Data management and analysis for the Earth System Grid

2008 J. Phys.: Conf. Ser. 125 012072

(http://iopscience.iop.org/1742-6596/125/1/012072)

Data management and analysis for the Earth System Grid

The Earth System Grid Center for Enabling Technologies Team:

D N Williams[3], R Ananthakrishnan[1], D E Bernholdt[7,9], S Bharathi[8], D Brown[5], M Chen[7], A L Chervenak[8], L Cinquini[5], R Drach[3], I T Foster[1,9], P Fox[5] , S Hankin[6], V E Henson[3], P Jones[4], D E Middleton[5,9], J Schwidder[7], R Schweitzer[6], R. Schuler[8], A Shoshani[2], F Siebenlist[1], A Sim[2], W G Strand[5], N Wilhelmi[5] and M Su[8]

[1] Argonne National Laboratory, Argonne, IL 60439, USA
[2] Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[3] Lawrence Livermore National Laboratory, Livermore, CA 94550, USA
[4] Los Alamos National Laboratory, Los Alamos, NM 87545, USA
[5] National Center for Atmospheric Research, Boulder, CO 80307, USA
[6] National Oceanic and Atmospheric Administration (PMEL), Seattle, WA 98115, USA
[7] Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA
[8] University of Southern California, Information Sciences Institute, Marina del Ray, CA 90292, USA

E-mail: williams13@llnl.gov, don@ucar.edu, itf@mcs.anl.gov, bernholdtde@ornl.gov

Abstract. The international climate community is expected to generate hundreds of petabytes of simulation data within the next five to seven years. This data must be accessed and analyzed by thousands of analysts worldwide in order to provide accurate and timely estimates of the likely impact of climate change on physical, biological, and human systems. Climate change is thus not only a scientific challenge of the first order but also a major technological challenge. In order to address this technological challenge, the Earth System Grid Center for Enabling Technologies (ESG-CET) has been established within the U.S. Department of Energy's Scientific Discovery through Advanced Computing (SciDAC)-2 program, with support from the offices of Advanced Scientific Computing Research and Biological and Environmental Research. ESG-CET's mission is to provide climate researchers worldwide with access to the data, information, models, analysis tools, and computational capabilities required to make sense of enormous climate simulation datasets. Its specific goals are to (1) make data more useful to climate researchers by developing Grid technology that enhances data usability; (2) meet specific distributed database, data access, and data movement needs of national and international climate projects; (3) provide a universal and secure web-based data access portal for broad multi-model data collections; and (4) provide a wide-range of Grid-enabled climate data analysis tools and diagnostic methods to international climate centers and U.S. government agencies. Building on the successes of the previous Earth System Grid (ESG) project, which has enabled thousands of researchers to access tens of terabytes of data from a small number of ESG sites, ESG-CET is working to integrate a far larger number of distributed data providers, high-bandwidth wide-area networks, and remote computers in a highly collaborative problem-solving environment.

Journal of Physics: Conference Series **125** (2008) 012072 doi:10.1088/1742-6596/125/1/012072

1. Context: climate change and Earth system modeling

For several decades, scientists have been working to understand the Earth's delicate ecosystem and intricate climate balance. Notably, they have sought an understanding of potential changes in these systems induced by naturally and human-generated greenhouse gases (see figure 1). Naturally occurring greenhouse gases—such as carbon dioxide—help absorb the sun's rays, keeping the planet warm enough to support life. Over the past century, however, humans have released vast quantities of greenhouse gases into the atmosphere and have simultaneously disrupted Earth's natural carbon sequestration process by cutting down forests. Although calculating the CO_2 sequestration of various types of trees is complicated, it is estimated that one mature tree consumes between 40 and 50 pounds of CO_2 per year and releases enough oxygen into the atmosphere to support two human beings [1].

The burning of fossil fuels and deforestation will undoubtedly continue well into the 21st century. Moreover, the effects of these activities are already being felt worldwide. The ten warmest years ever-recorded have all occurred since 1990, and rapid reductions in the extent of mountain glaciers have been observed on five continents. Atmospheric CO_2 has increased over the past century from 280 parts per million (ppm) to over 360 ppm— easily the largest concentration of CO_2 in the atmosphere since humans evolved. Oceans are also on the rise as ocean temperature increases, causing thermal expansion of the oceans as well as the melting of both land ice and polar icecaps. Rising sea levels threaten disastrous inundation of low-lying coastal countries, many of which are densely populated.

Together, the atmosphere, oceans, land, ice, and biosphere constitute a complex system that is heavily influenced by greenhouse gas emissions. The importance of understanding greenhouse gas emissions and their potentially devastating impact on this earth system cannot be overstated. Indeed, the quality of human life on earth, and perhaps even human survival itself, may depend on reaching that understanding.

To help understand these climatic imbalances and potential impacts, scientists rely on powerful computers and complicated computer codes to simulate the Earth's climate. Numerical climate models allow scientists to improve their understanding of the earth system by running exhaustive computational experiments aimed at studying the behaviors of individual chemical, biological, and geophysical processes, as well as relationships among these individual processes. Computational experiments performed with these models allow scientists to model and study all aspects of climate change to help understand future climate. Simulations can also be used to study the effectiveness of proposed approaches to mitigating the impact of climate change.

Climate models are also used to study regional climate changes as well as high-impact characteristics of climate such as typhoons, droughts, and floods. Accurately resolving these and other processes and discovering long-term patterns such as temperature distribution or the melting of land and sea ice, requires the application of intricate mathematical formulas to data at points on ever finer grid resolutions. Hence, as computing power increases, so does the opportunity to improve the computer models (see figure 2).

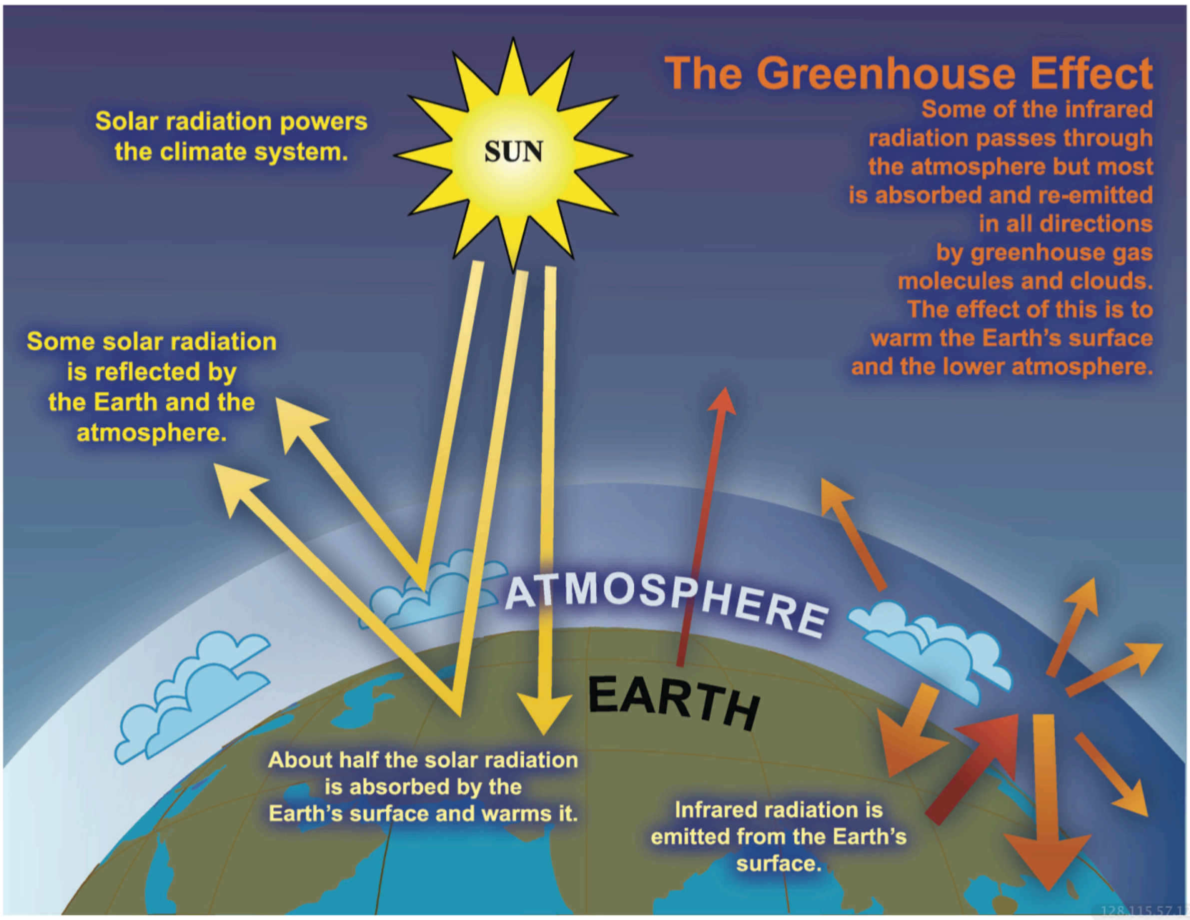

Figure 1. Depiction of the natural greenhouse effect. (Image courtesy of the Intergovernmental Panel on climate Change (IPCC) Fourth Assessment Report [2].)

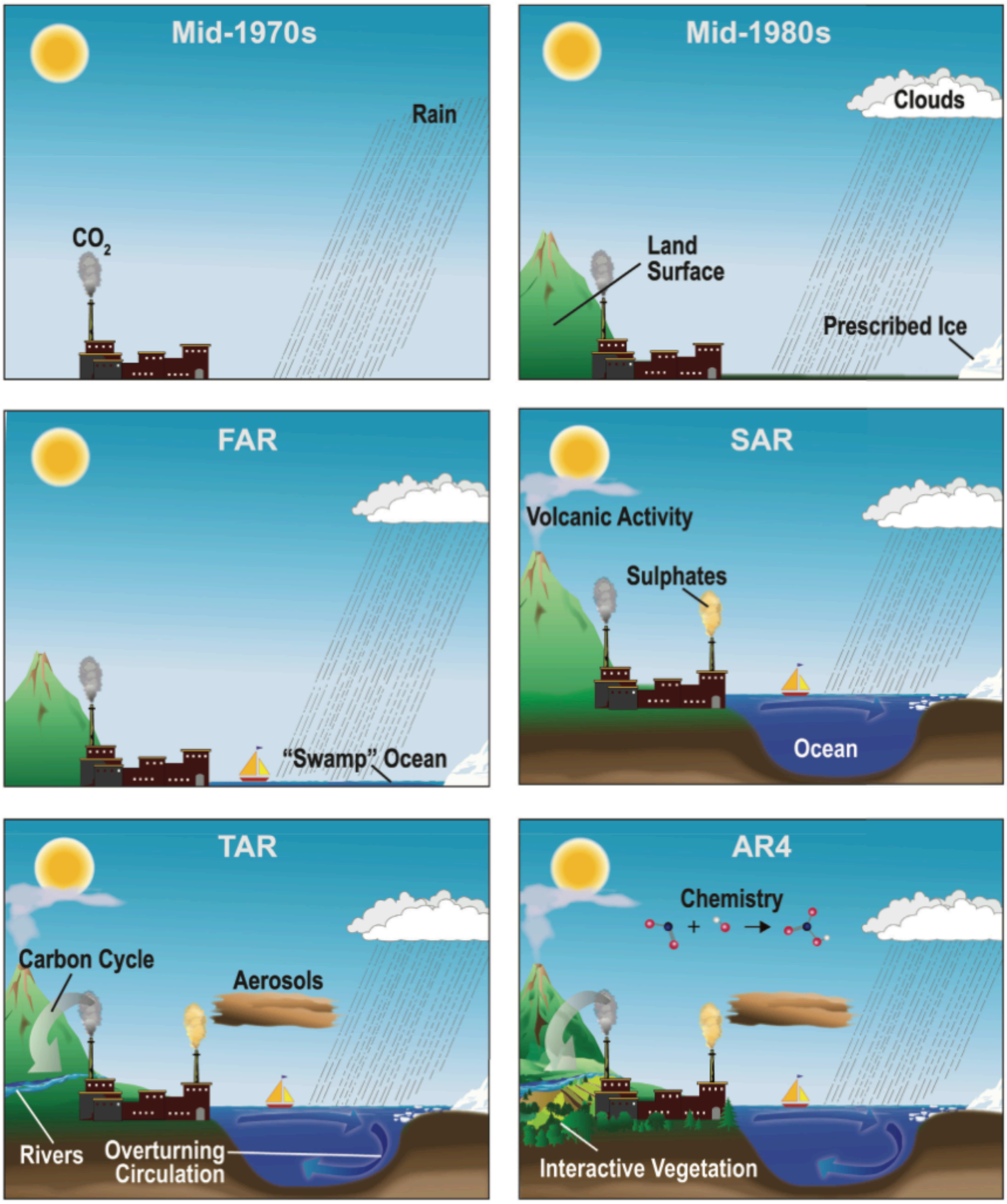

Figure 2. Progression of the complexity in the coupled models over the past few decades, where FAR represents the Intergovernmental Panel on Climate Change (IPCC) First Assessment Report, SAR represents the IPCC Second Assessment Report, TAR represents the IPCC Third Assessment Report, and AR4 represents the IPCC Fourth Assessment Report. (Image courtesy of the Intergovernmental Panel on climate Change (IPCC) Fourth Assessment Report [2].)

2. The climate data challenge and the Earth System Grid

A consequence of increased computing power and more comprehensive climate models is a dramatic increase in data output describing the more complex spatial and temporal Earth system model. Over the past decade, model output has increased exponentially—moving from megabytes, to gigabytes, to terabytes, and now petabytes. The worldwide climate community is expected to generate hundreds of petabytes of simulation data within the next five to seven years.

Using these data will not be easy. Indeed, merely examining or sorting data will be problematic, partly due to the enormous quantity of data involved, and partly due to the fact that the storage for all the data will be distributed worldwide. Far more than just climate model output data will be involved. For accurate study of the climate, in situ, satellite, biogeochemistry, and ecosystems data will also be required. Because of the high cost to generate the data (in both computer and human time), they must be carefully managed to maximize and preserve their value to the community.

Data providers want to make their data available to others but will do so only if they have confidence in distribution processes they understand and (to the extent necessary) control. Data users need to be able to find and access data of interest, and must have confidence in the origins and integrity of the data and the associated metadata. Because most users will have considerably less storage (and network connectivity) than the major centers where simulation data are generated and archived, data users will need assistance from the archive sites in performing common data reduction, analysis, and visualizations, all of which must be done on the "server side" to reduce the volume of data sent over the network and stored at the user's location.

The Earth System Grid (ESG) Center for Enabling Technologies (ESG-CET) [3, 4] was established to address these technical challenges. Specifically, it was charged with addressing the management and use of extremely large and diverse sets of climate data, aiming to build a "science gateway" to climate resources. Funded by the SciDAC program through the offices of Advanced Scientific Computing Research (ASCR) and Biological and Environmental Research (BER), ESG-CET intends to develop approaches and technologies that also benefit other scientific application areas that entail management of petascale data.

The remainder of this paper examines some of the issues associated with data management at this scale, and how the ESG-CET team is addressing those issues.

3. The ESG-CET

The ESG-CET project did not begin from scratch: it is built on the considerable existing infrastructure of the first-generation Earth System Grid [5], developed over the past several years, and on the associated suite of open-source software that ESG provides to access and manage climate data, information, models, analysis tools, and visualization tools. ESG-CET will also benefit from the considerable experience and user-feedback obtained through use of the first-generation system. ESG users now number nearly 10,000, and ESG data holdings total more than 235 terabytes.

The climate modeling community have become enthusiastic supporters and users of ESG, which has led in turn to requests and expectations that it will play an expanded role in data dissemination in the future. Hence, the ESG-CET project is the "next phase" in ESG development. The ESG-CET team is now designing new products and services that will enable users to find and obtain data from spinning disk or tertiary storage, migrate data from one location to another, and enable a number of workflow capabilities across a worldwide network of climate data archives. Further, this will occur at a scale several orders of magnitude larger than the community has seen to date.

Working closely with key customers and major collaborators, the ESG-CET team has established requirements for the next phase of the project. These customers and collaborators include the following:

- The Intergovernmental Panel on Climate Change (IPCC) Fifth Assessment Report (AR5)
- The Computational Climate End Station (CCES) at the DOE Leadership Computing Facility at Oak Ridge National Laboratory (ORNL)
- Groups addressing the data management and dissemination needs of the Community Climate System Model (CCSM) and related model development efforts
- A variety of regional and other related efforts within the climate community requiring similar data management and dissemination capabilities

This broad set of target users represents a significant change from the earlier project, for which CCSM development was the primary focus, and supporting the IPCC AR4 activity was added mid-project.

3.1. Distributed partners and virtual organizations

The requirements that have had the most significant impact on the architecture of the next-generation ESG environment have to do with the breadth and number of sites in the distributed system.

In the current production ESG environment, two primary portals service data archives at five sites. The two portals are located at the National Center for Atmospheric Research (NCAR) and the Program for Climate Model Diagnosis and Intercomparison (PCMDI) at Lawrence Livermore National Laboratory. Users connecting to these portals can access data archives at Los Alamos National Laboratory (LANL), Lawrence Berkeley National Laboratory (LBNL), LLNL, NCAR, and ORNL. Historically, NCAR and ORNL have served as the primary computational sites for development and science runs of CCSM and its precursors and collectively hold more than 198 TB of simulation data. Over 35 TB of data from the IPCC AR4 was collected from modeling centers in the 13 participating countries and consolidated at the PCMDI (at LLNL) for publication through ESG.

Planning for AR5, which is the major driver of requirements for ESG-CET, is still very much in progress, but the expectation is that the core simulation data will exceed 2 PB, with perhaps an order of magnitude more supplementary data. Rather than trying to consolidate all of this data at a single site, AR5 planners expect data to be hosted at the participating modeling centers, or perhaps consolidated within the participating nations. Assuming that AR5 participation will be similar to AR4, ESG-CET will need to support between one and two-dozen data archives in the next-generation ESG. Discussions have been held with several other groups interested in publishing specialized data collections through ESG; this has led to a plan for 20–30 data nodes, with the possibility of more. These sites will be located throughout the world, and only a handful will be at institutions where the ESG-CET team has a presence.

Each of these sites and special collections may be a significant research and data management operation in its own right, much more than a mere partner in ESG. Therefore, extending the ESG "virtual organization" (VO) will often be a matter of *federating* the ESG VO with other, established VOs. Indeed, these other groups may already have in place their own data services, user bases, and other organizational aspects. Though federation of VOs and data systems is an area of active research, a goal of ESG-CET is that ESG should be able to use identities and data archives managed by other VOs to the maximum extent possible, rather than requiring that users maintain multiple accounts, replicate data archives, etc. At the same time, federating external VOs must provide the "owners" of the data published in ESG with adequate information about (and control over) who has access to their data.

3.2. Data management

Recent trends in climate simulation are to use higher resolution grids (see figure 3), to include additional components in the simulation (for example, incorporating biogeochemical effects), and to rely more on ensembles of simulations rather than individual simulation runs. All these trends will increase both the number and sizes of the files comprising individual datasets and will increase the number of datasets to be published in ESG.

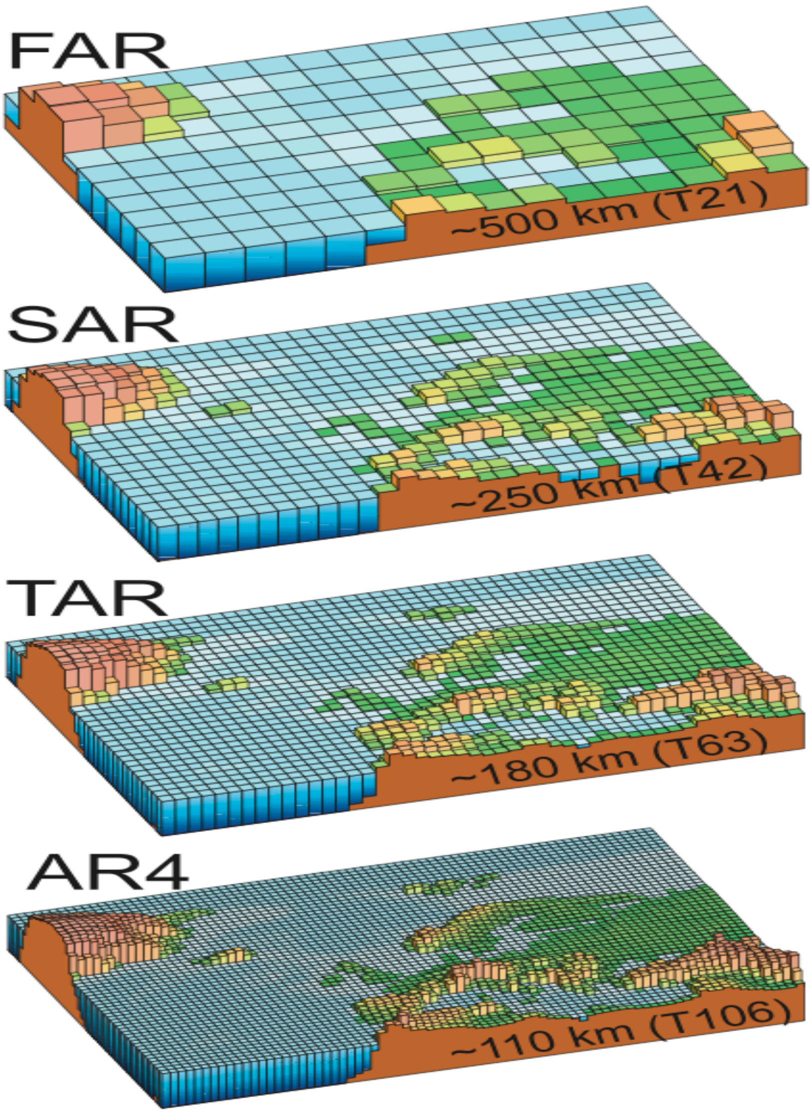

Figure 3. Progression of climate model grid resolution used in the IPCC Assessment Reports: FAR (1990), SAR (1996), TAR (2001), and AR4 (2007). (Image courtesy of the Intergovernmental Panel on Climate Change Fourth Assessment Report [2].)

The current ESG holdings, which include the entire IPCC AR4 archive and the past six years of joint DOE/NSF climate modeling simulation data, total roughly 235 TB. A reasonable estimate anticipates that the AR5 modeling campaign alone will involve several petabytes of data. In addition, holders of other collections of a similar magnitude have expressed interest in joining ESG. Thus, ESG-CET planning for the next generation system includes the capability of handling one to two orders of magnitude more data than the current system, easily reaching into the tens of petabytes. Planning for this scale has implications throughout ESG, from search and discovery through processing, analysis, and delivery of the results, and provides an important motivation for ESG-CET's emphasis on server-side analysis and subsetting capabilities.

The technology that will be used to store all of this data is also an important factor in the design of the next-generation ESG. While some sites will be able to field dedicated disk farms or use large-scale filesystems, (such as are being deployed at many HPC centers) many sites can accommodate these quantities of data only on hierarchical mass storage systems. The ESG design must allow for the fact that such data may take time to retrieve from the mass storage system before it can be processed and delivered to the user. Also, the organization of the data may make some user requests extremely expensive to carry out when that data comes from deep storage. One such example might be the extraction of the complete time history of a single variable from a dataset that is organized as one time step per file. These factors will have affects across the ESG system; planning to deal with them has led to discussions with model development teams about the organization of the output produced directly from the simulations (which also has a significant impact on I/O performance during the simulations).

3.3. New user communities
As ESG has become established in the climate modeling community as a widely used resource, the project has sought to broaden the range of users served. ESG-CET is focusing particularly on the needs of two types of users.

At one end of the spectrum are relatively small numbers of "power" users. These are researchers working with cutting-edge models that might produce huge volumes of data, use specialized grids, or have other novel characteristics. Such users may require especially detailed metadata for their models. They also may be particularly interested in high-end analyses and visualizations of their data, using sophisticated "thick" clients on their desktop computers, or perhaps more powerful computer resources for those tasks.

At the other end of the spectrum are the analysts who examine the impacts of climate change on the local and regional environments, populations, and infrastructures. This so-called impacts community is estimated to be an order of magnitude (or more) larger than the developers and immediate users of global climate models. Users in the impacts community tend employ digested or simplified information derived from one or many global climate simulations. They generally do not have a deep understanding of the science behind the global climate data, which primarily serves as inputs or boundary conditions for their models. They need the ability to discover and access model data in ways specific to their applications, including mapping to relevant coordinate systems, delivering data compatible with widely used geographic information systems, and creation of an appropriate and carefully documented suite of analyses and visualizations.

Both of these communities, as well as users who fall somewhere between the power user and the local analyst, have historically been part of ESG's primary user base and have strong needs for powerful data management and analysis capabilities, which are key areas of development in the ESG-CET project.

4. ESG-CET architecture overview
Given the ambitious nature of ESG-CET's future work, it is evident that the design and architecture of the ESG must be modified to enable the manipulation, management, and access of petabyte-scale datasets. A useful step is to review the current ESG architecture and then introduce the principal features of the new ESG-CET architecture.

4.1. Current ESG architecture
The current ESG architecture (see figure 4) went into production in late 2004. It consists of four major components designed to allow the climate community to share model data sets and enhances the scientist's ability to discover and access data sets of interest. The design also affords straightforward access to the tools required to analyze and interpret the data. The four major components of the architecture are as follows:

1. *Client applications*. The most important client application is the ESG portal, which provides a convenient, browser-based interface to ESG services. Other applications include user-level tools for data publication and assorted tools for data analysis and visualization.
2. *High-level and ESG-specific services*. These services span ESG resources and provide capability such as site-to-site data movement, distributed, reliable metadata access, and data aggregation and filtering.

3. *Globus/Grid infrastructure*. This provides remote, authenticated access to the shared ESG resources, enabling access to credentials, data and metadata, replica location, and submission and management of computational jobs.
4. *Database and application servers.* These servers include the basic resources on which ESG is constructed and includes the computational resources used for creation of virtual data, disk caches and mass storage systems used to hold data assets, database servers used to store metadata, and application servers for hosting ESG portal services.

The current ESG production data services span multiple domains including metadata, security, data transport, data aggregation and subsetting, usage metrics, and services monitoring. The cohesive integrated architecture is built in Java upon the Tomcat servlet engine and Struts web framework technologies. The production web portals allow data users and data providers to access the full spectrum of ESG high-level data services for publication, search and discovery, description, download, aggregation, and subsetting.

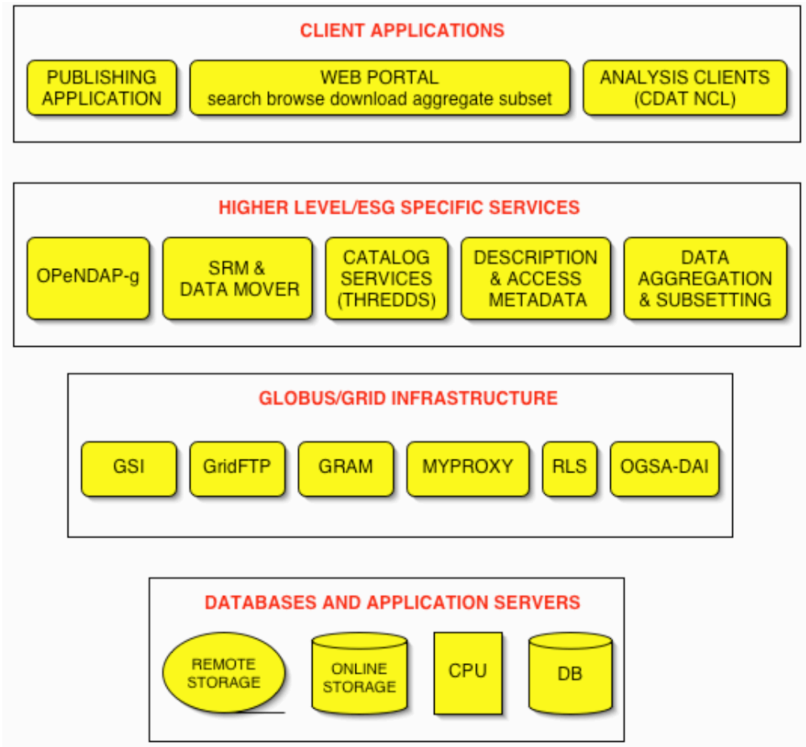

Figure 4. The four major components in the current ESG architecture.

4.2. New ESG-CET architecture

The current ESG architecture was designed to support a fixed (and small) set of well-known U.S. archive sites operating ESG nodes with essentially the same capabilities as a federated system, fronted by a single web-based portal as the primary interface to the user. In order to meet coming needs, this architecture must be greatly generalized. The next generation of the ESG must provide for a larger number of distributed sites located throughout the world. Further, the sites will have widely differing capabilities, and ESG must allow them to selectively federate, cooperate, or operate in standalone fashion, accommodating a variety of means for user access, including multiple portals and service- or API-based access, and data delivery mechanisms. The architecture must also be robust in the face of system and network failures at the participating sites.

In order to meet these expanded scientific and technical requirements, the new software framework (see figure 5) is designed to support a rich combination of interacting system components, while at the same time remaining fairly simple, lightweight, highly customizable, and easily deployable.

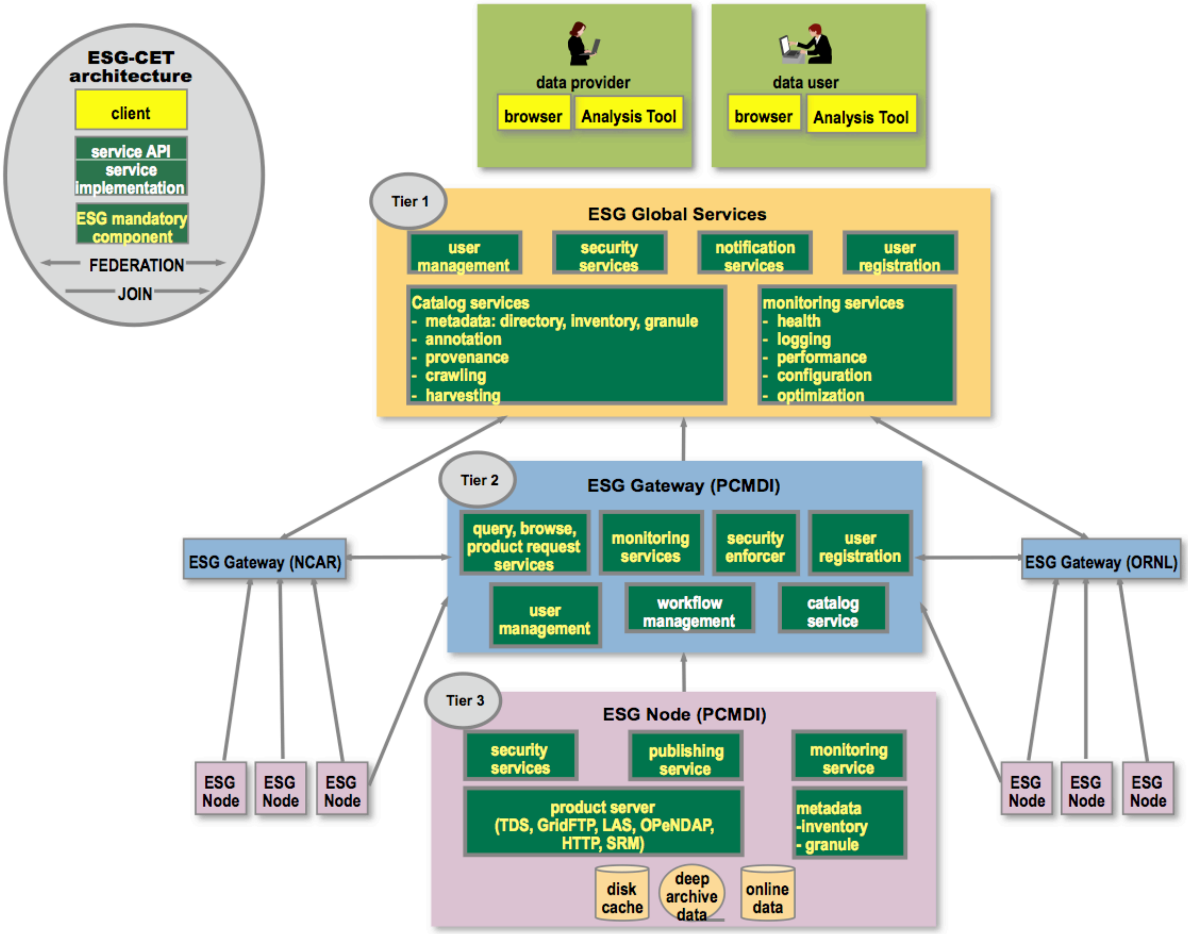

Figure 5. The new ESG-CET architecture, showing the three-tier data services.

The new ESG-CET framework will support both individual "ESG nodes" and collaborating, thematic ESG Gateways. ESG nodes are configured to offer a core set of ESG functionality for local data, metadata and computational resources, ESG Gateways federate multiple ESG nodes and provide access to distributed resources. In order to prepare for the ever-larger climate archives—expected to be in the petabytes—the new infrastructure is based on a three tier data service model:

1. *Tier 1 (Global metadata services for search and discover):* comprises a set of services providing shared functionality across the world-wide ESG-CET federation. An overall single sign-on authentication and authorization scheme will allow a registered user to access resources across the whole system, and to find data throughout the federation, independent of the site at which a search is launched. Tier 1 services will be managed by the ESG-CET staff.
2. *Tier 2 (Data gateways as data-request brokers)* comprises a limited number of ESG Data Gateways which act as brokers handling data requests to serve specific user communities. Services deployed on a Gateway include the user interface for searching and browsing metadata, for requesting data (including analysis and visualization) products, and for orchestrating complex workflows. At least initially, gateways will be managed directly by ESG-CET engineering staff.
3. *Tier 3 (ESG nodes with actual data holdings and metadata accessing services)* includes the actual data holdings and reside on a (potentially large) number of federated ESG Nodes, which host those data and metadata services needed to publish data onto ESG and execute data-product requests through an ESG Gateway. A single ESG Gateway serves data requests to many associated ESG nodes: for example, more than 20 institutions are expected to set

up ESG nodes hosting data as part of the IPCC Fifth Assessment. Personnel at local institutions will operate ESG nodes.

4.3. ESG-CET rollout plans

To reach the project goals, a distributed testbed based on the new architecture is scheduled to be in place by late 2008 to early 2009. The minimum testbed requirements include secure and reliable access to federated metadata, federated portals, a minimal set of server-side analysis functions, the ability to distribute data aggregation, publish data, and deliver data. Five U.S. institutions will participate as part of the initial ESG testbed: LLNL (ESG gateway, ESG node); NCAR (ESG gateway, ESG node); ORNL (ESG gateway, ESG node); LANL (ESG node); and LBNL (ESG node). In preparation for a truly global network of climate data services, the testbed will be expanded rapidly to include ESG nodes at several international sites operated by ESG collaborators.

5. Meeting user needs for data discovery and access

In the early days of global climate modeling, data management was *ad hoc* and required a great deal of manual intervention. Datasets were stored wherever the group that generated them could find space, in formats specific to the group or modeling code that generated them. Outsiders who wished to access this data needed to obtain special access to the storage resources and needed to acquire knowledge of the data location, metadata information, and appropriate data translators. Often users became discouraged by either not finding data or finding data that was unintelligible to them. In many instances, data providers would not provide data to outsiders, fearing that the requesting user would not know how to interpret their data correctly. Consequently, a majority of the data sets never received proper scientific scrutiny.

If data was acquired from the data providers, accessing and using it were often painful. Users generally had to develop special code to put the data in a useable format; often the formats were only good for that specific user's needs. Hence many scientists ended up writing their own programs to ingest, manipulate and display their data; and the software produced was generally not user-friendly, reusable, or portable. Often, metadata was kept in files separate from the data, which led to data loss and/or replicated data. Just locating metadata and associated data could be a tedious manual process. With no formal process and far too much done by hand or documented only in the heads of data providers, data management was virtually nonexistent. The communitywide inefficiency and redundancy in these efforts were a significant diversion of time away from the actual science.

As the community recognized the need to facilitate the exchange of data, they began to standardize on storage formats and metadata conventions. The Climate and Forecast (CF) [9] metadata conventions data standards have been well accepted in the community, as has the netCDF file format. The CF conventions define metadata that provide a definitive description of what each variable represents, and the spatial and temporal properties of the data. It enables extraction and comparison of different data sources from different facilities, and facilitates quality assurance.

While these developments helped with the interpretation of data, they did not address the problems of access to data. This was the primary motivation for the initial development of ESG. Taking advantage of the metadata conventions, the first generation of ESG was able to provide search, discovery, and access to a wide range of datasets across key U.S. climate data archives without requiring users to have individual accounts at all of the participating sites.

Arguably, the biggest problem facing climate scientists in the future is not the absence of data but rather its abundance. Current ESG-CET development is making improvements to every aspect of the data management process for both data providers and users. Rather than relying on separate prepublication quality control procedures (which were the responsibility of the providers), the new data publishing system provides built-in quality control on aspects of the metadata, which have been found essential to its effective use within the community. Through appropriate hooks additional project-specific automated quality assurance procedures can be incorporated. ESG-CET and several major data providers (i.e., the CCSM modelling team and the IPCC AR5) are collaborating closely to provide more useful metadata. For data users, the search and discovery capabilities are being expanded greatly. Users must be informed when published data they have downloaded has been removed or changed, and may also register interest in the publication of new datasets meeting certain criteria.

Other important aspects of data management are less visible to the user. In order to provide server-side processing capabilities that allow reduced download volumes, much more detailed metadata is required than for user-level search and discovery. These metadata, related to the structure and contents of individual files within the datasets (in some cases thousands or tens of thousands of files), must be harvested and catalogued by ESG. This is essential if ESG is to offer a reasonable user experience for server-side processing requests (regardless of where the data is actually stored). The increased emphasis on server-side processing also brings out the need to manage provenance information for any derived data.

6. Importance of server-side analysis

To address looming petascale data scenarios effectively, ESG must greatly enhance server-side capabilities for data analysis and visualization and provide for integration with existing community tools. By shifting processing away from the user workstation and into the ESG environment, it will not be necessary for every user to have petascale resources locally to process the data. Delivering the data needed by the user more precisely makes better use of limited wide-area network bandwidth. Both of these factors will contribute to a better overall experience for the user and facilitate more effective community access to large-scale climate data archives. Investigations are under way that will enable the ESG to cache or even preprocess and publish as separate datasets, some of the most popular requests.

6.1. Server-side analysis

The initial focus for server-side analysis is on integrating the Live Access Server (LAS) into the ESG environment. LAS provides a highly flexible web interface for a wide range of visualization, subsetting, analysis, and comparison operations. By default, it acts as a front-end for the Ferret analysis tool, though it can be configured for additional back-ends as well· Fourth-generation scripting languages are extremely popular in the climate community for analysis and visualization of model and observational data. The ESG-CET leadership includes parties responsible for three popular packages: the Climate Data Analysis Tools (CDAT) [10], the NCAR Command Language (NCL) [11] and its Python companion PyNGL, and Ferret [12], which enjoys a particular following in the ocean sciences community.

6.2. Enabling client-side interactions with ESG

The real power of the ESG system will become apparent when ESG is extended with APIs to support client-side applications, based on scripting languages and other tools. This will provide remote access to catalogs, data, and higher-level services for client applications. These applications will serve as drivers for ESG service interface design; the applications will be rewritten so that they can call on ESG services for metadata, catalog information, data, and analysis.

Like ESG itself, these applications have user communities in the thousands, so the combined potential for impact is believed to be quite large. However, such scenarios pose many challenges in terms of workflow definition and management, security, distributed services, and policy issues. The delivery of powerful access and analysis capabilities is a nontrivial task for several reasons:

- Many analysis procedures are written in powerful scripting language systems such as CDAT, NCL, and Ferret. These represent a considerable investment in software development that should be retained; however, also it is necessary to be able to apply the analysis procedures to large aggregated datasets that may be many terabytes in size. Thus it is essential to develop the ability to apply scripting language procedures to large aggregated datasets.

- These aggregated datasets may be located in archival storage, and indeed may be too large to move onto disk in their entirety at one time. Thus it is necessary to develop techniques and tools for managing the movement of data between archival storage and disk.
- As analysis services become more powerful, it is likely that the number of requests for these services will grow. The ESG must be capable of managing such workloads, and must coordinate the use of potentially many computing and storage resources.
- While it is relatively easy to provide a set of commonly used analysis procedures (e.g., "zonal mean"), it is probable that different communities and indeed different individual users will want to define their own procedures. It is important to enable users (at least on some computers) to define and upload their own analysis procedures, in a way that does not compromise the security or reliability of the server on which those procedures run.
- As more analysis procedures become available, it becomes increasingly important that users be able to discover their availability and determine their properties.

The basic analysis service framework, which will be developed in collaboration with the CEDPS SciDAC project, will provide for the wrapping of the analysis function code; the receipt, authorization, and dispatch of client requests; monitoring and control of execution, for purposes of troubleshooting, accounting, audit, management and so forth; and interactions with the back-end resources used to execute analysis functions.

A management interface will allow ESG administrators to monitor service delivery and control quality of service parameters such as the priorities given to different users and tasks. The backend provisioning component allows for the set of resources used for computational tasks to be varied over time, in response to changing load and priorities. The work here will focus on integration with climate data analysis frameworks (CDAT, NCL, and Ferret), the rest of the ESG software suite, and the mechanisms used to manage interactions with archival storage.

6.3. Example future analysis scenario

One example of how these capabilities will be used is illustrated in figure 6. A user searches the ESG-CET portal from a remote site. Required data are found at the distributed data node sites (e.g., NCAR deep storage archive and ORNL fast access disk). Using popular climate analysis tools (such as CDAT, NCL, Ferret), the user regrids the data where they physically reside before transferring the reduced data subset to the PCMDI gateway at LLNL, where further intercomparison diagnostics (such as computing a zonal mean) can be performed on the disparate data sets. Once the diagnostics are complete, the desired end products are delivered to the user's platform.

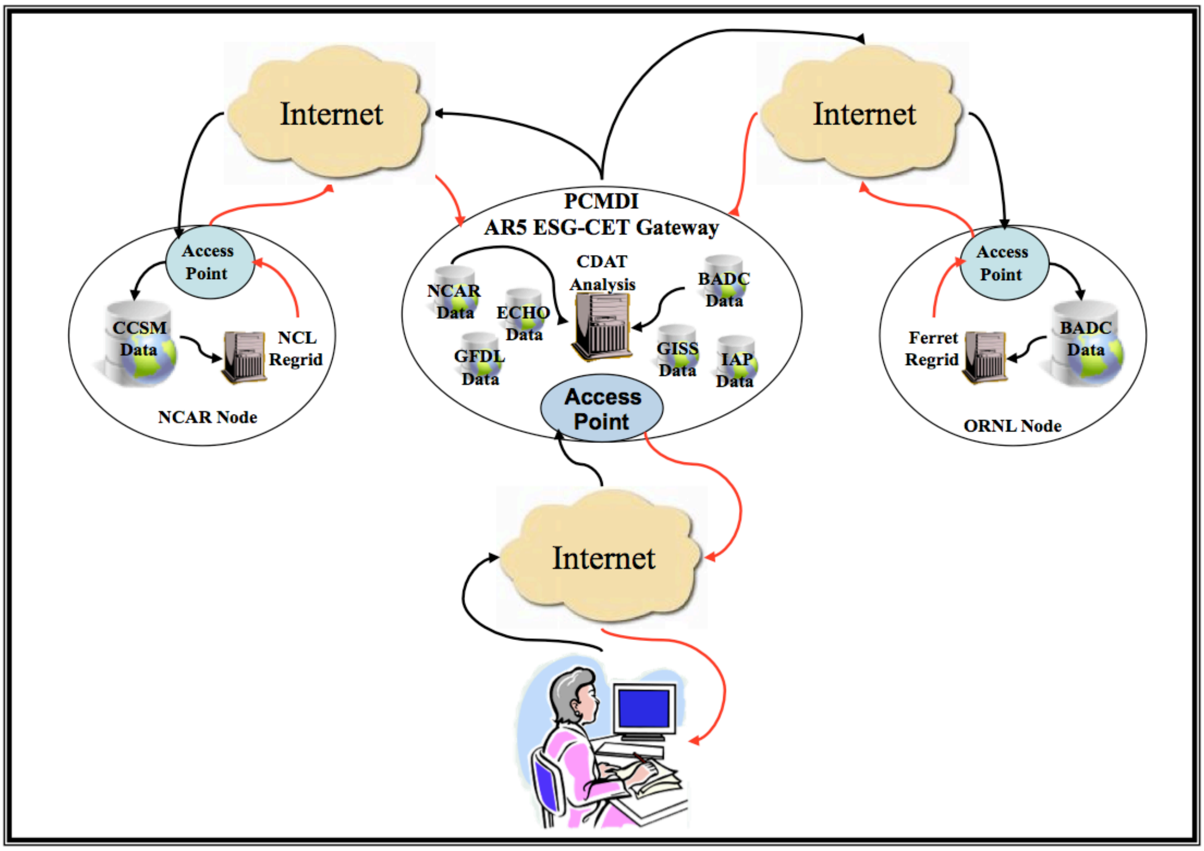

Figure 6. The user accessing and analyzing data via the new ESG-CET architecture.

7. Conclusions

ESG-CET's goal is ambitious—to meet, in a timely manner, the needs of a worldwide community for rapid and reliable access to, and analysis of, many petabytes of climate data. This goal can be attained, but only through a process of careful design and requiring considerable effort. In the process ESG users and designers will make important contributions to understanding how and why climate is changing and how best humanity can mitigate and/or adapt to that change. In order to be useful, the data and software that underpin climate change research must be made freely available to global change researchers worldwide, in a manner that allows convenient access, analysis, evaluation, discussion, intercomparison, and application. This requires a plan for an infrastructure and collaborative environment that links data centers, high-performance computing groups, users, models, data, and resources on a global scale. The creation of such an infrastructure and environment is vital to the success of climate change research. It demands continued investment in data management, analysis software, networking, and collaboration technologies. The methods and software developed in this work are likely to have some relevance to other projects in which large quantities of data must be shared and analyzed by distributed research teams, and the ESG can prove a valuable tool to scientific communities well outside the climate community.

Acknowledgments

This work is supported through the U.S. Department of Energy Office of Science, Offices of Advanced Scientific Computing Research and Biological and Environmental Research, through the SciDAC program. Argonne National Laboratory is managed by Argonne Chicago LLC under Contract DE-AC02-06CH11357. Information Sciences Institute is a research institute of the Viterbi School of Engineering at the University of Southern California. Lawrence Berkeley National Laboratory is managed by the University of California for the U.S. Department of Energy under contract No. DE-AC02-05CH11231. Lawrence Livermore National Laboratory is managed by the University of California for the U.S. Department of Energy under contract No. W-7405-Eng-48. Los Alamos National Security is managed by LLC (LANS) for the U.S. Department of Energy under the contract No. DE-AC52-06NA25396. National Center for Atmospheric Research is managed by the University Corporation for Atmospheric Research under the sponsorship of the National Science Foundation. Oak Ridge National Laboratory is managed by UT-Battelle, LLC for the U.S. Dept. of Energy under contract DE-AC-05-00OR22725. Pacific Marine Environmental Laboratory is under the National Oceanic and Atmospheric Administration's line office of Ocean and Atmosphere Research, lies within the U.S. Department of Commerce.

The ESG-CET executive committee consists of David Bernholdt, ORNL; Ian Foster, ANL; Don Middleton, NCAR; and Dean Williams, LLNL.

References

[1] McAliney M 1993 *Arguments for Land Conservation: Documentation and Information Sources for Land Resources Protection* (Trust for Public Land: Sacramento, CA)

[2] Intergovernmental Panel on Climate Change (IPCC) 4th Assessment Report 2007 http://www.ipcc.ch/ipccreports/ar4-wg1.htm

[3] Williams D N, Bernholdt D E, Foster I T and Middleton D E 2007 The Earth System Grid Center for Enabling Technologies: Enabling community access to petascale climate datasets *CTWatch) Quarterly* **3**(4)

[4] Bernholdt D, Bharathi S, Brown D, Chanchio K, Chen M, Chervenak A, Cinquini L, Drach B, Foster I, Fox P, Garcia J, Kesselman C, Markel R, Middleton D, Nefedova V, Pouchard L, Shoshani A, Sim A, Strand G and Williams D 2005 The Earth System Grid: supporting the next generation of climate modeling research *Proc. IEEE* **93**(3):485-495

[5] ESG. The Earth System Grid - turning climate model datasets into community resources 2004-2008 http://www.earthsystemgrid.org

[6] CCSM Carbon LAnd Model intercomparison Project (C-LAMP) 2007 http://www.climatemodeling.org/c-lamp/

[7] The Parallel Climate Model 2007 http://www.cgd.ucar.edu/pcm/

[8] The Parallel Ocean Program (POP) ocean circulation model 2007 http://climate.lanl.gov/Models/POP/

[9] CF. NetCDF climate and forecast metadata convention 2005-2008 http://cf-pcmdi.llnl.gov

[10] CDAT 2008 The Climate Data Analysis Tools (CDAT) http://www-pcmid.llnl.gov/software-portal/cdat

[11] NCL 2008 The NCAR Command Language (NCL) http://ngwww.ucar.edu/ncl/index.html

[12] Ferret: an analysis tool for Gridded and non-Gridded data 2008 http://ferret.wrc.noaa.gov/Ferret

[13] Fox P, Garcia J and West P 2007 OPeNDAP for the Earth System Grid *Data Science J.* (in preparation)

[14] Foster I, Nefefobva V, Ahsant M, Ananthakrishnan R, Liming L, Madduri R, Mulmo O, Pearlman L and Siebenlist F 2006 Streamlining Grid operations: definition and deployment of a portal-based user registration service *J. Grid Computing* **4**(2):135-144

[16] Shoshani A, Sim A and Gu J 2003 Storage resource managers: essential components for the Grid In Nabrzyski J, Schopf J M and Weglarz J eds. *Grid Resource Management: State of the Art and Future Trends* (Kluwer Academic Publishers)