Adv. Geosci., 8, 69–78, 2006
www.adv-geosci.net/8/69/2006/

Advances in Geosciences

A new generation of cyberinfrastructure and data services for earth system science education and research

M. K. Ramamurthy

Unidata, University Corporation for Atmospheric Research, Boulder, Colorado, USA

Received: 15 November 2005 – Revised: 30 January 2006 – Accepted: 14 March 2006 – Published: 6 June 2006

Abstract. A revolution is underway in the role played by cyberinfrastructure and modern data services in the conduct of research and education. We live in an era of an unprecedented data volume from diverse sources, multidisciplinary analysis and synthesis, and active, learner-centered education emphasis. Complex environmental problems such as global change and water cycle transcend disciplinary and geographic boundaries, and their solution requires integrated earth system science approaches. Contemporary education strategies recommend adopting an Earth system science approach for teaching the geosciences, employing pedagogical techniques such as enquiry-based learning. The resulting transformation in geoscience education and research creates new opportunities for advancement and poses many challenges. The success of the scientific enterprise depends heavily on the availability of a state-of-the-art, robust, and flexible cyberinfrastructure, and on the timely access to quality data, products, and tools to process, manage, analyze, integrate, publish, and visualize those data.

Concomittantly, rapid advances in computing, communication, and information technologies have revolutionized the provision and use of data, tools and services. The profound consequences of Moore's Law and the explosive growth of the Internet are well known. On the other hand, how other technological trends have shaped the development of data services is less well understood. For example, the advent of digital libraries, web services, open standards and protocols have been important factors in shaping a new generation of cyberinfrastructure for solving key scientific and educational problems.

This paper presents a broad overview of these issues, along with a survey of key information technology trends, and discuses how those trends are enabling new approaches to applying data services for solving geoscientific problems.

1 Introduction

Cyberinfrastructure and data services are transforming the conduct of research and education in the geosciences. We live in an era of an unprecedented data volume from diverse sources, multidisciplinary analysis and synthesis, and active, learner-centered education emphasis. For example, current day weather and coupled climate system prediction models and a new generation of remote-sensing systems like hyperspectral satellite instruments and rapid scan, phased-array radars are capable of generating massive amounts of data each day. Complex environmental problems such as global change and water cycle transcend disciplinary and geographic boundaries, and it is widely recognized that their solution requires integrated earth system science approaches. Contemporary education strategies recommend adopting an Earth system science approach for teaching the geosciences, employing new pedagogical techniques such as enquiry-based learning and hands-on activities. The resulting transformation in today's education and research enterprise creates new opportunities for advancement, but also many new challenges. For example, the success of this enterprise depends heavily on the availability of a state-of-the-art, robust, flexible, and scalable cyberinfrastructure, and on the timely, open and easy access to quality data, products, and tools to process, manage, analyze, integrate, publish, and visualize those data.

Concomittantly, rapid advances in computing, communication, and information technologies have also revolutionized the provision and use of data, tools and services in education and research. The profound consequences of Moore's Law, an empirical observation that the number of transistors on a chip doubles every 18 months, in the information revolution are well known. Similarly, the explosive growth in the use of the Internet in education and research, largely due to the advent of the World Wide Web, is also well documented. On the other hand, how other technological, social and cultural trends have shaped the development of data services is

Correspondence to: M. K. Ramamurthy
(mohan@ucar.edu)

Published by Copernicus GmbH on behalf of the European Geosciences Union.

somewhat less well understood. For example, the advent of digital libraries, web services, grid computing, open standards, protocols and frameworks, open-source models for software, and community models have been important factors, both individually and collectively, in shaping the use of a new generation of modern, end-to-end cyberinfrastructure for solving some of the most challenging scientific and educational problems.

The purpose of this survey paper is to present a broad overview of these and related issues largely from the author's perspective, along with a brief discussion of the how the above changes are enabling new approaches to applying data services for solving integrative, multidisciplinary geoscientific problems. To that end, the paper focuses on documenting some of the changes in the conduct of geoscience and science education, highlighting the revolution in cyberinfrastructure and documenting how the resulting technological advances and approaches are leading to an evolution from once proprietary and centralized data systems to open, distributed and standards-based data services that facilitate easier data integration and greater interoperability.

The layout of the paper is as follows. In Sect. 2, we present a few key scientific and education drivers, while Sect. 3 documents important information technology trends that have shaped new approaches to providing data services in the geosciences. Specific issues related to data services, including ideal data service attributes, data categories, and data analysis and integration methods are presented in Sect. 4. Section 5 offers a few concluding remarks on how today's cyberinfrastructure and data services are reshaping the science and education landscape.

2 Key drivers

Data, information, and embedded knowledge are central to the advancement of science and education, as articulated in NSF Geosciences Beyond 2000: Understanding and Predicting Earth's Environment and Habitability (NSF, 2000). The aforementioned report recognizes that 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 and access, distribution, analysis and management. Conversely, contemporary data services need to be firmly grounded in not only scientific and education drivers, and community needs, but are also greatly influenced by the myriad technological and sociological trends.

The following sections describe the key drivers and trends that have transformed data provision and access from centralized systems that were once based on proprietary architectures to modern, distributed data services. In the process, the new generation of data services has also reshaped their use in research and education in new and innovate ways.

2.1 Science driver

Numerous national and international reports underscore the importance of interdisciplinary environmental research and education. Among them are *Grand Challenges in Environmental Science* (NRC, 2001) and *Complex Environmental Systems: Synthesis for Earth, Life, and Society in the 21st Century* (NSF, 2003). The NRC report points to a growing recognition that "natural systems – ecosystems, oceans, drainage basins, including agricultural systems, the atmosphere, and so on – are not divided along disciplinary lines." Two of the grand challenges identified by the NRC, biogeochemical cycles and climate variability, depend heavily on integration of data from several disciplines. Another excellent example is hydrologic forecasting, one of the four challenges prioritized as deserving immediate investment.

According to the former NSF director Rita Colwell (Colwell, 1998), "Interdisciplinary connections are absolutely fundamental. They are synapses in this new capability to look over and beyond the horizon. Interfaces of the sciences are where the excitement will be the most intense…." For example, studies on societal impact of and emergency management during hurricane-related flooding involve integrating data from atmospheric sciences, oceanography, hydrology, geology, geography, and social sciences with data bases in the social sciences.

The NSF decadal plan for Environmental Research and Education (ERE) also echoes the need for improving our understanding of the natural and human processes that govern quantity, quality and availability of freshwater resources in the world. While recent advances in remote sensing, combined with a new generation of coupled models, are driving a new revolution in hydrometeorological predictions, future research and education in this area will require finding and integrating observational and model data from the oceans, the atmosphere, the cryosphere, and the lithosphere, crossing the traditional disciplinary boundaries.

Similar multidisciplinary needs are emerging to solve certain disaster/crisis management problems. Two highly topical examples are fire-weather forecasting and environmental modeling for homeland security. In homeland security, for instance, there is a need to forecast the dispersal of hazardous radioactive, biological, and chemical materials that may be released (accidentally or deliberately by terrorism) into the atmosphere. For the latter scenario, detailed, four-dimensional information on transport and dispersion of hazardous materials through the atmosphere, and their deposition to the ground are needed at a resolution of individual community scales. Moreover, this information needs to be linked in real-time to databases of population, evacuation routes, medical facilities, and so on to predict the consequences of various release scenarios (e.g., number of people may be exposed to or will be injured by potentially dangerous concentrations of those materials).

In addition to identifying national priorities and computational grand challenges in the sciences, many of the NSF and NAS reports cited above have also documented infrastructure needs, including comprehensive data collection, management and archival systems and new methods of data mining and knowledge extraction. For example, the NSF ERE Advisory Committee calls for building infrastructure and technical capacity with a new generation of cyberinfrastructure "to support local and global research and to disseminate information to a diverse set of users including environmental professionals, the public, and decision makers at all levels." Toward building the cyberinfrastructure, the ERE agenda foresees the need for a comprehensive suite of data services that will facilitate synthesis of datasets from diverse fields and sources, information in digital libraries, data networks, and web-based materials so that they can serve as essential tools for educators, students, scientists, policy-makers and the general public. Similar needs for web-based real-time and archived data services, including digital library integration and fusion of scientific information systems (SIS) with geographic information systems (GIS), were expressed at the NSF-sponsored Workshop on Cyberinfrastructure for Environmental Research and Education (CIERE, 2003).

Growing numbers of universities are engaged in real-time modeling activities, and this number is expected to increase as advances in computing and communication technologies facilitate local atmospheric modeling. A new generation of models (e.g., the Weather Research and Forecasting model, (Michalakes et al., 2001)) can predict weather on the sub 1-km scale, with the potential to address community-scale concerns. Providing initial and boundary condition data along with analysis and visualization tools for these efforts requires an extensive cyberinfrastructure.

The recent decades have also been marked by a revolution in our ability to survey, probe, map, and profile our global environment. For instance, a plethora of instruments and digital sensors mounted on geostationary and polar orbiting satellites scan vast areas of the earth's surface round the clock. With their powerful ability to continuously and remotely monitor the global environment, observations from satellite platforms are increasingly replacing in-situ surface and upper air observations. Today, dozens of satellites are rapidly scanning and measuring the global environment and in the process generating an ever-expanding range of geoscience data to help us manage and solve some of the most vexing and complex multidisciplinary problems of the society.

This revolution in remote sensing technology and techniques and their many geoscience applications have had a profound impact on geoscience operations. At the same time, the complexity and explosive growth in the volume of remotely sensed has also transformed the provision and use of data from remote sensing platforms such as satellites, radars, and lidars.

Modern environmental studies rely on diverse datasets, requiring tools to *find* and use the data. The data discovery process has become an important dimension of the scientific method, complementing theory, experimentation, and simulation as the tools of the trade.

2.2 Education driver

Challenges facing science education have been well articulated in a number of documents (e.g., *Shaping the Future* (AGU, 1997) and *Geoscience Education: A Recommended Strategy* (NSF, 1997)). They recommend adopting an Earth system science (ESS) approach for teaching the geosciences, integrating research experiences into curricula, employing contemporary pedagogies, and making appropriate use of educational technologies. Science education should also be about teaching students the language of science and providing students with opportunities to engage in scientific inquiry and investigation (Lemke, 1990).

Shaping the Future also calls for an inquiry-based approach to science education. For example, hands-on, learner-centered education in meteorology depends on the availability of meteorological data and analysis and display tools of high quality. By supplying these data and tools, programs like Unidata have been instrumental in transforming learning in the atmospheric sciences. Digital libraries (exemplified by efforts like the National Science Digital Library (NSDL) and the Digital Library for Earth System Education (DLESE)) augment web-based learning resources with high-quality data resources that can be embedded in interactive educational materials. Internet-based tools also open data access for faculty and students at small colleges where little system administration support is available for the installation of advanced data systems and applications. Engaging students with real-world data is a powerful tool not only for motivating students but also helping them learn both scientific content and principles and the processes of inquiry that are at the heart of science (Manduca, 2002). Earth science education is uniquely suited to drawing connections between the dynamic earth system and important societal issues and making science relevant to students. Recent catastrophic events like the 2004 Indian Ocean tsunami, Hurricane Katrina, and the October 2005 earth-quake in Northern Pakistan are three stark examples that drive home this point. These events also heavily underscore the importance of multidisciplinary integration and synthesis of data from the various Earth science disciplines. Working with such real-world events and actual data can place learning in a context that is both exciting and relevant. Another example is providing connections between classroom instruction and students' experience with their local environment (e.g., diurnal temperature changes and seasons), major weather events (e.g., tornadoes, hurricanes, blizzards) and climate events (e.g., global warming).

In essence, significant strides in advancing Earth science education can be made by incorporating new teaching techniques, active learning strategies, information technology, and integrating real-world Earth and space science data into our curriculum. To accomplish these objectives, students will need to have opportunities for genuine inquiry and hands-on experience, so that the excitement of discovery is infused into all courses while students gain experience in the process of science. A critical component of successful scientific inquiry includes learning how to collect, process, analyze, and integrate data. Innovative data services that promote this perspective on student learning are needed and should be integrated into Earth science education at all levels.

The richness of students' exploration and experience depends, among other things, on the quality of the data available and the tools and technology they use. To that end, cyberinfrastructure provided by organizations like Unidata allows students to access the very databases and tools that are used by the scientific and operational communities, and provides an important pathway toward the pursuit of the long-sought goal of the National Science Foundation to integrate research and education. Distributed computing, data access and collaboration are rapidly becoming the de facto means for learning and doing science, leading to the pervasive use of the World Wide Web in every day life of a scientist, an educator, or a student. This reliance on the Internet and other technologies and applications is only expected to increase greatly in the future.

3 Information technology trends

Computers and information technologies are now playing a central role in this complex and ever-changing world in which we live and work, with the World Wide Web reshaping almost every aspect of our work, including education, research, and commerce. Computing, communications and information technology trends of recent years have not only had a democratizing effect on daily life, but they have also changed the very nature of data services for education and research. For example, below is a partial list of key technologies and trends that have enabled a new generation of end-to-end data services in the scientific community:

- Internet
- Commodity microprocessors, enabled by Moore's Law
- World Wide Web
- Open source model
- Object-oriented programming
- Open standards, frameworks, and conventions
- Extensible Markup Language (XML)
- Web services
- Digital libraries
- Collaboratories
- Grid computing
- Data portals and federated, distributed servers
- Geographic Information Systems
- Ontologies and Semantic web
- Data mining and knowledge discovery

This section highlights a few of the above key information technology (IT) advances and trends that have revolutionized the provision and use of data in the geosciences. The highlights selected for further discussion are meant neither to provide a comprehensive overview of all key advances, nor are the specific examples cited sole exemplars of their myriad implementations. Taken together, however, the above technologies have enhanced the ability of data providers to better serve their communities, lower the costs for the users, and allowed a greater participation in the data activities in a new networked world.

The introduction of microprocessor based computer systems in the 1980s, combined with the increased connectivity of college campuses to the Internet, led to a transition from large scale, mainframe based technologies to low-cost distributed systems, making it possible for widespread access to and use of scientific data. The wiring of universities for Internet connectivity was a prerequisite for receiving data via, for example, the Unidata Internet Data Distribution system and the Local Data Manager, which use TCP/IP communication standards for data transport.

The advent of the World Wide Web (or simply the Web) in the 1990s brought about a revolution in information services. It was directly responsible not only for the explosive growth of the Internet and increasing its numbers of users, it also accommodated the ability to provide interactive, remote services. In the process, the Web radically transformed the sharing of data and information and resulted in greater use of communication infrastructures to create and store information and then to deliver it from providers to end users. The Web also brought with it a massive proliferation of online educational materials, many of them based around extensive use of interactive services. Services and tools were created to help one communicate, search for information and data, and make information and data available on the Internet. In the process, library services evolved from local traditional collections to global resources provided on demand via the Web, ushering in the era of digital libraries.

The 1995 NSF-sponsored Digital Libraries Workshop entitled "Interoperability, Scaling, and the Digital Library Research Agenda" defined digital libraries as: "An organized collection of multimedia data with information management methods that represent the data as information and knowledge."

Even though efficient retrieval of information is arguably the most important role of digital libraries, a potentially even more valuable contribution of digital libraries is their ability to preserve, catalog, and curate information, extend discourse, build communities that provide richer contexts for people to interact with information and each other, all toward the creation of new knowledge. According to Griffin (1998), the real value of digital libraries may ultimately prove to be their ability to "alter the way individuals, groups, organizations etc, behave, communicate, collaborate, and conduct business." In essence, digital libraries, much like other aspects of the Web, are becoming powerful instruments of change in education and research.

The digital library era has also spawned a movement toward open access to scholarly literature. Suber (2003) defines open-access literature as one which is digital, online, free of charge, and free of most copyright and licensing restrictions, whereas the Budapest Open Access Initiative (BOAI), an international effort to make research articles in all fields freely available on the internet, provides a slightly different definition: "literature with free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the Internet itself. The only constraint on reproduction and distribution, and the only role for copyright in this domain, is to give authors control over the integrity of their work and the right to be properly acknowledged and cited." The core belief of BOAI is that the removal of access barriers to literature will have a democratizing effect and will result in "accelerated research, enriched education, and make scholarly literature as useful as it can be, laying the foundation for uniting humanity in a common intellectual conversation and quest for knowledge." The open access concept has parallels in the data world. In the atmospheric sciences community, an important consequence of the open sharing ideal has been the free flow of meteorological data across geographic boundaries, per World Meteorological Organization (WMO) Resolution 40, which commits the WMO to broadening and enhancing the free and unrestricted international exchange of meteorological and related data and products. The resulting sharing and free flow of data meteorological data has had a noticeable impact on education and research in atmospheric and related sciences, as illustrated in a companion article by Yoksas et al. (2005).

Another notable information technology trend is the desire to integrate all information, including data and a variety of services behind a single entry point or a portal. Portals often include personalization features allowing users a tailored view into the information. The customization permits: a) a single point of authentication to validate access permissions and enable links to available resources, and b) the ability to design a customized view of available information.

The open access, open source and open standards are inter-related concepts that are gaining momentum and developers of data service are aggressively rethinking how they might both contribute to and benefit from these trends toward "openness." The benefits of open access, open source and open standards are numerous and when they are combined the benefits can be even greater.

Open source software is software that includes source code and is usually available at no charge. The open source model for software has many benefits, as articulated in a collection of essays by Raymond (1999). For instance, it has the advantage of harnessing the collective wisdom, experiences, expertise and requirements of large communities. Drawing upon the Linux development experience and based on his successful open source software project, *fetchmail*, Raymond makes a compelling argument for the proposition that, "Given enough eyeballs, all bugs are shallow" and for the importance of treating users as co-developers. Additional features and benefits include scalability, extensibility, and customizability. For example, people using a wide variety of hardware platforms, operating systems, and software environments can test, modify, and run software on their system to test for portability. Successful open source development efforts also do not start from scratch but rather try to adapt and build on top of existing code base, using the community process for refinement and reuse.

In the data services area, many excellent examples of open source software that are highly reliable and supported by a large community exist. They include Linux, Apache, MySQL, and similar projects. Network Common Data Form (netCDF), Open-source Project for a Network Data Access Protocol (OPeNDAP), Thematic Realtime Environmental Distributed Data Services (THREDDS, Domenico et al., 2002) are leading examples of open source software in the geosciences data infrastructure area. Because of its free and open source nature, netCDF software has been incorporated into over 50 other open source software packages and 15 commercial packages, resulting in its widespread use and status as a de facto standard for data format in atmospheric and related sciences. Likewise, the OPeNDAP software, which was originally called Distributed Ocean Data System, has found wide use outside the core oceanography data community where it originated. Open source software also increases opportunities for software reuse, adaptation to different hardware and software environments, and customization to user needs. The best example, perhaps, is the use of Linux in a wide range of electronic and computer systems, including videogame consoles, mobile phones, Personal Digital Assistants, and personal, mainframe computers and massively parallel high-performance computing systems. The large-scale availability of access to the Internet and Internet applications, coupled with widespread use of Linux in academia and the availability of inexpensive, commodity microprocessors and storage devices, has had a democratizing effect on data provision to and access by the geosciences community. As an example, today over 160 colleges and universities worldwide are participating in the Unidata Internet data distribution system and they are receiving, sharing and distributing data and integrating them into their education and research using inexpensive computers and freely-available, Linux-based open source applications.

The use of open standards models for middleware, a special kind of software between client and server processes to ensure consistency and interoperability, is particularly important for developing new data services. For example, an open standards-based middleware provides opportunities for the provision of a stable, consistent interface to a wide variety of applications, on a broad set of platforms and enable their inter-operability. In the process, it decouples data service providers from users, allowing end users with multiple clients to access the same services. This can accelerate the migration of data services to new and diverse platforms. Furthermore, it facilitates the "wrapping" of legacy systems in standard interfaces, giving them the ability to integrate with other distributed components and systems. Given the demand for standards-based, open systems that easily integrate, the open source development process provides a significant advantage over proprietary approaches to software development and use.

Interfaces based on open standards are by definition publicly documented and based on an explicit or de facto standard. There is evidence that well developed open standards for data formats are less likely to become quickly obsolete and are more reliable and stable than proprietary formats. Having access to the file format also allows users and developers to create data conversion utilities into other formats. File formats that use open standards can assist in long-term archiving because they allow for software and hardware independence. Open standards also allow for greater flexibility and easy migration to different systems and interoperability of diverse systems. Open access, open source software models, and open standards each offer a number of significant benefits in the provision of data services.

Extensible Markup Language, XML, is a simple, highly flexible, text-based framework for defining mark up languages. This standard for classifying, structuring, and encoding data allows organizations and services to exchange information more easily and efficiently. Although originally developed to facilitate Web-based publishing in a large scale, XML has since rapidly gained acceptance and usage in the exchange of a wide variety of data on the Web. An important emerging standard for interoperability of data systems is in the metadata area, which can use XML to share descriptions of underlying datasets.

The ability of XML to organize data into a computer-interpretable format that is also easy to code and read by humans is quickly making XML the lingua franca for business services and electronic commerce and also rapidly becoming a widely used standard in the data services world. Because of its simplicity and elegance, XML has radically transformed the provision of data services in the scientific community. Some of its principal benefits include: a) ability to delineate syntactic information from semantic information[1]; b) allows the creation of customizable markup languages for different use cases and application domains; c) platform independence. For example, XML makes it possible for providers of data services to send information about data sets, metadata, in a form completely separate from the presentation of the underlying data. Furthermore, service providers can present the same information in multiple forms or views using XML style sheets, customized to the needs of particular users. For example, Really Simple Syndication (RSS) is a lightweight XML format for sharing news and bulletins, and it has been used successfully by the U.S. National Weather Service to disseminate weather information such as local forecasts, watches, and warnings to Internet users. The same technology can also be used in the data services context to notify users when new data becomes available in a data system.

Web services, based on XML and HTTP, the two open standards that have become ubiquitous underpinnings of the Web, are emerging as tools for creating next generation distributed systems. Besides recognizing the heterogeneity as a fundamental ingredient, web services, independent of platform and development environment, can be bundled, published, shared, discovered, and invoked as needed to accomplish specific tasks. Because of their building-block nature, web services can be deployed to perform either simple, individual tasks or they can be chained to perform complicated business or scientific processes. As a result, web services, implemented in a Service Oriented Architecture (SOA) or framework, are quickly becoming a technology of choice for deploying cyberinfrastructure for data services. By wrapping existing applications and their components as web services in a SOA, the traditional obstacles to interfacing legacy and packaged applications with data systems are being overcome through loosely coupled integration. Such an approach to lightweight integration affords an easier pathway to interoperability amongst disparate systems and distributed services. The new software architectures based largely on web services standards are enabling whole new service-oriented and event-driven architectures that is challenging traditional approaches to data services. In a series of articles, Channabasavaiah et al. (2003) present a persuasive case for developing and deploying a SOA, as the level of complexity of traditional architectures increases and approaches the limit of their ability. They also provide a realistic plan for migrating existing applications to a SOA, one that leverages existing assets and allows for incremental implementation and migration of those assets. Several efforts are underway

[1] Syntactic metadata describes what the data *looks* like and how it is organized; semantic metadata describes what it really *means*.

within the geosciences community to apply web services and service oriented architectures to both migrate existing, stove-piped data systems as well as in the development of common architectures for future data systems. For example, the strategic plan for the U.S. Integrated Earth Observation System, the U.S. contribution to Global Earth Observation System of Systems (GEOSS), calls for the implementation of GEOSS services within a web-enabled, component-based architecture in its overall data management strategy so that the value of Earth observations data and information resources is maximized (IWGEO, 2005; Hood, 2005). Likewise, the Integrated Ocean Observing System and the NOAA Group on Earth Observations Integrated Data Environment (GEOIDE) are both planning to use a SOA/web services approach for providing data services to their respective communities.

Another computing model that is beginning to transform how resources are applied to solve complex scientific problems is Grid computing, a term that originated in the 1990s as a metaphor for making distributed computer power as easy to use as an electrical power Grid. While many definitions of Grid computing exist (Wikepedia: The Free Encyclopedia, 2006), the most definitive and widely used one is by Foster and Kesselman (1997). According to them, the Grid refers to "an infrastructure that enables the integrated, collaborative use of high-end [and distributed] computers, networks, databases, and scientific instruments owned and managed by multiple organizations." Grid applications often involve large amounts of data and/or computing and often require secure resource sharing across organizational boundaries. Grid computing and the science enabled by it, eScience, are two major trends in distributed computing. A key advantage of grid computing over historical distributed computing systems is that the Grid concept permits the virtualization of computing resources such that end-users have the illusion of using a single source of "computing power" without knowing the actual location where their computations are performed. The use of digital certificates to access systems on behalf of a user and third-party file transfer between grid nodes authenticated via certificates are specific examples of how Grid technology enables resource allocation and virtualization. Grid Services, which implement web services in a Grid architecture, are in still in their infancy, although several proof-of-concept testbeds have been deployed in a number of disciplines, including earth and atmospheric sciences, high energy physics, and biomedical informatics. Although the distinction between traditional web services and Grid services is subtle, Grid services, ideally, should enable virtualization for building and running applications that span organizations and share resources and infrastructure in a seamless way. Gannon et al. (2005) and Foster (2002) provide the distinguishing characteristics of a Grid service and specify what is needed for a web service to qualify as a Grid service.

4 Data service attributes

As articulated by Cornillon (2003), the ultimate objective of a data system or service is to provide requested data to the user or user's application (e.g., analysis or visualization tool) in a transparent, consistent, readily useable form. The users do not care as much about the technology behind those systems or services, but do about transparency and usability. The key to achieving Cornillon's two objectives is through interoperability of components, systems and services, via the use of standards.

In the opinion of this author, an ideal data service should have the following attributes:

- User-friendly interface
- Transparency (format, protocol, etc.)
- Customization and personalization of services
- Capability for server-side operations (e.g., subsetting, sub-sampling, etc)
- Aggregation of data and products
- Provision of rich metadata
- Integration across data types, formats, and protocols
- Intelligent client-server approaches to data access and analysis
- Interoperability across components and services
- Flexibility, extensibility, and scalability
- Ability to chain services via workflows
- Support an array of tools for access, processing, management, and visualization

As a result of the aforementioned trends, the last decade has seen an evolution of data systems like EOSDIS (Earth Observing System Data and Information System) towards a more layered and open architecture, while new data systems have been built and deployed using many open source and standards-based technologies (e.g., the NOAA National Operational Model Archive and Distribution System [NOMADS; Rutledge et al., 2002)], Community Data Portal (Middleton, 2001) and Earth System Grid (Foster et al., 2002); and data system at the British Atmospheric Data Centre (Lawrence, 2003) However, the transition has not been without challenges for a number of reasons, including:

- Heterogeneity and complexity of distributed observing, modeling, data, and communication systems
- Nature of data coverage: diversity and multiple spatial and temporal scales

- Data systems using both legacy components alongside contemporary applications, creating integration challenges
- A lack of standards and interoperability
- Non-monolithic user community
- Political, technological, and cultural and regulatory barriers, especially in global sharing of and access to data

Given the very high data rates from current and future generation observing systems such as GOES-R and NPOESS satellites, the user community will need a hybrid solution that couples a satellite-based data reception system with a terrestrial, Internet-based data access system. Both local and remote data access mechanisms will be required to deal with the large volumes of data. Both push systems for distributing data (e.g., Unidata Local Data Manager) or just notifications (using RSS feeds) and pull systems for remote access (e.g., THREDDS and OPeNDAP) will be required.

4.1 Broad data categories

While far too many data categories exist to describe in detail, typical data systems in atmospheric sciences must provide a seamless, end-to-end services for accessing, utilizing, and integrating data across the following data types:

- Real-time data
- Archived data
- Field and demonstration project data
- Episodic or case study data
- Data from related disciplines (hydrology, oceanography, cryosphere, chemical and biosphere – soil, vegetation, canopy, evapotranspiration)
- GIS databases

The first four categories, and to a lesser extent the fifth one, include data from in-situ and remote sensing observations, and output from models. Even though each discipline within the geosciences is unique and has different data needs depending on the use or application, the geoscience disciplines do share a common interest in accessing data of the types listed above. For instance, the first four data types are important for many applications in atmospheric sciences, oceanography, hydrology, geologic subdisciplines of seismology and volcanology. Another common attribute is the need for georeferencing and integration with information contained in GIS databases. This brings up an important area of ongoing research, namely, the development of a common data model for the geosciences, as they share a common representation of data in their spatial and temporal representations. A discussion of data models for geosciences is beyond the scope of this article, however.

4.2 Data deluge, data mining, and knowledge discovery

Advances in computing, modeling, and observational systems have resulted in a veritable increase in the volume of data. These data volumes will continue to see exponential growth in the coming years. For example, data from current and future observing systems will result in a 100 fold increase in volume in the next decade. The GOES-R satellite, scheduled for launch in 2012, will have a hyperspectral sounder with approximately 1600 channels. In contrast, the current generation GOES satellite sounders have 18 thermal infrared channels. Similarly, each NPOESS satellite when fully deployed will have raw data rates of nearly 1 Terabyte each day. Hey (2003) previews the imminent data deluge from the next generation of simulations, sensors, and modeling systems and experiments, and discusses the importance of metadata and the need to automate the process of converting raw data to useful information and knowledge and implications for Grid middleware architecture.

The data deluge clearly requires extraction of higher level information useful to users. The process of extracting higher level information is referred to as data mining. Data mining is a key step toward data reduction and knowledge discovery. Graves (1996) and Ramachandran et al. (1999) offer a methodology to efficiently mine and extract content-based metadata from Earth Science datasets and describe the capabilities of the ADaM (A Data Miner) tool, which enables phenomena-oriented data mining by incorporating knowledge of phenomena and detection algorithms in the system. Their methodology provides a meaningful solution needed by users to convert data to knowledge and cope with the data deluge. An ideal data system or service should include algorithms and facilities for data mining that can be applied to data sets as needed by users. Future success will depend on how well users are served by such discovery and mining tools and services pertaining to data integration.

4.3 Geographic Information Systems

Geographic Information Systems (GIS) have become indispensable tools for geoscientific exploration, commerce, and for decision making in environmental and social sciences (Morss, 2002). In general, GIS involves a broad array of computer tools for mapping, and managing geographically referenced data, and for spatial analysis. Regardless of the application, all the geographically related data can be input and prepared in a GIS such that users can display the specific information of interest, or combine data contained within the system to produce additional value-added information that might help answer a specific problem. 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 cyberinfrastructure and data services in the geosciences.

GIS' powerful capability is to integrate spatially referenced data from disparate sources into a single environment. An important application of GIS is linking 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. The ability to integrate observations and model data from several geoscience disciplines with socioeconomic and biophysical data in a common framework permits not only multidisciplinary analysis and synthesis, but also provides a pathway to approach geoscience problems and processes from an Earth system science perspective.

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 reason 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 of proprietary GIS applications, 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 applications shifting from operational support tools to strategic decision support systems, 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 evolution of data services.

Another recent trend of GIS is to make it accessible via the Internet, allowing easier 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 applications 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 (OGC) and International Standards Organization are aiming to addresses 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. For instance, the OGC "Geo-interface for Atmosphere, Land, Earth, and Ocean netCDF" (GALEON) Interoperability Experiment supports open access to atmospheric and oceanographic modeling and simulation outputs and it will implement a geo-interface to netCDF datasets via the OpenGIS Web Coverage Server protocol specification. The interface will provide interoperability among netCDF, OPeNDAP, ADDE, and THREDDS client/server and catalog protocols.

5 Concluding remarks

It is important to recognize that we are in the midst of a revolution in data services and the underlying information technologies. This revolution is far from complete. The data services we see today, though advanced, are still evolving and their evolution toward more complex and sophisticated systems is expected to continue for the foreseeable future.

This article has presented a brief overview of many issues reshaping geoscience education and research and it provided a survey of many of the technological trends that have contributed to new approaches to data provision and their integration in Earth system science education and research. The new approaches and services are transforming how students, faculty, and scientists use data services in their daily work The imminent data deluge from a new generation of remote sensing satellite instruments, next generation models, and experiments will have a profound impact on the scientific community and data infrastructure, and it calls for new ways to exploit these and other information technology trends for the development of new approaches to scientific data services.

Acknowledgement. This work was supported by the National Science Foundation under grants ATM-0317610.

Edited by: E. Cutrim, M. Ramamurthy, S. Nativi, and L. Miller
Reviewed by: anonymous referees

References

AGU: Shaping the Future of Undergraduate Earth Science Education: Innovation and Change Using an Earth System Approach, http://www.agu.org/sci_soc/spheres/⟩http://www.agu.org/sci_soc/spheres/, 1997.

Channabasavaiah, K., Holley, K., and Tuggle, E.: Migrating to a service-oriented architecture. Parts I & II, IBM DeveloperWorks white paper, http://www-128.ibm.com/developerworks/webservices/library/ws-migratesoa/, http://www-128.ibm.com/developerworks/library/ws-migratesoa2/, 2003.

CIERE: Cyberinfrastructure for Environmental Research and Education, Workshop sponsored by the National Science Foundation, Boulder, Colorado, 2002, urlhttp://www.ncar.ucar.edu/cyber/, 2003.

Colwell: The National Science Foundation's Role in the Arctic, Dr. Rita R. Colwell, Opportunities in Arctic Research: A Community Workshop, Arlington, Virginia, 1998, http://www.nsf.gov/od/lpa/forum/colwell/rc80903.htm, 1998.

Domenico, B., Caron, J., Davis, E., Kambic, R., and Nativi, S.: Thematic Real-time Environmental Distributed Data Services (THREDDS): Incorporating Interactive Analysis Tools into NSDL, Journal of Digital Information, 2, 4, http://jodi.ecs.soton.ac.uk/Articles/v02/i04/Domenico/, 2002.

Foster, I.: What is the Grid? A three point checklist, Grid Today, 1, 6, http://www.gridtoday.com/02/0722/100136.html, 2002.

Foster, I. and Kesselman, C.: Globus: A metacomputing intrastructure toolkit, International Journal of Supercomputer Applications, 11(2), 115–128, 1997.

Foster, I., Alpert, E., Chervenak, A., Drach, B., Kasselman, C., Nefedova, V., Middleton, D., Shoshani, A., Sim, A., and Williams, D.: The Earth System Grid II: Turning climate datasets into community resources, 19th Conference on Interactive Information Processing Systems, American Meteorological Society, Long Beach, CA, 2002.

Gannon, D., Plale, B., Christie, M., Fang, L., Huang, Y., Jensen, S., Kandaswamy, G., Marru, S., Lee Pallickara, S., Shirasuna, S., Simmhan, Y., Slominski, A., and Sun, Y.: Service Oriented Architectures for Science Gateways on Grid Systems, International Conference on Service Oriented Computing 2005, edited by: Benatallah, B., Casati, F., Traverso, P., LNCS 3826, Springer-Verlag Berlin Heidelberg, pp. 21–32, 2005.

Graves, S. J., Hinke, T., and Kansal, S.: Metadata: The golden nuggets of data mining, First IEEE Metadata Conference, Bethesda, MD, April 16–18, http://www.cs.uah.edu/~thinke/Mining/metapaper.html, 1996.

Griffin, S. M.: NSF/DARPA/NASA Digital Libraries Initiative: A Program Manager's Perspective, D-Lib Magazine, July/August, http://www.dlib.org/dlib/july98/07griffin.html, 1998.

Hey, A. J. G. and Trefethen, A.: The data deluge: An e-Science perspective. Grid Computing, Making the Global Infrastructure a Reality, edited by: Berman, F., Fox, G., Hey, A. J. G., John Wiley and Sons, 1060pp., 2003.

Hood, C. A.: Implementation of GEOSS: A review of all-hazards warning system and its benefits to public health, energy, and the environment, Written testimony, U.S. House Committee on Energy and Commerce, 9 March 2005.

IWGEO: Strategic Plan for the U.S. Integrated Earth Observation System, National Science and Technology Council Committee on Earth and Natural Resources, Washington, D.C., http://iwgeo.ssc.nasa.gov/, 2005.

Lawrence, B.: Experiences with archiving databases in British Atmospheric Data Center, 6th DPC Forum focussed on Open Source Software and Dynamic Databases, London, U.K., 24 June 2003.

Lemke, J. L.: Talking Science: Language, Learning, and Values, Norwood, NJ: Ablex Publishing, 1990.

Manduca, C.: Using Data in the Classroom, Workshop sponsored by NSF, Carleton College, April 2002, http://dlesecommunity.carleton.edu/research_education/usingdata/index.html, 2002.

Middleton, D.: The Community Data Portal: Sustainable strategies for enabling both providers and consumers of Earth System data, http://www.ncar.ucar.edu/stratplan/communitydata.html, 2001.

Michalakes, J., Chen, S., Dudhia, J., Hart, L., Klemp, J., Middlecoff, J., and Skamarock, W.: Development of a Next Generation Regional Weather Research and Forecast Model, Developments in Teracomputing: Proceedings of the Ninth ECMWF Workshop on the Use of High Performance Computing in Meteorology, edited by: Zwieflhofer, W. and Kreitz, N., World Scientific, Singapore, pp. 269–276, 2001.

Morss, R.: Working Group Report, Workshop on GIS in Weather, Climate, and Impacts Workshop, Boulder, http://www.esig.ucar.edu/gis/02workshop/working_groups/hazardsB.pdf, 2002.

NRC: Grand Challenges in Environmental Sciences, National Research Council, National Academy Press, 96 pp., http://www.nap.edu/books/0309072549/html/, 2001.

NSF: Geoscience Education: A Recommended Strategy, Arlington, Virginia, http://www.geo.nsf.gov/adgeo/geoedu/97_171.htm, 1997.

NSF: NSF Geosciences Beyond 2000: Understanding and Predicting Earth's Environment and Habitability, NSF Report No. 00-27, 54 pp., http://www.geo.nsf.gov/adgeo/geo2000/geo_2000_summary_report.htm, 2000.

NSF: Complex Environmental Systems: Synthesis for Earth, Life, and Society in the 21st Century, NSF Report No. 03-27, 80 pp., http://www.nsf.gov/geo/ere/ereweb/ac-ere/acere_synthesis_rpt_full.pdf, 2003.

OPeNDAP: Open-source Project for a Network Data Access Protocol, http://www.opendap.org/.

Ramachandran, R., Conover, H., Graves, S. J., Keiser, K., and Rushing, J.: The role of data mining in Earth Science data interoperability, ASPRS Annual Conference, Conference on Remote Sensing Education (CORSE), Portland, OR, May 17–22, 1999.

Raymond, E. S.: The Cathedral & the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary, O'Reilly and Associates, 268 pp, 1999.

Rutledge, G. K., Alpert, J., Stouffer, R. J., and Lawrence, B.: The NOAA Operational Archive and Distribution System (NOMADS), Proceedings of the Tenth ECMWF Workshop on the use of High Performance Computing in Meteorology, edited by: Zwiefhofer, W. and Kreitz, N., World Scientific, 106–129, 2002.

Suber, P.: Removing barriers to research: An introduction to open access for Librarians, College & Research Libraries News, 64 (February 2003), 92–94, 2003.

Yoksas, T., Almeida, W., Garrana, D., Castro, V., and Spangler, T.: Internet Data Distribution – Extending real-time data sharing throughout the Americas, 2005 European Geosciences Union General Assembly, Education Symposium, Session on Earth System Science Data Access, Distribution and User for Education and Research, Vienna, Austria, 24–29 April 2005.

Wikepedia: The Free Encyclopedia, http://en.wikipedia.org/wiki/Grid_computing, 2006.