction system. A variety of assimilation techniques exist, ranging from simple data insertion to complex and computationally expensive four-dimensional variational techniques. The latter, based upon optimal control theory, involve fitting a model to known observations subject to certain constraints. This includes both the development of the software for assimilating observations and the design of observing networks (what observations should be taken where and when). Finally, since it is impossible to estimate with certainty the state of the atmosphere at any point in time, an “ensemble” of forecasts can be averaged to produce the mean forecast as well as statistics about its likely quality. The latter are important aids to decision-making.

The next major advance in atmospheric prediction will come with the application of storm-resolving models (grid spacing of 1 km or less) that are initialized with observations of comparable resolution, and that are operated not on fixed schedules or in fixed configurations, as is now the case for coarser-grid operational systems, but in a manner that responds rapidly to the weather itself and to decision-driven inputs from the user. For example, assume that the finest storm-resolving grid of a numerical weather prediction model will be run over an area 200 km x 200 km. Based upon experience, the most sophisticated and computationally intensive storm-scale variational data assimilation system might require computing power equivalent to 25 short-term forecasts of the full dynamic model in order to obtain an “optimal” estimate of the initial condition. Accounting for data assimilation and 20-member ensemble forecasts, this forecast system would today run approximately 10 times slower than real time on a typical 1000-processor system currently available commercially.

Accurate prediction of hurricane tracks may require resolving the flow both within and around the storm. Since the spatial scales in these two regions differ substantially, uniform resolution is inherently inefficient: the grid needs to be refined near the storm but coarser resolution is sufficient outside the storm’s immediate vicinity. Adaptive multi-grid schemes, which automatically refines the mesh around the storm as it moves, are now available that make this possible.

To be useful, forecasts should be produced 10 times faster than the weather evolves. Thus, a total increase by a factor of 100 in sustained performance compared to what is available today on ~1000 processors would be required. The ability to produce such forecasts for several parts of the country or world simultaneously would introduce an additional multiplier.

### Climate Modeling

Research into climate system dynamics will aid in developing a fundamental understanding of the Earth's climate system and the capacity to predict the climate on seasonal, interannual, decadal and climate change time scales. Climate models, consisting of sub-models representing the oceans, atmosphere, land surface, and sea-ice, all coupled together, must have the capability to predict

- the likelihood of extreme climate fluctuations and abrupt climate change
- the response of climate and ecosystems to changes in climate forcing over the coming decades
- the availability of natural water resources, especially the likelihood, intensity, and duration of droughts
- the influence of different land-use scenarios on regional climate
- the distribution and intensity of atmospheric dust
- the impact of emission reduction strategies on regional climate scales
- the viability of different carbon sequestration scenarios
- the frequency of severe weather events such as hurricanes and droughts

Such research will also allow an overall reduction in the uncertainty in climate change projections. The information extracted from these types of predictions can be used by the policy community, economists, and land use planners and agricultural experts to develop adaptation strategies that permit the minimization of economic disruption from adverse climate fluctuations.

At seasonal to interannual time scales, models of moderate resolution can be used, and assimilation systems are required because the memory of initial conditions, particularly over land and oceans, is retained. Assimilation brings observations, typically sampled at different times and disparate spatial locations, into the model forecast framework to create a dynamically consistent, unified representation of the atmosphere and upper ocean – consistent with the model's dynamics – to improve subsequent forecasts. The goals for seasonal to interannual research include improving prediction skill on the six-month timescale and eventually provide routine 6-12 month seasonal predictions within the next ten years.

To be useful for S-I applications, single-image model performance must be roughly 1,000 simulated days per wall-clock day (1000 d/d), and since climate forecasts need to be based on ensembles of runs, typical systems run 8 to 10 concurrent coupled images (or 15 to 20 uncoupled images) now. It is assumed that, as computer power increases, the size of the calculation and the number of concurrent images will also increase while roughly maintaining these performance goals (1,000 d/d single image, 20,000 – 30,000 d/d aggregate throughput).

During the next decade, the magnitude of the calculation will increase due to the greater model complexity and increases in resolution. Performance projections for a prototype calculation – a single atmospheric model with uniformly increasing global horizontal resolution – suggest that the next decade will provide computing power sufficient to resolve mesoscale phenomena such as convective complexes and gravity wave-trains and non-hydrostatic dynamics. The transition between hydrostatic and non-hydrostatic dynamics occurs at about 12 km, while at 25 km we begin to resolve mesoscale phenomena, a transition that may result in qualitative improvements in results. For these reasons, a realistic goal for S-I for the next decade is to begin to resolve the mesoscale, at or near the hydrostatic limit.

Due primarily to limitations in computer resources, current climate models used for decadal and longer time scales and climate change simulation have much coarser resolution in which physically important processes such as cloud dynamics and ocean mesoscale dynamics are parameterized instead of being resolved. Although the fruit of considerable research, the best parameterizations still fail to do justice to the physics that they are trying to represent and produce results that are at variance with observations. Similarly, low resolution is recognized as a significant problem in the land surface component. It is generally believed that resolving these processes will be the basis of the next generation of climate modeling research.

The numerical models currently used for climate research at decadal-centennial scales are characterized by modest horizontal resolution of about 100 km and use about 50 vertical levels throughout the depth of the atmosphere and ocean. Land and ice models use comparable horizontal resolutions with fewer vertical levels. Interactive ecosystem models are just beginning to be integrated into the coupled physical models of the Earth system. Solving some of the basic problems in modeling the ocean, land surface and atmosphere that are important for making the

sorts of predictions listed above will require increases in computational power of 1000 to 30,000 current capabilities.

### Space Weather

“Space weather” refers to conditions on the Sun and in the solar wind, magnetosphere, ionosphere, and thermosphere that can influence the performance and reliability of space-borne and ground-based technological systems, and can affect human life or health.

In space plasmas very small-scale processes affect the large-scale behavior of the system. For example, in the Earth's magnetosphere one of the most important dynamic processes is magnetic reconnection, in which processes that occur on scales less than 10-100 km play a central role. Auroral arcs can have a thickness of tens or hundreds of meters. Of course, we know that even smaller scale atomic and nuclear processes have effects throughout the universe. However, most atomic and nuclear processes can be parameterized in terms of cross sections and recombination coefficients. Ways have not yet been found to similarly parameterize small-scale plasma processes. One of the great challenges for solar-terrestrial physics is learning how to understand these small-scale, non-fluid processes operating within a much larger physical system.

We need to address the national need for production-ready high-performance computational tools that model, analyze, and interpret space environment observational data. There is a more specific objective to meet the national need to produce first-principles, physics-based predictive (i.e. faster than real time) model of the Sun-Earth system. We will reach this goal by engineering existing codes to interoperate in common frameworks while simultaneously achieving high performance. The combination of interoperating models and the underlying software framework will serve the entire atmospheric science community, and will use observational data to test and validate the models.

This very challenging computational task will involve all aspects of modern high-end scientific computing. Scalability, single-node performance, latency, and load balancing are all critical elements. In addition, the range of dynamic scales requires the use of efficient grid adaptation technologies, as well as interfacing regions described by different physical approximations. In short, space weather modeling requires an adaptive physics approach with underlying adaptive grid and efficient parallel computing technologies.

## APPENDIX E – COMMUNITY MODELS

### 1. Need, Definition and Examples

We define herein a “community model” in very broad terms as a model or, more appropriately, a system that: a) is developed and continually enhanced with significant and broad-based involvement from the user community (which may include non-disciplinary experts); b) is used by a large community that typically extends well beyond the atmospheric sciences, e.g., to the hydrologic, solid earth, bio, environmental, and even medical sciences; c) receives stable, long-term funding from multiple agencies having various stakes in the resource; and d) is managed by one or more organizations having the capability to coordinate a broad user base, facilitate training, maintain a complex software infrastructure and associated data resources, develop and make available documentation, and ensure that the overall system operates at the leading edge of scientific and computational capability.

Based upon this definition, contemporary examples of community models/systems in the atmospheric and ocean sciences include the following:

- Models-3/Community Multiscale Air Quality (CMAQ)
- Regional Ocean Modeling System (ROMS)
- Community Climate System Model (CCSM) and Community Climate Model (CCM)
- NCAR/Penn State Mesoscale Model (MM5) and the emerging Weather Research and Forecast (WRF) Model

During the past several years, commercial modeling packages, such as STELLA, have emerged and provide a wide variety of simulative capabilities across many disciplines for research and education. Other examples of community models that eventually were commercialized include Gaussian, Charm, and Amber (chemistry, molecular dynamics), Fluent (fluid dynamics in complex geometries), AIPS++ (astronomy), and Abaqus (structural mechanics).

### 2. Management

The growth in community models/systems during the past two decades has fostered a variety of management methodologies. For example, management of the NCAR CCSM includes a Scientific Steering Committee, Advisory Board, several Working Groups, and a Change Review Board. Annual workshops are held to ensure continued engagement of all stakeholders, and a well-maintained web site provides reliable access to the very latest resources. So important is the management of community systems that the Environmental Protection Agency established a Community Modeling and Analysis System (CMAS) center to foster the development and distribution of its community resources.

It is extremely important that adequate attention, and funding, be directed toward community model/system management in cyberinfrastructure planning. To date, such support has been woefully inadequate, even for systems (e.g., MM5) that are used worldwide by thousands of scientists and students. One of the most important value propositions of community models is the ability to incorporate new capabilities developed by a wide variety of interests, yet doing so – which includes merging, validating, releasing, and maintaining code – is extremely time-intensive and delicate. Furthermore, in today’s parallel computing environments, code that runs “correctly” on one machine and its compiler may produce vastly different results when used with a different compiler on the same machine, or with the same compiler on a different architecture. Such issues can thwart the effectiveness of a community model/system, and thus we recommend that adequate resources be directed toward code testing, documentation, and user support. Further, we recommend that NSF/GEO closely examine existing modalities for managing community models and establish a set of best practices for the future.

### 3. Multi-Disciplinary Considerations and Standards

The atmospheric sciences community has performed admirably in involving other disciplines in its community modeling efforts. For example, the CCSM includes representation from geology, hydrology, ocean, biology, and ecology, as well as computer and computational sciences. Furthermore, atmospheric science is among the leaders in coupling models and developing standards for model (e.g., the Earth System Modeling Framework) and data (e.g., Earth Science Markup Language) structure and interoperability. Nevertheless, as noted in the recent NSF Cyberinfrastructure Advisory Committee report, much of the exciting research in the future will occur at the boundaries between traditional disciplines, and thus atmospheric science needs to position itself to take full advantages of emerging opportunities.

For example, considerable attention needs to be given to the development of coupled or linked models that are even broader in scope and that include dynamic representations of human activity and their associated impacts on built infrastructure, ecosystems, and physiography. Furthermore, community models must be able to operate in highly heterogeneous computing environments, be adaptively responsive to and able to control remote observations, be fault tolerant, and be able to accommodate data from other disciplines. Architecture plans for such models must consider these and other issues, including code documentation using hyperlinks, parallel constructs that can be adapted to multiple systems, and efficient strategies for real time, on-demand operation involving massive I/O. Further, such models must be able to transport seamlessly from the largest parallel computers in existence to individual desktop workstations.

Finally, standards for data interchange and model interoperability must be finalized, in collaboration with other disciplines, to maximize the effectiveness of model coupling. We do not recommend a universal standard – which in practice likely would be impossible, but rather a set of standards within the atmospheric sciences that can interoperate effectively with those being developed within other disciplines.

### 4. Middleware and Effective Linkages with the Operational Community

Owing to the availability of powerful, affordable computing resources at the campus, department, and individual investigator levels, along with real time data delivered via high bandwidth networks, numerous community models are being run on a continuous basis, in near real time, to produce experimental weather and climate forecasts. Many of these efforts involve collaborations between academia and the operational community, e.g., results from a community weather model run locally at a university meteorology department and used jointly by local National Weather Service forecasters and students in a forecasting class. Such efforts are extremely valuable as they bring a sense of operational functionality to the classroom, and also provide to practitioners various experimental products that may not be formally available for several years.

Unfortunately, most community models/systems do not provide the middleware, or orchestration software, needed to manage the complexities of real time operation, and in particular the assimilation of streaming observational data with its frequent outages, quality control issues, and management overhead. As a result, of the dozens of academic institutions in the U.S. that presently operate mesoscale forecast models in near real time, for example, usually at grid spacings substantially finer than are available from the National Weather Service operational models, only two or three have developed middleware that allows them to assimilate local and regional observations – and thus to justify the use of finer grids. Such middleware tends to be very application- and platform-specific and thus cannot effectively be shared.

Because academic institutions increasingly are unable to support the type of professional needed to develop and maintain such middleware, future efforts in community models must look beyond the models themselves to the orchestration capabilities (often referred to as “wrappers”) needed to operate the system in increasingly sophisticated ways – indeed, in a manner similar to that within a fully operational environment. This holds true for weather, climate, and coupled

models, and will allow academia to conduct more realistic and useful experiments, educate students more effectively, and transfer results more quickly to the operational sector.

Likewise, community modeling efforts should increase their interactions with the operational community. The WRF is perhaps the best recent example, and one of its principal goals is to bridge the cultural/philosophical gap that has long existed between the academic research and formal operational communities.

#### 5. Software Availability and Commercialization

All of the community atmospheric and ocean models cited earlier reside in the public domain and may be used free of charge and without restriction. In light of the trend toward intellectual property commercialization by both academia and the Federal government, great care must be taken to ensure that no restrictions exist on community model components developed within either sector, or jointly. It is quite possible, and in fact likely, that a particular academic institution, funded by a Federal agency to develop a specific component of a community model, might wish to license that component for commercial application and thus restrict its use. Such ownership and licensing issues must carefully be evaluated in the context of community models.

As the economic value of weather and climate information continues to grow, and as models continue to improve, so does the prospect that a private company will seek to commercialize a community system that resides in the public domain. Such activities should not be discouraged, and in fact represent an opportunity for investment in research by the private sector, provided that an equitable relationship can be constructed that in no manner restricts the research, education, or operational communities.

#### 6. Prototype Applications

Community models already are providing inestimable value in scientific research, education, operations, and policy making, and the future holds even greater promise. Areas in which community models are envisioned to make a substantial impact during the next decade – in terms of scientific advances, societal impacts, and policy – include energy, transportation (water, surface and air), space flight, agriculture, urban and regional planning, communication and power infrastructure, national defense/security, air and water quality/stewardship, species and habitat preservation, and human health.

We envision a greatly expanded application of community models as cross-disciplinary coupled systems to study challenging problems such as the health of coastal ecosystems, bio- and anthropogenic impacts on climate variability, and space weather impacts on surface and satellite communication and tracking systems. As a largely new direction, the linking of economic models into physically-based community models/systems will greatly expand the utility of both, and provide for a deeper understanding of the mutual impacts of the atmosphere and society.