The PLANET at Their Fingertips: CLIMATE Modeling Data Heats Up

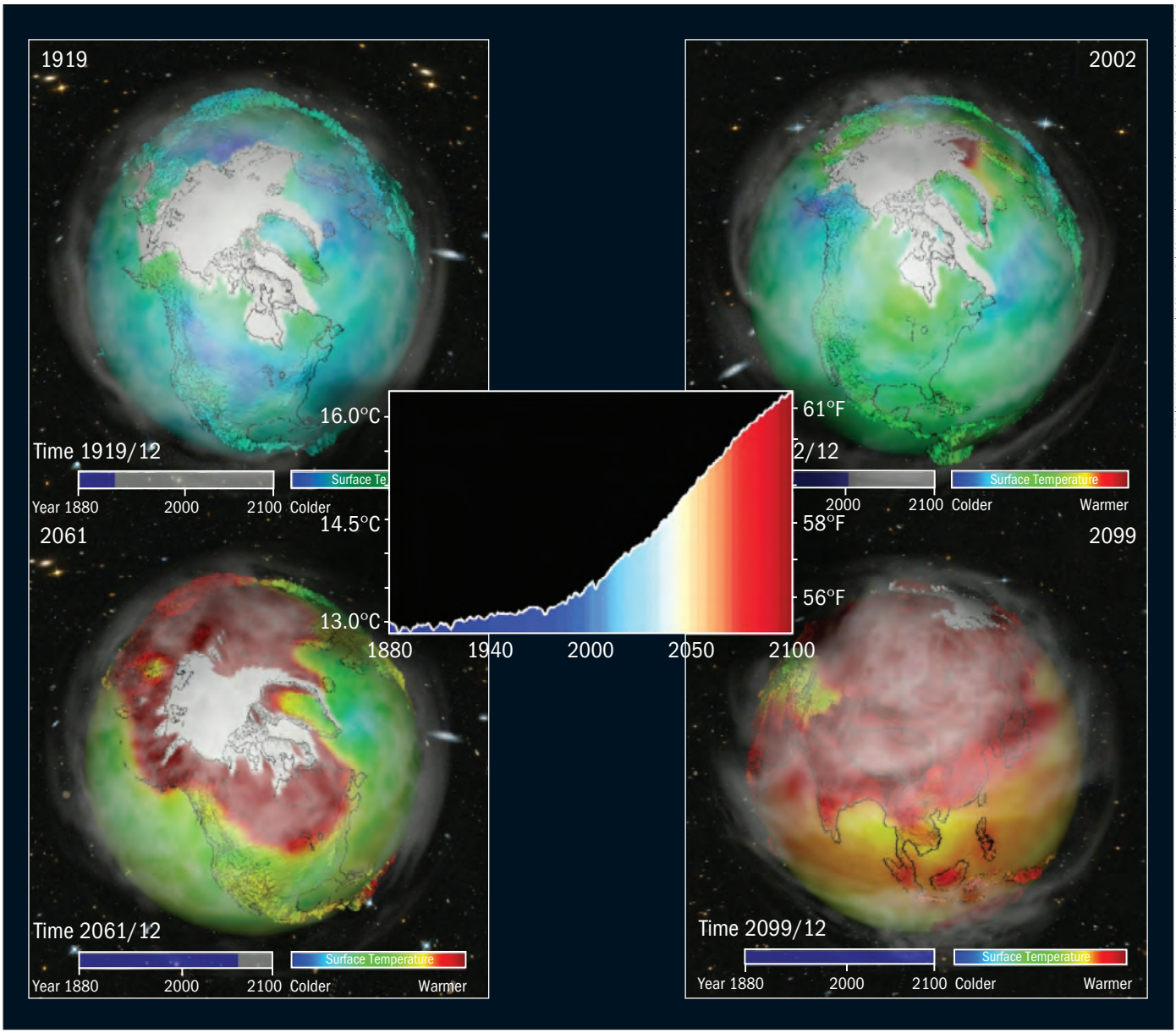
In one of the largest-ever worldwide collaborative efforts in computational science, climate scientists participated in the Nobel Prize-winning Intergovernmental Panel on Climate Change (IPCC). Their contribution was the completion and analysis of hundreds of massively-scaled coupled climate simulations, making the virtual Earth into a planet-sized laboratory experiment. The simulations produced an avalanche of data—scores of terabytes in almost a hundred thousand files. Housing, managing, searching, and disseminating data on this scale was critical to the effort and had never before been attempted. The SciDAC-2 Earth System Grid Center for Enabling Technologies devised a revolutionary technology to accomplish this data facilitation and has since continued creating a worldwide repository and access system that handles data on a scale that dwarfs even the IPCC effort.

The Earth's climate is changing due to fossil fuel use and other human activities. The surface temperature of the Earth has risen 1°C during the last 100 years and is expected to increase another 2–4°C over the next 100 years as atmospheric greenhouse gas levels continue to rise (figure 1). The consequences of such warming would be extreme: because temperature increases are disproportionately larger at high latitudes (figure 2, p44), rapid melting of Arctic sea ice and the large continental ice sheets on Greenland and Antarctica would cause a dramatic rise in sea levels. The effect would be flooding of vast coastal and low-lying regions worldwide, inundating major population centers, and forcing large-scale migrations of huge populations to continental interiors that are ill suited for rapid urbanization. Other predicted consequences include altered global precipitation patterns, with pervasive drought in some regions and flooding rainstorms in others.

Humans are changing the climate; the scientists must determine and explain how.

Roles of the Climate Scientist

Modern climate scientists play a crucial role in ascertaining the validity of climate predictions by performing three vital and interrelated tasks. First, they identify real trends in climate change and explain the nature of those trends. For example, they might explain which trends are natural cycles of Earth history that may reverse themselves in time, and which are permanent. A second role for the scientists has become both controversial and highly politicized:



D. Bader, C. Doutriaux, R. Gaunt, B. Santer, K. Taylor, B. Whitlock, and D. N. Williams, LLNL

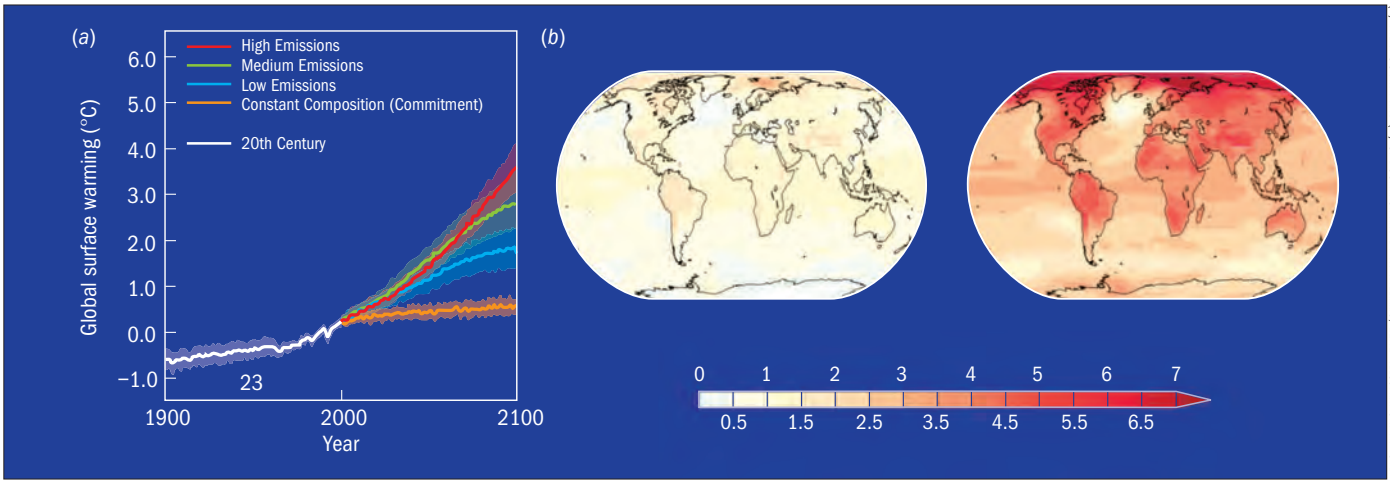
Figure 1. *The surface of the Earth is warming. Four maps of the northern hemisphere demonstrate the trend, with average surface temperature for: top left, 1919; top right, 2002; bottom left, 2061; and bottom right, 2099. Center, the global average, as a function of time, is shown. The computations were based on observed temperature (pre-2000) and a climate model assuming a continued high rate of greenhouse gas emission.*

to determine what climate changes have anthropogenic causes. Humans are changing the climate; the scientists must determine and explain how. The third role is to analyze and explain the efficacy of proposed methods of adapting to or mitigating the effects of such changes. Adaptation describes things man can do to cope with changes, such as constructing seawalls to protect low-lying areas from rising sea levels. Mitigation describes things humans can do to avoid, minimize, slow, or even reverse climate changes (for example, dispersing aerosols into the atmosphere to reflect solar radiation, thereby cooling the planet). Many schemes will be proposed; the climate scientists will be called upon to predict and explain their effects.

Modeling the Earth's Climate: A Laboratory the Size Of the Planet

Climate changes in the recent past can be deduced from recorded observations of temperature, precipitation, and even wind. Going further back into the past, similar inferences can be made from ice cores, lake sediments, tree rings, and other indicators.

Until recently, future changes to the climate could only be hypothesized employing small computational tools. As computing power has grown, large, global-scale climate modeling has expanded. Previous limits on computing capability and the size of datasets had constrained the resolution and scope of the effective modeling capability. With the ongoing revolution in high-capacity, high-performance



(a) The IPCC AR4 and (b) C. Doutriaux and D. N. Williams, LLNL. Illustration: A. Tovey

Figure 2. *(a) Two centuries of global surface temperature are shown. Historical observations from 1900–2000, with an average plotted in white. On the right are four simulated scenarios (plotted within a confidence interval), each representing a different rate of greenhouse gas emissions. (b) Climate models predict that the greatest rise in temperature will occur in the high latitudes. Shown is the average increase (deg K) in surface temperature as predicted using the CMIP3 archive, relative to the pre-industrial period. Left, early 21st century; right, late 21st century.*

computing (HPC) that situation has changed and continues to evolve rapidly. HPC machines now have sufficient power to build much more accurate and detailed simulations of long-term climate evolution using computer models of the climate system (figure 3). Such models have become the primary tools scientists use for predicting the future climate of Earth. These models are also used to study the causes of climate change. Scientists can vary the input parameters, alter the virtual climate system, and perform computational "experiments" to determine the effect on climate of individual physical processes; they can examine in detail the effects of variations in the biogeochemistry of the system, fluctuations in regional water budgets, or changes to the ecosystem. These experiments can also be run over long periods of virtual time (hundreds or thousands of years). Such experimentation cannot be performed in nature with its large and open systems and real-time constraints.

Early climate models were generally single-component, including only the atmosphere or the ocean. As computing has become more powerful, models are growing dramatically more complex. Modern climate models attempt to simulate the entire climate picture, including interactions among the four major climate systems: atmosphere, ocean, land surface, and sea ice. Each of these systems is itself a sophisticated representation of the governing physical processes. A model of this sort, coupling the effects of two, three, or all four of these Earth systems, is known as a coupled general-circulation model (GCM). GCMs are the workhorses of modern climate modeling, and more than a score of such high-fidelity systems are available to the modern climate scientist. Two examples are the Geophysical Fluid Dynamics Laboratory at the National Oceanic and Atmospheric Administration and the Community Climate System Model (CCSM) operating at the National Center for Atmospheric Research (NCAR).

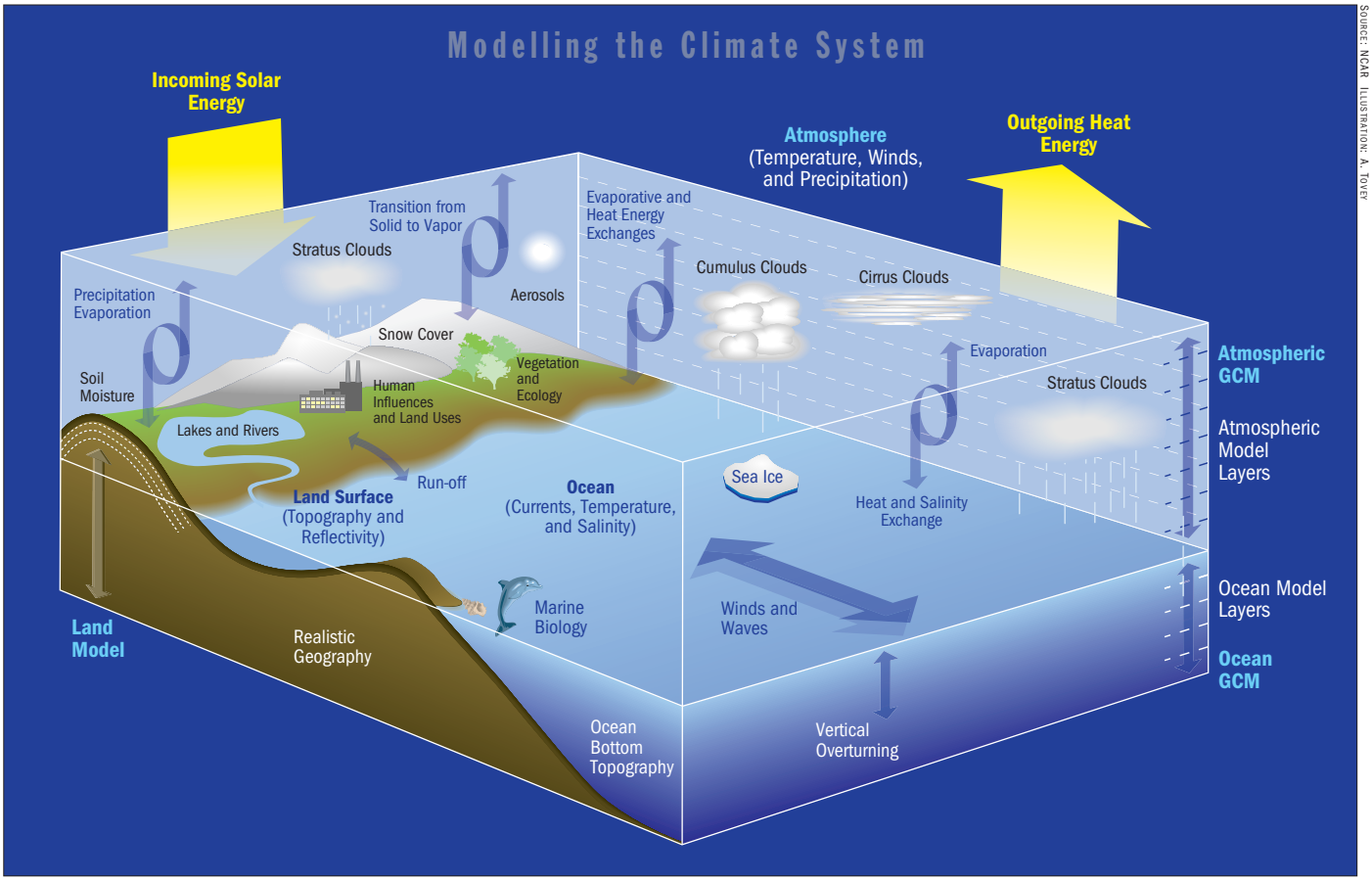
Usage of CCSM is widespread in the university community and at some national laboratories. NCAR and its Department of Energy (DOE) partners developed CCSM as a coupled model for the four major climate systems. The coupling was designed to avoid "flux adjustments," a common technique of introducing artificial numerical effects to compensate for physics that is too complicated to be handled directly. A novel structure enables interested groups in the scientific community to participate in development and guidance of CCSM so that it is very much a true "community climate model."

The explosive growth of supercomputer development in the 1990s brought accessibility to these GCMs within the reach of even very small, independent climate research groups. This accessibility has opened the field of climate modeling to a vastly increased number of practitioners, leading to greater concentration of effort and brainpower for climate analysis. The plethora of models, each developed independently over many years and having its own conventions, formats, input/output (I/O), data handling, and independent analysis tools, has made the task of comparing and evaluating results extremely difficult.

Models have become the primary tools scientists use for predicting the future climate of Earth. These models are also used to study the causes of climate change.

Data, Data, Everywhere…

Indeed, while HPC has been following Moore's Law and doubling computational power every two years or so, advances in hardware and software for the management, manipulation, storage, and retrieval of large datasets, while dramatic, have not kept pace. This discrepancy is keenly felt in the climate



SOURCE: NCAR ILLUSTRATION: A. TOVEY

Figure 3. *The global climate system comprises many interrelated processes and is described in the most sophisticated models as numerous interactive systems. Each can be modeled alone, but an accurate depiction of the climate necessarily employs them together.*

modeling community. Global climate models are often run using a surface spacing of 100–200 km over the entire two-dimensional surface of the Earth and may include applying a grid to the third dimension into the ocean and/or atmosphere. At every gridpoint a number of quantities (often as many as a hundred or more) must be calculated at each time step, and the models often simulate decades or even centuries of time. Such models produce extraordinarily large datasets. There are some two dozen major GCMs, and many, many single-component models, all representing the efforts of both large groups and thousands of individual contributors. These models are generating data at a collective rate that is truly staggering: some models generate multiple terabytes (a terabyte is 10^{12}, or a trillion, bytes of data) with each run. Even regional models, covering far less than the entire globe, may generate huge datasets, because they are often run at a finer grid spacing.

All these data must be stored, retrieved, processed, explored, visualized, and shared. The climate modeling community is extremely collaborative in nature. Even the largest simulations take long to produce and generate so much data to be analyzed that it is impractical to try to comprehend or interpret it alone. Hence, knowledge in the field advances through the sharing of models and data, with many groups wishing to examine the output from major simulation runs or to repeat experiments to compare different models, codes, or climate phenomena.

Complications in working with the data make interpreting the output from GCMs difficult. The data for runs of a given model exist in extremely large datasets, perhaps many terabytes in size, and are stored at diverse centers around the world, each with its own access requirements and protocol. For example, large datasets from the Geophysical Fluid Dynamics Laboratory are stored at NCAR and made available through a specialized data portal. Another complication is that coupled climate modeling may require the assimilation of real-world data to bring reality to a simulation. Surface temperatures, for example, from many thousands of weather stations are available; however, manipulation of these data into a useable format and incorporating it into a model is arduous.

Dissemination of and access to these enormous datasets pose a major obstacle to rapid and effective collaborative science. A group wishing to use data generated by others may encounter great

Modern climate models attempt to simulate the entire climate picture, including interactions among the four major climate systems: atmosphere, ocean, land surface, and sea ice.

Building the Earth System Grid from Scratch

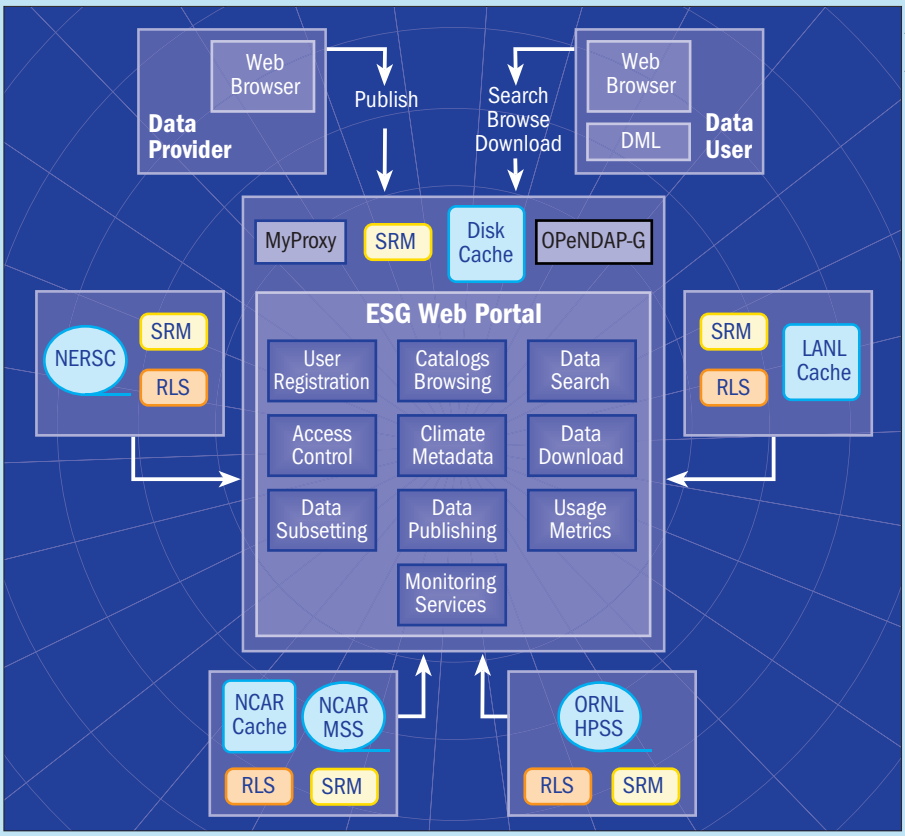
With the funding in 2000 by DOE's Next Generation Internet Program of a project entitled "Prototyping an Earth System Grid," the seeds of ESG were sown. For the prototype, the project team developed data grid technologies for managing the movement and replication of large datasets and applied these technologies in a practical setting: an ESG-enabled data browser based on the Climate Data Analysis Tools developed by the PCMDI at Lawrence Livermore National Laboratory.

The fledgling ESG began to grow and develop, transitioning from seed funding to continued support from both the DOE's Office of Biological and Environmental Research and the Office of Advanced Scientific Computing Research.

Having demonstrated a proof of the ESG concept, the project obtained support from SciDAC in 2002, becoming the Earth System Grid II project under the title "Turning Climate Datasets into Community Resources." Then, ESG stopped being a prototype or a proof-of-concept and became a reality. These efforts focused on developing metadata technologies (tools that facilitate standardizing the data formats, descriptions, cataloging, tracking data location, and so on); security technologies such as web-based user registration, authentication, and community authorization; data transport technologies (tools for high-performance access, robust multiple file transport, integration with mass storage systems, and support for dataset aggregation and division); and web portal technologies to provide interactive access to climate data holdings.

The first ESG II architecture (figure 4) integrated a wide-range of grid and standard IT technologies with the goal of facilitating data publishing by climate data providers and data access by the climate user community.

Since 2004, when the first production version was released, ESG has become an integral part of the way climate scientists carry out their work. Currently supporting more than 9,000 registered global users and managing more than 200 terabytes of data, ESG handles between 400 and 600 gigabytes of data downloads each day (figure 7, p48).



LLNL, ORNL, AND THE UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH. ILLUSTRATION: A. TOVEY

Figure 4. *The architecture of ESG II included both grid technology and standard information technology tools, facilitating browser-based publication, search, and downloading of remote data through a web portal.*

Further Information
PCMDI
http://cdat.sf.net

difficulty in locating the data, searching through a collection to find the specific data desired, determining the format, obtaining the data themselves (until rather recently, transferring multiple-terabyte datasets was only feasible by mailing computer disks, and many disks were required for larger output files), and translating the data into the format used by the new group.

The Earth System Grid

The Earth System Grid (ESG) is designed to address the management and use of extremely large, diverse datasets. Begun with DOE support in 2000 (sidebar "Building the Earth System Grid from Scratch"), ESG has since become a SciDAC-2 Center for Enabling Technologies (CET). The ESG-CET aims to build a "science gateway" to climate resources. Open to everyone, the ESG houses and provides access to and management of climate data, information, models, analysis tools, and visualization tools. ESG is built on the concept of grid computing—an approach that pools hardware and software assets of many distinct networks and locations into a centralized virtual system and provides access to distributed resources. The resulting virtual system is more powerful and has much greater capacity than

IPCC AR4

Figure 5. *Each succeeding IPCC assessment has found increasing evidence for anthropogenic causes for global warming. The latest assessment describes a human cause as "very likely."*

any of the individual systems or networks that it comprises. The goals of the ESG project are:

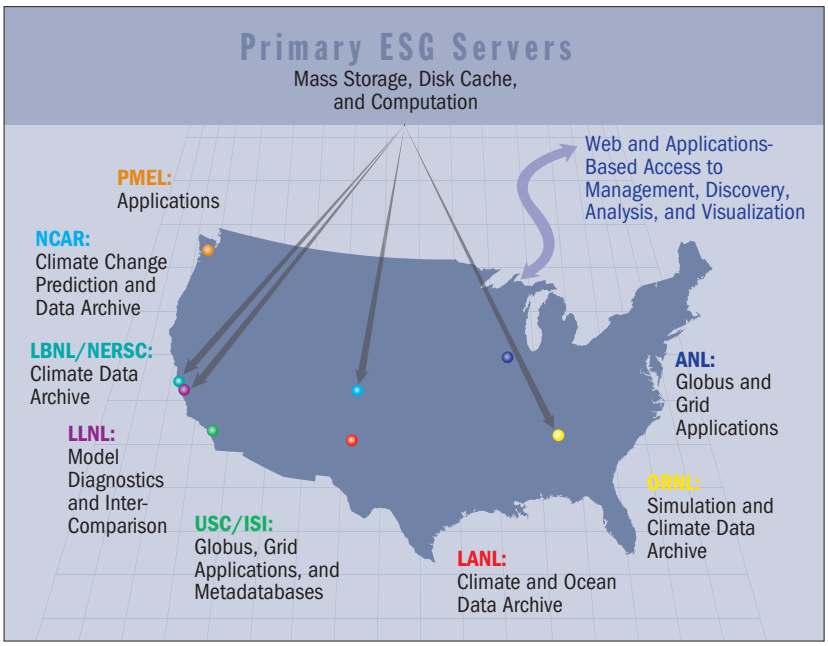
- To develop grid technology that specifically enhances usability of climate data
- To provide tools for distributed data access and data movement to support disbursed national and international climate projects
- To provide a universal and secure web-based data access portal for broad-based multi-model data collections
- To provide a wide-range of grid-enabled climate data analysis tools and diagnostic methods to international climate centers and U.S. government agencies

To achieve this, ESG-CET works to integrate distributed data and computers, high-bandwidth wide-area networks, and remote computing using climate data analysis tools in a highly collaborative problem-solving environment.

ESG and IPCC

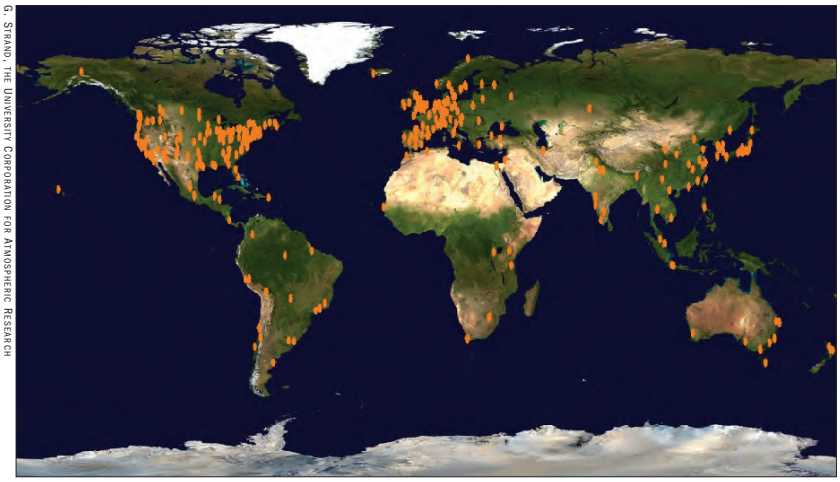
Nowhere is the need for grid-based access and tools better illustrated than in the work of IPCC, an organization of climate science research groups chartered in 1988 by the United Nations Environment Programme and the World Meteorological Organization. IPCC is charged with providing worldwide decision-makers with an objective source of information about climate change. IPCC does no research; its role is to "assess on a comprehensive, objective, open and transparent basis the latest scientific, technical, and socio-economic literature produced worldwide relevant to the understanding of the risk of human-induced climate change, its observed and projected impacts and options for adaptation and mitigation." There have been four IPCC assessments, one each in 1990, 1995, 2001, and 2007. Each assessment has edged closer to asserting anthropogenic causes for the global warming trend. The Fourth Assessment in 2007 describes a human cause as "very likely" (figure 5). IPCC was awarded the Nobel Peace Prize in 2007 (jointly with former U.S. Vice President Al Gore) for the role it has played in focusing the world's attention on the problem of climate change, underscoring the importance of this scientific issue to society.

To understand the scope of the data problem it represents, it is useful to understand what the IPCC assessment entails. IPCC developed several scenarios, each assuming specific rates of energy use and related emission of greenhouse gases. Research groups throughout the international climate community were tasked with running several future climate simulations for each of these scenarios. Each was a global simulation using a multi-component coupled GCM with a surface spacing on the order of 150 km (at the equator) and carrying the simulation through many years



SOURCE: NCAR ILLUSTRATION: A. TOVEY

Figure 6. *The ESG-CET consortium is composed of seven laboratories—Argonne, Los Alamos, Lawrence Berkeley, Lawrence Livermore, and Oak Ridge national laboratories; the National Center for Atmospheric Research; and the Pacific Marine Environmental Laboratory—and one university, the Information Sciences Institute at the University of Southern California.*



G. STRAND, THE UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

Figure 7. *ESG makes data a community resource, accessible worldwide. This image shows the institutions that accessed ESG data during 2007.*

of simulated time. Each simulation run generated huge volumes of data to be housed, managed, analyzed, and shared. A given IPCC assessment of future climate requires the analysis of many such datasets—a huge body of data generated from many different models examining many different future scenarios. The role of ESG is to provide scientists worldwide with the ability and tools to access, examine, and interpret these data.

For the Fourth Assessment, one of the largest and most important collections of models and data was housed at the Program for Climate Model Diagnosis and Intercomparison (PCMDI), located at Lawrence Livermore National Laboratory. PCMDI serves as one of ESG's three primary storage and access locations (or "portal gateways"), and houses data and models contributed to IPCC by numerous research groups, part of which is known as the Coupled Model Intercomparison Project Phase 3 (CMIP3).

CMIP3 had become the largest international global coupled climate model experiment and multi-model analysis effort ever attempted. Ultimately, a total of 17 modeling groups from 12 countries participated, employing 24 different models. More than 35 terabytes of model data were collected and housed in the archive; it is available to the climate community through ESG.

How ESG Makes It Work

The data produced in the Fourth Assessment were first transferred to, and then distributed from, a central database archive maintained by PCMDI. Future coupled climate simulations will produce datasets so large that this "centralized" approach will be impracticable, and it quickly became apparent that the climate community needed a more complex data distribution architecture enabling simultaneous participation of multiple data centers. To accommodate this new paradigm, ESG-CET began developing and implementing grid technologies employing the Internet to link climate centers and users across the globe with models, data, and other resources. The ESG-CET consortium comprises seven laboratories and one university (figure 6) that today manage some 250 terabytes of data for seven different climate-modeling efforts. The most important data collection is the CMIP3 (the data used in the Fourth Assessment), located at PCMDI, which is one of the portals. Among the other six model data archives managed by ESG-CET are the climate system archive of the CCSM (the largest data collection on ESG), as well as the archive of the North American Regional Climate Change Assessment Program, an international program with U.S., Canadian, and European participation. More than 9,000 users currently make use of ESG capabilities (figure 7).

ESG-CET developed a vision of a virtual collaborative environment providing remote users with the sense of "being there" with the data and computational resources required to perform their work. To this end, ESG employs a wide range of grid technologies to build an interface to the large and distributed data it manages, so that scientists and other users can easily download, combine, and analyze model data to develop projections of future climate and its impacts (sidebar "Under the Hood: Grid Software Makes ESG Function"). Each of the eight members of the consortium is a node on the "grid" and functions as a primary server;

Under the Hood: Grid Software Makes ESG Function

To its worldwide community of users, ESG looks and feels rather like an Internet shopping site (figure 8). ESG users, like Amazon.com users, search for the products they want, place those products in their shopping cart, and, when they are done, have their selections delivered. The big difference, of course, is that the products ESG offers are not books or electronics but hundreds of terabytes of unique and valuable data, generated on some of the world's fastest computers and entailing massive storage and computing systems to maintain and analyze. To harness these resources, ESG operates a distributed collection of services, constructed with sophisticated grid software.

ESG grid services and software make it possible, for example, for ESG users to register with an ESG portal, log on once, and then search for or retrieve data that may be located in one of several distributed storage servers. Grid software makes it possible for users to download data at high speed from ESG servers over wide area network links. Other grid software is used within ESG to monitor the health and status of the various services that constitute the ESG system. As ESG expands to a federated system, grid software will also be used to maintain replicated metadata catalogs and to federate authentication services.

Among the many challenges faced by ESG is security, which poses difficulties at two levels. First, ESG nodes must adhere to institutional computer and network security requirements and policies, and interface with shared resources such as mass storage systems that are not operated or owned by ESG. Second, ESG must correctly and securely implement policies required by its users: for example, access control to datasets, which in turn demands authentication and authorization.

The grid software used within ESG comes from a number of sources. One is the SciDAC Center for Enabling Distributed Petascale Science (CEDPS), which provides mechanisms such as GridFTP and the Data Replication Service for high-speed, reliable, and secure data movement and replication, among other things. Another is the SciDAC Scientific Data Management (SDM) Center, which provides the Storage Resource Manager software used to manage access to tertiary storage. Many software components used within ESG are distributed within the Globus Toolkit software distribution.

THE UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

Figure 8. *Visiting the ESG portal is not unlike an Internet shopping trip, where scientists can browse the climate model collections, locate the data they want, and order them with a few clicks of the mouse. Behind the curtain, grid technology software is handling a host of necessary functions for locating, packaging, and shipping the data, as well as handling authentication, authorization, accounting, and security over a grid-based system whose actual machines may be located all over the world.*

Further Information

CEDPS
http://www.cedps.net

SDM Center
http://sdmcenter.lbl.gov/

Globus Toolkit
http://www.globus.org/

that is, each houses some of the datasets or hosts grid software centers. Three of these nodes are designated as portal gateways (or simply portals; figure 9, p50): PCMDI hosts the CMIP3 collection; NCAR hosts the CCSM collections; and Oak Ridge National Laboratory (ORNL), a DOE Leadership Computing Facility, hosts the Climate Science Computational End Station, an ESG portal that is a main access point to the system.

Here is how the nodes work. Users submit requests for data to the portal. The portal determines to which of the data nodes the request should be sent. It forwards the request to that data node, and it directs the resulting data stream back to the user. Under this configuration, the portals facilitate access to the rest of the grid in addition to housing datasets and climate modeling tools of their own. The portal also provides a central location for authentication, authorization, and accounting services. This last function is crucial—because each model run can require hundreds of hours on extreme supercomputers and generate datasets many terabytes in

SOURCE: R. GAUNT, LLNL ILLUSTRATION: A. TOVEY

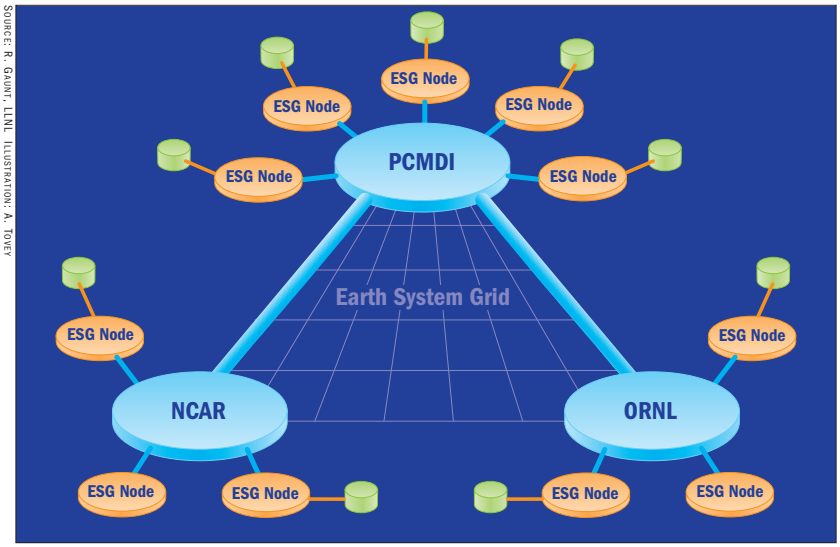


Figure 9. *The three major ESG-CET gateways: PCMDI, NCAR, and ORNL. ESG nodes with data storage devices are federated throughout the system using grid technology. In the future, ESG will allow users to access all data from the grid regardless of which gateway is used.*

SOURCE: G. STRAND, THE UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH, AND B. DRACH, LLNL ILLUSTRATION: A. TOVEY

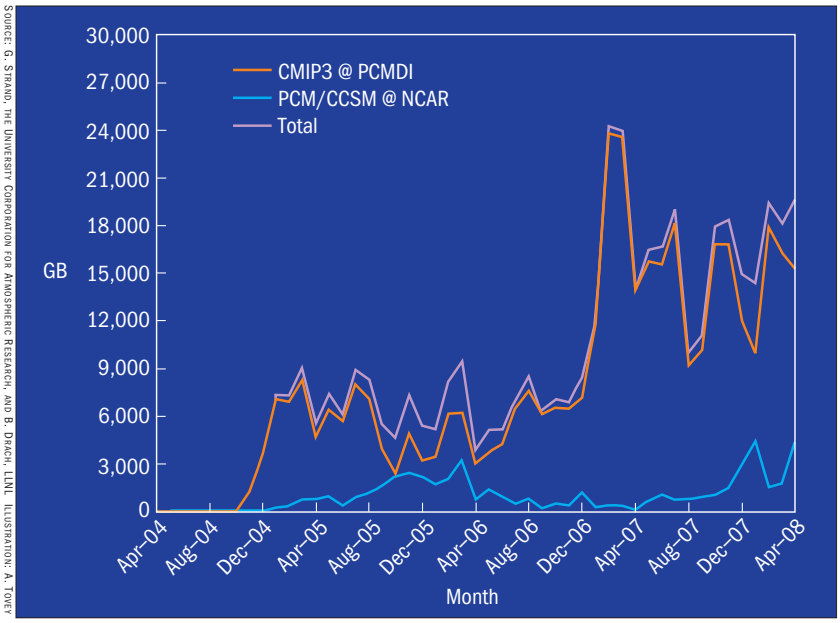


Figure 10. *Monthly volume of data downloaded (in gigabytes) from the two main portals on ESG, from April 2004 through April 2008. The blue line shows download rates for the collections through the NCAR portal, while the orange line indicates data movement through the PCMDI portal, which houses the CMIP3 data used in the IPCC Fourth Assessment.*

size, the cost of producing (and therefore replacing) the data is considerable, and protection of the datasets is a major ESG function.

ESG must provide several basic functions, too. The three most obvious are mechanisms for:

- Placing model data into the collections (known as publishing the model)
- Exploring the collections to ascertain what data are present or where specific data reside
- Retrieving data from the collections

Considering the enormous size and dispersed nature of the collection as a whole, as well as the huge size of individual datasets, all of these tasks are difficult.

Fundamental to the three tasks is the ability to move data from place to place on the grid. Initially, data movement relied on the tried-and-true File Transfer Protocol (FTP), but this simple method is very limiting, and ESG has included or inspired the development of several newer, more powerful methods. The ESG data portal, the Live Access Server, and the open-source project for a Network Data Access Protocol Grid are two newer methods. Online files are served via a lightweight authorized HTTP file server, while files in "deep storage" (that is, not stored online) are requested and served via the Storage Resource Manager (SRM), which retrieves them from the archive and transfers them to a central disk cache, where they are made available online. In some cases, individual files or groups of files are too large to be transferred via the ESG portal. For such cases, ESG developed a DataMover tool to effect robust large-scale data movement by interacting with the SRM and replicating thousands of files between specified mass storage systems. A client-side version of this tool, DataMover-Lite, automates multi-file data transfers from the SRM to the client's file system.

ESG also developed tools enabling scientists to publish their climate model data, thus making their data available to other scientists. Again, this is done through the ESG portal. As part of publishing into an archive, model data providers must register appropriate metadata—that is, a description of the data, how they were generated, and their format. Hence, data providers can manage all information related to generating, defining, archiving, and retrieving model simulation runs. Providers may also restrict access to their data, according to a variety of criteria.

Model providers wishing to add their data to the CMIP3 collection at PCMDI must, before transmitting, transform the model output into a format specified as netCDF (network Common Data Form) and must use the Climate and Forecast (CF) metadata convention. To facilitate this conversion, PCMDI provides participating modeling centers with the Climate Model Output Rewriter (known as CMOR and pronounced "Seymour") software, which produces CF-compliant netCDF files.

There are many other tools and grid software packages necessary for the operation of ESG, including codes and frameworks for database management, computer security, system monitoring, and a host of analytical and visualization tools, all technologies that are at the core of ESG functionality.

The value of ESG for publishing and providing access to model data is evident from the statistics

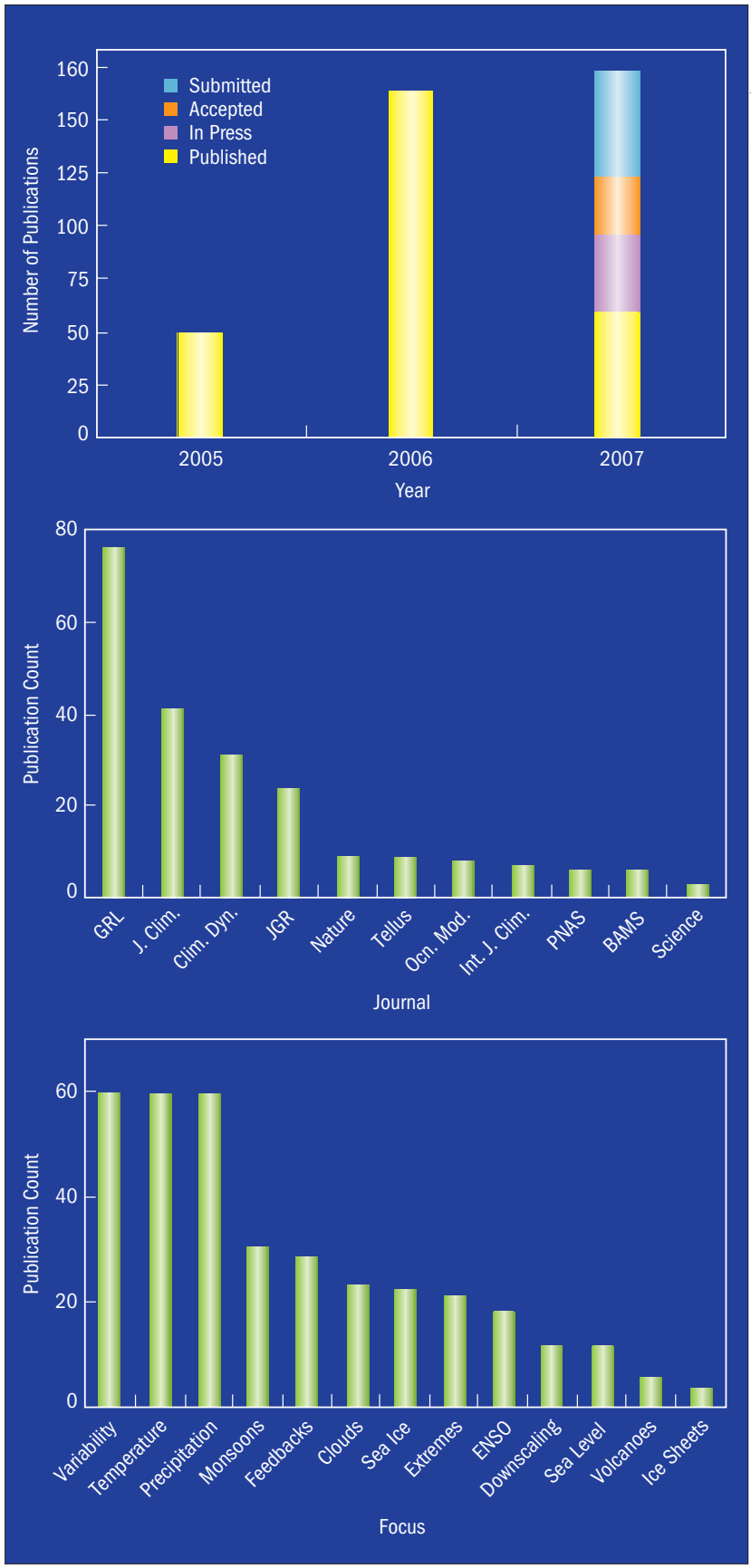
of the CMIP3 collection. Since December 2004, the IPCC model runs published in the archive have totaled some 35 terabytes in just over 78,000 individual files. Access to the collection has been enormous: more than 1,900 scientific subprojects have registered to obtain data for analysis, and more than 1.25 million files, totaling over 425 terabytes of data, have been downloaded. The daily volume of downloaded data averages more than 500 gigabytes (figure 10).

The access ESG provides has translated into an impressive volume of new research. In 2005, just after the data were made available, 50 peer-reviewed journal articles were published. Since then, an average of more than 150 papers per year, based on ESG-archived CMIP3 data, have appeared. The total now stands at 400 papers and continues to grow. The publications appear in a broad range of journals and span many topics central to climatology (figure 11). Many of the analyses in these publications contributed directly to the IPCC Fourth Assessment report and were recognized in the Nobel Committee's statement that the award was "for efforts to build up and disseminate greater knowledge about man-made climate change, and to lay the foundations for the measures that are needed to counteract such change." The IPCC Fourth Assessment demonstrated what can be effectively accomplished with a grid-enabled archive.

Serving a Broader Climate Community

ESG-CET supports numerous major initiatives throughout the broader climate community. As representative examples, NCAR employed ESG to publish data from the CCSM and its predecessor, the Parallel Climate Model. These two major data collections became the first to enjoy on-demand access to distributed collections using grid technology and established ESG-CET as a leader in developing such technologies. The ESG portal at NCAR provides access to approximately 160 terabytes of data stored in the CCSM, Parallel Climate Model, Parallel Ocean Program, Community Atmospheric Model, Community Land Model, and Community Sea Ice Model collections. More than 8,000 users, including climate scientists, analysts, educators, private industry, and government officials (both domestic and foreign), are registered through the NCAR portal gateway. Collectively, they have downloaded more than 35 terabytes of data.

ESG-CET is working to employ coupled Earth system models, adding sophistication, complexity, and utility beyond that of contemporary GCMs. For example, CCSM, through its Carbon-Land Model Intercomparison Project (C-LAMP), is incorporating biogeochemical models that integrate terrestrial and ocean ecosystems and atmospheric chemical processes to model the effect of



SOURCE: K. TAYLOR, LLNL ILLUSTRATION: A. TOVEY

Figure 11. *The history of publications based on the CMIP3 collections. Top, number of publications, 2005–2007. Center, distribution of publications appearing in high-visibility journals. Bottom, the focus of the journal articles covers topics from variability, temperature, and precipitation to sea level, ice sheets, and volcanoes.*

ESG-CET Tiered Architecture: A Design for the Future

ESG-CET is preparing to accommodate ever-larger climate archives that are expected to come online within the next four years. Specific focus is on preparing for the IPCC's Fifth Assessment Report archive, scheduled for early 2013, readying tools for publishing and processing massive amounts of data produced by the Climate Science Computational End Station ORNL, and supporting a wide range of improved climate model evaluation activities. To date, the ESG story has been told in terabytes (trillions of bytes), but scientists must soon manage, analyze, and visualize datasets counted in petabytes (quadrillions of bytes).

Today, ESG stores data at three primary repositories in the United States, but the climate community is moving toward petabyte-scale datasets, physically located at many sites throughout the world. For example, "core" data for the IPCC Fifth Assessment are expected to be one petabyte stored at PCMDI and another 25 to 50 petabytes housed at 20 or more ESG satellite nodes.

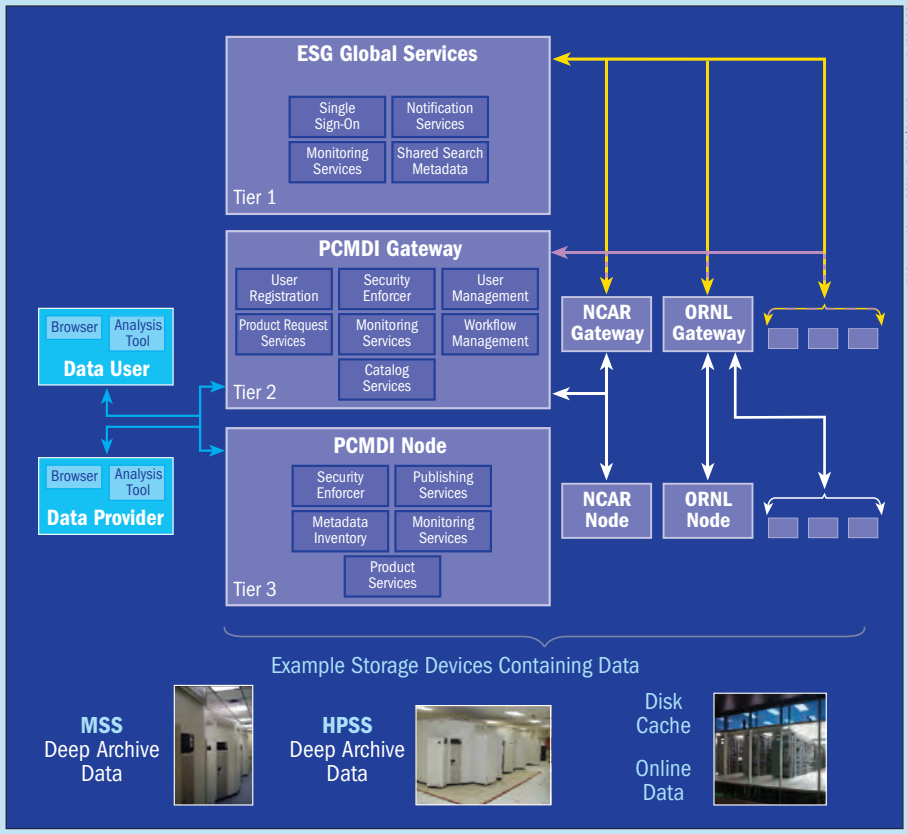
The extremely large sizes of the individual collections will mean that much more of the analysis will need to be performed on the data where they are stored; there will not be resources or bandwidth to constantly transfer data to and from analysis locations. Also, much of the analyses require access to multiple data repositories for comparisons between models or datasets. Hence, by 2011, ESG will be a full data-sharing environment and will provide users with open access to a broad range of data assets, spanning models, satellites, field data, biogeochemistry, ecosystems, and more.

This comprehensive infrastructure will federate many data centers into a virtual data repository, provide a full suite of analysis and data manipulation tools, integrate model and observational data, and provide model inter-comparison metrics, user support, and life-cycle maintenance.

The new architecture is based on three tiers of data services (figure 12):

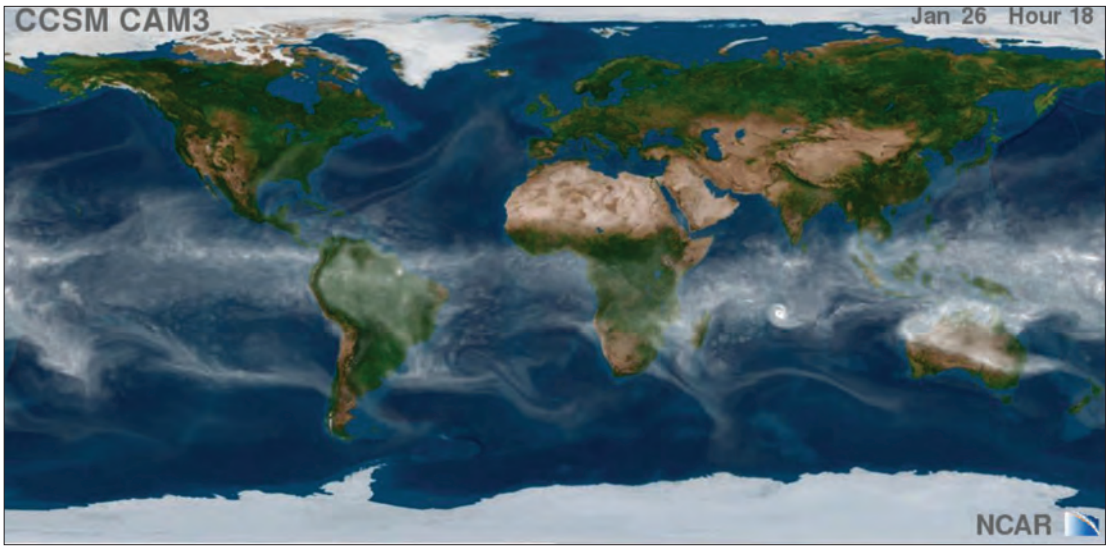
- **Tier 1** Global Metadata Services for Search and Discovery: Comprises a set of services providing shared functionality across the worldwide ESG-CET federation. The exact specifications of the global services have not been finalized but will likely include services for sharing search metadata, exchanging user attributes and resource policies, and overall monitoring of the ESG-CET system. 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 2** Data Gateways as Data-Request Brokers: Comprises a limited number of ESG data gateways that 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, requesting data (including analysis and visualization) products, and orchestrating complex workflows. Gateways will be operated directly by ESG-CET engineering staff.
- **Tier 3** ESG Nodes With Actual Data Holdings and Metadata Accessing Services: Includes the actual data holdings and resides 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. Personnel at local institutions will operate ESG nodes. A single ESG gateway serves data requests to many associated ESG nodes. For example, more than 20 institutions are expected to set up ESG data nodes as part of the IPCC Fifth Assessment.

To be ready for the expected data onslaught, a distributed testbed must be in place by mid-2009. To accomplish this end, five institutions have been selected for the test bed: PCMDI, NCAR, and ORNL will function as ESG gateways (and also as nodes), while Los Alamos and Lawrence Berkeley national laboratories will function as ESG nodes. Strong international interest, particularly in IPCC, will quickly bring expansion of the test bed to include several ESG nodes at international locations, paving the way for a truly global network of climate data services.



Source: D. N. Williams, LLNL. Illustration: A. Tovey

Figure 12. *The ESG Phase 2 tiered architecture showing the three levels of data services represented as Tier 1–Global; Tier 2–Gateway; and Tier 3–Data Node. Three ESG gateways are planned initially, at PCMDI, ORNL, and NCAR. The picture also shows where data users and data providers gain access.*



The University Corporation for Atmospheric Research

Figure 13. *Next-generation climate models are under development, and feature greatly refined resolution. This visualization depicts atmospheric water vapor (white) from a CCSM model computed using sample spacing of 37 km.*

The climate modeling community of tomorrow will have a tremendous need for ESG (or its descendents), and for that reason ESG must be fully functional, lightweight, fast, and accurate enough to do the necessary job.

the carbon cycle. Initially providing an access-controlled environment so that members of the C-LAMP Working Group may publish and exchange data exclusively among themselves, a C-LAMP-specific ESG portal has been established at ORNL. The intention is eventually to integrate the C-LAMP archive into the globally distributed ESG, making these data available to the public.

ESG—A Work in Progress

ESG is by no means a finished product because climate modeling is an ever-developing activity of increasing complexity and sophistication. There are numerous questions that today's climate models cannot answer. For example, modern trends in computational science are toward multiscale processes. In climate modeling, this now means coupling global phenomena to regional phenomena, which is not yet perfected. Beyond that is the fact that processes occurring at local scales, even down to extremely small spatial or temporal scales, may actually play significant roles in getting accurate predictions regionally and much remains to be done in these areas. The IPCC Fourth Assessment was done using global models with sample resolutions between 140 km and 250 km; obviously, a lot of "climate" can take place in the used kilometers between samples. The next IPCC assessment is expected to be accomplished using significantly finer resolution than the last one, and next-generation climate models that run at much higher resolution are currently in development. For example, the visualization shown in figure 13 depicts a CCSM experiment with a sample spacing of 37 km. Additionally, one ongoing problem is that of data assimilation. How can observations, which exist in great abundance and are being collected at breakneck speed throughout the world, be incorporated into the models?

ESG-CET is active in anticipating the needs of the future. Among its activities in this area is the design of next-generation architecture for ESG (sidebar "ESG-CET Tiered Architecture: A Design for the Future"), designated ESG Phase II. This design foresees vastly enlarged data collections in which ESG must make available tens of petabytes (a petabyte is 10^{15} bytes—a quadrillion bytes) of data, housed across a federation of perhaps two dozen data centers throughout the world.

A Promise for Tomorrow

Climate modeling will become increasingly important as societies are forced to adapt to changing climates and to mitigate their effects. Pressure on policy makers to take action in the face of climate changes will only increase with time. It is crucial, if dislocations and human suffering is to be minimized, that policy makers employ an effective, fully informed decision process. For this, they must have information from scientists that is as accurate as possible, which means increasing the complexity, sophistication, and rigor of climate models. The climate modeling community of tomorrow will have a tremendous need for ESG (or its descendents), and for that reason ESG must be fully functional, lightweight, fast, and accurate enough to do the necessary job. ESG-CET exists to ensure that this will be the case. ●

Contributor Dean Williams, LLNL

Acknowledgments This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344